06E:200:001 Lecture Notes[1]

Zhengyuan Gao

Fall 2013

[1]This version: 12/05/13. This sloppy note is compensating for my ugly handwriting and my poor English. I should warn you that it must contain numerous meaningless arguments and many typos. Read at your own risk!

Part I

Real Analysis

Chapter 1

Topologies

1.1 Topological Spaces

Definition. A "set" refers to a collection of objects. From now on, we will use calligraphical letter to denote a set, for example $\mathcal{X}$. The collection is written as $\mathcal{X} = \{x_1, \ldots, x_n\}$.

When we say "objects", it means that they should be logically distinguishable. That is, if x and y are two objects, $x = y$ and $x \neq y$ cannot hold simultaneously. The definition of an object makes sense is known as showing that the object is well defined.

Remark 1. A set may have some properties. So people often use these properties to define a set. A set associates with a property, say x which makes a function $f(x)$ less than zero, we should define the set in the following way:

$$\mathcal{X} = \{x : f(x) \leq 0\}.$$

Then $\mathcal{X}$ consists of all xs having the property $f(x) \leq 0$.

Definition. We denote $|\mathcal{X}|$ the total number of elements that $\mathcal{X}$ contains. This number is call cardinality. The cardinality of a singleton is one. The class of all subsets of a given set $\mathcal{X}$ as

$$2^{\mathcal{X}} := \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{X}\},$$

which is called the power set of $\mathcal{X}$.

Definition. The Cartersian product or product is a product of two non-empty sets $\mathcal{A}$ and $\mathcal{B}$, denoted as $\mathcal{A} \times \mathcal{B}$. The Cartesian product $\mathcal{A} \times \mathcal{B}$ is a set of all ordered pairs (a, b) where a comes from $\mathcal{A}$ and b comes from $\mathcal{B}$

$$\mathcal{A} \times \mathcal{B} := \{(a, b) : a \in \mathcal{A} \quad b \in \mathcal{B}\}.$$

One can easily extend the definition to n-fold product, $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ or $\oplus_{i=1}^{n} \mathcal{A}_i$ in short. The Cartesian product is not associative [1], so one can think the n-fold product as n-vector.

[1] Any distinct objects a and b, we have $\{a\} \times \{b\} = \{(a, b)\} \neq \{(b, a)\} = \{b\} \times \{a\}$.

Definition. An order pair is an ordered list (a, b) where we distinguish between the first and second elements. A relation is any set of ordered pairs. A function is a special kind of relation.

Example 1. The order list is ordered in the sense that: For any two ordered pairs (a, b) and (a', b'), we have $(a, b) = (a', b')$ iff $a = a'$ and $b = b'$. If one defines an *order list* using only the concept of sets

$$(a, b) = \{\{a\}, \{a, b\}\},$$

then $\{\{a\}, \{a, b\}\} = \{\{a'\}, \{a', b'\}\}$ iff $\{a\} = \{a'\}$ and $\{a, b\} = \{a', b'\}$. Without distinguishing between the first and second elements, $\{\{a\}, \{a, b\}\} = \{\{a, b\}, \{a\}\}$ so $\{a\} = \{a, b\}$ which contradicts the fact that $\{a\} \subset \{a, b\}$.

Definition. A relation is any set of *ordered pairs*. Let $\mathcal{S}$ be a subset of $\mathcal{X} \times \mathcal{X}$, then $\mathcal{S}$ is a relation on $\mathcal{X}$. $\mathcal{S}$ is an *equivalence relation* ($\sim$, indifferent relation) [2] if it is

(1) reflexive: $(x, x) \in \mathcal{S}$ for all $x \in \mathcal{X}$.
(2) symmetric: if $(x, y) \in \mathcal{S}$ for any $x, y \in \mathcal{X}$, then $(y, x) \in \mathcal{S}$.
(3) transitive: if $(x, y) \in \mathcal{S}$ and $(y, z) \in \mathcal{S}$ for any $x, y, z \in \mathcal{X}$, then $(x, z) \in \mathcal{S}$.

Beside these three properties, if for all $x, y \in \mathcal{X}$, $(x, y) \in \mathcal{S}$ and $(y, x) \in \mathcal{S}$ implies that $x = y$, then $\mathcal{S}$ is called antisymmetric.

Definition. A function (or a mapping): f from $\mathcal{X}$ to $\mathcal{Y}$, denoted by

$$f : \mathcal{X} \mapsto \mathcal{Y}$$

is a relation from $\mathcal{X}$ to $\mathcal{Y}$ i.e. $f \subset \mathcal{X} \times \mathcal{Y}$, in which every element of $\mathcal{X}$ appears exactly once as the first component of an ordered pair in the relation.

Remark 2. We often write $y = f(x)$ when (x, y) is an ordered pair in the function. In this case, y is called the image (or the value) of x under f and x the preimage of y. We call $\mathcal{X}$ the domain of f and call $\mathcal{Y}$ the codomain of f. The subset of $\mathcal{Y}$ consisting of those elements that appears as second components in the ordered pairs of f is called the range of f and is denoted by $f(\mathcal{X})$. A function is a rule that transforms one set into another, and refer to the set of all ordered pairs $(x, f(x))$ as the graph of the function $\mathrm{Gr}(f)$:

$$\mathrm{Gr}(f) := \{(x, f(x)) \in \mathcal{X} \times \mathcal{Y} : x \in \mathcal{X}\}.$$

Definition. A function f is called injective, if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. A function $f : \mathcal{X} \mapsto \mathcal{Y}$ is called onto, or surjective, if $f(\mathcal{X}) = \mathcal{Y}$. Thus, $f : \mathcal{X} \mapsto \mathcal{Y}$ is onto iff for any $y \in \mathcal{Y}$, there is $x \in \mathcal{X}$ such that $f(x) = y$. $f : \mathcal{X} \mapsto \mathcal{Y}$ is called bijective if f is both injective and surjective.

Definition. A set $\mathcal{T}$ is called an open set if it does not contain any of its boundary points.

[2] In economics, we usually write $x \sim y$ if $(x, y) \in \mathcal{S}$. So the following expressions are simplified to for $x, y, z \in \mathcal{X}$: (1)$x \sim x$ (2) if $x \sim y$ then $y \sim x$ (3) if $x \sim y$ and $y \sim z$ then $x \sim z$.

Definition. A topology on Ω is a collection $\mathcal{T}$ of subsets of Ω, and satisfies the following conditions:
(1) $\emptyset \in \mathcal{T}$ and $\Omega \in \mathcal{T}$,
(2) If $U_\alpha \in \mathcal{T}$ for all $\alpha \in A$, then $\cup_{\alpha \in A} U_\alpha \in \mathcal{T}$.
(3) If $U_j \in \mathcal{T}$ for all $1 \leq j \leq n$, then $\cap_{j=1}^{n} U_j \in \mathcal{T}$.

Definition. The pair $(\Omega, \mathcal{T})$ is call a topological space.

Definition. A topological space $(\Omega, \mathcal{T})$ is called compact if, whenever we have a collection U_α ($\alpha \in A$) of open sets with $\cup_{\alpha \in A} U_\alpha = \Omega$, we can find a finite subcollection $U_{\alpha_1}, U_{\alpha_2}, \ldots, U_{\alpha_n}$ with $\alpha_i \in A$ such that $U_{i=1}^{n} U_{\alpha_i} = \Omega$.

In other words, a set is compact if any cover by open sets has a finite subcover.

Remark. Memebers of Ω are called points. The members of $\mathcal{T}$ are open sets. Their complements, $\mathcal{F} = \Omega \backslash \mathcal{U}$, $\mathcal{U} \in \mathcal{T}$, are called closed set. There are sets that are niether closed nor open, i.e. $\emptyset$ and Ω are both open and closed since they are complements of each other.

The continuity really depends only on topology.

Definition. (Continuity-topological language) Given topological spaces $(\Omega, \mathcal{T})$ and $(\Upsilon, \mathcal{U})$, a function f from Ω into Υ is called continuous iff for all $u \in \mathcal{U}$, $f^{-1}(u) \in \mathcal{T}$.

We know that many complicated mathematical structures can be considered as a space which locally looks like a simpler space. Hausdorff defined topologies in terms of neighbourhoods, it appears to be technically easier to define topologies in terms of open sets as we have done in this course. However, topologists still use the notion of neighbourhoods. Loosely speaking, an open neighbourhood of x is an open set containing x.

Definition. Given topological spaces $(\Omega, \mathcal{T})$, if $x \in \Omega$, we say that $\mathcal{N}$ is a neighbourhood of x if we can find $U \in \mathcal{T}$ with $x \in U \subseteq \mathcal{N}$.

Definition. (Continuity-topological language 2) Given topological spaces $(\Omega, \mathcal{T})$ and $(\Upsilon, \mathcal{U})$, a function f from Ω into Υ is continuous iff given $x \in \Omega$ and $\mathcal{M}$ a neighbourhood of $f(x)$ in Υ, we can find a neighbourhood $\mathcal{N}$ of x with $f(\mathcal{N}) \subseteq \mathcal{M}$.

Is it possible to define convergence in terms of neighbourhoods? In set theory, sequences are inadequate tools for the study of topologies which have neighbourhood systems which are 'large in the set theoretic sense'. It turns out that the deeper study of set theory reveals not only the true nature of the problem but also solutions via nets (a kind of generalised sequence) or filters.

1.2 Metric Spaces

In this section, I will use bold letter $\mathbf{v}$ for vectors and natural letter v for real numbers.

Definition. A (real) vector space consists of a set $\mathcal{V}$ with elements called vectors and two operations with the following properties:

Vector addition: for each pair $\mathbf{u}, \mathbf{v} \in \mathcal{V}$, there is a vector $\mathbf{u} + \mathbf{v} \in \mathcal{V}$. This is

(1) commutativity: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

(2) associativity: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{v} + \mathbf{u}) + \mathbf{w}$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$.

(3) zero: there is a vector $\mathbf{0} \in \mathcal{V}$ such that $\mathbf{0} + \mathbf{u} = \mathbf{u} = \mathbf{u} + \mathbf{0}$ for all $\mathbf{u} \in \mathcal{V}$.

(4) inverses: for each $\mathbf{u} \in \mathcal{V}$, there is a vector $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.

Scalar Multiplication: for each pair $\mathbf{v} \in \mathcal{V}$ and real number $r \in \mathbb{R}$, there is a vector $r\mathbf{v} \in \mathcal{V}$. This satisfies

$$(1)(r+s)\mathbf{u} = r\mathbf{u} + s\mathbf{u} \qquad (4)1\mathbf{v} = \mathbf{v}$$
$$(2)r(s\mathbf{v}) = (rs)\mathbf{v} \qquad (5)0\mathbf{v} = \mathbf{0}$$
$$(3)r(\mathbf{u} + \mathbf{v}) = r\mathbf{u} + r\mathbf{v} \qquad (6)(-1)\mathbf{v} = -\mathbf{v}$$

Let $\mathcal{V}$ be a vector space over the real numbers $\mathbb{R}$.

Definition. A norm on $\mathcal{V}$ is a nonnegative real-valued function $\|v\|$ defined for $v \in \mathcal{V}$ such that

(1) $\|v\| = 0$ if and only if $v = 0$,

(2) $\|t\, v\| = |t|\, \|v\|$ for every $v \in \mathcal{V}$ and $t \in \mathbb{R}$, as appropriate, and

(3) $\|v + w\| \leq \|v\| + \|w\|$ for every $v, w \in \mathcal{V}$. Here $|t|$ is the absolute value of t when $t \in \mathbb{R}$.

More generally, let n be a positive integer, and let V be the space $\mathbb{R}^n$ of n-tuples $\mathbf{v} = (v_1, \ldots, v_n)$ of real numbers. As usual, this is a vector space with respect to coordinatewise addition and scalar multiplication.

Definition. The following norms are the other most common norms:

$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$, which is known as l_∞-norm.

$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$, which is known as l_2-norm.

Or more generally, $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ which is known as l_p-norm.

The metrics on $\mathbb{R}^n$ associated to the norms $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_\infty$ determine the same topology as the standard Euclidean metric. The main point is that $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$ are equivalent norms on $\mathbb{R}^n$ for each positive integer n, in the sense that each is bounded by constant multiples of the others.

Definition. Let $\mathcal{V}$ be a set and $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and satifies:

(1) $d(\mathbf{v}, \mathbf{w}) \geq 0$ for all $\mathbf{v}, \mathbf{w} \in \mathcal{V}$.

(2) $d(\mathbf{v}, \mathbf{w}) = 0$ iff $\mathbf{v} = \mathbf{w}$.

(3) $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$ for every $\mathbf{v}, \mathbf{w} \in \mathcal{V}$.

(4) $d(\mathbf{v}, \mathbf{w}) + d(\mathbf{w}, \mathbf{z}) \geq d(\mathbf{v}, \mathbf{z})$ for every $\mathbf{v}, \mathbf{w}, \mathbf{z} \in \mathcal{V}$. (triangle inequality)

then $(\mathcal{V}, d)$ is a metric space and $d(\cdot, \cdot)$ is a metric.

Example 2. If $\|\mathbf{v}\|$ is a norm on a norm vector space $\mathcal{V}$, then $d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|$ defines a metric on $\mathcal{V}$.

For instance, the standard Euclidean metric on $\mathbb{R}^n$ is the same as the metric associated to the standard Euclidean norm in this way. The vector length $\|\mathbf{x}\| = (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}}$, is called l_2-norm and is denoted as $\|\cdot\|_2$

Example 3. (l_2 space) It is not difficult to see the length of $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_2 = (x_1^2 + \cdots + x_n^2)^{\frac{1}{2}}$, is a metric space.

(1) $\|\mathbf{x}\|_2 \geq 0$ and $\|x\|_2 = 0$ iff $\mathbf{x} = (0, \cdots, 0)$.

(2) For all $\lambda \in \mathbb{R}$, $\|\lambda\mathbf{x}\|_2 = |\lambda|\|\mathbf{x}\|_2$.

(3) For any $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$, triangle inequality tells that $\|\mathbf{x} + \mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$ (Cauchy-Schwartz inequality).

Definition. Let $\mathcal{H}$ be a vector space over $\mathbb{R}$ with a mapping $\langle\cdot,\cdot\rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfying

(1) $\langle \alpha u + \beta v, w\rangle = \alpha \langle u, w\rangle + \beta \langle v, w\rangle$.

(2) $\langle v, w\rangle = \langle w, v\rangle$

(3) $\langle u, u\rangle \geq 0$

(4) If $\langle u, u\rangle = 0$, then $u = 0$.

Because the mapping $\langle\cdot,\cdot\rangle$ is called inner product, the space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ or simply wirtten as $\mathcal{H}$ is called inner product space. $\mathcal{H}$ is a normed space with $\|h\|_2 = \sqrt{\langle h, h\rangle}$ for any $h \in \mathcal{H}$.

The distance function imposes further features for balls.[3] We now give the definition of balls as follows.

Definition. A ball at $\mathbf{a}$ in $\mathbb{R}^n$ of radius r is the set

$$B_r(\mathbf{a}) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\| < r\}.$$

If $(\mathcal{X}, d)$ is a general metric space and $r > 0$,

$$B_r(a) = \{x \in \mathcal{X} : \, d(x, a) < r\}.$$

A ball $B_r(a)$ consists of all v such that $d(a, v) < \delta$.

Example 4. Usually, the ball is used for describing an open set in metric space. For example, a subset $\mathcal{U}$ of $\mathbb{R}^n$ is open if for every $\mathbf{a} \in \mathcal{U}$, there is some $r = r(\mathbf{a}) > 0$ such that $B_r(\mathbf{a})$ is contained in $\mathcal{U}$.

Definition. A limit point x can be then thought as a point $x \in \mathcal{X}$ if every $B_\delta(x)$ with a positive δ contains a point $x' \in \mathcal{X}$ and $x \neq x'$.

Definition. We say that $\mathcal{X}$ is dense in a subset $\mathcal{E}$ if every point of $\mathcal{E}$ is a limit point of $\mathcal{X}$.

Now we can define the continuity in our "new" space.

Definition. (Contuinity-metric language) Let $(\mathcal{X}, d)$ and $(\mathcal{Y}, \rho)$ be metric spaces. A function $f : \mathcal{X} \to \mathcal{Y}$ is called continuous if, given $t \in \mathcal{X}$ and $\epsilon > 0$, we can find a $\delta(t, \epsilon) > 0$ such that

$$\rho(f(t), f(s)) < \epsilon \quad \text{whenever } d(t, s) < \delta(t, \epsilon).$$

[3](OK) calls these balls *δ-neighborhood*.

Similarly, we can define openness by using the metric. We say a subset $\mathcal{E}$ in $\mathcal{X}$ is open if whenever $e \in \mathcal{E}$, we can find a $\delta(e) > 0$ such that $x \in \mathcal{E}$ whenever $d(x, e) < \delta(e)$.

The second class of well behaved sets identified by Cantor were the closed sets. In order to define closed sets in metric spaces, we need a notion of limit. Fortunately, the classical definition generalizes without difficulty.

Definition. (Convergence) Consider a sequence $(x_n)_{n=1}^{\infty}$ in a metric space (X, d). If $x \in X$ and given $\epsilon > 0$, we can find an integer $N \geq 1$ (depending on ϵ) such that

$$d(x_n, x) < \epsilon \text{ for all } n \geq N,$$

then we say $x_n \to x$ as $n \to \infty$ and that x is the limit of the sequence $(x_n)_{n=1}^{\infty}$.

Definition. Let (X, d) be a metric space, A set $\mathcal{F}$ in X is said to be closed if, whenever $x_n \in \mathcal{F}$ and $x_n \to x$ as $n \to \infty$, it follows that $x \in \mathcal{F}$.

Example 5. In $\mathbb{R}^n$ with Euclidean metric, then one point set $\{\mathbf{x}\}$ is not open. However if (X, d) is a discrete metric space, then $\{x\} = B_{1/2}(x)$ and all subsets of X are both open and closed.

Definition. Let X be a metric space and $S \subseteq X$. The largest open set in X that is contained in S is called the interior of S ($\text{int}_X(S)$). The smallest closed set in X that contains S is called the closure of S ($\text{cl}_X(S)$).

Definition. Let $\mathcal{V}$ be a real vector space. A set $\mathcal{E} \subseteq \mathcal{V}$ is said to be convex if for every $x, y \in \mathcal{E}$ and real number t with $0 < t < 1$, we have that $t\,x + (1 - t)\,y \in \mathcal{E}$.

Example 6. If $\|\cdot\|$ is a norm on $\mathcal{V}$, then the closed unit ball $\bar{B}_1 := \{\mathbf{v} \in \mathcal{V} : \|\mathbf{v}\| \leq 1\}$ is a convex set in $\mathcal{V}$. Similarly, the open unit ball $B_1 := \{\mathbf{v} \in \mathcal{V} : \|\mathbf{v}\| < 1\}$ is also a convex set in $\mathcal{V}$.

Definition. Let $(x_n)_{n=1}^{\infty}$ be a sequence in the metric space (X, d). For every ϵ, there is an integer $N(\epsilon)$ such that

$$d(x_n, x_m) < \epsilon$$

for all $m, n \geq N(\epsilon)$. Then $(x_n)_{n=1}^{\infty}$ is called Cauchy sequence.

Definition. A metric space (X, d) is said to be complete if every Cauchy sequence in X converges to an element of X. Note that completeness is not a topological property

Remark 3. If $\mathcal{M}$ is complete and $\mathcal{N} \subseteq \mathcal{M}$, then $\mathcal{N}$ is complete with respect to the restriction of $d(x, y)$ to $x, y \in \mathcal{N}$ if and only if $\mathcal{N}$ is a closed set in $\mathcal{M}$.

Let $\mathcal{V}$ be a real vector space equipped with a norm and hence a metric. If $\mathcal{V}$ is complete with respect to this metric, then $\mathcal{V}$ is said to be a Banach Space. If the norm is also associated to an inner product, then $\mathcal{V}$ is said to be a Hilbert Space. In other words, a Hilber space is a complete inner product space.

1.3 Compactness and Completeness in $\mathbb{R}^n$

Theorem 1. *(OK p.164 Example 11) Every Cauchy sequence in $\mathbb{R}^n$ converges. Thus, $\mathbb{R}^n$ is complete.*

Proposition 1. *Let $(\mathbf{x}_n)$ be a seuqence in a metric space X,*

(1) If $(\mathbf{x}_n)$ is convergent, then it is Cauchy,

(2) If $(\mathbf{x}_n)$ is Cauchy, then $(\mathbf{x}_n)$ is bounded, but $(\mathbf{x}_n)$ need not converge in X. (i.e. $(1/n)$ as a sequence in the metric space $(0, 1]$).

(3) If $(\mathbf{x}_n)$ is Cauchy and has a subsequence that converges in X, then it converges in X as well.

(4) If $x \in X$, $x' \in X$, and if $(\mathbf{x}_n)$ converges to x and to x', then $x' = x$.

Remark 4. While $(0, 1]$ is not a complete metric space, $[0, 1]$ is a complete metric subspace of $\mathbb{R}$. This suggests a tight connection between the closedness of a set and its completeness as a metric subspace. Indeed, a complete subspace of a metric space is closed.

Theorem 2. *(Bolzano–Weierstrass Theorem) Every bounded sequence of real numbers has a convergent subsequence.*

Definition. (In $\mathbb{R}^n$) A subset $X \subset \mathbb{R}^n$ is compact if every sequence $(\mathbf{x}_n)_{n=1}^{\infty}$ of points in X has a convergent subsequence $(\mathbf{x}_{n_i})_{i=1}^{\infty}$ with limit $\mathbf{x} = \lim_{i\to\infty} \mathbf{x}_{n_i}$ in X. (Actually it is a proposition, please see p.134 OK for the proof)

Remark 5. Bolzano–Weierstrass Theorem states that every bounded sequence has a convergent subsequence. Using this new language, we may deduce that every subset of $\mathbb{R}^n$ that is both closed and bounded is compact.

Lemma 1. *A compact subset of $\mathbb{R}^n$ is closed and bounded.*[4]

Proof. Let $\mathcal{C}$ be a compact subset of $\mathbb{R}^n$. Suppose that $\mathbf{x}$ is a limit point of $\mathcal{C}$, say $\mathbf{x} = \lim_{n\to\infty} \mathbf{c}_n$ for a sequence $(\mathbf{c}_n)$ in $\mathcal{C}$. Then this sequence has a subsequence $(\mathbf{c}_{n_k})$ converging to a point $\mathbf{c}$ in $\mathcal{C}$. Therefore,

$$\mathbf{x} = \lim_{n\to\infty} \mathbf{c}_n = \lim_{i\to\infty} \mathbf{c}_{n_i} = \mathbf{c} \in \mathcal{C}.$$

Thus $\mathcal{C}$ is closed.

To show that $\mathcal{C}$ is bounded, suppose that it were unbounded. That means that there is a sequence $\mathbf{c}_n \in \mathcal{C}$ such that $\|\mathbf{c}_n\| > n$ for each $n \geq 1$. Consider the sequence $(\mathbf{c}_n)$. If there were a convergent subsequence $(\mathbf{c}_{n_i})$ with limit $\mathbf{c}$, it would follow that

$$\|\mathbf{c}\| = \lim_{i\to\infty} \|\mathbf{c}_{n_i}\| \geq \lim_{i\to\infty} n_i = +\infty.$$

$\mathcal{C}$ must be bounded. □

[4] A subset $\mathcal{X}$ of $\mathbb{R}^n$ is called *bounded* if there is a real number R such that $\mathcal{X}$ is contained in the ball $B_R(0)$.

Theorem 3. *(Heine-Borel) A subset of $\mathbb{R}^n$ is compact iff it is closed and bounded.*

In every analysis related course, especially economic theory and theoretical econometrics, a lot of effort is spent finding the maximum or minimum of various functions. Sometimes there were theoretical reasons why such a point should exist. However, generally it was taken on blind faith (as our blind faith in utility maximization). A positive news is that the function may attain its maximum value even the function is quite bad, in the sense that it is not differentiable. To conclude this section, we will see how our new topological tools can explain this phenomenon.

Theorem 4. *If f is a continuous mapping of a compact metric space $\mathcal{X}$ into a metric space $\mathcal{Y}$. Then $f(\mathcal{X})$ is compact.*

Theorem 5. *(Extreme Value Theorem) Let $\mathcal{C}$ be a compact subset of $\mathbb{R}^n$, and let f be a continuous function from $\mathcal{C}$ into $\mathbb{R}^n$. Then there are points $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$ attaining the minimum and maximum values of f on $\mathcal{C}$. That is*

$$f(\mathbf{a}) \leq f(\mathbf{x}) \leq f(\mathbf{b})$$

for all $\mathbf{x} \in \mathcal{C}$.

(Think about the hyperbolic cotangent function.)
How is the connection between compactness and completeness in metric spaces?

Theorem 6. *(p.171 OK) A metric space is compact iff it is complete and totally bounded.*

1.4 Exercises

1. Prove the following statement: A set $A \subset \mathbb{R}^n$ is open iff if the complement of A, $A^c = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \notin A\}$, is closed. [answer: Let A be open. $(\mathbf{x}_n)$ is a sequence in A^c with limit $\mathbf{x}$. If $\mathbf{a}$ if any point in A, there is $r > 0$ such that $B_r(\mathbf{a})$ is contained in A. Hence $\|\mathbf{a} - \mathbf{x}_n\| \geq r$ for all $n \geq 1$. Therefore

   $$\|\mathbf{a} - \mathbf{x}\| = \lim_{n\to\infty} \|\mathbf{a} - \mathbf{x}_n\| \geq r.$$

   In particular, $\mathbf{x} \neq \mathbf{a}$. This is true for every point in A and hence $\mathbf{x} \in A^c$. This is A^c is closed. Conversely, suppose A is not open. Then there is some $\mathbf{a} \in A$ such that for every $r > 0$, the ball $B_r(\mathbf{a})$ is not contained in A. In particular, if $r = 1/n$, we can define $\mathbf{x}_n \in A^c$ such that $\|\mathbf{a} - \mathbf{x}_n\| < 1/n$. Then $\mathbf{a} = \lim_{n\to\infty} \mathbf{x}_n$ is a limit point of A^c belonging to A. Hence $A^c$is not closed. Contradiction.]

2. Prove Proposition 1.

3. Prove Extreme value theorem.

4. In l_∞, we mean the set of all bounded real sequences, that is

$$l_\infty := \{(x_m) \in \mathbb{R} : \sup\{|x_m| : m \in \mathbb{N}\} < \infty\}.$$

It is implicitly understood that this set is endowed with the metric $d_\infty : l_\infty \times l_\infty \to \mathbb{R}^+$ with

$$d_\infty((x_m), (y_m)) := \sup\{|x_m - y_m| : m \in \mathbb{N}\}.$$

The metric is called sup-metric on the set of all bounded real sequences. Prove (l_∞, d_∞) is s a complete metric space. Hint: You can refer to the proof of Example 11 (4) in (OK).

Chapter 2

Continuity, Convergence in Functions

You must be familiar with Riemann integrals. But in this course, what we will use is the Lebesgue integrals. Riemann integral, which forms the curical part in introductory analysis courses, has many deficiencies and thus it does not suffice for more advanced applications. To realize these deficiencies, we now have a tour in continuity and convergence for functions. It seems that the standard continuity and convergence requirements are too "relaxing" for Riemann integral. We need stronger requirements for both the continuity and convergence.

A trivial step is to assume f bounded and continuous on $[a, b]$ except at many points.

Example 7. (Dirichlet) Consider a function which is continuous at "almost all" points of $[a, b]$.

$$f(x) = \begin{cases} \frac{1}{n} & \text{if } x = \frac{m}{n} \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

This function is Riemann-integrable.

Above example is to show that f is Riemann-integrable iff it is continuous at "almost all" points of $[0, 1]$. What about the function is "too discontinuous"? In other words, what if f is discontinuous at "almost all" points of $[0, 1]$.

