

## Lecture 4: Basics of Probability Theory 1

Lecturer: Yiqiao Zhong

Last update: 15:24:36, 2024-09-20

*Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.*

In this lecture, we introduce random variables, basic inequalities, and their use in some problems.

## 1 Expectation

For a random variable  $X$  on a probability space  $(\Omega, \mathcal{F}, P)$ , we say  $X$  is integrable (or has finite expectation) iff  $\int |X| dP < \infty$ , and we denote  $\mathbb{E} X = \int X dP$ .

**Change of variables.** Recall that a random variable  $X$  induces a measure, or distribution/law, defined by

$$P \circ X^{-1}(B) = P(X^{-1}(B)), \quad \text{for any } B \in \mathcal{B}.$$

In the context of measures, we may use either  $dP$  or  $d\mu_X$ : for a Borel function  $f$ , we have

$$\mathbb{E}[f(X)] = \int f d\mu_X$$

whenever either integral exists. The proof is a simple [exercise](#).

As a consequence, we have the following observation: a distribution  $\mu_X$  and a cdf  $F_X$  are like twins,

$$F_X(b) - F_X(a) = \int_{(a,b]} \mathbf{1}(a < X \leq b) dP = \mu_X((a, b])$$

**Example 1.** A mix of discrete measure and absolutely continuous measure. Suppose that  $\mu_0$  is a discrete measure supported on  $\{0, 1, 2, \dots\}$  with  $\mu_0(a) = p_a$  where  $p_0, p_1, p_2, \dots$  are positive numbers satisfying  $\sum_{j \geq 0} p_j = 1$ . Also suppose  $h \geq 0$  a pdf. Consider the following measure

$$\mu(B) = \frac{1}{2} \mu_0(B) + \frac{1}{2} \int_B h(x) dx.$$

We know that there is no pdf or pmf for this mixed measure. But the cdf always exists and in this case, the cdf is continuous except for locations at  $\{0, 1, 2, \dots\}$  where cdf has jumps.

**Decomposition of measures.** The above example raises the following question—is there a principled way of decomposing any measure defined on the Borel sigma-algebra? Let  $(\mathbb{R}, \mathcal{B}, \mu)$  be a probability space, then we have the following decomposition of measure.

$$\mu = \mu_{\text{dis}} + \mu_{\text{cont}} + \mu_{\text{sing}}, \tag{1}$$

where

- $\mu_{\text{dis}}$  is the discrete measure, which is supported on a finite set or countable set  $D$ , i.e.,  $\mu_{\text{dis}}(D^c) = 0$ . This measure causes the jumps (discontinuity points) in the cdf  $F_X$ ; conversely, if there are jumps in the cdf, then we can find a discrete measure in the decomposition.
- $\mu_{\text{cont}}$  is the absolutely continuous part w.r.t. Lebesgue measure; so by RN derivative, we have a (Lebesgue) pdf  $f$  with  $\mu_{\text{cont}}(A) = \int_A f \, dx$  for any Borel set  $A$ . Most continuous random variables in statistics have this continuous measure solely.
- $\mu_{\text{sing}}$  is the singular part. Formally, it is a *singular measure* w.r.t.  $\mu_{\text{dis}} + \mu_{\text{cont}}$ . We rarely encounter this measure in statistics, but it is good to know that there exist such measures. As a consequence, even when we have a continuous cdf  $F_X$  (no discrete measure), we cannot guarantee a Lebesgue pdf  $f$  such that  $F_X(A) = \int_A f(x) \, dx$ .

**Useful identities.** If  $f : [0, \infty)$  is a nondecreasing and continuously differentiable function with  $f(0) = 0$ , we may calculate  $\mathbb{E}[f(|X|)]$  in a different way.

$$\mathbb{E}[f(|X|)] = \int_0^\infty f'(t)P(|X| > t) \, dt. \quad (2)$$

In particular,

$$\mathbb{E}[|X|] = \int_0^\infty P(|X| > t) \, dt,$$

which connects the expectation with the tail probability of  $|X|$ . The identity (2) can be proved by Fubini's theorem ([exercise](#)).

**Basic examples.** If  $X \sim N(0, 1)$ , then  $X^2$  is chi-squared distributed denoted by  $\chi_1^2$  see [2][Example 1.14].

Moreover, if  $X \sim N(0, I_2)$  is a two-dimensional standard normal variable, then  $\|X\|^2$  is an exponential random variable ([exercise](#)).

