Indian Institute of Technology Kanpur
CS773 Online Learning and Optimization

Scribe: Kalpant Pathak
Instructor: Purushottam Kar
Date: January 7, 2016

LECTURE
3

Preliminaries: Convex Analysis and Probability Theory

1 Introduction

In this lecture, we will continue our discussion on vector spaces and convex analysis and look at results such as the *Riesz representation theorem* and convex projections. In addition, we will discuss some fundamentals of *Probability Theory* including σ-fields, probability measures, notion of independence, expectations etc. We will present a formal definition of random variables and distribution functions associated with them. Also, we will discuss various inequalities associated with random variables.

2 Riesz Representation Theorem

For any inner product space $(V, \langle \cdot, \cdot \rangle)$[1], consider its dual space $\mathcal{F}_{\text{lin}}$. Recall that $\mathcal{F}_{\text{lin}}$ consists of all linear functionals over V. Now, let us define another set, that of linear functions $\mathcal{G}_{\text{lin}}$ that can be represented as follows:

$$\mathcal{G}_{\text{lin}} = \{ f_{\mathbf{v}} : \mathbf{u} \mapsto \langle \mathbf{u}, \mathbf{v} \rangle \, , \, \mathbf{v} \in V \}.$$

Now, from the definition it is clear that $\mathcal{G}_{\text{lin}} \subseteq \mathcal{F}_{\text{lin}}$. Riesz representation theorem tells us for *Hilbert spaces*, the reverse is true as well. Hilbert spaces are inner product spaces that are *complete* with respect to the norm induced by the inner product[2].

Theorem 3.1. (*Riesz Representation Theorem*) Let $\mathcal{F}_{\text{lin}}$ and $\mathcal{G}_{\text{lin}}$ be be defined as above for a Hilbert space $\mathcal{H}$. Then we have $\mathcal{G}_{\text{lin}} = \mathcal{F}_{\text{lin}}$.

This result tells us that the method used to define $\mathcal{G}_{\text{lin}}$ gives us the only way linear functionals can be defined. Proving this result in all its generality is beyond the scope of this course. However, to give a flavor of the result, we prove the result below for Hilbert spaces that have a countable orthonormal basis. For simplicity, we give the proof for finite basis, but the same may be extended to the case of a countably infinite basis as well.

Note, however, that the Riesz representation theorem is a much more elegant result that does not require a basis to be established over the Hilbert space.

Proof. To prove the equality, we need to prove $\mathcal{G}_{\text{lin}} \supseteq \mathcal{F}_{\text{lin}}$. Let $V = \mathbb{R}^n$ be the n-dimensional Euclidean space with a finite orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ i.e. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \mathbb{I}\{i = j\}$. Then

[1] An inner product space is a vector space over which an inner product has been established

[2] An inner product space is said to be complete if every Cauchy sequence in that space converges, with respect to the norm induced by the inner product, to an element in the space itself.

every vector $\mathbf{u} \in V$ can be written as a unique linear combination of basis vectors of V as

$$\mathbf{u} = u_1 \cdot \mathbf{e}_1 + u_2 \cdot \mathbf{e}_2 + \ldots + u_n \cdot \mathbf{e}_n,$$

where $u_i = \langle \mathbf{u}, \mathbf{e}_i \rangle$. Now, given a linear functional $f \in \mathcal{F}_{\text{lin}}$, define the vector

$$\mathbf{v}_f = f(\mathbf{e}_1) \cdot \mathbf{e}_1 + f(\mathbf{e}_2) \cdot \mathbf{e}_2 + \ldots + f(\mathbf{e}_n) \cdot \mathbf{e}_n$$

Consider any vector $\mathbf{u} \in V$. We have

$$\begin{aligned} f(\mathbf{u}) &= f(u_1 \cdot \mathbf{e}_1 + u_2 \cdot \mathbf{e}_2 + \ldots + u_n \cdot \mathbf{e}_n) \\ &= f(u_1 \cdot \mathbf{e}_1) + f(u_2 \cdot \mathbf{e}_2) + \ldots + f(u_n \cdot \mathbf{e}_n) \\ &= u_1 \cdot f(\mathbf{e}_1) + u_2 \cdot f(\mathbf{e}_2) + \ldots + u_n \cdot f(\mathbf{e}_n) \\ &= \langle \mathbf{u}, \mathbf{e}_1 \rangle \cdot f(\mathbf{e}_1) + \langle \mathbf{u}, \mathbf{e}_2 \rangle \cdot f(\mathbf{e}_2) + \ldots + \langle \mathbf{u}, \mathbf{e}_n \rangle \cdot f(\mathbf{e}_n) \\ &= \langle \mathbf{u}, \mathbf{v}_f \rangle , \end{aligned}$$

which establishes the result. □

Definition 3.1 (Lipschitz Function)**.** A function $f : V \to \mathbb{R}$ on a normed space V, is said to be L-Lipschitz if

