ma414l4.tex
Lecture 4. 2.2.2012

II. STOCHASTIC PROCESSES

1. Conditional expectations.

Suppose that X is a random variable, whose expectation exists (i.e. $E|X| < \infty$, or $X \in L_1$). Then EX, the expectation of X, is a scalar (a number) – non-random. The expectation operator E averages out all the randomness in X, to give its mean (a weighted average of the possible value of X, weighted according to their probability, in the discrete case). It often happens that we have *partial information* about X – for instance, we may know the value of a random variable Y which is associated with X, i.e. carries information about X. We may want to average out over the remaining randomness. This is an expectation conditional on our partial information, or more briefly a conditional expectation. This idea will be familiar already from elementary courses, in two cases:
1. *Discrete case*, based on the formula

$$P(A|B) := P(A \cap B)/P(B) \text{ if } P(B) > 0.$$

If X takes values $x_1, \cdots, x_m$ with probabilities $f_1(x_i) > 0$, Y takes values $y_1, \cdots, y_n$ with probabilities $f_2(y_j) > 0$, (X, Y) takes values (x_i, y_j) with probabilities $f(x_i, y_j) > 0$, then
(i) $f_1(x_i) = \Sigma_j f(x_i, y_j)$, $\quad f_2(y_j) = \Sigma_i f(x_i, y_j)$,
(ii) $P(Y = y_j|X = x_i) = P(X = x_i, Y = y_j)/P(X = x_i) = f(x_i, y_j)/f_1(x_i)$

$$= f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

This is the *conditional distribution* of Y *given* $X = x_i$, written

$$f_{Y|X}(y_j|x_i) = f(x_i, y_j)/f_1(x_i) = f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

Its expectation is

$$E(Y|X = x_i) = \Sigma_j y_j f_{Y|X}(y_j|x_i) = \Sigma_j y_j f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

But this approach only works when the events on which we condition have *positive* probability, which only happens in the *discrete* case.
2. *Density case.* If (X, Y) has density $f(x, y)$, X has density $f_1(x) :=$

$\int_{-\infty}^{\infty} f(x,y)dy$, Y has density $f_2(y) := \int_{-\infty}^{\infty} f(x,y)dx$. We *define* the *conditional density* of Y *given* $X = x$ by the continuous analogue of the discrete formula above:

$$f_{Y|X}(y|x) := f(x,y)/f_1(x) = f(x,y)/\int_{-\infty}^{\infty} f(x,y)dy.$$

Its expectation is

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y f(x,y)dy/\int_{-\infty}^{\infty} f(x,y)dy.$$

Example: Bivariate normal distribution, $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

$$E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),$$

the familiar *regression line* of statistics (linear model). See e,g. [BF].

The problem here is that joint densities need not exist – do not exist, in general. One of the great contributions of Kolmogorov's classic book of 1933 was the realization that measure theory – specifically, the Radon-Nikodym theorem – provides a way to treat conditioning in general, without making assumptions that we are in one of the two cases – discrete case and density case – above. Recall that the probability triple is $(\Omega, \mathcal{A}, P)$. Suppose that $\mathcal{B}$ is a sub-σ-field of $\mathcal{A}$, $\mathcal{B} \subset \mathcal{A}$ (recall that a σ-field represents information; the big σ-field $\mathcal{A}$ represents 'knowing everything', the small σ-field $\mathcal{B}$ represents 'knowing something').

