

Lectures 23–24. Conditional probabilities. Markov chains.

Let B be an event such that $P(B) > 0$. The *conditional probability* of an event A *under the condition* B (or *under the condition that* B *occurred*, or *occurs*, or *takes place*, or some other synonym) is defined (and denoted) as

$$P(A|B) = P_B(A) = \frac{P(A \cap B)}{P(B)}. \quad (23-24.1)$$

The first notation, $P(A|B)$, is the traditional and customary one, but it leads some people to think that it is the probability, $P(\quad)$, of the ‘conditional event’, $A|B$ (and it is not only beginning students, but also some engineering-minded people who are speaking about ‘conditional random variables’, etc.). Nothing of the kind! the composite sign $P(A|B)$ is divided into two signs not this way, but: it is the *conditional probability under condition* B , $P(\quad|B)$, – and of what? of *the event* A . The example $\Omega = \{1, 2, 3\}$, $P\{1\} = P\{2\} = P\{3\} = 1/3$, $B = \{1, 2\}$, $A = \{1\}$ shows that it can happen that no event at all has the probability equal to the conditional probability $P(A|B)$ ($= 1/2$ in this example).

The notation $P_B(A)$ is less likely to be misinterpreted as the probability, $P(\quad)$, of some ‘conditional event’, which is denoted, exotically, as ${}_B A$, and one part of which is put inside the parentheses, and the other, outside.

Note that when speaking about conditional probabilities and just probabilities, the latter are sometimes called *unconditional* ones. Note also that sometimes instead of the expression *the conditional probability of* A *under the condition* ... the expression *probability of* A *under the condition* ... is used.

Of course, we can write the definition (23–24.1) of the conditional probability in another form:

$$P(A \cap B) = P(B) \cdot P(A|B); \quad (23-24.2)$$

and also, say,

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B|A) \cdot P(C|A \cap B) \cdot P(D|A \cap B \cap C), \quad (23-24.3)$$

provided all conditional probabilities in the right-hand side are meaningful (i. e., $P(A) \neq 0$, $P(A \cap B) \neq 0$, $P(A \cap B \cap C) \neq 0$).

Theorem 23–24.1 (the Total Probability Formula). *Let* A_k , $k = 1, 2, \dots, n$ ($, \dots$) *be a finite or a countable number of mutually exclusive (i. e. disjoint) events, whose union is the whole* Ω : $\bigcup_k A_k = \Omega$.

Then for every (other) event B

$$P(B) = \sum_k P(A_k) \cdot P(B|A_k). \quad (23-24.4)$$

Here, if $P(A_k) = 0$, *the conditional probability* $P(B|A_k)$ *makes no sense, so we have to make precise how the corresponding summands in (23–24.4) are understood: we agree to take* $0 \cdot \text{nonsense} = 0$.

Proof. The k -th summand in (23–24.4) is equal to $P(A_k \cap B)$ if $P(A_k) \neq 0$ by (23–24.2); if $P(A_k) = 0$, it is also equal to $P(A_k \cap B)$, but by a completely other reason: by our convention above, and by the inclusion $A_k \cap B \subseteq A_k$ and the inequality $P(A_k \cap B) \leq P(A_k) = 0$. So we have to prove that

$$P(B) = \sum_k P(A_k \cap B). \quad (23-24.5)$$

This follows from the fact that $B = \Omega \cap B = (\bigcup_k A_k) \cap B = \bigcup_k (A_k \cap B)$ is a union of disjoint summands, and the countable additivity axiom.

If we fix an event B with $P(B) \neq 0$, we can consider a new probability space $(\Omega, \mathcal{F}, P_B)$ (the first two elements Ω, \mathcal{F} are the same as in (Ω, \mathcal{F}, P)). The only thing here to be checked is that $P_B(A)$, as a function of A , is a *measure* satisfying $P_B(\Omega) = 1$. The latter is obvious, and that P_B is a measure is easy: it is clearly nonnegative, and for disjoint A_k

$$\begin{aligned} P_B\left(\bigcup_k A_k\right) &= \frac{P(B \cap \bigcup_k A_k)}{P(B)} = \frac{\sum_k P(B \cap A_k)}{P(B)} \quad (\text{the sets } B \cap A_k \text{ are disjoint}) \\ &= \sum_k \frac{P(B \cap A_k)}{P(B)} = \sum_k P_B(A_k) \end{aligned} \quad (23-24.6)$$

(in the lecture I formulated this as a theorem; but I can see no need now to attract so much attention to a very simple thing).