Example 8. Consider the upper and lower sums for the indicator function $\mathbf{1}_{\mathbb{Q}}$ of $\mathbb{Q}$ over $[0, 1]$. When we partition $[0, 1]$, each subinterval must contain both rational and irrational pointsl; thus each upper sum is 1 and each lower sum 0. Hence we cannot calculate the Riemann integral of f over the $[0, 1]$.

Remark 6. The example shows that we have no easy way of integrating over more general sets, or of integrating functions whose values are distributed 'awkwardly' over sets that differ greatly from intervals.

Apart from the problem of continuity, there is a more serious issue for Riemann integral: Riemann integral doesn't interact well with taking the limit of a sequence of functions. The difficulties arise if the function (f_n) converge to f pointwise, i.e. $f_n(x) \to f(x)$ for all x. 1. The limit need not be Riemann integrable and so the convergence question does not even

make sense. 2. Even the limit is Riemann integrable, the convergence of Riemann integrals does not hold.

Example 9. Take $f = \mathbf{1}_{\mathbb{Q}}$, $f_n = \mathbf{1}_{A_n}$ where $A_n = \{q_1, \ldots, q_n\}$ and the sequence (q_n), $n \geq 1$ is an enumeration of the rationals, so that (f_n) is monotone increasing. We know $f_n(x) \to f(x)$ for all x. However, as in the above example, f is not Riemann integrable.

Example 10. Let $f = 0$. Consider the interval $[0, 1]$ and $f_n(x)$ such that

$$f_n(x) = \begin{cases} 4n^2 x & \text{if } 0 \leq x < \frac{1}{2n}, \\ 4n - 4n^2 x & \text{if } \frac{1}{2n} \leq x < \frac{1}{n}, \\ 0 & \text{if } \frac{1}{n} \leq x \leq 1. \end{cases}$$

This is a continuous function with integral 1. On the other hand, the sequence $f_n(x)$ converges to $f = 0$ for all x on $[0, 1]$.

Remark 7. To avoid the convergence problem, we can use the idea of uniform convergence.

2.1 Uniform Continuity

Definition. Let $(\mathcal{M}, d(x, y))$ and $(\mathcal{N}, \rho(u, v))$ be metric spaces. A mapping $f : \mathcal{M} \to \mathcal{N}$ is said to be *uniformly continuous* if for every $\epsilon > 0$ there is a $\delta > 0$ such that $\rho(f(x), f(y)) < \epsilon$ for every $x, y \in \mathcal{M}$ with $d(x, y) < \delta$.

Uniformly continuous mappings are continuous in particular. One can check that $f : \mathcal{M} \to \mathcal{N}$ is uniformly continuous iff for every pair of sequences $\{x_j\}_{j=1}^{\infty}$, $\{y_j\}_{j=1}^{\infty}$ of elements of $\mathcal{M}$ such that $\lim_{j\to\infty} d(x_j, y_j) = 0$,

$$\lim_{j\to\infty} \rho(f(x_j), f(y_j)) = 0.$$

It is easy to see that the composition of two uniformly continuous mappings is uniformly continuous, using the definition of uniform continuity in terms of ϵ's and δ's, or the characterization of uniform continuity in terms of sequences.

Remark. If $f : \mathcal{M} \to \mathcal{N}$ is uniformly continuous and $\{w_l\}_{l=1}^{\infty}$ is a Cauchy sequence of elements of $\mathcal{M}$, then $\{f(w_l)\}_{l=1}^{\infty}$ is a Cauchy sequence in $\mathcal{N}$. If f is uniformly continuous and $\mathcal{E} \subseteq \mathcal{M}$ is totally bounded, then $f(\mathcal{E})$ is totally bounded in $\mathcal{N}$. The sum of two uniformly continuous functions is uniformly continuous, as is the product of such a function and a constant. The product of two bounded uniformly continuous functions is uniformly continuous.

Proposition 2. *Let $(\mathcal{M}, d(x, y))$ and $(\mathcal{N}, \rho(u, v))$ be metric spaces. A mapping $f : \mathcal{M} \to \mathcal{N}$ is continuous. If $\mathcal{M}$ is compact, then f is uniformly continuous.*

Proof. To see this, let $\epsilon > 0$ be given. For each $x \in \mathcal{M}$, there is a $\delta(x) > 0$ such that

$$\rho(f(y), f(x)) < \frac{\epsilon}{2}$$

when $y \in \mathcal{M}$ and $d(y, x) < \delta(x)$, by continuity. If $B_{\delta(x)/2}(x)$ is the open ball in M with center x and radius $\delta(x)/2$, then the open balls $B_{\delta(x)/2}(x)$, $x \in \mathcal{M}$, cover $\mathcal{M}$. By compactness, there are finitely many elements $x_1, \ldots, x_k$ of $\mathcal{M}$ such that $\mathcal{M} \subseteq \bigcup_{i=1}^{k} B_{\delta_i/2}(x_i)$. Put $\delta = \min(\delta(x_1)/2, \ldots, \delta(x_k)/2)$, and let x, y be arbitrary elements of $\mathcal{M}$ such that $d(x, y) < \delta$. There is an i, $1 \leq i \leq k$, such that $x \in B(x_i)$, and for which $d(y, x_i) < \delta(x_i)/2 + \delta \leq \delta(x_i)$, by the triangle inequality. It follows that

$$\rho(f(x), f(y)) \leq \rho(f(x), f(x_i)) + \rho(f(x_i), f(y)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

as desired. □

A continuous map from a metric space $(\mathcal{X}, d)$ into another metric $(\mathcal{Y}, \rho)$ remains continuous if we remetrize $\mathcal{X}$ ($\mathcal{Y}$) by a equivalent metric to d (ρ). This is in general not true for uniform continuity.

Proposition 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces. $f : \mathcal{X} \to \mathcal{Y}$ is uniformly continuous. If (x_n) is a Cauchy sequence in $\mathcal{X}$, then $f(x_n)$ is Cauchy.*

Let's see an application of uniform continuity.

Definition. Let $(\mathcal{M}, d(x, y))$ be a metric space. $\varphi : \mathcal{M} \to \mathcal{M}$ is a strict contraction in the sense that there is a positive real number $c < 1$ such that

$$d(\varphi(x), \varphi(y)) \leq c\, d(x, y)$$

for every $x, y \in \mathcal{M}$. If $x, x' \in \mathcal{M}$ are fixed by φ, which is to say that $\varphi(x) = x$ and $\varphi(x') = x'$, then $d(x, x') = d(\varphi(x), \varphi(x')) \leq c\, d(x, x')$, which implies that $d(x, x') = 0$ since $c < 1$ and hence that $x = x'$.

Theorem 7. *(Contraction Mapping Theorem) Let $(\mathcal{M}, d(x, y))$ be a metric space and $\varphi : \mathcal{M} \to \mathcal{M}$ is a strict contraction. If $\mathcal{M}$ is complete, then there is an $x \in \mathcal{M}$ such that $\varphi(x) = x$.*

Proof. Let z be any point in $\mathcal{M}$, and consider the sequence $\{z_n\}_{n=1}^{\infty}$ of elements of $\mathcal{M}$ defined recursively by $z_1 = z$, $z_{n+1} = \varphi(z_n)$. Thus $d(z_{n+2}, z_{n+1}) \leq c\, d(z_{n+1}, z_n)$ for each n. By repeating this, we get

$$d(z_{n+1}, z_n) \leq c^{n-1}\, d(z_2, z_1)$$

for every $n \geq 1$, and hence

$$\begin{aligned} d(z_{n+l}, z_n) &\leq \sum_{i=0}^{l-1} d(z_{n+i+1}, z_{n+i}) \leq \sum_{i=0}^{l-1} c^{n+i-1}\, d(z_2, z_1) \\ &\leq c^{n-1} \Big(\sum_{i=0}^{\infty} c^i \Big)\, d(z_2, z_1) = \frac{c^{n-1}}{1-c}\, d(z_2, z_1) \end{aligned}$$

for $l, n \geq 1$. This implies that $\{z_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathcal{M}$, which converges because $\mathcal{M}$ is complete. Moreover,

$$\varphi\Big(\lim_{n\to\infty} z_n\Big) = \lim_{n\to\infty} \varphi(z_n) = \lim_{n\to\infty} z_{n+1} = \lim_{n\to\infty} z_n$$

since φ is continuous. □

2.2 Uniform convergence

Definition. Let $(\mathcal{X}, d)$, $(\mathcal{Y}, \rho)$ be metric spaces. A mapping $f : \mathcal{X} \to \mathcal{Y}$ is said to be *bounded* if $f(\mathcal{X})$ is a bounded set in $\mathcal{Y}$. Let $B(\mathcal{X})$ be the space of bounded functions. Let $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ be the space of bounded continuous mappings from $\mathcal{X}$ into $\mathcal{Y}$. The *supremum metric* on $\mathcal{C}_b(\mathcal{X}, \mathcal{Y})$ is defined by

$$d_\infty(f_1, f_2) = \sup\{\rho(f_1(x), f_2(x)) : x \in \mathcal{X}\}$$

for $f_1, f_2 \in \mathcal{C}_b(\mathcal{X}, \mathcal{Y})$.

Definition. Let $(\mathcal{X}, d)$ be a metric space, and let $\mathcal{C}_b(\mathcal{X})$ be the space of bounded continuous real-valued functions on $\mathcal{X}$, i.e., $\mathcal{C}_b(\mathcal{M}) = \mathcal{C}_b(\mathcal{M}, \mathbb{R})$.

Definition. Let $\mathcal{X}$ be a set, let $(\mathcal{Y}, \rho(u, v))$ be a metric space, and let f_n, $n \geq 1$, and f be functions on $\mathcal{X}$ with values in $\mathcal{Y}$. If

$$\lim_{n\to\infty} |f_n(x) - f(x)| = 0$$

in $\mathcal{Y}$ for all $x \in \mathcal{X}$, then we say that the sequence of functions f_n converges *pointwise* to f on $\mathcal{X}$.

Definition. For any bounded function $f \in \mathcal{B}(\mathcal{X})$ and continuous bounded functions $f_n \in \mathcal{C}_b(\mathcal{X})$, uniform converges means

$$\lim_{n\to\infty} \sup\{|f_n(x) - f(x)| \, : \, x \in \mathcal{X}\} = 0.$$

where $\sup\{|f_n(x) - f(x)| \, : \, x \in \mathcal{X}\}$ is called the supremum metric.

Remark 8. In the case of bounded functions, uniform convergence is identical to convergence with respect to d_∞. The following is a general definition of uniform convergence.

Definition. Let $\mathcal{X}$ be a set, let $(\mathcal{Y}, \rho(u, v))$ be a metric space, and let f_n, $n \geq 1$, and f be functions on $\mathcal{X}$ with values in $\mathcal{Y}$. If for every $\epsilon > 0$ there is a positive integer N such that

$$\rho(f_n(x), f(x)) < \epsilon$$

when $n \geq N$ in $\mathcal{Y}$ for all $x \in \mathcal{X}$, then we say that the sequence of function f_n uniformly converges to f on $\mathcal{X}$.

It is easy to see that uniform convergence implies pointwise convergence.

Proposition 4. *For the $d_\infty(\cdot,\cdot)$ metric on $\mathcal{C}_b(\mathcal{X},\mathcal{Y})$, a sequence (f_n) of elements of $\mathcal{C}_b(\mathcal{X},\mathcal{Y})$ converges to $f \in \mathcal{C}_b(\mathcal{X},\mathcal{Y})$ in the supremum metric iff (f_n) converges to f uniformly.*

Proposition 5. *Suppose that $(\mathcal{X},d)$, $(\mathcal{Y},\rho)$ are metric spaces, and let (f_n) be a sequence of continuous mappings from $\mathcal{X}$ to $\mathcal{Y}$. If f_n converges uniformly to a mapping $f:\mathcal{X}\to\mathcal{Y}$, then f is continuous too.*

Proof. For let $x \in \mathcal{X}$ and $\epsilon > 0$ be given. Since f_n converges uniformly to f, there is a positive integer L such that

$$\rho(f_n(y), f(y)) < \frac{\epsilon}{3} \quad \text{for every } y \in \mathcal{X}$$

when $n \geq N$. Because f_N is continuous at x, there is a $\delta > 0$ such that $\rho(f_N(y), f_N(x)) < \epsilon/3$ when $y \in \mathcal{X}$ and $d(y,x) < \delta$. Therefore,

$$\begin{aligned} \rho(f(y), f(x)) &\leq \rho(f(y), f_N(y)) + \rho(f_N(y), f_N(x)) + \rho(f_N(x), f(x)) \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \end{aligned}$$

when $y \in \mathcal{X}$ and $d(y,x) < \delta$, as desired. □

The same argument shows that f is uniformly continuous if the f_n's are. As another variant, if (x_n) is a sequence of elements of $\mathcal{X}$ that converges to $x \in \mathcal{X}$, then one can show that $(f_n(x_n))$ converges to $f(x)$ in $\mathcal{Y}$ under these conditions.

Definition. $\mathcal{C}(\mathcal{X})$ is a space of all continuous functions $f:\mathcal{X}\to\mathbb{R}$.

A general $\mathcal{C}(\mathcal{X})$ cannot be metrized by sup-metric, because a continuous function need not be bounded. However, when $\mathcal{X}$ is compact, the sup-metric can metrize $\mathcal{C}(\mathcal{X})$.

Remark 9. Sup-metric will induce the completeness of Riemann integrals on $\mathcal{C}([0,1])$. A sequence (f_n) in the space $\mathcal{C}([0,1])$ converges uniformly to f if $a_n = \sup\{|f_n(x) - f(x)| : 0 \leq x \leq 1\}$ converges to 0. In this case one can easily prove the convergence of the Riemann integrals e.g. $\int_0^1 f_n(x)dx \to \int_0^1 f(x)dx$.

Remark 10. Sup-metric $\sup\{|f(x) - g(x)| : 0 \leq x \leq 1\}$ has nothing to do with integration and the uniform convergence is too restrictive for many application. A more natural concept of metric is given by $\|f-g\|_1 := \int_0^1 |f(x)-g(x)|dx$ called $L_1([0,1])$, leads to another problem. Consider

$$g_n(x) = \begin{cases} 0 & \text{if } 0 \leq x < \frac{1}{2}, \\ n(x-\frac{1}{2}) & \text{if } \frac{1}{2} \leq x < \frac{1}{2} + \frac{1}{n}, \\ 1 & \text{otherwise.} \end{cases}$$

It can be shown that $\int_0^1 |g_n(x) - g_m(x)|dx \to 0$ as $n, m \to \infty$ (as a Cauchy sequence). However this sequence by no means will converge to a continuous function f, since the pointwise limit is $f(x) = 1$ for $x > \frac{1}{2}$ and 0 otherwise, so that $f = \mathbf{1}_{(\frac{1}{2},1]}$. So the space $\mathcal{C}([0,1])$ is too small for $L_1([0,1])$ from this point view.

Proposition 6. *Suppose that $\mathcal{N}$ is complete, then $\mathcal{C}_b(\mathcal{M}, \mathcal{N})$ is complete with respect to the supremum metric $d_\infty(\cdot, \cdot)$.*

Proof. Let $\{f_j\}_{j=1}^\infty$ be a Cauchy sequence in $\mathcal{C}_b(\mathcal{M}, \mathcal{N})$, so that for every $\epsilon > 0$ there is a positive integer $L(\epsilon)$ such that

$$d_\infty(f_j, f_l) < \epsilon$$

when $j, l \geq L(\epsilon)$. In particular, $\{f_j(x)\}_{j=1}^\infty$ is a Cauchy sequence in $\mathcal{N}$ for every $x \in \mathcal{M}$, which converges to an element $f(x)$ of $\mathcal{N}$ since $\mathcal{N}$ is complete. One can check that

$$\rho(f_j(x), f(x)) \leq \epsilon \quad \text{for every } x \in \mathcal{M}$$

when $j \geq L(\epsilon)$, which means that $\{f_j\}_{j=1}^\infty$ converges uniformly to f and that f is continuous, as desired. □

Thus $\mathcal{C}_b(\mathcal{M})$ is a vector space over the real numbers with respect to the usual operations of pointwise addition of functions and multiplication of functions by constants, and moreover $\mathcal{C}_b(\mathcal{M})$ is a commutative algebra with respect to the operation of pointwise multiplication of functions.

Definition. The *supremum or L_∞-norm* of a function $f \in \mathcal{C}_b(\mathcal{M})$ is

$$\|f\|_\infty = \sup\{|f(x)| : x \in \mathcal{M}\},$$

and it satisfies the commutative algebra

$$\|f_1 + f_2\|_\infty \leq \|f_1\|_\infty + \|f_2\|_\infty$$

and

$$\|f_1 \, f_2\|_\infty \leq \|f_1\|_\infty \, \|f_2\|_\infty$$

for every $f_1, f_2 \in \mathcal{C}_b(\mathcal{M})$.

It is easy to see that $\|f_1 - f_2\|_\infty$ is the same as the supremum metric on $\mathcal{C}_b(\mathcal{M})$. Using the triangle inequality, one can check that $f_p(x) = d(x, p)$ is a continuous function on $\mathcal{M}$ for every $p \in \mathcal{M}$. These functions are bounded when $\mathcal{M}$ is bounded, and otherwise $\min(f_p(x), r)$ is a bounded continuous function on $\mathcal{M}$ for every $r \geq 0$. This shows that there are always a lot of bounded continuous real-valued functions on any metric space, and in particular these functions separate elements of $\mathcal{M}$, in the sense that for every $x, y \in \mathcal{M}$ with $x \neq y$ there is an $f \in \mathcal{C}_b(\mathcal{M})$ such that $f(x) \neq f(y)$.

Corollary 1. *(OK p. 249) For any metric space $\mathcal{M}$, $\mathcal{C}_b(\mathcal{M})$ is complete with respect to the supremum metric $d_\infty(\cdot, \cdot)$.*

Finally, we conclude this section with a discussion about interchange the operations of taking limits by uniform convergence.

Example 11. Let (x_k) be a sequence in $\mathcal{X}$ and $x_k \to x$. Let (φ_m) be a sequence in $\mathcal{C}(\mathcal{X})$ such that $\varphi_m \to \varphi$ uniformaly.

$$\lim_{k\to\infty} \lim_{m\to\infty} \varphi_m(x_k) = \lim_{k\to\infty} \varphi(x_k) = \varphi(x) = \lim_{m\to\infty} \varphi_m(x) = \lim_{m\to\infty} \lim_{k\to\infty} \varphi_m(x_k)$$

2.3 Compactness for $\mathcal{C}(\mathcal{X})$ (Optional)

Does a compact set $\mathcal{X}$ give a compact $\mathcal{C}(\mathcal{X})$?

Theorem 8. *(Sequential Compact in metric spaces) A subset $\mathcal{S}$ of a metric space $\mathcal{X}$ is compact iff every sequence in $\mathcal{S}$ has a convergent sequence to a point in $\mathcal{S}$.*

Example 12. $\mathcal{C}([0,1])$ is not compact. For example, $f_n(t) := t^n$, $n \in \mathbb{N}$, then f_m is a sequence in $C([0,1])$ without a convergent subsequence.

To maintain compactness, we need an additional condition.

Definition. Let $\mathcal{F} \subset \mathcal{C}(\mathcal{X})$. $\mathcal{F}$ is equi-continuous at $x \in \mathcal{X}$ if, for any given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$|\varphi(x) - \varphi(y)| < \varepsilon$$

for all $\varphi \in \mathcal{F}$ and $y \in B_\delta(x) \subset \mathcal{X}$.

Definition. Let $\mathcal{F} \subset \mathcal{C}(\mathcal{X})$. $\mathcal{F}$ is uniform equi-continuous if, for any given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$|\varphi(x) - \varphi(y)| < \varepsilon$$

for all $\varphi \in \mathcal{F}$ and any $x, y \in \mathcal{X}$ with $d(x,y) < \delta$.

Just as a continuous real function on a compact set is uniformly continuous, an equicontinuous $\mathcal{F} \subset \mathcal{C}(\mathcal{X})$ is uniformly equicontinuous whenver $\mathcal{T}$ is compact.

Theorem 9. *(Arzela-Ascoli Theorem, OK p. 264) Let $\mathcal{X}$ be a compact metric psace, and $\mathcal{F} \subseteq \mathcal{C}(\mathcal{X})$. Then $\mathcal{F}$ is compact iff it is closed, bounded and equicontinuous.*
/

[Connection between Arzela-Ascoli and Heine-Borel: from wiki In view of Ascoli's theorem, a sequence in C(X) converges uniformly if and only if it is equicontinuous and converges pointwise. The hypothesis of the statement can be weakened a bit: a sequence in C(X) converges uniformly if it is equicontinuous and converges pointwise on a dense subset to some function on X (not assumed continuous).[6] This weaker version is typically used to prove Ascoli's theorem for separable compact spaces. Another consequence is that the limit of an equicontinuous pointwise convergent sequence of continuous functions on a metric space, or on a locally compact space, is continuous. (See below for an example.) In the above, the hypothesis of compactness of X cannot be relaxed. To see that, consider a compactly supported continuous function g on R with g(0) = 1, and consider the equicontinuous sequence of functions {fn} on R defined by fn(x) = g(x − n). Then, fn converges pointwise to 0 but does not converge uniformly to 0.]

2.4 Exercises

1. Give another proof of Proposition 1. Answer: [Suppose that $\{x_j\}_{j=1}^{\infty}$, $\{y_j\}_{j=1}^{\infty}$ are sequences of elements of $\mathcal{M}$ such that $\lim_{j\to\infty} d(x_j, y_j) = 0$, but $\rho(f(x_j), f(y_j))$ does not converge to 0. This means that there is an $\epsilon > 0$ such that $\rho(f(x_j), f(y_j)) \geq \epsilon$ for infinitely many j. Without loss of generality, we may suppose that this holds for all j, since otherwise we can replace our sequences with the subsequences where it does hold. By compactness, there is a strictly increasing sequence $\{j_l\}_{l=1}^{\infty}$ of positive integers such that $\{x_{j_l}\}_{l=1}^{\infty}$ converges to a point $x \in \mathcal{M}$, and we also get that $\{y_{j_l}\}_{l=1}^{\infty}$ converges to x too, since $d(x_{j_l}, y_{j_l}) \to 0$ as $l \to \infty$. Continuity of f at x implies that $\{f(x_{j_l})\}_{l=1}^{\infty}$ and $\{f(y_{j_l})\}_{l=1}^{\infty}$ both converge to $f(x)$ in N, and hence that $\lim_{l\to\infty} \rho(f(x_{j_l}), f(y_{j_l})) = 0$, a contradiction.]

Chapter 3

Linear Functionals and Representation Theorem

3.1 Linear Operators and Linear Functionals

Definition. (L_p-norm on $[0,1]$) Let f be a continuous real valued function on the closed unit interval $[0,1]$ in the real line. For each positive real number p, $\|f\|_p = \Big(\int_0^1 |f(x)|^p \, dx \Big)^{1/p}$.

This obviously satisfies the triangle inequality when $p = 1$, and one can show that it also holds for $p > 1$ using the same argument as for finite sums. Thus $\|f\|_p$ defines a norm on the vector space of continuous functions on $[0,1]$ when $p \geq 1$. We also have that $\|f\|_p \leq \|f\|_\infty$ where $\|f\|_\infty$ is the supremum norm of f. If $0 < p < q < \infty$, then one can check that $\|f\|_p \leq \|f\|_q$.

Remark. L_p-norm on the real line with compact support is $\|f\|_p = \Big(\int_{-\infty}^{\infty} |f(x)|^p \, dx \Big)^{1/p}$. The integral here can be reduced to one on a bounded interval, since f has compact support. However, $\|f\|_p$ is normally neither monotone increasing nor decreasing in p in this case.

Remark. If $p = 2$, then these norms associated to suitable inner products. On the unit interval, these inner products are given by $\langle f_1, f_2 \rangle = \int_0^1 f_1(x) \, f_2(x) \, dx$ in the real case. The inner products on continuous functions with compact support on $\mathbb{R}$ are defined similarly.

Definition. A norm vector space (or called norm linear space) $(X, \|\cdot\|)$ is a vector space X equipped with a norm $\|\cdot\|$.

Definition. A Banach space is a normed linear space that is a complete metric space with respect to the metric derived from its norm.

Definition. Let $T : X \to Y$ be a linear map between linear spaces X, Y. The kernel (or null space) of T, denoted by $\ker T$ is the subset of X defined by $\ker T := \{x \in X : Tx = 0\}$. The range of T, denoted by $\mathrm{ran} T$, is the subset of Y defined by $\mathrm{ran} T := \{y \in Y : \text{ exists } x \in X \text{ s.t. } Tx = y\}$.

Example 13. $C([a,b])$ equipped with sup-norm is a Banach space. $C(K)$ on a compact space K equipped with sup-norm is a Banach space. l_p and $L_p([a,b])$ are both Banach spaces for $1 \leq p \leq \infty$.

Definition. A linear operator (or called map) $T : X \to Y$ between linear spaces X, Y is a function satisfying

$$T(ax + by) = aT(x) + bT(y)$$

for all $x, y \in X$ and $a, b \in \mathbb{R}$.

If X, Y are norm space, then we can define a bounded linear map.

Definition. Let V and W be two normed linear spaces, with norm $\|\cdot\|_V$ and $\|\cdot\|_W$. A linear map $T : V \to W$ is bounded if there is a constant $M \geq 0$ such that $\|T(x)\|_W \leq M\|x\|_V$ for all $x \in V$. The operator norm (or uniform norm) $\|T\|$ of T is given by

$$\|T\| = \inf \{M \,:\, \|T(x)\|_W \leq M\|x\|_V\}.$$

It is easy to see that every linear mapping is bounded when $V \in \mathbb{R}^n$. If $V = W$ with the same norm, then the identity mapping $I(v) = v$ is bounded, with $M = 1$.

Definition. The space $\mathcal{L}(V, W)$ is the set of all linear maps from V into W. And the set of all bounded linear maps is deonoted by the space $\mathcal{BL}(V, W)$ which is a linear subspace of $\mathcal{L}(V, W)$.

Remark 11. Equivalent expressions for $\|T\|$ are:

$$\|T\| = \sup_x \frac{\|T(x)\|}{\|x\|};\ \|T\| = \sup_{\|x\|\leq 1} \|T(x)\|;\ \|T\| = \sup_{\|x\|=1} \|T(x)\|.$$

For linear maps, boundedness is equivalent to continuity. (Note that the previous definition is taken infimum over constant M but supremum is taken over x.)

Proposition 7. *If T is a bounded linear map from V into W, then $\|T(v_1) - T(v_2)\|_W \leq A\|v_1 - v_2\|_V$ for every $v_1, v_2 \in V$, and it follows that $T : V \to W$ is uniformly continuous. Conversely, if a linear map $T : V \to W$ is continuous at the origin, then it is bounded.*

Proof. ($\Rightarrow$) The linear map implies that $\|T(v) - T(w)\|_W = \|T(v - w)\|_W \leq A\|v - w\|_V$ for every $v, w \in V$.

($\Leftarrow$) Suppose T is continuous at 0. $\|x\| = 0$ if $x = 0$ in the norm vector space. $\|T(x)\| \leq A\|x\| = 0$ if $x = 0$. Since T is linear in a norm space, $T(0) = 0$. Then there is a $\delta > 0$ such that $\|T(v)\|_W < 1$ satisfies $\|v\|_V < \delta$. For any non-zero $x \in V$, we define

$$\tilde{x} = \delta \frac{x}{\|x\|}.$$

Then $\|\tilde{x}\| \leq \delta$, so $\|T\tilde{x}\| \leq 1$. It follows from the linearity of T that

$$\|Tx\| = \frac{\|x\|}{\delta}\|T\tilde{x}\| \leq M\|x\|,$$

where $M = 1/\delta$. Thus T is bounded. □

The proof shows that if a linear map is continuous at zero, then it is continuous at every point.