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq L \|\mathbf{u} - \mathbf{v}\| , \ \forall \mathbf{u}, \mathbf{v} \in V$$

where L is a constant. In other words, a Lipschitz function can not change values too abruptly. For differentiable functions, this happens if the gradient is norm bounded.

Lemma 3.2. Let $f : V \to \mathbb{R}$ be a differentiable function such that $\|\nabla f(\mathbf{x})\| \leq L$. Then f is L-Lipschitz.

Proof. The mean value theorem states that if f is differentiable, then for any $\mathbf{u}, \mathbf{v}$, there exists a $\lambda \in [0, 1]$ such that

$$f(\mathbf{u}) - f(\mathbf{v}) = \langle \nabla f((\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v})), \mathbf{u} - \mathbf{v} \rangle .$$

Applying the Cauchy-Schwartz inequality and the fact that $\|\nabla f((\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v}))\| \leq L$ proves the result. □

3 Convex Projections

Let us consider a constrained convex optimization problem

$$\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$$

where $f(\cdot)$ is a convex function and $\mathcal{C}$ is the convex set which arises due to various convex and/or affine constraints. A popular way solving these constrained optimization problems is the *projected gradient descent method* which involves taking steps opposite to the direction of the gradient and projecting back onto the constraint.

The convex projection step is crucial to this procedure. We analyze some properties of convex projections now.

Definition 3.2 (Convex Projection)**.** Let $\mathcal{C} \subset V$ be a convex set and $\|\cdot\|$ be a (non-degenerate) norm on the vector space V. Then the convex projection of a vector $\mathbf{z} \in V$ onto the set $\mathcal{C}$ is defined as:

$$\Pi_{\mathcal{C}}(\mathbf{z}) := \arg\min_{\mathbf{x}\in\mathcal{C}} \|\mathbf{x} - \mathbf{z}\| \tag{1}$$

The problem of fining a convex projection is a convex optimization problem itself but in most of the cases, (1) has closed form solution.

Exercise 3.1. Prove that all norms are convex functions.

Example 3.1. Let $\mathcal{C} = \mathcal{B}_2(\mathbf{0}, r)$ be the ball corresponding to the ℓ_2 norm, then the projection of a vector $\mathbf{z} \in V$ onto $\mathcal{C}$ is given as:

$$\Pi_{\mathcal{C}}(\mathbf{z}) = \begin{cases} \mathbf{z}, & \text{if } \mathbf{z} \in \mathcal{C} \\ \frac{\mathbf{z}}{\|\mathbf{z}\|} \cdot r, & \text{if } \mathbf{z} \notin \mathcal{C}. \end{cases}$$

