PROBABILITY 2017
HANDOUT 3: CONDITIONING

GIOVANNI PISTONE

Contents

1. Conditional expectation 1
2. Conditional distribution 4
References 5

Conditioning is one among the core concepts in reasoning about uncertainty in Probability, in Statistics, in Economics, in Machine Learning. In this notes we refer mainly to the textbook by D. Williams [2, Ch. 9]. A concise and fully rigorous review of the basic mathematics is in the monograph by C. Dellacherie and P.-A. Meyer [1, Ch. I-III].

1. Conditional expectation

1 (Definition)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, X a real random variable with finite expectation, $\mathrm{E}_\mu(|X|) < +\infty$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. A random variable $\widehat{X}$ is a *version of the conditional expectation of X given $\mathcal{G}$* if, and only if,

(1) $\widehat{X}$ is integrable and $\mathcal{G}$-measurable;
(2) for all bounded and $\mathcal{G}$-measurable random variable it holds

$$\mathrm{E}_\mu\left(G\widehat{X}\right) = \mathrm{E}_\mu(GX) \ .$$

As the equation $\mathrm{E}_\mu\left(G(\widehat{X} - X)\right) = 0$, $G \in \mathcal{L}^\infty(\mathcal{G})$, is linear in G and continuous under bounded pointwise convergence, it is enough to check it for random variables of the for $\mathbf{1}_C$, $C \in \mathcal{C}$, $\mathcal{C}$ π-system generating $\mathcal{G}$. [Monotone-Class Theorem [2, ¶3.14].]

2 (Almost sure equivalence)**.** If $\widehat{X}_1$, $\widehat{X}_2$, are two versions of the conditional expectation of X, then $\mathrm{E}_\mu\left(G(\widehat{X}_1 - \widehat{X}_2)\right) = 0$ i.e. $\widehat{X}_1 = \widehat{X}_2$ μ-almost-surely. [Take $G = \mathrm{sign}\left(\widehat{X}_1 - \widehat{X}_2\right)$ to get $\mathrm{E}_\mu\left(\left|\widehat{X}_1 - \widehat{X}_2\right|\right) = 0$.] More generally, if $X_1 = X_2$ μ-almost-surely, then $\widehat{X}_1 = \widehat{X}_2$ μ-almost-surely. We write $\mathrm{E}_\mu(X|\mathcal{G})$ to denote the μ-class of versions and, with abuse of notation, $\widehat{X} = \mathrm{E}_\mu(X|\mathcal{G})$. If $L^1(\mathcal{F}, \mu)$ is the vector space of classes μ-equivalent real random variables, there exists a mapping

$$L^1(\mathcal{F}, \mu) \ni X \mapsto \mathrm{E}_\mu(X|\mathcal{G}) \in L^1(\mathcal{G}, \mu) \ .$$

3. The *existence issue* i.e., the fact that the previous mapping is actually defined on all of $L^1(\mathcal{F}, \mu)$, is discussed in [2, ¶9.5]. We skip this topic, together with a related issue namely, the notion of μ-complete σ-algebra. Many proofs of existence are actually available, either based on some result of Functional Analysis, or based on results from advanced Measure Theory such as the Radon-Nikodým Theorem. Here, we are mainly

Date: Moncalieri, March 2017.

focused on either *computing* a version of the conditional expectation of a given random variable, or *checking* that a random variable is a version of the conditional expectation of some random variable. We have defined the conditional expectation for integrable random variables. It is possible to define the conditional expectation for positive random variables, see the comments below about properties of the conditional expectation.

4. *Projection property* Let $\mathcal{H}$ be a sub-σ-field of $\mathcal{G}$. Then $\mathrm{E}_\mu\left(\mathrm{E}_\mu\left(X|\mathcal{G}\right)|\mathcal{H}\right) = \mathrm{E}_\mu\left(X|\mathcal{H}\right)$. In fact, the conditional expectation $X \mapsto \mathrm{E}_\mu\left(X|\mathcal{F}\right)$ operator is a projection operator on $L^1(\mathcal{F},\mu)$. It is the transposed operator of the injection operator $\mathcal{L}^\infty(\mathcal{G}) \to \mathcal{L}^\infty(\mathcal{F})$.