Definition. A linear map $T : V \to \mathbb{R}$ from a vector space V into $\mathbb{R}$ satisfying

$$T(ax + by) = aT(x) + bT(y)$$

for all $x, y \in V$ and $a, b \in \mathbb{R}$ is also known as a *linear functional* on V. A linear functional on a vector space V with a norm $\|\cdot\|_V$ is bounded if it is bounded with respect to the standard norm on $\mathbb{R}$ such that $|T(x)| \leq M\|x\|_V$.

Arrow and Debreu (1954) define a price system to be a continuous linear functional.

Definition. (Finite dimension) Every finite-dimensional normed linear space is a Banach space. Every linear operator on a finite-dimensional space is continuous, and that all norms on a finite-dimensional space are equivalent, e.g. $c\|\cdot\|_V \leq \|\cdot\|_W \leq C\|\cdot\|_V$. None of these statements is true for infinite-dimensional linear spaces.

Example 14. Any $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$, $T(x) = Ax$ is an linear operator $T : \mathbb{R}^n \to \mathbb{R}^m$ and $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

Example 15. Let $L(x) := \int_0^1 f(x)dx$. Then $L : C[0,1] \to \mathbb{R}$ is a linear functional. The space of all continuously differentiable functions on $[0,1]$ is $C^1[0,1]$. Differenitalbe operator $D : C[0,1] \to C[0,1]$ by $D(f) := f'$ is a linear operator.

Linear operators play an important role in identifying the basic algebraic relation between two linear spaces.

Proposition 8. *(OK p.394) A subset W of a linear space X is a hyperplane in X iff $W = \{x \in X : T(x) = a\}$ for some $a \in \mathbb{R}$ and nonzero linear functional T on X.*

A hyperplane divides the entire space into two parts. For linear space, it divides the space into two half spaces. The half space could be closed e.g. $\{x \in X : T(x) \geq a\}$ and $\{x \in X : T(x) \leq a\}$ or open (replace inequalities with strictly inequalities).

3.2 Duality

Duality intuitively means that the maps into and out of one type of mathematical object X can be naturally associated to the maps out of and into a dual object X^*. The dual space of a linear space consists of the scalar-valued linear maps on the space. Duality methods play a crucial role in many parts of analysis.

Definition. Given a linear space X, the space of linear functionals on X is called the dual space of X and the space of continuous linear functionals on X is called the topological dual space of X.

In our context, the dual space of X is $\mathcal{L}(X, \mathbb{R})$ and the topological dual space is $\mathcal{BL}(X, \mathbb{R})$. If X is finite dimensional, then $\mathcal{L}(X, \mathbb{R}) = \mathcal{BL}(X, \mathbb{R})$, so both duals are the same. If X is infinite dimensional, then $\mathcal{L}(X, \mathbb{R})$ is much larger than $\mathcal{BL}(X, \mathbb{R})$. However, the dual of $\mathcal{BL}(X, \mathbb{R})$ is a Banach space. Thus from now on, when we say dual, we implicitly refer to the topological dual space.

Definition. Two linear spaces X, Y are linearly isomorphic if there is a one- to-one, onto linear map $T : X \to Y$. If T also preserves norms, meaning that $\|T(x)\| = \|x\|$ for all $x \in X$, then X, Y are isometrically isomorphic.

Definition. A Hamel basis, or algebraic basis, of a linear space is a maximal linearly independent set of vectors. Each element of a linear space may be expressed as a unique finite linear combination of elements in a Hamel basis. Every linear space has a Hamel basis, and any linearly independent set of vectors may be extended to a Hamel basis by the repeated addition of linearly independent vectors to the set until none are left.

A Hamel basis of an infinite-dimensional space is frequently very large. In a normed space, we have a notion of convergence, and we may therefore consider various types of topological bases in which infinite sums are allowed.

Example 16. The dual space X^* of a finite-dimensional space X is linearly isomorphic to X. To see this, pick a basis $\{e_1, \ldots, e_n\}$ of X. The map $T_i : (x_1, \ldots, x_n) \to x_i$ defines $T_i(\sum_{j=1}^n x_j e_j) = x_i$ which is an element of dual space X^*. The linearity of T_i is obvious. For any $\varphi : X \to \mathbb{R}$ on finite-dimensional X,

$$\varphi\left(\sum_{j=1}^{n} x_j e_j\right) = \sum_{j=1}^{n} \varphi(e_j) x_j$$

or $\varphi = \sum_{i=1}^n \varphi(e_i) T_i$. $\{T_1 \ldots, T_n\}$ is a basis of X^* and is called the dual basis of $\{e_1, \ldots, e_n\}$. The dual basis has the property that $T_i(e_j) = \delta_{ij}$ where δ_{ij} is the Kronecker delta function.

Remark 12. One way to obtain a linear functional on a linear space is to start with a linear functional defined on a subspace, extend a Hamel basis of the subspace to a Hamel basis of the whole space and extend the functional to the whole space, by use of linearity and an arbitrary definition of the functional on the additional basis elements.

All linear functionals on a finite-dimensional linear space are bounded. It is not obvious that this extension procedure can be used to obtain bounded linear functionals on an infinite-dimensional linear space, because the extension need not be bounded. In fact, it is possible to maintain boundedness of an extension by a suitable choice of its values at the original subspace, as stated in the following version of the Hahn-Banach theorem.

Theorem 10. *(Hahn-Banach) If Y is a linear subspace of a norm linear space X and $u : Y \to \mathbb{R}$ is a bounded linear functional on Y with $\|u\| = M$, then there is a bounded linear functional $\varphi : X \to \mathbb{R}$ on X such that φ restricted to Y is euqal to u and $\|\varphi\| = M$.*

One consequence of this theorem is that there are enough bounded linear functionals to separate X, meaning that if $\varphi(x) = \varphi(y)$ for all $\varphi \in X^*$, then $x = y$.

Example 17. Linear functions connect Pareto-optimal allocations and competitive equilibria. Second welfare theorem says Pareto-optimal allication under suitable choice of prices can be supported as a competitive euqilibrium. This basically is a mathematical idea of using Hahn-Banach Theorem (or its revised version "separation theorem" for convex sets).

3.3 Representation Theorem

Although a finite-dimensional space is linearly isomorphic with its dual space, there is no canonical way to identify the space with its dual; there are many isomorphisms, depending on an arbitrary choice of a basis. Hilbert spaces are of our interests for a special reason that the topological dual space of a Hilbert space can be identified with the original space in a natural way through the inner product (Riesz representation theorem). The dual of an infinite-dimensional Banach space is, in general, different from the original space.

Definition. A linear functional φ on a Hilbert space $\mathcal{H}$ is bounded or continuous if there exists $M \in \mathbb{R}$ such that $|\varphi(x)| \leq M\|x\|$ for all $x \in \mathcal{H}$. The norm on $\mathcal{H}$ is associated to an inner product.

Some important classes of bounded linear operators are on Hilbert spaces, including projections, unitary operators, and self-adjoint operators. We, however, will only visit projection briefly and then go to the representation theorem.

The Cauchy-Schwarz inequality implies

$$-1 < \frac{\langle x, y\rangle}{\|x\|\|y\|} < 1.$$

Therefore there is a unique θ, $0 < \theta < \pi$, such that $\cos\theta = \langle x, y\rangle / (\|x\|\|y\|)$. We define this θ to be the angle between x and y. On the other hand,

$$\langle x, y\rangle = \|x\|\|y\|\cos\theta.$$

Thus when x is orthogonal to y, the inner product is zero.

Definition. If M and N are subspaces of a linear space X such that every $x \in X$ can be written as $x = y + z$ with $y \in M$ and $z \in N$, then we say that $X = M \oplus N$ is the direct sum of M and N, and we call N as complementary subspace of M in X. The decomposition $x = y + z$ is unique iff $M \cap N = \{0\}$.

Example 18. A given subspace M has many complementary subspaces. $X = \mathbb{R}^3$ and M is a plane through the origin, then any line through the origin that does not lie on M is a complementary subspace.

Definition. If a linear space $X = M \oplus N$, then a projection is $P : X \to X$ of X onto M along N by $Px = y$ where $x = y + z$ with $y \in M$ and $z \in N$. Then $X = \text{ran}P + \text{ker}P$. If M and N are linear subspaces of X, $\text{ran}P = M$ and $\text{ker}P = N$.

Definition. (Another definition of projection) A projection on a linear space X is a linear operator $P : X \to X$ such that $P^2 = P$.

When using Hilbert spaces, we are particularly interested in orthogonal subspaces. Suppose that M is a closed subspace of a Hilbert space $\mathcal{H}$. Then by definition $\mathcal{H} = M \oplus M^\perp$.

Example 19. If $x = y + z$ and $x' = y' + z'$, where $y, y' \in M$ and $z, z' \in M^\perp$. The the orthogonality of M and $M^\perp$ implies that

$$\langle Px, x' \rangle = \langle y, y' + z' \rangle = \langle y, y' \rangle = \langle y + z, y' \rangle = \langle x, Px' \rangle .$$

This equation states that an orthogonal projection is self-adjoint.

Example 20. If $x \in \mathcal{H}$ and $Px \neq 0$, then Cauchy-Schwarz inequality implies that

$$\|Px\| = \frac{\langle Px, Px \rangle}{\|Px\|} = \frac{\langle x, P^2 x \rangle}{\|Px\|} = \frac{\langle x, Px \rangle}{\|Px\|} \leq \|x\|.$$

Therefore $\|P\| \leq 1$. If $P \neq 0$, then there is an $x \in \mathcal{H}$ with $Px \neq 0$ and $\|P(Px)\| = \|Px\|$ so that $\|P\| \leq 1$. Combine this result and Cauchy-Schwarz inequality's result we have $\|P\| = 1$.

Theorem 11. *Each $y \in \mathcal{H}$ determines a linear functional φ_y on $\mathcal{H}$, defined by*

$$\varphi_y(x) = \langle y, x \rangle.$$

The Cauchy-Schwarz inequality implies that φ_y is a bounded linear functional on $\mathcal{H}$, and that $\|\varphi_y\| = \|y\|$.

Remark 13. If $\mathcal{H}$ has finite dimension, then every linear functional on $\mathcal{H}$ is of the form φ_y for some $y \in \mathcal{H}$. If $\mathcal{H}$ is a Hilbert space, then every bounded linear functional on $\mathcal{H}$ is of this form.

Proof. Let φ be a bounded linear functional on $\mathcal{H}$, which is not identically equal to 0.

kerφ is a proper closed subspace of $\mathcal{H}$. There is a nonzero vector $z \in \mathcal{H}$ such that $z \perp \text{ker}\varphi$. We define a linear map $P : \mathcal{H} \to \mathcal{H}$ by

$$Px = \frac{\varphi(x)}{\varphi(z)} z.$$

Then $P^2 = P$. As we know $\mathcal{H} = \text{ran}P \oplus \text{ker}P$. Moreover,

$$\begin{aligned} \text{ran}P &= \{az : a \in \mathbb{R}\} \\ \text{ker}P &= \text{ker}\varphi \end{aligned}$$

so that ran$P \perp \ker P$. It follows that P is an orthogonal projection and

$$\mathcal{H} = \{az : a \in \mathbb{R}\} \oplus \ker\varphi$$

is an orthogonal direct sum $(\ker\varphi)^{\perp} = \text{ran}P = \{az : a \in \mathbb{R}\}$. We therefore can write $x \in \mathcal{H}$ as

$$x = az + n$$

for $a \in \mathbb{R}$ and $n \in \ker\varphi$.

Take the least square argument, we get

$$a = \frac{\langle z, x\rangle}{\|z\|^2}$$

and then we have

$$\varphi(x) = a\varphi(z) + \varphi(n) = a\varphi(z).$$

If we let $y = z/\|z\|^2$, then it follows that $\varphi(x) = \langle y, x\rangle = \varphi_y(x)$ for every $x \in \mathcal{H}$, as desired. □

Th theorem says that $\varphi_y(x) = \langle y, x\rangle$ defines a bounded linear functional on $\mathcal{H}$ for every $y \in \mathcal{H}$. Riesz representation theorem basically states that this φ_y is unique.

Theorem 12. *(Riesz representation) If φ is a bounded linear functional on $\mathcal{H}$, there is a unique vector $y \in \mathcal{H}$ such that*

$$\varphi_y(x) = \langle y, x\rangle.$$

Proof. Suppose $\varphi_y = \varphi_{y'}$. Then $\varphi_y(x) = \varphi_{y'}(x)$ when $x = y - y'$ which implies that $\|y - y'\|^2 = 0$ so $y = y'$. □

Riesz representation theorem characterizes the bounded linear functionals on a Hilbert space.

Example 21. Consider the commodity space $(S, \|\cdot\|)$. S is often determined by economic model itself. But one needs an appropriate norm to make the Hahn-Banach theorem applicable (the norm determines the class of continuous linear functionals on S). It is the usual case that l_2-norm is chosen since every continuous linear functional has an inner product representation. Then Riesz representation guarantees the existence of a set of prices in the usual sense.

Chapter 4

Applications

4.1 Ergodic Theorem

This section is based on Aliprantis and Border (Chapter 20).

Definition. From Riesz representation we know for any bounded operator A in a Hilbert space $\mathcal{H}$, there must exist an adjoint operator, which we denote by A^* (don't be confused this notation with the dual space.) . In fact if $A \in \mathcal{BL}(\mathcal{H}, \mathcal{H})$, then $A^* \in \mathcal{BL}(\mathcal{H}, \mathcal{H})$.

$$\langle x, Ay \rangle = \langle A^* x, y \rangle$$

for any $x, y \in \mathcal{H}$.

To see this, for every $x \in \mathcal{H}$, define a map $\varphi_x(y) = \langle x, Ay \rangle$. By the Riesz representation, there is a unique $z \in \mathcal{H}$ such that $\varphi_x(y) = \langle z, y \rangle$. Let $z = A^* x$. The linearity of A^* comes from the uniquness in the Riesz representation theorem and the linearity of the inner product.

Example 22. For finite dimension cases, the matrix of the adjoint of a linear map on $\mathbb{R}^n$ with matrix A is A^T. That is $x(Ay) = (A^T x)y$.

Example 23. The adjoint relates to the solution of a linear equation $Ax = y$. If A is a bounded linear operator, we can consider an adjoint equation $A^* z = 0$ such that

$$\langle Ax, z \rangle = \langle x, A^* z \rangle = 0.$$

Then a necessary condition for a solution x of $Ax = y$ is that $\langle y, z \rangle = 0$ for all $z \in \ker A^*$. Namely $y \in (\ker A^*)^\perp$ for any $y \in \overline{\text{ran} A}$.

Definition. A bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ is self-adjont if $A^* = A$ such that $\langle x, Ay \rangle = \langle Ax, y \rangle$ for any $x, y \in \mathcal{H}$.

Example 24. For finite dimension cases, the matrix A is self-adjoint iff A is symmetric such that $A = A^T$.

Definition. A linear operator $U : \mathcal{H}_1 \to \mathcal{H}_2$ is said to be unitary if

$$\langle Ux, Uy \rangle_{\mathcal{H}_2} = \langle x, y \rangle_{\mathcal{H}_1}$$

for any $x, y \in \mathcal{H}_1$. Two Hilbert spaces are isomorphic if there is a unitary linear map between them. In other words, unitary operator is one-to-one and onto, and perserves the inner product. Especially if $U : \mathcal{H} \to \mathcal{H}$, then $U^*U = U^*U = I$.

Example 25. You may think unitary operator for matrix U is U^{-1}. Not exactly. Only if U is a complex matrix $U^* = U^{-1}$. For real valued matrces, we say Q is orthogonal if $Q^T = Q^{-1}$.

Ergodic theorem is about equivalence between time averages and probabilistic averages. It is crucial for understanding the dynamics.

Theorem 13. *(von Neumann) A unitary operator U on Hilbert space $\mathcal{H}$. Let $\mathcal{M} = \{x \in \mathcal{H} | Ux = x\}$ be the subspace of vectors and let P be the orthogonal projection onto $\mathcal{M}$. Then for all $x \in \mathcal{H}$, we have*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n x = Px.$$

Proof. We know that for Hilbert space $\mathcal{H} = \ker P \oplus \mathrm{ran} P$.

If $x \in \mathrm{ran} P$, then $\mathcal{M} = \mathrm{ran} P$. Since $Ux = x$ and $Px = x$, the result is trivial.

Suppose $x \in \ker P$. $\mathcal{M} = \{x \in \mathcal{H} | (I - U)x = 0\} = \ker(I - U)$. Because P is a projection onto $\mathcal{M}$, $\mathrm{ran} P = \ker(I - U)$. $Ux = x = U^*x$ by the unitary property. Then

$$\ker P = \ker(I - U)^\perp = \ker(I - U^*)^\perp = \overline{\mathrm{ran}(I - U)}.$$

Thus any $x \in \ker P$ may be approximated by $(I - U)y$. If $x = (I - U)y$, then

$$\frac{1}{N+1} \sum_{n=0}^{N} U^n x = \frac{1}{N+1} \sum_{n=0}^{N} (U^n - U^{n+1})y = \frac{1}{N+1}(y - U^{N+1}y) \to 0.$$

If $x \in \ker P$, then there is a sequence $x_k = (I - U)y_k$ with $x_k \to x$, we have

$$\lim_{N\to\infty} \left\| \frac{1}{N+1} \sum_{n=0}^{N} U^n x \right\| = \lim_{N\to\infty} \left\| \frac{1}{N+1} \sum_{n=0}^{N} U^n (x - x_k) \right\| + \underbrace{\lim_{N\to\infty} \left\| \frac{1}{N+1} \sum_{n=0}^{N} U^n x_k \right\|}_{\to 0} \leq \|x - x_k\|.$$

Note that k is arbitrary and $x_k \to x$, so $\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n x = Px = 0$ for every $x \in \ker P$. □

The result basically says that the averages of U^n converge to P.

Usually the ergodic theorem is applied to probability measure. Let P now be a probability measure on a probability space $\mathcal{P} := (\Omega, \mathcal{F}, P)$.

Definition. A measure preserving map $T : \mathcal{P} \to \mathcal{P}$ is $P(T^{-1}(A)) = P(A)$ for all measurable subsets $A \in \mathcal{F}$.

Remark 14. If you still remember in lecture one we define random variable as a "function", the measure preserving map basically is a self-mapping function $T^{-1}(A) = \{\omega \in \Omega | T(\omega) \in A\}$. Now let the random variable be f (a measurable function on Ω), if T is a measure preserving map, then

$$\mathbb{E}f = \int_\Omega f(\omega)dP(\omega) = \int_\Omega (f \circ T)(\omega)dP(\omega) = \mathbb{E}f \circ T.$$

Then f is invariant under T.

Definition. A one-to-one and onto measure preserving map T on a probability space $\mathcal{P} := (\Omega, \mathcal{F}, P)$ is ergodic iff $f \in L^2(\mathcal{P})$ such that $f = f \circ T$.

Similary as von Neumann's idea, one can consider a unitrary operator $U : L^2(\mathcal{P}) \to L^2(\mathcal{P})$ where $L^2(\mathcal{P})$ is a Hilbert space of second-order random variables on $\mathcal{P}$ so that $Uf = f \circ T$. To see this

$$\langle Uf, Ug \rangle = \int_\Omega f(T(\omega))g(T(\omega)dP(\omega) = \int_\Omega f(\omega)g(\omega)dP(\omega) = \langle f, g \rangle .$$

Theorem 14. *(Probability) A one-to-one and onto measure preserving map $T : \mathcal{P} \to \mathcal{P}$ is ergodic iff for every $f \in L^2(\mathcal{P})$*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} f \circ T^n = \int_\Omega f(\omega)dP(\omega).$$

Example 26. In a dynamical systems, $x_{t+1} = Tx_t$. $\frac{1}{T+1}\sum_{t=1}^{T} x_t$ is equivalent to left side of the above equation. While the right hand side is the probability average of x_0.

4.2 Convex Optimization on Function Spaces

In order to obtain existence of solutions for a general optimization problem, two basic properties are needed: compactness and lower semicontinuity. We know compactness. Now we consider the latter one.

Definition. For a metric space $\mathcal{X}$, let $M : \mathcal{X} \to \mathbb{R}$. The functional M is called lower semicontinuous (lsc) at $x \in \mathcal{X}$ if

$$M(x) \leq \lim_{k\to\infty} \inf M(x_k)$$

for all sequence x_k converging to x.

Theorem 15. *If $M : \mathcal{X} \to \mathbb{R}$ is lsc and the level set $\{x \in \mathcal{X} | M(x) \leq C\}$ is non-empty and compact for some $C \in \mathbb{R}$, then there exists a global minimum such that*

$$M(x^*) = \min_{x\in\mathcal{X}} M(x).$$

Proof. Let $\alpha = \inf_{x\in\mathcal{X}} M(x)$. There exists a sequence (x_k) such that $M(x_k) \to \alpha$. For k sufficiently large, $M(x_k) \leq C$ and hence (u_k) is contained in a compact set. Consequently, there exists a subsequence (x_{k_l}) such that $x_{k_l} \to \tilde{x}$ for some $\tilde{x} \in \mathcal{X}$. By lsc, we have

$$\alpha \leq M(\tilde{x}) \leq \lim_{k\to\infty} \inf M(x_k) = \alpha.$$

Thus, $\tilde{x}$ is a global minimizer. □

Convex problems are an important and interesting class of minimization problems, they have several advantageous properties.

Definition. For a metric space $\mathcal{X}$, a functional $M : \mathcal{X} \to \mathbb{R}$ is convex if for all $\alpha \in [0, 1]$, $u, v \in \mathcal{X}$:

$$M(\alpha u + (1-\alpha)v) \leq \alpha M(u) + (1-\alpha)M(v)$$

A set $\mathcal{C} \subset \mathcal{X}$ is called convex

$$\alpha u + (1-\alpha)v \in \mathcal{C}.$$

An optimization problem is called convex, if both $M(\cdot)$ and $\mathcal{C}$ are convex.

A fundamental property of convex problems is that any local minimizer is also a global one. If $M(\cdot)$ is not smooth, we can employ convexity to prove some fundamental properties

Theorem 16. *Let $\mathcal{X}$ be a Banach space. If $M : \mathcal{X} \to \mathbb{R}$ is convex, locally bounded around x, then $M(\cdot)$ is lsc at x.*

Proof. Let $x_k \to x$. For $\epsilon > 0$, we can find a sequence (a_k) such that $\|(x - x_k)/a_k\| \leq \epsilon$ and $a_k \to 0$ as $k \to \infty$. If k is sufficiently large, $\|x_k - x\| \leq \epsilon$. Let ϵ is large enough so that M is bounded in the closed ball $\overline{B_{2\epsilon}(x)}$ and define

$$v_k = x_k + \frac{x - x_k}{a_k}$$

so that $v_k \in \overline{B_{2\epsilon}(x)}$. By convexity

$$M(x) \leq a_k M(v_k) + (1 - a_k)M(x_k) \leq 2a_k c + M(x_k)$$

where c is a bound for M in $\overline{B_{2\epsilon}(x)}$. Thus

$$M(x) \leq \lim_{k\to\infty} \inf(2a_k c + M(x_k)) = \lim_{k\to\infty} \inf M(x_k).$$

□

Remark 15. The above result of convexity and local boundedness implying lower semicontinuity is similar to a classical result for linear operators, where local boundedness implies continuity. In general, roughly speaking convexity in optimization plays the same role as linearity in solving equations.

Another advantageous property of convex functionals is the possibility to define a generalized gradient.

Definition. Define a continuous linear operator $D(u;\cdot) : \mathcal{U} \to \mathcal{V}$ where $\mathcal{U}$ and $\mathcal{V}$ are Banach spaces.

$$D(u;v) = \lim_{t\to 0^+} \frac{F(u+tv) - F(u)}{t} = F'(u)v$$

for $u, v \in \mathcal{U}$

Assume first that M is twice continuously Frechet-differentiable and the second derivative $M^{(2)}$ is positive definite,

$$M(w) = M(u)+M^{(1)}(u)(w-u)+\int M^{(2)}(u+t(w-u))(w-u,w-u)dt \geq M(u)+M^{(1)}(u)(w-u).$$

Definition. Let $\mathcal{X}$ be a Banach space. The subgradient ∂M at a point $x \in \mathcal{X}$ is

$$\partial M(x) := \{p \in \mathcal{X}^* | M(w) \geq M(x) + p \cdot (w-x), \forall w \in \mathcal{X}\}.$$

where $\mathcal{X}^*$ is the dual space of $\mathcal{X}$.

Note that the subgradient is now a set of elements in $\mathcal{X}^*$ instead of a single element.

The subgradient can be used to obtain a local optimality condition, which is necessary and sufficient for convex problems.

Theorem 17. *Let $\mathcal{X}$ be a Banach space and let $M : \mathcal{X} \to \mathbb{R}$ be convex. Then each local minimum is a global minimum. In addition, $x^* \in \mathcal{X}$ is a minimizer iff $0 \in \partial M(x^*)$.*

A frequently used technique for convex optimization is duality. The idea is to replace the optimization problem by an equivalent problem in the dual space $\mathcal{X}^*$ involving a dual functional which is called convex conjugate or Fenchel transform.

Definition. Let $\mathcal{X}$ be a Banach space and any $M : \mathcal{X} \to \mathbb{R}$ (not need to be convex). The convex conjugate function $M^* : \mathcal{X}^* \to \mathbb{R}$ is given by

$$M^*(p) = \sup_{x\in\mathcal{X}} (p \cdot x - M(x)).$$

Proposition 9. *Let $\mathcal{X}$ be a Banach space and any $M : \mathcal{X} \to \mathbb{R}$ (not need to be convex). Then M^* is convex.*

Proof.

$$\begin{aligned}
M^*(ap + (1-a)q) &= \sup_{x\in\mathcal{X}} (a(p\cdot x) + (1-a)(q\cdot x) - M(x)) \\
&= \sup_{x\in\mathcal{X}} (a[(p\cdot x) - M(x)] + (1-a)[(q\cdot x) - M(x)]) \\
&\leq \sup_{x,y\in\mathcal{X}} (a[(p\cdot x) - M(x)] + (1-a)[(q\cdot y) - M(y)]) \\
&= a\sup_{x\in\mathcal{X}}[(p\cdot x) - M(x)] + (1-a)\sup_{y\in\mathcal{X}}[(q\cdot y) - M(y)] \\
&= aM^*(p) + (1-a)M^*(q).
\end{aligned}$$

□

Part II

Probability and Measure Theory

Chapter 5

Elementary Concepts

5.1 Preliminaries

5.1.1 Data and Models

We assume the data is given, and concern ourselves only with how these data should be analyzed. Data X, possibly vector-valued and belong to an set Ω. Model describes the mechanism that produced this data.

Example 27. For example, if $X = (X_1, \ldots, X_n)$ is a vector consisting of the recorded heights of n students, then the model might say that these individuals were sampled completely at random from the entire population of students, and that heights of students in the population are normally distributed. In short, we would write something like $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$; here "iid" stands for independent and identically distributed.

Inference we shall say that $X \sim P_\theta$ where, for each $\theta \in \Theta$, P_θ is a probability distribution. Then inference problem can be stated as follows: Use data X to determine which population in $\{P_\theta : \theta \in \Theta\}$ was the one that produced the observed X. We shall refer to θ as the *parameter* and Θ the *parameter space*.

Example 28. In the heights example, it shall be assumed that at least one of μ and σ^2 are unknown, and we want to use the observed data X to learn about these unknown quantities. So, in some sense, the population in question is actually just a class/family of distributions—in the heights example this is the collection of all (univariate) normal distributions.

To summarize, the statistical inference problem consists of data X taking values in a sample space Θ and a family of probability distributions $\{P_\theta : \theta \in \Theta\}$. The ultimate goal is to identify the particular P_θ which produced the observed X.

5.1.2 Statistics and Probability

Statistics and probability are closely related. Probability can be used directly to describe characteristics of a sample taken from a fixed populations. The statistics problem, on the other hand, has a known sample but an unknown population.

Example 29. The general sampling model "$X \sim P_\theta$" is a probabilistic statement for known θ. For example, if $X \sim \mathcal{N}(0,1)$ then we know from introductory probability courses that $\Pr\{X \leq 1\} = \Phi(1)$, where $\Phi(\cdot)$ is the standard normal distribution function. Statistical inference is to use this information when θ is *unknown*.

