Mathematical Methods for Statistics
STAT:5120 (22S:190)

Luke Tierney

Spring 2016

Week 1

Wednesday, January 20, 2016

Course Outline

- This course covers the basics of mathematical analysis at a rigorous level.
- The basic results of this are heavily used in statistics research.
- Methods of careful reasoning and proof are also essential to statistics research.
- Readings will be assigned for each class; you should read these *before* class.
- In class we will discuss the material in the readings, review key points, and work problems.

Basic Set Theory

- Basic set operations
  - unions, intersections
  - set difference, symmetric difference, complements
  - complement notation: E^c
  - commutative, distributive properties,
  - De Morgan's laws
- Functions and mappings
  - domain and range
  - image and inverse image (or preimage)
  - basic properties, like

$$f^{-1}\left(\bigcup_\alpha A_\alpha\right) = \bigcup_\alpha f^{-1}(A_\alpha)$$

- Equivalence relations
  - reflexivity, symmetry, transitivity
  - equivalence classes

Basic Properties of Real Numbers

- The real numbers $\mathbb{R}$ are an ordered set.
- The real numbers $\mathbb{R}$ form a field.
- The real numbers $\mathbb{R}$ have the upper bound property:
  - For any set $E \subset \mathbb{R}$ that is not empty and bounded above there exists a number $\alpha \in \mathbb{R}$ such that:

    (i) α is an upper bound on E.

    (ii) if $\gamma < \alpha$ then γ is not an upper bound on E.

    α is the unique number with these properties.
  - α is the *least upper bound* of E, or the *supremum* of E,

    $$\alpha = \sup E.$$

  - A non-empty set F bounded below has a *greatest lower bound*, or *infimum*, β, denoted

    $$\beta = \inf F.$$

- The rational numbers $\mathbb{Q}$ are a *dense subset* of the reals:
  - For and real numbers $a < b$ there exists a rational number q such that $a < q < b$.
- The *extended real numbers* are

  $$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}.$$

  Conventions:

  (a) If $x \in \mathbb{R}$ then

  $$x + \infty = \infty \qquad x - \infty = -\infty \qquad \frac{x}{\infty} = \frac{x}{-\infty} = 0.$$

  (b) If $x > 0$, then $x \cdot (+\infty) = \infty$ and $x \cdot (-\infty) = -\infty$.

  (c) If $x < 0$, then $x \cdot (+\infty) = -\infty$ and $x \cdot (-\infty) = +\infty$.

  Sometimes it is also useful to define $0 \cdot (\pm\infty) = 0$.

Finite, Countable, and Uncountable Sets

- Two sets A and B are *equivalent* if there exists a one-to-one mapping from A onto B.
- Equivalence of sets is an *equivalence relation*.
- Two finite sets are equivalent if and only if they have the same number of elements.
- A set is *countably infinite* if it is equivalent to the set of positive integers.
- Rudin uses countable to mean countably infinite, and *at most countable* to mean finite or countably infinite.
- An infinite set that is not countable is said to me *uncountable*.
- Every subset of a countably infinite set is at most countable.
- If A is at most countable and for every $\alpha \in A$ the set B_α is at most countable, then the union
$$\bigcup_{\alpha \in A} B_\alpha$$
is at most countable.
- The set $\mathbb{Q}$ of rational numbers is countably infinite.
- The set of all sequences whose elements are the digits 0 and 1 is not countable.
- This can be used to show that the set of real numbers is not countable.
- Let F be a CDF and let D be the set of discontinuities of F. Then D is at most countable.

Friday, January 22, 2016

Metric Spaces

Basic Definitions

- One way to express ideas of approximation and convergence is to use a *distance* or *metric*
- Distance between two points x and y in the plane or in space is define as

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2};$$

  for the plane $n = 2$; for 3-dimensional space $n = 3$.
- The definition makes sense for any finite n.
- The key properties:
  1. Non-negativity: $d(x,y)$ is non-negative for all x,y;
  2. Definiteness: $d(x,y) = 0$ if and only if $x = y$;
  3. Symmetry: $d(x,y) = d(y,x)$ for all x,y;
  4. Triangle inequality: $d(x,z) \leq d(x,y) + d(y,z)$ for all x,y,z.
- A function with these properties is a metric, or distance.
- A metric space is a pair (X,d) with X a set and d a metric on X.
- If (X,d) is a metric space and $Y \subset X$ then (Y,d) is also a metric space, sometimes called a sub-space.

Examples

- $X = \mathbb{R}$ and $d(x,y) = |x-y|$.
- X is any set, and
$$d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y. \end{cases}$$
This is sometimes called the *discrete metric*.
- $X = \mathbb{R}^n$ and $d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$.
- $X = \mathbb{R}^n$ and $d(x,y) = \sum_{i=1}^n |x_i - y_i)^2$.
- These last two are of the form $d(x,y) = \|x-y\|$ where $\|x\|$ is a *norm*.
- Other norms on $\mathbb{R}^n$:
  - $\|x\|_p = \sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$ for $p \geq 1$.
  - $\|x\|_\infty = \max_i |x_i|$.
- X is the set of all continuous functions on $[0,1]$, and
$$\|x\| = \max_{0 \leq t \leq 1} |x(t)|.$$
This metric space is sometimes denoted $C_{[0,1]}$.
- If $d(x,y)$ is a metric then
$$d_2(x,y) = \frac{d(x,y)}{1 + d(x,y)}$$
$$d_3(x,y) = \min(d(x,y), 1)$$
are also metrics. They are bounded, which can be useful.
- Let X, Y be two random variables. Then
$$d(X,Y) = E[\min(|X-Y|, 1)]$$
is a metric on (equivalence classes of) random variables.
  - This is called the Ky Fan metric.
  - Convergence in this metric is convergence in probability.

Metric Space Topology

Basic Definitions

- A *neighborhood* of a point x_0, or an *open ball* or *open sphere* centered at a point x_0 is the set

$$N_r(x_0) = \{x : d(x_0, x) < r\};$$

  r is the radius of the neighborhood/ball/sphere.

- A point x is a *limit point* of a set A if every neighborhood of x contains a point in $y \in A$ with $y \neq x$.

- Equivalently: x is a limit point of A if every neighborhood of x contains infinitely many points from A.

- A point x is a *contact point* of a set A if every neighborhood of x contains at least one point from A.

- If $x \in A$ is not a limit point of A then it is an *isolated point*.

- A point x is an *interior point* of A if it has a neighborhood N with $N \subset A$.

- A set A is *open* if every element of A is an interior point of A.

- A set is closed if it contains all its limit points.

- A set is *perfect* if it is closed and every point is a limit point, i.e. is has no isolated points.

- The *interior* of a set A, denoted A°, is the set of all interior points of A.

- The *closure* of a set A, denoted $\overline{A}$, is the union of A and its limit points.

- A set A is *dense* in X if $X = \overline{A}$.

- If $A \subset Y \subset X$ then A is an open subset relative to Y if for every $x \in A$ has a neighborhood N with $N \cap Y \subset A$.

Examples

- The interval $(0,1)$ is an open subset of $\mathbb{R}$.
- The interval $[0,1]$ is a perfect subset of $\mathbb{R}$.
- The interior of $[0,1]$ is $(0,1)$.
- The set $A = \{1/n : n = 1,2,\dots\}$ has one limit point, $x = 0 \notin A$. All elements of A are isolated points.
- For any metric space (X,d) the set X and the empty set $\emptyset$ are both open and closed.
- The rationals $\mathbb{Q}$ are a dense subset of the reals $\mathbb{R}$.
- Let $X = \mathbb{R}^2$, let

$$Y = \{(a,b) \in X : a = b\}$$
$$A = \{(a,b) \in X : a = b, 0 < a < 1\}.$$

  Then A is not open as a subset of X but is open as a subset of Y.

Week 2

Monday, January 25, 2016

Metric Space Topology

Basic Results

- A set is open if and only if its complement is closed. [R Thm 2.23]
- A set is closed if and only if its complement is open. [R Thm 2.23 Cor.]
- For any collection $\{A_\alpha\}$ of open sets the union $\bigcup_\alpha A_\alpha$ is open. [R Thm 2.24]
- For any finite collection $A_1, \ldots, A_n$ of open sets the intersection $\bigcap_{i=1}^n A_i$ is open. [R Thm 2.24]
- Analogous results hold for closed sets with union/intersection switched.
- $\overline{A}$ is the intersection of all closed sets containing A; it is the smalles closed set containing A. [R Thm 2.27]
- A° is the union of all open subsets of A; it is the largest open subset of A. [R Prob 2.9]
- The closure $\overline{A}$ of as set A is the set of all contact points of A. [KF §9.1]
- The set A' of limit points of A is closed. [R Prob 2.6]
- A and $\overline{A}$ have the same limit points. [R Prob 2.6]

- Suppose $A_1, A_2, \ldots$ are subsets of a metric space. [R Prob 2.7]
  - If $B_n = \bigcup_{i=1}^{n} A_i$ then $\overline{B}_n = \bigcup_{i=1}^{n} \overline{A}_i$.
  - If $B = \bigcup_{i=1}^{\infty} A_i$ then $\overline{B} \supset \bigcup_{i=1}^{\infty} \overline{A}_i$.

Some General Topology

Basic Definitions

- A *topology* τ on a set X is a collection of subsets of X, called open sets, such that
  - $X \in \tau$ and $\emptyset \in \tau$;
  - For any collection $\{A_\alpha\}$ of sets in τ the union $\bigcup_\alpha A_\alpha$ is also in τ.
  - For any finite collection $A_1, \ldots, A_n$ of sets in τ the intersection $\bigcap_{i=1}^n A_i$ is also in τ.
- Complements of open sets are called closed sets.
- A neighborhood of a point $x \in X$ is any open set $A \in \tau$ with $x \in A$.
- The definitions of contact point and limit points carry over.
- The definitions of closure and interior also carry over.
- A collection $\{V_\alpha\}$ of open sets is a *base* for a topology if every open set is a union of a subcollection of $\{V_\alpha\}$

Examples

- The open sets in a metric space (X, d) are a topology on X.
- The smallest topology on a set X is $\{X, \emptyset\}$.
- The largest topology on X is the collection of all subsets of X. This corresponds to the discrete metric.
- Pointwise convergence of functions $f : \mathbb{R} \to \mathbb{R}$ can be expressed as convergence in a topology, but not using a metric.
- In a metric space the collection of open balls is a base for the topology.

Properties and Specializations

- General topology is often too general to capture what we need.
- Some additional properties of metric spaces need to be brought over.
- The most important are *axioms of separation*: [KF §9.5]
  1. A topology satisfies the *first axiom of separation* (is a T_1 topology) if for any $x \in X$ and $y \in X$ with $x \neq y$ there exist neighborhoods A_x of x and A_y of y such that $x \notin A_y$ and $y \notin A_x$.
  2. A topology satisfies the *second axiom of separation* (is a T_2 topology, or a *Hausdorff* topology) if for any $x \in X$ and $y \in X$ with $x \neq y$ there exist neighborhoods A_x of x and A_y of y such that A_x and A_y are disjoint.
  3. A T_1 space is *normal* if for every pair of disjoint closed sets B_1 and B_2 there exist disjoint open sets A_1 and A_2 such that $B_1 \subset A_1$ and $B_2 \subset A_2$.
- In a T_1 topology every single point set is closed.
- Every normal space is Hausdorff. [KF §9 p. 86]
- Every metric topology is normal. [KF §9 Thm 9]
- A topological space is *separable* if it contains a countable dense subset.
- A topological space with a countable base is said to satisfy the *second axiom of countability*. [KF §9 p. 82]
  - A collection of open sets $\mathscr{B}$ in a topology τ is a *base* for τ if every open set in τ is the union of sets in $\mathscr{B}$.
- A topological space with a countable neighborhood base at every point is said to satisfy the *first axiom of countability*. [KF §9 p. 83]
  - A collection of open neighborhoods $\mathscr{B}$ of a point x is a *neighborhood base at x* if for every open set G containing x there is a set $A \in \mathscr{B}$ such that $A \subset G$.

Wednesday, January 27, 2016

Compactness

Basic Definitions

- Compactness is a key for establishing existence of maxima and minima.
- An open cover of a set K is a collection $\{G_\alpha\}$ of open sets such that $K \subset \bigcup_\alpha G_\alpha$.
- A set K is compact if and only if every open cover contains a finite subcover.
- These definitions work for metric spaces as well as topological spaces.
- If the entire space X is compact it is called a *compactum*.

Properties

- Closed subsets of compact sets are compact.
- Compact subsets of metric spaces are closed.
- If the sets $\{K_\alpha\}$ are compact and the intersection of every finite subcollection is non-empty, then $\bigcap_\alpha K_\alpha$ is non-empty. [R Thm 2.36]
- Corollary: If $\{K_n\}$ are compact, nonempty, and satisfy $K_n \subset K_{n+1}$ for all n, then $\bigcap_{i=1}^{\infty} K_n$ is non-empty. [Cor to R Thm 2.35]
- Every infinite subset of a compact set K has a limit point in K. [R Thm 2.37]
- Heine-Borel Theorem: A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. [R Thm 2.41]
- Weierstrass Theorem: Every bounded infinite set in $\mathbb{R}^n$ has a limit point in $\mathbb{R}^n$. [R Thm 2.42]
- Every separable metric space has a countable base. [R Prob 2.23]
- If every infinite subset of a metric space has a limit point then the metric space is separable. [KF §11 Cor 1]
- A subset K of a metric space is compact if and only if every infinite subset of K has a limit point in K. [KF §11 Cor 2 + R Thm 2.37, R Prob 2.26]

Compact subsets of metric spaces are closed: Some Details

- Strategy: Show that the complement is open.
- Let x be a point in K^c.
- For each $y \in K$ let

$$A_y = \left\{ z : d(y,z) < \frac{1}{2} d(x,y) \right\}$$
$$B_y = \left\{ z : d(x,z) < \frac{1}{2} d(x,y) \right\}.$$

- Then
  - For each y the sets A_y and B_y are disjoint, so $B_y \subset A_y^c$.
  - $\{A_y : y \in K\}$ is an open cover of K.
- Since K is compact there is a finite subcover $A_{y_1}, \ldots, A_{y_n}$.
- The intersection of the corresponding sets $B_{y_1}, \ldots, B_{y_n}$ satisfies

$$\bigcap_{i=1}^{n} B_{y_i} \subset \bigcap_{i=1}^{n} A_{y_i}^c = \left(\bigcup_{i=1}^{n} A_{y_i} \right)^c \subset K^c,$$

  and

$$\bigcap_{i=1}^{n} B_{y_i} = \{ z : d(x,z) < \min_i \frac{1}{2} d(x,y_i) \}$$

  is an open neighborhood of x.
- So x is an interior point of K^c.
- This is true for any $x \in K^c$, so K^c is open and K is closed.

