CHAPTER 1: MEASURE THEORY

September 7, 2021

Contents

1 Probability Space $(\Omega, \mathcal{F}, P)$

1.1 Definition; Properties

Definition. A **sample space**, denoted by Ω, is a set (of "outcomes").

Definition. A collection of subsets of Ω, denoted by $\mathcal{F}$, is called a σ**-field** or σ**-algebra** if
(i) $\Omega \in \mathcal{F}$,
(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,
(iii) If $A_1, A_2, \cdots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
We often refer to elements of $\mathcal{F}$ as **events**.

Example. The smallest σ-field is $\{\emptyset, \Omega\}$; The largest σ-field is $\{\text{All subsets of } \Omega\}$.

Fact. If $\mathcal{F}_i, i \in I$ are all σ-fields, then $\cap_{i \in I} \mathcal{F}_i$ is a σ-field.

Definition. The above $(\Omega, \mathcal{F})$ is called a **measurable space**.

Definition. $\mu : \mathcal{F} \to \mathbb{R}$ is called a **measure** if
(1) $\mu(A) \geq 0, \ \forall \ A \in \mathcal{F}$,
(2) $\mu(\emptyset) = 0$,
(3) If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint, then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Properties of Measures. Let μ be a measure on a measurable space $(\Omega, \mathcal{F})$. Then:
(a) If $A \subset B$, then $\mu(A) \leq \mu(B)$. (Monotonicity.)
(b) $\forall \ A, B$, $\mu(A) + \mu(B) = \mu(A \cup B) + \mu(A \cap B)$. (Addition law.)
(c) If $A \subset \cup_{i=1}^{\infty} A_i$, then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$. (Sub-additivity.)
(d) If $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$. (Continuity from below.)
(e) If $A_n \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_n) \downarrow \mu(A)$. (Continuity from above.)

Proof. The basic idea is to consider disjoint events and use (1)–(3) in the definition of measure.

Proof of (a): Note that A and $B \backslash A$ are disjoint. We have

$$\mu(B) = \mu(A \cup (B \backslash A)) \overset{(2)}{=} \mu(A) + \mu(B \backslash A) \overset{(1)}{\geq} \mu(A).$$

Proof of (b): Write each term as a sum involving measures of the disjoint events $B \backslash A$, $A \cap B$ and $A \backslash B$ and use (3).

Proof of (c): Write A as a disjoint union of events

$$A = A \cap (\cup_{i=1}^{\infty} A_i) = (A \cap A_1) \cup (A \cap (A_2 \backslash A_1)) \cup (A \cap (A_3 \backslash (A_1 \cup A_2))) \cup \dots.$$

From (3) and (a), we have

$$\begin{aligned} \mu(A) =& \mu(A \cap A_1) + \mu(A \cap (A_2 \backslash A_1)) + \mu(A \cap (A_3 \backslash (A_1 \cup A_2))) + \dots \\ \leq& \mu(A_1) + \mu(A_2) + \mu(A_3) + \dots. \end{aligned}$$

Proof of (d): Let $B_1 = A_1$ and $B_i = A_i \backslash A_{i-1}$ for $i \geq 2$. Note that B_i's are disjoint and their union is A. Therefore,

$$\mu(A) = \mu(\cup_{i=1}^{\infty} B_i) \overset{(3)}{=} \sum_{i=1}^{\infty} \mu(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(B_i) = \lim_{n\to\infty} \mu(A_n).$$

Proof of (e): Consider $(A_1 \backslash A_n) \uparrow (A_1 \backslash A)$ and use (d).

$\square$

Definition. If $\exists A_i \uparrow \Omega$ with $\mu(A_i) < \infty$, then μ is called a **σ-finite measure**.
If $\mu(\Omega) < \infty$, then μ is called a **finite measure**.
If $\mu(\Omega) = 1$, then μ is called a **probability measure**.