5. Examples.

(1) If $\mathcal{G} = \{\varnothing, \Omega\}$, then $\mathrm{E}_\mu\left(X|\mathcal{G}\right) = \mathrm{E}_\mu\left(X\right)$.
(2) If $\mathcal{G} = \mathcal{F}$, then $\mathrm{E}_\mu\left(X|\mathcal{G}\right) = X$.
(3) Let $\{A_1, \ldots, A_n\}$ be a measurable partition of Ω and let $\mathcal{G} = \sigma(A_1, \ldots, A_n)$. Assume $\mu(A_j) \neq 0$, $j = 1, \ldots, n$. It holds

$$\mathrm{E}_\mu\left(X|\mathcal{G}\right) = \sum_{j=1}^{n} \frac{\int_{A_j} X \; d\mu}{\mu(A_j)} \mathbf{1}_{A_j} = \sum_{j=1}^{n} \mathrm{E}_\mu\left(X|A_j\right) \mathbf{1}_{A_j} \; .$$

(4) Let μ be a probability measure $(\Omega, \mathcal{F})$ and $p \cdot \mu$ a μ-absolutely-continuous finite measure. Then the restriction of the measure to $\mathcal{G}$ is $(p \cdot \mu)|_{\mathcal{G}} = \mathrm{E}_\mu\left(p|\mathcal{G}\right) \cdot (\mu|_{\mathcal{G}})$.

6 (Conditioning to a random variable)**.** Let $(S, \mathcal{S})$ be a measurable space, $Y \colon \Omega \to S$ a measurable mapping, and $\mathcal{Y} = \sigma(Y) = Y^{-1}(\mathcal{S})$. A real random variable is $\mathcal{Y}$-measurable if, and only if, it is of the form $\phi \circ Y$, where ϕ is a real random variable on $(S, \mathcal{S})$. [The "if" part follows from $(\phi \circ Y)^{-1}(\mathcal{B}) \subset \mathcal{Y}$; the "only if" part is true because of the Monotone Class Theorem, see [2, ¶3.14].] In this situation, the definition of conditional expectaion is rephrased as follows. A version of the conditional expectation of X given $\sigma(Y)$ is a μ-integrable real random variable of the form $\widehat{\phi}_{\mu,X} \circ Y$ such that for all bounded measurable $\phi \colon S \to \mathbb{R}$ it holds $\mathrm{E}_\mu\left(\phi(Y)\widehat{\phi}_{\mu,X}(Y)\right) = \mathrm{E}_\mu\left(\phi(Y)X\right)$. Notice that we could write this in terms of the joint distribution of the random variables X and Y as $\int \phi(y)\widehat{\phi}_{\mu,X}(y) \; \mu_Y(dy) = \int \phi(y)x \; \mu_{X,Y}(dxdy)$. An imprecise, but widely used, notation is $\phi_{\mu,X}(y) = \mathrm{E}_\mu\left(X|Y = y\right)$, which is called the *expected value of X, given $Y = y$.*

7. Special cases.

(1) If $X \perp\!\!\!\perp Y$ then $\mathrm{E}_\mu\left(X|\sigma(Y)\right) = \mathrm{E}_\mu\left(X\right)$. in fact,

$$\int \phi(y)x \; \mu_{X,Y}(dxdy) = \int \phi(y) \left(\int x \; \mu_X(dx) \right) \; \mu_Y(dy) \; .$$

(2) If $X \perp\!\!\!\perp Y$ then $\mathrm{E}_\mu\left(f(X,Y)|\sigma(Y)\right) = \int f(x,Y) \; \mu_X(dx)$. In this case we have

$$\int \phi(y)f(x,y) \; \mu_X \otimes \mu_Y(dxdy) = \int \phi(y) \left(\int f(x,y) \; \mu_X(dx) \right) \; \mu_Y(dy) \; .$$

