Review of Basic Probability

Bibliography

- Bertsekas, D. P., & Tsitsiklis, J. N. (2002). Introduction to Probability, Athena Scientific, Massachusetts Institute of Technology.
- ROSS, Sheldon. A first course in probability. Pearson, 2014.
- Shiryaev, A. N. (1996). Probability, Graduate texts in mathematics, Springer

LECT 1

1 Sets

A set is a collection of objects. If x is an element of a set S then we write $x \in S$. If S contains countably many elements, then we can write S as a list $\{x_1, x_2, \ldots\}$ where the x_i's represent the elements of S. For example, the set of odd integers can be written as $\{-1, +1, -3, +3, -5, +5, \ldots\}$ and is therefore countable—in this case it is countably infinite. Note here that we do not put any restriction on the collection of objects, which can virtually be anything. Sets that are 'often used in practice' are those that specify a collection of sequences. For example,

$$S = \left\{ x = (x_1, x_2, \ldots) | x_i \in \mathbb{R} \text{ for all i, and} \sum_i x_i = 1 \right\}.$$

In this example, each element of S is a sequence (that is an ordered collection) of real numbers that satisfy a certain property P (that their sum equals one). This example shows an alternative way to specify a set S through a property that is shared by all its elements

$$S = \{x | x \text{ satisfies P}\}$$

for a certain property P.

1.1 Set operations

1. $S \cup T = T \cup S$, and $S \cap T = T \cap S$ (symmetry)
2. $S \cap (T \cup U) = (S \cap U) \cup (S \cap U)$ and $S \cup (T \cap U) = (S \cup U) \cap (S \cup U)$ (distributivity)
3. $(\cup_n S_n)^c = \cap_n S_n^c$ and $(\cap_n S_n)^c = \cup_n S_n^c$ (De Morgan's law, notice here that these identities also hold for uncountable family of sets)

Notation. *Given sets S, T we write $S \backslash T$ (read 'S minus T') to denote $S \cap T^c$*

Exercise 1. *Prove Properties 1.-3.*

2 Probabilistic model

A probabilistic model is a mathematical description of an uncertain situation or a random experiment. It has two main ingredients, the sample space Ω which represents all possible outcomes, and a probability law $\mathbb{P}$ which assigns probabilities to events that may or may not happen depending on the outcome of the random experiment. For example, consider a fair coin toss. The two possible outcomes are $\{H, T\}$, and their probabilities are

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2.$$

The fact that these probabilities sum to one captures the assumption that the result of the coin toss will, with certainty, be either H or T, that is $\mathbb{P}(\{H\} \cup \{T\}) = \mathbb{P}(\{H, T\}) = 1$.

The above example hints that the probability law should assign probabilities to events so that to satisfy certain consistency properties. For instance, the probability of 'seeing more' should not be smaller than the probability of 'seeing less'. This is formalized in the following two definitions.

Definition 1 (Measurable space)**.** *A measurable space $(\Omega, \mathscr{F})$ is a set Ω together with a family of subsets $\mathscr{F}$, referred to as 'events' ,called a σ-*algebra*, that satisfies the following properties:*

- $\Omega \in \mathscr{F}$
- *if* $A \in \mathscr{F}$ *then* $A^c \in \mathscr{F}$
- *if* $A_n \in \mathscr{F}$ $n = 1, 2, \ldots$ *then* $\cup_{n=1}^{\infty} A_n \in \mathscr{F}$

Subsets that satisfy the above properties are referred to as 'events.'

Remark 1.

- $\mathscr{F} = \{\Omega, \emptyset\}$ *this is the 'poorest' σ-algebra, not very useful.*
- $\mathscr{F} = \{\Omega, \emptyset, A, A^c, \}$, $A \in \Omega$ *(σ-algebra 'generated by A'). For instance, for the coin toss example we have say* $A = \{H\}$ *and* $A^c = \{T\}$.
- *More generally, given a set of subsets $\mathscr{S}$, we can define the 'σ-algebra generated by $\mathscr{S}$' as being the smallest σ-algebra that contains $\mathscr{S}$*
- *if* $|\Omega| < \infty$ *then* $\mathscr{F} = \{$*all subsets of* $\mathscr{F}\}$ *is the 'richest' σ-algebra of* Ω. $|\mathscr{F}| = 2^{|\Omega|}$
- Important, 'Borel algebra'*:* $\Omega = \mathbb{R}$ *and* $\mathscr{F}$ *is the σ-algebra generated by the sets of* $\mathbb{R}$ *that can be written as finite sums of intervals* $(a, b]$, $a < b$. *Notation:* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$

Exercise 2. *Show that any singleton* $\{a\}$ *belongs to* $\mathscr{B}(\mathbb{R})$. *Deduce that the Borel algebra can alternatively be defined by replacing the semi-open set condition* $(a, b]$, *by* $[a, b)$, *by* $[a, b]$, *or by* (a, b).

A random experiment is mathematically modelled by the notion of 'probability space.'

Definition 2 (Probability space)**.** *A probability space is a tuple* $(\Omega, \mathscr{F}, P)$ *where* $(\Omega, \mathscr{F})$ *is a measurable space and* P *a function* $\mathscr{F} \longrightarrow [0, 1]$ *that satisfies*

i. $P(A) \geq 0$ *for any* $A \in \mathscr{F}$

ii. if $A_n \in \mathscr{F}$ $n = 1, 2, \ldots$ *are disjoint then*

$$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$$

iii. $P(\Omega) = 1$.