So on this new probability space we can consider everything we have considered for the space (Ω, \mathcal{F}, P) , using in their names the epithet ‘conditional’: *conditional distributions, conditional probability densities, conditional expectations*, etc. E.g., the *conditional distribution* of a random variable ξ taking values in a measurable space (X, \mathcal{X}) under the condition B is, by definition, the set function

$$\mu_B(C) = \mu_{\xi|B}(C) = P\{\xi \in C|B\} \quad (23-24.7)$$

($P\{\xi \in C|B\}$ is the short notation instead of the cumbersome $P(\{\xi \in C\}|B)$).

We do not need to prove that this set function is a *measure*: we *know* that the distribution of a random variable with respect to an arbitrary probability measure is a measure, and $\mu_{\xi|B}$ is just a distribution with respect to the probability measure P_B .

Let A_1, A_2, \dots, A_n ($, \dots$) be mutually exclusive (disjoint) events, $\bigcup_i A_i = \Omega$ (as in Theorem 23–24.1); and let ξ be a random variable taking values in a measurable space (X, \mathcal{X}) . If μ is the distribution of ξ (the *unconditional* distribution), and μ_{A_i} is its conditional distribution under the condition A_i , we have for every $C \in \mathcal{X}$, by the total probability formula:

$$\mu(C) = P\{\xi \in C\} = \sum_i P(A_i) \cdot P\{\xi \in C|A_i\} = \sum_i P(A_i) \cdot \mu_{A_i}(C). \quad (23-24.8)$$

This formula means that the unconditional distribution μ is *the mixture of the conditional distributions* μ_{A_i} with the weights $q_i = P(A_i)$ (what the *mixture* means, see Lecture 5).

Another “conditional” concept: the *conditional expectation* of a random variable ξ under the condition B is defined as

$$E_B \xi = E(\xi|B) = \int_{\Omega} \xi(\omega) P_B(d\omega). \quad (23-24.9)$$

It is easy to prove that

$$E(\xi|B) = \frac{\int_B \xi(\omega) P(d\omega)}{P(B)} = \frac{E I_B \cdot \xi}{P(B)}. \quad (23-24.10)$$

The proof is carried out like this: first for all random variables taking a finite number of values; then, by definition, for nonnegative random variables, taking their nondecreasing approximations by those taking finite numbers of values; then for arbitrary ones, as the difference of the integral of the positive part and the negative part.

Theorem 23 – 24.2. *If the unconditional expectation $E\xi$ exists, then the conditional expectation $E_B \xi$ also exists for every event B with $P(B) > 0$.*

Proof. If the integral $\int_{\Omega} \xi dP$ exists, then the integral over every smaller set $B \in \mathcal{F}$ also exists. And then we use formula (23–24.10)

Theorem 23 – 24.3 (the Total Expectation Formula). *Let A_k , $k = 1, 2, \dots, n$ (\dots) be a finite or a countable number of mutually exclusive (i.e. disjoint) events, whose union is the whole Ω : $\bigcup_k A_k = \Omega$.*

Let ξ be a random variable for which the expectation $E\xi$ makes sense. Then all conditional expectations $E(\xi|A_i)$ also exist for all i for which $P(A_i) > 0$, and

$$E\xi = \sum_i P(A_i) \cdot E(\xi|A_i). \quad (23-24.11)$$

Proof. Formula (23–24.11) follows from the fact that $\int_{\bigcup_i A_i} \xi dP = \sum_i \int_{A_i} \xi dP$ and formula (23–24.10).

We can consider not only conditional distributions, conditional expectations, conditional *variances*, etc., but also the *conditional conditional probability* (iterated conditional probability):

$$P_B(A|C) = \frac{P_B(A \cap C)}{P_B(C)}, \quad (23-24.12)$$

provided that not only $P(B) \neq 0$, but also $P_B(C) \neq 0$.

It is easy to see that

$$P_B(A|C) = \frac{P((A \cap C) \cap B)/P(B)}{P(C \cap B)/P(B)} = \frac{P(A \cap B \cap C)}{P(B \cap C)} = P(A|B \cap C). \quad (23-24.13)$$

So no really new concept arises; but in return we need no separate proof of the following

Theorem 23–24.4 (the Total Conditional Probability Formula). *Under the conditions and conventions of Theorem 23–24.1, if $P(B) \neq 0$,*

$$P(C|B) = \sum_k P(A_k|B) \cdot P(C|B \cap A_k). \quad (23-24.14)$$

We can also write the Total Conditional Expectation Formula (a cross between (23–24.11) and (23–24.14)): $E(\xi|B) = \sum_k P(A_k|B) \cdot E(\xi|B \cap A_k)$.