Suppose that Y is a non-negative random variable whose expectation exists: $EY < \infty$. The set-function

$$Q(B) := \int_B Y dP \qquad (B \in \mathcal{B})$$

is non-negative (because Y is), σ-additive – because

$$\int_B Y dP = \Sigma_n \int_{B_n} Y dP$$

if $B = \cup_n B_n$, B_n disjoint – and defined on the σ-algebra $\mathcal{B}$, so is a *measure* on $\mathcal{B}$. If $P(B) = 0$, then $Q(B) = 0$ also (the integral of anything over a null set is zero), so $Q << P$. By the Radon-Nikodym theorem, there exists a Radon-Nikodym derivative of Q with respect to P on $\mathcal{B}$, which is $\mathcal{B}$-measurable (in the RN theorem earlier, we had 'measurable', meaning '$\mathcal{A}$-measurable; here

replace $\mathcal{A}$ by $\mathcal{B}$). Following Kolmogorov (1933), we call this RN derivative the *conditional expectation* of Y *given* (or *conditional on*) $\mathcal{B}$, $E(Y|\mathcal{B})$: this is $\mathcal{B}$-measurable, integrable, and satisfies

$$\int_B Y dP = \int_B E(Y|\mathcal{B}) \lceil \mathcal{P} \qquad \forall \mathcal{B} \in \mathcal{B}. \tag{$*$}$$

In the general case, where $E|Y| < \infty$ but Y can change sign,

$$Y = Y_+ - Y_-$$

and define $E(Y|\mathcal{B})$ by linearity as

$$E(Y|\mathcal{B}) := E(Y_+|\mathcal{B}) - E(Y_-|\mathcal{B}).$$

Suppose now that $\mathcal{B}$ is the σ-field generated by a random variable X: $\mathcal{B} = \sigma(X)$ (so $\mathcal{B}$ represents the information contained in X, or what we know when we know X). Then $E(Y|\mathcal{B}) = E(Y|\sigma(X))$, which is written more simply as $E(Y|X)$. Its defining property is

$$\int_B Y dP = \int_B E(Y|X) dP \qquad \forall B \in \sigma(X).$$

Similarly, if $\mathcal{B} = \sigma(X_1, \cdots, X_n)$ ($\mathcal{B}$ is the information in $(X_1, \cdots, X_n)$) we write $E(Y|\sigma(X_1, \cdots, X_n)$ as $E(Y|X_1, \cdots, X_n)$:

$$\int_B Y dP = \int_B E(Y|X_1, \cdots, X_n) dP \qquad \forall \mathcal{B} \in \sigma(X_1, \cdots, X_n).$$

Note. 1. To check that something is a conditional expectation: we have to check that it integrates the right way over the right sets [i.e., as in (*)].
2. From (*): if two things integrate the same way over all sets $B \in \mathcal{B}$, they have the same conditional expectation given $\mathcal{B}$.
3. We shall pass between the notations $E(Y|\mathcal{B})$ and $E_{\mathcal{B}}Y$ at will.
4. The conditional expectation thus defined coincides with any we may have already encountered - in regression or multivariate analysis, for example. However, this may not be immediately obvious. The conditional expectation defined above – via σ-fields and the Radon-Nikodym theorem – is rightly called by Williams ([**W**], p.84) 'the central definition of modern probability'. It may take a little getting used to. As with all important but non-obvious definitions, it proves its worth in action: see below.

2. Properties of conditional expectations.
1. $\mathcal{B} = \{\emptyset, \Omega\}$. Here $\mathcal{B}$ is the *smallest* possible σ-field (*any* σ-field of subsets of Ω contains $\emptyset$ and Ω), and represents 'knowing nothing'.

$$E(Y|\{\emptyset, \Omega\}) = EY.$$

Proof. We have to check (*) for $B = \emptyset$ and $B = \Omega$. For $B = \emptyset$ both sides are zero; for $B = \Omega$ both sides are EY. //
2. $\mathcal{B} = \mathcal{A}$. Here $\mathcal{B}$ is the *largest* possible σ-field, and represents 'knowing everything'.