Remark 2. *The previous definition implies that the probability of 'seeing more' is not smaller than the probability of 'seeing less', that is* $A \subset B \Rightarrow P(A) \leq P(B)$. *In fact,* $A \subset B \Rightarrow B^c \subset A^c$, *hence* A^c *can be written as the disjoint union of* B^c *and* $A^c \setminus B^c$. *Therefore*

$$P(A) = 1 - P(A^c) = 1 - (P(B^c \cup (A^c \setminus B^c))) = 1 - (P(B^c) + P(A^c \setminus B^c) \leq 1 - P(B^c) = P(B)$$

where the third equality holds by ii. *and where the inequality holds by* i.

Example 1 (Coin tosses)**.** *The random experiment consists of a coin being tossed* n *times and for which we record the results* $(a_1, a_2, \ldots, a_n)$. *The sample space is therefore*

$$\Omega = \{\omega : \omega = (a_1, a_2, \ldots, a_n), a_i = 0, 1\}$$

As a $\sigma-$*algebra, consider* $\mathscr{F}$ *= all subsets of* Ω.*To specify a probability over* $(\Omega, \mathscr{F})$ *we need to assign a probability to each* $\omega = (a_1, a_2, \ldots, a_n)$. *For instance,*

$$P(\{\omega\}) = p^{\sum_i a_i}(1-p)^{n - \sum_i a_i}$$

which satisfies $\sum_\omega P(\{\omega\}) = 1$. *Note that here one random experiment gives* n *results. We could alternatively view each of these outcomes as independent experiments (see Section 3.4).*

The following theorem provides a very useful bound:

Theorem 1 (Union bound)**.** *If* $\{A_j\}_{j \geq 1}$ *is a (countable) set of disjoint events*

$$P(\cup_{j \geq 1} A_j) \leq \sum_{j \geq 1} P(A_j) .$$

The usefulness of the theorem stems from the fact that it is often easier to compute the probabilities of the individual events than their union.

Proof of Theorem 1. Since $\cup_{i=1}^n A_i$ can be written as a union of disjoint events

$$A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_2 \cup A_1)) \cup \ldots \cup (A_n \setminus \cup_{i=1}^{n-1} A_i)$$

we have

$$P(\cup_{i=1}^{n} A_i) = P(A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_2 \cup A_1)) \cup \ldots (A_n \setminus \cup_{i=1}^{n-1} A_i))$$
$$\overset{Def.2+Rem.2}{\leq} \sum_{i=1}^{n} P(A_i)$$

□

Definition 3. *Given a probability space* $(\Omega, \mathscr{F}, P)$*, the conditional probability of an event* $A \in \mathscr{F}$ *given an event* $B \in \mathscr{F}$ *is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

whenever $P(B) > 0$*. If* $P(A|B) = P(A)$*, A and B are said* independent.

The conditional probability can alternatively be computed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

whenever $P(B) > 0$. This is Bayes' rule.

Exercise 3. *Prove Bayes' rule.*

Exercise 4. *Out of the students in a class, 60% are geniuses, 70% love chocolate, and 40% fall into both categories. Determine the probability that a randomly selected student is neither a genius nor a chocolate lover.*

Exercise 5. *A six-sided die is loaded in a way that each even face is twice as likely as each odd face. Construct a probabilistic model for a single roll of this die, and find the probability that a 1, 2, or 3 will come up.*

Exercise 6. *We roll two fair 6-sided dice. Each one of the 36 possible outcomes is assumed to be equally likely.*

1. *Find the probability that doubles are rolled.*
2. *Given that the roll results in a sum of 4 or less, find the conditional probability that doubles are rolled.*

3. *Find the probability that at least one die roll is a 6.*

4. *Given that the two dice land on different numbers, find the conditional probability that at least one die roll is a 6.*

3 Random variable

Definition 4 (Measurable function and random variable)**.** *Given a measurable space* $(\Omega, \mathscr{F})$*, a function* $X : \Omega \to \mathbb{R}$ *such that*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F} \tag{1}$$

for any $B \in \mathscr{B}(\mathbb{R})$ *is called 'measurable' or* $\mathscr{F}$*-measurable. When* $(\Omega, \mathscr{F})$ *is the underlying measurable space of a probability space* $(\Omega, \mathscr{F}, P)$*, it is custom to call* X *'random variable' because its argument is random. In this case we also define the 'cumulative distribution function' (cdf) of* X *as*

$$F_X(x) \triangleq P(\{\omega \in \Omega : X(\omega) \leq x\}) \triangleq P(X \leq x)$$

Example 2.

- $\Omega = \{H, T\}^n$
- $P(\omega) = p^{\#T}(1-p)^{\#H}$, $\omega \in \Omega$,
- $X(\omega) =$ *first place where a 'H' appears (if* $\omega = TTHTHHT$ *then* $X(\omega) = 3$*). If* $\omega = \underbrace{TT\ldots T}_{n\times}$, $X(\omega) \triangleq n+1$.

