ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 01/30/20

Lecture 4: Random Variables (cont.) and Conditional Probabilities

Lecturer: Prof. Ziv Goldfeld

Scriber: Sueda Taner, *Net ID*: st939

Assistant Editor: Kia Khezeli

4.1 Expectation and Variance

We now focus on integration of measurable functions (random variables) with respect to their probability law.

Definition 4.1 (Expectation) *Let $X : \Omega \to \mathbb{R}^d$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The expectation of X is*

$$\mathbb{E}[X] := \int_{\mathbb{R}^d} x \mathrm{d}\mathbb{P}_X(x).$$

Example 4.1 (Special Cases) *For a discrete random variable X, the expected value is given by*

$$\mathbb{E}[X] = \sum_{x \in \mathrm{supp}(\mathbb{P}_X)} x p_X(x).$$

For a continuous random variable Y, its expected value is given by

$$\mathbb{E}[X] = \int_{\mathrm{supp}(\mathbb{P}_X)} x f_X(x) \mathrm{d}x.$$

Consider computing the expected value of a function of a random variable. Specifically, given a Borel measurable function g, we want to evaluate $\mathbb{E}[g(X)]$. The ad hoc approach here is to define $Y = g(X)$, obtain its probability law $\mathbb{P}_Y$, and calculate $\mathbb{E}[Y]$. The following proposition gives a simpler way to compute $\mathbb{E}[g(X)]$ with the original probability law $\mathbb{P}_X$ as the underlying measure.

Proposition 4.1 (Expectation of a function) *Let $X : \Omega \to \mathbb{R}^d$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a Borel measurable function. Then*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x) \mathrm{d}\mathbb{P}_X(x).$$

Exercise 4.1 *Show that $Y : \Omega \to \mathbb{R}^{d'}$ defined as $Y = g(X)$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.*

Using Proposition 4.1, we define the variance of random variables.

Definition 4.2 (Variance) *Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The variance of X is*

$$\mathsf{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

4.2 Law of Large Numbers

The law of large numbers (LLN) quantifies the fact that averaging many identically and independently distributed copies of a random variable gives a good approximation of its expectation.

Proposition 4.2 *Let $X_1, \ldots, X_n$ be a sequence of $\mathbb{R}$-valued i.i.d. random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following hold:*

1. *Weak LLN: If* $\mathbb{E}[|X_1|] < \infty$*, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| > \epsilon\right) = 0, \forall \epsilon > 0.$$

2. *Weak LLN for Functions: Let* $f : \mathbb{R} \to \mathbb{R}$ *be Borel a measurable function such that* $\mathbb{E}[|f(X_1)|] < \infty$*. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_1)]\right| > \epsilon\right) = 0, \forall \epsilon > 0.$$

3. *Uniform Weak LLN: Let* $f_1, \dots, f_k : \mathbb{R} \to \mathbb{R}$ *be measurable functions with* $\mathbb{E}[|f_l(X_1)|] < \infty, \forall l \in \{1, \dots, k\}$*. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{l=1}^{k} \left\{\left|\frac{1}{n}\sum_{i=1}^{n} f_l(X_i) - \mathbb{E}[f_l(X_1)]\right|\right\} > \epsilon\right) = 0, \forall \epsilon > 0.$$

Remark 4.1 (Weak vs. Strong LLN) *The name 'weak' LNN refers to the fact that the above results account for convergence in probability. There are stronger versions of the LLN, such as the* strong LLN*, which guarantees almost sure convergence.*

4.3 Conditional Probability and Transition Kernels

Transition kernels formulate the notion of a random transformation of a random variable. Later, such kernels will serve as our models from noisy communication channels.