$$E(Y|\mathcal{A}) = Y \qquad P - a.s.$$

Proof. We have to check (*) for *all* sets $B \in \mathcal{A}$. The only integrand that integrates like Y over *all* sets is Y itself (or a function $= Y$ a.s.)
Note. When we condition on $\mathcal{A}$ ('knowing everything'), we *know* Y (because we know everything). There is thus no uncertainty left in Y to average out, so taking the conditional expectation (averaging out remaining randomness) has no effect, and leaves Y unaltered.
3. If Y is $\mathcal{B}$-measurable, $E(Y|\mathcal{B}) = Y$ P-a.s.
Proof. Recall that Y is *always* $\mathcal{A}$-measurable (this is the definition of Y being a random variable). For $\mathcal{B} \subset \mathcal{A}$, Y may not be $\mathcal{B}$-measurable, but if it is, the proof above applies with $\mathcal{B}$ in place of $\mathcal{A}$.
Note. If Y is $\mathcal{B}$-measurable, when we are given $\mathcal{B}$ (that is, when we condition on it), we *know* Y. That makes Y effectively a constant, and when we take the expectation of a constant, we get the same constant.
4 (Pull-out property). If Y is $\mathcal{B}$-measurable, $E(YZ|\mathcal{B}) = YE(Z|\mathcal{B})$ P-a.s.
Proof. We need to show

$$\int_B YZdP = Y\int_B ZdP \qquad (B \in \mathcal{B}).$$

If $Y = I_{B'}$ is the indicator of a set $B' \in \mathcal{B}$, this holds, as both sides are $\int_{B\cap B'} ZdP$. By linearity, it holds for simple $\mathcal{B}$-measurable functions. It then extends to non-negative integrable $\mathcal{B}$-measurable functions by approximation as usual, and to the general case by taking positive and negative parts. //
Note. Williams calls this property 'taking out what is known'. To remember it: if Y is $\mathcal{B}$-measurable, then given $\mathcal{B}$ we know Y, so Y is effectively a constant, so can be taken out through the integration signs.
5 (Tower property). If $\mathcal{C} \subset \mathcal{B}$, $E[E(Y|\mathcal{B})\ |\mathcal{C}] = E[Y|\mathcal{C}]$ a.s.

Proof. $E_{\mathcal{C}}E_{\mathcal{B}}Y$ is $\mathcal{C}$-measurable, and for $C \in \mathcal{C} \subset \mathcal{B}$,

$$\int_C E_{\mathcal{C}}[E_{\mathcal{B}}Y]dP = \int_C E_{\mathcal{B}}Y dP \qquad \text{(definition of } E_{\mathcal{C}} \text{ as } C \in \mathcal{C})$$

$$= \int_C Y dP \qquad \text{(definition of } E_{\mathcal{B}} \text{ as } C \in \mathcal{B}).$$

So $E_{\mathcal{C}}[E_{\mathcal{B}}Y]$ satisfies the defining relation for $E_{\mathcal{C}}Y$. Being also $\mathcal{C}$-measurable, it *is* $E_{\mathcal{C}}Y$ (a.s.). //

5' (Tower property). If $\mathcal{C} \subset \mathcal{B}$, $E[E(Y|\mathcal{C}) \ |\mathcal{B}] = E[Y|\mathcal{C}]$ a.s.

Proof. $E[Y|\mathcal{C}]$ is $\mathcal{C}$-measurable, so $\mathcal{B}$-measurable as $\mathcal{C} \subset \mathcal{B}$, so $E[.|\mathcal{B}]$ has no effect, by 3. //

Corollary. $E[E(Y|\mathcal{C}) \ |\mathcal{C}] = E[Y|\mathcal{C}]$ a.s.

Thus the operation $E[.|\mathcal{C}]$ is linear and *idempotent* (doing it twice is the same as doing it once), so is a *projection*. So we can use what we know about projections, from Linear Algebra, Functional Analysis etc.

Note. The tower property (in either form) is also known as the *iterated conditional expectations property* or *coarse-averaging property*. When conditioning on two σ-fields, one larger (finer), one smaller (coarser), the coarser rubs out the effect of the finer, either way round.