(3) Let X, Y, be random variables in $\mathbb{R}^m$ such that $(X - Y) \perp\!\!\!\perp Y$. Then

$$\mathrm{E}_\mu\left(f(Y)|\sigma(Y)\right) = \mathrm{E}_\mu\left(f((X-Y)+Y)|\sigma(Y)\right) = \int f(s,Y) \; \mu_{(X-Y)}(ds) \; .$$

Cf. the Gaussian case.

(4) If $\mu_{X,Y}(dx,dy) = p_{X,Y}\cdot \nu_X \otimes \nu_Y$, then $\mu_Y = \left(\int p(x,y)\ \nu_X(dx)\right)\cdot \nu_Y(dy)$ and the characteristic equality becomes

$$\int \phi(y)\phi_X(y)\ \left(\int p(x,y)\ \nu_X(dx)\right)\cdot \nu_Y(dy) = \int \phi(y)\left(\int x\ p_{X,Y}\ \nu_X(dx)\right)\ \nu_Y(dy)\ ,$$

hence we can take

$$\widehat{\phi}_X(y) = \int x\ p_{X|Y}(x|y)\ \nu_X(dx), \quad p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_X(x)}\ .$$

8. *Properties.* All random variables are defined on the probability space $(\Omega,\mathcal{F},\mu)$ and $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$

(1) *Normalization.* $\mathrm{E}_\mu(\mathbf{1}|\mathcal{G}) = \mathbf{1}$.

(2) $\mathcal{G}$*-Linearity.* If $\mathrm{E}_\mu(X|\mathcal{G}) = \widehat{X}$ and $\mathrm{E}_\mu(Y|\mathcal{G}) = \widehat{Y}$, then $\mathrm{E}_\mu(AX+BY|\mathcal{G}) = A\widehat{X}+B\widehat{Y}$ μ-almost-surely if $A,B\in\mathcal{L}^\infty(\mathcal{G})$.

(3) *Positivity.* If $X\geqslant 0$ and $\mathrm{E}_\mu(X|\mathcal{G}) = \widehat{X}$, then $\widehat{X}\geqslant 0$. Linearity and positivity together imply monotonicity. [Hint: take $G = \mathbf{1}_{\{\widehat{X}\leqslant 0\}}$ in the characteristic property]

(4) Normalization, linearity and monotonicity together imply *Jensen inequality.* Assume $\Phi\colon \mathbb{R}\to\mathbb{R}$ and assume both X and $\Phi(X)$ are integrable. Let $x\mapsto a+bx\leqslant \Phi(x)$. Then $a+b\,\mathrm{E}_\mu(X|\mathcal{G})\leqslant \mathrm{E}_\mu(\Phi(X)|\mathcal{G})$. Chose a version $\widehat{X} = \mathrm{E}_\mu(X|\mathcal{G})$ Because of the convexity, for each $\omega\in\Omega$, there exists coefficients $a(\omega),b(\omega)$ such that $a(\omega)+b(\omega)\widehat{X}(\omega) = \Phi(\widehat{X}(\omega))$. We have shown that $\Phi(\mathrm{E}_\mu(X|\mathcal{G}))\leqslant \mathrm{E}_\mu(\Phi(X)|\mathcal{G})$. In particular, $\mathrm{E}_\mu(|X||\mathcal{G})^\alpha\leqslant \mathrm{E}_\mu(|X|^\alpha|\mathcal{G})$ if $\alpha\geqslant 1$.

