Math 525: Lecture 5

January 18, 2018

1 Series (review)

Definition 1.1. A sequence $(a_n)_n \subset \mathbb{R}$ *converges* to a point $L \in \mathbb{R}$ (written $a_n \to L$ or $\lim_{n\to\infty} a_n = L$) if for each $\epsilon > 0$, we can find N such that $|a_n - L| < \epsilon$ for all $n \geq N$. If the sequence does not converge to any point in $\mathbb{R}$, we say it *diverges.*

In our case, we always use "converge" to mean "converges to a point in $\mathbb{R}$". Depending on the context, sometimes people will be talking about convergence in other spaces (e.g., if $a_n \to \infty$, one might say the sequence converges to a point in $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$: this is a perfectly valid use of the terminology).

Definition 1.2. Let $(a_n)_{n=1}^{\infty} \subset \mathbb{R}$ be a sequence. Then...

- We say the series $\sum_{n=1}^{\infty} a_n$ *converges* (written $\sum_{n=1}^{\infty} a_n < \infty$) if the sequence of partial sums $(s_N)_{N=1}^{\infty}$ defined by $s_N = \sum_{n=1}^{N} a_n$ converges. Otherwise, we say it *diverges.* In the convergent case, we define $\sum_{n=1}^{\infty} a_n = \lim_{N\to\infty} s_N$.
- The series $\sum_{n=1}^{\infty} a_n$ *converges absolutely* if the series $\sum_{n=1}^{\infty} |a_n|$ converges.

Proposition 1.3. *If a series converges* ***absolutely*** *to $L \geq 0$, the series converges to a number in $[-L, L]$.*

Proof. First, note that if $\sum_n |a_n| \to L$, then

$$-L \leftarrow -\sum_{n=1}^{N} |a_n| \leq \sum_{n=1}^{N} a_n \leq \sum_{n=1}^{N} |a_n| \to L. \qquad \square$$

The remainder of the proof requires knowledge of Cauchy sequences (if you are not familiar with them, you can safely skip this proof). Suppose $\sum_{n=1}^{\infty} a_n$ converges absolutely. Define $s_N = \sum_{n=1}^{N} a_n$ and $S_N = \sum_{n=1}^{N} |a_n|$. Then, for $N > M$,

$$|s_N - s_M| = |a_N + a_{N+1} + \cdots + a_{M+1}| \leq |a_N| + |a_{N+1}| + \cdots + |a_{M+1}| = |S_N - S_M|.$$

Since $(S_N)_N$ is convergent, it is a Cauchy sequence. From the above, we see that $(s_N)_N$ is a Cauchy sequence and hence convergent.

Rearranging the terms in a series may change its value. However, in some cases, we can safely rearrange the terms in a series.

Proposition 1.4. *If a series is made up only of positive terms, it can be rearranged without changing its sum.*

2 Expectation (discrete case)

In this lecture, we define expectations for discrete random variables. Handling the discrete case separately serves the purpose of building our intuition of expectations before we handle the more difficult case of non-discrete random variables.

Recall that for a discrete random variable X, we can find a countable set $\{x_n\}_n \subset \mathbb{R}$ such that

$$\sum_n \mathbb{P}(\{X = x_n\}) = 1.$$

Note that this does not necessarily imply that the range of X is $\{x_n\}_n$ (remember, random variables are functions from Ω to $\mathbb{R}$). However, we can define a new random variable, call it Y, as follows:

$$Y(\omega) = \sum_n x_n I_{\{X=x_n\}}(\omega).$$

Note that

$$\mathbb{P}(\{Y = x_n\}) = \mathbb{P}(\{I_{\{X=x_n\}} = 1\}) = \mathbb{P}(\{X = x_n\}),$$

and hence the random variables X and Y are, for all intents and purposes, identical. Therefore, for the remainder, we will always assume—without loss of generality–that a discrete random variable X has the form

$$X(\omega) = \sum_n x_n I_{\Lambda_n}(\omega)$$

for some partition $\Lambda_1, \Lambda_2, \ldots$ of the sample space (i.e., $\Omega = \Lambda_1 \cup \Lambda_2 \cup \cdots$ and $\Lambda_i \cap \Lambda_j = \emptyset$ whenever $i \neq j$).

Definition 2.1. A discrete random variable X is *integrable* if

$$\sum_n |x_n| \, \mathbb{P}(\Lambda_n) < \infty.$$

Definition 2.2. The *expectation* of an integrable discrete random variable X is

$$\mathbb{E}X = \sum_n x_n \mathbb{P}(\Lambda_n).$$

Example 2.3. Toss a coin $N \geq 1$ times. Let X_N be the number of heads. If the probability of heads is p, the expectation of X_N is

$$\mathbb{E}X_N = \sum_{n=0}^{N} n\mathbb{P}(\{X_n = n\}) = \sum_{n=1}^{N} n \binom{N}{n} p^n (1-p)^{N-n} = pN.$$

That is, you are expected to see pN heads "on average". If the coin is fair, for example, $pN = N/2$ (half of the tosses should, on average, be heads).