Definition. Let $\mathcal{A}$ be a collection of subsets of Ω, $\sigma(\mathcal{A})$ denotes the **smallest σ-field containing** $\mathcal{A}$, or equivalently,

$$\sigma(\mathcal{A}) = \cap_{\mathcal{A}\subset\mathcal{F}, \mathcal{F} \text{ is a } \sigma\text{-field}} \mathcal{F}.$$

Example. If $\mathcal{A} = \{A\}$, then $\sigma(\mathcal{A}) = \{\emptyset, \Omega, A, A^c\}$.

Definition. A collection of subsets of Ω, $\mathcal{F}$, is called a **field** or **algebra** if
(i) $\Omega \in \mathcal{F}$,
(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,
(iii) If $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then $\cup_{i=1}^{n} A_i \in \mathcal{F}$.

Fact. Any σ-field is a field, but not vice versa. Consider the counter-example that $\Omega = \mathbb{Z}$, $\mathcal{F} = \{A \subset \mathbb{Z} :$ either A or A^c is a finite set$\}$.

1.2 Measures on $\mathbb{R}^d$; π-λ Theorem

Now we focus on sample space being the Euclidean space $\Omega = \mathbb{R}^d$.
Definition. **Borel σ-field** on $\mathbb{R}$, denoted by $\mathcal{B}$ or $\mathcal{R}$, is defined to be

$$\mathcal{B} = \sigma(\{(a, b] : -\infty < a < b < \infty\}).$$

Fact. $\mathcal{B}$ can be equivalently defined to be

$$\mathcal{B} = \sigma(\{(a, b) : -\infty < a < b < \infty\}) = \sigma(\{\text{Open sets in } \mathbb{R}\}).$$

Definition. **Borel σ-field** on $\mathbb{R}^d$, denoted by $\mathcal{B}$ or $\mathcal{R}^d$, is defined to be

$$\mathcal{B} = \sigma(\{(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d] : -\infty < a_i < b_i < \infty\}).$$

Next, we focus on probability measures on $\mathbb{R}$.

Definition. $F : \mathbb{R} \to \mathbb{R}$ is called a Stieltjes measure function if

(i) F is nondecreasing,

(ii) F is right-continuous, i.e., $\lim_{y \downarrow x} F(y) = F(X)$.

Fact. Every measure μ on $(\mathbb{R}, \mathcal{R})$ s.t. $\mu((a,b]) < \infty$ for any $-\infty < a < b < \infty$ determines a Stieltjes measure function F (up to constants) $F(0) = c$ and

$$F(x) = \begin{cases} c + \mu((0,x]) & \text{if } x > 0 \\ c - \mu((x,0]) & \text{if } x < 0. \end{cases}$$

The main result in this subsection is to show that probability measures on $\mathbb{R}$ are determined by distribution functions. This means the cumulative distribution function (cdf) we learned in the elementary probability course actually determines a probability measure on $\mathbb{R}$. This is a special case of the following theorem.

Theorem. Every Stieltjes measure function F determines a unique measure μ on $(\mathbb{R}, \mathcal{R})$ such that

$$\mu((a,b]) = F(b) - F(a), \quad \forall \; -\infty < a < b < \infty. \tag{1.1}$$

We only prove that such a measure is unique if $\mu((-\infty, \infty)) < \infty$. We need Dynkin's π-λ theorem for this purpose. We first state the π-λ theorem, then use it to prove the uniqueness, finally prove the π-λ theorem.

Definition. $\mathcal{P}$ is a **π-system** if

$$A, B \in \mathcal{P} \Longrightarrow A \cap B \in \mathcal{P}.$$

Example. $\{(a,b] : -\infty < a \le b < \infty\}$ is a π-system.

Definition. $\mathcal{L}$ is a **λ-system** if

(1) $\Omega \in \mathcal{L}$,

(2) If $A, B \in \mathcal{L}$ and $A \subset B$, then $B \backslash A \in \mathcal{L}$,

(3) If $A_1, A_2, \cdots \in \mathcal{L}$ and $A_i \uparrow A$, then $A \in \mathcal{L}$.

Fact. If $\mathcal{F}$ is both a π-system and a λ-system, then $\mathcal{F}$ is a σ-field.

