§ 1.4. Independence, conditional expectation, Borel-Cantelli lemma and Kolmogorov 0-1 laws.

(i). Conditional probability and independence of events.

For any two events, say $A$ and $B$, the conditional probability of $A$ given $B$ is defined as

$$P(A|B) = P(A \cap B) / P(B), \text{ if } P(B) \neq 0.$$ 

This is the chance of $A$ to happen, given $B$ has happened.

In common sense, the independence between events $A$ and $B$ should be, information about event $B$ happens/or not, does not change the chance of $A$ to happen/or not, and vice versa. In other words, whether $B$ (A) happens or not does not contain any information about whether $A$ (B) happens. Therefore the definition of independence should be $P(A|B) = P(A)$ or $P(B|A) = P(B)$. But to include that case of $P(A) = 0$ or $P(B) = 0$, the mathematical definition of independence is $P(A \cap B) = P(A)P(B)$, which is equivalent to $P(A^c \cap B) = P(A^c)P(B)$ or $P(A \cap B^c) = P(A)P(B^c)$ or $P(A^c \cap B^c) = P(A^c)P(B^c)$. The definition is extended in the following to independence between events.

**Definition** Events $A_1, ..., A_n$ are called independent if $P(\bigcap_{i=1}^n B_i) = \prod_{i=1}^n P(B_i)$ where $B_i$ is $A_i$ or $A_i^c$. Events $A_1, ..., A_n$ are called pairwise independent if any pair of two events are independent.

The above definition implies, if $A_1, ..., A_n$ are independent (pairwise independent), then $A_{i_1}, ..., A_{i_k}$ are independent (pairwise independent). (Please DIY).

The $\sigma$-algebras generated by a single set $A$, denoted as $\sigma(A)$ is $\{\emptyset, A, A^c, \Omega\}$. Independence between $A_1, ..., A_n$ can be interpreted as independence between the $\sigma$-algebras: $\sigma(A_i), i = 1, ..., n$.

(ii). Borel-Cantelli Lemma.

The Borel-Cantelli Lemma is considered as *sine qua non* of probability theory and is instrumental in proving the law of large numbers. Please note in the proof below the technique of using the indicator functions to handle probability of sets,

**Theorem 1.1. (BOREL-CANTELLI LEMMA)** For events $A_1, A_2, ...$,

1. $\sum_{n=1}^{\infty} P(A_n) < \infty \implies P(A_n, i.o.) = 0$;
2. If $A_n$ are independent, $\sum_{n=1}^{\infty} P(A_n) = \infty \implies P(A_n, i.o.) = 1$.

Here $A_n, i.o.$ means $A_n$ happens infinitely often, i.e., $\bigcap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$.

**Proof.** (1): Let $1_{A_n}$ be the indicator function of $A_n$. Then, $A_n, i.o.$ is the same as $\sum_{n=1}^{\infty} 1_{A_n} = \infty$.

Hence,

$$E(\sum_{i=1}^{\infty} 1_{A_n}) = \sum_{n=1}^{\infty} E1_{A_n} = \sum_{n=1}^{\infty} P(A_n) < \infty.$$ 

It implies $\sum_{i=1}^{n} 1_{A_n} < \infty$ with probability 1. This is equivalent to $P(A_n, i.o.) = 0$.

(2). $\sum_{n=1}^{\infty} P(A_n) = \infty$ implies $\prod_{k=n}^{\infty} (1 - P(A_k)) = 0$ since $\log(1-x) \leq -x$ for $x \in [0,1]$, for all $n \geq 1$. By dominated convergence theorem

$$E(\lim inf_{n} 1_{A_k}) = E(\lim_{n} \prod_{k=n}^{\infty} 1_{A_k}) = \lim_{n} E(\prod_{k=n}^{\infty} 1_{A_k}) = \lim_{n} \prod_{k=n}^{\infty} (1 - P(A_k)) = 0.$$ 