5.1.3 Large Samples

The general strategy is to embed, in one way or another, the particular problem into a hypothetical sequence of infinitely many "similar" problems. In so doing, tools from probability can be introduced. For example, in frequentist statistics, one looks for procedures which perform well on average across this hypothetical sequence. In Bayesian statistics, a super-population is introduced from which the unknown θ is believed to be sampled from; this allows application of Bayes' theorem to incorporate the observed data.

We, however, will avoid such philosophical concerns in this course.

5.2 Mathematical preliminaries

5.2.1 Measure and integration

Definition. A *space* is the set with some added structure..

Definition. The *sample space* (denoted by Ω) is the set of all possible outcomes of an experiment.

Definition. An *event* (denoted by A) is a collection of possible outcomes of an experiment, that is, a subset of the sample space.

Definition. 4 If the sets $A_1, A_2, \ldots$ are pairwise disjoint and their union $\cup_i A_i$ is equal to the sample space, the collection $A_1, A_2, \ldots$ forms a *partition* of the sample space.

Example 30. . There are six outcomes in the sample space, corresponding to the number on top of the die, so we can take $\Omega = \{1,2,3,4,5,6\}$. Possible events include "an odd number", $A_1 = \{1,3,5\}$, "an even number", $A_2 = \{2,4,6\}$. A_1 and A_2 are disojint $A_1 \cap A_2 = \emptyset$ and form a partition.

Definition. A *measure* is a generalization of the concept of length, area, volume, etc. More specifically, a measure μ is a non-negative set-function, i.e., μ assigns a non-negative number to subsets A of an abstract set $\mathcal{A}$, and this number is denoted by $\mu(A)$. Similar to lengths, μ is assumed to be *additive*:

$$\mu(A \cup B) = \mu(A) + \mu(B), \quad \text{for each disjoint } A\text{and } B.$$

This extends by induction to any finite set $A_1, \ldots, A_n$ of disjoint sets. But a stronger assumption is σ-*additivity*:

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i), \quad \text{for all disjoint } A_1, A_2, \ldots.$$

Note that finite additivity does not imply σ-additivity. All of the (probability) measures we're familiar with are σ-additive. But there are some peculiar measures which are finitely additive but not σ-additive. The classical example of this is the following.

Example 31. Take $\Omega = \{1, 2, \ldots\}$ and define a measure μ as

$$\mu(A) = \begin{cases} 0 & \text{if } A\text{is finite} \\ 1 & \text{if } A\text{is the complement of a finite set,} \end{cases}$$

It is easy to see that μ is additive. Taking a disjoint sequence $A_i = \{i\}$ we find that $\mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\Omega) = 1$ but $\sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} 0 = 0$. Therefore, μ is not σ-additive.

In general, a measure μ cannot be defined for all subsets $A \subseteq \Omega$. But the class of subsets on which the measure can be defined is, in general, a *σ-algebra*, or *σ-field*.

Definition. An *algebra* is the set with some added algebraic structures.

Definition. A σ-algebra $\mathcal{A}$ is a collection of subsets of Ω that satisfies the following properties:
(a) Ω is in $\mathcal{A}$;
(b) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$;
(c) and if $A_1, A_2, \ldots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Definition. The sets $A \in \mathcal{A}$ are said to be *measurable*. We refer to $(\Omega, \mathcal{A})$ as a measurable space and $(\Omega, \mathcal{A}, \mu)$ as a measure space.

Definition. A measure μ is *finite* if $\mu(\Omega)$ is a finite number. Probability measures are special finite measures where $\mu(\Omega) = 1$. A measure μ is said to be σ-finite if there exists a sequence of sets $\{A_i\} \subset \mathcal{A}$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $\mu(A_i) < \infty$ for each i.

Example 32. Let Ω be a countable set (e.g. a set of countable points) and $\mathcal{A}$ the class of *all* subsets of Ω; then clearly $\mathcal{A}$ is a σ-algebra. Define μ according to the rule

$$\mu(A) = \text{number of points in } A, \quad A \in \mathcal{A}.$$

Then μ is a σ-finite measure which is refered to as *counting measure*.

Example 33. Let Ω be a subset of d-dimensional Euclidean space $\mathbb{R}^d$. Take $\mathcal{A}$ to be the smallest σ-algebra that contains the collection of open rectangles

$$A = \{(x_1, \ldots, x_d) : a_i < x_i < b_i, i = 1, \ldots, d, \quad a_i < b_i\}.$$

Then $\mathcal{A}$ is the Borel σ-algebra on Ω, which contains all open and closed sets in Ω; but there are subsets of Ω that do not belong to $\mathcal{A}$! Then the (unique) measure μ, defined by

$$\mu(A) = \prod_{i=1}^{d} (b_i - a_i), \quad \text{for rectangles } A \in \mathcal{A}$$

is called *Lebesgue measure*, and it's σ-finite.

Our last result has something to do with constructing new measures from old. It also allows us to generalize the familiar notion of probability densities which, in turn, will make our lives easier when discussing the general statistical inference problem. Suppose f is a non-negative integrable function. Then

$$\nu(A) = \int_A f \, d\mu \tag{5.1}$$

defines a new measure ν on $(\Omega, \mathcal{A})$. An important property is that $\mu(A) = 0$ implies $\nu(A) = 0$; the terminology is that ν is *absolutely continuous with respect to* μ, or ν is *dominated* by μ, written $\nu \ll \mu$. But it turns out that, if $\nu \ll \mu$, then there exists f such that (5.1) holds. This is the famous Radon–Nikodym theorem.

Theorem 18. *(Radon–Nikodym) Suppose $\nu \ll \mu$. Then there exists a non-negative μ-integrable function, such that (5.1) holds. The function f, often written as $f = d\nu/d\mu$ is the Radon–Nikodym derivative of ν with respect to μ.*

The function f satisfying the above equality is uniquely defined up to a μ-null set, that is, if g is another function which satisfies the same property, then $f = g$ almost everywhere on $(\Omega, \mathcal{A}, \mu)$.

We'll see later that, in statistical problems, the Radon–Nikodym derivative is the familiar density or, perhaps, a likelihood ratio. The Radon–Nikodym theorem also formalizes the idea of change-of-variables in integration. For example, suppose that μ and ν are σ-finite measures defined on Ω, such that $\nu \ll \mu$, so that there exists a unique Radon–Nikodym derivative $f = d\nu/d\mu$. Then, for a ν-integrable function φ, we have

$$\int \varphi \, d\nu = \int \varphi \, f \, d\mu;$$

symbolically this makes sense: $d\nu = (d\nu/d\mu)\, d\mu$. The probability density function of a random variable is the Radon–Nikodym derivative of the induced measure with respect to some base measure (usually the Lebesgue measure for continuous random variables).

(A rigorous proof of the theorem is beyond current scope. You could refer to the following note for details. Note that the proof uses sign measure which generalizes standard measure function.

The Hahn-Banach and Radon-Nikodym theorem
http://math.bu.edu/people/mkon/MA779/RadonNykodim.pdf)

5.2.2 Probability basics

Mathematical probability is just a special case of the measure theory stuff presented above. Probabilities are finite measures, random variables are the measurable functions, expected values are just integrals.

Start with an essentially arbitrary measurable space $(\Omega, \mathcal{A})$, and introduce a probability measure P; that is $P(\Omega) = 1$. Then $(\Omega, \mathcal{A}, P)$ is called a *probability space*. The idea is that Ω contains all possible outcomes of the random experiment.

Remark. $\sigma([0,1])$ is characterized as the minimal σ-field generated by: (a) the open intervals (a,b) on $[0,1]$; (b) the closed intervals $[a,b]$; (iii) the closed half-lines $[a,b)$, and so on. It is also the minimal σ-field containing all the open sets in $[0,1]$. (It can be generalized to any metric space).

Definition. (Kolmogorov Axioms) Given measurable space $(\Omega, \mathcal{A})$, a probability function is a function P from $\mathcal{A}$ to $\mathbb{R}$ satisfying:

(a) For all $A \in \mathcal{A}$, $P(A) \geq 0$,

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

(c) If $A_1, \ldots,$ are pairwise disjoint, then $P(\cup_i A_i) = \sum_i P(A_i)$.

Example 34. An immediate implication of the Kolmogorov axioms is that (a) $P(A^c) = 1 - P(A)$. Because $1 = P(\Omega) = P(A) + P(A^c)$. And (b) $P(\emptyset) = 0$ and (c) $P(A) \leq 1$ for any $A \in \mathcal{A}$. The other useful results are for any A_1 and A_2 in $\mathcal{A}$, we have

$$\begin{aligned} &(d) \qquad P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2); \\ &(e) \qquad P(A_1 \cap A_2^c) = P(A_1) - P(A_1 \cap A_2); \\ &(f) \qquad P(A) \leq P(B) \text{ if } A \subset B. \end{aligned}$$

Example 35. Based on above results, we can have the following:

(a) $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, \ldots$.

(b) $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any events $A_1, A_2, \ldots$

The following theorem establishes the connection between limits of probabilities and sequences of events.

Theorem 19. *For a given probability space* $(\Omega, \mathcal{A}, P)$

(a) If $\{A_1, A_2, \ldots\}$ *is an increasing sequence of events, i.e.* $A_1 \subseteq A_2 \subseteq \ldots$, *then* $\lim_{n\to\infty} P(A_n) = P(\cup_{i=1}^{\infty} A_i)$.

(b) If $\{A_1, A_2, \ldots\}$ *is a decreasing sequence of events, i.e.* $A_1 \supseteq A_2 \supseteq \ldots$, *then* $\lim_{n\to\infty} P(A_n) = P(\cap_{i=1}^{\infty} A_i)$.

Example 36. $([0,1], \sigma([0,1]), \mu)$. The sample space is the real interval $[0,1]$. $\sigma([0,1])$ is Borel σ-algebra on $[0,1]$. This is the minimal σ-algebra generated by the elementary events, e.g. $\{[0,b), 0 \leq b \leq 1\}$:

$$\begin{aligned} \lim_{n\to\infty} \mu([0,1/n)) &= \mu(\cap_{n=1}^{\infty}[0,1/n)) = \mu(\{0\}) = 0 \\ \lim_{n\to\infty} \mu((0,1/n)) &= \mu(\cap_{n=1}^{\infty}(0,1/n)) = \mu(\emptyset) = 0 \\ \lim_{n\to\infty} \mu([a-1/n, b+1/n]) &= \mu(\cap_{n=1}^{\infty}[a-1/n, b+1/n]) = \mu([a,b]) = b-a \end{aligned}$$

This collection contains things like $[1/4, 2/5]$, $\{1/5\}$ etc. $\mu(\cdot)$ for all $A \in \sigma([0,1])$ is Lebesgue measure, defined as the sum of the lengths of the intervals onctained in A.

Definition. A random variable X is a measurable function that assigns one and only one numerical value to each outcome of an experiment such that $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}$.

Example 37. Toss a coin three times. The sample space is

$$\Omega = \{\text{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}\}.$$

Define a random variable X by taking the number of heads in the three tosses. Thus the random variable can take on values $0, 1, 2, 3$. So $X(\{HHT\}) = 2$.

Remark. It's important to understand that X, as a mapping, is not random; instead, X is a function of a randomly chosen element ω in Ω. So when we are discussing probabilities that X satisfies such and such properties, we're actually thinking about the probability (or measure) on the set of ω's for which $X(\omega)$ satisfies the particular property. To make this more precise we write

$$P(X \in A) = P(\{\omega : X(\omega) \in A\}) = PX^{-1}(A).$$

To simplify notation, etc, we will often ignore the underlying probability space, and work simply with the probability measure $P_X(\cdot)$. This is what we're familiar with from basic probability and statistics. Namely, $P(X \in A)$ is a probability function on $\mathcal{X}$, defined in terms of the probability function on Ω.

Example 38. In treatment experiments, the effect is either "effective" or "non-effective". Here we have $\Omega = \{\omega_1, \omega_2\}$, where ω_1 is the effective outcome with probability $P(\omega_1) = q$ and ω_2 is non-effective outcome with $P(\omega_2) = 1 - q$ probability. It is natural to define a random variable X as $X(\omega_1) = X_1 = 1$ and $X(\omega_2) = X_2 = 0$. Here $\mathcal{X} = \{0, 1\}$.

Example 39. The statement $X \sim \mathcal{N}(0, 1)$ means that the probability measure induced on $\mathbb{R}$ by the mapping X is a standard normal distribution.

Remark. If $\Omega = \{\omega_1, \omega_2, \dots\}$ is a countable set, then we will observe $X = x_i$ iff the outcome of the random experiment is an $\omega_i \in \Omega$ such that $X(\omega_i) = x_i$. Note that random variables that take a countable number of values are called discrete. Random variables that take values from an interval of real numbers are called continuous.

Example 40. For a continuous random variable $X : \Omega \to \mathbb{R}$. We can define the following probability space:

(a) Sample space is real line $\mathbb{R}$.
(b) Event space is $\sigma(\mathbb{R})$, the Borel σ-algebra on the real line.
(c) Probability P_X such that for $A \in \sigma(\mathbb{R})$

$$P_X(A) = P_\omega(\omega \in \Omega : X(\omega) \in A) = P_\omega(X^{-1}(A)).$$

A consequence of this construction is for all $A \in \sigma(\mathbb{R})$, $X^{-1}(A) \in \sigma(\mathbb{R})$. Otherwise, $P_\omega(X^{-1}(A))$ may not be well-defined, since the domain of the $P(\cdot)$ function is $\sigma(\mathbb{R})$. This is the requirement that the random variable $X(\cdot)$ is Borel-measurable.

Example 41. $X(\omega) = |\omega|$ with ω from the probability space $([-1,1], \sigma([-1,1]), \mu/2)$ where μ is the Lebesgue measure. Then

$$P_x\left([\frac{1}{3}, \frac{2}{3}]\right) = \mu\left([\frac{1}{3}, \frac{2}{3}]\right)/2 + \mu\left([-\frac{1}{3}, -\frac{2}{3}]\right)/2 = \mu\left([\frac{1}{3}, \frac{2}{3}]\right).$$

When there is no possibility of confusion, we will drop the "X" subscript and simply write P for P_X.

Definition. The cumulative distribution function (CDF) of a random variable X is defined as $F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$. It is often wirtten as $P(X \leq x)$.

Proposition 10. *The function $F_X(x)$ is a CDF iff*
(a) $F(x)$ is non-decreasing.
(b) $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.
(c) $F(x)$ is right-continuous. (such that for any $\delta \geq 0$, $\lim_{\delta\to 0} F_X(x+\delta) = F_X(x)$).

Proof. Note that a CDF can be equivalently written as

$$F_X(x) = \Pr(\{\omega \in \Omega : X(\omega) \leq x\}) = \Pr(A_x)$$

where $A_x = \{\omega \in \Omega : X(\omega) \leq x\}$.

(a) For any $x_i < x_j$ we have $A_{x_i} \subseteq A_{x_j}$, thus $P(A_{x_i}) \leq P(A_{x_j})$ and thus $F_X(x_i) \leq F_X(x_j)$.

(b) Define a decreasing sequence x_n such that $x_n \to -\infty$ as $n \to \infty$. Then for $x_n \geq x_{n+1}$ we have $A_{x_n} \supseteq A_{x_{n+1}}$ and

$$\cap_{n=1}^{\infty} A_{x_n} = \emptyset.$$

Hence, by Theorem 2 (b), we have

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P(A_{x_n}) = P(\cap_{n=1}^{\infty} A_{x_n}) = 0.$$

Since $x_n \to -\infty$, we have the result. Similarly argument holds for $\lim_{x\to\infty} F_X(x) = 1$.

(c) Homework. □

Definition. A random variable X is discrete if $F_X(x)$ is a step function of x. A random variable X is continuous if $F_X(x)$ is a continuous function of x.

Definition. The random variable X and Y are identically distributed if for every set $A \in \mathcal{A}$, $P_X(X \in A) = P_Y(Y \in A)$.

Proposition 11. *X and Y are identically distributed iff $F_X(z) = F_Y(z)$ for every z.*

5.3 Exercises

1. If $\mathcal{A}$ is σ-algebra, show that for any $A_i \in \mathcal{A}$,

$$\cap_{i=1}^{\infty} A_i \in \mathcal{A}.$$

[Answer: Properties of (b) and (c) together with DeMorgan's laws $(A_1 \cap A_2) = (A_1^c \cup A_2^c)^c$ assure that $\cap_{i=1}^{\infty} A_i \in \mathcal{A}$.]

2. Show the following result holds for any A_1 and A_2 in a probability space $(\Omega, \mathcal{A}, P)$:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

by using the implication of the Kolmogorov axioms. [Hint: The proof relies on creating pairwise disjoint sets for which one can add up the probabities by the third axiom]

3. If P is a probability function, then the following inequality holds

$$P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

for any events $A_1, A_2 \ldots$. [Answer: http://en.wikipedia.org/wiki/Boole's_inequality]

4. Prove proposition 1 (c). [Answer: Hint: Define a decreasing sequence x_n such that $x_n \to x$ as $n \to \infty$ and $x_1 > x_2 \cdots > x_n$. So $A_x \subseteq A_{x_n}$ for all n. Note that

$$\lim_{n\to\infty} P(A_{x_n}) = P(\cap_{i=1}^{\infty} A_{x_i}).$$

By above argument,

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P(A_{x_n}) = P(\cap_{i=1}^{\infty} A_{x_i}) = P(A_x) = F_X(x).$$

]

Chapter 6

Lebesgue Measure and its Convergence

6.1 Construction of Lebesgue Measures

Remember that a measure μ on a set $\mathcal{X}$ associates to a subset $A \subset \mathcal{X}$ a nonnegative number $\mu(A)$, called the measure of A. A set is a measurable set if it has a well-defined measure. We require that the measurable sets form a σ-algebra, meaning that complements, countable unions, and countable intersections of measurable sets are measurable.

We have not explained why Lebesgue measure should exist at all. Now it is time to do that. Before showing the construction, we look at an idea about negligibility. The idea of a 'negligible' set relates to one of the limitations of the Riemann integral, as we saw in the previous lecture (lecture on continuity). Since the function $f = \mathbf{1}_{\mathbb{Q}}$ takes a non-zero value only on $\mathbb{Q}$, then "area under its graph" must be very closely linked to the 'length' of the set $\mathbb{Q}$. Because the sets $\mathbb{R}$ and $\mathbb{R}\backslash\mathbb{Q}$ are so different from intervals, we cannot integrate f in the Riemann sense. Then how should we define this concept for more general sets?

A finite set is not an interval but since a single point has length 0, adding finitely many such lengths together should still give 0. The underlying concept here is that if we decompose a set into a finite number of disjoint intervals, we compute the length of this set by adding the lengths of the pieces. The following is more general definition of sets of 'zero length'

Definition. A null set $A \subset \mathbb{R}$ is a set that may be covered by a sequence of intervals of arbitrarly small total length. That is, given $\epsilon > 0$, exists $\{I_n\}$ such that $A \subseteq \cup_{n=1}^{\infty} I_n$ and $\sum_{n=1}^{\infty} l(I_n) < \epsilon$ where $l(I)$ is the length of set I.

Example 42. Any one-element set (singleton) is a null set. Any countable set $A = \{x_1, x_2, \dots\}$ is null. A countable union of null sets is also null.

Example 43. A uncountable sets can be null but not always, i.e. Cantor set (Start with an interval $[0, 1]$, then remove the middle $(1/3, 2/3)$, and then do the removals for $[0, 1/3]$ and $[2/3, 1]$ and do these sequentially. At nth stage you will have a set C_n consisting of 2^n disjoint closed intervals, each of length $1/3^n$. Thus the total length of C_n is $(2/3)^n$ which converges to 0 as $n \to \infty$.

The simple concept of null sets provides the key to our idea of length, since it tells us what we can ignore. A general notion of length is given by Lebesgue outer measure.

Definition. The Lebesgue outer measure of a set A is

$$m^*(A) = \inf\left\{\sum_{n=1}^{\infty} l(I_n) : \ I_n \text{ are intervals, } A \subseteq \cup_{n=1}^{\infty} I_n\right\}$$

where $\{I_n\}$ cover the set A.

The outer measure is the infimum of lengths of all possible covers of A. Note that $m^*(A) \geq 0$ for any $A \subseteq \mathbb{R}$.

Example 44. $A \subseteq \mathbb{R}$ is a null set iff $m^*(A) = 0$.

Proposition 12. *(i) m^* is monotone: the bigger the set, the greater its outer measure. (ii) The outer measure of an interval equals its length. (iii) For any sequence of sets $\{A_n\}$, outer measure satisfies*

$$m^*\left(\cup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} m^*(A_n)$$

which is countable sub-additive.

Yes, you are right. Now we just need another inequality to bound the outer measure so that we have the additivity property of the Lebesgue measure.

Definition. (Continuity property for sets) If pairwise disjoint sets $\{A_n\}$ have union A, then the lengths of the set $B_n = A \backslash \cup_{k=1}^{n} A_n$ is expected to decrease to 0 as $n \to \infty$, or say B_n tends to be a null set as $n \to \infty$.

With outer measure, sub-additivity, we wish to ensure that if sets A_n are pairwise disjoint, then the inequality of sub-additivity becomes an equality. However, it turns out that this will not in general be true for outer measure. Then how about decompose a set into finitely many disjoint pieces? The answer is: with continuity property and finite additivity, one expect that length of a set should be countably additive.

Definition. A set $A \subseteq \mathbb{R}$ is Lebesgue measurable if for every set $B \subseteq \mathbb{R}$ we have

$$m^*(B) = m^*(B \cap A) + m^*(B \cap A^c)$$

where $A^c = \mathbb{R} \backslash A$

Obviously, $B = (B \cap A) \cup (B \cap A^c)$ by the sub-additivity we have

$$m^*(B) \leq m^*(B \cap A) + m^*(B \cap A^c).$$

So the rest task is to verify

$$m^*(B) \geq m^*(B \cap A) + m^*(B \cap A^c).$$

This is your homework (2).

Example 45. We summarize the properties of the family of all Lebesgue measurable sets $\mathcal{M}$ as follows: $\mathcal{M}$ is closed under countable unions, countable intersections, and complements. It contains intervals and all null sets.

Definition. For any A in the family of all Lebesgue measurable sets $\mathcal{M}$, we write $\mu(A)$ and call $\mu(A)$ the Lebesgue measure of the set A. Lebesgue measure $\mu : \mathcal{M} \to [0.\infty]$ is a countably additive set function defined on the σ-field $\mathcal{M}$ of measurable sets. Lebesgue measure of an interval is equal to its length. Lebesgue measure of a null set is zero.

The countable additivity of μ and on the definition of the sum of a series in $[0, \infty]$ allows that

$$\sum_{i=1}^{\infty} \mu(A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(A_i).$$

Proposition 13. *μ is continuous at $\emptyset$, i.e. if (B_n) decrease to$\emptyset$, then $\mu(B_n)$ decrease to 0.*

Proof. Let $A_n = B_n \backslash B_{n+1}$ define a disjoint sequence in $\mathcal{M}$. Note that $\cup_n A_n = B_1$. Then

$$\mu(A_n) = \mu(B_n) - \mu(B_{n+1}) \geq 0$$

and hence

$$\mu(B_1) = \sum_{n=1}^{\infty} \mu(A_n) = \lim_{k\to\infty} \sum_{n=1}^{k} [\mu(B_n) - \mu(B_{n+1})] = \mu(B_1) - \lim_{n\to\infty} \mu(B_n)$$

which implies that $\mu(B_n) \to 0$. □

Theorem 20. *Suppose that $\mathcal{A}$ is the σ-algebra on $\mathcal{X}$ generated by the collection of sets $\mathcal{F}$. Let μ and υ be two measures on $\mathcal{A}$ such taht $\mu(B) = \upsilon(B)$ for every $B \in \mathcal{F}$. If there is a countable family of sets $\{B_i\} \subset \mathcal{F}$ such that $\cup_i B_i = \mathcal{X}$ and $\mu(B_i) < \infty$, then $\mu = \upsilon$.*

A more general result about Lebesgue measurable in $\mathbb{R}^n$ is as follows. We just state it with any proof. But you have the hints from the case in $\mathbb{R}$.

Theorem 21. *A set $A \subseteq \mathbb{R}^n$ is Lebesgue measurable iff for every $\epsilon > 0$, there is a closet set F and an open set G such that $F \subset A \subset G$ and $\mu(G\backslash F) < \epsilon$. Moreover*

$$\begin{aligned}\mu(A) =& \inf\{\mu(U) : U \text{ is open and } A \subset U\} \\ =& \inf\{\mu(K) : K \text{ is compact and } K \subset A\}.\end{aligned}$$

Thus a Lebesgue measurable set may be approximated from the outside by open sets, and from the inside by compact sets.

Example 46. (Geometric properties: Translation invariant) For every measurable set A, $\mu(\tau(A)) = \mu(A)$ where

$$\tau(A) = \{y \in \mathbb{R}^n : y = x + \tau \text{ for some } x \in A\}.$$

6.2 Measurable Functions and Convergences

Measurable functions are the natural mapping between measurable spaces.

Definition. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A measurable function is a mapping $f : X \to Y$ such that $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$.

The measurability of $f : X \to Y$ depends only on the σ-algebras on X and Y and not on what measure is defined on X or Y.

Definition. Two measurable functions $f : X \to Y$ and $g : X \to Y$ are equal almost everywhere (a.e. in short) means $\mu(\{x \in X | f(x) \neq g(x)\}) = 0$.

Example 47. Every continuous function between topological spaces is Borel measurable. A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is measurable w.r.t. Lebesgue σ-algebra on the domain $\mathbb{R}^n$ and the Borel σ-algebra on the range $\mathbb{R}$.

Proposition 14. *Let $(X, \mathcal{A})$ be a measurable space. A function $f : X \to \mathbb{R}$ is measurable iff the set $\{x \in X |\, f(x) < c\}$ belongs to $\mathcal{A}$ for every $c \in \mathbb{R}$.*

Definition. A sequence of functions (f_n) from a measure space $(X, \mathcal{A}, \mu)$ to $\mathbb{R}$ converges pointwise a.e. to a function $f : X \to \mathbb{R}$ if $\lim_{n\to\infty} f_n(x) = f(x)$ for every $x \in X \backslash N$ where N is the null set.

Definition. a complete measure space

Theorem 22. *If (f_n) is a sequence of measurable functions that converges pointwise to f, then f is measurable. If $(X, \mathcal{A}, \mu)$ is a complete measure space and (f_n)converges pointwise a.e. to f, then f is measurable.*

Definition. A function $\varphi : X \to \mathbb{R}$ on a measurable space $(X, \mathcal{A})$ is a simple function if there are meausrable sets $A_1, \ldots, A_n$ and real numbers $c_1, \ldots, c_n$ such that $\varphi = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$.

The representation of a simple function as a sum of indicator functions is not unique. A standard representation uses disjoint sets A_i and distinct values c_i.

Theorem 23. *Let function $f : X \to [0, \infty]$ be non-negative and measurable. There is a monotone increasing sequence $\{\varphi_n\}$ of simple functions that converges pointwise to f.*

Proof. For $n \in \mathbb{N}$, we sub-divide the range of f into $2^{2n} + 1$ intervals such that

$$I_{n,k} = \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)$$

for $k = 1, 2, \ldots, 2^{2n}$, $I_{n,2^{2n}+1} = [2^n, \infty]$ of length 2^{-n}. The measurable sets are $A_{n,k} = f^{-1}(I_{n,k})$ for $k = 1, 2, \ldots, 2^{2n} + 1$. Then the increasing sequence of simple functions are

$$\varphi_n = \sum_{k=1}^{2^{2n}+1} (\frac{k-1}{2^n}) \mathbf{1}_{A_{n,k}}$$

which converges pointwise to f as $n \to \infty$. □

The result can be easily extended to $f : X \to \mathbb{R}$. Just re-write the function as two non-negative measurable functions $f = f_+ - f_-$ where $f_+ = \max\{f, 0\}$ and $f_- = \max\{-f, 0\}$. Then approximate each part by simple functions.