Proof. We need to verify that if $A_1, A_2, \cdots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$, which is (iii) in the definition of σ-fields. This following by observing

$$\cup_{i=1}^{\infty} A_i = A_1 \cup (A_1 \cup A_2) \cup (A_1 \cup A_2 \cup A_3) \cup \cdots,$$

$A_1 \cup A_2 = (A_1^c \cap A_2^c)^c$, and using the definitions of λ-system and π-system. □

Dynkin's π-λ Theorem. If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$. Recall $\sigma(\mathcal{P})$ is the smallest σ-field containing $\mathcal{P}$.

Proof of uniqueness by the π-λ theorem. Note that the Stieltjes measure function F determines the value of the measure on

$$\mathcal{P} := \{(a, b] : -\infty < a \le b < \infty\} \qquad \text{(this is a } \pi\text{-system as discussed above)}$$

through (1.1). It suffices to show that if two measures μ_1 and μ_2 agree on $\mathcal{P}$, then they agree on $\mathcal{R} = \sigma(\mathcal{P})$. To this end, we define

$$\mathcal{L} := \{A \in \mathcal{R} : \mu_1(A) = \mu_2(A)\}.$$

By the π-λ theorem, we are only left to show that $\mathcal{L}$ is a λ-system. (1)–(3) in the definition of λ-system follows by the addition law and the continuity from below properites of measures. $\square$

Sketch of the proof for the π-λ theorem. The π-λ theorem follows from

(a): If $\lambda(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$, then $\lambda(\mathcal{P})$ is a σ-field.

To prove (a), it suffices to show that

(b): $\lambda(\mathcal{P})$ is closed under intersection.

To prove (b), we let

$$g_A = \{B \in \lambda(\mathcal{P}) : A \cap B \in \lambda(\mathcal{P})\}$$

and prove

(c): If $A \in \lambda(\mathcal{P})$, then g_A is a λ-system.

(c) can be verified directly by checking (1)–(3) in the definition of the λ-system. $\square$

2 Random Variables X and their Distributions $\mathcal{L}(X)$

2.1 Random Variable

Definition. Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. $f : \Omega_1 \to \Omega_2$ is called **measurable** if for any $A \in \mathcal{F}_2$, $f^{-1}(A) \in \mathcal{F}_1$, where $f^{-1}(A) = \{w_1 \in \Omega_1 : f(w_1) \in A\}$.

Fact. $\{f^{-1}(A) : A \in \mathcal{F}_2\}$ is a σ-field in Ω_1; $\{A \subset \Omega_2 : f^{-1}(A) \in \mathcal{F}_1\}$ is a σ-field in Ω_2.

As a consequence, if $\mathcal{F}_2 = \sigma(\mathcal{A}_2)$, then to check f is measurable, we only need to check $\forall\ A \in \mathcal{A}_2,\ f^{-1}(A) \in \mathcal{F}_1$.

Proposition. Let $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2), (\Omega_3, \mathcal{F}_3)$ be measurable spaces. If $f_1 : \Omega_1 \to \Omega_2$ and $f_2 : \Omega_2 \to \Omega_3$ are both measurable, then $f_2 \circ f_1 : \Omega_1 \to \Omega_3$ is measurable.

Definition. If there is a measure μ_1 on $(\Omega_1, \mathcal{F}_1)$, through a measurable function $f : \Omega_1 \to \Omega_2$, we define a measure on $(\Omega_2, \mathcal{F}_2)$ by

$$\mu_2(A) = \mu_1(f^{-1}(A)).$$

Such μ_2 is called the **induced measure**.

Definition. Let $(\Omega, \mathcal{F})$ be a measurable space. Recall that $(\mathbb{R}, \mathcal{R})$ and $(\mathbb{R}^d, \mathcal{R}^d)$ are Euclidean spaces equipped with Borel σ-fields. If $f : \Omega \to \mathbb{R}$ is measurable, then f is called a real-valued (or one-dimensional) **random variable**, usually denoted by X. If $f : \Omega \to \mathbb{R}^d$, $d \geq 2$, is measurable, then f is called a d-dimensional random variable (or a **random vector**), usually denoted by $X = (X_1, \ldots, X_d)^\top$.