Then, $P(\lim inf_n A_n^c) = 0$ and hence $P(\lim sup_n A_n) = 1$. □

As an immediate consequence,
Example 1.2 Suppose \( A_1, ..., A_n, ... \) are independent events with \( \sum_n p_n = \infty \) where \( p_n = P(A_n) \). Then,

\[
X_n = \sum_{i=1}^{\infty} \frac{1}{n} A_i \rightarrow 1 \quad \text{a.s.}
\]

Proof Since

\[
E(X_n - 1)^2 = \sum_{i=1}^{\infty} \frac{1}{n} p_i (1 - p_i) \leq \frac{1}{\sum_{i=1}^{\infty} p_i} \rightarrow 0,
\]

it follows that \( X_n \rightarrow 1 \) in \( L_2 \) and therefore also in probability by the Chebyshev inequality:

\[
P(|X_n - 1| > \epsilon) \leq \frac{E(X_n - 1)^2}{\epsilon^2} \leq \frac{1}{\epsilon^2 \sum_{i=1}^{\infty} p_i} \rightarrow 0.
\]

Consider \( n_k \uparrow \infty \) as \( k \rightarrow \infty \), such that

\[
\sum_{k=1}^{\infty} \frac{1}{n_k} \sum_{i=1}^{n_k} p_i < \infty \quad \text{and} \quad \sum_{i=1}^{n_k+1} p_i \rightarrow 1.
\]

Then,

\[
\sum_{k=1}^{\infty} P(|X_{n_k} - 1| > \epsilon) < \infty.
\]

The Borel-Cantelli lemma implies \( X_{n_k} \rightarrow 1 \) a.s.. Observe that, for \( n_k \leq n \leq n_{k+1}, \)

\[
1 \leftarrow \sum_{i=1}^{n_k} \frac{1}{n_{k+1}} A_i \leq X_n = \sum_{i=1}^{n} \frac{1}{n_i} A_i \leq \frac{\sum_{i=1}^{n_k+1} 1}{\sum_{i=1}^{n_k+1} p_i} \rightarrow 1, \quad \text{a.s.}
\]

The desired convergence holds.

Remark. The trick of bracketing \( X_n \) by the two quantities in the above inequality is also used in proving the uniform convergence of the empirical distribution to the population distribution:

\[
|F_n(x) - F(x)| \rightarrow 0, \quad \text{a.s.},
\]

where \( F_n(x) = (1/n) \sum_{i=1}^{n} 1_{x_i \leq x} \) and \( \xi_i \) are iid with cdf \( F \). The idea is further elaborated in the context of empirical approximation in terms of bracketing/packing numbers.

Example 1.3. Repeatedly toss a coin, which has probability \( p \) to be head and \( q = 1 - p \) to be tail on each toss. Let \( X_n = H \) or \( T \) when \( n \)-th toss is a head or tail. Let

\[
l_n = \max\{m \geq 0 : X_n = H, X_{n+1} = H, ..., X_{n+m-1} = H, X_{n+m} = T\}
\]

be the length of run of heads starting from \( n \)-th toss. Then,

\[
\limsup_n l_n / \log n = 1 / \log(1/p).
\]

Proof. \( l_n \) follows a geometric distribution, i.e.,

\[
P(l_n = k) = pq^k, \quad P(l_n \geq k) = P(X_n = 1, ..., X_{n+k-1} = 1) = p^k \quad k = 0, 1, 2, ...
\]
For any $\epsilon > 0$, 
\[
\sum_{n=1}^{\infty} P \left( l_n > (1 + \epsilon) \frac{\log n}{\log(1/p)} \right) \leq \sum_{n=1}^{\infty} p^{(1+\epsilon) \frac{\log n}{\log(1/p)}} \leq \sum_{n=1}^{\infty} e^{-(1+\epsilon) \log n} = \sum_{n=1}^{\infty} n^{-(1+\epsilon)} < \infty
\]

By the Borel-Cantelli lemma, 
\[
\limsup_{n} \frac{l_n}{\log n/\log(1/p)} \leq 1.
\]