(5) *Monotone convergence.* If $0\leqslant X_n\uparrow X$ and $\widehat{X}_n = \mathrm{E}_\mu(X_n|\mathcal{G})$, $n\in\mathbb{N}$, then random variable $\widehat{X}$ defined by $\widehat{X}_n\uparrow\widehat{X}$ is such that $\mathrm{E}_\mu\left(G\widehat{X}\right) = \mathrm{E}_\mu(GX)$ if $0\leqslant G\in\mathcal{L}^\infty(\mathcal{G})$. It follows immediatly from the monotone convergence for the expectation [Notice that here we are assuming each X_n to be 'integrable so that the conditional expectation is defined. This is not necessary if we define conditional expectation for non-negative random variable as it was for che expectation. We do not consider this generalization in this notes.] If moreover X happens to be integrable, then $\widehat{X} = \mathrm{E}_\mu(X|\mathcal{G})$.

(6) *Fatou lemma.* If $0\leqslant X_n$ and $\widehat{X}_n = \mathrm{E}_\mu(X_n|\mathcal{G})$, $n\in\mathbb{N}$, then $\wedge_{m\geqslant n}X_m\leqslant X_m$ if $m\geqslant n$, so that $\mathrm{E}_\mu(\wedge_{m\geqslant n}X_m|\mathcal{G})\leqslant \wedge_{m\geqslant n}\mathrm{E}_\mu(X_m|\mathcal{G})$. From the monotone convergence it follows $\mathrm{E}_\mu(G(\liminf_{n\to\infty}X_n))\leqslant \mathrm{E}_\mu(G(\liminf_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G})))$ if $G\in\mathcal{L}^\infty(\mathcal{G})$ and $G\geqslant 0$. If $\liminf_{n\to\infty}X_n$ is integrable, then we can write $\mathrm{E}_\mu(\liminf_{n\to\infty}X_n|\mathcal{G})\leqslant \liminf_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G})$.

(7) *Dominated convergence.* If in the fatou lemma we assume that the sequence $(X_n)_{n\in\mathbb{N}}$ is dominated by the integrable random variable Y, by considering the non-negative sequence $(Y-X_n)_{n\in\mathbb{N}}$ we can obtain the inequality

$$\mathrm{E}_\mu\left(\liminf_{n\to\infty}X_n\middle|\mathcal{G}\right)\leqslant \liminf_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G})\leqslant \limsup_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G})\leqslant \mathrm{E}_\mu\left(\limsup_{n\to\infty}X_n\middle|\mathcal{G}\right)\ .$$

If the sequence is convergent, then $\liminf_{n\to\infty}X_n = \lim_{n\to\infty}X_n = \limsup_{n\to\infty}X_n$ hence $\liminf_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G}) = \limsup_{n\to\infty}\mathrm{E}_\mu(X_n|\mathcal{G})$ and the sequence of conditional expectations is convergent to the expectation of the limit. The condition of positivity can be dropped by decomposing the positive and negative part of the sequence and the limit.

2. Conditional distribution

9 (Transition probability measure)**.** Given a product measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ a *transition* is a mapping $\mu_{1|2} \colon \mathcal{F}_1 \times \Omega_2$ such that

(1) for each $x_2 \in \Omega_2$ tha mapping $\mathcal{F}_1 \ni A_1 \mapsto \mu_{1|2}(A_1|x_2)$ is a probability measure on $(\Omega_1, \mathcal{F}_1)$ and
(2) for each $A_1 \in \mathcal{F}_1$ the mapping $\Omega_2 \ni x_2 \mapsto \mu_{1|2}(A_1|x_2)$ is $\mathcal{F}_2$-measurable.

10 (Integration of probability measures)**.** Given a transition $\mu_{1|2}$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ and a probability measure μ_2 on $(\Omega_2, \mathcal{F}_2)$, there exists a unique probability measure $\mu = \int \mu_{1|2}\, d\mu_2$ on the product measurable space such that for each positive or μ-integrable function $f \colon \Omega_2 \times \Omega_2 \ni (x_1, x_2) \mapsto f(x_1, x_2)$ it holds

$$\int f \, d\mu = \int \left(\int f(x_1, x_2) \, \mu_{1|2}(dx_1|x_2) \right) \, \mu_2(dx_2) \, .$$

The measure μ is characterised on functions of the form $f(x_1, x_2) = f_1(x_1) f_2(x_2)$ by

$$\int f_1 f_2 \, d\mu = \int \left(\int f_1(x_1) \, \mu_{1|2}(dx_1|x_2) \right) f_2(x_2) \, \mu_2(dx_2) \, .$$

[The proof is a simple variation of the argument for Fubini theorem.]