Proposition. $X = (X_1, \ldots, X_d)^\top$ is a random vector if and only if X_i is a random variable for all $1 \leq i \leq d$.

Proof.

"$\Longrightarrow$":

$$X_i^{-1}((a, b]) = X^{-1}(\mathbb{R} \times \cdots \times \mathbb{R} \times (a, b] \times \mathbb{R} \times \cdots \times \mathbb{R}) \in \mathcal{F}.$$

"$\Longleftarrow$":

$$X^{-1}((a_1, b_1] \times \cdots \times (a_d, b_d]) = [X_1^{-1}((a_1, b_2])] \cap \cdots \cap [X_d^{-1}((a_d, b_d])] \in \mathcal{F}.$$

$\square$

As a consequence: If $X_1, \ldots, X_n$ are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \to (\mathbb{R}, \mathcal{R})$ is a measurable function, then $f(X_1, \ldots, X_n)$ is a random variable.

Therefore, the usual algebraic operations of random variables results in a random variable. For example, $X_1 + \cdots + X_n$ is a random variable. This also applies to limits, as shown in the next theorem.

Theorem 1.3.5. If $X_1, X_2, \dots$ are random variables, then so are

$$\inf_{n\geq 1} X_n, \quad \sup_{n\geq 1} X_n, \quad \limsup_{n\to\infty} X_n, \quad \liminf_{n\to\infty} X_n,$$

regarded as functions from Ω to the extended real line $([-\infty, \infty], \mathcal{R}^*)$ equipped with the σ-algebra generated by $\mathcal{R} \cup \{-\infty\} \cup \{\infty\}$.

Proof.

$$\{\inf_{n\geq 1} X_n < a\} = \cup_{n\geq 1}\{X_n < a\} \in \mathcal{F}.$$

$$\{\sup_{n\geq 1} X_n > a\} = \cap_{n\geq 1}\{X_n > a\} \in \mathcal{F}.$$

$$\limsup_{n\to\infty} = \inf_{n\geq 1}(\sup_{m\geq n} X_m).$$

$$\liminf_{n\to\infty} = \sup_{n\geq 1}(\inf_{m\geq n} X_m).$$

□

Note that

$$\Omega_0 := \{\omega \in \Omega : \lim_{n\to\infty} X_n \text{ exists}\} = \{\omega \in \Omega : \limsup_{n\to\infty} X_n - \liminf_{n\to\infty} X_n = 0\} \in \mathcal{F}.$$

Definition. If $\mu(\Omega_0) = \mu(\Omega)$, then we say X_n converges **almost everywhere (a.e.)**. If $\mu(\Omega_0) = \mu(\Omega) = 1$, then we say X_n converges **almost surely (a.s.)**.

2.2 Distribution

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X : \Omega \to \mathbb{R}$ be a real valued random variable. The induced measure

$$\mu(A) := P(\{w \in \Omega : X(w) \in A\}) =: P(X \in A)$$

is called the **probability measure** (or **probability distribution**) of X.

Definition. The **distribution function (d.f.)** of X is defined to be $F : \mathbb{R} \to [0,1]$,

$$F(x) = F_X(x) = P(X \leq x).$$

Properties of d.f.. (a) F is non-decreasing.
(b) F is right-continuous.
(c) $\lim_{x\to-\infty} F(x) = 0$; $\lim_{x\to\infty} F(x) = 1$.

These properties are inherited from the properties of measures.

Example. If

$$F(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < 1 \\ 1, & x \ge 1, \end{cases}$$

then it is called the uniform distribution.

Proposition. If X has a continuous d.f. F, then $Y := F(X)$ has the uniform distribution.

Proof. For $0 < y < 1$ (Here F^{-1} denotes the largest value among the preimage) :

$$P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y)) = F(F^{-1}(y)) \overset{\text{by continuity}}{=} y.$$

□

Next theorem provides a way of constructing a random variable with an arbitrary distribution.