We next try to find a subsequence with limit as large as 1. Choose $r_n = n^\gamma$ (we need a sequence going fast to infinity so that the following $A_n$ are independent). Let $d_n$ be the integer part of $\log n/\log(1/p)$ and let 
\[
A_n = \{X_{r_n} = H, X_{r_n+1} = H, \ldots, X_{r_n+d_n-1} = H\}
\]

Then $A_n, n \geq 1$ are independent, and 
\[
P(A_n) = p^{d_n} = e^{d_n \log p} \approx 1/n
\]

Therefore, $\sum_n P(A_n) = \infty$. It then follows from the Borel Cantelli lemma that $P(A_n, i.o.) = 1$. Since $A_n = \{l_n \geq d_n\}$, we have 
\[
\limsup_{n} \frac{l_n}{\log n/\log(1/p)} \geq \limsup_{n} \frac{l_n}{d_n} \geq 1.
\]

Remark. An analogous problem occurs in the setting of Poisson processes. Consider a Poisson process with intensity $\lambda > 0$. The sojourn times (time between two consecutive events) $\xi_0, \xi_1, \ldots$ are iid $\sim$ exponential distribution with mean $1/\lambda$. Then, $\limsup_{x \to \infty} l_x/x = 1/\lambda$, where $l_x$ the time period between $x$ and the time of the event right after $x$.

(iii). Independence between $\sigma$-algebras and between random variables.

Definitions. Let $A_1, \ldots, A_n$ be $\sigma$-algebras. They are called independent if $A_1, \ldots, A_n$ are independent for any $A_j \in A_j, j = 1, \ldots, n$. Random variables $X_1, \ldots, X_n$ are called independent, if the $\sigma$-algebras generated by $X_j, 1 \leq j \leq n$, are independent, i.e.,
\[
P(\bigcap_{j=1}^{n} X_j^{-1}(B_j)) = \prod_{j=1}^{n} P(X_j^{-1}(B_j)) \quad \text{or} \quad P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{j=1}^{n} P(X_j \in B_j)
\]

for any Borel sets $B_1, \ldots, B_n$ in $(-\infty, \infty)$.

There are several equivalent definition of the independence of random variables:

Two r.v.s $X$ and $Y$ are called independent, if $E(g(X)f(Y)) = E(g(X))E(f(Y))$ for all bounded (measurable) functions $g$ and $f$. or, equivalently, if 
\[
P(X \leq t, \text{ and } Y \leq s) = \prod_{i=1}^{n} P(X_i \leq t_i) \quad \text{for all } t_j \in (-\infty, \infty), j = 1, \ldots, n.
\]

i.e., in terms of cumulative distribution functions. 
\[
F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{for all } x, y.
\]

If the joint density exists, This is the same as $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Roughly speaking, independence between two r.v.s $X$ and $Y$ is interpreted as $X$ taking any value “has nothing to do with” $Y$ taking any value, and vice versa.
(iv). Conditional expectation.

(1). Conditional distribution and conditional expectation with respect to a set $A$.

Suppose $A$ is a set with $P(A) > 0$, and $X$ is a random variable. Then, the *conditional expectation* is

$$E(X|A) = E(X1_A)/P(A).$$

The *conditional distribution* of $X$ given $A$ is

$$P(X \leq t|A) = P(\{X \leq t\} \cap A)/P(A).$$

Then, $E(X|A) = \int t \, dP(X \leq t|A)$, if exist.

As a simple example, let $X \sim Unif[0, 1]$. Let $A_i = \{ i - 1/n < X \leq i/n \}$ for $i = 1, ..., n$.

$$E(X|A_i) \equiv E(X1_{A_i})/P(A_i) = (i - 1/2)/n.$$ 

Similarly $E(X|A_i^c) \equiv E(X1_{A_i^c})/P(A_i^c)$.

Interpretation: $E(X|A)$ is the weighted “average” (expected value) of $X$ over the set $A$.

(2). Conditional expectation with respect to a r.v.