## 2 Functional perspective and inequalities

For  $p \in [1, \infty]$ , we can define the norm of a random variable as long as the expectation exists<sup>1</sup>.

$$\|X\|_p = \begin{cases} \{\mathbb{E}[|X|^p]\}^{1/p}, & 1 \leq p < \infty \\ \inf\{M : P(|X| > M) = 0\}, & p = \infty \end{cases}$$

It is easy to check that this definition indeed qualifies as a norm<sup>2</sup> by Lecture 1 ([exercise](#)).

The standard inequalities such as Hölder's inequality holds. The following shows the celebrated Cauchy-Schwarz inequality.

$$\mathbb{E}[XY] = \|XY\|_1 \leq \|X\|_2 \cdot \|Y\|_2$$

**Example 2.** As a consequence of Hölder's inequality, if  $X$  has finite  $q$ -th moment, that is  $\mathbb{E}[|X|^q] < \infty$ , then we have  $\mathbb{E}[|X|^p] < \infty$  for  $1 \leq p < q$ .

**Example 3.** Recall that  $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$ . As a consequence of Cauchy-Schwarz inequality, we have

$$|\text{Cov}(X, Y)| \leq [\text{Var}(X)\text{Var}(Y)]^{1/2}.$$

<sup>1</sup>When  $p = \infty$ , we look for *essential supremum* since changing a zero-measure set should not change the value of the norm

<sup>2</sup>Rigorously speaking, we view  $X = Y$  a.s. as the same random variable (using the notion of equivalent classes), so a zero random variable in this sense is the zero element in functional analysis.

**Theorem 1** (Jensen's inequality). Suppose that  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  is a convex function and  $\mathbb{E}X, \mathbb{E}[\varphi(X)]$  both exist. Then

$$\mathbb{E}[\varphi(X)] \geq \varphi[\mathbb{E}X]$$

Moreover, if  $\varphi$  is strictly convex, then  $\mathbb{E}[\varphi(X)] = \varphi[\mathbb{E}X]$  implies  $X = \mathbb{E}[X]$  a.s.

The next result is very simple but widely used.

**Theorem 2** (Markov's inequality). Let  $X$  be a nonnegative random variable and  $a > 0$  be a scalar. Then,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

A basic fact about expectation (or abstract integral) is that if  $X \leq Y$ , then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$  (check from definition; [exercise](#)). Then, the proof is based on the simple observation

$$aP(X \geq a) \leq \mathbb{E}[X\mathbf{1}\{X \geq a\}] \leq \mathbb{E}[X].$$

**Almost sure convergence.** We say  $X_n$  converges almost surely to  $X$  if

$$P(\lim_n X_n(\omega) = X(\omega)) = 1$$

or simply  $\lim_n X_n(\omega) = X(\omega)$  a.s., or even simpler  $\lim_n X_n = X$ . There is little difference between almost sure convergence and pointwise convergence, since we can always modify a set (or countably many sets) of zero measure without changing the probabilities.

The inequalities from abstract integration apply here. For example, if  $\lim_{n \rightarrow \infty} X_n = X$  a.s. and  $|X_n| \leq Y$  where  $\mathbb{E}Y < \infty$ , then  $\lim_n \mathbb{E}[X_n] = \mathbb{E}X$ .

### 3 Useful examples

**Gradient of log density.** Suppose that  $f(x) \geq 0$  is a pdf with a bounded support (the set  $\{x : f(x) > 0\}$  is bounded) in  $\mathbb{R}^d$ . Let  $X$  be a random variable with such a pdf. Then

$$\mathbb{E}[\nabla \log f(X)] = 0.$$

This identity follows from the exchangeability of  $\int$  and  $\nabla$  ([exercise](#): what theorem we call for this exchangeability?).

**Moment method for random graphs.** In combinatorics and discrete probability, a simple analysis involves the calculation of the first moment. Consider the **Erdős-Rényi** graph from  $(G, E) \sim \mathcal{G}(n, p)$ , where we have  $n$  vertices and  $A_{ij} \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$  is an independent Bernoulli random variable representing whether there is each edge  $A_{ij} = 1$  or not  $A_{ij} = 0$ . There is a triangle in  $G$  with vertices  $i, j, k$  iff  $A_{ij} = A_{jk} = A_{ki} = 1$ .