Definition 4.3 (Transition Kernel) *Let* $(\mathcal{X}, \mathcal{F})$ *and* $(\mathcal{Y}, \mathcal{G})$ *be two measurable spaces. A function* $\kappa(\cdot|\cdot) : \mathcal{G} \times \mathcal{X} \to \mathbb{R}$ *is a transition kernel from* $(\mathcal{X}, \mathcal{F})$ *to* $(\mathcal{Y}, \mathcal{G})$ *if*

1. $\kappa(\cdot|x)$ *is a probability measure on* $(\mathcal{Y}, \mathcal{G})$*, i.e.,* $\kappa(\cdot|x) \in P(\mathcal{Y}), \forall x \in \mathcal{X}$*.*

2. $\kappa(B|\cdot) : \mathcal{X} \to \mathbb{R}$ *is* $\mathcal{F}$*-measurable, i.e. a random variable with respect to* $(\mathcal{X}, \mathcal{F}), \forall B \in \mathcal{G}$*.*

Remark 4.2 (Notation) *We often denote a transition kernel from* $(\mathcal{X}, \mathcal{F})$ *to* $(\mathcal{Y}, \mathcal{G})$ *by* $P_{Y|X}$*, to highlight the underlying spaces.*

Remark 4.3 (Transition Kernel as Transformation) *Transition kernels from* $\mathcal{X}$ *to* $\mathcal{Y}$ *can be thought of random as transformations of a distribution* $P_X \in \mathcal{P}(\mathcal{X})$ *into another distribution* $P_Y \in \mathcal{P}(\mathcal{Y})$*. Indeed, defining,*

$$P_Y(B) := \mathbb{E}_{P_X}[\kappa(B|X)] = \int_{\mathcal{X}} \kappa(B|x) \mathrm{d}P_X(x) \quad B \in \mathcal{G},$$

the following proposition shows that $\kappa(\cdot|\cdot)$ *transforms* P_X *into* P_Y*.*

Proposition 4.3 (Induced Probability Measure) P_Y *as defined above is a probability measure on* $(\mathcal{Y}, \mathcal{G})$*.*

Proof: It suffices to show the following:

(i) Normalization: $P_Y(\mathcal{Y}) = \int_{\mathcal{X}} \underbrace{\kappa(\mathcal{Y}|x)}_{=1 \text{ by definition}} \mathrm{d}P_X(x) = \int_{\mathcal{X}} \mathrm{d}P_X(x) = 1$.

(ii) σ-additivity: Let $\{B_n\}_{n=1}^{\infty}$ be a sequence of disjoint $\mathcal{G}$-measurebale sets. We have

$$\begin{aligned} P_Y\left(\bigcup_{n=1}^{\infty} B_n\right) &= \int \kappa\left(\bigcup_{n=1}^{\infty} B_n \middle| x\right) \mathrm{d}P_X(x) \\ &= \int_{\mathcal{X}} \sum_{n=1}^{\infty} \kappa(B_n|x) \mathrm{d}P_X(x) \\ &\overset{(*)}{=} \sum_{n=1}^{\infty} \int_{\mathcal{X}} \kappa(B_n|x) \mathrm{d}P_X(x) \\ &= \sum_{n=1}^{\infty} P_Y(B_n). \end{aligned}$$

where $(*)$ is the Fubini-Tonelli theorem.

■

Proposition 4.4 (Transition Kernels vs. Joint Probability Measures) *Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces, $P_X \in \mathcal{P}_{\mathcal{F}}(\mathcal{X})$ and $\kappa(\cdot|\cdot)$ be a transition kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$. Then there exists a unique probability measure P_{XY} on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$, where $\mathcal{F} \otimes \mathcal{G} := \sigma\big(\{A \times B : \ A \in \mathcal{F}, \ B \in \mathcal{G}\big)$ is the product σ-algebra, such that*

$$P_{XY}((A, B)) = \int_A \kappa(B|x) \mathrm{d}P_X(x), \quad \forall (A, B) \in \mathcal{F} \otimes \mathcal{G}.$$

Conversely, given a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G}, P_{XY})$, there exists a unique pair comprising $P_X \in \mathcal{P}_{\mathcal{F}}(\mathcal{X})$ and a transition kernel $\kappa(\cdot|\cdot)$ such that the above equation holds.