Relative Compactness

- A subset of a topological space is *relatively compact*, or *precompact* if its closure is compact. [KF §10 Def 4]
- Let (X,d) be a metric space. A set $B \subset X$ is said to be an ε-net for a set $A \subset X$ and a value $\varepsilon > 0$ if for every $x \in A$ there exists a $y \in B$ such that $d(x,y) < \varepsilon$. [KF §11 Def 1]
- A set $A \subset X$ is said to be *totally bounded* if there exists a finite ε-net B_ε for A and every $\varepsilon > 0$. [KF §11 Def 2]
- If A is totally bounded and $B \subset A$ then B is also totally bounded.
- Every precompact subset of a metric space is totally bounded. [KF §11 Thm 2]
- This will help develop a characterization of the compact subsets of the set $C_{[0,1]}$ of continuous functions on $[0,1]$ with she distance

$$d(x,y) = \max_t |x(t) - y(t)|.$$

Connected Sets

- Two sets A and B are *separated* if $A \cap \overline{B} = \emptyset$ and $\overline{A} \cap B = \emptyset$. [R Def 3.45]
- A set is *connected* if it is *not* the union of two non-empty separated sets.
- Separated sets are disjoint. [R Rem 3.46]
- Disjoint sets need not be separated.
- Examples on the real line:
  - $A = (0,1)$ and $B = (1,2)$ are separated.
  - $A = (0,1)$ and $B = [1,2)$ are not separated.
- On the real line a set is connected if and only if it is an interval.

Vector Spaces and Norms

Basic Definitions

[KF §22]

- A *vector space*, or *linear space*, is a set together with a scalar field and
  - a vector addition operation
  - a scalar multiplication operation

  Every vector space includes a zero vector.
- The most commonly used scalars fields are the real numbers or the complex numbers.
- A *norm*, or length, is a real-valued function $\|x\|$ defined on a vector space L such that
  1. $\|x\| \geq 0$ for all $x \in L$, and $\|x\| = 0$ if and only if $x = 0$.
  2. $\|\alpha x\| = |\alpha|\|x\|$ for all vectors x and scalars α.
  3. $\|x+y\| \leq \|x\| + \|y\|$ for all $x, y \in L$.
- A vector space with a norm is called a *normed space*.
- A normed space becomes a metric space by defining

$$d(x,y) = \|x-y\|.$$

Examples

- Some vector spaces:
  - $\mathbb{R}^n$.
  - The set of polynomials of degree at most n.
  - The set of polynomials of any (finite) degree.
  - The set of continuous functions on a particular interval.
- Some norms:
  - $\|x\| = \sqrt{\sum x_i^2}$ for $x \in \mathbb{R}^n$.
  - $\|x\| = \max_i |x_i|$ for $x \in \mathbb{R}^n$.
  - $\|x\| = \max_t |x(t)|$ for $x \in C_{[0.1]}$.
  - $\|x\| = \sqrt{\int_0^1 x(t)^2 dt}$ for $x \in C_{[0.1]}$.

Friday, January 29, 2016

Distances for Probabilities

- The *total variation distance* between probabilities P and Q is

$$d_{TV}(P,Q) = \sup_{A \in \mathscr{B}} |P(A) - Q(A)|.$$

- If P and Q have densities p and p, then

$$d_{TV}(P,Q) = \frac{1}{2} \int |p(x) - q(x)| dx.$$

- The *Hellinger distance* between two probability densities p and q is defined as

$$d_H(p,q) = \sqrt{\frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx}.$$

- Suppose (M,d) is a metric space and P and Q are probabilities on M. For $p \geq 1$ the p-th *Wasserstein metric* is

$$d_{W,p}(P,Q) = \inf \sqrt[p]{\int E[d(X,Y)^p]},$$

where the infimum is taken over all joint distributions of M-valued random variables X and Y with marginal distributions $X \sim P$ and $Y \sim Q$.

- The *Lèvy* distance between two CDFs F and G is

$$d_L(F,G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x \in \mathbb{R}\}.$$

The Lèvy Distance: Some Details

- Symmetry: Fix $\varepsilon > 0$.
  - If for all $x \in \mathbb{R}$

$$F(x-\varepsilon) - \varepsilon \le G(x)$$
$$G(x) \le F(x+\varepsilon) + \varepsilon$$

    then, taking $y = x - \varepsilon$,

$$F(y) \le G(y+\varepsilon) + \varepsilon$$

    and, taking $y = x + \varepsilon$,

$$G(y-\varepsilon) - \varepsilon \le F(y)$$

    for all y.
  - So $d_L(F,G)$ and $d_L(G,F)$ are infima over the same sets and are identical.
- Definiteness: By right continuity $d(F,G) = 0$ implies $G(x) \le F(x)$ for all x, and $d_L(G,F) = 0$ implies $F(x) \le G(x)$ for all x. So $d_L(F,G) = 0$ implies $G(x) = F(x)$ for all x.
- Triangle inequality:
  - Suppose $\varepsilon_1 > d_L(F,H)$ and $\varepsilon_2 > d_L(H,G)$.
  - Then for all x

$$F(x-\varepsilon_1-\varepsilon_2) - \varepsilon_1 - \varepsilon_2 \le H(x-\varepsilon_2) - \varepsilon_2 \le G(x)$$

    and

$$G(x) \le H(x+\varepsilon_2) + \varepsilon_2 \le F(x+\varepsilon_1+\varepsilon_2) + \varepsilon_1 + \varepsilon_2.$$

  - So $d_L(F,G) \le \varepsilon_1 + \varepsilon_2$.
  - Taking the infimum over all $\varepsilon_1 > d_L(F,H)$ and $\varepsilon_2 > d_L(H,G)$ then gives

$$d_L(F,G) \le d_L(F,H) + d_L(H,G).$$

Convergence of Sequences

Basic Definitions

- A *sequence* x_n in X is a function defined on the set $\mathscr{J}$ of positive integers with values in X.
- The *range* of the sequence is the image of this function, the set of values.
- Convergence of numeric sequences:
  - A sequence of numbers x_n converges to a number x if for every $\varepsilon > 0$ there exists an integer N_ε such that $|x_n - x| < \varepsilon$ for all $n \geq N_\varepsilon$.
  - A sequence of numbers x_n converges to a number x if for any $\varepsilon > 0$ all but finitely many elements of the sequence satisfy $|x_n - x| < \varepsilon$. [R Thm 3.3 a]
- Convergence of sequences in a metric space:
  - A sequence of points x_n in a metric space converges to a point x if $d(x_n, x)$ converges to zero.
  - A sequence of points in a metric space x_n converges to a point x if for every $\varepsilon > 0$ there exists an integer N_ε such that $d(x_n, x) < \varepsilon$ for all $n \geq N_\varepsilon$.
  - A sequence of points x_n in a metric space converges to a point x if every open neighborhood of x contains all but finitely many elements of the sequence.
- Notations:

$$x_n \to x$$
$$x_n \to x \text{ as } n \to \infty$$
$$\lim_{n\to\infty} x_n = x$$

- A sequence of points x_n in a metric space *converges*, or *is convergent*, if there exists a point x in the metric space such that $x_n \to x$.
- A sequence that does not converge is said to *diverge*, or to *be divergent*.

Examples

- Let $X = [0,\infty)$ with the usual metric, and $x_n = 1/n$. Then $x_n \to 0$.
- Let $X = (0,\infty)$ with the usual metric, and $x_n = 1/n$. Then x_n does not converge.
- Let $X = \mathbb{R}^2$ with the usual metric, and $x_n = (1 + 1/n, 2 - 1/n)$. Then $x_n \to (1,2)$.
- $X = \overline{\mathbb{R}}$, let $f(x)$ be the logistic CDF

$$f(x) = \begin{cases} 0 & \text{if } x = -\infty \\ \frac{1}{1+e^{-x}} & \text{for finite } x \\ 1 & \text{if } x = +\infty, \end{cases}$$

  and let

$$d(x,y) = |f(x) - f(y)|.$$

  Then

  - d is a metric on X;
  - $x_n = 1/n \to 0$;
  - $x_n = n \to \infty$.

- The metric space (X,d) is compact; this is sometimes called the *two-point compactification* of the reals.

Basic Properties

Suppose (X,d) is a metric space.

- If $x_n \to x$ and $x_n \to x'$ then $x = x'$ [R Thm 3.3 b]
  - Suppose $x \neq x'$.
  - Then there exist open neighborhoods G_x of x and $G_{x'}$ of x' with
  $$G_x \cap G_{x'} = \emptyset.$$
  - Since $x_n \to x$ all but finitely many elements of the sequence are in G_x.
  - So at most finitely many elements of the sequence can be in $G_{x'}$, and x_n cannot converge to x'.
  - This result is true in any Hausdorff space.
- A point x is a contact point of a set A if and only if there exists a sequence x_n of points in A such that $x_n \to x$. [KF §6 Thm 2]
- A point x is a limit point of a set A if and only if there exists a sequence x_n of distinct points in A such that $x_n \to x$. [KF §6 Thm 2']

More Results

- Suppose $x_n \to x$ and $y_n \to y$. Then $d(x_n, y_n) \to d(x, y)$. [KF Prob 6.3]
  - From the triangle inequality: For all $x, y, u, z \in X$

$$|d(x,z) - d(u,y)| \le d(x,y) + d(z,u).$$

  - So

$$|d(x_n,y_n) - d(x,y)| \le d(x_n,x) + d(y_n,y).$$

- In a normed space, if $x_n \to x$ and $y_n \to y$ then $x_n + y_n \to x + y$:

$$\begin{aligned} \|(x_n + y_n) - (x + y)\| &= \|(x_n - x) + (y_n - y)\| \\ &\le \|x_n - x\| + \|y_n - y\|. \end{aligned}$$

- Analogous results hold for scalar multiplication.
- Analogous results hold for basic real or complex arithmetic operations. [R Thm 3.4]
- Let $X = \mathbb{R}^k$ with d the Euclidean distance. Then

$$x_n = (a_{1,n}, \ldots, a_{k,n}) \to x = (a_1, \ldots, a_k)$$

  if and only if $a_{i,n} \to a_i$ for all i:

  - Suppose $a_{i,n} \to a_i$ for all i. Then

$$d(x_n, x)^2 = \sum_{i=1}^{k} (a_{i,n} - a_i)^2 \to 0,$$

    so $d(x_n, x) \to 0$ and $x_n \to x$.
  - Suppose $x_n \to 0$. Then $d(x_n, x) \to 0$ and for each i

$$|a_{i,n} - a_i| \le d(x_n, x).$$

Distance to a Set

Let (X,d) be a metric space, $x \in X$ a point, and $A \subset X$ a subset. The distance between A and x is defined as [KF Prob 6.10]

$$d(A,x) = \inf_{a \in A} d(a,x).$$

- $d(A,x) = 0$ if and only if $x \in \overline{A}$.
- For $x, y \in X$ the distance to A satifies

$$|d(A,x) - d(A,y)| \leq d(x,y).$$

- As a result, if $x_n \to x$ then $d(A,x_n) \to d(A,x)$.

The distance between two sets A and B can be defined as [KF Prob 6.10]

$$d(A,B) = \inf_{\substack{a \in A \\ b \in B}} d(a,b).$$

If $A \cap B \neq \emptyset$ then $d(A,B) = 0$, but the converse is not true.

Details for $|d(A,x) - d(A,y)| \le d(x,y)$

- By the triangle inequality, for any $a \in A$

$$d(a,x) \le d(x,y) + d(a,y).$$

- Therefore

$$d(A,x) = \inf_{b \in A} d(b,x) \le d(a,y) + d(x,y)$$

  for all $a \in A$.

- So $d(A,x) - d(x,y)$ is a lower bound for the set $\{d(a,y) : a \in A\}$
- Therefore $d(A,x) - d(x,y)$ is less than or equal to the greatest lower bound,

$$d(A,y) - d(x,y) \le \inf_{a \in A} d(a,y) = d(A,y).$$

- So $d(A,x) \le d(A,y) + d(x,y)$, or

$$d(A,x) - d(A,y) \le d(x,y).$$

- Reversing the roles of x and y,

$$d(A,y) - d(A,x) \le d(y,x) = d(x,y).$$

- So

$$|d(A,x) - d(A,y)| \le d(x,y).$$

Week 3

Monday, February 1, 2016

Subsequences

- Given a sequence x_n and a sequence of positive integers n_k with $n_1 < n_2 < n_3 < \dots$ the sequence $y_k = x_{n_k}$ is a *subsequence* of x_n.
- In a metric space x_n converses to x if and only if every subsequence converges to x. [R after Def 3.5]
- In a metric space x_n converges to $x \in X$ if and only if every subsequence x_{n_k} has a further subsequence $x_{n_{k_i}}$ that converges to x.

- If x_n is a sequence in a compact metric space then some subsequence converges to a point $x \in X$. [R Thm 3.6 a]
  - If the range of the sequence is finite, then at least one value x is repeated infinitely often; use the subsequence where $x_n = x$.
  - If the range is infinite, then it has a limit point in X.
    * Otherwise we could find for each $x \in X$ a set G_x containing at most one point in the range;
    * $\{G_x : x \in X\}$ is an open cover of X;
    * but it does not contain a finite subcover of the range;
    * therefore it does not contain a finite subcover of X.
  - Let x be such a limit point.
  - Let $n_0 = 0$ and let

$$n_k = \min\{i > n_{k-1} : d(x_i, x) < 1/2^k\}.$$

  - This is possible since x is a limit point of the range.
  - Then $d(x_{n_k}, x) < 1/2^k \to 0$ and thus $x_{n_k} \to x$.
- Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence. [R Thm 3.6b]
- The set of all *subsequential limits* of a sequence is closed. [R Thm 3.7]

The Set of Subsequential Limits is Closed: Some Details

- Let x be a limit point of the set of subsequential limits of x_n.
- Let $n_0 = 0$.
- Suppose $n_0, \ldots, n_{k-1}$ have been chosen.
- Set $\delta = 1/2^{k+1}$.
- Since x is a limit point of the subsequential limits, there exists a point x' that is a limit of some subsequence and satisfies $d(x', x) < \delta$.
- Since x' is a limit of some subsequence, there are infinitely many points in the original sequence with distance to x' less than δ and indices larger than n_{k-1}.
- Choose n_k as one of these indices, say the smallest.
- Then $d(x_{n_k}, x) < d(x_{n_k}, x') + d(x', x) < 2\delta = 1/2^k$.
- So the subsequence x_{n_k}converges to x.