Suppose that  $p = p(n) = n^{-\alpha}$  where  $\alpha > 0$  is a constant. What is the chance that we can find at least one triangle in a random graph drawn from  $\mathcal{G}(n, p)$ ? We can answer this question in the large  $n$  regime. Let  $B_{i,j,k}$  be the event that there is a triangle formed by vertices  $i, j, k$ . Let  $N(G)$  be the number of triangles in  $G$ . Then

$$\begin{aligned} P(N(G) \geq 1) &\stackrel{(i)}{\leq} \mathbb{E}[N(G)] = \mathbb{E}\left[\sum_{1 \leq i < j < k \leq n} \mathbf{1}\{i, j, k \text{ form a triangle}\}\right] \\ &= \sum_{1 \leq i < j < k \leq n} P(i, j, k \text{ form a triangle}) \\ &\stackrel{(ii)}{=} \binom{n}{3} p_n^3 \\ &= \frac{1 + o(1)}{6} (np_n)^3 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

where (i) is due to Markov's inequality and (ii) is due to independence. The last term converges to 0 if  $\alpha > 1$  and diverges to  $\infty$  if  $\alpha < 1$ . We may conclude that if  $\alpha > 1$ ,

$$P(\text{there exists no triangle in } G) = 1 - o(1) \quad \text{as } n \rightarrow \infty.$$

Actually,  $\alpha_c = 1$  is a critical threshold: further analysis by *second moment calculation* can show that

$$P(\text{there exists no triangle in } G) = \begin{cases} 1 - o(1) & \alpha > 1 \\ o(1) & \alpha < 1 \end{cases}$$

**KL divergence.** Suppose that  $P, Q$  are two probability measures on  $(\mathbb{R}^d, \mathcal{B}^d)$  and  $P \ll Q$ . A fundamental notion in information theory is Kullback–Keibler (KL) divergence, defined by

$$D_{\text{KL}}(P\|Q) = \int \log \frac{dP}{dQ} dP.$$

Here  $\frac{dP}{dQ}$  is the RN derivative. In the special case where both  $P$  and  $Q$  have Lebesgue pdf  $f$  and  $g$ , we have

$$D_{\text{KL}}(P\|Q) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

(i) We can show that  $D_{\text{KL}}(P\|Q) \geq 0$ . Denote  $h(x) = x \log x$  and the RN derivative as  $\varphi = dP/dQ$ . By convention, the composition of functions  $h$  and  $\varphi$  is denoted by  $h \circ \varphi$ . So, by change of variable,

$$\begin{aligned} D_{\text{KL}}(P\|Q) &= \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ = \int h \circ \varphi dQ \\ &\stackrel{(i)}{\geq} h\left(\int \varphi dQ\right) \stackrel{(ii)}{=} h\left(\int dP\right) = h(1) = 0 \end{aligned}$$

where we used Jensen's inequality in (i) and change of variable again in (ii). Moreover,  $D_{\text{KL}}(P\|Q) = 0$  iff  $P = Q$  a.s. KL divergence measures how close two distributions are, though it is not a metric (asymmetric).

(ii) If  $P$  and  $Q$  are supported on a finite set  $\{1, 2, \dots, N\}$ , and  $Q$  is a uniform distribution over  $\{1, 2, \dots, N\}$ , then, denoting the pmf of  $P$  by  $p$

$$D_{\text{KL}}(P\|Q) = \log N - H(P), \quad \text{where } H(P) = \sum_{i=1}^N p(i) \log \frac{1}{p(i)}.$$

The quantity  $H(P)$  is known as the **entropy** of  $P$ . If  $P$  is the Bernoulli distribution with parameter  $\gamma$ , we have seen in Lecture 1 that  $H(P) = h(\gamma) \log(1/\gamma) + (1 - \gamma) \log(1/(1 - \gamma))$ .

(iii) In machine learning, a widely used loss function for classification tasks is **cross-entropy**. Given a target classification probability  $p$  over  $K$  classes  $\{1, 2, \dots, K\}$ , we want to find a classifier with output probability  $q$  being close to  $p$ , as measured by the loss:

$$H(p, q) = \mathbb{E}_{X \sim p}[-\log q(X)] = H(p) + D_{\text{KL}}(p\|q)$$

Equivalently, we want to find a prediction probability  $q$  so that  $D_{\text{KL}}(p\|q)$  is as small as possible. Usually, we obtain  $q$  by applying the **Softmax function** to real-valued predictions: given the output of an algorithm (e.g., neural network)  $x \in \mathbb{R}^K$ ,

$$q(k) = \frac{e^{x_k}}{\sum_{k'=1}^K e^{x_{k'}}}, \quad k = 1, 2, \dots, K$$

Cross-entropy with softmax is actually (slightly generalized) logistic loss for binary classification ([exercise](#)).