For two random variables $X, Y$, $E(X|Y)$ is a function of $Y$, i.e., measurable to $\sigma(Y)$, such that, for any $A \in \sigma(Y)$,

$$E(X|A) = E[E(X|Y)1_A].$$

Interpretation: $E(X|Y)$ is the weighted “average” (expected value) of $X$ over the set $\{ Y = y \}$ for all $y$. It is a function of $Y$ and therefore is a r.v. measurable to $\sigma(Y)$.

If their joint density $f(x, y)$ exists, then the conditional density of $X$ given $Y = y$ is $f_{X|Y}(x|y) \equiv f(x, y)/f_Y(y)$. And

$$E(X|Y = y) \equiv \int x f_{X|Y}(x|y)dx.$$

(3). Conditional expectation with respect to a $\sigma$-algebra $\mathcal{A}$.

Conditional expectation w.r.t. a $\sigma$-algebra is the most fundamental concept in probability theory, especially in martingale theory in which the very definition of martingale depends on conditional expectation.

Recall that a random variable, say $X$, is measurable to a $\sigma$-algebra $\mathcal{A}$ is that for any interval $(a, b)$, $\{ \omega : X(\omega) \in (a, b) \} \in \mathcal{A}$. In other words, $\sigma(X) \subseteq \mathcal{A}$ is interpreted as all information about $X$, (which is $\sigma(X)$), is contained in $\mathcal{A}$.

If $\mathcal{A} = \sigma(A_1, ..., A_n)$ where $A_i \cap A_j = \emptyset$, then $X$ measurable to $\mathcal{A}$ implies $X$ must be constant over each $A_i$. If $\mathcal{A}$ is generated by a r.v. $Y$, then $X$ measurable to $\mathcal{A}$ implies $X$ must be a function of $Y$.

A heuristic understanding is that if $Y$ is known, then there is no uncertainty of $X$, or if $Y$ assumes one value, $X$ cannot assume more than one values.

**Definition** For a random variable $X$ and a completed $\sigma$-algebra $\mathcal{A}$, $E(X|A)$ is defined as an $\mathcal{A}$-measurable random variable such that, for any $A \in \mathcal{A}$,

$$E(X1_A) = E(E(X|A)1_A),$$

i.e. $E(X|A) = E(E(X|A)|A)$ for every $A \in \mathcal{A}$ with $P(A) > 0$.

If $\mathcal{A} = \sigma(A_1, ..., A_n)$ where $A_i \cap A_j = \emptyset$, then

$$E(X|A) = \sum_{j=1}^{n}E(X|A_i)1_{A_i},$$

which is a r.v. that, on each $A_i$, takes the conditional average of $X$, i.e., $E(X|A_i)$, as its value.

Motivated from this simple case, we may obtain an important understanding of the conditional
Conditional mean/expectation with respect to σ algebra shares many properties just like the ordinary expectation.

Properties:

1. \( E(aX + bY | \mathcal{A}) = aE(X | \mathcal{A}) + bE(Y | \mathcal{A}) \)
2. If \( X \in \mathcal{A} \), then \( E(X | \mathcal{A}) = X \).
3. \( E(E(X | \mathcal{F}) | \mathcal{A}) = E(X | \mathcal{A}) \) for two σ-algebras \( \mathcal{A} \subseteq \mathcal{F} \).

Further properties, such as the dominated convergence theorem, Fatou’s lemma and monotone convergence theorem also hold for conditional mean w.r.t. a σ-algebra. (See DIY exercises.)

(v). Kolmogorov’s 0-1 law.

One of the most important theorem in probability theory is the martingale convergence theorem. In the following, we provide a simplified version, without a rigorous introduction of martingale and without giving a proof.

**Theorem 1.2 (Simplified Version of Martingale Convergence Theorem)** Suppose \( \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \) for \( n \geq 1 \). Let \( \mathcal{F} = \sigma(\cup_{n=1}^{\infty} \mathcal{F}_n) \). For any random variable \( X \) with \( E(|X|) < \infty \),

\[
E(X | \mathcal{F}_n) \rightarrow E(X | \mathcal{F}), \quad \text{a.s.}
\]

The martingale convergence theorem, even with the simplified version, has broad applications. For example, one of the most basic 0-1 laws: the Kolomogorov 0-1 law, can be established upon it.