Cauchy Sequences

Basic Definitions

- A sequence x_n in a metric space (X,d) is a *Cauchy sequence* if for every $\varepsilon > 0$ there is an integer N_ε such that

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

  for all $n, m \geq N_\varepsilon$.

- A metric space is *complete* if every Cauchy sequence converges.
- The *diameter* of a subset A of a metric space is defined as

$$\text{diam}(A) = \sup\{d(x,y) : x, y \in A\}$$

- The diameter can be finite or infinite.
- A set A is *bounded* if its diameter is finite.

Basic Properties

- Every convergent sequence in a metric space is a Cauchy sequence. [R Thm 3.11 a]
- Every Cauchy sequence in a compact metric space converges. [R Thm 3.11 b]
- $\mathbb{R}^k$ with the Euclidean metric is complete. [R Thm 3.11 c]
- A sequence x_n is a Cauchy sequence if and only if [R Def 3.9]

$$\lim_{n\to\infty} \text{diam}(\{x_n, x_{n+1}, x_{n+2}, \dots\}) = 0.$$

- For any subset A of a metric space [R Thm 3.10 a]

$$\text{diam}(A) = \text{diam}(\overline{A}).$$

Examples

- Let $X = (0,\infty)$ with the usual metric, and $x_n = 1/n$. Then x_n does not converge, so X is not complete.
- The diameter of $[-1,1]$ is $\text{diam}([-1,1]) = 2$.
- The diameter of $(0,\infty)$ is $\text{diam}((0,\infty)) = \infty$.
- For a value $p \geq 1$ let X be the set of infinite sequences

$$X = \left\{ x = (a_1, a_2, \dots) : a_i \in \mathbb{R}, \sum_{i=1}^{\infty} |a_i|^p < \infty \right\}.$$

  This is a vector space, and

$$\|x\| = \sqrt[p]{\sum_{i=1}^{\infty} |a_i|^p}$$

  is a norm on this space. This space is called ℓ_p and is a complete metric space.

- A normed vector space that is complete is called a *Banach space*.

Completeness of ℓ_p

- Showing completeness usually involves these steps:
  1. Assume we have a Cauchy sequence x_n.
  2. Figure out what the limit x would have to be.
  3. Verify that x is in the metric space.
  4. Show that x_n converges to x.
- Steps 1, 2:
  - Suppose $x_n = (a_{1,n}, a_{2,n}, \dots)$ is a Cauchy sequence in ℓ_p.
  - This means for each ε there is an integer N such that $\|x_n - x_m\| < \varepsilon$ for all $n, m \geq N$.
  - Since $|a_{i,n} - a_{i,m}| \leq \|x_n - x_m\|$ for each i, this means that for each i the sequence $a_{i,n}$ is a Cauchy sequence of real numbers.
  - Since $\mathbb{R}$ is complete, this means that for each i there exists a real number a_i such that $a_{i,n} \to a_i$.
  - So the limit would have to be $x = (a_1, a_2, \dots)$.

- Steps 3, 4:
  - For Step 3 we need to show that $\sum_{i=1}^{\infty} |a_i|^p < \infty$.
  - For Step 4 we need to show that $\|x_n - x\| \to 0$.
  - Fix $\varepsilon > 0$ and choose N such that $\|x_n - x_m\| < \varepsilon$ for $n, m > N$.
  - Then for any fixed positive integer M

$$\sum_{i=1}^{M} |a_{i,n} - a_{i,m}|^p \le \|x_n - x_m\|^p < \varepsilon^p.$$

    Taking the limit as $m \to \infty$ gives

$$\sum_{i=1}^{M} |a_{i,n} - a_i|^p \le \varepsilon^p.$$

    Since this holds for all finite M we also have

$$\sum_{i=1}^{\infty} |a_{i,n} - a_i|^p \le \varepsilon^p.$$

    This shows that $\|x_n - x\| \le \varepsilon < \infty$ for some n and therefore

$$\|x\| \le \|x_n\| + \|x_n - x\| < \infty,$$

    so $x \in X$, which is Step 3.
  - This also shows that, given any $\varepsilon > 0$, there exists an integer N such that $\|x_n - x\| \le \varepsilon$ for all $n \ge N$, i.e. that $x_n \to x$, which is Step 4.

More Properties

- A sequence of sets A_n is said to be *nested* if

$$A_1 \supset A_2 \supset \ldots A_n \supset \ldots$$

- Suppose K_n is a nested sequence of compact subsets of a metric space. If [R Thm 3.10 b]

$$\lim_{n\to\infty} \text{diam}(K_n) = 0$$

then $\cap_{n=1}^{\infty} K_n$ consists of exactly one point.

- Suppose A_n is a nested sequence of closed subsets of a complete metric space. If [KF §7 Prob 4]

$$\lim_{n\to\infty} \text{diam}(A_n) = 0$$

then $\cap_{n=1}^{\infty} A_n$ consists of exactly one point.

- A *closed sphere* of radius r centered at a point x in a metric space is a set of the form

$$S_r(x) = \{y : d(x,y) \leq r\}.$$

- A metric space is complete if and only if every nested sequence of closed spheres $S_{r_n}(x_n)$ with $r_n \to 0$ has a nonempty intersection, which will consist of exactly one point. [KF §7 Thm 2]

- A metric space is compact if and only if it is totally bounded and complete. [KF §11 Thm 2]

- A subset of a complete metric space is relatively compact if and only if it is totally bounded. [KF §11 Thm 3]

$C_{[0,1]}$ is complete

- The metric space $C_{[0,1]}$ is the set of continuous real-valued functions on $[0,1]$ with the norm [KP §7 Ex 4]
$$\|x\| = \sup_{0 \le t \le 1} |x(t)|.$$
- Suppose x_n is a Cauchy sequence in this space.
- Then for each $t \in [0,1]$ and any integers n, m
$$|x_n(t) - x_m(t)| \le \|x_n - x_m\|.$$
- So the sequence of real numbers $x_n(t)$ is a Cauchy sequence and thus has a limit $x(t)$.
- Given $\varepsilon > 0$ there is an integer N such that
$$|x_n(t) - x_m(t)| \le \|x_n - x_m\| \le \varepsilon$$
for all $t \in [0,1]$.
- Taking the limit as $m \to \infty$ this gives
$$|x_n(t) - x(t)| \le \varepsilon$$
for all $t \in [0,1]$ and all $n \ge N$
- Therefore
$$\|x_n - x\| = \sup_{0 \le t \le 1} |x_n(t) - x(t)| \le \varepsilon.$$
- Still need Step 3: The function $x(t)$ is continuous.

- To show continuity at $t \in [0,1]$ we need to show that for every $\varepsilon > 0$ there is a $\delta > 0$ such that
$$|x(s) - x(t)| < \varepsilon$$
when $|s-t| \le \delta$.

- Since $\|x_n - x\| \to 0$ we can find an integer N such that $\|x_N - x\| < \varepsilon/3$, and therefore
$$|x_N(u) - x(u)| < \varepsilon/3$$
for all $u \in [0,1]$.

- Since x_N is continuous at t, there exists a $\delta > 0$ such that
$$|x_N(s) - x_N(t)| < \varepsilon/3$$
when $|s-t| < \delta$.

- So when $|s-t| < \delta$
$$\begin{aligned} |x(s) - x(t)| &\le |x(s) - x_N(s)| + |x_N(s) - x_N(t)| + |x_N(t) - x(t)| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 \\ &= \varepsilon. \end{aligned}$$

- So x is continuous on $[0,1]$.

Wednesday, February 3, 2016

Cauchy Sequences

$C^2_{[0,1]}$ is not complete

- The metric space $C^2_{[0,1]}$ is the set of continuous real-valued functions on $[0,1]$ with the norm [KP §7 Ex 6]

$$\|x\| = \sqrt{\int_0^1 x(t)^2 dt}.$$

- For $n \geq 2$ Let

$$x_n(t) = \begin{cases} 0 & \text{for } 0 \leq t \leq \frac{1}{2} - \frac{1}{n} \\ \frac{1}{2} + \frac{n}{2}\left(t - \frac{1}{2}\right) & \text{for } \frac{1}{2} - \frac{1}{n} \leq t \leq \frac{1}{2} + \frac{1}{n} \\ 1 & \text{for } \frac{1}{2} + \frac{1}{n} \leq t \leq 1 \end{cases}$$

and

$$x(t) = \begin{cases} 0 & \text{for } 0 \leq t < \frac{1}{2} \\ 1 & \text{for } \frac{1}{2} \leq t \leq 1 \end{cases}$$

- Then
  - x_n is continuous for all n
  - $x_n(t) \to x(t)$ for all $t \neq \frac{1}{2}$
  - $x(t)$ has a discontinuity at $t = \frac{1}{2}$.
  - $\|x_n - x\| = 1/\sqrt{6n}$.
  - So for any n, m we have $\|x_n - x_m\| \leq 1/\sqrt{6n} + 1\sqrt{6m}$, and thus x_n is a Cauchy sequence.

- Suppose y is a continuous function on $[0,1]$ and $\|x_n - y\| \to 0$.
- Because of the nature of the discontinuity of x at $t = \frac{1}{2}$ we have $x(t) \neq y(t)$ on some open interval in $[0,1]$ with one endpoint at $t = \frac{1}{2}$.
- This implies that $M = \|x - y\| > 0$.
- By the triangle inequality

$$\|x_n - y\| \geq \|x - y\| - \|x_n - x\| = M - \|x_n - x\|.$$

- Since $\|x_n - x\| \to 0$,

$$\|x_n - y\| \geq M/2 > 0.$$

  for all large n

- So x_n does not converge to any continuous function in $C^2_{[0,1]}$, and therefore $C^2_{[0,1]}$ is not complete.

Convergence of Random Variables

Almost Sure Convergence

- Let $X, X_1, X_2, \ldots$ be random variables on a probability space $(S, \mathscr{B}, P)$.
- $X_n \to X$ almost surely if

$$P(\{s \in S : X_n(s) \to X(s)\}) = 1,$$

or

$$P(\{s \in S : X_n(s) \not\to X(s)\}) = 0.$$

- For $\varepsilon > 0$ let

$$\begin{aligned} A_\varepsilon &= \{s \in S : |X_n(s) - X(s)| \geq \varepsilon \text{ for infinitely many } n\} \\ &= \{s \in S : |X_n(s) - X(s)| \geq \varepsilon \text{ infinitely often}\} \\ &= \{s \in S : |X_n(s) - X(s)| \geq \varepsilon \text{ i.o.}\}. \end{aligned}$$

- Then the set of points s where convergence fails can be written as

$$\{s \in S : X_n(s) \not\to X(s)\} = \bigcup_{\varepsilon > 0} A_\varepsilon.$$

- For $\delta > \varepsilon$ we have $A_\delta \subset A_\varepsilon$, so

$$\bigcup_{\varepsilon > 0} A_\varepsilon = \bigcup_{n=1}^{\infty} A_{1/n},$$

and $A_{1/n} \subset A_{1/m}$ for $n \leq m$.

- $X_n \to X$ almost surely if and only if $P(A_\varepsilon) = 0$ for all $\varepsilon > 0$.

Convergence in Probability

- X_n converges in probability to X if for any $\varepsilon > 0$

$$\lim_{n\to\infty} P(\{s \in S : |X_n(s) - X(s)| < \varepsilon\}) = 1,$$

  or, equivalently,

$$\lim_{n\to\infty} P(\{s \in S : |X_n(s) - X(s)| \geq \varepsilon\}) = 0.$$

- $X_n \to X$ in probability if and only if $X_n \to$ in the Ky Fan metric

$$d_{\mathrm{KF}}(X,Y) = E[\min(|X - Y|, 1)].$$

  Several other metrics can be used as well.

- If $X_n \to X$ almost surely, then $X_n \to X$ in probability as well.

- The converse does not hold: it is possible to have $X_n \to X$ in probability and not almost surely.

- There is a partial converse: If $X_n \to X$ in probability, then there is a subsequence X_{n_k} such that $X_{n_k} \to X$ almost surely.

Application: The Continuous Mapping Theorem

Suppose $X_n \to$ in probability and f is a real-valued function. Let D be the set of points where f is not continuous. If $P(X \in D) = 0$, then $f(X_n) \to f(X)$ in probability.

- Since convergence in probability is convergence in a metric space, it is enough to show that every subsequence of $f(X_n)$ has a further subsequence that converges to $f(X)$ in probability.
- Given a subsequence X_{n_k}, which converges in probability to X, find a further subsequence $X_{n_{k_i}}$ that converges to X almost surely.
- Let
$$A = \left\{ s \in S : X_{n_{k_i}}(s) \to X(s) \text{ as } i \to \infty \text{ and } X(s) \notin D \right\}.$$
- Then $P(A) = 1$.
- Furthermore, for every $s \in A$
$$\lim_{i \to \infty} f(X_{n_{k_i}}(s)) = f(X(s)).$$
- So $f(X_{n_{k_i}}) \to f(X)$ almost surely, and therefore in probability as well.

The Borel-Cantelly Lemma

Let B_n be a sequence of events such that

$$\sum_{i=1}^{\infty} P(B_n) < \infty.$$

Then the number of these events that occur is finite with probability one, or $P(B_n \text{ i.o}) = 0$.

- The number N of these events that occur is

$$N = \sum_{i=1}^{\infty} 1_{B_n}.$$

- The expected value of N is

$$E[N] = \sum_{i=1}^{\infty} P(B_n).$$

- By hypothesis this expected value is finite, so $P(N < \infty) = 1$.

Existence of an Almost Surely Convergent Subsequence

- Suppose $X_n \to X$ in probability.
- Let $n_0 = 0$ and for every positive integer k define

$$n_k = \min\left\{n > n_{k-1} : P\left(|X_n - X| \geq \frac{1}{k}\right) \leq \frac{1}{2^k}\right\}$$

- For any $\varepsilon > 0$

$$P(|X_{n_k} - X| \geq \varepsilon \text{ i.o}) \leq P\left(|X_{n_k} - X| \geq \frac{1}{k} \text{ i.o}\right)$$

- By the choice of the subsequence

$$\sum_{k=1}^{\infty} P\left(|X_{n_k} - X| \geq \frac{1}{k}\right) \leq \sum_{k=1}^{\infty} \frac{1}{2^k} < \infty.$$

- So $P(|X_{n_k} - X| \geq \varepsilon \text{ i.o}) = 0$ for every $\varepsilon > 0$, and $X_{n_k} \to X$ almost surely.