Theorem. Let $\Omega = (0,1)$, $\mathcal{F} = \{\text{Borel sets}\}$, $P =$Lebesgue measure. Define $X : \Omega \to \mathbb{R}$ to be

$$X(\omega) = F^{-1}(\omega),$$

where

$$F^{-1}(\omega) := \inf\{y : F(y) \ge \omega\} = \sup\{y : F(y) < \omega\}.$$

Then the d.f. of X is F.

Proof. Note that

$$P(X \le x) = P(\{\omega : F^{-1}(\omega) \le x\}),$$
$$F(x) = P(\{\omega : \omega \le F(x)\}).$$

The right-hand-sides are equal by the definition of F^{-1}; hence $P(X \le x) = F(x)$. □

Definition. X and Y are said to be **equal in distribution** if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

Definition. The **support** of a random variable X with d.f. F is defined to be

$$\{x \in \mathbb{R} : F(x + \varepsilon) - F(x - \varepsilon) > 0,\ \forall\, \varepsilon > 0\}.$$

Definition. Denote the set of discontinuity points of F (which must be countable) by

$$\{a_1, a_2, \cdots\}.$$

Let $b_j = F(a_j) - F(a_j-) > 0$.

If $\sum_{j=1}^{\infty} b_j = 1$, then F is called a **discrete distribution**.

If $\sum_{j=1}^{\infty} b_j = 0$, then F is called a **continuous distribution**.

If $F(x) = \int_{-\infty}^{x} f(y)dy$, then F is called **absolutely continuous** and has **density function** f.

Theorem. Any distribution function F can be written as

$$F = c_1 F_d + c_2 F_a + c_3 F_s,$$

where $c_1, c_2, c_3 \geq 0$, $c_1+c_2+c_3 = 1$, F_d is a discrete d.f., F_a is an absolutely continuous d.f., and F_s is a singular distribution function, meaning that F_s' exists and equals to 0 almost everywhere.

Definition. Let $X = (X_1, \ldots, X_d)^\top$ be a $\mathbb{R}^d$-valued random vector. The **distribution function** of X is defined to be $F : \mathbb{R}^d \to [0,1]$ and for $x = (x_1, \ldots, x_d)^\top$,

$$F(x) = P(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Note that X and Y are allowed to be defined on different probability spaces; or be two different random variables on the same probability space.

2.3 Examples

- Normal distribution, denoted by $N(\mu, \sigma^2)$, has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

- Exponential distribution, denoted by $\exp(\lambda)$, has density function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

- Poisson distribution, denoted by $Poisson(\lambda)$, has probability mass function

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

- Lognormal, chi-square, Gamma, Cauchy, Beta, ...

Properties of $N(0,1)$. Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$. Then for $x > 0$,

$$(\frac{1}{x} - \frac{1}{x^3})\phi(x) \leq 1 - \Phi(x) \leq \min\{\frac{1}{x}\phi(x), \frac{1}{2}e^{-x^2/2}\}.$$

3 Expectation $E(X)$

3.1 Definition

Let X be a random variable defined on $(\Omega, \mathcal{F}, P)$. The **expectation** of X is defined in four steps.

Definition 1. Given a set $A \in \mathcal{F}$, define

$$X(\omega) = 1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A. \end{cases}$$

Such a random variable is called an **indicator random variable** and its expectation is defined to be

$$E(1_A) := P(A).$$

Definition 2. Let $X = \sum_{i=1}^n a_i 1_{A_i}$, where $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint and $a_1, \ldots, a_n \in \mathbb{R}$. Such a random variable is called a **simple random variable** and its expectation is defined to be

$$E(X) = \sum_{i=1}^n a_i P(A_i).$$

Definition 3. For a **nonnegative random variable**, i.e., $X(w) \geq 0 \ \forall \ w \in \Omega$, define

$$E(X) := \sup_{\substack{Y: 0 \leq Y \leq X \\ Y \text{ is a simple random variable}}} E(Y).$$

Note: It can be $+\infty$.