11 (Transition densities)**.** A simple case occurs when the transition has the form

$$\mu_{1|2}(A_1|x_2) = \int_{A_1} p_{1|2}(x_1|x_2) \, \nu_1(dx), \quad A_1 \in \mathcal{F}_1, x_2 \in \Omega_2$$

where $(x_1, x_2) \mapsto p_{1|2}(x_1|x_2)$ is measurable on the product space $(\Omega_1, \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ and $x_1 \mapsto p_{1|2}(x_!|x_2)$ is a ν_1-probability density for each $x_2 \in \Omega_2$. In such a case,

$$\int \left(\int f_1(x_1) \, \mu_{1|2}(dx_1|x_2) \right) f_2(x_2) \, \mu_2(dx_2) = \int \left(\int f_1(x_1) p_{1|2}(x_1|x_2) \nu_1(dx_1) \right) f_2(x_2) \, \mu_2(dx_2) = \iint f_1(x_1) f_2(x_2) p_{1|2}(x_1|x_2) \, \nu_1(dx_1) \mu_2(dx_2) \, ,$$

that is, $\mu = p_{1|2} \cdot \nu_1 \otimes \mu_2$. If moreover the second measure has itself a density, $\mu_2 = p_2 \cdot \nu_2$, then $\mu = (p_{1|2} \otimes p_2) \cdot \nu_1 \otimes \nu_2$

12 (Examples)**.**
(1) Let T_1, T_2 be independent and Exp(1). Then the distribution of T_1 given $T_1 + T_1 = t$ is uniform on $]0, t[$.
(2) If $(Y_1, Y_2) \sim \mathrm{N}_{n_1+n_2}(0, \Sigma)$, $\det \Sigma \neq 0$, find the conditional distribution of Y_1 given Y_2.
(3) If Y_1, Y_2 are independent and $\mathrm{N}(0, 1)$, find the distribution of (y_1, Y_2) given $Y_1^2 + Y_2^2$.

13 (Regular version of the conditional expectation)**.** With the notations above, denoting with X_1, X_2 the coordinate projection, the random variable $\widehat{f}(X_2) = \int f(x_1, X_2) \, \mu_{1|2}(dx_1|X_2)$ is a version of the conditional expectation $\mathrm{E}_\mu \left(f(X_1, X_2) | \sigma(X_2) \right)$, namely a *regular version*. In fact,

$$\mathrm{E}_\mu \left(f(X_1, X_2) g(X_2) \right) = \int \left(\int f(x_1, x_2) \, \mu_{1|2}(dx_1|x_2) \right) g(x_2) \, \mu_2(dx_2) = \mathrm{E}_\mu \left(\widehat{f}(X_2) g(X_2) \right) \, .$$

14 (Conditional independence (CI) of events)**.** Conditional independence is a key property in Statistics e.g., in Graphical Models, in Stochastic Processes, in Markov processes, in Random Fields, in Machine Learning.

The nonzero events A, B, C are such that A and C are independent given B, $A \perp\!\!\!\perp C \,|\, B$, if each one of the following equivalent conditions are satisfied:

$$\begin{aligned}
\mathrm{P}\,(A \cap C|B) &= \mathrm{P}\,(A|B)\,\mathrm{P}\,(C|B) \quad \text{(CI)} \\
\mathrm{P}\,(A \cap B \cap C) &= \mathrm{P}\,(A|B)\,\mathrm{P}\,(C|B)\,\mathrm{P}\,(B) \quad \text{(J-CI)} \\
\mathrm{P}\,(A|B \cap C) &= \mathrm{P}\,(A|B) \quad \text{(M)} \\
\mathrm{P}\,(A \cap B \cap C) &= \mathrm{P}\,(A|B)\,\mathrm{P}\,(B|C)\,\mathrm{P}\,(C) \quad \text{(J-M)} \\
\mathrm{P}\,(A \cap B \cap C)\,\mathrm{P}\,(B) &= \mathrm{P}\,(A \cap B)\,\mathrm{P}\,(B \cap C) \quad \text{(A)}
\end{aligned}$$