Metric Space Valued Random Variables

- Let $(S, \mathscr{B}, P)$ be a probability space and (M, d) a metric space.
- A function $X : S \to M$ is an *M-valued random variable* if

$$X^{-1}(A) \in \mathscr{B}$$

  for all open sets A in the metric space.

- All results for convergence of random variables carry over with the obvious modifications to use

$$d(a,b)$$

  instead of $|a-b|$.

Friday, February 4, 2016

Contraction Mappings

Basic Definitions and Results

- Let (X,d) be a metric space and $A : X \to X$ a mapping from X into X.
- A is a *contraction mapping* if there exists a number $\alpha < 1$ such that

$$d(A(x),A(y)) \leq \alpha d(x,y).$$

.

- Notation:

$$Ax = A^1x = A(x)$$
$$A^nx = A(A^{n-1}x).$$

- If A is a contraction mapping and $x_n \to x$ then $Ax_n \to Ax$ since

$$d(Ax_n,Ax) < d(x_n,x).$$

- A point x is a *fixed point* of A if $Ax = x$.
- If x and y are fixed points then

$$d(x,y) = d(Ax,Ay) \leq \alpha d(x,y).$$

  - Since $\alpha < 1$ this is only possible if $d(x,y) = 0$.
  - So $x = y$.
  - If a fixed point exists it is unique.

- *The fixed point theorem:* Every contraction mapping A on a complete metric space has a unique fixed point, and for any point in the metric space x_0 the sequence $x_n = A^nx_0$ converges to the fixed point of A. [KF §8 Thm 1]
- The result also holds if A^k is a contraction mapping for some integer $k > 1$. [KF §8 Thm 1′]

The Fixed Point Theorem: Some Details

- Given a point x_0 let $x_n = A^n x_0$.
- For any $n \le m$

$$\begin{aligned} d(x_n, x_m) &= d(A^n x_0, A^n (A^{m-n} x_0)) \\ &\le \alpha^n d(x_0, A^{m-n} x_0) \\ &\le \alpha^n [d(x_0, x_1) + d(x_1, x_2) + \cdots + d(x_{m-n-1}, x_{m-n})] \\ &\le \alpha^n d(x_0, x_1)[1 + \alpha + \alpha^2 + \cdots + \alpha^{m-n-1}] \\ &\le \alpha^n d(x_0, x_1) \frac{1}{1-\alpha}. \end{aligned}$$

- So x_n is a Cauchy sequence and converges to a point x.
- Since A is continuous x is a fixed point:

$$Ax = A \lim_{n\to\infty} x_n = \lim_{n\to\infty} Ax_n = \lim_{n\to\infty} x_{n+1} = x$$

- Error bounds:

$$\begin{aligned} d(x_n, x) &\le \frac{\alpha}{1-\alpha} d(x_n, x_{n-1}) \\ d(x_n, x) &\le \frac{\alpha^n}{1-\alpha} d(x_1, x_0) \end{aligned}$$

Applications

- Showing existence and of solutions to differential and integral equations. [KF §8]
- Newton's method for computing square roots:
  - Suppose $a > 0$ and
$$A(x) = \frac{1}{2}\left(x + \frac{a}{x}\right).$$
  - Then
$$A(\sqrt{a}) = \frac{1}{2}\left(\sqrt{a} + \frac{a}{\sqrt{a}}\right) = \sqrt{a}.$$
  - So $\sqrt{a}$ is a fixed point of A.
  - Let $X = [\sqrt{a}, \infty)$.
  - For $x \in X$ the derivative is
$$A'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right) \geq 0.$$
  - So A maps X into X.
  - For $x, y \in X$
$$\begin{aligned} |A(x) - A(y)| &= \frac{1}{2}\left|x - y + \frac{a}{x} - \frac{a}{y}\right| = \frac{1}{2}\left|x - y + a\frac{y-x}{xy}\right| \\ &= \frac{|x-y|}{2}\left(1 - \frac{a}{xy}\right) \leq \frac{1}{2}|x-y|. \end{aligned}$$
  - So A is a contraction on X.
  - If we start with $x_0 > \sqrt{a}$ then $A^n x_0$ converges to $\sqrt{a}$ very quickly:

```
> mkA <- function(a) function(x) 0.5 * (x + a / x)
> A <- mkA(2)
> A(2)
[1] 1.5
> A(A(2))
[1] 1.416667
> A(A(A(2)))
[1] 1.414216
> sqrt(2)
[1] 1.414214
```

Iterative Solution of Linear Equation Systems

- Suppose we want to solve the system of linear equations

$$Bx - b.$$

- Let C be non-singular and write $B = B - C + C$.
- The equation $Bx = b$ can be written as

$$(B-C)x + Cx = b$$

or

$$x = C^{-1}(C-B)x + C^{-1}b.$$

So the solution x is a fixed point for the mapping

$$Ax = C^{-1}(C-B)x + C^{-1}b.$$

- C is chosen to make solving $Cy = c$ easy.
- Choosing C equal to the diagonal part of B produces the *Jacobi* method for iteratively solving linear equations.
- Another iterative method of this type is the *Gauss-Seidel* method.
- These methods work if their mappings are contraction mappings.
- Choosing the right norm can make verifying this easier. [KF §8 Ex 2]
- The Jacobi method is a contraction for the norm

$$\|(a_1, \ldots, a_k)\| = \max_{1 \le i \le k} |a_i|$$

if B is *strongly row diagonally dominant*:

$$\sum_{\substack{j=1 \\ j \neq i}}^{k} |b_{ij}| < |b_{ii}|$$

for all $i = 1, \ldots, k$.

Monotonic Sequences

- A sequence of real numbers x_n is monotonically increasing, or non-decreasing, if $x_n \leq x_{n+1}$ for all n.
- Monotonically decreasing is defined analogously.
- A monotonic sequence converges to a finite limit if and only if it is bounded. [R Thm 3.14]
- Any sequence converging in a metric space is bounded.
- If x_n is bounded and non-decreasing let x be its least upper bound.
- For any $\varepsilon > 0$ the value $x - \varepsilon$ is not an upper bound on the sequence.
- So there is an N such that $x_N > x - \varepsilon$, and since the sequence is monotone this means $s_n > x - \varepsilon$ for all $n \geq N$.
- Since $x_n \leq x$ for all n this means

$$|x_n - x| < \varepsilon$$

  for all $n \geq N$.
- So x_n converges to x.

Upper and Lower Limits

Basic Definitions and Properties

- Every sequence in $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$ has a subsequence that converges to a limit in $\overline{\mathbb{R}}$.
- The *upper limit* of a sequence of x_n in $\overline{\mathbb{R}}$ is least upper bound on the set of all subsequential limits.
- The *lower limit* is analogous.
- Notation:

$$\limsup_{n\to\infty} x_n$$
$$\liminf_{n\to\infty} x_n$$

- The upper and lower limits satisfy

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup\{x_m : m \geq n\}$$
$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf\{x_m : m \geq n\}$$

- If $x^* = \limsup_{n\to\infty} x_n$ then [R Thm 3.17]
  - x^* is a limit of a subsequence of x_n.
  - If $x > x^*$ then there exists an integer N such that $x_n < x$ for all $n \geq N$.

  x^* is the only number with these properties.
- $x_n \to x$ if and only if

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$$

- If $x_n \leq y_n$ for all n then

$$\liminf_{n\to\infty} x_n \leq \liminf_{n\to\infty} y_n$$
$$\limsup_{n\to\infty} x_n \leq \limsup_{n\to\infty} y_n$$

Examples

- For $x_n = (-1)^n(1-1/n)$

$$\limsup_{n\to\infty} x_n = 1$$
$$\liminf_{n\to\infty} x_n = -1$$

- For $x_n = (-1)^n(1+1/n)$

$$\limsup_{n\to\infty} x_n = 1$$
$$\liminf_{n\to\infty} x_n = -1$$

- For $x_n = (-1)^n n$

$$\limsup_{n\to\infty} x_n = +\infty$$
$$\liminf_{n\to\infty} x_n = -\infty$$

- For $x_n = (-1)^n(1/n)$

$$\limsup_{n\to\infty} x_n = 0$$
$$\liminf_{n\to\infty} x_n = 0$$

Some Special Sequences

- If $p > 0$, then $\lim_{n\to\infty} \sqrt[n]{p} = 1$:

$$\log(\sqrt[n]{p}) = \frac{1}{n}\log(p) \to 0.$$

- $\lim_{n\to\infty} \sqrt[n]{n} = 1$:
  - Set $x_n = \sqrt[n]{n} - 1$.
  - Then $x_n \geq 0$, and

$$n = (1 + x_n)^n = \sum_{k=0}^{n} \binom{n}{k} x_n^k \geq \frac{n(n-1)}{2} x_n^2.$$

  - So

$$0 \leq x_n \leq \sqrt{\frac{2}{n-1}}$$

    and $x_n \to 0$.

Week 4

Monday, February 8, 2016

Series

Basic Definitions and Properties

- Given a sequence $a_1, a_2, \ldots$ of numbers an associated series is the sequence of partial sums
$$s_n = \sum_{i=1}^{n} a_i.$$
- $\sum a_n$ is used as shorthand for the series or its limit.
- If s_n converges to s then the series is said to converge, and
$$s = \sum_{i=1}^{\infty} a_i.$$
- Sometimes other starting indices are used.
- From the Cauchy criterion a series $s_n = a_1 + \cdots + a_n$ converges if and only if for any $\varepsilon > 0$ there exists and integer N such that for any $n, m \geq N$ with $n \leq m$ [R Thm 3.22]
$$\left| \sum_{i=n}^{m} a_i \right| < \varepsilon.$$
- This means that $|a_n| \to 0$ if s_n converges.
- $|a_n| \to 0$ is not sufficient for s_n to converge.

Convergence Criteria

- A series of non-negative terms converges if and only if the sequence of partial sums is bounded. [R Thm 3.24]
  - The partial sum sequence is monotonic.
- If $|a_n| \le c_n$ for all large n and $\sum c_n$ converges, then $\sum a_n$ converges. [R Thm 3.25 a]
  - This follows from the Cauchy criterion.
- For a sequence of real numbers a_n, if $a_n \ge d_n \ge 0$ and $\sum d_n$ diverges then to does $\sum a_n$. [R Thm 3.25 a]
- A theorem due to Cauchy: If $a_1 \ge a_2 \ge \cdots \ge 0$ then $\sum a_n$ converges if and only if [R Thm 3.27]

$$\sum_{k=1}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + 8a_8 + \ldots$$

  converges.
- Alternating series: Suppose [R Thm 3.43]
  - $|a_1| \ge |a_2| \ge |a_3| \ge \ldots$.
  - $a_{2m-1} \ge 0$ and $a_{2m} \le 0$ for $m = 1, 2, 3, \ldots$
  - $a_n \to 0$.

  Then $\sum a_n$ converges.

- Root test: For a series $\sum a_n$ let [R Thm 3.33]

$$\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

  - If $\alpha < 1$ then $\sum a_n$ converges.
  - If $\alpha > 1$ then $\sum a_n$ diverges.

- Ratio test: The series $\sum a_n$ [R Thm 3.34]

  - converges if

$$\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1;$$

  - diverges if

$$\left|\frac{a_{n+1}}{a_n}\right| \geq 1$$

    for all large n.

- The ratio test is often easier to use.

- The root test is more discriminating, since [R Thm 3.37]

$$\liminf_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| \leq \liminf_{n\to\infty} \sqrt[n]{|a_n|} \leq \limsup_{n\to\infty} \sqrt[n]{|a_n|} \leq \limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|$$

- If

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1$$

  or

$$\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| > 1$$

  then $|a_n| \not\to 0$.

Examples

- For $0 \le x < 1$ [R Thm 3.26]

$$\sum_{n=1}^{\infty} x^{n-1} = \frac{1}{1-x}$$

since

$$\sum_{n=1}^{m} x^{n-1} = \frac{1-x^m}{1-x} \to \frac{1}{1-x}.$$

- For $x \ge 1$

$$\sum_{n=1}^{m} x^{n-1} \ge \sum_{n=1}^{m} 1 = m \to \infty,$$

so the series diverges.

- The number e is defined as

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}.$$

- This is finite since

$$\sum_{n=0}^{\infty} \frac{1}{n!} < 1 + \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} = 1 + 2 = 3.$$

- Alternatively,

$$\frac{a_{n+1}}{a_n} = \frac{n!}{(n+1)!} = \frac{1}{n+1} \to 0.$$

The number e can also be obtained as [R Thm 3.31]

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n.$$

Define

$$s_n = \sum_{k=0}^{n}\frac{1}{n!} \qquad\qquad t_n = \left(1+\frac{1}{n}\right)^n.$$

By the binomial theorem

$$\begin{aligned} t_n &= \sum_{k=0}^{n}\binom{n}{k}\frac{1}{n^k} = \sum_{k=0}^{n}\frac{1}{k!}\left(\frac{n(n-1)\dots(n-k+1)}{n^k}\right) \\ &\le \sum_{k=0}^{n}\frac{1}{k!} = s_n. \end{aligned}$$

So

$$\limsup_{n\to\infty} t_n \le \limsup_{n\to\infty} s_n.$$

For each $m \le n$

$$t_n \ge \sum_{k=0}^{m}\frac{1}{k!}\left(\frac{n(n-1)\dots(n-k+1)}{n^k}\right) = \sum_{k=0}^{m}\frac{1}{k!}\prod_{j=1}^{k}\left(1-\frac{i-1}{n}\right)$$

Fixing m and letting $n\to\infty$ gives

$$\liminf_{n\to\infty} t_n \ge s_m$$

for all m, and therefore

$$\liminf_{n\to\infty} t_n \ge \liminf_{m\to\infty} s_m.$$

Since

$$\liminf_{m\to\infty} s_m = \limsup_{n\to\infty} s_n = e$$

this implies

$$\liminf_{n\to\infty} t_n = \limsup_{n\to\infty} t_n = e.$$

- $\sum \frac{1}{n^p}$ converges if $p > 1$ and diverges if $p \le 1$. [R Thm 3.28]
  - If $p \le 0$ then $\frac{1}{n^p} \ge 1$, so the series diverges.
  - For $p > 1$ the terms $a_n = \frac{1}{n^p}$ are decreasing, so the Cauchy theorem applies:
    $$\sum 2^k a_{2^k} = \sum 2^k \frac{1}{2^{kp}} = \sum 2^{(1-p)k}.$$
    $2^{(1-p)} < 1$ if and only if $p > 1$, so this series converges if and only if $p > 1$.
- The series
  $$\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}$$
  converges if and only if $p > 1$.
  - For $p \le 0$ the series is bounded below by $\sum_{n=2}^{\infty} \frac{1}{n}$, which diverges.
  - For $p > 0$ the terms are decreasing, so the Cauchy theorem applies:
    $$\sum_{k=1}^{\infty} 2^k \frac{1}{2^k (\log 2^k)^p} = \sum_{k=1}^{\infty} \frac{1}{(k \log 2)^p} = \frac{1}{(\log 2)^p} \sum_{k=1}^{\infty} \frac{1}{k^p}$$
  - The series $\sum \frac{1}{k^p}$ converges if and only if $p > 1$.