Then

$$F_X(x) = \begin{cases} 0 & x < 1 \\ \sum_{j=1}^{k} p^{j-1}(1-p) & k \leq x < k+1 \quad k = 1, 2, \ldots, n \\ 1 & x \geq n \end{cases} .$$

Definition 5 (Indicator function)**.** *Consider a measurable space* $(\Omega, \mathscr{F}, P)$*. The indicator function* $\mathbb{1}_A(\omega)$ *with respect to an event* $A \in \mathscr{F}$ *is defined to be equal to* 1 *if* $\omega \in A$ *and zero otherwise.*

Example 3. *Referring to the previous example define*

$$A_i = \{\omega | T \text{ appears in the } i\text{-th position}\}.$$

Then the number of T's that appear in a sequence ω can be written as

$$\#T(\omega) = \sum_{i=1}^{n} \mathbb{1}_{A_i}(\omega).$$

LECT 2

3.1 Discrete vs continuous random variables

Definition 6 (Discrete r.v.)**.** *A random variable is 'discrete' if its range is (at most) countably infinite. In this case, the r.v. is characterised by its probability mass function (pmf) P_X defined as*

$$P_X(x) \triangleq P(X^{-1}(x)) \triangleq P(\{\omega : X(w) = x\}) .$$

Notation. *Whenever clear from context we shall simply write $P(x)$ instead of $P_X(x)$.*

Example 4 (Famous discrete r.v.'s)**.**

- *Bernoulli*(p)*:* $P_X(1) = 1 - P_X(0) = p$
- *Binomial*(n, p)*:* $P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x \in \{0, 1, 2, \ldots, n\}$, $p \in [0, 1]$

  Interpretation: counts number of successes out of n trials
- *Geometric*(p)*:* $P_X(x) = p(1-p)^{k-1}$, $x \in \{1, 2, \ldots\}$, $p \in [0, 1]$

  Interpretation: records time of first success.
- *Poisson*(λ)*:* $P_X(x) = e^{-\lambda} \frac{\lambda^k}{x!}$, $x \in \{0, 1, 2, \ldots\}$, $\lambda > 0$.

Exercise 7. *Suppose that $p_n \in [0, 1]$ is such that $\lim_{n\to\infty} n \cdot p_n = \lambda$. Show that $P_{X^{Bin(n,p_n)}}(x) \to P_{X^{Poisson(\lambda)}}(x)$ as $n \to \infty$. In other words, the Poisson distribution approximates the binomial r.v. in the limit $n \to \infty$ in the small probability of success regime.*

Definition 7 (Continuous r.v.)**.** *A random variable is (absolutely) continuous if its cdf can be written as*

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy$$

for some nonnegative function f_X called the 'probability density function' (pdf) of X.

Example 5. *Gaussian r.v.:* $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}$

Remark 3. *In the sequel we will refer to pmf, pdf, or cdf of a r.v. X generically as 'probability law.'*

Remark 4. *Note that the probability law of a random varible X completely specifies it—in the sense that $P(X \in B)$ can be computed for any Borel set B. In particular there is no need (except for interpretation purposes) to refer to the underlying probability space.*

3.2 Multiple random variables

For simplicity of the presentation we will focus on pairs of random variables, the case of $n \geq 3$ will then be straightforward.

A pair of discrete random variables X_1, X_2 is characterized by its joint p.m.f.

$$P_{X_1,X_2}(x_1, x_2) \quad \text{for all} \quad x_1, x_2\,.$$

A pair of absolutely continuous random variables random variables X_1, X_2 is characterized by its joint pdf

$$f_{X_1,X_2}(x_1, x_2) \quad \text{for all} \quad x_1, x_2\,.$$

From the joint probability law one derives the law of each individual random variable (and the marginals) by 'marginalizing'. For the discrete case

$$P_{X_1}(x_1) = \sum_{x_2} P_{X_1,X_2}(x_1, x_2)$$

and for the continuous case

$$f_{X_1}(x_1) = \int f_{X_1,X_2}(x_1, x_2)\, dx_2.$$

Note that marginalizing just says that the probability that '$X_1 = x_1$' is the same as the probability that '$X_1 = x_1$ without restriction on X_2.'

3.3 Deducing the law of $Y = g(X)$ knowing the law of X

For the discrete case we have

$$P_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P_X(x)$$

Instead, for the continuous case we use the following receipe. Compute first the cdf

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y))$$

then take its derivative:

$$f_Y(y) = dF_Y(y)/dy$$

(Recall that $\frac{d}{dy}\int_{-\infty}^{h(y)} f(x)dx = (\frac{d}{dy}h(y))f(h(y)))$

Example 6 (Linear transformation)**.** *Knowing $f_X(x)$what is $f_Y(y)$ if $Y = aX + b$ ($a \neq 0$)? Using the receipe we have*

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(aX + b \le y) \\ &= P(aX + b \le y) \\ &= P(X \le (y-b)/a) \quad a > 0 \\ &= \int_{-\infty}^{(y-b)/a} f_X(y)dy \end{aligned}$$

Hence,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a} f_X((y-b)/|a|)\,.$$

We note that if X is Gaussian(μ, σ^2), then $Y = aX+b$ is Gaussian$(a\mu+b, a^2\sigma^2)$. I.e., the Gaussian law is preserved under linear transformation.