Lebesgue integral provides an extension of the Riemann integral which applies to highly discontinuous and unbounded functions, and which behaves very well with respect to limiting operations. To construct the Lebesgue integral, we first define integral of a simple function.

Definition. If $\varphi = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$ is a simple function on a measure space $(X, \mathcal{A}, \mu)$. The integral of φ w.r.t. μ is $\int \varphi d\mu = \sum_{i=1}^n c_i \mu(A_i)$.

The value of the sum on the RHS is independent of how φ is represented. Because of the way the approximating simple functions are constructed, the Lebesgue approach to integration is sometimes contrasted with the Riemann approach since it sub-divides the range of the function instead of the domain.

Definition. Let $f : X \to [0, \infty]$ be a non-negative measurable function on a measure space $(X, \mathcal{A}, \mu)$. Define

$$\int f d\mu = \sup \left\{ \int \varphi d\mu : \varphi \text{ is simple and } \varphi \leq f \right\}.$$

If $f : X \to \mathbb{R}$ and $f = f_+ - f_-$, then define

$$\int f d\mu = \int f_+ d\mu - \int f_- d\mu,$$

provided that at least one of the integrals on RHS is finite. If A is a measurable subset of X, $\int_A f d\mu = \int f \mathbf{1}_A d\mu$.

Lebesgue integral does not assign a value to the integral of a highly osciallatory function f which have infinities for $\int f_+ d\mu$ and $\int f_- d\mu$.

6.3 Exercises

1. For any sequence of sets $\{A_n\}$, prove the following statement

$$m^* (A_1 \cup A_2) \leq m^*(A_1) + m^*(A_2).$$

(Answer: Take $\epsilon = 1/n$. Find covering sequence (I_k^1) of A_1 and (I_k^2) of A_2 such that

$$\sum_{k=1}^{\infty} l(I_k^j) \leq m^*(E_1) + \frac{\epsilon}{2},$$

where $j = 1, 2$. Then adding up

$$\sum_{k=1}^{\infty} l(I_k^1) + \sum_{k=1}^{\infty} l(I_k^2) \leq m^*(E_1) + m^*(E_1) + \epsilon.$$

Because the sequence $(I_1^1, I_1^2, I_2^1, I_2^2, \dots)$ covers $A_1 \cup A_2$, we have

$$m^*(A_1 \cup A_2) \leq \sum_{k=1}^{\infty} l(I_k^1) + \sum_{k=1}^{\infty} l(I_k^2).$$

Then let $n \to \infty$, the result is desired.)

2. Suppose $A_1 \cap A_2 = \emptyset$, A_1 and A_2 are both Lebesgue measurable and in $\mathbb{R}$. Show that

$$m^*(A_1 \cup A_2) = m^*(A_1) + m^*(A_2).$$

[Answer: Let $B \subset \mathbb{R}$. By definition of measurabilities of A_1 and A_2, we have

$$m^*(B) = m^*(B \cap A_i) + m^*(B \cap A_i^c) \qquad (*)$$

for $i = 1, 2$. Replace A_2 with $B \cap A_1^c$

$$\begin{aligned} m^*(B \cap A_1^c) =& m^*((B \cap A_1^c) \cap A_2) + m^*((B \cap A_1^c) \cap A_2^c) \\ =& m^*(B \cap (A_1^c \cap A_2)) + m^*(B \cap (A_1^c \cap A_2^c)) \end{aligned}$$

Since A_1 and A_2 are disjoint, $A_1^c \cap A_2 = A_2$. By de Morgan's law $A_1^c \cap A_2^c = (A_1 \cup A_2)^c$.

$$m^*(B \cap A_1^c) = m^*(B \cap A_2) + m^*(B \cap (A_1 \cup A_2)^c).$$

Substitue this into the (*) equation,

$$m^*(B) = m^*(B \cap A_1) + m^*(B \cap A_2) + m^*(B \cap (A_1 \cup A_2)^c).$$

Now by the sub-additivity of m^*, we have

$$m^*(B \cap A_1) + m^*(B \cap A_2) \geq m^*((B \cap A_1) \cup (B \cap A_2)) = m^*(B \cap (A_1 \cup A_2)).$$

Then

$$m^*(B) \geq m^*(B \cap (A_1 \cup A_2)) + m^*(B \cap (A_1 \cup A_2)^c).$$

The inverse inequality is always ture. The result is desired.]

Chapter 7

Integration and Expectation

7.1 Integration and Expectation

Definition. Suppose that X is a random variable with PDF or PMF $p_X(x)$. Let $\mathcal{X} = \{x : p_X(x) > 0\}$. This is called the support set or support of the distribution of X, and intuitively is the set of values that the random variable X can take on.

Remark 16. When X is a continuous random variable, its PDF is not unique. A more precise definition is that X should be the set of all points x such that every open neighborhood of x has positive probability. A consequence of this definition is that supports are closed sets.

When P_X, a measure on the space $\mathbb{X}$, is dominated by a σ-finite measure μ, the Radon-Nikodym theorem says there is a density $dP_X/d\mu = p_X$, and

$$P_X(A) = \int_A p_X \, d\mu.$$

Remark. When μ is counting measure, p_X is a probability mass function (PMF)

$$p_X(x) = P(X = x)$$

So, for any set A, the probability

$$\Pr(\{\omega \in \Omega : X(\omega) \in A\}) = P(X \in A) = \sum_{x \in A} p_X(x).$$

When μ is Lebesgue measure, p_X is a probability density function (PDF)

$$\Pr(\{\omega \in \Omega : X(\omega) \in A\}) = P(X \in A) = \int_A p_X(x) dx.$$

One of the benefits of the measure-theoretic formulation is that we do not have to handle these two important cases separately. Note that the PDF is not unique. Because we only care about integrals over the PDF, we can change its value at a countable number of points without changing any of the associated probabilities, and thus without changing the distribution of the random variable.

Theorem 24. *(Radon-Nikodym Theorem*) A necessary and sufficient condition for the existence of PDF p_X is that the probability measure P_X of a real-valued random variable X be absolutely continuous with respect to Lebesgue measure.*

Remark. Absolutely continuous w.r.t. Lebesgue measure means that all sets in the support of X (which is a part of the real line) which have zero Lebesgue measure must also have zero probability under P_X. Namely, for all $A \in \mathbb{R}$ such that $\mu(A) = 0$ implies $P_X(A) = 0$. Since only singletons (and countable sets of singletons) have zero Lebesgue measure, this condition essentially rules out random variables which have a "point mass" at some points.

Example 48. $([0,1], \sigma([0,1]), \mu)$ and random variable

$$X(\omega) = \begin{cases} \frac{1}{2} & \text{if } \frac{1}{4} \leq \omega \leq \frac{1}{2} \\ \omega & \text{otherwise} \end{cases}$$

Now $\mu(1/2) = 0$, but $\Pr(X = 1/2) = \Pr(\omega \in [1/4, 1/2]) = 1/4$. So P_X is not absolutely continuous w.r.t. Lebesgue measure, and thus it has no density function.

Definition. Let φ be a real-valued measurable function defined on $\mathbb{X}$. Then the expected value of $\varphi(X)$ is

$$\mathbb{E}_X\{\varphi(X)\} = \int_{\mathbb{X}} \varphi(x)\, dP_X(x) = \int_{\mathbb{X}} \varphi(x) p_X(x)\, d\mu(x),$$

the latter expression holding only when $P_X \ll \mu$ for a σ-finite measure μ on $\mathbb{X}$.

The usual properties of expected value hold in this more general case; the same tools we use in measure theory to study properties of integrals of measurable functions are useful for deriving such things.

Definition. The mean is another name for the expected value of X, usually is denoted as $\mu = \mathbb{E}[X]$. (Don't be confused with the μ we used for Lebesgue measure.) The k-th moment of X is $\mu_k = \mathbb{E}[X^k]$. The variance of X is $\mathrm{Var}(X) = \sigma^2 = \mathbb{E}[(X-\mu)^2]$. Or $\sigma^2 = \mu_2 - \mu_1^2$. The expectation of a linear function of a random variable is the linear function of the expectation:

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X].$$

The variance of a linear function of a random variable is the variance of the random variable multiplied by the square of the slope coefficient:

$$\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X).$$

If $g_1(x) \geq g_2(x)$ for all x, then $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$.

Remark. Expectation can be treated as one of the best predictors. Suppose we want to choose a single value b as a prediction for the random outcome X. We might measure the quality of the prediction by $(X-b)^2$. This is the squared prediction error, and presumably we want this to be as small as possible. We might try to choose b to minimize the expectation $\mathbb{E}[(X-b)]^2$. Note that

$$\begin{aligned}\mathbb{E}[(X-b)]^2 =& \mathbb{E}[(X-\mu)^2 + (\mu-b)^2 + 2(X-\mu)(\mu-b)] \\ =& \mathbb{E}[(X-\mu)^2] + (\mu-b)^2 + 2(\mu-b)\mathbb{E}[X-\mu].\end{aligned}$$

Note however that $\mathbb{E}[X-\mu] = 0$ by definition. So the first term is equal to the variance, and is the same regardless of the choice of b. The second term is clearly minimized by setting $b = \mu$.

In probability and statistics, product spaces are especially important. The reason, as we eluded to before, is that independence of random variables is connected with product spaces and, in particular, product measures. If $X_1, \ldots, X_n$ are iid P_X, then their joint distribution is the product measure

$$P_{X_1} \times P_{X_2} \times \cdots \times P_{X_n} = P_X \times P_X \cdots \times P_X = P_X^n.$$

The first term holds with only "independence;" the second requires "identically distributed;" the last term is just a short-hand notation for the middle term.

When we talk about convergence theorems, such as the law of large numbers, we say something like: for an infinite sequence of random variables $X_1, X_2, \ldots$ some event happens with probability 1. But what is the measure being referenced here? In the iid case, it turns out that it's an *infinite product measure*, written as P_X^∞. We'll have more to say about this when the time comes.

7.2 Conditional Distributions

Conditional distributions in general are rather abstract. When the random variables in question are discrete (μ = counting measure), however, things are quite simple; the reason is that events where the value of the random variable is fixed have positive probability, so the ordinary conditional probability formula involving ratios can be applied.

When one or more of the random variables in question are continuous (dominated by Lebesgue measure), then more care must be taken.

Definition. Suppose random variables X and Y have a joint distribution with density function $p_{X,Y}(x,y)$, with respect to some dominating (product) measure $\mu \times \nu$. Then the corresponding marginal distributions have densities with respect to μ and ν, respectively, given by

$$p_X(x) = \int p_{X,Y}(x,y)\, d\nu(y) \quad \text{and} \quad p_Y(y) = \int p_{X,Y}(x,y)\, d\mu(x).$$

Moreover, the conditional distribution of Y, given $X = x$, also has a density with respect to ν, and is given by the ratio

$$p_{Y|X}(y \mid x) = p_{X,Y}(x, y)/p_X(x).$$

As a function of x, for given y, this is clearly μ-measurable since the joint and marginal densities are measurable. Also, for a given x, $p_{Y|X}(y \mid x)$ defines a probability measure Q_x, called the *conditional distribution of Y, given $X = x$*, through the integral $Q_x(B) = \int_B p_{Y|X}(y \mid x)\, d\nu(y)$; that is, $p_{Y|X}(y \mid x)$ is the Radon–Nikodym derivative for the conditional distribution Q_x.

For us, conditional distribution can always be defined through its conditional density though, in general, a conditional density may not exist even if the conditional distribution Q_x does exist. There are real cases where the most general definition of conditional distribution is required. Also, I should mention that conditional distributions are not unique; but, we shall not dwell on this point here.

Definition. Given conditional distribution with density $p_{Y|X}(y \mid x)$, define conditional probabilities:

$$P(Y \in B \mid X = x) = \int_B p_{Y|X}(y \mid x)\, d\nu(y).$$

The law of total probability then allows us to write

$$P(Y \in B) = \int P(Y \in B \mid X = x) p_X(x)\, d\mu(x),$$

in other words, marginal probabilities for Y may be obtained by taking expectation of the conditional probabilities.

Definition. For any ν-integrable function φ, we may write the conditional expectation

$$\mathbb{E}\{\varphi(Y) \mid X = x\} = \int \varphi(y) p_{Y|X}(y|x)\, d\nu(y).$$

We may evaluate the above expectation for any x, so we actually have defined a (μ-measurable) function, say, $g(x) = \mathbb{E}(Y \mid X = x)$; here I took $\varphi(y) = y$ for simplicity. Now, $g(X)$ is a random variable, to be denoted by $\mathbb{E}(Y \mid X)$, and we can ask about its mean, variance, etc. The corresponding versions of the law of total probability for conditional expectations are

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}\{\mathbb{E}(Y \mid X)\}, \\ \mathrm{Var}(Y) &= \mathrm{Var}\{\mathbb{E}(Y \mid X)\} + \mathbb{E}\{\mathrm{Var}(Y \mid X)\}, \end{aligned}$$

where $\mathrm{Var}(Y \mid X)$ is the conditional variance, i.e., the variance of Y relative to its conditional distribution. The first formula above is called a law of iterated expectation.

7.3 Some Inequalities

Theorem 25. *(Cauchy–Schwarz inequality) If f^2 and g^2 are measurable, then*

$$\Big(\int fg\,d\mu\Big)^2 \le \int f^2\,d\mu \cdot \int g^2\,d\mu.$$

Proof. Take any λ; then $\int (f+\lambda g)^2\,d\mu \ge 0$. In particular,

$$\underbrace{\int g^2\,d\mu}_{a}\cdot\lambda^2 + \underbrace{2\int fg\,d\mu}_{b}\cdot\lambda + \underbrace{\int g^2\,d\mu}_{c} \ge 0 \quad \forall\ \lambda.$$

In other words, the quadratic (in λ) can have at most one real root. Using the quadratic formula,

$$\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a},$$

it is clear that the only way there can be fewer than two real roots is if b^2-4ac is ≤ 0. Using the definitions of a, b, and c we find that

$$4\Big(\int fg\,d\mu\Big)^2 - 4\int f^2\,d\mu \cdot \int g^2\,d\mu \le 0,$$

and from this the result follows immediately. □

Definition. The function f is said to be convex on $\mathbb{X}$ if, for any $x, y \in \mathbb{X}$ and any $\alpha \in [0,1]$, the following inequality holds:

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y).$$

In the case where f is twice differentiable, there is an alternative characterization of convexity. A twice-differentiable function f, defined on p-dimensional space, is convex iff

$$\nabla^2 f(x) = \Big(\Big(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\Big)\Big)_{i,j=1,\ldots,p},$$

the matrix of second derivatives, is positive semi-definite for each x.

Example 49. Examples of convex (univariate) functions include e^x, $-\log x$, x^r for $r > 1$.

Convexity is important in optimization problems (maximum likelihood, least squares, etc) as it relates to existence and uniqueness of global minima.

Theorem 26. *(Jensen's inequality) Suppose φ is a convex function on an open interval $\mathbb{X} \subseteq \mathbb{R}$, and X is a random variable taking values in $\mathbb{X}$. Then*

$$\varphi[\mathbb{E}(X)] \le \mathbb{E}[\varphi(X)].$$

If φ is stricly convex, then equality holds if and only if X is constant.

Proof. First, take x_0 to be any fixed point in $\mathbb{X}$. Then there exists a linear function $\ell(x) = c(x - x_0) + \varphi(x_0)$, through the point $(x_0, \varphi(x_0))$, such that $\ell(x) \leq \varphi(x)$ for all x. To prove our claim, take $x_0 = \mathbb{E}(X)$, and note that

$$\varphi(X) \geq c[X - \mathbb{E}(X)] + \varphi[\mathbb{E}(X)].$$

Taking expectations on both sides gives the result. □

Theorem 27. *(Markov's Inequality) Suppose X is a random variable and h is a non-decreasing non-negative function. The expectation $\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$ exists. Then for any $a > 0$,*

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[h(X)]}{h(a)}.$$

Proof. We can write

$$\int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \geq \int_{a}^{\infty} h(x)f(x)\mathrm{d}x \geq h(a)\int_{a}^{\infty} f(x)\mathrm{d}x = h(a)\Pr\{X \geq a\}.$$

This leads directly to the following inequality called Markov inequality:

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[h(X)]}{h(a)}.$$

If we set $h(x) = |x|$, we have the standard version of Markov inequality:

$$\Pr(|X(\omega)| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}$$

for $a > 0$. □

Theorem 28. *(Chebyshev's Inequality) For any r.v. Y with mean μ and variance σ^2, and for any $k > 0$,*

$$\Pr(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. Apply Markov's Inequality with $X = (Y - \mu)^2$ and $a = k^2\sigma^2$. The result follows. □

7.4 Exercises

1. Show the conditional CDF for $X \sim U[0, 1]$ with conditioning event $X \geq z$. $U[0, 1]$ is the uniform distribution.

2. Conditional expectationoften appears in econometrics as the Best approximation. The idea is as follows. Suppose that X and Y are two random variables. We wish to find the function f^* such that $f*(X)$ is the minimizer of $\mathbb{E}[Y - f(X)]^2$ over all functions f i.e.

   $$\mathbb{E}[Y - f^*(X)]^2 = \min_f \mathbb{E}[Y - f(X)]^2.$$

   Show that $f^*(X)$ as $\mathbb{E}[Y|X]$, the conditional expectation of Y given X.

3. Change of variables: Let X be a random variable with distribution μ on $(S, \mathcal{S})$. If f is a measurable function from $(S, \mathcal{S})$ to $(\mathbb{R}, \mathcal{R})$ such that $f \geq 0$ or $\mathbb{E}|f(X)| < \infty$, then

$$\mathbb{E}[f(X)] = \int_S f(y)\, \mu(dy).$$

To prove this result, we need four steps: Step 1 use indicator Functions. If $B \in \mathcal{S}$ and $f = 1_B$, then

$$\mathbb{E}[1_B(X)] = \Pr(X \in B) = \mu(B) = \int_S 1_B(y)\, \mu(dy).$$

Prove step 2: use Simple Functions. Let $f(x) = \sum_{i=1}^n c_i 1_{B_i}(x)$ to prove the argument.

4. Prove step 3: extend the result in step 2 to Nonnegative Functions. Can you construct a simple function f_n such that $f_n \uparrow f$ as $n \to \infty$ and $f \geq 0$? Combining result in step 2, can you prove the following statement?

$$\mathbb{E}[f(X)] = \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \lim_{n\to\infty} \int_S f_n(y)\mu(dy) = \int_S f(y)\mu(dy).$$

5. Prove step 4: extend the result in step 3 to Integrable Functions. [Hint write $f(x) = f^+(x) - f^-(x)$ and use the result in step 3.]

Chapter 8

Integration of Limits and Distributions

8.1 Integration of Limits

Recalled that if $X \geq 0$ is a random variable on $(\Omega, \mathcal{F}, P)$, then we define its expected value to be $\mathbb{E}X = \int X dP$. Notice that this quantity may be ∞. For general X, we say that $\mathbb{E}X$ exists if the difference

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$$

is well-defined, which it will be if either $\mathbb{E}X^+ < \infty$ or $\mathbb{E}X^- < \infty$. These integrals are taken over all of Ω. If we wish to integrate over a measurable subset $A \subset \Omega$, we will write

$$\mathbb{E}[X|\mathcal{A}] \equiv \int_{\mathcal{A}} X \, dP \equiv \int X \mathbf{1}_{\mathcal{A}} \, dP.$$

Notice that $\mathbb{E}X$ inherits all of the properties of the Lebesgue integral. In particular,

Theorem 29. *Suppose that $X, Y \geq 0$ or $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$. Then 1) $\mathbb{E}[X+Y] = \mathbb{E}X + \mathbb{E}Y$. 2) $\mathbb{E}[aX+b] = a\mathbb{E}[X] + b$ for any $a, b \in \mathbb{R}$. 3) If $\Pr\{X \geq Y\} = 1$, then $\mathbb{E}X \geq \mathbb{E}Y$.*

We are interested in conditions that guarantee that if $X_n \to X$, then $\mathbb{E}X_n \to \mathbb{E}X$. The following example shows that this does not hold in general.

Example 50. Take $\Omega = (0,1)$, $\mathcal{F}$ are the Borel sets and P is Lebesgue measure on $(0,1)$. If $X_n = n1_{(0,1/n)}$, then $X_n \to X \equiv 0$, but $\mathbb{E}X_n = 1 > 0 = \mathbb{E}X$.

We begin by recalling three classical results from analysis.

Lemma 2. *(Fatou's Lemma) If $X_n \geq 0$, then $\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]$.*

Theorem 30. *(Monotone Convergence Theorem) If $0 \leq X_n \uparrow X$, then $\mathbb{E}X_n \uparrow \mathbb{E}X$.*

Proof. Since $X_n \leq X$ for all n, we know that $\limsup_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X$. However, since $X = \liminf_{n\to\infty} X_n$, Fatou's Lemma implies that $\mathbb{E}X \leq \liminf_{n\to\infty} \mathbb{E}X_n$. Combining these two results shows that

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n.$$

□

Theorem 31. *(Dominated Convergence Theorem) If $X_n \to X$ a.s. and $|X_n| \le Y$ for all n, where $\mathbb{E}Y < \infty$, then $\mathbb{E}X_n \to \mathbb{E}X$.*

The special case where Y is constant is called the bounded convergence theorem.

The following theorem can handle some cases that are not covered by either the monotone or the dominated convergence theorems.

Theorem 32. *Suppose that $X_n \to X$ a.s. Let g, h be continuous functions such that 1) $g \ge 0$ and $g(x) \to \infty$ as $|x| \to \infty$; 2) $|h(x)|/g(x) \to 0$ as $|x| \to \infty$; 3) $\mathbb{E}[g(X_n)] \le K < \infty$ for all n. Then $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$.*

Proof. By subtracting a constant from h, we can assume wlog that $h(0) = 0$. Choose M so that $g(x) > 0$ whenever $|x| \ge M$. Given a random variable Y, let $\bar{Y} = Y 1_{(|Y| \le M)}$.[1] Then $\bar{X}_n \to \bar{X}$ a.s. If $|X_n| < M$ for all n sufficiently large then $\bar{X}_n = X_n \to X = \bar{X}$, if $|X_n| > M$ for all n sufficiently large then $\bar{X}_n = 0 \to \bar{X} = 0$. Since $h(\bar{X}_n)$ is bounded and h is continuous, the bounded convergence theorem implies that

$$\mathbb{E}[h(\bar{X}_n)] \to \mathbb{E}[h(\bar{X})].$$

(The rest of the proof is going to be a homework.) □

Corollary 2. *Suppose that $X_n \to X$ a.s. and that there exists a $K < \infty$ and a $p > 1$ such that $\mathbb{E}[X_n^p] \le K$ for all $n \ge 1$. Then $\mathbb{E}X_n \to \mathbb{E}X$.*

8.2 Special Distributions

In most of what we do in this course, here in particular, we will ignore the underlying probability space and work just with probability measures on the X-space. In a statistical problem, there is not just one probability measure in question, but a whole family of measures P_θ indexed[2] by a parameter $\theta \in \Theta$. You're already familiar with this setup; $X_1, \ldots, X_n$ iid $\mathcal{N}(\theta, 1)$ is one common example. Here are some others.

Example 51. (Binomial Distribution) The probability mass function with parameters n and p is

$$p_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{n-x}$$

for $x = 0, 1, \ldots n$. The mean of X is np. A single Bernoulli trial is $p^x(1-p)^{1-x}$ for $x = 0, 1$. Its mean is p and its variance is $p(1-p)$.

Example 52. (Uniform Distribution) A r.v. X on the interval (a, b) has PDF $p_X(x) = 1/(b-a)$ for $x \in (a, b)$. The mean and variance are $(a+b)/2$ and $(b-a)^2/12$ respectively.

[1] Check the (*) inequality in your exercise to understand the role of truncation.

[2] Note that the subscript in P_θ serves a different purpose than the subscript P_X.

Example 53. (Exponential Distribution) A r.v. X with parameter $\lambda > 0$ has PDF $p_X(x) = \lambda \exp(-\lambda x)$ for $x > 0$. Sometimes, it is written in terms of $\beta = 1/\lambda$ and its PDF becomes $p_X(x) = (1/\beta)\exp(-x/\beta)$.

Example 54. One of the most important distributions is the normal distribution. It does not have as easy a motivation as some of the other distributions, but it is of fundamental importance as an approximation to a large number of statistics through the central limit theorem. A r.v. X has a normal distribution with parameters μ and σ^2, denoted by $\mathcal{N}(\mu, \sigma^2)$ has PDF

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

for $-\infty < x < \infty$, with the parameter space $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

8.3 Moment Generating Functions and Characteristic Functions

The moments of a random variable are summarized in the moment generating function.

Definition. The moment generating function of X is $\mathbb{E}e^{tX}$ provided that the expectation exists in some neighborhood $t \in [-h, h]$ of zero. Thi is also called the Laplace transform.

Example 55. Standard Normal distribution:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(tx - \frac{x^2}{2}\right) dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-t)^2 - t^2\right) dx \\ &= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-t)^2\right) dx \\ &= \exp\left(\frac{1}{2}t^2\right) \cdot 1.\end{aligned}$$

First moment: $\mathbb{E}X = \partial e^{\frac{t^2}{2}}/\partial t|_{t=0} = t \cdot \exp(t^2/2)|_{t=0} = 0$. Second moment: $\mathbb{E}X^2 = \partial^2 e^{\frac{t^2}{2}}/\partial t^2|_{t=0} = e^{\frac{t^2}{2}} + t^2 e^{\frac{t^2}{2}} = 1$. kth-moment: $\partial^k \mathbb{E}[Xe^{tX}]/\partial t^k|_{t=0} = \mathbb{E}X^k$.

In many cases, the moment generating function can characterize a distribution. But problem is that it may not exist (eg. Cauchy distribution) For a r.v. X, is its distribution uniquely determined by its moment generating function?

Theorem 33. *For $X \sim P_X$ and $Y \sim P_Y$, if their moment generating functions exist, and they are equivalent for all t in some neighborhood of zero, then $P_X(u) = P_Y(u)$ for all u.*

One of the most important properties of the normal distribution is that linear transformations of normal random variables are also normally distributed. Consider a random

variable $X$with a $\mathcal{N}(\mu, \sigma^2)$ distribution, and consider the transformation $Y = a + bX$. Then, through the moment generating function

$$\begin{aligned}\mathbb{E}(e^{tY}) &= \mathbb{E}(e^{(a+bX)t}) = e^{at} \cdot \mathbb{E}(e^{btX}) \\ &= \exp\left(at + \mu bt + \frac{(\sigma bt)^2}{2}\right) = \exp\left(\tilde{\mu}t + \frac{\tilde{\sigma}^2 t^2}{2}\right)\end{aligned}$$

Hence Y has a normal distribution with $\tilde{\mu} = a + b\mu$ and $\tilde{\sigma}^2 = b^2\sigma^2$.

Note that if the moment generating function exists, then it characterizes a random variable with an infinite number of moments (because themoment generating function is infinitely differentiable).

There is a connection between the normal distribution and the Chi-squared distribution. If X has a standard normal distribution $\mathcal{N}(0, 1)$, then $Y = X^2$ has a Chi-squared distribution with degrees of freedom equal to one. One argument goes as follows

$$\begin{aligned}\mathbb{E}(e^{tY}) = \mathbb{E}(e^{X^2 t}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2 + tx^2\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1+2t)x^2}{2}\right) dx \\ &= \frac{1}{(1-2t)^{1/2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi/(1-2t)}} \exp\left(-\frac{x^2}{2/(1-2t)}\right) dx = \frac{1}{(1-2t)^{1/2}}\end{aligned}$$

which is the moment generating function for a Chi-square distribution with degrees of freedom one.

Definition. If X is a random variable, then its characteristic function is defined to be

$$\varphi(t) = \mathbb{E}e^{itX} = \int_{-\infty}^{\infty} \exp(itx)p(x)dx$$

This is also called the Fourier transform.