Conditional probabilities are used to define a very important type of dependence between random variables: *Markov-type* dependence.

Let X be a finite or a countable set; \mathcal{X} , as usual, the σ -algebra of all its subsets.

A *Markov chain* in X is a sequence $\xi_0, \xi_1, \dots, \xi_n, \dots$ of random elements of this space, such that for all n , $x_0, x_1, \dots, x_n, x_{n+1} \in X$

$$P\{\xi_{n+1} = x_{n+1} | \xi_0 = x_0, \xi_1 = x_1, \dots, \xi_{n-1} = x_{n-1}, \xi_n = x_n\} = P\{\xi_{n+1} = x_{n+1} | \xi_n = x_n\}, \quad (23-24.15)$$

provided that the probability of the condition is positive:

$$P\{\xi_0 = x_0, \xi_1 = x_1, \dots, \xi_{n-1} = x_{n-1}, \xi_n = x_n\} > 0. \quad (23-24.16)$$

If the probability of the condition is equal to 0, we require *nothing*.

If we interpret the integer-valued variable k in ξ_k as time, and take n to be *the present*, then 0, 1, ..., $n - 1$ are the past, and $n + 1$ the future – and (23–24.15) is interpreted thus: *the future depends on the past only through the present*. Of course, Markov chains (and Markov *stochastic processes*, for which this kind of dependence is also characteristic) are mathematical models of some ‘real-world’ processes, but there are real-world processes for which the mathematical model of Markov chains is inappropriate; e. g., the sequence of letters of a text (the space X is the alphabet): as you can understand, there are important long-time dependencies between the later parts and the earlier parts of natural texts.

The distribution $q_x = P\{\xi_0 = x\}$, $x \in X$, is called *the initial distribution* of the Markov chain. The right-hand side in the formula (23–24.15) depends only on n , x_n , and x_{n+1} , so it can be written as $p_{x_n x_{n+1}}^{n+1}$, where the numbers p_{xy}^{n+1} , $x, y \in X$, form a matrix P_{n+1} . Some rows in this matrix are not filled – if the probability $P\{\xi_n = x\} = 0$; then we will fill it arbitrarily with nonnegative numbers whose sum is equal to 1. The matrix $P_k = (p_{xy}^k)_{x, y \in X}$ is called *the matrix of transition probabilities of the Markov chain at the k -th step*, or the *k -th transition matrix*. The entries of P_k are, in fact, *conditional probabilities*, but such is the tradition.

Let us call a matrix $(p_{xy})_{x, y \in X}$ a *stochastic matrix* if its entries are nonnegative, and their sum in every row is equal to 1: $p_{xy} \geq 0$, and $\sum_y p_{xy} = 1$ for every $x \in X$. A transition matrix of a Markov chain is necessarily a stochastic one.

Let us consider some examples.

A sequence of independent random variables ξ_k taking values $1, 2, \dots, m$, with common distribution given by $P\{\xi_k = i\} = p_i$ is a Markov chain whose transition matrix P_k is the same at every step: $P_k = P$,

$$P = \begin{pmatrix} p_1 & p_2 & \dots & p_m \\ p_1 & p_2 & \dots & p_m \\ \dots & \dots & \dots & \dots \\ p_1 & p_2 & \dots & p_m \end{pmatrix} \quad (23-24.17)$$

(with identical rows).

*

Another example:

The simplest symmetric random walk in \mathbb{Z}^1 (the set of all integers). As the set X we take the set \mathbb{Z}^1 of all integers. Let a particle move in discrete time like this: if at some step it is at a point $x \in \mathbb{Z}^1$, then, independently of how it came to this point, it goes over with probability $1/2$ to the nearest point to the right, $x + 1$, and with probability $1/2$ to the nearest point to the left, $x - 1$. The transition matrix for this chain is also independent on after how many steps it is, and this matrix has the form

$$P = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & \dots \\ \dots & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & \dots \\ \dots & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & \dots \\ \dots & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & \dots \\ \dots & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix} \quad (23-24.18)$$

\triangle

(since the matrix is infinite in both directions, I am telling you that the 0-th row is that is opposite the equality sign, and the zeroth column that above the sign \triangle). We can express the same by a formula:

$$p_{x, x-1} = p_{x, x+1} = 1/2, \quad (23-24.19)$$

and all other are equal to 0.