In the multivariate setup the receipe remains the same. Let $X = (X_1, X_2, \ldots, X_n)$ have density $f_X(x)$, $g : \mathbb{R}^n \to \mathbb{R}^n$, and let

$$Y = (Y_1, Y_2, \ldots, Y_n) = (g_1(X), g_2(X), \ldots, g_n(X)) = g(X)$$

and

$$X = (X_1, X_2, \ldots, X_n) = (g_1^{-1}(Y), g_2^{-1}(Y), \ldots, g_n^{-1}(Y)) = g^{-1}(X)$$

This step assumes that the inverse map $g^{-1}(X)$ exists. One first computes the Jacobian of g^{-1}

$$\frac{Dg^{-1}}{dy}(y) \triangleq \det \begin{pmatrix} \frac{\partial g_1^{-1}(y)}{\partial y_1} & \ldots\ldots & \frac{\partial g_1^{-1}(y)}{\partial y_n} \\ \frac{\partial g_2^{-1}(y)}{\partial y_1} & \ldots\ldots & \frac{\partial g_2^{-1}(y)}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial g_n^{-1}(y)}{\partial y_1} & \ldots\ldots & \frac{\partial g_n^{-1}(y)}{\partial y_n} \end{pmatrix},$$

again assuming that it is well defined—that is that each partial derivative exists and is finite. The density of the vector Y is then given by

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{Dg^{-1}}{dy}(y) \right|$$

Example 7. *Let*

$$f_X(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}, \quad x_1, x_2 \in \mathbb{R}$$

and let $y_1 = \sqrt{x_1^2 + x_2^2} \in \mathbb{R}_+$ *and* $y_2 = Arctan(x_2/x_1) \in [0, 2\pi]$.

$x_1 = y_1 \cos(y_2) (= g_1^{-1}(y_1, y_2))$ *and* $x_2 = y_1 \sin(y_2) (= g_2^{-1}(y_1, y_2))$

One verifies that

$$\frac{Dg^{-1}}{dy}(y) = y_1$$

and thus

$$f_Y(y_1, y_2) = \frac{1}{2\pi} e^{-y_1^2/2} y_1 \, .$$

Exercise 8. *Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?*

The expectation and the variance of a random variable capture the average and the spread around the average:

Definition 8 (Expectation, variance, standard deviation)**.**

- *Discrete r.v.:* $E(X) \triangleq \sum_x x P_X(x)$
- *Absolutely continuous r.v.:* $E(X) = \int_{\mathbb{R}} x f_X(x) dx$.

The variance of a r.v. X *is defined as* $Var(X) = E(X - E(X))^2$, *and the standard deviation, denoted as* σ *is defined as* $\sqrt{Var(X^2)}$.

Example 8.

	$E(X)$	$Var(X)$
$Binomial(n,p)$	np	$np(1-p)$
$Geometric(p)$	$1/p$	$(1-p)/p^2$
$Poisson(\lambda)$	λ	λ
$Gaussian(m,\sigma^2)$	m	σ^2

The following property, the linearity of expectation, is a very important one and can easily be derived from the definition of expectation: for any finite weighted sum of (dependent or independent) random variables $\sum_{i=1}^k \alpha_i X_i$ (with finite constants α_i's) we have

$$E(\sum_{i=1}^k \alpha_i X_i) = \sum_{i=1}^k \alpha_i E(X_i) \, . \qquad \text{⚠}$$

Example 9. *Consider Example 2. The expected number of T's in a sequence* ω *is*

$$E(\#T) = E(\sum_{i=1}^n \mathbb{1}_{A_i}) = \sum_{i=1}^n E(\mathbb{1}_{A_i})$$

Now,

$$E(\mathbb{1}_{A_i}) = 1 \cdot P(\textit{there is an T in the i-th position}) + 0 \cdot P(\textit{there is a H in the i-th position})$$
$$= p$$

An therefore $E(\#T) = np$.

The following example, taken from the book "Introduction to Geometric Probability" by Daniel Klein and Giancarlo Rota, is meant to elegantly illustrate how useful linearity of expectation is.

Example 10 (Buffon's needle)**.** *Consider a floor made of parallel strips of wood, each of width* w*. We drop a needle of length* $\ell \leq w$ *on the floor. What is the probability that the needle crosses at least one line between two strips?*

Let $f(\ell)$ *denote the expected number of crossings when a needle of length* ℓ *is dropped on the floor—*$f(\ell)$ *is unknown. Now consider two needles of length* ℓ_1 *and* ℓ_2 *and let* X_1 *be the number of lines crossed by the first needle and let* X_2 *be the number of lines crossed by the second one. Assuming that the two needles are aligned, to form a big needle of length* ℓ*, by linearity of expectation (!) we have*

$$f(\ell) = f(\ell_1 + \ell_2) = E(X_1 + X_2) = E(X_1) + E(X_2) = f(\ell_1) + f(\ell_2).$$

We note here that as long as a the big needle can be decomposed into two needles, with an arbitrary angle, of lengths ℓ_1 *and* ℓ_2*, the above equalities hold. This implies two things. The first is that the number of crossing of all big needles of length* ℓ *is* $f(\ell_1 + \ell_2) = f(\ell_1) + f(\ell_2)$*, irrespective of the angle of the two sub-needles, and irrespective of whether or not they are attached (which potentially introduces dependency). Because of this and the fact that* $f(0)$ *is 'obviously'* 0 *we deduce that* f *is a linear function, that is* $f(\ell) = c \cdot \ell$ *for some constant* c*. More precisely, we have linearity for any combination of needles, whether attached or not. Now consider a circle of diameter* w *(why not?). Such a circle when thrown on the floor always crosses the lines exactly twice. We then deduce that*

$$f(w\pi) = 2 = c \cdot w \cdot \pi$$

and therefore $c = 2/\pi w$. *Hence,*

$$E(\#crossings) = \frac{2}{\pi w}\ell.$$

Now, if $\ell \leq w$ *then the needle crosses either one line or no line (if* $\ell = w$ *the needle can be positioned so that it touches two lines, but this never happens). So in this case the expected number of crossings is equal to the probability that the needle crosses one line. Note that for the argument to be completed there is a limiting argument needed to approximate a circle with a polygon since* $f(\ell)$ *refers to straight lines.*