Remark. The characteristic function always exists. It completely determines the distribution of X. This follows from the equality

$$\varphi(t) = \mathbb{E}e^{itX} = \mathbb{E}\cos(tX) + i\,\mathbb{E}\sin(tX).$$

Characteristic functions satisfy the following properties: 1) $\varphi(0) = 1$; 2) $\varphi(-t) = \mathbb{E}\cos(-tX) + i\,\mathbb{E}\sin(-tX) = \overline{\varphi(t)}$; 3) $|\varphi(t)| = |\mathbb{E}e^{itX}| \leq \mathbb{E}|e^{itX}| \leq 1$; 4) $\mathbb{E}e^{it(aX+b)} = e^{itb}\varphi(at)$.

Remark. If X and Y are independent with characteristic functions φ_1 and φ_2, then the characteristic function of $X + Y$ is

$$\mathbb{E}e^{it(X+Y)} = \mathbb{E}\left[e^{itX}e^{itY}\right] = \mathbb{E}e^{itX}\mathbb{E}e^{itY} = \varphi_1(t)\varphi_2(t).$$

Notice that this relationship extends to arbitrary finite sums of independent random variables.

Example 56. Bernoulli distribution: If $P(X=1)=P(X=-1)=1/2$, then

$$\mathbb{E}e^{itX}=\left(e^{it}+e^{-it}\right)/2=\cos(t).$$

Example 57. Poisson distribution: If $P(X=k)=e^{-\lambda}\lambda^k/k!$, then

$$\mathbb{E}e^{itX}=\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k e^{itk}}{k!}=e^{\lambda(e^{it}-1)}$$

Example 58. Normal distribution: Suppose that X is a standard normal random variable. Then

$$\mathbb{E}e^{itX} \;=\; \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{itx-x^2/2}dx=e^{-t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-it)^2/2}dx=e^{-t^2/2}.$$

In general, if X is a normal random variable with mean μ and variance σ^2, then $Z=(X-\mu)/\sigma$ is a standard normal random variable and so

$$\mathbb{E}e^{itX}=e^{it\mu}e^{-\sigma^2t^2/2}.$$

The corresponding density $p_X(x)$ is available by the inverse Fourier transform, which is

$$p_X(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi_X(t)\exp(-itx)dt.$$

Take as given that the characteristic function is $e^{-t^2/2}$. The inversion formula yields

$$\begin{aligned}p_X(x)=&\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-t^2/2)\exp(-itx)dt\\=&\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-itx-x^2/2)dt\end{aligned}$$

Now substituting $z=-t$, we have

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(itz-z^2/2)dz=\frac{1}{\sqrt{2\pi}}e^{z^2/2}.$$

Example 59. Uniform distribution on (a,b): If X has density $1_{(a,b)}(x)/(b-a)$, then

$$\mathbb{E}e^{itX}=\frac{1}{b-a}\int_a^b e^{itx}dx=\frac{e^{itb}-e^{ita}}{it(b-a)}.$$

In the special case where $a=-b=l$, the characteristic function is $\sin(lt)/lt$.

Example 60. Exponential distribution: If X has density $\lambda e^{-\lambda x}$ on $[0,\infty)$, then

$$\mathbb{E}e^{itX}=\int_0^{\infty}\lambda e^{(it-\lambda)x}dx=\frac{\lambda}{\lambda-it}.$$

Characteristic function also summarizes the moments of a random variable. Specifically, note that the h-th derivative of $\varphi_X(t)$ is

$$\varphi_X^{(h)}(t)=\int_{-\infty}^{\infty}i^h x^h\exp(-itx)p_X(x)dx.$$

8.4 Exercises

1. Theorem 4: To control the truncation error, let

$$\epsilon_M \equiv \sup\{|h(x)|/g(x) : |x| \geq M\}.$$

and observe that for any random variable Y we have

$$(\star) \qquad \begin{aligned} \big|\mathbb{E}[h(\bar{Y})] - \mathbb{E}[h(Y)]\big| &\leq \mathbb{E}\big|h(\bar{Y}) - h(Y)\big| \\ &= \mathbb{E}[|h(Y)|\,|\,|Y| > M] \;\leq\; \epsilon_M \mathbb{E}[g(Y)]. \end{aligned}$$

Can you use this argument to finish the proof? [Hint: take $Y = X_n$ and take $Y = X$ in $(\star)$, and see what could you have.] [Answer: Taking $Y = X_n$ in $(\star)$ and using condition (3) in the theorem, we have

$$\big|\mathbb{E}h(\bar{X}_n) - \mathbb{E}h(X_n)\big| \leq K\epsilon_M.$$

To estimate the remaining truncation error, notice that because $g \geq 0$ and g is continuous, Fatou's lemma implies that

$$\mathbb{E}[g(X)] \leq \liminf_{n\to\infty} \mathbb{E}[g(X_n)] \leq K.$$

Then, taking $Y = X$ in $(\star)$ gives

$$\big|\mathbb{E}h(\bar{X}) - \mathbb{E}h(X)\big| \leq K\epsilon_M.$$

Finally, by the triangle inequality, we have

$$\begin{aligned} \big|\mathbb{E}h(X_n) - \mathbb{E}h(X)\big| &\leq \big|\mathbb{E}[h(X_n)] - \mathbb{E}[h(\bar{X}_n)]\big| \\ &+ \big|\mathbb{E}[h(\bar{X}_n)] - \mathbb{E}[h(\bar{X})]\big| + \big|\mathbb{E}[h(\bar{X})] - \mathbb{E}[h(X)]\big|. \end{aligned}$$

Letting $n \to \infty$, we obtain

$$\limsup_{n\to\infty} \big|\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]\big| \leq 2K\epsilon_M$$

which can be made arbitrarily close to 0 since $\epsilon_M \to 0$ as $M \to \infty$.]

2. Can you prove that the characteristic function φ is uniformly continuous on $\mathbb{R}$? [Answer: Since

$$\begin{aligned} \big|\varphi(t+h) - \varphi(t)\big| &= \big|\mathbb{E}\big(e^{i(t+h)X} - e^{itX}\big)\big| \\ &\leq \mathbb{E}\big|e^{i(t+h)X} - e^{itX}\big| = \mathbb{E}\big|(e^{ihX} - 1)e^{itX}\big|, \end{aligned}$$

and the bounded convergence theorem shows that the last quantity converges to 0 as $h \to 0$, it follows that φ is uniformly continuous on $\mathbb{R}$.]

Chapter 9

LLN and CLT

9.1 Weak Law of Large Numbers

We begin by defining some modes of convergence for random variables. Suppose that $X_n, n \geq 1$ and X are random variables defined on the same probability space.

Definition. (L^2 weak law) We say that X_n converges to X in L^p if $\mathbb{E}|X - X_n|^p \to 0$ as $n \to \infty$.

Definition. We say that X_n converges to X **in probability** and write $X_n \xrightarrow{p} X$ if for every $\epsilon > 0$, we have $\Pr\{|X_n - X| > \epsilon\} \to 0$.

Lemma 3. *If $r > 0$ and $\mathbb{E}|X_n|^r \to 0$, then $X_n \to 0$ in probability.*

Proof. The result follows from Chebyshev's inequality which shows that

$$\Pr\{|X_n| > \epsilon\} \leq \epsilon^{-r}\mathbb{E}|X_n|^r \to 0.$$

□

We say that a family of random variables, $X_i, i \in I$, is **uncorrelated** if $\mathbb{E}X_i^2 < \infty$ for every $i \in I$ and $\mathbb{E}X_iX_j = 0$ whenever $i \neq j$.

Lemma 4. *Let $X_1, \cdots, X_n$ be uncorrelated. Then*

$$Var\,(X_1 + \cdots + X_n) = \sum_{i=1}^{n} Var(X_i).$$

Let $X_1, X_2, \cdots$ be uncorrelated random variables with $\mathbb{E}X_i = \mu$ and $Var(X_i) \leq C < \infty$. If $S_n = X_1 + \cdots + X_n$, then as $n \to \infty$, $S_n/n \to \mu$ in L^2 and in probability.

Proof. To prove L^2 convergence, observe that $\mathbb{E}[S_n/n] = \mu$, so

$$\mathbb{E}\big(S_n/n - \mu\big)^2 = \mathrm{Var}(S_n/n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \leq \frac{Cn}{n^2} \to 0.$$

Convergence in probability then follows from Lemma 1. □

Many limit theorems in probability address the asymptotic behavior of the row sums of arrays $X_{n,k}, 1 \leq k \leq n$ of random variables.

Theorem 34. *Let $\mu_n = \mathbb{E}S_n$ and $\sigma_n^2 = Var(S_n)$. If $\sigma_n^2/b_n^2 \to 0$, then $(S_n - \mu_n)/b_n \to 0$ in probability.*

Proof. The result follows from Lemma 2 since

$$\mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 = b_n^{-2}\mathrm{Var}(S_n) \to 0.$$

□

We can use truncation to extend the weak law to random variables without a second moment.

Theorem 35. *(Weak law for triangular arrays) For each $n \geq 1$, let $X_{n,k}, 1 \leq k \leq n$ be a collection of independent random variables. Let $b_n, n > 1$ be a collection of real numbers with $b_n \to \infty$ and let $\bar{X}_{n,k} = X_{n,k}1_{(|X_{n,k}| \leq b_n)}$. Suppose that as $n \to \infty$*
(1) $\sum_{k=1}^{n}\Pr(|X_{n,k}| > b_n) \to 0$, and
(2) $b_n^{-2}\sum_{k=1}^{n}\mathbb{E}\bar{X}_{n,k}^2 \to 0$.
If $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{k=1}^{n}\mathbb{E}\bar{X}_{n,k}$, then $(S_n - a_n)/b_n \to 0$ in probability.

Proof. Let $\bar{S}_n = \bar{X}_{n,1} + \cdots + \bar{X}_{n,n}$. Then

$$\Pr\left(\left|\frac{S_n - a_n}{b_n}\right| > \epsilon\right) \leq \Pr(S_n \neq \bar{S}_n) + \Pr\left(\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon\right).$$

To estimate the first term on the RHS, note that

$$\Pr(S_n \neq \bar{S}_n) \leq \Pr\left(\cup_{k=1}^{n}\{\bar{X}_{n,k} \neq X_{n,k}\}\right) \leq \sum_{k=1}^{n}\Pr(|X_{n,k}| > b_n) \to 0$$

by condition (1). For the second term, we can use condition (2) along with Chebyshev's inequality and the fact that $\mathrm{Var}(X_n) \leq \mathbb{E}X_n^2$ to show that

$$\begin{aligned}\Pr\left(\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon\right) &\leq \epsilon^{-2}\mathbb{E}\left|\frac{\bar{S}_n - a_n}{b_n}\right|^2 = \epsilon^{-2}b_n^{-2}\mathrm{Var}(\bar{S}_n) \\ &= (b_n\epsilon)^{-2}\sum_{k=1}^{n}\mathrm{var}(\bar{X}_{n,k}) \leq (b_n\epsilon)^{-2}\sum_{k=1}^{n}\mathbb{E}(\bar{X}_{n,k})^2 \to 0.\end{aligned}$$

□

Lemma 5. *If $Y \geq 0$ and $p > 0$, then*

$$\mathbb{E}Y^p = \int_0^\infty py^{p-1}P(Y > y)dy.$$

Proof. Using Fubini's Theorem, we can calculate

$$\begin{aligned}\int_0^\infty py^{p-1}P(Y > y)dy &= \int_0^\infty \int_\Omega py^{p-1}1_{(Y>y)}dPdy \\ &= \int_\Omega \int_0^\infty py^{p-1}1_{(Y>y)}dydP \\ &= \int_\Omega \int_0^Y py^{p-1}dydP \;=\; \int_\Omega Y^p dP \;=\; \mathbb{E}Y^p.\end{aligned}$$

□

Theorem 36. *(Weak Law of Large Numbers) Let $X_1, X_2, \cdots$ be i.i.d. with*

$$(\star) \qquad xP(|X_1| > x) \to 0 \text{ as } x \to \infty.$$

Let $S_n = X_1 + \cdots + X_n$ and $\mu_n = \mathbb{E}\big(X_1 1_{(|X_1|\leq n)}\big)$. Then $S_n/n - \mu_n \to 0$ in probability.

Proof. We will apply Theorem 2 with $X_{n,k} = X_k$ and $b_n = n$. To check condition (1) in that theorem, observe that $(\star)$ implies that

$$\sum_{k=1}^n P(|X_{n,k}| > n) = nP(|X_1| > n) \to 0.$$

For condition (2), we need to show that $n^{-2} \cdot n\mathbb{E}\bar{X}_{n,1}^2 \to 0$. First observe that

$$\frac{1}{n}\mathbb{E}\bar{X}_{n,1}^2 = \frac{1}{n}\int_0^\infty 2yP(|\bar{X}_{n,1}| > y)dy \leq \frac{1}{n}\int_0^n 2yP(|X_1| > y)dy$$

since $P(|\bar{X}_{n,1}| > y) = 0$ for $y \geq n$. To show that the last term tends to 0 as $n \to \infty$, let $g(y) = 2yP(|X_1| > y)$ and notice that $0 \leq g(y) \leq 2y$ and $g(y) \to 0$ as $y \to \infty$ imply that $M = \sup g(y) < \infty$. If we define $\epsilon_K = \sup\{g(y) : y > K\}$, then for $n > K$

$$\int_0^n 2yP(|X_1| > y)dy \leq KM + (n - K)\epsilon_K.$$

Dividing by n and letting $n \to \infty$ gives

$$\limsup_{n\to\infty} \frac{1}{n}\int_0^n 2yP(|X_1| > y)dy \leq \epsilon_K.$$

The result then follows upon noting that $\epsilon_K \to 0$ as $K \to \infty$. □

The familiar form of the weak law of large numbers is:

Corollary 3. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}|X_i| < \infty$. Let $S_n = X_1 + \cdots + X_n$ and $\mu = \mathbb{E}X_1$. Then $S_n/n \to \mu$ in probability.*

Proof. Chebyshev's inequality and the dominated convergence theorem imply that

$$xP(|X_1| > x) \le \mathbb{E}\big(|X_1| 1_{(|X_1|>x)}\big) \to 0 \quad \text{as } x \to \infty$$
$$\mu_n = \mathbb{E}\big(X_1 1_{(|X_1|\le n)}\big) \to \mathbb{E}X_1 = \mu \quad \text{as } n \to \infty.$$

Using Theorem 3, we see that if $\epsilon > 0$, then $P(|S_n/n - \mu_n| > \epsilon/2) \to 0$. Since $\mu_n \to \mu$, it follows that $P(|S_n/n - \mu| > \epsilon) \to 0$. □

Definition. A sequence of distribution functions $(P_{X_n}; n \ge 1)$ is said to **converge weakly** to a limit P if $P_{X_n}(x) \to P(x)$ for all x that are points of continuity of P. In this case, we write $P_{X_n} \rightsquigarrow P$ or $P_{X_n} \Rightarrow P$ or $P_{X_n} \xrightarrow{w} P$. Similarly, a sequence of random variables X_n is said to **converge in distribution** to a limit X, written $X_n \rightsquigarrow X$ or $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$, if the distribution functions $P_{X_n}(x) = \Pr(X_n \le x)$ converge weakly to the distribution function of X.

Theorem 37. *(Central Limit Theorem) Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2 < \infty$. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \rightsquigarrow Z$$

where $Z \sim \mathcal{N}(0,1)$.

Proof. Step 1: The following inequality holds for all x:

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \le \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right). \tag{9.1}$$

Step 2: If $\mathbb{E}|X|^2 < \infty$, then

$$\varphi(t) = 1 + it\mathbb{E}X - \frac{1}{2}t^2\mathbb{E}X^2 + r.$$

the error term r is bounded by

$$\mathbb{E}\left[\frac{|tX|^3}{6} \wedge \frac{2|tX|^2}{2} \right] = t^2\mathbb{E}\left[\frac{|t||X|^3}{6} \wedge |X|^2 \right].$$

Since the expression inside the expectation tends to 0 as $t \to 0$, from the dominated convergence theorem and the fact that X^2 is integrable we know $r \to 0$.

Step 3: It suffices to consider the case $\mu = 0$. Since X_1 has finite variance, step 2 implies that

$$\varphi(t) = \mathbb{E}\left[e^{itX_1}\right] = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$$

so that

$$\mathbb{E}\left[\exp\left(itS_n/\sigma\sqrt{n}\right)\right] = \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n \to e^{-t^2/2}.$$

Since the limit on the right-hand side is the characteristic function of Z, the result follows. □

Theorem 38. *(Delta Method): Let X_n be a sequence of random vectors such that*

$$\sqrt{n}(X_n - \mu) \rightsquigarrow \mathcal{N}(0, \Sigma)$$

where Σ is positive definite and finite. Let g denote a continuous differentiable function from $\mathbb{R}^d$ into $\mathbb{R}^k$, and let $G(x) = \partial g/\partial x$ denote the $k \times d$ matrix of partial derivative. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \rightsquigarrow \mathcal{N}(0, G(\mu)\Sigma G(\mu)^T).$$

Appendix

Theorem. *(Fubini) Let $f(x, y)$ be a non-negative measurable function on $\mathbb{X} \times \mathbb{Y}$. Then*

$$\int_{\mathbb{X}} \left[\int_{\mathbb{Y}} f(x,y)\, d\nu(y)\right] d\mu(x) = \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} f(x,y)\, d\mu(x)\right] d\nu(y).$$

The common value above is the double integral, written $\int_{\mathbb{X}\times\mathbb{Y}} f\, d(\mu \times \nu)$.

Theorem. *(Dominated convergence) Given measurable $\{f_n\}$, suppose that*

$$f(x) = \lim_{n\to\infty} f_n(x) \quad \mu\text{-almost everywhere,}$$

and $|f_n(x)| \le g(x)$ for all n, for all x, and for some integrable function g. Then f_n and f are integrable, and

$$\int f\, d\mu = \lim_{n\to\infty} \int f_n\, d\mu.$$

Proof. of Equation (1) in CLT. Integration by parts gives

$$\int_0^x (x-s)^n e^{is} ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1} e^{is} ds,$$

which for $n = 0$ says

$$\int_0^x e^{is} ds = x + i\int_0^x (x-s) e^{is} ds.$$

Since $\int_0^x e^{is}ds = (e^{ix}-1)/i$, rearranging gives

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s)e^{is}ds.$$

Next, taking $n = 1$, we have

$$e^{ix} = 1 + ix + \frac{i^2x^2}{2} + \frac{i^3}{2}\int_0^x (x-s)^2 e^{is}ds,$$

and iterating leads to

$$(a) \qquad e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} = \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds.$$

To estimate the magnitude of the remainder term on the right-hand side, we can use the fact that $|e^{is}| \le 1$ for all s to see that

$$\left|\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds\right| \le \frac{|x|^{n+1}}{(n+1)!}.$$

This estimate is good when $|x|$ is small relative to n. To cope with large $|x|$, we again integrate by parts

$$\begin{aligned}\frac{i}{n}\int_0^x (x-s)^n e^{is}ds &= -\frac{x^n}{n} + \int_0^x (x-s)^{n-1}e^{is}ds \\ &= -\int_0^x (x-s)^{n-1}ds + \int_0^x (x-s)^{n-1}e^{is}ds.\end{aligned}$$

Multiplying through by $i^n/(n-1)!$ then gives

$$\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds = \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1)ds$$

and since $|e^{is}-1| \le 2$ for all s, it follows that

$$(b) \qquad \left|\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds\right| \le \left|\frac{2}{(n-1)!}\int_0^x (x-s)^{n-1}ds\right| \le \frac{2|x|^n}{n!}.$$

Thus the conclusion follows upon combining (a) and (b). □

Excercise

1. Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}|X_i| < \infty$. Let $S_n = X_1 + \cdots + X_n$ and $\mu = \mathbb{E}X_1$. Use the previous result to show that $S_n/n \to \mu$ in probability. [Refer to the main text.]
2. To show that the last term of $(\star\star)$ tends to 0 as $n \to \infty$. The result of the theorem follows. [Refer to the main text.]

Part III

Statistics

Chapter 10

Likelihood and MLE

10.1 Likelihood

Likelihood is surely one of the most important concepts in statistical theory. The likelihood function establishes a preference among the possible parameter values given data $X = x$. That is, a parameter values θ_1 with larger likelihood is better than parameter value θ_2 with smaller likelihood, in the sense that the model P_{θ_1} provides a better fit to the observed data than P_{θ_2}. This leads naturally to procedures for inference which select, as a point estimator, the parameter value that makes the likelihood the largest, or rejects a null hypothesis if the hypothesized value has likelihood too small. The likelihood function is also of considerable importance in Bayesian analysis.

Likelihood function provides *the* basis for statistical inference. That is, all "good" statistical methods are driven by the likelihood function (or some variation thereof). There is a formal, and somewhat controversial version of this claim, called the *likelihood principle*, which says something like the following: all the relevant information in data about the parameter is contained in the (shape of the) likelihood function and, furthermore, if two data sets give rise to likelihood functions with the same shape, then the same conclusions should be reached for both data sets.

> What has now appeared is that the mathematical concept of probability is ... inadequate to express our mental confidence or indifference in making ... inferences, and that the mathematical quantity which usually appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "likelihood" to designate this quantity; since both words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship. -by Fisher 1973

We use θ^* to denote the true value of the parameter that generated the data, and use θ to denote any element of the parameter space.

Definition. (Specification) Let $(\mathbb{X}, \mathcal{A})$ be a measurable space equipped with a family $\{P_\theta : \theta \in \Theta\}$ of probability measures (models) indexed by a parameter $\theta \in \Theta$. Data $X_1, \ldots, X_n$

are independent and identically distributed according to some P_{θ^*}. The goal is to estimate the unknown parameter θ^*.

Since $\{P_\theta : \theta \in \Theta\}$ is defined on the measurable space $(\mathbb{X}, \mathcal{A})$. If P_θ is absolutely continuous with respect to a dominating σ-finite measure μ, then for each θ, the Radon-Nikodym derivative $(dP_\theta/d\mu)(x)$ is the usual probability density function for the observable X, written as $p_\theta(x)$. For fixed θ, we know that $p_\theta(x)$ characterizes the sampling distribution of X.

Definition. A point estimator is a statistic used to provide a guess about θ.

Definition. Given $X = x$, the likelihood function is $L(\theta) = p_\theta(x)$.

Definition. θ is *identifiable*, that is, $\theta \mapsto P_\theta$ is one-to-one.

Remark. This just means that it is possible to estimate θ based on sample data. An example of a model that's not identifiable is $\mathcal{N}(\theta_1 + \theta_2, 1)$, we can estimate the sum $\theta_1 + \theta_2$ but not the individual components.

If we have more than one random variable, say $X_1, \ldots, X_n$, the likelihood function is based on the joint probability density/mass function: the likelihood is $L_n(\theta) = p_\theta(X_1, \ldots, X_n)$. If the random variables are i.i.d., with common density function

$$\mathcal{L}(\theta) = p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i)$$

Often we prefer to work with the logarithm of the likelihood function, the log likelihood function:

$$L_n(\theta) = \ln \mathcal{L}(\theta) = \ln p_\theta(X_1, \ldots, X_n) = \sum_{i=1}^{n} \ln p_\theta(X_i).$$

Definition. Given a class of potential models P_θ indexed by $\theta \in \Theta$, we observe $X = x$ and we'd like to know which model is the most likely to have produced this x. This defines an optimization problem:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta), \tag{10.1}$$

is the *maximum likelihood estimate* (MLE) of θ.

$P_{\hat{\theta}}$ is then considered the most likely model, that is, among the class $\{P_\theta : \theta \in \Theta\}$, the model $P_{\hat{\theta}}$ provides the best fit to the observed $X = x$. In terms of "ranking" intuition, $\hat{\theta}$ is ranked the highest.

- *Bayesian estimation* (optional). In a Bayesian context, there is also a prior probability measure Π on the parameter space Θ. Then, according to Bayes' theorem, the posterior distribution $\Pi_n = \Pi_{n,x}$, given $X = x$, satisfies $\Pi_n \ll \Pi$ and

$$\frac{d\Pi_n}{d\Pi}(\theta) = \frac{L_n(\theta)}{\int_\Theta L_n(u)\, d\Pi(u)} \propto L_n(\theta). \tag{10.2}$$

When Π has a density π with respect to some dominating measure ν, then so does Π_n and $\pi_n(\theta) = (d\Pi_n/d\nu)(\theta) \propto L_n(\theta)\pi(\theta)$. If a Bayesian is forced to produce a point estimate of $\theta^\star$, then he/she might choose the posterior mean or posterior mode.

Remark. MLE is $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$. If we use the likelihood function as a relative measure of plausibility for candidate θ values, then the MLE $\hat{\theta}_n$ is the "most plausible." The intuition is that when lots of data are available, the likelihood function will look like a spike around $\theta = \theta^*$, so the MLE will be close to the true value.

Consider the random variable Y defined as the ratio of the density function at some arbitrary value of θ to the density function at θ^*, both evaluated at the random variable X:

$$Y = p_\theta(X)/p_{\theta^*}(X).$$

The logrithmic function gives the convexity. By Jensen's inequality

$$\mathbb{E}[-\ln Y] \geq -\ln \mathbb{E}[Y],$$

implying

$$\mathbb{E}\left[-\ln\left(\frac{p_\theta(X)}{p_{\theta^*}(X)}\right)\right] \geq -\ln\left(\mathbb{E}\left[\frac{p_\theta(X)}{p_{\theta^*}(X)}\right]\right),$$

where the expectation is over the distribution of X, that is the density $p_{\theta^*}(X)$. Then

$$\mathbb{E}\left[\frac{p_\theta(X)}{p_{\theta^*}(X)}\right] = \int \frac{p_\theta(x)}{p_{\theta^*}(x)} p_{\theta^*}(x)dx = \int p_\theta(x)dx = 1,$$

for all θ. So $\mathbb{E}\left[-\ln\left(\frac{p_\theta(X)}{p_{\theta^*}(X)}\right)\right] \geq 0$ implying

$$\mathbb{E}\left[-\ln p_\theta(X)\right] \leq \mathbb{E}\left[\ln p_{\theta^*}(X)\right]$$

for all θ. This implies that the expected value of the log likelihood is maximized at the true value θ^*, and therefore there is some hope that the actual log likelihood function is maximized at a value close to θ^*.

Example 61. (Normal Distribution with unknown μ and variance 1) The likelihood is

$$\mathcal{L}(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \mu)^2\right),$$

the log-likelihood is $-\ln(2\pi)/2 - \sum_i (X_i - \mu)^2/2$. The value of μ that maximizes the log likelihood function is $\hat{\mu} = \sum_i X/n$.

Example 62. (Normal Distribution with unknown μ and unknown variance) The likelihood is

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(X_i - \mu)^2\right),$$

the log-likelihood is

$$L(\mu, \sigma^2) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(X_i - \mu\right)^2.$$

To maximize this over μ and σ^2, we need to solve a system of equations:

$$\begin{aligned}\frac{\partial L}{\partial \mu}(\mu, \sigma^2) &= \sum_{i=1}^{n} \frac{1}{\sigma^2}(X_i - \mu) \\ \frac{\partial L}{\partial \sigma^2}(\mu, \sigma^2) &= \sum_{i=1}^{n} -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X_i - \mu)^2\end{aligned}$$

Setting both to zero gives us

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \hat{\mu})^2/n$$

where $\hat{\mu} = \sum_i X/n$.

Remark. An important property of maximum likelihood estimators is their invariance under reparametrization: If $\hat{\theta}$ is the maximum likelihood estimator for θ^*, then $\hat{\pi} = g(\hat{\theta})$ is the MLE for any one-to-one transformation $g(\cdot)$.

10.2 Hypothesis Testing

For two competing hypotheses H_0 and H_1 about the parameter θ, the likelihood ratio is often used to make a comparison. For example, for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the likelihood ratio is $L(\theta_0)/L(\theta_1)$, and large (resp. small) values of this ratio indicate that the data x favors H_0 (resp. H_1). A more difficult and somewhat more general problem is $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, where Θ_0 is a subset of Θ.