6. *Role of independence.* If Y is independent of $\mathcal{B}$,

$$E(Y|\mathcal{B}) = EY \qquad a.s.$$

Proof. We require

$$E[Y]P(B) = E[Y]\int_B dP = \int_B Y dP \qquad (B \in \mathcal{B}).$$

If $Y = I_A$ is an indicator, I_A, I_B are independent, so

$$P(A \cap B) = E[I_{A\cap B}] = E[I_A.I_B] = E[I_A].E[I_B] = P(A)P(B),$$

by the Multiplication Theorem. This gives the result for indicators; we extend to simple functions by linearity, and thence to the non-negative integrable case and the general case as usual. //

7. *Conditional Mean Formula.*

$$E[E(Y|\mathcal{B})] = EY \quad P-a.s.$$

Proof. Take $\mathcal{C} = \{\emptyset, \Omega\}$ in 5 and use 1. //
Example. Check this for the bivariate normal distribution considered above.
8. *Conditional Variance Formula.*

$$varY = E_X var(Y|X) + var_X E(Y|X).$$

Recall $varX := E[(X - EX)^2]$. Expanding the square,

$$varX = E[X^2 - 2X.(EX) + (EX)^2] = E(X^2) - 2(EX)(EX) + (EX)^2 = E(X^2) - (EX)^2.$$

Conditional variances can be defined in the same way. Recall that $E(Y|X)$ is constant when X is known ($= x$, say), so can be taken outside an expectation over X, E_X say. Then

$$var(Y|X) := E(Y^2|X) - [E(Y|X)]^2.$$

Take expectations of both sides over X:

$$E_X var(Y|X) = E_X[E(Y^2|X)] - E_X[E(Y|X)]^2.$$

Now $E_X[E(Y^2|X)] = E(Y^2)$, by the Conditional Mean Formula, so the right is, adding and subtracting $(EY)^2$,

$$\{E(Y^2) - (EY)^2\} - \{E_X[E(Y|X)]^2 - (EY)^2\}.$$

The first term is $varY$, by above. Since $E(Y|X)$ has E_X-mean EY, the second term is $var_X E(Y|X)$, the variance (over X) of the random variable $E(Y|X$ (random because X is). Combining, the result follows.
Interpretation. $varY =$ total variability in Y,
$E_X var(Y|X) =$ variability in Y not accounted for by knowledge of X,
$var_X E(Y|X) =$ variability in Y accounted for by knowledge of X.
Example: the bivariate normal.

$$Y|X = x \text{ is } N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)), \qquad varY = \sigma_2^2,$$

$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \qquad E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1),$$

which has variance $(\rho\sigma_2/\sigma_1)^2 varX = (\rho\sigma_2/\sigma_1)^2\sigma_1^2 = \rho^2\sigma_2^2$,

$var(Y|X = x) = \sigma_2^2$ for all $x, var(Y|X) = \sigma_2^2(1-\rho^2), E_X var(Y|X) = \sigma_2^2(1-\rho^2)$.

Corollary. $E(Y|X)$ has the same mean as Y and smaller variance (if anything).
Proof. From the Conditional Mean Formula, $E[E(Y|X)] = EY$. Since $var(Y|X) \geq 0$, $E_X var(Y|X) \geq 0$, so $varE[Y|X] \leq varY$ from the Conditional Variance Formula.

This result has important applications in estimation theory. Suppose we are to estimate a parameter θ, and are considering a statistic X as a possible estimator (or basis for an estimator) of θ. We would naturally want X to contain all the information on θ contained within the entire sample. What (if anything) does this mean in precise terms? The answer lies in the concept of *sufficiency* ('data reduction') – one of the most important contributions to statistics of the great English statistician R. A. (Sir Ronald) Fisher (1880-1962). In the language of sufficiency, the Conditional Variance Formula is seen as (essentially) the Rao-Blackwell Theorem, a key result in the area (see the index in your favourite Statistics book if you want more here).