LECT 3: The following theorem provides a handy expression for computing expectation of random variables taking values over the nonnegative integers:

Theorem 2. *For a r.v.* X *taking values in* $\mathbb{N}$ *we have*

$$E(X) = \sum_{n=1}^{\infty} P(X \geq n)$$

From this theorem, we immediately get the lower bound $E(X) \geq k \cdot P(X \geq k)$ for any nonnegative integer k, or $P(X \geq k) \leq E(X)/k$, which is Markov inequality, which we will revisit later.

Proof of Theorem 2.

$$\begin{aligned} E(X) &= 1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3) + \ldots \\ &= P(X=1) + \left\{ \begin{array}{l} P(X=2) \\ P(X=2) \end{array} \right. + \left\{ \begin{array}{l} P(X=3) \\ P(X=3) \\ P(X=3) \end{array} \right. + \ldots \\ &= \sum_{n=1}^{\infty} P(X \geq n) \end{aligned}$$

□

Properties of $E(X)$ and $\mathrm{Var}(X)$:

- If $X = cste$, $E(X) = c$, $\mathrm{Var}(X) = 0$.

- if $Y = aX + b$: $E(Y) = aE(X) + b$, $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$
- $0 \le \mathrm{Var}(X) = E(X^2) - (E(X))^2$
- If $g(x) \ge h(x)$ then $E(g(X)) \ge E(h(X))$

Theorem 3 (Jensen's inequality)**.** *$E(g(X)) \ge g(E(X))$ whenever g is convex.*

Proof. Let $c = E(X)$ and let $\phi(x)$ be the line that is tangent to $g(x)$ at $x = c$, i.e., $\phi(x) = g(c) + \alpha(x - c)$ for some constant α.

By convexity $g(x) \ge \phi(x)$ and therefore

$$Eg(X) \ge E(\phi(X)) = E(g(c) + \alpha(X - c)) = g(E(X))$$

by linearity of the expectation. □

Example 11. *Jensen's inequality applied to the function $x \to x^2$ immediately yields $E(X^2) \ge (E(X))^2$, the nonnegativity of the variance.*

Exercise 9. *Consider a sequence of independent tosses of a biased coin at times $t = 0, 1, 2, \ldots$ On each toss, the probability of a 'head' is p, and the probability of a 'tail' is $1 - p$. A reward of one unit is given each time that a tail follows immediately after a 'head.' Let R be the total reward paid in times $1, 2, ..., n$. Find $E[R]$ and $Var(R)$.*

Solution: Let A_k be the event 'there is a reward at time k' and let $\mathbb{1}_{A_k}$ denote the corresponding indicator function, that is $\mathbb{1}_{A_k}(\omega) = 1$ if the sequence of coin tosses w has a T at position k and an H at position $k-1$. Then $E(\mathbb{1}_{A_k}) = p(1-p)$ and we have

$$E(R) = E(\sum_{k=1}^{n} \mathbb{1}_{A_k}) = np(1-p)$$

by linearity of expectation.

For the variance observe that

$$E(\mathbb{1}_{A_k}^2) = p(1-p) \tag{2}$$

$$E(\mathbb{1}_{A_k}\mathbb{1}_{A_{k+1}}) = 0 \qquad \textit{(it is impossible to get a reward at position k and k+1)} \tag{3}$$

$$E(\mathbb{1}_{A_k}\mathbb{1}_{A_{k+\ell}}) = p^2(1-p)^2 \qquad \textit{for } \ell \ge 2 \tag{4}$$

We then get by linearity of expectation

$$E(R^2) = \sum_{k=1}^{n} \sum_{m=1}^{n} E(\mathbb{1}_{A_k} \mathbb{1}_{A_m}). \tag{5}$$

In this sum, the n terms where $k = m$ contribute each to $p(1-p)$. The $2(n-1)$ terms where $|k-m| = 1$ contribute to 0. Finally, the remaining $n^2 - n - 2(n-1)$ terms contribute to $p^2(1-p)^2$. Therefore,

$$E(R^2) = np(1-p) + (n^2 - 3n + 2)p^2(1-p)^2.$$

Finally,

$$\begin{aligned} Var(R) &= E(R^2) - (E(R))^2 \\ &= np(1-p) + (n^2 - 3n + 2)p^2(1-p)^2 - n^2p^2(1-p)^2 \\ &= np(1-p) - (3n-2)p^2(1-p)^2. \end{aligned}$$

Exercise 10. *If the weather is good, which happens with probability 0.6, Alice walks the 2km to her class at a speed V of 5km/h, and otherwise rides her motorcycle at a speed V of 30km/h. What is the mean of the time T to get to her class? Consider the following two options:*

i. Option 1: Compute $E(T)$ by first computing the pmf of T.

ii. Option 2: Compute the mean of the speed $E(V)$ to get to class, then claim that the mean time $E(T) = 2/E(V)$.

Which one is correct? Why?