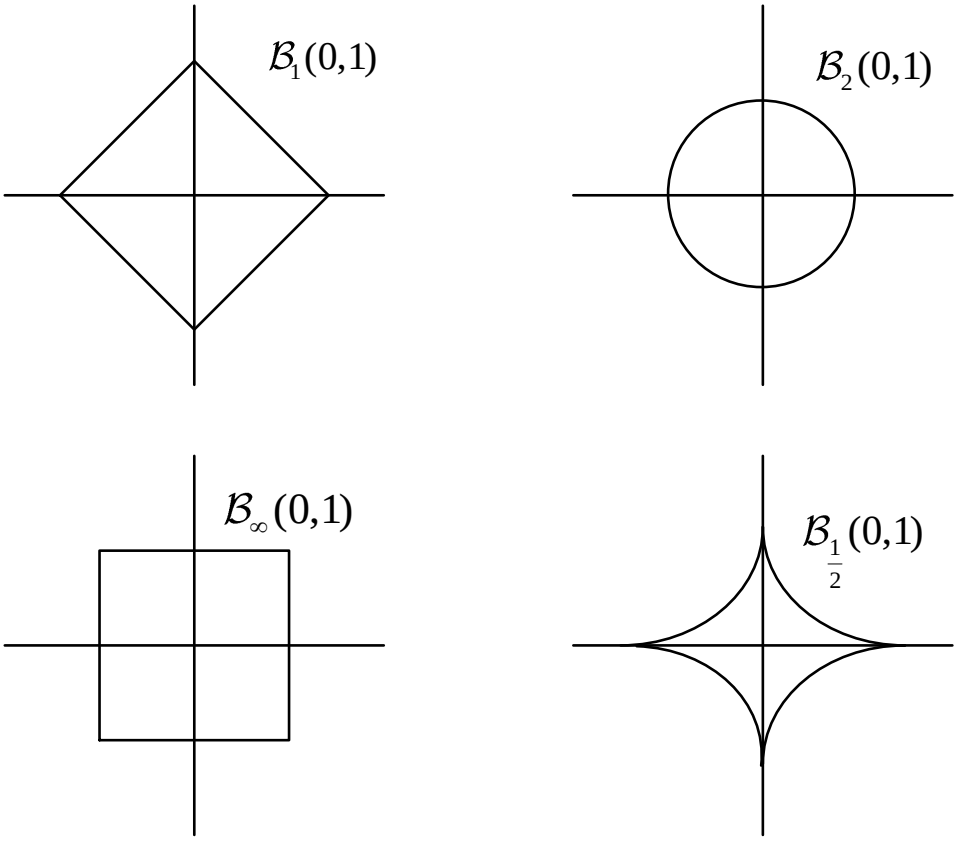

Figure 1: Graphical representation of two-dimensional norm balls associated with different norms. Note that $\ell_{1/2}$ (and in general ℓ_p for any $p < 1$) is not a norm although it is often referred to as one.

3.1 Properties of Convex Projections

In the following, we will look at projections with respect to the ℓ_2 norm.

3.1.1 Property I

If $\hat{\mathbf{z}}$ is the projection of $\mathbf{z} \in V$ onto a convex set $\mathcal{C} \subset V$, then

$$\langle \mathbf{x} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle \leq 0, \ \forall \mathbf{x} \in \mathcal{C} \tag{2}$$

in other words, the angle between the vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ is always greater than 90°. Intuitively, the case when angle between vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ becomes less than 90° is only possible when $\mathcal{C}$ is a non-convex set.

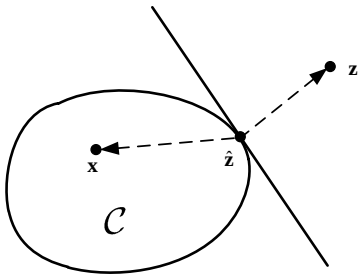

Figure 2: Graphical representation of projection property-I. The angle between the vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ is less greater 90°.

Proof. (Bertsekas, 2010) The supporting hyperplane for the convex set $\mathcal{C}$ passing through vector $\hat{\mathbf{z}}$ is given as:

$$\langle \mathbf{w} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle = 0$$

We note that proving that the convex set is contained in the halfspace $\mathcal{C} \subseteq \mathcal{H} = \{\mathbf{w} | \langle \mathbf{w} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle \leq 0\}$ will establish the claimed result. Now, the projection $\hat{\mathbf{z}}$ is obtained by minimizing the following convex function:

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 .$$

However, a point $\hat{\mathbf{z}}$ minimizes a function f over a convex set $\mathcal{C}$ if and only if

$$\langle \nabla f(\hat{\mathbf{z}}), \mathbf{x} - \hat{\mathbf{z}} \rangle \geq 0, \quad \forall \mathbf{x} \in \mathcal{C}$$

Since $\nabla f(\hat{\mathbf{z}}) = \hat{\mathbf{z}} - \mathbf{z}$, this condition becomes equivalent to (2). This completes the proof. □

3.1.2 Property II

If $\hat{\mathbf{z}}$ be the projection of $\mathbf{z} \in V$ onto a convex set $\mathcal{C} \subset V$, then

$$\|\mathbf{x} - \mathbf{z}\|_2 \geq \|\mathbf{x} - \hat{\mathbf{z}}\|_2 , \ \forall \mathbf{x} \in \mathcal{C} \tag{3}$$

i.e., $\hat{\mathbf{z}}$ is closer to all the points in $\mathcal{C}$ than $\mathbf{z}$.