3. Filtrations.

The Kolmogorov triples $(\Omega, \mathcal{F}, P)$, and the Kolmogorov conditional expectations $E(X|\mathcal{B})$, give us all the machinery we need to handle *static* situations involving randomness. To handle *dynamic* situations, involving randomness which unfolds with *time* – the essence of Stochastic Processes – we need further structure.

Suppose time evolves in integer steps, $t = 0, 1, 2, \cdots$ (so we start at time $t = 0$; we postpone continuous time). We suppose, for simplicity, that information is never lost (or forgotten): thus, as time increases we learn more. Recall that σ-fields represent information or knowledge. We thus need an increasing sequence of σ-fields $\{\mathcal{F}_n : n = 0, 1, 2, \cdots\}$, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ $(n = 0, 1, 2, \cdots)$, where $\mathcal{F}_n$ represents what we know at time n. As usual, we take the σ-fields to be *complete*, i.e., to contain all subsets of null sets as null sets. Thus $\mathcal{F}_0$ represents the initial information (if there is none, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, the trivial σ-field). On the other hand, $\mathcal{F}_\infty := lim_{n\to\infty}\mathcal{F}_n$ represents all we ever will know (the 'Doomsday σ-field'). Often, $\mathcal{F}_\infty$ will be $\mathcal{F}$, but not always; see e.g. [**W**], 15.8 for an interesting example.

Such a family $\{\mathcal{F}_n : n = 0, 1, 2, \cdots\}$ is called a *filtration*; a probability space endowed with such a filtration, $\{\Omega, \{\mathcal{F}_n\}, \mathcal{F}, \mathcal{P}\}$ is called a *filtered probability space*. (These definitions are due to P. A. MEYER (1934-2003) of Strasbourg; Meyer and the Strasbourg (and more generally, French) school

of probabilists have been responsible for the 'general theory of [stochastic] processes', and for much of the progress in stochastic integration, since the 1960s). Since the filtration is so basic to the definition of a stochastic process, the more modern term for a filtered probability space is a *stochastic basis*.

We take a stochastic basis $(\Omega, \{\mathcal{F}_t, \}, \mathcal{F}, P)$ (or *filtered probability space*), which following Meyer we assume satisfies the *usual conditions* (conditions habituelles):
a. completeness: each $\mathcal{F}_t$ contains all P-null sets of $\mathcal{F}$;
b. the filtration is right-continuous, i.e. $\mathcal{F}_t = \mathcal{F}_{t+} := \cap_{s>t}\mathcal{F}_s$.

4. Finite-dimensional distributions

A *stochastic process* $X = (X(t))_{t\geq 0}$ is a family of random variables defined on a stochastic basis. We say X is *adapted* if $X(t) \in \mathcal{F}_t$ (i.e. $X(t)$ is $\mathcal{F}_t$-measurable) for each t: thus $X(t)$ is known when $\mathcal{F}_t$ is known, at time t.

If $\{t_1, \cdots, t_n\}$ is a finite set of time points in $[0, \infty)$, $(X(t_1), \ldots, X(t_n))$ is a random n-vector, with a distribution, $\mu(t_1, \ldots, t_n)$ say. The class of all such distributions as $\{t_1, \ldots, t_n\}$ ranges over all finite subsets of $[0, \infty)$ is called the class of all *finite-dimensional distributions* of X. These satisfy certain obvious consistency conditions:
DK1. deletion of one point t_i can be obtained by 'integrating out the unwanted variable', as usual when passing from joint to marginal distributions;
DK2. permutation of the times t_i permutes the arguments of the measure $\mu(t_1, \ldots, t_n)$ on $\mathbf{R}^n$ in the same way.
Conversely, a collection of finite-dimensional distributions satisfying these two consistency conditions arises from a stochastic process in this way (this is the content of the *Daniell-Kolmogorov theorem*). This classical result (due to P.J. Daniell in 1918 and A.N. Kolmogorov in 1933) is the basic existence theorem for stochastic processes. For the proof, see e.g. [K].