3.4 Expectation and dependencies

The probability of X_2 given X_1 is defined as

$$P_{X_2|X_1}(x_2|x_1) = P_{X_1,X_2}(x_1, x_2)/P_{X_1}(x_1)$$

for the discrete case, and as

$$f_{X_2|X_1}(x_2|x_1) = f_{X_1,X_2}(x_1, x_2)/f_{X_1}(x_1)$$

for the continuous case.

Definition 9 (Independence)**.** *X and Y are 'independent' if $P_{X|Y}(x|y) = P_X(x)$ for all x, y ($f_{X,Y}(x,y) = f_X(x) f_Y(y)$).*

Definition 10 (Conditional expectation and variance)**.** *The conditional expectation of Y given $X = x$ is defined as*

$$E(Y|X = x) = \sum_y y P_{Y|X}(y|x) .$$

The conditional variance of Y given $X = x$ is defined as

$$Var(Y|X = x) = E((Y - E(Y|X = x))^2 | X = x) .$$

$E(Y|X)$ and $\mathrm{Var}(Y|X)$ are random variables since functions of the random variable X.

Properties:

- $E(Y) = E(E(Y|X))$ (to check)
- $\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X))$ (to check)
- if X and Y are independent,
  $E(XY) = E(X)E(Y)$ and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Definition 11 (Covariance)**.** *The 'covariance' of X and Y is defined as*

$$cov(X, Y) = E(XY) - E(X)E(Y) .$$

More generally, given a vector $X = (X_1, X_2, \ldots, X_n)$ we define the covariance matrix

$$K = \Big(\{Cov(X_i, X_j)\} \Big) .$$

A random vector is said uncorrelated *if its covariance matrix is diagonal.*

Remark 5. *Independence implies uncorrelated but the opposite is not true in general. Example: (X, Y) take the values $(1, 0)$, $(0, -1)$, $(-1, 0)$, $(0, 1)$ each with probability $1/4$. We have that $E(XY) = E(X)E(Y) = 0$ but X and Y are not independent.*

Definition 12 (Jointly Gaussian r.v.'s)**.** *A random vector* $(X_1, X_2, \ldots, X_n)$ *is Gaussian (or, has jointly Gaussian components) if for any real vector* $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ *the random variable*

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_n X_n$$

has a Gaussian distribution.

Theorem 4. *A random vector* $X = (X_1, X_2, \ldots, X_n)$ *with mean* μ *is Gaussian iff its pdf can be written as*

$$f_X(x) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}(x-\mu)^t K^{-1}((x-\mu))}$$

for some symetric and positive matrix K*.*

Remark 6. *It can be checked that* $K = \left(Cov(X_i, X_j)\right)$ *which justifies the terminology 'covariance matrix' for* K*.*

Remarkably, for gaussian random variables 'uncorrelated' and 'independence' are equivalent:

Theorem 5. *Let* $\vec{X} = (X_1, X_2, \ldots, X_n)$ *be a Gaussian vector. Then* $X_1, X_2, \ldots, X_n$ *are independent if and only if they are uncorrelated.*

Exercise 11. *Find two random variables, each Gaussian, but that are not not jointly Gaussian.*

Exercise 12. *Random variables* X *and* Y *have joint pmf*

$$P_{X,Y}(x, y) = c(x^2 + y^2)$$

if $x \in \{1, 2, 3\}$ *and* $y \in \{1, 3\}$ *and* 0 *otherwise. Compute:*

1. *the value of* c
2. $P(Y < X)$
3. $P(Y > X)$
4. $P(Y = X)$
5. $P(Y = 3)$

6. *the marginals* $p_X(x)$ *and* $p_Y(y)$
7. *the expectations* $E(X)$, $E(Y)$, *and* $E(X+Y)$
8. *the variances* $Var(X)$, $Var(Y)$, $Var(X+Y)$
9. *Let* A *denote the event that* $X \geq Y$. *Compute* $E(X|A)$ *and* $Var(X|A)$.

Exercise 13. *Prove that two events A and B are independent if and only if their corresponding indicator functions are independent.*

3.5 Sum of random variables

Definition 13. *Let* X *and* Y *be two independent random variables with pmf's* P_X *and* P_Y. *The pmf of their sum* $W = X + Y$ *is called the convolution of* P_X *and* P_Y *and is given by*

$$P_W(w) = \sum_x P_X(x) P_Y(w - x)$$

for the discrete case, and by

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) dx$$

for the continuous case.

Justification
Discrete case:

$$\begin{aligned} P_W(w) &\triangleq P(X + Y = w) \\ &= \sum_{(x,y):x+y=w} P_{X,Y}(X = x, Y = y) \\ &= \sum_{(x,y):x+y=w} P_X(X = x) P_Y(Y = y) \\ &= \sum_x P_X(X = x) P_Y(Y = w - x) \end{aligned}$$

For the continuous case as usually we proceed through the cdf:
We have

$$F_W(w) = P(X + Y \leq w) = \int_{x+y \leq w} f_{X,Y}(x, y) dx dy.$$

Since X and Y are independent we have $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

$$\int_{x+y\leq w} f_{X,Y}(x,y)dxdy = \int_{\mathbb{R}} dx f_X(x) \int_{-\infty}^{w-x} dy f_Y(y) \qquad (6)$$

Differentiating with with respet to w we then deduce that

$$f_W(w) = \int_{\mathbb{R}} f_X(x) f_Y(w-x)dx.$$

Remark 7. *The expressions for $P_W(w)$ and $f_W(w)$ are referred to as the convolution of the pdfs (pmfs) of X and Y.*

LECT 4

4 Inequalities

Theorem 6.