Note, in particular, that we only define the expectation of random variables that are integrable, and integrability has to do with absolute convergence.

Remark 2.4. We have, so far, ignored a technical issue. Earlier, we characterized a random variable in terms of the sets $\Lambda_1, \Lambda_2, \ldots$ However, the choice of these sets is not unique. For example, the constant random variable

$$X(\omega) = 1$$

can be written in many ways. Two possibilities are $X(\omega) = 1I_{\{\Omega\}}(\omega)$ and $X(\omega) = 1I_{\{\Lambda\}}(\omega) + 1I_{\{\Lambda^c\}}(\omega)$ where Λ is any subset of the sample space Ω. Since the definition of expectation depends on a particular choice of the sets $\Lambda_1, \Lambda_2, \ldots$, it is not clear that the expectation will remain the same if we change our choice of $\Lambda_1, \Lambda_2, \ldots$ This technicality is handled on page 55 of Walsh, John B. *Knowing the odds: an introduction to probability.* Vol. 139. American Mathematical Soc., 2012.

Example 2.5. Let X be an integer-valued nonnegative random variable (i.e., $\sum_{n \geq 0} \mathbb{P}(\{X = n\}) = 1$). If X is integrable,

$$X(\omega) = \sum_{n \geq 0} n I_{\{X=n\}}(\omega).$$

Therefore,

$$\mathbb{E}[X] = \sum_{n \geq 0} n\mathbb{P}(\{X = n\}) = 0\mathbb{P}(\{X = 0\}) + 1\mathbb{P}(\{X = 1\}) + 2\mathbb{P}(\{X = 2\}) + \cdots$$
$$= (\mathbb{P}(\{X = 1\}) + \mathbb{P}(\{X = 2\}) + \cdots) + (\mathbb{P}(\{X = 2\}) + \mathbb{P}(\{X = 3\}) + \cdots) = \sum_{n \geq 1} \mathbb{P}(\{X \geq n\}).$$

Proposition 2.6. *Let X and Y be discrete random variables and $a, b \in \mathbb{R}$. Then,*

1. *If X and Y are integrable, so is $aX + bY$ and*

$$\mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y.$$

2. *If $|X| \leq |Y|$ and Y is integrable, then X is integrable.*
3. *If X and Y are integrable and $X \leq Y$, then $\mathbb{E}X \leq \mathbb{E}Y$.*
4. *If X is integrable, $\mathbb{E}X \leq \mathbb{E}|X|$.*

Proof. Recall that we can partition Ω into events $\Lambda_1^X, \Lambda_2^X, \ldots$ on which X is constant. We can do the same for Y, obtaining $\Lambda_1^Y, \Lambda_2^Y, \ldots$ This allows us to define $\Lambda_{ij} = \Lambda_i^X \cap \Lambda_j^Y$, on which **both** X and Y are constant. Since $(\Lambda_{ij})_{i,j}$ is a countable sequence, let's relabel it $(\Lambda_n)_n$ and take $X = x_n$ and $Y = y_n$ on Λ_n.

1. Suppose X and Y are integrable. Then,

$$\begin{aligned} \sum |ax_n + by_n| \, \mathbb{P}(\Lambda_n) &\leq |a| \sum |x_n| \, \mathbb{P}(\Lambda_n) + |b| \sum |y_n| \, \mathbb{P}(\Lambda_n) \\ &= |a| \sum |x_n| \, \mathbb{P}(\Lambda_n^X) + |b| \sum |y_n| \, \mathbb{P}(\Lambda_n^Y) \\ &= |a| \, \mathbb{E}[|X|] + |b| \, \mathbb{E}[|Y|] \end{aligned}$$

and hence $aX + bY$ is integrable. Repeating almost the exact same computation as above without the absolute value signs yields the desired result.

2. This follows from $\sum |x_n|\mathbb{P}(\Lambda_n) \le \sum |y_n|\mathbb{P}(\Lambda_n)$.

3. Exercise.

4. Take $Y = |X|$ in (3). $\square$

Most importantly, the above proposition tells us that the expectation is a linear function. That is, let $\mathcal{X}$ be the set of all random variables. Define $T\colon \mathcal{X} \to \mathbb{R}$ as the mapping from a random variable to its expectation:

$$T(X) = \mathbb{E}X.$$

Then, T is linear function (i.e., $T(aX + bY) = aT(X) + bT(Y)$).