- For $0 < p \le 1$ and an integer k consider the series

$$\sum_{n=1}^{\infty} n^k p(1-p)^{n-1}.$$

Then

$$\frac{(n+1)^k(1-p)^n x}{n^k(1-p)^{n-1}} = \left(1+\frac{1}{n}\right)^k (1-p) \to 1-p < 1.$$

So the series converges by the ratio test.

- For $a \ge 0$ consider the series

$$\sum_{n=1}^{\infty} x_n = \sum_{n=1}^{\infty} \frac{(-a)^n}{n}.$$

  - For $a < 1$

$$\left|\frac{x_{n+1}}{x_n}\right| \le a\frac{n}{n+1} = \frac{a}{1+1/n} \to a < 1.$$

    So the series converges by the ratio criterion
  - For $a = 1$ the series converges by the alternating series criterion.
  - For $a > 1$ the series diverges by the ratio criterion.

- Consider the series

$$\sum_{n=0}^{\infty} x_n = \sum_{n=0}^{\infty} \frac{n^n}{2^{(n^2)}}.$$

Then

$$\sqrt[n]{x_n} = \frac{n}{2^n} \to 0,$$

so the series converges by the root test.

Power Series

Basic Definitions and Properties

- For a sequence of complex numbers c_n and a complex number z the series

$$\sum_{n=0}^{\infty} c_n z^n$$

  is called a *power series*.

- For a power series $\sum c_n z^n$ the *radius of convergence* is

$$R = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|c_n|}},$$

  with $1/\infty = 0$.

- Then

$$\limsup_{n\to\infty} \sqrt[n]{|c_n z^n|} = |z| \limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{|z|}{R}.$$

- So by the ratio criterion [R Thm 3.39]
  - the series converges if $|z| < R$
  - the series diverges if $|z| > R$

Wednesday, February 10, 2016

Power Series

Examples

- The series
$$\sum n^n z^n$$
has $R = 0$.

- The series
$$\sum \frac{z^n}{n!}$$
has $R = \infty$.

- The series
$$\sum \frac{z^n}{n}$$
has $R = 1$.

- Suppose $p_n \geq 0$ and $\sum p_n = 1$; i.e. p_n is the PMF of a nonnegative integer-valued random variable.
  - Then the radius of convergence of $\sum p_n z^n$ satisfies $R \geq 1$.
  - $R > 1$ if and only if the MGF
$$M(t) = \sum_{n=0}^{\infty} e^{tn} p_n$$
is finite for some $t > 0$, i.e, the MGF is finite in a neighborhood of the origin.

Summation by Parts

Basic Result

- Suppose a_n, b_n are sequences for $n = 0, 1, \ldots$. [R Thm 3.41]
- Let
$$A_n = \begin{cases} \sum_{i=0}^{n} a_n & \text{for } n \geq 0 \\ 0 & \text{n = -1.} \end{cases}$$
- Then for $0 \leq n \leq m$
$$\sum_{k=n}^{m} a_k b_k = \sum_{k=n}^{m-1} A_k (b_k - b_{k+1}) + A_m b_m - A_{n-1} b_n.$$
- To verify, write
$$\begin{aligned} \sum_{k=n}^{m} a_k b_k &= \sum_{k=n}^{m} (A_k - A_{k-1}) b_k \\ &= \sum_{k=n}^{m} A_k b_k - \sum_{k=n}^{m} A_{k-1} b_k \\ &= \sum_{k=n}^{m} A_k b_k - \sum_{k=n-1}^{m-1} A_k b_{k+1} \\ &= \sum_{k=n}^{m-1} A_k b_k - \sum_{k=n}^{m-1} A_k b_{k+1} + A_m b_m - A_{n-1} b_n \\ &= \sum_{k=n}^{m-1} A_k (b_k - b_{k+1}) + A_m b_m - A_{n-1} b_n. \end{aligned}$$

Examples

- Find $\sum_{k=1}^{m} kx^{k-1}$ for $x \neq 1$:
  - Let $a_n = x^{n-1}$ and $b_n = n$.
  - Then $A_n = \frac{1-x^n}{1-x}$ for $n \geq 0$.
  - So

$$\begin{aligned}
\sum_{k=1}^{m} kx^{k-1} &= \sum_{k=1}^{m-1} \frac{1-x^k}{1-x}(-1) + \frac{1-x^m}{1-x}m \\
&= \frac{1}{1-x}\left(1-m+\frac{x-x^m}{1-x}+m(1-x^m)\right) \\
&= \frac{1}{1-x}\left(1+\frac{x-x^m}{1-x}-mx^m\right) \\
&= \frac{1}{(1-x)^2}\left(1-x^m-mx^m(1-x)\right) \\
&= \frac{1-(m+1)x^m+mx^{m+1}}{(1-x)^2}
\end{aligned}$$

- Find $\sum_{k=1}^{m} k^2$.
  - Let $a_n = 1$ and $b_n = n^2$.
  - Then $A_n = n$ for $n \geq 1$ and $A_0 = 0$.
  - So

$$\begin{aligned}
\sum_{k=1}^{m} k^2 &= \sum_{k=1}^{m-1} k(k^2 - (k+1)^2) + m^3 - 0 \\
&= -\sum_{k=1}^{m-1} (2k^2 + k) + m^3 \\
&= -2\sum_{k=1}^{m-1} k^2 - \sum_{k=1}^{m-1} k + m^3 \\
&= -2\sum_{k=1}^{m} k^2 + 2m^2 - \frac{m(m-1)}{2} + m^3 \\
&= -2\sum_{k=1}^{m} k^2 + \frac{3m^2 + m + 2m^3}{2} \\
&= -2\sum_{k=1}^{m} k^2 + \frac{(2m+1)(m+1)m}{2}
\end{aligned}$$

  - Solving for $\sum_{k=1}^{m} k^2$ then gives

$$\sum_{k=1}^{m} k^2 = \frac{(2m+1)(m+1)m}{6}.$$

- Suppose [R Thm 3.42]

  (a) The partial sums A_n are bounded.

  (b) $b_1 \geq b_2 \geq b_3 \geq \cdots \geq 0$.

  (c) $\lim b_n = 0$.

  Then $\sum a_n b_n$ converges.

  - Choose M so $|A_n| \leq M$ for all n.
  - Given $\varepsilon > 0$ choose N such that $b_N < \frac{\varepsilon}{2M}$.
  - For $m \geq n \geq N$

$$\begin{aligned}\left|\sum_{k=n}^{m} a_k b_k\right| &= \left|\sum_{k=n}^{m-1} A_k(b_k - b_{k+1}) + A_m b_m - A_{n-1} b_n\right| \\ &\leq M\left(\sum_{k=n}^{m-1}(b_k - b_{k+1}) + b_m + b_n\right) \\ &= 2Mb_n \\ &\leq 2Mb_N < \varepsilon.\end{aligned}$$

  - So $\sum a_n b_n$ is a Cauchy sequence and thus converges.

- Alternating series: Suppose [R Thm 3.43]

  (a) $|c_1| \geq |c_2| \geq |c_3| \geq \ldots$.

  (b) $c_{2m-1} \geq 0$ and $c_{2m} \leq 0$ for $m = 1, 2, 3, \ldots$

  (c) $c_n \to 0$.

  Then $\sum c_n$ converges.

  – Let $a_n = (-1)^{n+1}$ and $b_n = |c_n|$ in Rudin's Theorem 3.42.

- Suppose [R Thm 3.44]

  (a) $c_1 \geq c_2 \geq c_3 \geq \ldots$.

  (b) the radius of convergence of $\sum c_n z^n$ is $R = 1$

  (c) $c_n \to 0$.

  Then $\sum c_n z^n$ converges for every complex z with $|z| = 1$ except possibly $z = 1$.

  – Let $a_n = z^n$ and $b_n = c_n$.

  – Then for z with $|z| \leq 1$ and $z \neq 1$

$$|A_n| = \left| \sum_{k=0}^{n} z^k \right| = \left| \frac{1 - z^{m+1}}{1 - z} \right| \leq \left| \frac{2}{1 - z} \right|.$$

  – So Rudin's Theorem 3.42 applies.

Absolute Convergence

- The series $\sum a_n$ converges absolutely if $\sum |a_n|$ converges.
- If $\sum |a_n|$ converges then so does $\sum a_n$.
- For nonnegative series convergence and absolute convergence are equivalent.
- Most standard convergence tests (e.g. root, ratio) are test of absolute convergence.
- Power series converge absolutely on the interior of the circle of convergence.
- A *rearrangement* of a series $\sum a_n$ is a series $\sum b_n$ with
  - $b_n = a_{k_n}$
  - k_n is a sequence of positive integers in which every positive integer occurs exactly once.
- If $\sum a_n$ converges absolutely then every rearrangement converges and to the same limit.
- If $\sum a_n$ is a series of real numbers that converges but not absolutely, then for any α, β with
$$-\infty \le \alpha \le \beta \le +\infty$$
there exists a rearrangement b_n with partial sums B_n such that
$$\liminf_{n\to\infty} B_n = \alpha$$
$$\limsup_{n\to\infty} B_n = \beta$$

Addition and Multiplication of Series

Basic Definitions and Properties

- If $\sum a_n = A$ and $\sum b_n = B$, then $\sum(a_n + b_n) = A + B$.
- The product, or *Cauchy product*, of two series $\sum a_n$ and $\sum b_m$ is the series $\sum c_n$ with
$$c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$
- c_n is also called the *convolution* of a_n and b_n.
- Motivation: For power series $\sum a_n z^n$ and $\sum b_n z^n$
$$\left(\sum a_n \sum b_n\right) = \sum \left(\sum_{k=0}^{n} a_k b_{n-k} \right) z^n.$$
- Suppose $\sum a_n = A$ and $\sum b_n = B$ and at least one of these series converges absolutely. Then [R 3.50]
$$\sum c_n = AB.$$
- If neither $\sum a_n$ nor $\sum b_n$ converges absolutely then $\sum c_n$ may diverge.

Example

Suppose a_n, b_n are non-negative, $\sum a_n = \sum b_n = 1$, and both $\sum a_n e^{tn}$ and $\sum b_n e^{tn}$ converges for some t. Then

$$\left(\sum a_n e^{tn}\right)\left(\sum b_n e^{tn}\right) = \sum c_n e^{tn}$$

with

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

Friday, February 12, 2016

Convergence of Functions

Basic Definitions and Properties

- Let [R Def 4.1]
  - (X, d_X), (Y, d_Y) be metric spaces,
  - A a subset of X,
  - $f : A \to Y$,
  - $p \in X$ a limit point of A, and $q \in Y$.

  Then $f(x) \to q$ as $x \to p$, or

  $$\lim_{x \to p} f(x) = q$$

  if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

  $$d_Y(f(x), q) < \varepsilon$$

  whenever $x \in A$ and $d_X(x, p) < \delta$.
- The definition does not require $p \in A$, or that f be defined at p.
- If f is defined at p we can have $f(x) \to q$ as $x \to p$ but $q \neq f(p)$.
- For general topological spaces, $f(x) \to q$ as $x \to p$ if for any open neighborhood V of q there exists an open neighborhood U of p such that

  $$f(x) \in V$$

  whenever $x \in U \cap A$.
- For metric spaces, $\lim_{x \to p} f(x) = q$ if and only if [R Thm 4.2]

  $$\lim_{n \to \infty} f(p_n) = q$$

  for every sequence $p_n \in A$ with $p_n \neq p$ and $p_n \to p$.
- For metric spaces, $f(x) \to q$ and $f(x) \to q'$ as $x \to p$ implies $q = q'$.

Example

- The function
$$F(x) = \begin{cases} 1 - \frac{1}{2}e^{-x} & \text{for } x \le 0 \\ 0 & \text{otherwise} \end{cases}$$
as a function on $A = (-\infty, 0)$ has a limit as $x \to 0$
$$\lim_{x<0, x\to 0} F(x) = \lim_{x\uparrow 0} F(x) = 0.$$

- With $A = (0, \infty)$ or $A = [0, \infty)$.
$$\lim_{x>0, x\to 0} F(x) = \lim_{x\downarrow 0} F(x) = \frac{1}{2} = F(0).$$

- With (X, d) a two-point compactification of $\mathbb{R}$ and $A = \mathbb{R}$
$$\lim_{x\to\infty} F(x) = 1.$$

- With $A = \mathbb{R}$, for any $p \neq 0$
$$\lim_{x\to p} F(x) = F(p).$$

Continuous Functions

Basic Definitions and Properties

- Let [R Def 4.5]
  - (X, d_X), (Y, d_Y) be metric spaces,
  - $f : X \to Y$,

  Then $f(x)$ is continuous as a point $p \in X$ if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that
  $$d_Y(f(x), f(p)) < \varepsilon$$
  whenever $d_X(x, p) < \delta$.

- For general topological spaces, f is continuous at p if for any open neighborhood V of $f(p)$ there exists an open neighborhood U of p such that [KF §9.6]
  $$f(x) \in V$$
  whenever $x \in U$, or $f(U) \subset V$.

- f is continuous on X if it is continuous at every $p \in X$.

- It p is a limit point of X then f is continuous at p if and only if [R Thm 4.6]
  $$\lim_{x \to p} f(x) = f(p).$$

- If p is an isolated point of X then f is considered to be continuous at p.

- Rudin defines continuity for a function defined on a subset ut the says this isn't necessary. [R Rem 4.12]

- Suppose X, Y, Z are metric spaces, $f : X \to Y, g : Y \to Z$ are functions, and $h : X \to Z$ is defined as [R Thm 4.5]

$$h(x) = g(f(x))$$

  - h is the composition of f and g.
  - Common notation: $h = g \circ f$.