Definition. Define the likelihood ratio as

$$T_n = T_n(X, \Theta_0) = \frac{\sup_{\theta\in\Theta_0} L_n(\theta)}{\sup_{\theta\in\Theta} L_n(\theta)}. \tag{10.3}$$

Remark. The interpretation of this likelihood ratio is if the ratio is small, then data lends little evidence to the null hypothesis.

If the model is suitably "regular," then there is a nice asymptotic distribution theory for the MLE and likelihood ratio statistic. The presentation will be kept informal here; precise statements and proofs will come later.

Consider the case where $X = (X_1, \ldots, X_n)$ is an iid sample with a common density $p_\theta(x)$, $\theta \in \Theta \subseteq \mathbb{R}^d$, so that the likelihood function is $L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$. If the data were not iid, then the likelihood would still just be the joint density of data, treated as a function of θ. However, the theory would generally need some revision.

The MLE $\hat{\theta}$ in (10.1), a d-vector, is a solution to the likelihood equation $\nabla \log L_n(\theta) = 0$. Let $I(\theta)$ denote the $d \times d$ matrix. Suppose a consistent solution $\hat{\theta} = \hat{\theta}_n$ to the likelihood equation and it satisfies $\{nI(\theta^*)\}^{1/2}(\hat{\theta}_n - \theta^*) \to \mathcal{N}(0, I)$ in distribution under P_{θ^*} as $n \to \infty$. This *asymptotic normality* result says that, when n is large, the sampling distribution of $\hat{\theta}_n$ is approximately normal with mean θ^* and covariance matrix $\{nI(\theta^*)\}^{-1}$. This approximate distribution, together with a suitable estimate/approximation for $I(\theta)$, can be used to construction asymptotically correct confidence regions and tests for θ.

Let Θ be an open subset of $\mathbb{R}^d$, consider the testing problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, where Θ_0 is a subset of Θ that specifies the values $\theta_{01}, \ldots, \theta_{0m}$ of $\theta_1, \ldots, \theta_m$, $1 \leq m \leq d$. If H_0 is true, and if certain regularity conditions hold, then $W_n = -2 \log T_n$, with T_n in (10.3), satisfies $W_n \to \chi^2(m)$ in distribution, as $n \to \infty$. Such a result provides an approximate size-α test of H_0 when n is large, i.e., by rejecting H_0 iff W_n is more than $\chi^2_{m,1-\alpha}$, the $100(1-\alpha)$ percentile of the $\chi^2(m)$ distribution.

Chapter 11

Large Sample Theorem in MLE (I)

11.1 Introduction

Large-sample theory was and is crucial to the development of statistical methods. Before the availability of high-power computing, the only way to solve many statistical problems was via asymptotic approximations. Nowadays, bootstrap and Markov chain Monte Carlo methods are available to get finite-sample approximate inference, but there are still needs for asymptotic theory.

In cases where the MLE or Bayes estimates are available in closed-form, standard tools from probability theory (e.g., law of large numbers, central limit theorem, etc) can be used to develop large-sample properties. But in most interesting cases, computation of the estimators is non-trivial. The likelihood could be too complicated to solve analytically, in which case, some numerical optimization procedure is needed. Popular techniques include the Newton–Raphson method and the Expectation–Maximization (EM) algorithm. Another challenge is that the likelihood may have many local maxima so, in such cases, there are uniqueness concerns. The point here is that we cannot, in general, rely on a formulae to derive asymptotic properties since, in most cases, there will be none. Instead we must rely completely on asymptotic properties of the likelihood function itself.

11.2 Likelihood-based Asymptotics

11.2.1 Consistency

Suppose $\hat{\theta}_n$ is some estimate—the MLE in this case—of the unknown parameter $\theta = \theta^\star$. Then the estimate is consistent if it converges to the true value when the sample size n increases to infinity.

Definition. (Consistency) An estimate $\hat{\theta}_n$ is consistent for $\theta^\star$ if $\hat{\theta}_n \to \theta^\star$ in $P_{\theta^\star}$-probability as $n \to \infty$. More precisely, $\hat{\theta}_n$ is consistent for $\theta^\star$ if

$$\lim_{n\to\infty} P_{\theta^\star}\big\{|\hat{\theta}_n - \theta^\star| > \epsilon\big\} = 0 \quad \forall\, \epsilon > 0.$$

The estimate is strongly consistent if convergence is in $P_{\theta^\star}$-probability 1.

Towards consistency of the MLE, we start with a preliminary results which says that, for any $\theta \neq \theta^\star$, $L_n(\theta^\star)$ exceeds $L_n(\theta)$ for all but finitely many n with probability 1.

Definition. If $A_n, n \geq 1$ is a sequence of subsets of Ω, then we define

$$\limsup_{n\to\infty} A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\omega \text{ that are in infinitely many } A_n\}$$

$$\liminf_{n\to\infty} A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m = \{\omega \text{ that are in all but finitely many } A_n\}.$$

It is common to write $\limsup A_n = \{\omega : \omega \in A_n \text{ i.o.}\}$.

Definition. (Almost Surely Convergence) Let $X_n \to X$ a.s. denote X_n converges to X almost surely. Then for all $\epsilon > 0$, $\Pr(|X_n - X| > \epsilon \text{ i.o.}) = 0$.

Our first key result is:

Lemma 6. *(Borel-Cantelli Lemma) If* $\sum_{n=1}^{\infty} \Pr(A_n) < \infty$, *then* $\Pr(A_n \text{ i.o.}) = 0$.

Proof. Let $N = \sum_{k=1}^{\infty} 1_{A_k}$ be the number of events that occur. $\Pr(A_n, \text{i.o}) = 0$ iff $\Pr(N < \infty) = 1$. By Fubini's theorem, $\mathbb{E}[N] = \sum_{k=1}^{\infty} \Pr(A_k) < \infty$. $\mathbb{E}[N] < \infty$ is equivalent to $\Pr(N < \infty) = 1$. So we must have $\Pr(A_n \text{ i.o.}) = 0$. □

Theorem 39. *Suppose* $X_1, \ldots, X_n$ *are iid with density* $p_{\theta^\star}$. *For any fixed* $\theta \neq \theta^\star$,

$$P_{\theta^\star}\big\{L_n(\theta) > L_n(\theta^\star) \text{ i.o.}\big\} = 0.$$

The intuition goes as follows. Let

$$R_n(X, \theta) = \frac{1}{n} \log \frac{L_n(\theta)}{L_n(\theta^\star)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}.$$

By the law of large numbers, $R_n(X, \theta) \to -K(\theta^\star, \theta)$ almost surely, where

$$K(\theta^\star, \theta) = \mathbb{E}_{\theta^\star}\Big\{\log \frac{p_{\theta^\star}(X)}{p_\theta(X)}\Big\} = \int \log \frac{p_{\theta^\star}(x)}{p_\theta(x)} p_\theta(x)\, d\mu(x)$$

is the Kullback-Leibler divergence of p_θ from $p_{\theta^\star}$, which is positive for for $\theta \neq \theta^\star$ by Jensen's inequality. Therefore, $R_n(X, \theta)$ should be negative for large n, so the event $\{L_n(\theta^\star) > L_n(\theta)\}$, which is equivalent to $\{R_n(X, \theta) < 0\}$, should have high probability.

Proof. Take any $\delta > 0$. Since $\mathbb{E}_{\theta^\star}\{L_n(\theta)/L_n(\theta^\star)\} = 1$, it follows from Markov's inequality that $P_{\theta^\star}\{R_n(X, \theta) \geq \delta\} = P_{\theta^\star}\{L_n(\theta)/L_n(\theta^\star) \geq e^{n\delta}\} \leq e^{-n\delta}$. Since $\sum_n e^{-n\delta} < \infty$, the Borel–Cantelli lemma implies that $P_{\theta^\star}\{R_n(X, \theta) \geq \delta \text{ i.o.}\} = 0$ for all $\delta > 0$. By letting $\delta \to 0$, it follows that $P_{\theta^\star}\{R_n(X, \theta) > 0 \text{ i.o.}\} = 0$. But the event $\{R_n(X, \theta) > 0\}$ is equivalent to $\{L_n(\theta) > L_n(\theta^\star)\}$, so we're done. □

Theorem 39 shows that $L_n(\theta^\star)$ is likely to be greater than $L_n(\theta)$ for any other fixed θ when n is large. This suggests that the MLE is consistent, but doesn't prove it, unless Θ is finite. The trouble is that the result only describes behavior of the likelihood ratio $R_n(X, \theta)$ for fixed θ. There are basically two approaches to get a consistency result. The first technique relaxes the definition of MLE consistency, only asking if there is a sequence of solutions to the likelihood equation that's consistent.

Theorem 40. *Let $X_1, \ldots, X_n$ are iid P_θ, and assume that $p_\theta(x)$ is the μ-density of P_θ, and that the support of P_θ does not depend on θ. Suppose that, for μ-almost all x, $p_\theta(x)$ is differentiable in $\theta \in \Theta_0$, with derivative $p'_\theta(x)$. Then, with probability tending to 1, there exists a consistent sequence of solutions $\hat{\theta}_n$ to the likelihood equation*

$$0 = \frac{\partial}{\partial \theta} \log L_n(\theta \mid X) = \sum_{i=1}^{n} \frac{p'_\theta(X_i)}{p_\theta(X_i)},$$

i.e., $\hat{\theta}_n \to \theta^\star$ in probability under $P_{\theta^\star}$ for all interior points $\theta^\star$ of Θ.

Proof. Take $a > 0$ small enough that $(\theta^\star - a, \theta^\star + a)$ is contained in the interior of Θ. Write $\ell_n(\theta \mid x) = \log L_n(\theta \mid x)$ and define

$$S_n(a) = \{x : \ell_n(\theta^\star \mid x) > \ell_n(\theta^\star - a \mid x) \text{ and } \ell_n(\theta^\star \mid x) > \ell_n(\theta^\star + a \mid x)\}.$$

Then for any $x \in S_n(a)$, there exists $\hat{\theta}_n(a) \in (\theta^\star - a, \theta^\star + a)$ at which $\ell_n(\theta)$ has a local maximum, i.e., $\ell'_n(\hat{\theta}_n(a)) = 0$. We also know by Theorem 39 that $P_{\theta^\star}(S_n(a)) \to 1$ as $n \to \infty$ for any fixed $a > 0$, and from this it follows that there exists $a_n \downarrow 0$ such that $P_{\theta^\star}\{S_n(a_n)\} \to 1$ as $n \to \infty$. Let $\hat{\theta}_n^+(x)$ equal $\hat{\theta}_n(a_n)$ if $x \in S_n(a_n)$ and equal to some arbitrary constant otherwise. Then,

$$P_{\theta^\star}\{\ell'_n(\hat{\theta}_n^+(X)) = 0\} \geq P_{\theta^\star}\{S_n(a_n)\} \to 1.$$

Therefore, for any $a > 0$ and n sufficiently large,

$$P_{\theta^\star}\{|\hat{\theta}_n^+(X) - \theta^\star| < a\} \geq P_{\theta^\star}\{|\hat{\theta}_n^+(X) - \theta^\star| < a_n\} \geq P_{\theta^\star}\{S_n(a_n)\} \to 1,$$

completing the proof. □

This theorem indicates that the process of estimating θ by solving the likelihood equation is a reasonable one in the sense that there will be a consistent solution. If the solution to the likelihood equation is unique, then that solution is the MLE and it's consistent. However, if the likelihood equations have more than one solution, then the theorem is useless because it doesn't say which sequence of solutions to pick. Wald's consistency (see optional section) gives another way of showing the result.

11.3 Wald's Consistency (Optional)

The second approach of proving the consistency is to show, directly, that the global maximizer of the likelihood function is consistent—this is a genuine consistency theorem for the MLE. One can imagine, however, that some stronger conditions are required to get this stronger result. In fact, what is needed is uniform control of fluctuations in $R_n(X, \theta)$. These kind of conditions are generally referred to as *Wald conditions* after the famous mathematician Abraham Wald.

Theorem 41. *Let $X_1, \ldots, X_n$ be iid with density p_θ wrt μ. Fix $\theta^\star$ and define, for each $B \subseteq \Theta$ and each $x \in \mathbb{X}$,*

$$Z(B, x) = \inf_{\theta \in B} \log \frac{p_{\theta^\star}(x)}{p_\theta(x)}. \tag{11.1}$$

Assume that for each $\theta \neq \theta^\star$, there is an open set B_θ such that $\theta \in B_\theta$ and $\mathbb{E}_{\theta^\star}\{Z(B_\theta, X)\} > 0$. If Θ is not compact, assume further that there exists compact $K \subseteq \Theta$ such that $\theta^\star \in K$ and $\mathbb{E}_{\theta^\star}\{Z(K^c, X)\} > 0$. Then $\hat{\theta}_n \to \theta^\star$ with $P_{\theta^\star}$-probability 1.

Proof. If Θ is compact, take $K = \Theta$. We shall prove that, for every $\epsilon > 0$,

$$P_{\theta^\star}\big\{\limsup_{n\to\infty} |\hat{\theta}_n - \theta^\star| \geq \epsilon\big\} = 0. \tag{11.2}$$

For fixed $\varepsilon > 0$, take B to be the open interval centered at $\theta^\star$ of length 2ε. Since $K \setminus B$ is a compact set, and $\{B_\theta : \theta \in K \setminus B\}$ is an open cover, we may extract a finite sub-cover, say, $B_{\theta_1}, \ldots, B_{\theta_k}$. For notational simplicity, rename K^c and these sets as $\Theta_1, \ldots, \Theta_m$, so that $\Theta = B \cup (\bigcup_{j=1}^m \Theta_j)$ and $\mathbb{E}_{\theta^\star} Z(\Theta_j, X) > 0$.

Write $c_j = \mathbb{E}_{\theta^\star} Z(\Theta_j, X)$. Then by the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n Z(\Theta_j, X_i) \to c_j, \quad \text{with } P_{\theta^\star}\text{-probability 1 for each } j.$$

Let A_j denote the set of data sequences for which this convergence holds and set $A = \bigcap_{j=1}^m A_j$. Then $P_{\theta^\star}(A) = 1$ and $n^{-1}\sum_{i=1}^n Z(\Theta_j, x_i) \to c_j > 0$ for all $x = (x_1, x_2, \ldots,) \in A$. If "i.o." stands for "infinitely often," then we have:

$$\begin{aligned}
\big\{x : \limsup_{n\to\infty} |\hat{\theta}_n(x_1, \ldots, x_n) - \theta^\star| \geq \epsilon\big\} &\subseteq \bigcup_{j=1}^m \big\{x : \hat{\theta}_n(x_1, \ldots, x_n) \in \Theta_j \text{ i.o.}\big\} \\
&\subseteq \bigcup_{j=1}^m \Big\{x : \inf_{\theta\in\Theta_j} \frac{1}{n}\sum_{i=1}^n \log \frac{p_{\theta^\star}(x_i)}{p_\theta(x_i)} \leq 0 \text{ i.o.}\Big\} \\
&\subseteq \bigcup_{j=1}^m \Big\{x : \frac{1}{n}\sum_{i=1}^n Z(\Theta_j, x_i) \leq 0 \text{ i.o.}\Big\} \\
&\subseteq \bigcup_{j=1}^m A_j^c.
\end{aligned}$$

Since the last set is A^c and $\mathsf{P}_{\theta^\star}(A^c) = 0$, the result (11.2) follows. □

The Wald-type theorem above is quite powerful, but the conditions are difficult to check. Here is one example for a uniform distribution, which is outside the nice regular exponential family.

Example. Suppose $X_1, \ldots, X_n$ are iid $U(0, \theta)$, so that $p_\theta(x) = \theta^{-1}$ for $0 \le x \le \theta$. In this case, the MLE is $\hat{\theta} = X_{(n)}$, the largest of the X_i's.

To apply Theorem 41, first observe that

$$\log \frac{p_{\theta^\star}(x)}{p_\theta(x)} = \begin{cases} \log(\theta/\theta^\star) & \text{if } x \le \min\{\theta, \theta^\star\} \\ \infty & \text{if } \theta \le x \le \theta^\star \\ -\infty & \text{if } \theta^\star < x \le \theta \\ \text{undefined} & \text{if } x > \max\{\theta, \theta^\star\}. \end{cases}$$

The last two cases have $P_{\theta^\star}$-probability zero, so we may choose $B_\theta = (\frac{\theta+\theta^\star}{2}, \infty)$ when $\theta > \theta^\star$. In this case

$$Z(B_\theta, x) = \log \frac{\theta + \theta^\star}{2\theta^\star} > 0 \quad \text{with } P_{\theta^\star}\text{-probability 1.}$$

When $\theta < \theta^\star$, choose $B_\theta = (\frac{\theta}{2}, \frac{\theta+\theta^\star}{2})$. In this case, $Z(B_\theta, x) = \infty$ if $x > (\theta + \theta^\star)/2$. Hence, $\mathsf{E}_{\theta^\star} Z(B_\theta, x) > 0$ in either case. We also need a compact set K such that $\mathsf{E}_{\theta^\star} Z(\Theta \setminus K, X) > 0$. Let $K = [\theta^\star/a, a\theta^\star]$ for some $a > 1$. Then

$$\inf_{\theta \in \Theta \setminus K} \log \frac{p_{\theta^\star}(X)}{p_\theta(X)} = \begin{cases} \log(X/a) & \text{if } X < \theta^\star/a \\ \log a & \text{if } X \ge \theta^\star/a. \end{cases}$$

Taking expectation we get

$$\frac{1}{\theta^\star}\Big(\int_0^{\theta^\star/a} \log(x/\theta^\star)\, dx + \int_{\theta^\star/a}^{\theta^\star} \log a\, dx \Big).$$

The first integral goes to 0 and the second goes to ∞ as $a \to \infty$. This means that there is some $a > 1$ such that the expectation is positive. It now follows from Theorem 41 that the MLE is consistent.

11.4 Some Examples

Example 63. (No unique MLE of Laplace distribution) The likelihood function is

$$L(\beta, \sigma) = \prod_{i=1}^{n} \left(\frac{\sigma}{2} e^{-\sigma |y_i - \beta|} \right),$$

while taking the log, we have $\ell(\beta, \sigma) = n \log \sigma - \sigma \sum_{i=1}^{n} |y_i - \beta|$.

To take the derivative of ℓ, we have

$$\frac{\partial |y_i - \beta|}{\partial \beta} = \begin{cases} -1 & \text{if } y_i > \beta \\ 1 & \text{if } y_i < \beta \\ \text{undefined} & \text{if } y_i = \beta \end{cases}$$

Therefore,

$$\frac{\partial \ell}{\partial \beta} = \#\{y_i : y_i < \beta\} - \#\{y_i : y_i > \beta\},$$

where $\#\{\cdot\}$ means the number of this event. Note that $\partial \ell / \partial \beta$ is zero if the same number of y_i are less than β as greater than β. If n is even, $\hat{\beta}$ will be any value between the $n/2$-th and $n/2+1$-th of the sorted values of y, and if n is odd, $\hat{\beta}$ is the middle value of that y.

Example 64. (MLE of Mixture normal) Mixture normal.

$$\begin{aligned} x \sim & \mathcal{N}(\mu_1, \sigma_1^2) \text{ with probability } p \\ x \sim & \mathcal{N}(\mu_2, \sigma_2^2) \text{ with probability } 1-p \end{aligned}$$

The the mixture density is

$$f(x; p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right].$$

Taking derivative w.r.t five parameters $p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, we will have a system of five equations. The solution of this system is the MLE of mixture normal model.

Here I just list two of five:

$$\begin{aligned} \frac{\partial \ln f}{\partial p} &= \sum_{i=1}^{n} \frac{1}{\Delta} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] - \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] \right) = 0 \\ \frac{\partial \ln f}{\partial \mu_1} &= \sum_{i=1}^{n} \frac{1}{\Delta} \left(\frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] \times \frac{(x-\mu_1)}{\sigma_1^2} \right) = 0 \end{aligned}$$

where $\Delta = \left(\frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] \right)$.

Chapter 12

Large Sample Theorem in MLE (II)

12.1 Asymptotic Efficiency

While consistency is obviously a desirable property, its practical use is limited in the sense that it is never a good idea to conclude that $\hat{\theta}_n = \theta^\star$, which amounts to assuming $n = \infty$. A better idea is to study the asymptotic fluctuations of a *normalized* version of $\hat{\theta}_n$, which allows for the construction of asymptotically correct confidence intervals, hypothesis tests, etc. The theorems presented below will pinpoint the asymptotic distributions of the MLE and posterior means. The next definition is relevant for this task.

Definition 1. (Asymptotic normality) A consistent estimator $\hat{\theta}_n$ of θ is asymptotically normal if, for some sequence deterministic c_n that converges to infinity, $c_n(\hat{\theta}_n - \theta) \to \mathcal{N}(0, v(\theta))$ in distribution, where $v(\theta) > 0$ for all θ.

In what follows, the sequence c_n will be the usual $n^{1/2}$; however, this is not the case in general. For example, there are certain problems where the rate is slower than root-n. For instance, in pointwise kernel density estimation at a boundary, c_n is $n^{1/3}$.

For the case with $c_n = n^{1/2}$, one could ask what the asymptotic variance $v(\theta)$ might be. This too can vary. However, there is something of a lower bound, called the *Cramer–Rao inequality*. Under some regularity conditions, the variance function $v(\theta)$ is no less than $I(\theta)^{-1}$, the inverse of the Fisher information. An estimator $\hat{\theta}_n$ with corresponding variance function $v(\theta)$ equal to the lower bound $I(\theta)^{-1}$ is called *asymptotically efficient*.

As we demonstrate below, the MLE is asymptotically efficient. But there is a general phenomenon called *super-efficiency* where, for certain θ values, $v(\theta) < I(\theta)^{-1}$. Fortunately, this somewhat problematic super-efficiency can occur only for θ in a topologically small set, in particular, a set of Lebesgue measure zero.

First, let's look at the Cramer-Rao lower bound.

Definition. An *unbiased estimator* is one such that $\mathbb{E}_\theta[\hat{\theta}] = \theta$.

Cramér and Rao proved that one can give a lower bound for the variance of unbiased estimators.

We work in terms of the log-likelihood $L(\theta \,|\, x) = \log \ell(\theta \,|\, x)$.

Definition. We define the *score* as $q(\theta \,|\, x) = \frac{\partial}{\partial\theta} L(\theta \,|\, x)$ and we define the *Hessian* as

$$H(\theta \,|\, x) = \frac{\partial^2}{\partial\theta^2} L(\theta \,|\, x).$$

Observe that $H(\theta \,|\, X)$ is a random variable which is a function of the random variable X, so that we can evaluate

$$I(\theta) = -\mathbb{E}_\theta\, H(\theta \,|\, X),$$

which R A Fisher define as the *information*.

Lemma 7. $\mathbb{E}_\theta[\, q(\theta \,|\, X)] = 0$.

Proof. From the definition

$$\begin{aligned}
\mathbb{E}_\theta\, q = \mathbb{E}_\theta \partial L/\partial\theta &= \int \{\partial(\log \ell)/\partial\theta\} p_\theta(x) dx = \int \{\partial(\log p)/\partial\theta\} p_\theta(x) dx \\
&= \int \{(\partial p/\partial\theta)/p\} p_\theta(x) dx = \int (\partial f/\partial\theta) dx \\
&= \frac{\partial}{\partial\theta} \int p(x) dx = \frac{\partial}{\partial\theta} 1 = 0.
\end{aligned}$$

since in any reasonable case it makes no difference whether differentiation with respect to θ is carried out inside or outside the integral with respect to x. □

Lemma 8. $I(\theta) = \mathbb{E}_\theta(\partial L/\partial\theta)^2$.

Proof. Again differentiating under the integral sign

$$\begin{aligned}
I(\theta) &= -\mathbb{E}_\theta \partial^2(\log \ell)/\partial\theta^2 = -\int \{\partial^2(\log p_\theta)/\partial\theta^2\} p_\theta(x) dx \\
&= -\int \frac{\partial}{\partial\theta} \left(\frac{\partial p_\theta/\partial\theta}{p_\theta} \right) p_\theta(x) dx \\
&= -\int \left(\frac{\partial^2 p_\theta/\partial\theta^2}{p_\theta} \right) p_\theta(x) dx + \int \left(\frac{(\partial p_\theta/\partial\theta)^2}{p_\theta^2} \right) p_\theta(x) dx \\
&= -\int (\partial^2 p_\theta/\partial\theta^2) dx + \int \{\partial(\log p_\theta)/\partial\theta\}^2 p_\theta(x) dx \\
&= -\frac{\partial}{\partial^2\theta} 1 + \int (\partial L/\partial\theta)^2 p_\theta(x) dx \\
&= \mathbb{E}_\theta(\partial L/\partial\theta)^2.
\end{aligned}$$

□

Lemma 9. *Var*$[\, q(\theta \,|\, X)] = I(\theta)$.

Proof. Immediate since $\mathbb{E}_\theta\, q(\theta \,|\, X) = 0$ and so $\mathrm{Var}[\, q(\theta \,|\, X)] = \mathbb{E}_\theta\, (q(\theta \,|\, X)^2$. □

Lemma 10. *The covariance Cov$(\hat{\theta},\, q)$ of $\hat{\theta}$ and $q(\theta \,|\, X)$ is unity.*

Proof. We note that since $\mathbb{E}_\theta[\hat{\theta}] = \theta$, that is,

$$\theta = \int \hat{\theta} p_\theta(x) dx$$

we can differentiate with respect to θ to get

$$\begin{aligned} 1 &= \frac{\partial}{\partial\theta} \int \hat{\theta}\, p_\theta(x) dx = \int \hat{\theta} \left(\frac{\partial p_\theta / \partial\theta}{p_\theta} \right) p_\theta(x) dx = \int \hat{\theta}\, q_\theta(x) dx \\ &= \mathbb{E}_\theta[\hat{\theta}\, q] \end{aligned}$$

so that as $\mathbb{E}_\theta\, \hat{\theta} = \theta$ and $\mathbb{E}_\theta\, q = 0$

$$\mathrm{Cov}(\hat{\theta},\, q) = \mathbb{E}_\theta(\hat{\theta} - \theta)(q - 0) = \mathbb{E}_\theta[\,\hat{\theta}\,(q - \theta)]\, \mathbb{E}_\theta[\,q] = 1.$$

□

The required bound for the variance of $\hat{\theta}$ now follows simply:

Theorem 42. *(Cramér-Rao bound) The variance of an unbiased estimator $\hat{\theta}$ satisfies*

$$\mathit{Var}\, \hat{\theta} \geqslant \frac{1}{I(\theta)}.$$

Proof. We simply use the well-known inequality

$$\mathrm{Cov}(U, V)^2 \leqslant (\mathrm{Var}\, U)(\mathrm{Var}\, V)$$

with $U = \hat{\theta}$, $V = q(\theta \,|\, X)$, so that by Lemma 9 we have $\mathrm{Var}\, V = I(\theta)$ and by Lemma 10 we have $\mathrm{Cov}(U, V) = 1$. □

12.2 Likelihood-based Asymptotics Normality

Theorem 43. *(Slutsky's Theorem): Consider random variables X_n, Y_n, and X, such that X_n converges in distribution to X and Y_n converges in probability to a constant c with probability 1, then (a) $X_n + Y_n$ converges in distribution to $X + c$. (b) X_n / Y_n converges in distribution to X/c.*

Theorem 44. *Let $X_1, \ldots, X_n$ are iid P_θ, and assume that $p_\theta(x)$ is the μ-density of P_θ, and that the support of P_θ does not depend on θ. Suppose that, for μ-almost all x, $p_\theta(x)$ is differentiable in $\theta \in \Theta_0$, with derivative $p'_\theta(x)$.*

Let $\Theta \subseteq \mathbb{R}$, and $\hat{\theta}_n$ a consistent sequence of solutions to the likelihood equation. Assume $p_\theta(x)$ has continuous second partial derivatives wrt θ and that differentiation can be passed

under the integral sign. Assume that there exists a function $g_r(x,\theta)$ such that, for each interior point $\theta^\star$,

$$\sup_{\theta:|\theta-\theta^\star|\le r}\Big|\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)-\frac{\partial^2}{\partial\theta^2}\log p_{\theta^\star}(x)\Big|\le g_r(x,\theta^\star), \tag{12.1}$$

with $\lim_{r\to 0}\mathbb{E}_\theta\{g_r(X,\theta)\}=0$ for each θ. Assume the Fisher information $I(\theta)$ exists and is positive. Then under $P_{\theta^\star}$,

$$n^{1/2}(\hat\theta_n-\theta^\star)\to\mathcal{N}(0,I(\theta^\star)^{-1}),\quad \textit{in distribution.}$$

Proof. Let $\ell_n(\theta\mid x)=n^{-1}\log L_n(\theta\mid x)$ be the scaled log-likelihood. We assume $\theta^\star$ is in the interior of Θ, so there exists an open neighborhood of $\theta^\star$ also in the interior of Θ. By the assumed consistency of $\hat\theta_n$, the event that $\hat\theta_n$ is in this open neighborhood of $\theta^\star$ has probability converging to 1. Therefore, it suffices to consider the behavior of $\hat\theta_n$ only when it is in this open neighborhood where the log-likelihood is well-behaved, in particular, $\ell_n'(\hat\theta_n)=0$. Next, take a first-order (linear) Taylor approximation of $\ell_n'(\hat\theta_n)$ around $\theta^\star$:

$$0=\ell_n'(\theta^\star)+\ell_n''(\tilde\theta_n)(\hat\theta_n-\theta^\star),\quad (\text{for }\hat\theta_n\text{near }\theta^\star),$$

where $\tilde\theta_n$ is between $\hat\theta_n$ and $\theta^\star$. Then we get

$$n^{1/2}(\hat\theta_n-\theta^\star)=-\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\tilde\theta_n)},\quad (\text{for }\hat\theta_n\text{near }\theta^\star).$$

So, it remains to show that the right-hand side above has the stated asymptotically normal distribution. Let's look at the numerator and denominator separately.