Proof. We have, for any $\mathbf{x} \in \mathcal{C}$,

$$\begin{aligned}
\|\mathbf{x} - \mathbf{z}\|_2^2 &= \|\mathbf{x} - \hat{\mathbf{z}} + \hat{\mathbf{z}} - \mathbf{z}\|_2^2 \\
&= \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 + \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 - \underbrace{2 \langle \mathbf{x} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle}_{\leq 0 \text{ (Property I)}} \\
&\geq \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 + \underbrace{\|\mathbf{z} - \hat{\mathbf{z}}\|^2}_{\geq 0} \\
&\geq \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 .
\end{aligned}$$

□

4 Probability Theory

Definition 3.3 (σ-Field (Ash and Doléans-Dade, 2005))**.** Let Ω represent the possible set of outcomes of a random experiment and let $\mathcal{F}$ be a collection of subsets of Ω. Then $\mathcal{F}$ is called a *σ-field* (or a *σ-algebra*) if the following conditions hold,

1. (Surety of an event) $\Omega \in \mathcal{F}$.
2. (Closure under complementation) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. (Closure under countable unions) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

$\mathcal{F}$ is often referred to as the *event space* is is supposed to denote the set of events we are interested in analyzing. A few examples follow:

Example 3.2 (Roll of a dice)**.** $\Omega = \{1, 2, 3, 4, 5, 6\}$. A possible σ-field or event space over this set of outcomes can be the power set of Ω (collection of all subsets of Ω) $\mathcal{F} = 2^{\Omega}$. Another possible σ-field or event space is $\mathcal{F} = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$.

Note that the second event space does not contain several events contained in the power set event space – for instance the second event space does not allow us to talk about the event where the roll of dice results in a prime number. However, both are σ-fields in their own right, the second being a *sub-field* of the first.

Definition 3.4 (Probability Measure)**.** A probability measure on a σ-field $\mathcal{F}$ is a nonnegative, real-valued set function $\mathbb{P} : \mathcal{F} \to [0, 1]$ which satisfies following axioms:

1. $\mathbb{P}[A] \geq 0, \ \forall A \in \mathcal{F}$.
2. $\mathbb{P}[\Omega] = 1$.
3. If $\{A_n\}_{n \geq 1}$ is a countable collection of pairwise disjoint sets in $\mathcal{F}$, i.e. $A_i \cup A_j = \phi$ unless $i = j$, then

$$\mathbb{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \mathbb{P}[A_j].$$

4.1 Random Variables

Definition 3.5 (Probability Space)**.** A triplet $(\Omega, \mathcal{F}, \mathbb{P})$ of a set of outcomes, a valid σ-field thereupon, and a valid probability measure thereupon, constitutes a probability space.

Definition 3.6 (Random Variable)**.** A real-valued random variable X defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ satisfying the following properties

1. For every $r \in \mathbb{R}$, the set $A_r := \{\omega \in \Omega : X(\omega) \leq r\}$ satisfies $A_r \in \mathcal{F}$. i.e., the set $\{X \leq r\}$ is a *measureable event* for every $r \in \mathbb{R}$.
2. The probabilities of events $\{X = \infty\}$ and $\{X = -\infty\}$ is zero:

$$\mathbb{P}[X = \infty] = 0, \qquad \mathbb{P}[X = -\infty] = 0.$$

Some examples of commonly used random variables is given below.

Example 3.3 (Indicator Random Variable)**.** Consider an event $E \in \mathcal{F}$. The indicator random variable for event E is defined as:

$$\mathbb{I}[E] = \begin{cases} 1, & \text{if } E \text{ takes place, i.e. the outcome } \omega \in E. \\ 0, & \text{if } E \text{ does not take place, i.e. } \omega \notin E. \end{cases}$$

Indicator variables are immensely useful in proving results in an elegant manner, as we shall see soon.

Example 3.4 (Bernoulli Random Variable)**.** Given a binary set of outcomes, e.g. a toss of a coin $\Omega = \{H, T\}$, one defines a Bernoulli variable assigning numeral values to these events.

$$B_p = \begin{cases} 1, & \text{the first outcome takes place } \omega = H. \\ 0, & \text{the second outcome takes place } \omega = T. \end{cases}$$

If the coin lands heads with probability p, also known as the *bias* of the coin, then we have the following property for the Bernoulli variable.

$$B_p = \begin{cases} 1, & \text{with probability } p. \\ 0, & \text{with probability } 1 - p. \end{cases}$$

Example 3.5 (Rademacher Random Variable)**.** A Rademacher random variable R is defined to have the following property:

$$R = \begin{cases} 1, & \text{with probability } \frac{1}{2}. \\ -1, & \text{with probability } \frac{1}{2}. \end{cases}$$

A Rademacher random variable can, for instance, be realized by a fair coin.