**Collinearity.** In regression analysis, we have observations of  $p$  attributes or features. If we treat these features as random variables  $X = (X_1, X_2, \dots, X_p)$ , then our data are realized values of these random variables, which are themselves random variables. Suppose that  $a_1 X_1 + \dots + a_p X_p = 0$  for some coefficients  $a_1, \dots, a_p$ . This may happen if we collect the same feature twice or if we include all levels in categorical data analysis. If this happens, then the support of the measure  $\mu_X$  is in a linear subspace.

Denote the covariance matrix of  $X$  by  $\Sigma$  (assuming second moments exist). If  $\text{rank}(\Sigma) = r < p$ , then we may apply an eigendecomposition of  $\Sigma$ :  $\Sigma = U \Lambda U^\top$  where  $U \in O(p, r)$ , and then

$$\tilde{X} = U^\top X$$

This is a simple example of **dimension reduction** via PCA, where we remove uninformative directions of high-dimensional data. By definition, for any vector  $u$  orthogonal to  $U$ , i.e.,  $U^\top u = 0$ , we must have  $\text{Var}(u^\top X) = u^\top \Sigma u = 0$ .

Note that in practice, we don't have excess to the distribution of  $X$ , so we approximate  $\Sigma$  by the **empirical covariance** obtained from  $n$  observations  $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$  as column vectors,

$$\hat{\Sigma} = \frac{1}{n} X^{(i)} (X^{(i)})^\top.$$

Applying PCA to the data (eigendecomposition on  $\hat{\Sigma}$ ) helps us find the directions with the most variance and discard the directions with no or little variance.

## 4 Further readings

**More on measure decomposition.** Let  $\mu_1$  and  $\mu_2$  be two measures on  $\mathcal{B}(\mathcal{R})$  the collection of Borel sets. We say  $\mu_1$  is singular w.r.t.  $\mu_2$  (written  $\mu_1 \perp \mu_2$ ) iff there is a set  $E \in \mathcal{B}(\mathcal{R})$  s.t.  $\mu_1(E) = 0$  and  $\mu_2(E^c) = 0$ .

Take  $\mu_1 = \mu_{dis} + \mu_{cont}$ ,  $\mu_2 = \mu_{sing}$  in Eq.(1), we see that for any measurable set  $A = (A \cap E) \cup (A \cap E^c)$ ,

$$\mu(A) = (\mu_{dis} + \mu_{cont})(A \cap E^c) + \mu_{sing}(A \cap E).$$

This implies that  $E$  is the support for  $\mu_{sing}$ ,  $E^c$  is the support for  $\mu_{dis} + \mu_{cont}$ . The main part and the singular part are taking effect in mutually exclusive events.

- One simple consequence: note that for any  $\mu_{dis}(\{a\}) > 0$ ,  $\{a\} \subset E^c$  because  $E^c$  is the largest possible set for  $\mu_{dis}$  to have positive value,  $a \in E^c$  and thus  $\mu_{sing}(\{a\}) = 0$ .
- Eq. 1 is actually a by-product of the famous Radon-Nykodym theorem. One may refer to [1][Thm 6.10] for more details.

One example of singular measures is the Cantor measure; see the link for example. More details about measure decomposition can be found via the link or Chapter 6 of Rudin's book [1].

**$L_p$  convergence.** Given a probability space  $(\Omega, \mathcal{F}, P)$ , we can treat random variables on this space that differ only on a zero-measure set as elements in the same equivalence class. Under this view, we can check that  $\|\cdot\|_p$  is a norm so we have a normed space. Further, it is a complete space (thus Banach space) under this norm. Given a sequence  $(X_n)_{n \geq 1}$ , it converges to a limit  $X$  under  $L_p$  norm iff

$$\lim_{n \rightarrow \infty} \|X_n - X\|_p = 0, \quad \text{also denoted by } X_n \xrightarrow{L_p} X.$$

Some of the convergence properties can be found in Section 1.5.1 of [2].

**More on KL divergence and entropy.** A standard introduction to information theory is [3]. A quick reference for the results in this lecture note is the wikipedia.

## References

- [1] W. Rudin. *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Education, 1987.
- [2] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.
- [3] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.