Definition 4. For an arbitrary random variable X, write $X = X^+ - X^-$, where

$$X^+ = \max\{X, 0\}, \quad X^- = \max\{-X, 0\}.$$

$E(X^+)$ and $E(X^-)$ are defined as in Definition 3.

If $E(X^+) = E(X^-) = \infty$, then we say the expected value of X does not exist.

Otherwise, define

$$E(X) = E(X^+) - E(X^-).$$

If both $E(X^+)$ and $E(X^-)$ are finite, then $E(X)$ and $E(|X|)$ are also finite.

blueDefinitions 3 and 4 can be defined similarly for generalized random variables taking values on $[-\infty, \infty]$.

Note that according to the above definitions, set with measure 0 can be neglected in the expectation. For example, if

$$X = \begin{cases} 0 & \text{in } \Omega_0 \\ \infty & \text{in } \Omega_0^c \end{cases}$$

and $P(\Omega_0) = 1$, then $E(X) = 0$. For another example, if $X = Y$ a.s., then $E(X) = E(Y)$ if it exists.

3.2 Properties

Properties. Suppose $X, Y \geq 0$ or $E|X|, E|Y| < \infty$. We have:

(a) If $X \geq Y$ a.s., then $E(X) \geq E(Y)$ (monotonicity)

(b) $E(X + Y) = E(X) + E(Y)$ (linearity)

Proof. Monotonicity follows easily from Definitions 1–4 of expectations. In the following, we prove the linearity.

If X and Y are simple random variables, then (b) follows from the definition 2 of expectations and simple algebra. We omit the details.

We now consider the case $X, Y \geq 0$ and $X, Y \leq n$ (later we will send $n \to \infty$). Let M to an integer such that $M \geq 2n$. Divide the interval $[0, M]$ into equally distributed subintervals of length $1/2^M$. For any nonnegative random variable W and $W \leq 2n$, we define

$$W_M^{(l)} := \lfloor 2^M W \rfloor / 2^M, \quad W_M^{(u)} := W_M^{(l)} + \frac{1}{2^M},$$

where $\lfloor \cdot \rfloor$ denotes the integer part of a real number. It can be easily checked that both $W_M^{(l)}$ and $W_M^{(u)}$ are simple random variables, and moreover,

$$W_M^{(l)} \leq W \leq W_M^{(u)}.$$

Therefore,

$$\begin{aligned} E(X+Y) &\leq E[(X+Y)_M^{(u)}] \\ &\leq E[X_M^{(u)} + Y_M^{(u)}] && \text{(from the sub-additivity of the operator } (\cdot)_M^{(u)}) \\ &= E[X_M^{(u)}] + E[Y_M^{(u)}] && \text{(from the linearity for simple random variables)} \\ &\leq E(X) + \frac{1}{2^M} + E(Y) + \frac{1}{2^M}. \end{aligned}$$

This implies $E(X + Y) \leq E(X) + E(Y)$ by sending $M \to \infty$. Similarly, we can prove $E(X+Y) \geq E(X)+E(Y)$ by working with $(\cdot)_M^{(l)}$. Therefore, we have proved $E(X+Y) = E(X) + E(Y)$ for the case $X, Y \geq 0$ and $X, Y \leq n$.

Next, we consider the case $X, Y \geq 0$ (not necessarily bounded). For any positive number n, we can easily check that

$$(X \wedge n) + (Y \wedge n) \leq (X + Y) \wedge 2n \leq (X \wedge 2n) + (Y \wedge 2n).$$

Taking expectations and using (a) and (b) for the bounded case, we have

$$E(X \wedge n) + E(Y \wedge n) \leq E[(X + Y) \wedge 2n] \leq E(X \wedge 2n) + E(Y \wedge 2n). \tag{3.1}$$

From Definition 3 of the expectation, we have, for any nonnegative random variable W, $E(W \wedge n) \uparrow E(W)$ as $n \uparrow \infty$. Sending $n \to \infty$ in (3.1) yields the linearity for nonnegative random variables.