4.2 Cumulative Distribution Function

The cumulative distribution function $F_X(x)$ of a random variable X describes the probability that X takes on a value less than or equal to a number real number $x \in \mathbb{R}$, that is,

$$F_X(x) = \mathbb{P}\left[X \leq x\right]. \tag{4}$$

When the function $F_X(x)$ is differentiable for every x, it's derivative gives us the *probability density function* (often abbreviated as PDF) $f_X(x)$ of the random variable X.

$$f_X(x) = F'_X(x). \tag{5}$$

When a random variable possesses a PDF, we have

$$\mathbb{P}\left[a \leq X \leq b\right] = \int_a^b f_X(t)dt$$

4.2.1 Properties of Distribution Function

The distribution function $F_X(x)$ has following properties (Papoulis, 1991):

1. $\lim_{x\to+\infty} F_X(x) = 1 \qquad \lim_{x\to-\infty} F_X(x) = 0.$
2. It is a nondecreasing function of x:

$$\text{if} \quad x_1 < x_2 \quad \text{then} \quad F_X(x_1) \leq F_X(x_2)$$

3. If $F_X(x_0) = 0$ then $F_X(x) = 0$ for every $x \leq x_0$.
4. $\mathbb{P}(X > x) = 1 - F_X(x)$.
5. The function $F_X(x)$ is right continuous:

$$F_X(x^+) = F_X(x),$$

where $F_X(x^+) = \lim_{\epsilon \to 0^+} F_X(x + \epsilon)$.

6. $\mathbb{P}[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1)$.
7. $\mathbb{P}[X = x] = F_X(x) - F_X(x^-)$,
where $F_X(x^-) = \lim_{\epsilon \to 0^+} F_X(x - \epsilon)$.
8. $\mathbb{P}[x_1 \leq X \leq x_2] = F_X(x_2) - F_X(x_1^-)$.

4.3 Expectation

If Ω is discrete, then a random variable $X : \Omega \to \mathbb{R}$ can only take on values in a discrete set, that is, $X \in \{a_1, a_2, \ldots, \}$, $a_i \in \mathbb{R}$. Then the expectation of the random variable X is given as:

$$\mathbb{E}X = \sum_i a_i \cdot \mathbb{P}[X = a_i] \tag{6}$$

If Ω is continuous, the expectation of the random variable X is given as:

$$\mathbb{E}X = \int_{-\infty}^{\infty} \mathbb{P}[X \geq t]\, dt \tag{7}$$

Definition 3.7 (Variance)**.** For a random variable X with a finite expectation, we define its variance as its second centered moment

$$\mathrm{Var}[X] = \mathbb{E}(X - \mathbb{E}X)^2.$$

4.3.1 Properties of Expectation

1. If $X \equiv c$ be a constant random variable, then $\mathbb{E}[X] = c$.
2. (*Linearity*) If X and Y be two random variables, then $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$.
3. If X and Y be two random variables such that $X \leq Y$ a.s., then $\mathbb{E}X \leq \mathbb{E}Y$.
4. (*Law of iterated expectation*) If X and Y be two random variables, then $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X$.

Note that the above results *do not* require any independence or correlation assumptions on the random variables and hold universally.

4.3.2 Functions of random variables and Conditional Expectation

- (LOTUS – Law of the unconscious statistician) If $g : \mathbb{R} \to \mathbb{R}$ be a function and X be a discrete random variable over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then:

$$\mathbb{E}g(X) = \sum_i g(a_i) \cdot \mathbb{P}[X = a_i]. \tag{8}$$

- If X and Y be two discrete random variables such that $X \in \{a_1, \ldots, a_n, \ldots\}$ and $Y = \{b_1, \ldots, b_m, \ldots\}$, then the conditional expectation of random variable X given the event $\{Y = b\}$ is given as:

$$\mathbb{E}[X|Y = b] = \sum_i a_i \cdot \mathbb{P}[X = a_i|Y = b]. \tag{9}$$

4.3.3 Independence

Two events $E_1, E_2 \in \mathcal{F}$ are independent if

$$\mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_1] \cdot \mathbb{P}[E_2].$$