In fact, the first four equalities become the fifth one if the conditional probabilities are computed in terms of joint probabilities. The algebraic form (A) can be written in terms of indicator functions as

$$\mathrm{E}\,(\mathbf{1}_A \mathbf{1}_B \mathbf{1}_C)\,\mathrm{E}\,(\mathbf{1}_B) = \mathrm{E}\,(\mathbf{1}_A \mathbf{1}_B)\,\mathrm{E}\,(\mathbf{1}_B \mathbf{1}_C) \ .$$

which shows the bi-linearity in $\mathbf{1}_A$ and $\mathbf{1}_C$. For example, writing $\mathbf{1}_A = \mathbf{1} - \mathbf{1}_{A^c}$ one gets $A^c \perp\!\!\!\perp C \,|\, B$.

15 (Conditional Independence for random variables)**.** *Random variables Y_1, Y_3 are conditionally independent given the random variable Y_2,* $Y_1 \perp\!\!\!\perp Y_3 \,|\, Y_2$ if each one of the following equivalent conditions are satisfied. If f_i, $i = 1, \ldots, 3$, are bounded,

$$\begin{aligned}
\mathrm{E}\,(f_1(Y_1) f_3(Y_3)|Y_2) &= \mathrm{E}\,(f_1(Y_1)|Y_2)\,\mathrm{E}\,(f_3(Y_3)|Y_2) \\
\mathrm{E}\,(f_1(Y_1)|Y_2, Y_3) &= \mathrm{E}\,(f_1(Y_1)|Y_2)
\end{aligned}$$

Let us prove the equivalence. The second one holds if, and only if, for all bounded $f_2(Y_2), f_3(Y_3)$

$$\mathrm{E}\,(f_1(Y_1) f_2(Y_2) f_3(Y_3)) = \mathrm{E}\,(\mathrm{E}\,(f_1(Y_1)|Y_2)\, f_2(Y_2) f_3(Y_3))$$

The LHS is equal to

$$\mathrm{E}\,(\mathrm{E}\,(f_1(Y_1) f_3(Y_3)|Y_2)\, f_2(Y_2))$$

and the RHS is equal to

$$\mathrm{E}\,(\mathrm{E}\,(f_1(Y_1)|Y_2)\, f_2(Y_2)\,\mathrm{E}\,(f_3(Y_3)|Y_2))\,.$$

It follows that the first equation holds. By reversing the computation we get the other implication.

When a regular version of the conditional expectation given Y_2 is available, then conditional independence is equivalent to the product form of the transition.

16 (Markov process)**.** A stochastic process $Y_1, \ldots, Y_N$ is a *Markov Process* if $(Y_1, \ldots, Y_k) \perp\!\!\!\perp Y_k, \ldots, Y_N \,|\, Y$ $k = 1, 2, \ldots, N$. Equivalently,

$$\mathrm{E}\,(f(Y_k, \ldots, Y_N)|Y_1, \ldots, Y_k) = \mathrm{E}\,(f(Y_k, \ldots, Y_N)|Y_k)\,.$$

References

[1] Claude Dellacherie and Paul-André Meyer, *Probabilities and potential*, North-Holland Mathematics Studies, vol. 29, North-Holland Publishing Co., Amsterdam-New York; North-Holland Publishing Co., Amsterdam-New York, 1978. MR 521810

[2] David Williams, *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 1991.

Collegio Carlo Alberto
E-mail address: giovanni.pistone@carloalberto.org