Finally, we consider the case $E|X|, E|Y| < \infty$. Write

$$X = X^+ - X^-, \quad Y = Y^+ - Y^-, \quad X+Y = (X^+ + Y^+) - (X^- + Y^-) = (X+Y)^+ - (X+Y)^-.$$

From the latter equality, we have

$$E[(X+Y)^+ + (X^- + Y^-)] = E[(X+Y)^- + (X^+ + Y^+)]; \tag{3.2}$$

hence from the linearity of E for the previous case of nonnegative random variables, we have

$$E[(X+Y)^+] + E[(X^- + Y^-)] = E[(X+Y)^-] + E[(X^+ + Y^+)].$$

Therefore,

$$\begin{aligned}
E(X+Y) =& E(X+Y)^+ - E(X+Y)^- && \text{(Definition 4 of the expectation)} \\
=& E(X^+ + Y^+) - E(X^- + Y^-) && \text{(From (3.2))} \\
=& E(X^+) + E(Y^+) - E(X^-) - E(Y^-) \\
& && \text{(From the linearity of } E \text{ for nonnegative random variables)} \\
=& E(X) + E(Y). && \text{(Definition 4 of the expectation)}
\end{aligned}$$

□

Monotone Convergence Theorem (MCT). Let $\{X_n \geq 0, n = 1, 2, \dots\}$ be a sequence of nonnegative random variables. If $X_n \uparrow X$, then $E(X_n) \uparrow E(X)$.

Proof. By monotonicity, $\{E(X_n)\}_{n=1}^{\infty}$ is a sequence of nonnegative nonincreasing numbers. It must converge to a value a (possibly ∞). We need to show that $E(X) = a$. We consider two cases.

Case 1: $a = \infty$. Because $E(X) \geq E(X_n)$, for any n, if $a = \infty$, then $E(X)$ must also be ∞; hence in this case, $E(X) = a$.

Case 2: $a < \infty$. By the argument in Case 1, we have $E(X) \geq a$. We are left to show that $E(X) \leq a$. Recall Definition 3:

$$E(X) := \sup_{\substack{Y: 0 \leq Y \leq X \\ Y \text{ is a simple random variable}}} E(Y).$$

It suffices to show that $E(Y) \leq a$, or $E(Y) \leq a + \varepsilon$ for all $\varepsilon > 0$ and all Y in the supremum above. Fix $\varepsilon > 0$ and such a Y. Suppose

$$Y = \sum_{j=1}^{m} b_j 1_{B_j},$$

where $\{B_1, \dots, B_m\}$ are disjoint. Define

$$Y_\varepsilon = \sum_{j=1}^{m} (b_j - \frac{\varepsilon}{2}) 1_{B_j}.$$

Note that

$$\begin{aligned}
E(X_n) =& E[X_n 1(X_n \geq Y_\varepsilon)] + E[X_n 1(X_n < Y_\varepsilon)] \\
\geq& E[Y_\varepsilon 1(X_n \geq Y_\varepsilon)] + E[X_n 1(X_n < Y_\varepsilon)] \\
\geq& E(Y_\varepsilon) - E[Y_\varepsilon 1(X_n < Y_\varepsilon)] \\
\geq& E(Y_\varepsilon) - E[M 1(X_n < Y_\varepsilon)] && \text{(For a sufficiently large constant } M) \\
=& E(Y_\varepsilon) - M P(X_n < Y_\varepsilon) \\
\geq& E(Y_\varepsilon) - \frac{\varepsilon}{2}, && \text{(For sufficiently large } n)
\end{aligned}$$

where in the last inequality, we used $\{X_n < Y_\varepsilon\} \to \emptyset$ and convergence from above property of measures. Therefore,

$$E(Y) \overset{\text{Definition of } Y_\varepsilon}{\leq} E(Y_\varepsilon) + \frac{\varepsilon}{2} \overset{\text{Above inequality}}{\leq} E(X_n) + \varepsilon \leq a + \varepsilon.$$

□

Theorem (Fatou's Lemma). If $X_n \geq 0$, $\forall n$, then

$$\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n].$$

Proof. We have

$$\begin{aligned}
\liminf_{n\to\infty} E[X_n] &\geq \liminf_{n\to\infty} E[\inf_{k\geq n} X_k] \\
&= \lim_{n\to\infty} E[\inf_{k\geq n} X_k] \\
&\overset{MCT}{=} E[\lim_{n\to\infty} \inf_{k\geq n} X_k] \\
&= E[\liminf_{n\to\infty} X_n].
\end{aligned}$$

□

Dominated Convergence Theorem (DCT). If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some Y with $E[Y] < \infty$. Then $E[X_n] \to E[X]$.