Important though it is as a general existence result, however, the Daniell-Kolmogorov theorem does not take us very far. It gives a stochastic process X as a random function on $[0, \infty)$, i.e. a random variable on $\mathbf{R}^{[0,\infty)}$. This is a vast and unwieldy space; we shall usually be able to confine attention to much smaller and more manageable spaces, of functions satisfying regularity conditions. The most important of these is continuity: we want to be able to realize $X = (X(t, \omega))_{t\geq 0}$ as a random continuous function, i.e. a member of $C[0, \infty)$; such a process X is called path-continuous (since the map $t \to X(t, \omega)$ is called the sample path, or simply path, given by ω) – or more briefly, continuous. This is possible for the extremely important case of Brownian motion, for example, and its relatives. Sometimes we need to

allow our random function $X(t,\omega)$ to have jumps. It is then customary, and convenient, to require $X(t)$ to be right-continuous with left limits (RCLL), or càdlàg (*continu à droite, limite à gauche*) – i.e. to have X in the space $D[0,\infty)$ of all such functions (the Skorohod space). This is the case, for instance, for the Poisson process and its relatives (see below).

General results on realisability – whether or not it is possible to realize, or obtain, a process so as to have its paths in a particular function space – are known; see for example the Kolmogorov-Ĉentsov theorem. For our purposes, however, it is usually better to construct the processes we need directly on the function space on which they naturally live.

Given a stochastic process X, it is sometimes possible to improve the regularity of its paths without changing its distribution (that is, without changing its finite-dimensional distributions). For background on results of this type (separability, measurability, versions, regularization etc.) see e.g. the classic book [D].

There are several ways to define 'sameness' of two processes X and Y. We say
(i) X and Y have the *same finite-dimensional distributions* if, for any integer n and $\{t_1,\cdots,t_n\}$ a finite set of time points in $[0,\infty)$, the random vectors $(X(t_1),\ldots,X(t_n))$ and $(Y(t_1),\ldots,Y(t_n))$ have the same distribution;
(ii) Y is a *modification* of X if, for every $t\geq 0$, we have $P(X_t=Y_t)=1$;
(iii) X and Y are *indistinguishable* if almost all their sample paths agree:

$$P\left[X_t=Y_t;\forall 0\leq t<\infty\right]=1.$$

Indistinguishable processes are modifications of each other; the converse is not true in general. However, if both processes have right-continuous sample paths, the two concepts are equivalent. This will cover the processes we encounter in this course.

A process is called *progressively measurable* if the map $(t,\omega)\mapsto X_t(\omega)$ is measurable, for each $t\geq 0$. Progressive measurability holds for adapted processes with right-continuous (or left-continuous) paths – and so always in the generality in which we work.

Finally, a random variable $\tau:\Omega\to[0,\infty]$ is a *stopping time* if $\{\tau\leq t\}\in\mathcal{F}_t$ for all $t\geq 0$.

If $\{\tau<t\}\in\mathcal{F}_t$ for all t, τ is called an *optional time*. For right-continuous filtrations (as here, under the usual conditions) the concepts of stopping and optional times are equivalent.

For a set $A \subset \mathbf{R}^d$ and a stochastic process X, we can define the *hitting time* of A for X as

$$\tau_A := \inf\{t > 0 : X_t \in A\}.$$

For our usual situation (RCLL processes and Borel sets) hitting times are stopping times.

We will also need the *stopping time σ-algebra* $\mathcal{F}_\tau$ defined as

$$\mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap \{\tau \leq t\} \in \mathcal{F}_t.$$

Intuitively, $\mathcal{F}_\tau$ represents the events known at time τ.

The continuous-time theory is technically much harder than the discrete-time theory, for two reasons:
1. questions of path-regularity arise in continuous time but not in discrete time;
2. uncountable operations (such as taking the supremum over an interval) arise in continuous time. But measure theory is constructed using countable operations: uncountable operations risk losing measurability.
This is why discrete and continuous time are often treated separately.