A generalization of this: the simplest symmetric random walk in the two-dimensional integer lattice \mathbb{Z}^2 . This random walk exists in more that one version. The first one is that with the transition probabilities

$$p_{x, y; x+1, y} = p_{x, y; x-1, y} = p_{x, y; x, y+1} = p_{x, y; x, y-1} = 1/4, \quad (23-24.20)$$

and the remaining, zeros (with equally probable transitions at every step to the nearest point to the right, to the left, up, and down); the second one, with

$$p_{x, y; x\pm 1, y\pm 1} = 1/4, \quad (23-24.21)$$

and all remaining transition probabilities zero – with equally probable transitions to the four points that are the nearest in the diagonal direction. (We don't write down the transition matrix: the rows and the columns should be numbered by points of a two-dimensional lattice, and the natural arrangement of this matrix would be in a four-dimensional space). Both these versions are isomorphic to one another: one is obtained from the other by means of a rotation by 45° and a change of scale (make a picture or two).

On a three-dimensional lattice, there are several non-isomorphic versions of the simplest random walk, in particular, that with equally probable transitions to the six nearest points in the directions of the axes: right, left, forward, backward, up and down; or with equal transition probabilities to the eight points that are nearest in the diagonal direction: $p_{x, y, z; x\pm 1, y\pm 1, z\pm 1} = 1/8$ (the rest is 0).

One more example: **sums of independent random variables.** Let $\eta_1, \eta_2, \dots, \eta_n, \dots$ be independent identically distributed integer-valued random variables, $P\{\xi_k = i\} = p_i$. Let us make up the sums: $\xi_n = \sum_{k=1}^n \eta_k$ (in particular, $\xi_0 = 0$). Then $\xi_0, \xi_1, \xi_2, \dots$ is a Markov chain in \mathbb{Z}^1 .

Indeed, for x_i being integers ($x_0 = 0$) we have:

$$\begin{aligned} P\{\xi_{n+1} = x_{n+1} | \xi_0 = x_0, \xi_1 = x_1, \dots, \xi_{n-1} = x_{n-1}, \xi_n = x_n\} \\ = P\{\eta_{n+1} = x_{n+1} - x_n | \eta_1 = x_1 - x_0, \eta_2 = x_2 - x_1, \dots, \eta_n = x_n - x_{n-1}\} \quad (23-24.22) \\ = P\{\eta_{n+1} = x_{n+1} - x_n\} = p_{x_{n+1} - x_n}. \end{aligned}$$

That is, it is a Markov chain with transition matrix (at every step)

$$P = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & p_{-1} & p_0 & p_1 & p_2 & p_3 & \dots \\ \dots & p_{-2} & p_{-1} & p_0 & p_1 & p_2 & \dots \\ \dots & p_{-3} & p_{-2} & p_{-1} & p_0 & p_1 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}. \quad (23-24.23)$$

\triangle

A particular case is the simplest random walk on \mathbb{Z}^1 , where $p_{\pm 1} = 1/2$, $p_i = 0$ for $i \neq \pm 1$.

The following example was not given in Lecture 24, but I am including it here:

Let $\eta_1, \eta_2, \dots, \eta_n, \dots$ be a sequence of independent random variables with the same distribution, namely, $P\{\eta_i = 1\} = 1/3$, $P\{\eta_i = -2\} = 2/3$; and ξ_n is the smallest nonnegative remainder after dividing the sum $\sum_{i=1}^n \eta_i$ by 5 (e.g., if $\eta_1 = 1, \eta_2 = -2, \dots$, we have: $\xi_0 = 0, \xi_1 = 1, \xi_2 = 4$, because $\eta_1 + \eta_2 = -1 = 5 \cdot (-1) + 4; \dots$). The transition matrix of the Markov chain $\xi_0, \xi_1, \xi_2, \dots$ is

$$P = \begin{pmatrix} 0 & 1/3 & 0 & 2/3 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 0 & 1/3 & 0 \\ 0 & 2/3 & 0 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 & 0 \end{pmatrix}. \quad (23-24.24)$$

We'll go to the general theory in the next lecture.