Markov *Let X be a nonnegative random variable then $P(X \geq a) \leq E(X)/a$, $a > 0$.*

Chebyshev: *For any random variable $P(|X - E(X)| \geq a) \leq Var(X)/a^2$, $a > 0$*

Chernoff bound: *For any random variable $P(X > a) \leq \inf_{t>0} E(e^{t(X-a)})$*

Note that the three inequalities quantify 'rare events'. For the Markov inequality to be non-trivial, the constant a should scale at least linearly with the expectation $E(X)$. For instance, $P(X \geq 10 \cdot E(X)) \leq 1/10$. By contrast, the Chebyshev inequality is non-trivial when a scales linearly with the standard deviation. For instance, $P(|X - E(X)| \geq 3\sigma_X) \leq 1/9$.

Proof of Theorem 6. Markov:

$$\begin{aligned} E(X) &\geq \sum_{x\geq a} xP(X = x) \\ &\geq a \sum_{x\geq a} P(X = x)\,. \end{aligned}$$

Chebyshev: from Markov's inequality we get

$$\begin{aligned} P(|X - E(X)| \geq a) &= P((X - E(X))^2 \geq a^2) \\ &\leq \frac{\mathrm{Var}(X)}{a^2} \end{aligned}$$

Chernoff:

$$\begin{aligned} P(X \geq a) &= P(e^{tX} \geq e^{ta}) \quad t > 0 \\ &\leq E(e^{tX})/e^{ta} \quad \text{Markov} \end{aligned}$$

Therefore $P(X > a) \leq \inf_{t>0} E(e^{t(X-a)})$. □

The term $E(e^{tX})$ in Chernoff bound is called the moment generating function of X (see Definition 16 below).

5 Limit Theorems

Definition 14 (Convergence in probability and almost sure)**.** *Given a probability space $(\Omega, \mathscr{F}, P)$, a sequence of random variables $X_1, X_2, \ldots$ converges in probability, respectively almost surely, to a r.v. X if*

$$\lim_{n\to\infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0 \quad \textit{for any } \varepsilon > 0\,,$$

respectively

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1\,.$$

Note the difference between almost sure convergence and convergence in probability. Almost sure convergence means that there exists an event $A \in \mathscr{F}$ of probability 1 such that for any $\omega \in A$ the corresponding sequence $x_1(\omega), x_2(\omega), \ldots$ converges to $x(w)$ (in the classical sense). This, in particular means, that for any of these sequences and for any $\varepsilon > 0$, $|x_n(w) - x(w)| > \varepsilon$ at most finitely many times—for otherwise the sequence would not converge.

Theorem 7 (Fundamental criterion for almost sure convergence, Borel-Cantelli lemma)**.** *A sequence of r.v. $X_1, X_2, \ldots$ converges almost surely to a r.v. X if for any $\varepsilon > 0$*

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \varepsilon) < \infty$$

Example 12. *Let X be a an exponentially distributed random variable with parameter $\lambda = 1$, that is $f_X(x) = e^{-x}$. We now prove that $X_n = X/n$ converges to zero, in probability. For any $\varepsilon > 0$ we have*

$$P(|X_n - 0| \geq \varepsilon) = P(X \geq n\varepsilon) = 1 - F_X(n\varepsilon) = e^{-n\varepsilon} \overset{n\to\infty}{\to} 0. \qquad (7)$$

Example 13. *Consider a sequence of independent random variables X_n that are uniformly distributed in the interval $[0,1]$, and let*

$$Y_n = \min\{X_1, X_2, \ldots, X_n\}.$$

We have

$$\begin{aligned} P(|Y_n - 0| \geq \varepsilon) &= P(X_1 \geq \varepsilon)P(X_2 \geq \varepsilon)\ldots P(X_n \geq \varepsilon) \\ &= (1-\varepsilon)^n. \end{aligned} \qquad (8)$$

Hence, Y_n converges to zero in probability. Furthermore, since $\sum_{n=1}^{\infty} P(|Y_n - 0| \geq \varepsilon) < \infty$, we also have almost sure convergence by the Borel-Cantelli lemma.

Remark 8. *If X_n converges to a number a in probability, it doesn't mean that $E(X_n)$ converges to a as $n \to \infty$ (L_1 convergence). Example:*

$$P(X_n = y) = \begin{cases} 1 - 1/n & for\, y = 0 \\ 1/n & for\, y = n^2 \\ 0 & elsewhere \end{cases}.$$

We have that $\lim_{n\to\infty} P(X_n = 0) = 1$ but $E(X_n) \to \infty$ as $n \to \infty$.

Example 14 (Convergence in probability, not almost surely)**.** *Consider the uniform distribution over $\Omega = [0,1]$ and define $X(\omega) = \omega$ and*

$$\begin{array}{ccc} X_1(\omega) = \omega + \mathbb{1}_{[0,1]}(\omega) & X_2(\omega) = \omega + \mathbb{1}_{[0,1/2]}(\omega) & X_3(\omega) = \omega + \mathbb{1}_{[1/2,1]}(\omega) \\ X_4(\omega) = \omega + \mathbb{1}_{[0,1/3]}(\omega) & X_5(\omega) = \omega + \mathbb{1}_{[1/3,2/3]}(\omega) & X_6(\omega) = \omega + \mathbb{1}_{[2/3,1]}(\omega) \\ \cdots & \cdots & \cdots \end{array}$$

We have that $P(|X_n - X| \geq \varepsilon) = 0$ tends to zero as it is equal to the probability of an interval that goes to zero. So $X_1, X_2, \ldots$ converges to X in probability. However, for any ω, $X_n(\omega)$ oscillates between ω and $\omega+1$ infinitely often. Thus X_n does not converge to X (almost surely).