Similarly, two random variables X and Y are called independent if

$$\mathbb{P}[X \leq r \wedge Y \leq t] = \mathbb{P}[X \leq r] \cdot \mathbb{P}[Y \leq t], \quad \forall r, t \in \mathbb{R}.$$

Conditional Independence Consider three events $E_1, E_2, E_3 \in \mathcal{F}$. Given event E_3 has already occurred, the events E_1 and E_2 are independent if

$$\mathbb{P}[E_1 \cap E_2|E_3] = \mathbb{P}[E_1|E_3] \cdot \mathbb{P}[E_2|E_3].$$

Remark 3.1 ((Stoyanov, 1988)). The concept of independence can be extended to any finite number of events or classes. We say that the events $E_1, \ldots, E_n \in \mathcal{F}$ are **mutually independent** if for $k = 2, \ldots, n$ and $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ we have

$$\mathbb{P}[E_{i_1} E_{i_2} \ldots E_{i_k}] = \mathbb{P}[E_{i_1}] \cdot \mathbb{P}[E_{i_2}] \cdot \ldots \cdot \mathbb{P}[E_{i_k}].$$

If this relation is fulfilled in the particular case $k = 2$ we say that the given events $E_1, \ldots, E_n$ are **pairwise independent**.

Example 3.6. Suppose India is playing Australia in a series of two one-dayers. Assume that the teams are evenly matched, i.e. for an given game there is equal chance of any team winning the game. Then consider three events

$$\begin{aligned} E_1 &\Rightarrow \text{ India wins first match,} \\ E_2 &\Rightarrow \text{ India wins second match,} \\ E_3 &\Rightarrow \text{ The matches were not fixed.} \end{aligned}$$

Then, given event E_3, the events E_1 and E_2 are independent. However, if E_3 is not given, i.e. if there is a possibility of the matches being fixed, say in favor of a particular country, then the outcome of the first match reveals a significant amount of information about the expected outcome of the second match and the events are no longer independent.

Example 3.7 (Pairwise independence does not imply mutual independence (Stoyanov, 1988)). Suppose a box contains four tickets labelled 112, 121, 211, 222. Choose one ticket at random and consider the events $A_1 = \{1 \text{ occurs in the first place}\}$, $A_2 = \{1 \text{ occurs in the second place}\}$ and $A_3 = \{1 \text{ occurs in the third place}\}$. We have $\mathbb{P}[A_1] = \mathbb{P}[A_2] = \mathbb{P}[A_3] = \frac{1}{2}$ and

$$\mathbb{P}[A_1 A_2] = \mathbb{P}[A_1 A_3] = \mathbb{P}[A_2 A_3] = \frac{1}{4}$$

This means that the three events are pairwise independent. However,

$$\mathbb{P}[A_1 A_2 A_3] = 0 \neq \frac{1}{8} = \mathbb{P}[A_1]\,\mathbb{P}[A_2]\,\mathbb{P}[A_3]$$

and hence these events are not mutually independent.

5 Inequalities Involving random variables

5.1 Jensen's Inequality (Cover and Thomas, 2006)

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function and let X be a real valued random variable, then

$$f(\mathbb{E}X) \leq \mathbb{E}f(X). \tag{10}$$

Proof. For simplicity, we will assume that f is a differentiable function. Let us choose a $x_0 \in \mathbb{R}$ – we will fix x_0 later. Then, by convexity of f, we have, a.e.

$$f(X) \geq f(x_0) + \langle \nabla f(x_0), X - x_0 \rangle$$

Taking expectation of both sides:

$$\mathbb{E}f(X) \geq f(x_0) + \langle \nabla f(x_0), \mathbb{E}X - x_0 \rangle$$

where, the above inequality follows from the linearity of the expectation operator and the fact that $\mathbb{E}X \geq \mathbb{E}Y$ if $X \geq Y$ a.s. By choosing $x_0 = \mathbb{E}X$ we have

$$f(\mathbb{E}X) \leq \mathbb{E}f(X),$$

which proves the result. □

5.2 Markov's Inequality

Theorem 3.3. Let X be a positive valued random variable i.e. $X \geq 0$ a.s. then for any $t > 0$