Proof. Note that $X_n + Y \geq 0$. By Fatou's lemma:

$$\liminf_{n\to\infty} E(X_n + Y) \geq E[\liminf_{n\to\infty} (X_n + Y)] = E[X + Y];$$

hence $\liminf_{n\to\infty} E(X_n) \geq E[X]$. Similarly,

$$\limsup_{n\to\infty} E(X_n - Y) = -\liminf_{n\to\infty} E(-X_n + Y) \leq -E[\liminf_{n\to\infty}(-X_n + Y)] = -E[-X + Y];$$

hence $\limsup_{n\to\infty} E(X_n) \leq E(X)$.

$\square$

3.3 Useful Inequalities

Jensen's Inequality. If X is a random variable, φ is a convex function, $E|X| < \infty$ and $E|\varphi(X)| < \infty$, then

$$E[\varphi(X)] \geq \varphi[E(X)].$$

For example, $E[|X|^p] \geq [E|X|]^p$, for $p \geq 1$.

Proof. Let $c = E(X)$. By convexity, there exist a, b such that

$$\varphi(c) = ac + b, \quad \varphi(x) \geq ax + b.$$

Therefore,

$$E[\varphi(x)] \geq aE(X) + b = \varphi(c) = \varphi(E(X)).$$

$\square$

Hölder's Inequality. If $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$E[|XY|] \leq \|X\|_p \|Y\|_q,$$

where $\|X\|_p := (E|X|^p)^{1/p}$ and $\|X\|_\infty := \inf\{a : P(|X| > a) = 0\}$.

The case $p = q = 2$ is called the Cauchy-Schwarz inequality.

Proof. By appropriate scaling, we only need to consider the case $\|X\|_p = \|Y\|_q = 1$. From

$$xy \leq \frac{x^p}{p} + \frac{x^q}{q}, \ \forall\ x, y \geq 0,$$

we have

$$E|XY| \leq \frac{1}{p} + \frac{1}{q} = 1.$$

$\square$

Minkowski's Inequality. For $p \geq 1$, we have

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Proof. Let q be such that $\frac{1}{p}+\frac{1}{q}=1$. We have

$$\begin{aligned}(E|X+Y|^p)^{\frac{1}{p}} &= (E|X||X+Y|^{p-1}+E|Y||X+Y|^{p-1})^{\frac{1}{p}}\\ &\overset{\text{Hölder}}{\leq} \left[(E|X|^p)^{\frac{1}{p}}(E|X+Y|^{(p-1)q})^{\frac{1}{q}}+(E|Y|^p)^{\frac{1}{p}}(E|X+Y|^{(p-1)q})^{\frac{1}{q}}\right]^{\frac{1}{p}}\\ &= (\|X\|_p+\|Y\|_p)^{\frac{1}{p}}(E|X+Y|^p)^{\frac{1}{pq}}.\end{aligned}$$

Solving the recursive inequality proves the result.

□

Markov's Inequality. If X is a nonnegative random variable and $a>0$, then

$$P(X\geq a)\leq \frac{E(X)}{a}.$$

Proof.

$$P(X\geq a)=E[1(X\geq a)]\leq E[\frac{X}{a}1(X\geq a)]\leq \frac{E|X|}{a}.$$

□

Chebyshev's Inequality.

$$P(|X-E(X)|\geq a)\leq \frac{Var(X)}{a^2}.$$

Proof. Apply Markov's inequality to $[X-E(X)]^2$. □