As we saw above convergence in probability does not imply L_1 convergence. Almost sure convergence does not either imply L_1 convergence, but with an additional condition it does:

Theorem 8 (Dominated convergence)**.** *If (X_n) is a sequence of random variables such that*

- *$X_n \to X$ almost surely*
- *there exists a random variable Y such that $|X_n| \leq Y$ almost surely for all n and such that $E(|Y|) < \infty$*

Then

$$\lim_{n\to\infty} E(X_n) = E(X)\,.$$

There are many ways to define convergence for random variables. Another popular notion of convergence is convergence in distribution:

Definition 15. *A sequence of real-valued random variables $X_1, X_2, \ldots$ converges to X in distribution (or in law, or weakly) if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for every x at which F_X is continuous.*

Convergence in distribution represents an intermediate situation between the notions of almost sure convergence and in probability convergence:

Theorem 9. *The following implication holds: a.s. convergence $\Rightarrow$ weak convergence $\Rightarrow$ convergence in probability*

Theorem 10 (Strong Law of Large Numbers)**.** *Consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with finite mean μ and finite variance. Then, random variable $S_n = (1/n)\sum_{i=1}^{n} X_i$ converges almost surely to μ.*

The SLLN implies the Weak Low of Large Numbers where convergence almost sure is replaced by convergence in probability.

In a sense, the SLLN provides a bridge between the real world and probability theory. S_n is a variable that can be empirically observed. It corresponds, for instance, to the relative frequency of Heads in a sequence of coin tosses. On the other hand, μ is theoretically defined as the first moment of the distribution. The SLLN basically says that μ is not only a theoretical quantity, it has an operational interpretation as well since it matches the empirical frequency, asymptotically.

While the SLLN tells us that S_n/n approximates well the mean, it does not tell us how S_n/n may deviate from μ. The central limit theorem provides an approximate of the law of the deviation.

Theorem 11 (Central limit theorem (CLT))**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables such that $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$ (both assumed to be finite). Then*

$$\left(\lim_{n\to\infty} P\left(\frac{\sqrt{n}}{\sigma}(S_n - n\mu) \le x\right) =\right) \lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Hence, a proper scaling of $S_n - n\mu$ converges weakly (in distribution) to the normal random variable. Note the universality of the CLT: convergence happens for any law of the X_i's. The only requirement is that mean and variance are finite.

6 Moment Generating function and Characteristic function

The moment generating function and the characteristic function are of central importance for proving limit theorems for sums of independent random variables. We define them and list a few properties.

Definition 16 (Moment generating function)**.** *The moment generating function of a random variable X is defined as*

$$M_X(t) = E(e^{tX}).$$

Remark 9. *A few properties*

- $M_X(0) = 1$
- *for any, constants* $b, t > 0$ *we have the lower bound*

  $$M_X(t) = E(e^{tX}) \geq E(e^{t\min\{X,b\}}) \geq E(\mathbb{1}_{\{X \geq t\}} e^{t\min\{X,b\}}) = P(X \geq t)e^{tb}$$

- *The* $n-th$ *derivative evaluated at zero satisfies*

  $$M^{(n)}(t)|_{t=0} = E(X^n)$$

  (hence the term 'moment generating' function). Hence, knowing all the derivatives of M at a single point gives all the moments of the random variable. This can be verified by an immediate calculation, for instance,

  $$M^{(1)}(t) = \frac{d}{dt}E(e^{tX}) = E\left(\frac{d}{dt}e^{tX}\right) = E(Xe^{tX})$$

  and setting $t = 0$ *gives* $E(X)$*. Another way to see this is by first writing*

  $$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

  then taking expectation

  $$Ee^{tX} = 1 + tm_1 + \frac{t^2m_2}{2!} + \frac{t^3m_3}{3!} + \cdots$$

  where the m_i*'s are the moments. The* $n-th$ *derivative evaluated at* $t = 0$ *gives* m_n*.*

- *If* X *and* Y *are independent* $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$
- *If* $Y = aX$ *then* $M_Y(t) = M_X(a \cdot t)$

Exercise 14. *Compute the moment generating function of the binomial* $B(n, p)$ *(hint: try first for* $n = 1$ *and use a property above).*

The moment generating function has a disadvantage. To be well defined, the pdf $f_X(x)$ should decay sufficiently so that $f_X(x)e^{tx} \to 0$. The characteristic function, which we now define, is also a

random variable transform, but it is always well-defined, irrespectively of how fast the pdf/pmf decays. It has similar properties to the moment generating function.

Definition 17 (Characteristic function)**.** *The characteristic function of a random variable X is defined as*

$$\phi(t) = E(e^{itX}) ,$$

that is, the Fourier transform of the pmf/pdf.

Remark 10.

- $E(X^n) = i^n \cdot \phi^{(n)}(t)|_{t=0}$
- *if $Y = aX$ then $\phi_Y(t) = \phi_X(a \cdot t)$*
- *if X and Y are independent random variables, their sum $Z = X + Y$ has characteristic function*

$$\phi_Z(t) = \phi_X(t) \cdot \phi_Y(t) .$$