**Corollary (Kolomogorov 0-1 Law)** Suppose \( X_1, ..., X_n, ... \) are a sequence of independent r.v.s. Then all tails events have probability 0 or 1.

**Proof.** Suppose \( A \) is a tail event. Then \( A \) is independent of \( X_1, ..., X_n \) for any fixed \( n \). Therefore \( E(1_A | \mathcal{F}_n) = P(A) \) where \( \mathcal{F}_n \) is the σ-algebra generated by \( X_1, ..., X_n \). But, by Theorem 1.2, \( E(1_A | \mathcal{F}_n) \rightarrow 1_A \) a.s. Hence \( 1_A = P(A) \), and \( A \) can only be 0 or 1. \( \square \)

A heuristic interpretation of Kolmogorov’s 0-1 law could be in the perspective of information. When σ-algebras \( \mathcal{A}_1, ..., \mathcal{A}_n, ... \) are independent, the information carried by each \( \mathcal{A}_i \) are independent or unrelated or non-overlapping. Then, the information carried by \( \mathcal{A}_n, \mathcal{A}_{n+1}, ... \) shall shrink to 0 as \( n \rightarrow \infty \), as, if otherwise, \( \mathcal{A}_n, \mathcal{A}_{n+1}, ... \) would have something in common.

As straightforward applications of Kolmogorov’s 0-1 law:

**Corollary** Suppose \( X_1, ..., X_n, ... \) are a sequence of independent random variables. Then,

\[
\lim_{n \to \infty} \inf X_n, \quad \lim_{n \to \infty} \sup X_n, \quad \lim_{n \to \infty} \sup S_n/a_n \quad \text{and} \quad \lim_{n \to \infty} \inf S_n/a_n
\]

must be either a constant or \( \infty \) or \( -\infty \), a.s., where \( S_n = \sum_{i=1}^{n} X_i \) and \( a_n \uparrow \infty \).

**Proof.** Consider \( A = \{ \omega : \liminf X_n(\omega) > a \} \). Try to show \( A \) is a tail event. (DIY) \( \square \)

Remark. Without invoking martingale convergence theorem, Kolmogorov’s 0-1 law can be shown through \( \pi - \lambda \) theorem, which we do not plan to cover.

**DIY Exercises.**

**Exercise 1.13** ** Suppose \( X_n \) are iid random variables. Then \( X_n/n^{1/p} \to 0 \) a.s. if and only if \( E(|X_n|^p) < \infty \) for \( p > 0 \). Hint: Borel-Cantelli lemma.

**Exercise 1.14** ** Suppose \( X_n \) be iid r.v.s with \( E(X_n) = \infty \). Show that \( \limsup_n |S_n|/n = \infty \) a.s. where \( S_n = X_1 + \cdots + X_n \).

**Exercise 1.15** ** Suppose \( X_n \) are iid nonnegative random variables such that \( \sum_{k=1}^{\infty} kP(X_1 > a_k) < \infty \) for \( a_k \uparrow \infty \). Show that \( \limsup_n \max_{1 \leq i \leq n} X_i/a_n \leq 1 \) a.s.
Exercise 1.16  ⭐⭐⭐⭐ (Empirical Approximation) For every fixed \( t \in [0, 1] \), \( S_n(t) \) is a sequence of random variables such that, with probability 1 for some \( p > 0 \),

\[
|S_n(t) - S_n(s)| \leq n|t - s|^p,
\]

for all \( n \geq 1 \) and all \( t, s \in [0, 1] \). Suppose for every constant \( C > 0 \), there exists an \( c > 0 \) such that

\[
P(|S_n(t)| > C(n \log n)^{1/2}) \leq e^{-cn}
\]

for all \( n \geq 1 \) and \( t \in [0, 1] \).

Show that, for any \( p > 0 \),

\[
\frac{\max\{|S_n(t)| : t \in [0, 1]\}}{(n \log n)^{1/2}} \to 0 \quad \text{a.s.}
\]

Hint: Borel-Cantelli lemma.