Numerator. The numerator can be written as

$$n^{1/2}\ell_n'(\theta^\star)=n^{1/2}\cdot\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\theta}\log p_\theta(X_i)\Big|_{\theta=\theta^\star}.$$

The summands are iid with mean zero and variance $I(\theta^\star)$, by our assumptions about interchanging derivatives and integrals. Therefore, the standard Central Limit Theorem says that $n^{1/2}\ell_n'(\theta^\star)\to\mathcal{N}(0,I(\theta^\star))$ in distribution. (Delta method)

Denominator. The claim is that $-\ell_n''(\tilde\theta_n)\to I(\theta^\star)$ in $P_{\theta^\star}$-probability. If we can show this, then it follows from Slutsky's Theorem that

$$-\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\tilde\theta_n)}\to\frac{\mathcal{N}(0,I(\theta^\star))}{I(\theta^\star)}=\mathcal{N}(0,I(\theta^\star)^{-1}),\quad \text{in distribution,}$$

which is the desired result. The key to showing this is to first recognize that $-\ell_n''(\theta^\star)\to I(\theta^\star)$ in probability by the law of large numbers. So we can write $-\ell_n''(\tilde\theta_n)=-\ell_n(\theta^\star)+\Delta_n$, where

$$\Delta_n=\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\Big|_{\theta=\theta^\star}-\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\Big|_{\theta=\tilde\theta_n}.$$

By our assumptions, we have $|\Delta_n| \leq n^{-1}\sum_{i=1}^n g_r(X_i, \theta^\star)$ when $|\theta^\star - \tilde{\theta}_n| \leq r$. By the weak law of large numbers, $n^{-1}\sum_{i=1}^n g_r(X_i, \theta^\star) \to m_r(\theta^\star) := \mathbb{E}_{\theta^\star} g_r(X, \theta^\star)$ in probability. Given $\varepsilon > 0$, choose $r > 0$ small enough that $m_r(\theta^\star) < \epsilon/2$. Then

$$\begin{aligned} P_{\theta^\star}\{|\Delta_n| > \epsilon) &\leq P_{\theta^\star}\Big\{\frac{1}{n}\sum_{i=1}^n g_r(X_i, \theta^\star) > \epsilon\Big\} + P_{\theta^\star}\{|\theta^\star - \tilde{\theta}_n| \geq r\} \\ &\leq P_{\theta^\star}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n g_r(X_i, \theta^\star) - m_r(\theta^\star)\Big| > \frac{\epsilon}{2}\Big\} + P_{\theta^\star}\{|\theta^\star - \tilde{\theta}_n| \geq r\}. \end{aligned}$$

The last two inequalities go to zero, as $n \to \infty$: the first by the LLN, and the second by the fact that $\tilde{\theta}_n$ is consistent.[1] Therefore, $-\ell_n''(\tilde{\theta}_n) \to I(\theta^\star)$ in probability, as was to be shown. So we're done. □

When using these theorems on asymptotic normality, it is common to replace $I(\theta^\star)$ with a quantity that does not depend on the unknown parameter. Standard choices are the *expected* Fisher information $I(\hat{\theta}_n)$ and the *observed* Fisher information $-\ell_n''(\hat{\theta}_n)$, the negative Hessian.

Remark. Note that the Delta Theorem is actually more general, showing how to create new central limit theorems from existing ones; that is, the Delta Theorem is not specific to MLEs, etc.

12.2.1 Generalizations: empirical processes, M-estimators, etc

We can see that maximum likelihood estimation is done by maximizing a data-dependent function or, sometimes equivalently, finding a root of some other data-dependent function. In this light, one might think we could construct estimators by applying these ideas to other functions besides the likelihood or log-likelihood. It turns out that this is indeed possible. Estimators found by optimizing a data-dependent function are generically referred to as M-estimators.

Things are more interesting when the function has no formula for its optimizer. The general theory involves some rather sophisticated mathematics, namely, convergence of random functions. Important examples of these kinds of random functions are called *empirical processes*. These are general versions of the well-known empirical distribution function.

12.3 Bayesian analysis

12.3.1 Introduction

Define a sample (measurable) space $(\mathbb{X}, \mathcal{A})$ which is equipped with a family of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Suppose that there exists a σ-finite measure μ such that $P_\theta \ll \mu$ for all θ, so that we have Radon–Nikodym derivatives (densities) $p_\theta(x) = (dP_\theta/d\mu)(x)$

[1] $\tilde{\theta}_n \to \theta^\star$ since $\tilde{\theta}_n$ is between $\hat{\theta}_n$ and $\theta^\star$, and $\hat{\theta}_n \to \theta^\star$.

w.r.t. μ. Bayesians assume that some probability distribution Π on Θ is also available from somewhere. We call Π the *prior distribution*. Let the notation U stand for a random variable distributed according to Π, and u for the observed values. (When it relates to inference about the parameter, we go back to the θ notation.)

The Bayesian setup assumes the following hierarchical model:

$$U \sim \Pi \quad \text{and} \quad X \mid (U = u) \sim p_u(x). \tag{12.2}$$

The goal is to take the information from the observed $X = x$ and update the prior information about the "parameter" U. This is accomplished quite generally via Bayes' theorem. But before seeing the technical stuff, it helps to understand the reasoning behind this particular choice. If uncertainty about θ is described by the (subjective) probability distribution Π, then the uncertainty about θ *after seeing data* x should be described by the conditional distribution Π_x, the *posterior* distribution of U given $X = x$. This posterior distribution is used for inference.

A general measure-theoretic version Bayes theorem.

Theorem 45. *(Bayes Theorem) Under the setup described above, let Π_x denote the conditional distribution of U given $X = x$. Then $\Pi_x \ll \Pi$ for P_Π-almost all x, where $P_\Pi = \int P_u \, d\Pi(u)$ is the marginal distribution of X from model (12.2). Also, the Radon–Nikodym derivative of Π_x with respect to Π is*

$$\frac{d\Pi_x}{d\Pi}(u) = \frac{p_u(x)}{p_\Pi(x)},$$

for those x such that the marginal density $p_\Pi(x) = (dP_\Pi/d\mu)(x)$ is neither 0 nor ∞. Since the set of all x such that $p_\Pi(x) \in \{0, \infty\}$ is a P_Π-null set, the Radon–Nikodym derivative can be defined arbitrarily for such x.

The theorem provides a formula for the probability $P(A \mid B)$ in terms of the opposition conditional probability $P(B \mid A)$ and the marginal probabilities $P(A)$ and $P(B)$.

Remark. The posterior mean is defined as

$$\hat{\theta}_{\text{mean}} = \mathbb{E}(U|X = x) = \int u \, d\Pi_x(u),$$

and, in the case where Π_x has a density π_x, the posterior mode is defined as

$$\hat{\theta}_{\text{mode}} = \arg\max_u \pi_x(u),$$

which is similar to the maximum likelihood estimate.

12.3.2 Posterior consistency

From a Bayesian point of view, one can consider the same kind of consistency for the posterior mean or mode or whatever. But Bayesians have a complete posterior distribution Π_n supported on Θ to work with and they can consider asymptotic properties of the distribution itself, not just that of functionals.

Definition 2. (Posterior consistency) The posterior is said to be consistent at $\theta^\star$ if, with $P_{\theta^\star}$-probability 1, $\Pi_n(\Theta^\star) \to 1$ for any open neighborhood $\Theta^\star$ of $\theta^\star$.

Remark. Posterior consistency says that the posterior will assign arbitrarily large probability to any arbitrarily small neighborhood of the true $\theta^\star$, provided n is large enough. This suggests that the posterior mean or mode will also be close to $\theta^\star$.

Theorem 46. *(Doob's theorem) Let $\mathbb{X}$ and Θ be complete metric spaces equipped with their respective Borel σ-algebras, and assume $\theta \mapsto P_\theta$ is one-to-one. For a given prior Π, there exists a set $\Theta_0 \subseteq \Theta$ such that $\Pi(\Theta_0) = 1$ and the posterior Π_n is consistent at any $\theta^\star \in \Theta_0$.*

Remark. Doob's theorem is quite general. It says that there is a set with prior probability 1 on which the posteriors are consistent. The trouble is that there is typically only one parameter value—the "true value" $\theta^\star$—that's of interest, and simply knowing that consistency holds on a set of prior probability 1 generally says nothing about consistency at $\theta^\star$. It is in this sense that Doob's theorem can be unsatisfactory, so stronger theorems are desirable.

Chapter 13

Hypothesis Testing

13.1 Likelihood Ratio Test

Where do statistical tests come from? A particular discrepancy measure may seem sensible as a way to capture the relevant departure from H_0. AS MLE is very widely applicable to parametric estimation problems, the likelihood ratio test is very widely applicable to parametric testing problems.

13.1.1 Test $H_0 : \theta = \theta_0$

The likelihood function assigns to alternative values of θ their plausability in $L(\theta)$. We consider the value $L(\theta_0)$ and assess whether it is nearly the same as the maximal value $L(\hat{\theta})$. For a random sample $X_1, \ldots, X_n$, we may examine the likelihood ratio

$$LR = \frac{p_{\theta_0}(X_1, \ldots, X_n)}{p_{\hat{\theta}}(X_1, \ldots, X_n)}$$

and see how small it is. Because the MLE maximizes the likelihood function, we have $LR \leq 1$.

For a random sample $X_1, \ldots, X_n$ with joint pdf $p_\theta(x_1, \ldots, x_n)$, the likelihood ratio test of $H_0 : \theta = \theta_0$ evaluates the observed likelihood ratio statistic

$$LR_{obs} = \frac{p_{\theta_0}(x_1, \ldots, x_n)}{p_{\hat{\theta}}(x_1, \ldots, x_n)}$$

and assigns the p-value

$$p = \Pr\left(\frac{p_{\theta_0}(X_1, \ldots, X_n)}{p_{\hat{\theta}}(X_1, \ldots, X_n)} < LR_{obs}\right) \tag{13.1}$$

computed under the assumption that $H_0 : \theta = \theta_0$ is satisfied, i.e., the assumption that $X_1, \ldots, X_n$ have pdf $p_{\theta_0}(X_1, \ldots, X_n)$.

Note that it is equivalent to examine the log of the likelihood ratio: in (13.1) we may take logs to get

$$p = \Pr\left(\log\left[\frac{p_{\theta_0}(X_1,\ldots,X_n)}{p_{\hat{\theta}}(X_1,\ldots,X_n)}\right] < \log LR_{obs}\right).$$

As when maximizing a likelihood function, taking logs generally simplifies the expression.

Example 65. Suppose $X \sim B(n,q)$ and we wish to test $H_0 : q = q_0$. We would have $q_0 = .5$ and $\hat{q} = x/n$, with $n = 17$ and $x = 14$. The PDF is

$$p_q(x) = \binom{n}{x} q^x (1-q)^{n-x}$$

and the observed likelihood ratio statistic is

$$\begin{aligned} LR_{obs} &= \frac{q_0^x(1-q_0)^{n-x}}{\hat{q}^x(1-\hat{q})^{n-x}} \\ &= \frac{1}{2^n(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \\ &= \frac{1}{2^n(\frac{14}{17})^{14}(1-\frac{14}{17})^3}. \end{aligned}$$

The negative log likelihood ratio becomes

$$\begin{aligned} -\log LR_{obs} &= n\log 2 + x\log\frac{x}{n} + (n-x)\log(1-\frac{x}{n}) \\ &= 17\log 2 + 14\log\frac{14}{17} + 3\log(1-\frac{14}{17}). \end{aligned}$$

The advantage of the likelihood ratio test is that it gives a specific method that can be applied in many, many problems and, furthermore, like ML estimation, it turns out to have very good properties in large samples.

Example 66. (Test of $H_0 : (\omega,\theta) = (\omega,\theta_0)$) We now consider the case in which the parameter vector may be decomposed into two sub-vectors ω and θ, having respective dimensions m_1 and m_2. For example, in linear regression we would have a parameter vector (β_0,β_1) and we might decompose it as $\omega = \beta_0$ and $\theta = \beta_2$. We consider null hypotheses of the form $H_0 : \theta = \theta_0$ which now becomes a short-hand for $H_0 : (\omega,\theta) = (\omega,\theta_0)$. In linear regression, for example, we might consider whether there is a non-zero slope by introducing $H_0 : \beta_1 = 0$. This is short for $H_0 : (\beta_0,\beta_1) = (\beta_0,0)$, which means simply that H_0 does not put any restriction on $\omega = \beta_0$. A wide variety of statistical models that are submodels of larger models may be written in this form.

Result Under fairly general conditions, for large samples, if θ is a vector of length m then $-2\log LR$ has an approximate $\chi^2_{m_2}$ distribution, so that an approximate p-value may be obtained from the chi-squared distribution with m_2 degrees of freedom.

Remark. The likelihood ratio test may be used to derive the t test, and also other standard tests used in common situations. For testing independence of two traits, the likelihood ratio test is approximately equivalent to the χ^2 test of independence in large samples, meaning that in large samples it gives very nearly the same p-value as the χ^2 test of independence.

13.1.2 The likelihood ratio test is optimal for simple hypotheses.

Consider both H_0 and H_A : $H_0 : X \sim f(x)$ and $H_A : X \sim g(x)$ and consider the problem of testing H_0 versus the alternative H_A. This is often called the case of "simple versus simple" hypotheses. The likelihood ratio may be written

$$LR(x) = \frac{f(x)}{g(x)}.$$

Note that the likelihood ratio test will reject H_0 when $LR(x)$ is sufficiently small (which is equivalent to $-\log LR(x)$ being sufficiently large). In other words, the likelihood ratio test will reject H_0 when $LR(x) < c$ for some suitable number c. The type I error is

$$\alpha_{LR} = \Pr(LR(X) < c|H_0)$$

and the power is

$$\text{Power}_{LR} = \Pr(LR(X) < c|H_A).$$

Lemma 11. ***(Neyman-Pearson Lemma)*** *Let α be a positive number less than 1 and let $c = c(\alpha)$ be chosen so that*

$$\alpha_{LR} = \alpha.$$

Let $T(X)$ be another test statistic having type I error α_T such that

$$\alpha_T \leq \alpha.$$

Then the power of these two tests satisfies

$$Power_{LR} \geq Power_T.$$

Remark. Neyman-Pearson lemma says that the likelihood ratio test is the optimal test, in the sense of power, for testing H_0 versus H_A. More generally, likelihood ratio tests may be shown to be optimal for large samples (see Section 16.6 of van der Vaart, 1998).

13.2 The Neyman-Pearson Lemma (statistical decision version)

On a measurable space Ω two probabilities P_θ are given, $\theta = 0, 1$. After observing an outcome $x \in \Omega$ one action a is chosen between two possible values $a = 0$ or $a = 1$. A loss is associated to this decision:

if $\theta = 0$ and the action $a = 1$ is chosen, then the loss is $k_0 > 0$;

if $\theta = 1$ and the action $a = 0$ is chosen, then the loss is $k_1 > 0$;

in any other case the loss is zero.

Given an observed value x, the choice of the action is randomized: $a = 1$ is chosen with probability $\varphi(x)$, $a = 0$ is chosen with probability $1 - \varphi(x)$. The function $\varphi : \Omega \to [0, 1]$ is

called *test function* and it will be assumed to be measurable. It specifies the *decision rule* (or *strategy*). We define the *expected loss*

$$L_\varphi(\theta, x) = \begin{cases} k_0\varphi(x), & \text{if } \theta = 0, \\ k_1(1-\varphi(x)), & \text{if } \theta = 1, \end{cases}$$

and the *risk*

$$R(\theta, \varphi) = \int_\Omega L_\varphi(\theta, x)\, P_\theta(dx).$$

We denote by ν a finite measure such that $P_\theta << \nu$, $\theta = 0, 1$. We may take $\nu = P_0 + P_1$. Let f_θ denote the density $dP_\theta/d\nu$. Then

$$R(0, \varphi) = k_0 \int_\Omega \varphi f_0\, d\nu, \qquad R(1, \varphi) = k_1 \int_\Omega (1-\varphi) f_1\, d\nu.$$

The goal is to choose φ in a way to minimize the risk. We say that φ^* dominates φ if

$$R(0, \varphi^*) \le R(0, \varphi), \qquad R(1, \varphi^*) \le R(1, \varphi),$$

and at least one of the inequalities is strict. Strategies φ and φ^* are identified if $\varphi = \varphi^*$, $P_0 + P_1$-a.s.

We will denote $\mathcal{C}$ the following class of strategies.

(i) for $k \in (0, \infty)$,

$$\varphi_k(x) = \begin{cases} 1 & \text{if } f_1(x) > k f_0(x), \\ \text{arbitrary} & \text{if } f_1(x) = k f_0(x), \\ 0 & \text{if } f_1(x) < k f_0(x); \end{cases}$$

(ii) for $k = 0$,

$$\varphi_0(x) = \begin{cases} 1 & \text{if } f_1(x) > 0, \\ 0 & \text{if } f_1(x) = 0; \end{cases}$$

(iii) for $k = \infty$,

$$\varphi_\infty(x) = \begin{cases} 1 & \text{if } f_0(x) = 0, \\ 0 & \text{if } f_0(x) > 0. \end{cases}$$

Strictly speaking, if $k \in (0, \infty)$, each φ_k is a whole set of functions.

Theorem 47. *(Neyman-Pearson) For every $\varphi \notin \mathcal{C}$, there exists $\varphi^* \in \mathcal{C}$ that dominates φ. No element of $\mathcal{C}$ dominates any other element of $\mathcal{C}$.*

The Theorem will be a consequence of the following Propositions.

Proposition 15. *If $R(0, \varphi) \le R(0, \varphi^*)$, $R(1, \varphi) \le R(1, \varphi^*)$, and $\varphi^* \in \mathcal{C}$, then $\varphi = \varphi^*$, $P_0 + P_1$-a.s. In particular, $R(0, \varphi) = R(0, \varphi^*)$, $R(1, \varphi) = R(1, \varphi^*)$ and $\varphi \in \mathcal{C}$.*

Proposition 16. *If $0 \le \alpha \le R(0, \varphi_0)$, then there exists $\varphi^* \in \mathcal{C}$ such that $R(0, \varphi^*) = \alpha$.*

Proof of the Theorem.

Let φ be a strategy, $\varphi \notin \mathcal{C}$. If $R(0,\varphi) \le R(0,\varphi_0)$, then by Proposition 16 there exists $\varphi^* \in \mathcal{C}$ such that $R(0,\varphi^*) = R(0,\varphi)$. If we had, in addition, $R(1,\varphi^*) \ge R(1,\varphi)$, then by Proposition 15 we would conclude that $\varphi \in \mathcal{C}$, which is false. Therefore we must have $R(1,\varphi^*) < R(1,\varphi)$, and this implies that φ^* dominates φ.

If $R(0,\varphi) > R(0,\varphi_0)$, then φ_0 dominates φ, since

$$R(1,\varphi_0) = k_1 \int_\Omega (1-\varphi_0) f_1 \, d\nu = 0 \le R(1,\varphi).$$

The last statement follows immediately from Proposition 15.

Proof of Proposition 15. Notice that

$$\left\{ \begin{array}{l} R(0,\varphi) \le R(0,\varphi^*), \\ R(1,\varphi) \le R(1,\varphi^*). \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \int_\Omega \varphi f_0 \, d\nu \le \int_\Omega \varphi^* f_0 \, d\nu, \\ \int_\Omega (1-\varphi) f_1 \, d\nu \le \int_\Omega (1-\varphi^*) f_1 \, d\nu. \end{array} \right.$$

so that

$$\left\{ \begin{array}{l} \int_\Omega (\varphi^* - \varphi) f_0 \, d\nu \ge 0, \\ \int_\Omega (\varphi^* - \varphi) f_1 \, d\nu \le 0. \end{array} \right. \tag{13.2}$$

Before proceeding, let us notice that $f_0 + f_1 > 0$, $P_0 + P_1$-a.s. Indeed,

$$(P_0+P_1)\{f_0+f_1=0\} = \int_{\{f_0+f_1=0\}} d(P_0+P_1) = \int_{\{f_0+f_1=0\}} (f_0+f_1)\, d\nu = 0.$$

φ^* is one of the functions φ_k, for some $k \in [0,\infty]$. We distinguish three cases.

First case: $k \in (0,\infty)$**.**

From (13.2) we obtain

$$\int_\Omega (\varphi^* - \varphi)(f_1 - k f_0)\, d\nu \le 0. \tag{13.3}$$

If $f_1 > k f_0$ then $\varphi^* = 1$ and $\varphi^* \ge \varphi$. If $f_1 < k f_0$ then $\varphi^* = 0$ and $\varphi^* \le \varphi$. So the integrand function in (13.3) is nonnegative, and it follows that $(\varphi^* - \varphi)(f_1 - k f_0) = 0$, ν-a.s. So on the set $\{f_1 \ne k f_0\}$ we have $\varphi = \varphi^*$, ν-a.s. We conclude that φ is one of the functions φ_k.

Second case: $k = 0$**.**

In this case $\varphi^* = \varphi_0$. Then $\varphi^* = 1$ on $\{f_1 > 0\}$. It follows from (13.2) that

$$\int_{\{f_1>0\}} (1-\varphi) f_1 \, d\nu \le 0.$$

Since the integrand function is nonnegative, we deduce that $\varphi = 1 = \varphi^*$ on $\{f_1 > 0\}$, ν-a.s. By (13.2) again,

$$0 \le \int_{\{f_1=0\}} (\varphi^* - \varphi) f_0 \, d\nu = \int_{\{f_1=0\}} (-\varphi f_0)\, d\nu.$$

So we have $\varphi f_0 = 0$, ν-a.s. on $\{f_1 = 0\}$. Since $f_0 + f_1 > 0$, $P_0 + P_1$-a.s., then $\varphi = 0$, $P_0 + P_1$-a.s. on $\{f_1 = 0\}$ and we conclude that $\varphi = \varphi^*$, $P_0 + P_1$-a.s.

Third case: $k = \infty$**.**

In this case $\varphi^* = \varphi_\infty$. Then $\varphi^* = 0$ on $\{f_0 > 0\}$. It follows from (13.2) that

$$\int_{\{f_0>0\}} (-\varphi f_0)\, d\nu \geq 0.$$

We deduce that $\varphi = 0 = \varphi^*$, ν-a.s. on $\{f_0 > 0\}$. By (13.2) again,

$$0 \geq \int_{\{f_0=0\}} (\varphi^* - \varphi) f_1\, d\nu = \int_{\{f_0=0\}} (1-\varphi) f_1\, d\nu.$$

So we have $(1-\varphi)f_1 = 0$, ν-a.s. on $\{f_0 = 0\}$. Since $f_0 + f_1 > 0$, $P_0 + P_1$-a.s., then $\varphi = 1$, $P_0 + P_1$-a.s. on $\{f_0 = 0\}$ and we conclude that $\varphi = \varphi^*$, $P_0 + P_1$-a.s.

Proof of Proposition 16. If $\alpha = 0$, then $R(0, \varphi_\infty) = k_0 \int_\Omega \varphi_\infty f_0\, d\nu = 0 = \alpha$. So we will assume $0 < \alpha < R(0, \varphi_0)$. Let us define the function

$$g(k) = k_0 \int_{\{f_1 > k f_0\}} f_0\, d\nu, \qquad k \in [0, \infty).$$

Notice that if we set

$$\varphi_k(x) = \begin{cases} 1 & \text{if } f_1(x) > k f_0(x), \\ 0 & \text{if } f_1(x) \leq k f_0(x), \end{cases} \tag{13.4}$$

then $g(k) = R(0, \varphi_k)$. It is readily verified that g is decreasing on $[0, \infty)$ and

$$\lim_{k\to\infty} g(k) = 0, \qquad g(0) = k_0 \int_{\{f_1>0\}} f_0\, d\nu = R(0, \varphi_0) > \alpha.$$

Moreover, for any sequence k_n, with $k_n \neq k$, the implications

$$k_n \downarrow k \geq 0 \;\Rightarrow\; \{f_1 > k_n f_0\} \uparrow \{f_1 > k f_0\}, \quad k_n \uparrow k > 0 \;\Rightarrow\; \{f_1 > k_n f_0\} \downarrow \{f_1 \geq k f_0\} \cap \{f_1 > 0\}$$

show that g is right-continuous on $[0, \infty)$ and at any possible point of discontinuity $k > 0$ it has a jump equal to

$$\begin{aligned} &g(k^-) - g(k) = k_0 \textstyle\int_{\{f_1 \geq k f_0\} \cap \{f_1 > 0\}} f_0\, d\nu - k_0 \int_{\{f_1 > k f_0\}} f_0\, d\nu \\ &= k_0 \textstyle\int_{\{f_1 \geq k f_0\}} f_0\, d\nu - k_0 \int_{\{f_1 > k f_0\}} f_0\, d\nu = k_0 \int_{\{f_1 = k f_0\}} f_0\, d\nu. \end{aligned}$$

If there exists $k > 0$ such that $g(k) = \alpha$, then $\alpha = R(0, \varphi_k)$ and the proof is finished. Otherwise there exists $k > 0$ such that $g(k^-) \geq \alpha > g(k)$. Let us define

$$\varphi'_k(x) = \begin{cases} 1 & \text{if } f_1(x) > k f_0(x), \\ c & \text{if } f_1(x) = k f_0(x), \\ 0 & \text{if } f_1(x) < k f_0(x), \end{cases}$$

where $c \in [0, 1]$ is a suitable constant to be determined in such a way that $R(0, \varphi'_k) = \alpha$, i.e. we wish the following equality to hold:

$$\alpha = R(0, \varphi'_k) = k_0 \int_{\{f_1 > k f_0\}} f_0\, d\nu + k_0 \int_{\{f_1 = k f_0\}} c\, f_0\, d\nu = g(k) + c(g(k^-) - g(k)).$$

It suffices to choose $c = (\alpha - g(k))/(g(k^-) - g(k)) \in (0, 1]$.