- If f is continuous at x and g is continuous at $y = f(x)$, then h is continuous at x.

  - Let V be an open neighborhood of $h(x) = g(y)$.
  - Since g is continuous at y there is an open neighborhood W of y such that $g(W) \subset V$.
  - Since f is continuous at x there is an open neighborhood U of x such that $f(U) \subset W$.
  - So $h(U) = g(f(U)) \subset V$.

- Suppose $f_1, \ldots, f_k$ are real valued functions and $f : X \to \mathbb{R}^k$ is defined as [R Thm 4.10]

$$f(x) = (f_1(x), \ldots, f_k(x)).$$

  Then f is continuous if and only if each of the f_i is continuous.

- $f(x,y) = x + y$ is continuous on $\mathbb{R}^2$.
- $f(x,y) = xy$ is continuous on $\mathbb{R}^2$.
- $f(x,y) = x/y$ is continuous on $\{(x,y) \in \mathbb{R} : y \neq 0\}$.
- Combining these: if f, g are continuous then

  - $f(x) + g(x)$ is continuous
  - $f(x)g(x)$ is continuous
  - $f(x)/g(x)$ is continuous at x with $g(x) \neq 0$.

- These results hold for complex-valued functions as well.
- Analogous results hold for vector-valued functions.

Examples

- If $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, then the *coordinate functions*

$$f_i(x) = x_i$$

  are continuous.

- For any nonnegative integers $n_1, \ldots, n_k$

$$f(x) = x_1^{n_1} \ldots x_k^{n_k}$$

  is continuous.

- Any multinomial, i.e. any sum of finitely many terms of the form

$$\sum c_{n_1 \ldots n_k} x_1^{n_1} \ldots x_k^{n_k}$$

  is continuous.

Uniform Convergence and Power Series

- A sequence of functions $f_n : X \to Y$ *converges uniformly* to a function $f : X \to Y$ if
$$\lim_{n\to\infty} \sup_{x\in X} d_Y(f_n(x), f(x)) = 0$$

- As shown previously, if $f_n : X \to Y$ are continuous functions and f_n converges uniformly to a function $f : X \to Y$, then f is continuous.

- Suppose $\sum c_n z^n$ has radius of convergence R, and define
$$f(z) = \sum_{n=0}^{\infty} c_n z^n$$
for $|z| < R$. Then f is continuous.

- For any $R_0 < R$ and $z \le R_0$
$$\sum_{n=m}^{\infty} |c_n||z|^n \le \sum_{n=m}^{\infty} |c_n| R_0^n \to 0.$$

- So for any $R_0 < R$ the functions
$$f_n(z) = \sum_{k=1}^{n} c_k z^k$$
converge uniformly to $f(z)$ for $|z| < R_0$.

- So $f(z)$ is continuous on the interior of the circle of convergence.

- The function
$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$
is continuous.

A Useful Characterization of Continuity

A function $f : X \to Y$ is continuous if and only if for every open set $G \subset Y$ the inverse image $f^{-1}(G)$ is an open subset of X. [R Thm 4.8 KF §9 Thm 10]

- Suppose f is continuous and let G be an open subset of Y.
  - For any $x \in f^{-1}(G)$ let $y = f(x)$.
  - Then $y \in G$, and G is an open neighborhood of y.
  - Since f is continuous, there exists an open neighborhood U_x of x such that $f(U_x) \subset G$, i.e. $U_x \subset f^{-1}(G)$.
  - So every point $x \in f^{-1}(G)$ has an open neighborhood $U_x \subset f^{-1}(G)$ and therefore $f^{-1}(G)$ is open.
- Conversely, suppose $f^{-1}(G)$ is open whenever $G \subset Y$ is open.
  - For any $x \in X$ let $V \subset Y$ be an open neighborhood of $y = f(x)$.
  - Then $U = f^{-1}(V)$ is open and $x \in U$.
  - So U is an open neighborhood of x with $f(U) \subset V$.
  - So f is continuous at x, for any $x \in X$.

Corollary: A function $f : X \to Y$ is continuous if and only if for every closed set $F \subset Y$ the inverse image $f^{-1}(F)$ is a closed subset of X.

Continuity and Compactness

- Let X, Y be topological spaces, $f : X \to Y$ a continuous function, and suppose that X is compact. Then the image $f(X)$ is lso compact. [R Thm 4.14 KF §10 Thm 5]
  - Let $\{V_\alpha\}$ be an open cover of $f(X)$.
  - Let $U_\alpha = f^{-1}(V_\alpha)$
  - Then $\{U_\alpha\}$ is a cover of X, and continuity of f implies that the U_α are open.
  - Since X is compact, there is a finite subcover $\{U_{\alpha_1}, \ldots, U_{\alpha_n}\}$ with
    $$X \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}.$$
  - Since
    $$f(f^{-1}(A)) \subset A$$
    for any set A, and
    $$f(A \cup B) = f(A) \cup f(B)$$
    for any sets A, B, this implies that
    $$f(X) \subset f(U_{\alpha_1}) \cup \cdots \cup f(U_{\alpha_n}) \subset V_{\alpha_1} \cup \cdots \cup V_{\alpha_n}$$
  - So the open cover $\{V_\alpha\}$ contains the finite subcover $\{V_{\alpha_1}, \ldots, V_{\alpha_n}\}$.
- Suppose f is a continuous one-to-one mapping of a compact topological space X onto a topological space Y. Then the inverse function [R Thm 4.17 KF §10 Thm 6]
  $$g(y) = f^{-1}(y)$$
  is continuous.
  - Let F be a closed set subset of X.
  - Then $f(F) = g^{-1}(F)$ is compact and therefore closed.

- Suppose f is a continuous real-valued function on a compact space X. Then $f(X)$ is closed and bounded. [R Thm 4.16]
- This means there exist points $x_m, x_M \in X$ such that

$$f(x_m) = \inf_{x \in X} f(x)$$
$$f(x_M) = \sup_{x \in X} f(x)$$

- If a likelihood function is continuous and the parameter space is compact then a maximum likelihood estimator exists.
- If we only care about maxima (or minima) we only need half of continuity.
- A real-valued function is *upper semicontinuous* at x_0 if for any $\varepsilon > 0$ there exists an open neighborhood U of x_0 with [KF §12 Def 2]

$$f(x) < f(x_0) + \varepsilon$$

  for all $x \in U$.
- Lower semicontinuity is defined analogously.
- A finite upper semicontinuous function on a compact topological space is bounded from above.
- A finite upper semicontinuous function on a compact topological space attains its least upper bound.
- A common process in establishing the existence of a maximum:
  - establish that the function is upper semicontinuous;
  - argue that there is a compact set K with a function value in; K that is larger than all function values in K^c;
  - conclude that the maximum is attained in K and is a global maximum.

Example

- Let $X_1, \ldots, X_n$ be independent and uniformly distributed on the interval from zero to $\theta > 0$.
- Using the density
$$f_1(x|\theta) = \frac{1}{\theta} 1_{[0,\theta]}(x)$$
leads to the likelihood function
$$L_1(\theta|X) = \frac{1}{\theta^n} 1_{[X_{(n)},\infty)}(\theta).$$
- Using the density
$$f_2(x|\theta) = \frac{1}{\theta} 1_{(0,\theta)}(x)$$
leads to the likelihood function
$$L_2(\theta|X) = \frac{1}{\theta^n} 1_{(X_{(n)},\infty)}(\theta).$$
- L_1 is upper semicontinuous and attains its maximum at $\widehat{\theta} = X_{(n)}$.
- L_2 is not upper semicontinuous and does not attain its maximum — the MLE does not exist.

Week 5

Monday, February 15, 2016

Uniform Continuity

- A function f on a metric space X with values in a metric space Y is *uniformly continuous* if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that [R Def 4.18]

$$d_Y(f(x), f(y)) < \varepsilon$$

  for any $x, y \in X$ with $d_X(x, y) < \delta$.

- If $f : X \to Y$ is continuous and X is compact then f is uniformly continuous. [R Thm 4.19]

- Outline of a proof:
  - Fix $\varepsilon > 0$.
  - For each x find $\delta_x > 0$ such that $d_X(y,x) < \delta_x$ implies

$$d_Y(f(x), f(y)) < \frac{\varepsilon}{2}.$$

  - Let

$$V_x = \{y : d_X(y,x) < \delta_x/2\}.$$

  - $\{V_x\}$ is an open cover of X, so there is a finite subcover $\{V_{x_1}, \ldots, V_{x_n}\}$.
  - Let $\delta = \frac{1}{2}\min(\delta_{x_1}, \ldots, \delta_{x_n})$.
  - Suppose $d_X(x,y) < \delta$. Then for some k

$$d_X(x, x_k) \leq \frac{1}{2}\delta_{x_k}.$$

  - Then

$$d_X(y, x_k) \leq d_X(x,y) + d_X(x, x_k) < \delta + \frac{1}{2}\delta_{x_k} < \delta_{x_k}.$$

  - So

$$d_Y(f(x), f(y)) \leq d_Y(f(x), f(x_k)) + d_Y(f(x_k), f(y)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Equicontinuous Functions

Basic Definitions and Properties

- A family $\mathscr{F}$ of functions on a metric space X with values in a metric space Y is *equicontinuous* if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that [KF §11 Def 4]

$$d_Y(f(x), f(y)) < \varepsilon$$

for all $f \in \mathscr{F}$ and all $x, y \in X$ with

$$d_X(x, y) < \delta.$$

- The family $\mathscr{F}$ is *uniformly bounded* if [KF §11 Def 3]

$$\sup_{x \in X} \operatorname{diam}\{f(x) : f \in \mathscr{F}\} < \infty.$$

- A finite collection $\{f_1, \ldots, f_k\}$ of continuous functions on a compact metric space X is uniformly bounded and equicontinuous.
  - For each function $f_i(X)$ is bounded; the maximum of the bounds is a bound on all functions.
  - Each function is uniformly continuous; given $\varepsilon > 0$ take the minimum of the δ_i.

Arzelà-Ascoly Theorem

A family $\mathscr{F}$ of continuous in $C_{[0,1]}$ is relatively compact if and only if it is equicontinuous and uniformly bounded.

- The proof uses the equivalence of relative compactness and total boundedness.
- Outline of necessity proof:
  - Suppose $\mathscr{F}$ is relatively compact and hence totally bounded.
  - Given $\varepsilon > 0$ find a finite $\varepsilon/3$ net.
  - This finite set of functions is uniformly bounded and equicontinuous.
  - Every $f \in \mathscr{F}$ is within $\varepsilon/3$ of a function in the net, so $\mathscr{F}$ is uniformly bounded.
  - Given $\varepsilon > 0$ pick δ so that for every g in the net

    $$|g(x) - g(y)| < \varepsilon/3$$

    when $|x-y| < \delta$.
  - Then for any $f \in \mathscr{F}$ find g in the net that is within $\varepsilon/3$ of f.
  - Then for any x, y with $|x-y| < \delta$

    $$\begin{aligned} |f(x) - f(y)| &\le |f(x) - g(x)| + |g(x) - g(y)| + |g(y) - f(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon. \end{aligned}$$

- This direction holds for functions from a general compact metric space X into a general metric space Y.

- Outline of the sufficiency proof:
  - Suppose $\mathscr{F}$ is uniformly bounded and equicontinuous.
  - Let K satisfy $|f(x)| \le K$ for all $x \in [0,1]$ and $f \in \mathscr{F}$.
  - Given $\varepsilon > 0$ choose $\delta > 0$ such that

    $$|f(x) - f(y)| < \frac{\varepsilon}{5}$$

    for all $f \in \mathscr{F}$ when $|x - y| < \delta$.
  - Choose numbers $x_0, x_1, \ldots, x_n$ and $y_0, y_1, \ldots, y_m$ such that

    $$0 = x_0 < x_1 < x_2 < \cdots < x_n = 1$$
    $$-K = y_0 < y_1 < y_2 < \cdots < y_m = K$$

    with

    $$x_{i+1} - x_i < \delta$$
    $$y_{j+1} - y_j < \frac{\varepsilon}{5}$$

    for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.
  - The finite set of continuous piece-wise linear functions with vertices at the points (x_i, y_j) is an ε-net for $\mathscr{F}$.
- This argument depends on properties of the real line.
- Other approaches allow the result to be generalized to general metric spaces.

A More General Version

Suppose $\mathscr{F}$ is an equicontinuous collection of functions from a compact metric space X into a metric space Y such that for each $x \in X$ the set

$$V_x = \{f(x) : f \in \mathscr{F}\}$$

is a relatively compact subset of Y. Then there is a sequence $g_n \in \mathscr{F}$ that converges uniformly.

- A compact metric space is separable, i.e. it has a countable dense subset $\{x_1, x_2, \dots\}$
- Since V_{x_1} is relatively compact, there is a sequence of functions $f_{1,n} \in \mathscr{F}$ such that $f_{1,n}(x_1)$ converges to a value $y_1 \in Y$.
- From this sequence extract a subsequence $f_{2,n}$ such that $f_{2,n}(x_2)$ has a limit y_2, and so on.
- Let $g_n = f_{n,n}$ be the *diagonal subsequence*.
- For each n the sequence $\{g_i : i \geq n\}$ is a subsequence of $\{f_{n,j} : j \geq 1\}$, so

  $$\lim_{n\to\infty} g_n(x_k) = y_k$$

  for all k.

- Using equicontinuity, given $\varepsilon > 0$ chose $\delta > 0$ such that

$$d_Y(f(x), f(y)) < \frac{\varepsilon}{3}$$

for all $f \in \mathscr{F}$ when $d_X(x,y) < \delta$.

- Since $\{x_n\}$ is dense in X and X is compact there is an integer M such that the sets

$$U_i = \{x : d_x(x, x_i) < \delta\}$$

for $i = 1, \ldots, M$ is a cover of X.

- The sequences $g_n(x)$ are Cauchy sequences for each $x \in \{x_1, \ldots, x_M\}$, so there exists an integer N such that

$$d_Y(g_n(x), g_m(x)) < \frac{\varepsilon}{3}$$

when $n, m \geq N$ and all $x \in \{x_1, \ldots, x_M\}$.

- $X \subset U_1 \cup \cdots \cup U_M$, so for each $x \in X$ there is a $k \leq M$ such that $x \in U_k$, and for $n, m \geq N$

$$\begin{aligned} d_Y(g_n(x), g_m(x)) &\leq d_Y(g_n(x), g_n(x_k)) + d_Y(g_n(x_k), g_m(x_k)) + d_Y(g_m(x_k), g_m(x)) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

- Thus g_n is uniformly Cauchy and hence has a continuous limit to which it converges uniformly.