As an example of expectations, we introduce now *probability generating function* of a discrete random variable. We point out that our treatment is a bit cavalier for the time being, but we will come back to generating functions in a more principled manner. Before we move to this example, let's give a simple definition:

Definition 2.7. The *probability mass function* (PMF) of a discrete random variable X is $p\colon \mathbb{R} \to [0,1]$ defined by

$$p(x) = \begin{cases} \mathbb{P}(\Lambda_n) & \text{if } x = x_n \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.8. Let X be an integer-valued nonnegative random variable. Define G, the probability generating function of X, by

$$G(t) = \mathbb{E}\left[t^X\right].$$

Ignoring the integrability of X,

$$G(t) = \sum_{n=0}^{\infty} p(n)t^n = p(0) + \sum_{n=1}^{\infty} p(n)t^n$$

where p is the probability mass function of X. Being a power series, G has a radius convergence $0 \le R \le \infty$ which characterizes which values of t it converges for (i.e., converges for $|t| < R$ and diverges for $t > |R|$). Since

$$G(1) = \sum_{n=0}^{\infty} p(n)1^n = \sum_{n=0}^{\infty} p(n) = 1,$$

we know that the radius of convergence must be at least one (i.e., $R \ge 1$). Furthermore,

$$G(0) = p(0) + \sum_{n=1}^{\infty} p(n)0^n = p(0).$$

Now, if we take derivatives of G (for values of t inside the radius of convergence), we get

$$\begin{aligned}
G'(t) &= \sum_{n=1}^{\infty} np(n)t^{n-1} \\
G''(t) &= \sum_{n=2}^{\infty} n(n-1)\,p(n)t^{n-2} \\
&\vdots \\
G^{(k)}(t) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)\,p(n)t^{n-k}.
\end{aligned}$$

We conclude that $p(n) = \frac{1}{n!}G^{(n)}(0)$ for $n = 1, 2, \ldots$

In the previous example, we wrote $\mathbb{E}[t^X]$ even though t^X was not necessarily integrable for arbitrary t. We will often perform this abuse of notation by writing $\mathbb{E}Y$ for any random variable Y with the implicit understanding that the $\mathbb{E}Y$ is only well-defined when Y is integrable.

Proposition 2.9. *Let $f\colon \mathbb{R} \to \mathbb{R}$ and X be a discrete random variable with probability mass function p and support $\{x_n\}_n$. Then, $Y = f \circ X$ is integrable if and only if $\sum_n |f(x_n)|p(x_n) < \infty$, in which case*

$$\mathbb{E}\left[f(X)\right] = \sum_n f(x_n)p(x_n).$$

3 Variance

Definition 3.1. Let X be a discrete random variable. Its *variance* is defined as

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

(there is an implicit assumption about integrability in the definition of variance). Its *standard deviation* is $\sqrt{\operatorname{Var} X}$.

Once again interpreting $\mathbb{E}$ as an average, it is clear from the definition that variance is a measure of how far the random variable is from the expectation "on average". Note also that

$$\begin{aligned}
\operatorname{Var} X &= \mathbb{E}\left[(X - \mathbb{E}X)^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}X\mathbb{E}X + (\mathbb{E}X)^2 \\
&= \mathbb{E}\left[X^2\right] - 2\left(\mathbb{E}X\right)^2 + (\mathbb{E}X)^2 \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2,
\end{aligned}$$

which gives us a useful formula for the variance of a random variable. We will see next class that $\mathbb{E}X^2$ is referred to as the *second* ***raw*** *moment*, and another name for the variance is the *second* ***central*** *moment*.

Example 3.2. Toss a coin $N \geq 1$ times. Let X_N be the number of heads. Remember, we computed $\mathbb{E}X_N = pN$. Therefore, to get $\operatorname{Var} X_N$, it is sufficient to compute $\mathbb{E}[X_N^2]$:

$$\mathbb{E}\left[X_N^2\right] = \sum_{n=0}^{N} n^2 \mathbb{P}(\{X_N = n\}) = \sum_{n=1}^{N} n^2 \binom{N}{n} p^n (1-p)^{N-n} = p\left(\left(N-1\right)Np + N\right).$$

Exercise 3.3. Let X be a random variable with finite variance. Then

$$\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X).$$

Exercise 3.4. Show that a random variable with finite second raw moment also has finite expectation (i.e., it is not possible to have a random variable with finite second raw moment but infinite expectation).

Remark 3.5. For those familiar with measure theory, the above is an immediate consequence of the deeper fact that for a finite measure space, $\mathbb{L}^q \subset \mathbb{L}^p$ for $1 \leq p \leq q \leq \infty$.