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t}. \tag{11}$$

Proof. Let $\mathbb{I}$ be the indicator function, then

$$\begin{aligned} X &= \underbrace{X \cdot \mathbb{I}\{X < t\}}_{\geq 0} + \underbrace{X \cdot \mathbb{I}\{X \geq t\}}_{\geq t \cdot \mathbb{I}\{X \geq t\}} \\ &\geq t \cdot \mathbb{I}\{X \geq t\} \end{aligned}$$

taking expectation of both sides

$$\mathbb{E}X \geq t \cdot \mathbb{P}[X \geq t].$$

Note that we have used a very useful property of the indicator function here. For any event $A \in \mathcal{F}$, if we construct the indicator random variable $\mathbb{I}\{A\}$, then we have $\mathbb{E}\mathbb{I}\{A\} = \mathbb{P}[A]$. □

We will often use the notation μ to denote the expectation of a random variable i.e. $\mu = \mathbb{E}X$. The notation σ is often used to denote the standard deviation i.e. $\sigma^2 = \mathbb{E}(X - \mu)^2$.

5.3 Chebychev Inequality

Theorem 3.4. Let the random variable X have expectation μ and variance σ^2. Then $\forall t > 0$

$$\mathbb{P}[|X - \mu| > t] \leq \frac{\sigma^2}{t^2}.$$

Proof. Let $Y = (X - \mu)^2$. Notice that Y is a positive valued random variable. Applying Markov's inequality on the random variable Y we get:

$$\begin{aligned} \mathbb{P}\left[(X - \mu)^2 \geq a\right] &\leq \frac{\mathbb{E}[(X - \mu)^2]}{a} \\ \Rightarrow \mathbb{P}[|X - \mu| > t\sigma] &\leq \frac{1}{t^2}. \end{aligned}$$

where $t = \frac{a}{\sigma}$. □

5.4 Hoeffding's Inequality (Wasserman)

Let X be a random variable such that $\mathbb{E}X = \mu$ and $a \leq X \leq b$ a.s.. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) realizations of the random variable X. Also, let $\bar{X}_n = \frac{1}{n}\sum_i X_i$ denote the empirical expectation of the random variables. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

5.5 Chernoff's Inequality (Wikipedia, b)

Let X be a Bernoulli random variable with bias μ i.e. $\mathbb{E}X = \mu$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. realizations of X. As before, let $\bar{X}_n = \frac{1}{n}\sum_i X_i$ denote the empirical expectation. Then, for every $\epsilon > 0$,

$$\mathbb{P}\left[\bar{X}_n \geq \mu + \epsilon\right] \leq \left(\left(\frac{\mu}{\mu+\epsilon}\right)^{\mu+\epsilon}\left(\frac{1-\mu}{1-\mu-\epsilon}\right)^{1-\mu-\epsilon}\right)^n$$
$$\mathbb{P}\left[\bar{X}_n \leq \mu - \epsilon\right] \leq \left(\left(\frac{\mu}{\mu-\epsilon}\right)^{\mu-\epsilon}\left(\frac{1-\mu}{1-\mu+\epsilon}\right)^{1-\mu+\epsilon}\right)^n$$

5.6 Bernstein Inequality (Wikipedia, a)

Let X_i be centered i.i.d. random variables i.e. $\mathbb{E}X_i = 0$ such that $|X_i| \leq M$ a.s. for some $M > 0$. Also, let $\sigma^2 = \frac{1}{n}\sum_i \mathbb{E}X_i^2$. Then, for every $\epsilon > 0$,

$$\mathbb{P}\left[\bar{X}_n > \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2 + \frac{2}{3}M\epsilon}\right)$$

References

Robert B. Ash and C. A. Doléans-Dade. *Probability & Measure Theory.* Elsevier, 2005.

Dimitri P. Bertsekas. *Convex Optimization Theory.* Univ. Press, 2010.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2006.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1991.

Jordan M. Stoyanov. *Counterexamples in Probability.* John Wiley & Sons, 1988.

Larry Wasserman. Probability inequalities. `http://www.stat.cmu.edu/~larry/=stat705/Lecture2.pdf`.

Wikipedia. Bernstein inequalities. `https://en.wikipedia.org/wiki/Bernstein_inequalities_(probability_theory)`, a.

Wikipedia. Chernoff bound. `https://en.wikipedia.org/wiki/Chernoff_bound#Precise_statements_and_proofs`, b.