Wednesday, February 17, 2016

Consistency of the MLE

Information Inequality

- Suppose $X_1, X_2, \ldots$ are independent random variables with density $f(x|\theta_0)$
- By Jensen's inequality, for any θ

$$\begin{aligned} T(\theta) = E\left[\log\left(\frac{f(X_1|\theta)}{f(X_1|\theta_0)}\right)\right] &\le \log E\left[\frac{f(X_1|\theta)}{f(X_1|\theta_0)}\right] \\ &= \log \int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\,dx \\ &= \log \int f(x|\theta)\,dx \\ &= 0. \end{aligned}$$

- Since the logarithm is strictly concave, equality holds only if $f(X|\theta) = f(X|\theta_0)$ almost surely, i.e. if and only if the two densities are identical.
- θ is said to be *identifiable* if this cannot happen.
- If θ is identifiable, then $T(\theta) \le 0$ for all θ, and $T(\theta) = 0$ if and only if $\theta = \theta_0$.

Finite Parameter Space

- Suppose $\Theta = \{\theta_0, \theta_1, \ldots, \theta_k\}$.
- Let
$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i|\theta)}{f(X_i|\theta_0)}\right).$$
- By strong the law of large numbers
$$\begin{aligned} L_n(\theta_0) &\to T(\theta_0) = 0 \quad \text{almost surely} \\ L_n(\theta_i) &\to T(\theta_i) < 0 \quad \text{almost surely, for } i = 1, \ldots, k \end{aligned}$$
- Chose z such that
$$\max_{1 \le i \le k} T(\theta_i) < z < 0.$$
- Then almost surely for all large n
$$\max_{1 \le i \le k} L_n(\theta_i) < z < L_n(\theta_0).$$
- So almost surely for all large n the value among $\theta_0, \theta_1, \ldots, \theta_k$ that maximizes the likelihood is the correct value θ_0.
- The finite set of parameter values
  - allows a value z to be fit between $\max_{1 \le i \le k} T(\theta_i)$ and $T(\theta_0) = 0$;
  - to obtain a result that holds almost surely for all parameter values simultaneously.

  We can extend the result to an uncountable parameter space if we
  - assume the parameter space is compact;
  - assume the right level of continuity in T and L_n.

Compact Parameter Space

- Suppose the parameter space Θ is compact.
- Suppose there is a function $D(x)$ such that
  - for all x and all $\theta, \theta' \in \Theta$

$$|\log f(x|\theta) - \log f(x|\theta')| \le |\theta - \theta'| D(x)$$

  - $M = E[D(X_1)] = \int D(x) f(x|\theta_0) dx < \infty$.
- This assumption ensures that
  - the log likelihood functions $L_n(\theta)$ are continuous and

$$|L_n(\theta) - L_n(\theta')| \le |\theta - \theta'| \frac{1}{n} \sum_{i=1}^{n} D(X_i) = |\theta - \theta'| \overline{D}_n;$$

  - the function $T(\theta)$ is continuous, and

$$|T(\theta) - T(\theta')| \le |\theta - \theta'| M.$$

- Need to show: for any $\varepsilon > 0$ almost surely for all large n

$$\sup_{\theta : |\theta - \theta_0| \ge \varepsilon} L_n(\theta) \le \sup_{\theta : |\theta - \theta_0| < \varepsilon} L_n(\theta)$$

- We still have $L_n(\theta_0) \to 0$ almost surely.
- So it is enough to showy that, given $\varepsilon > 0$, there is a value $z < 0$ such that almost surely for all large n

$$\sup_{\theta : |\theta - \theta_0| \ge \varepsilon} L_n(\theta) < z.$$

- Fix $\varepsilon > 0$, and let

$$K = \{\theta \in \Theta : |\theta - \theta_0| \geq \varepsilon\}.$$

- K is a closed subset of the compact set Θ and therefore compact.
- T is continuous and $T(\theta) < 0$ for all $\theta \in K$, so

$$\sup_{\theta \in K} T(\theta) < 0$$

- Choose z and z' such that

$$\sup_{\theta \in K} T(\theta) < z' < z < 0.$$

- For each $\theta \in K$ let

$$V_\theta = \left\{ \theta' : |\theta' - \theta| < \frac{z' - T(\theta)}{M} \right\}.$$

- $\{V_\theta\}$ is an open cover of K, so it contains a finite subcover $\{V_{\theta_1}, \ldots, V_{\theta_l}\}$.
- Now for each $i = 1, \ldots, k$

$$\begin{aligned} \sup_{\theta \in V_{\theta_i}} L_n(\theta) &\leq L_n(\theta_i) + \frac{z' - T(\theta_i)}{M} \overline{D_n} \\ &\to T(\theta_i) + \frac{z' - T(\theta_i)}{M} M \\ &= z' < z \end{aligned}$$

  almost surely.

- So almost surely for all large n

$$\sup_{\theta \in K} L_n(\theta) \leq \max_{1 \leq i \leq k} \sup_{\theta \in V_{\theta_i}} L_n(\theta) < z.$$

An Alternative Approach

- Let S be the sample space, and for $s \in S$ let

$$L_n(\theta)(s) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i(s)|\theta)}{f(X_i(s)|\theta_0)}\right)$$
$$\overline{D}_n(s) = \frac{1}{n}\sum_{i=1}^{n} D(X_i(s)).$$

- Let $\{\theta_0, \theta_1, \theta_2, \dots\}$ be a dense subset of Θ.
- Let $A \subset S$ be the set of all points in the sample space such that

$$L_n(\theta_k)(s) \to T(\theta_k)$$
$$\overline{D}_n(s) \to M.$$

- By the strong law of large numbers $P(A) = 1$.
- For each $s \in A$
  - the sequence of real numbers $\overline{D}_n(s)$ converges and is therefore bounded;
  - this implies that the sequence of functions $L_n(\theta)(s)$ is uniformly bounded and equicontinuous;
  - Since the sequence $L_n(\theta)(s)$ also converges point-wise to $T(\theta)$ on a dense subset of a compact set this means that

$$\sup_{\theta\in\Theta} |L_n(\theta)(s) - T(\theta)| \to 0.$$

  - As before, given $\varepsilon > 0$ let

$$K = \{\theta \in \Theta : |\theta - \theta_0| \geq \varepsilon\},$$

    which is compact, and choose z with

$$\sup_{\theta\in K} T(\theta) < z < T(\theta_0) = 0.$$

- Since $L_n(\theta_0)(s) \to T(\theta_0) = 0$,

$$\sup_{\theta:|\theta-\theta_0|<\varepsilon} L_n(\theta)(s) > z$$

  for all large n.

- Uniform convergence implies that for all large n

$$\sup_{\theta \in K} L_n(\theta)(s) < z.$$

- So for all large n all maxima of $L_n(\theta)(s)$ are within ε of θ_0.

Some Notes

- These arguments to *not* guarantee uniqueness of the MLE.
- Showing convergence for an unbounded parameter space usually involves two steps:
  - show that for all large n the MLE will be in a compact set containing θ_0 in its interior;
  - use the compact parameter space result after that.
- Useful general sufficient conditions for unbounded parameter spaces are hard to come by.
- Conditions for special classes, like exponential families, are easier.
- There are a number of settings where the global MLE is *not* consistent.
- Under fairly reasonable conditions there exists a sequence of unique local maxima of the likelihood that converge to the true parameter value.

Continuity and Connectedness

- If $f : X \to Y$ is continuous and $E \subset X$ is connected, then $f(E)$ is connected. [R Thm 4.22]
- Corollary: If f is a continuous real valued function on $[a,b]$ and $f(a) < f(b)$ then for any c with $f(a) < c < f(b)$ there exists an $x \in (a,b)$ such that $f(x) = c$. This is called the *intermediate value theorem*. [R Thm 4.23]
- Outline of Proof:
  - Suppose $f(E)$ is not connected.
  - Then there exist non-empty sets $A, B \subset Y$ such that $f(E) = A \cup B$ and $\overline{A} \cap B = A \cap \overline{B} = \emptyset$.
  - Let

$$G = E \cap f^{-1}(A)$$
$$H = E \cap f^{-1}(B).$$

  - Then $E = G \cup H$ and both G, H are non-empty.
  - Since $A \subset \overline{A}$ we have $G \subset f^{-1}(\overline{A})$.
  - Since f is continuous, $f^{-1}(\overline{A})$ is closed.
  - This implies $\overline{G} \subset f^{-1}(\overline{A})$.
  - Since $f(H) = B$ and $\overline{A} \cap B = \emptyset$, this implies $\overline{G} \cap H = \emptyset$.
  - Similarly $G \cap \overline{H} = \emptyset$.
  - So E is not connected, a contradiction.

Discontinuities

- For a function $f : \mathbb{R} \to \mathbb{R}$ define the right hand and left hand limits at x, if they exist, as [R Def 4.25]

$$f(x+) = \lim_{y \downarrow x} f(y) = \lim_{y > x, y \to x} f(y)$$
$$f(x-) = \lim_{y \uparrow x} f(y) = \lim_{y < x, y \to x} f(y)$$

- f is continuous at x if and only if $f(x+)$ and $f(x-)$ both exist and

$$f(x+) = f(x-) = f(x).$$

- If f defined on an open interval containing x is discontinuous as x and both $f(x+)$ and $f(x-)$ exist, then the discontinuity is [R Def 4.26]
  - a *discontinuity of the first kind*, or
  - a *simple discontinuity*.
- Otherwise, it is a *discontinuity of the second kind*.

Examples

- The function

$$f(x) = \begin{cases} \sin(1/x) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

  has a discontinuity of the second kind at $x = 0$.

- The function

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

  has a discontinuity of the second kind at every x.

- A CDF can only have simple discontinuities.

Monotonic Functions

- A real-valued function f defined on an interval of the real line is *monotonically increasing* if $x < y$ implies $f(x) \leq f(y)$ for all x, y. [R Def 4.28]
- f is *strictly monotonically increasing* if $x < y$ implies $f(x) < f(y)$ for all x, y.
- Decreasing functions are analogous.
- If f is monotonic then $f(x+)$ and $f(x-)$ always exist, and for $x < y$ [R Thm 4.29]

$$f(x+) \leq f(y-).$$

- As a result,
  - monotonic functions have no discontinuities of the second kind;
  - monotonic functions have at most countably many discontinuities. [R Thm 4.30]

Week 6

Monday, February 22, 2016

Differentiation

Basic Definitions and Properties

- For a real-valued function f defined on an interval $[a,b]$ define its *derivative* at a point $x \in [a,b]$ as [R Def 5.1]

$$f'(x) = \frac{df(x)}{dx} = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}$$

  if the limit exists.

- f is said to be *differentiable* at x if the limit exists.

- If f is differentiable at x then f is continuous at x, since [R Thm 5.2]

$$f(y) - f(x) = \frac{f(y) - f(x)}{y - x}(y - x) \to f'(x) \times 0 = 0$$

  as $y \to x$.

- f can be continuous and not differentiable at isolated points.
  - For example, $f(x) = |x|$ is continuous but not differentiable at $x = 0$.

- It is possible for f to be continuous everywhere but differentiable nowhere.
  - Almost all sample paths of Brownian motion have this property.

Basic Differentiation Rules

- Suppose $f, g : [a,b[\to \mathbb{R}$ are differentiable at $x \in [a,b]$. Then [R Thm 5.3]

  addition rule: $f+g$ is differentiable at x with

  $$(f+g)'(x) = f'(x) + g'(x);$$

  multiplication rule: fg is differentiable at x with

  $$(fg)'(x) = f'(x)g(x) + f(x)g'(x);$$

  quotient rule: f/g is differentiable at x if $g(x) \neq 0$ with

  $$\left(\frac{f}{g}\right)'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{g(x)^2}.$$

- **Chain rule:** Suppose [R Thm 5.5]
  - $f : [a,b] \to \mathbb{R}$ is differentiable at $x \in [a,b]$;
  - $g : f([a,b]) \to \mathbb{R}$ is differentiable at $f(x)$.

  Then $h(x) = g(f(x))$ is differentiable at x with

  $$h'(x) = g'(f(x))f'(x).$$

- **Power Series**: If

  $$f(x) = \sum_{n=0}^{\infty} c_n x^n$$

  has radius of convergence R, then f is differentiable for $|x| < R$ with

  $$f'(x) = \sum_{n=1}^{\infty} n c_n x^{n-1};$$

  this power series also has radius of convergence R.

Examples

- For $f(x) = c$, a constant, $f'(x) = 0$.
- For $f(x) = x$, $f'(x) = 1$.
- For integers $n > 0$ let $f_n(x) = x^n$. Then

$$f_n'(x) = (f_{n-1}(x)x)' = f_{n-1}'(x)x + f_{n-1}(x).$$

So

$$\begin{aligned} f_2'(x) &= 2x \\ f_3'(x) &= 3x^2 \\ &\vdots \\ f_n'(x) &= nx^{n-1} \end{aligned}$$

- $f(x) = e^x = \sum \frac{x^n}{n!}$, so

$$f'(x) = \sum_{n=1}^{\infty} n\frac{x^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x.$$

- If $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, then

$$f'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}(-x) = -xf(x).$$

- For $x > 0$ let $f(x) = \log(x)$. Then $e^{\log(x)} = x$, so

$$1 = \frac{d}{dx}\left(e^{\log(x)}\right) = e^{\log(x)}\frac{d}{dx}\log(x) = x\frac{d}{dx}\log(x) = xf'(x),$$

and therefore $f'(x) = \frac{1}{x}$.

- For $x > 0$ and $\alpha \neq 0$ let $f(x) = x^\alpha = e^{\alpha \log(x)}$. Then

$$f'(x) = \frac{d}{dx}\left(e^{\alpha\log(x)}\right) = e^{\alpha\log(x)}\frac{d}{dx}(\alpha\log(x)) = x^\alpha\frac{\alpha}{x} = \alpha x^{\alpha-1}.$$

Maxima and Minima

- A real valued function f on a topological space X has a *local maximum* at x if there is an open neighbrhood U of x such that [R Def 5.7]

$$f(y) \le f(x)$$

for all $y \in U$.

- The local maximum is a *strict local maximum* if

$$f(y) < f(x)$$

for all $y \in U \setminus \{x\}$.

- Local minima are defined analogously.