Remark 4.4 (Notation) *Henceforth, we will write $P_{XY} = P_X \cdot P_{Y|X} = P_Y \cdot P_{X|Y}$ while understanding this decomposition with respect to the above proposition.*

4.4 Conditional Expectation

We next define conditional expectations (given a random variable or its realization). The underlying machinery in these definitions is the transition kernel. With respect to Remarks 4.2-4.4, we proceed with random variable notation.

Definition 4.4 (Conditional Expectation) *Let $P_{Y|X}(\cdot|\cdot)$ be a transition kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$. Define*

1. $\mathbb{E}\left[Y|X = x_0\right] := \int_{\mathcal{Y}} y \mathrm{d}P_{Y|X}(y|x_0), \quad x_0 \in \mathcal{X};$
2. $\mathbb{E}\left[Y|X\right] := \int_{\mathcal{Y}} y \mathrm{d}P_{Y|X}(y|\cdot).$

Proposition 4.5 (Relation Between Conditional Expectations) *Let $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that decomposes as $P_{XY} = P_X P_{Y|X}$ (see Remark 4.4). Then there exists a unique measurable function $h : \mathcal{X} \to \mathcal{Y}$ such that*

i) $h(x) = \mathbb{E}\left[Y|X = x\right], \quad \forall x \in \mathcal{X};$

ii) $h(X) = h \circ X = \mathbb{E}\left[Y|X\right]$ *almost surely.*

Example 4.2 *Let $X = \begin{cases} 1 & w.p. \quad 1/2 \\ -1 & w.p. \quad 1/2 \end{cases}$ be independent of $Z \sim \mathcal{N}(0, \sigma^2)$, and define $Y = X + Z$. What is $\mathbb{E}\left[X|Y\right]$? To compute it, we start from $\mathbb{E}\left[X|Y = y\right]$ and identify the function h. Then we can obtain $\mathbb{E}\left[X|Y\right]$ by composing h with Y.*

Exercise 4.2 *Show that* $\mathbb{E}\left[X|Y=y\right] = \tanh\left(\frac{y}{\sigma^2}\right)$.

Based on the above, we conclude that $\mathbb{E}\left[X|Y\right] = h(Y) = \tanh\left(\frac{Y}{\sigma^2}\right)$.

Theorem 1 (Law of Total Expectation) *Let* $(X,Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ *that decomposes as* $P_{XY} = P_X P_{Y|X}$ *(see Remark 4.4). Then*

$$\mathbb{E}_{P_Y}\left[Y\right] = \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[Y|X\right]\right] = \int_{\mathcal{X}} \mathbb{E}\left[Y|X=x\right] \mathrm{d}P_X(x)$$

Proof: Consider:

$$\begin{aligned}
\mathbb{E}_{P_Y}\left[Y\right] &= \int_{\mathcal{Y}} y \mathsf{d}P_Y(y) \\
&\overset{(a)}{=} \int_{\mathcal{Y}}\int_{\mathcal{X}} y \mathsf{d}P_{Y|X}(y|x)\mathsf{d}P_X(x) \\
&\overset{(b)}{=} \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} y \mathsf{d}P_{Y|X}(y|x)\right) \mathsf{d}P_X(x) \\
&\overset{(c)}{=} \int_{\mathcal{X}} \mathbb{E}\left[Y|X=x\right] \mathsf{d}P_X(x) \\
&= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[Y|X\right]\right],
\end{aligned}$$

where (a) is because $P_Y(B) := \int_{\mathcal{X}} P_{Y|X}(B|x)\mathsf{d}P_X(x)$ for any measurable set B, (b) is the Fubini-Tonelli theorem, and (c) is the definition of $\mathbb{E}\left[Y|X=x\right]$. This concludes the proof. ■