- If $f : [a,b] \to \mathbb{R}$ has a local maximum at $x \in (a,b)$ and if $f'(x)$ exists, then $f'(x) = 0$. [R Thm 5.8]

  - If $x_n \uparrow x$, then $x_n - x < 0$ for all n, and $f(x_n) - f(x) \le 0$ for all large n, so
$$f'(x) = \lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x} \ge 0.$$
  - If $x_n \downarrow x$, then $x_n - x > 0$ for all n, and $f(x_n) - f(x) \le 0$ for all large n, so
$$f'(x) = \lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x} \le 0.$$
  - So $f'(x) = 0$.

- An analogous result holds on more general domains, e.g. $\mathbb{R}^k$.

A common strategy for establishing a global maximum of $f : X \to \mathbb{R}$:

- Show that the equation $f'(x) = 0$ has a unique root x^*.
- Show that for some open set A with compact closure $\overline{A}$ and some $x_0 \in A$

$$\sup_{x \in A^c} f(x) < f(x_0).$$

Then

- Since there is a point in A with a higher value than any point in A^c,

$$\sup_{x \in X} f(x) \leq \sup_{x \in A} f(x).$$

- Since $\overline{A}$ is compact and f is continuous, the maximum is attained.
- Since $f(x) < f(x_0)$ for all x in the boundary $\overline{A} \backslash A$, the maximum must be attained on the open set A.
- Any global maximum is also a local maximum, and any interior local maximum x satisfies $f'(x) = 0$.
- Since x^* is the unique solution to this equation, it is the unique global maximum.

For a real-valued x the set A is often of the form $A = (-M, M)$ for a large M.

Mean Value Theorems

- A general result: Suppose $f, g : [a,b] \to \mathbb{R}$ are differentiable on (a,b). [R Thm 5.9] Then there exists an $x \in (a,b)$ such that

$$(f(b) - f(a))g'(x) = (g(b) - g(a))f'(x).$$

- The standard mean value theorem: Suppose $f : [a,b] \to \mathbb{R}$ is differentiable on (a,b). Then there exists an $x \in (a,b)$ such that [R Thm 5.10]

$$f(b) - f(a) = (b-a)f'(x).$$

  - Follows from the general result taking $g(x) = x$.

- Some simple consequences:
  - If $f'(x) \geq 0$ for all $x \in (a,b)$ then f is monotonically increasing.
  - If $f'(x) \leq 0$ for all $x \in (a,b)$ then f is monotonically decreasing.
  - If $f'(x) = 0$ for all $x \in (a,b)$ then f is constant.

Details for the general version

- For $y \in [a,b]$ define

$$h(y) = (f(b) - f(a))g(y) - (g(b) - g(a))f(y).$$

- Then
  - h is differentiable on (a,b) with derivative

$$h'(y) = (f(b) - f(a))g'(y) - (g(b) - g(a))f'(y).$$

  - $h(a) = h(b) = 0$.
- Need to show that $h'(x) = 0$ for some $x \in (a,b)$.
- If $h(y) = 0$ for all $y \in (a,b)$ we are done.
- If $h(y) > 0$ for some $y \in (0,1)$ then h attains its maximum over $[a,b]$ at a point $x \in (a,b)$, and that point satisfies $h'(x) = 0$.
- Similarly, if $h(y) < 0$ for some $y \in (a,b)$, then h has a local minimum at some $x \in (a,b)$ with $h'(x) = 0$.

Continuity of Derivatives

- Suppose $f : [a,b] \to \mathbb{R}$ is differentiable on $[a,b]$, and suppose [R Thm 5.12]

$$f'(a) < \lambda < f'(b).$$

  Then

$$f'(x) = \lambda$$

  for some $x \in (a,b)$.

- An analogous result holds when $f'(b) > f'(a)$.
- This does not imply that derivarives must be continuous; the function [R Ex 5.6 b]

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

  is differentiable everywhere. but the derivative is not continuous at $x = 0$.

- This does imply that a derivative cannot have any simple discontinuities.

Some Details

- Let $g(y) = f(y) - \lambda y$.
- Then $g'(a) < 0$, so for some $y_1 \in (a,b)$ we have $g(y_1) < g(a)$.
- Similarly, $g'(b) > 0$, so for some $y_2 \in (a,b)$ we have $g(y_2) < g(b)$.
- So
$$\inf_{a<y<b} g(y) < \min(g(a), g(b)),$$
and therefore g attains its minimum over $[a,b]$ at a point $x \in (a,b)$.
- That point is an interior local minimum, so
$$g'(x) = f'(x) - \lambda = 0.$$

L'Hospital's Rule

- Also known as L'Hôpital's rule.
- Suppose

$$\lim_{x\to a}\frac{f'(x)}{g'(x)} = A$$

for $\infty \le A \le +\infty$, and either

$$\lim_{x\to a} f(x) = 0 \qquad\qquad \lim_{x\to a} g(x) = 0$$

or

$$\lim_{x\to a} f(x) = \infty \qquad\qquad \lim_{x\to a} g(x) = \infty.$$

Then

$$\lim_{x\to a}\frac{f(x)}{g(x)} = A$$

- Rudin's proof uses the general form of the mean value theorem.

Examples

- $$\lim_{x\to 0}\frac{\sin x}{x} = \lim_{x\to 0}\frac{\cos x}{1} = 1.$$

- For $\beta > 0$
$$\lim_{x\to\infty}\frac{\log(x)}{x^\beta} = \lim_{x\to\infty}\frac{1/x}{\beta x^{\beta-1}} = \lim_{x\to\infty}\frac{1}{\beta x^\beta} = 0.$$

- For $\beta > 0$ and $x > 0$

$$\lim_{x\to 0} x^\beta \log(x) = \lim_{x\to 0}\frac{\log(x)}{x^{-\beta}} = \lim_{x\to 0}\frac{1/x}{-\beta x^{-\beta-1}} = \lim_{x\to 0} -\frac{x^\beta}{\beta} = 0.$$

- For $\beta > 0$
$$\lim_{x\to\infty} x^\beta e^{-x} = \lim_{x\to\infty}\frac{x^\beta}{e^x} = \lim_{x\to\beta}\frac{\beta x^{\beta-1}}{e^x}$$
Repeated application gives convergence to zero.

- Alternative approach: take logarithms to obtain
$$\lim_{x\to\infty}\beta\log(x) - x = \lim_{x\to\infty} x\left(\frac{\beta\log(x)}{x} - 1\right)$$
This limit is $-\infty$ since
$$\lim_{x\to\infty}\left(\frac{\beta\log(x)}{x} - 1\right) = -1.$$

Wednesday, February 24, 2016

L'Hospital's Rule

Some Details

- Suppose
  - f, g are differentiable on (a, b);
  - $f(x) \to 0$ and $g(x) \to 0$ as $x \to a$;
  - as $x \to a$
$$\frac{f'(x)}{g'(x)} \to A.$$
- If $A = +\infty$ then
$$\limsup_{y \to a} \frac{f(y)}{g(y)} \leq A$$
- Suppose $-\infty \leq A < +\infty$, and choose q, r such that $A < r < q < \infty$.
- There is a value $c \in (a, b)$ such that for $a < x < c$
$$\frac{f'(x)}{g'(x)} < r.$$
- If $a < x < y < c$ then by the generalized mean value theorem there exists a $t \in (x, y)$ such that
$$\frac{f(y) - f(x)}{g(y) - g(x)} = \frac{f'(t)}{g'(t)}.$$
Since the right hand side is bounded by r for $t \in (x, y) \subset (a, c)$ this means
$$\frac{f(y) - f(x)}{g(y) - g(x)} < r$$
for all x, y with $a < x < y < c$.
- Let $x \to a$ to obtain
$$\frac{f(y)}{g(y)} \leq r$$
for all $y \in (a, c)$.

- Therefore
$$\limsup_{y\to a}\frac{f(y)}{g(y)} \le r < q$$
for all $q > A$.

- This implies
$$\limsup_{y\to a}\frac{f(y)}{g(y)} \le A$$
for $-\infty \le A < \infty$ as well.

- An analogous argument implies
$$\liminf_{y\to a}\frac{f(y)}{g(y)} \ge A.$$

- Combined, these yield
$$\lim_{y\to a}\frac{f(y)}{g(y)} = A.$$

Taylor's Theorem

- Suppose f is a real-valued function on $[a,b]$.
- Suppose f is differentiable with derivative f' and f' is differentiable. [R Thm 5.14]
- The *second derivative* is $f'' = (f')'$.
- The n-th derivative, if it exists, is $f^{(n)} = (f^{(n-1)})'$.
- Alternate notation:
$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$$
- For $f^{(n)}$ to exists at x, $f^{(n-1)}$ has to exist in a neighborhood of x.
- The neighborhood can be one-sided if $x = a$ or $x = b$.
- Suppose
  - $f^{(n-1)}$ is continuous on $[a,b]$;
  - $f^{(n)}(x)$ exists for all $x \in (a,b)$.
- For $\alpha \in [a,b]$ define
$$P_{n,\alpha}(y) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(y-\alpha)^k.$$
- Then for $\beta \in [a,b]$ there exists an x between α and β such that [R Thm 5.15]
$$f(\beta) = P_{n,\alpha}(\beta) + \frac{f^{(n)}(x)}{n!}(\beta-\alpha)^k.$$
- This means
  - f can be approximated by a polynomial;
  - the error can be estimated if we have a bound on $|f^{(n)}(x)|$ over the interval between α and β.

Some Details

- The proof uses repeated applications of the mean value theorem.
- Let M be defined by

$$f(\beta) = P_{n,\alpha}(\beta) + M(\beta - \alpha)^n.$$

- Let

$$g(t) = f(t) - P_{n,\alpha}(t) - M(t - \alpha)^n.$$

- Then $g^{(n)}(t) = f^{(n)}(t) - n!M$.
- We need to show that $g^{(n)}(x) = 0$ for some x between α and β.
- Since $P^{(k)}_{n,\alpha}(\alpha) - f^{(k)}(\alpha)$ for $k = 0, 1, \ldots, n-1$,

$$g(\alpha) = g'(\alpha) = \ldots g^{(n-1)}(\alpha) = 0.$$

- The definition of M means $g(\beta) = 0$.
- So by the mean value theorem there is an x_1 between α and β such that $g'(x_1) = 0$.
- Using the mean value theorem again, there is and x_2 between α and x_1 such that $g''(x_2) = 0$.
- After n steps e conclude that there is an x_n between α and x_{n-1}, and therefore between α and β, such that $g^{(n)}(x_n) = 0$.

Differentiating Vector-Valued Functions

- Suppose X is a vector space with a norm.
- The definition of the derivative carries over to a function $f : [a,b] \to X$. [R Rem 5.16]
- If $X = \mathbb{R}^k$ and $f(x) = (f_1(x), \ldots, f_k(x))$, then
$$f'(x) = (f_1'(x), \ldots, f_k'(x)).$$
- Many basic results carry over.
- The mean value theorem and its corollaries do not.
- A variant is available:
  - Suppose $f : [a,b] \to \mathbb{R}^k$ is differentiable on (a,b).
  - Then there exists an $x \in (a,b)$ such that
$$\|f(b) - f(a)\| \le (b-a)\|f'(x)\|.$$
- Proof outline:
  - For vectors $x = (x_1, \ldots, x_k), y = (y_1, \ldots, y_k)$ let
$$\langle x, y \rangle = \sum_{i=1}^{k} x_i y_i.$$
  - $\langle x, y \rangle$ is the *inner product*, or *scalar product* of x, y.
  - Let $z = f(b) - f(a)$, and set $g(t) = \langle z, f(t) \rangle$.
  - g is real-valued and continuous on $[a,b]$ and differentiable on (a,b).
  - So by the mean value theorem there exists and $x \in (a,b)$ such that
$$g(b) - g(a) = (b-a)g'(x) = (b-a)\langle z, f'(x) \rangle.$$
  - Now
$$g(b) - g(a) = \langle z, f(b) \rangle - \langle z, f(a) \rangle = \langle z, z \rangle = \|z\|^2.$$
  - By the Cauchy-Schwartz inequality,
$$\|z\|^2 = (b-a)\langle z, f'(x) \rangle \le (b-a)\|z\|\|f'(x)\|$$
  - This implies
$$\|z\| = \|f(b) - f(a)\| \le (b-a)\|f'(x)\|$$

Euclidean Spaces

Basic Definitions and Properties

- An inner product on a real vector space is a function $\langle x, y\rangle$ such that [KF §16 Def 1]
  - $\langle x, x\rangle \geq 0$, with $\langle x, x\rangle = 0$ if and only if $x = 0$.
  - $\langle x, y\rangle = \langle y, x\rangle$ for all x, y.
  - $\langle cx, y\rangle = c\langle x, y\rangle$ for any vectors x, y and scalar c.
  - $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$ for any x, y, z.
- A real vector space with an inner product is an *inner product space* or a *Euclidean space*. [KF §16 Def 2]
- Let $\|x\| = \sqrt{\langle x, x\rangle}$. Then [KF §16 Thm 1]
  - $\|x\|$ is a norm;
  - for any x, y the inner product $\langle x, y\rangle$ satisfies the Cauchy-Schwartz inequality
$$|\langle x, y\rangle| \leq \|x\|\|y\|$$
- The angle $\theta \in [0, \pi]$ between two vectors x, y in a Euclidean space is defined by [KF §16 Def 3]
$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\|y\|}.$$
- In particular, x, y are *orthogonal* if $\langle x, y, =\rangle 0$.
- A normed space is Euclidean if and only if for all x, y
$$\|x+y\| + \|x-y\| = 2(\|x\| + \|y\|).$$
- In a Euclidean space
$$\langle x, y\rangle = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2).$$
- A Euclidean space that is complete is called a *Hilbert space*. [KF §16 Thm 15]

Examples

- $\mathbb{R}^k$ with the usual inner product is a Euclidean space.
- $\ell_2 = \{(x,.x_2,\dots) : \sum x_i^2 < \infty\}$ with

$$\langle x, y\rangle = \sum x_i y_i$$

  is a Euclidean space, and a Hilbert space.

- The set of random variables $L_2 = \{X : E[X^2] < \infty\}$ with $\langle X, Y\rangle = E[XY]$ is a Euclidean space, and a Hilbert space.
- Let $C^2_{[0,1]}$ be the set of all continuous functions on $[0,1]$ with the norm

$$\|x\| = \sqrt{\int_0^1 x(t)^2 dt}.$$

  Then $C^2_{[0,1]}$ is a Euclidean space with inner product

$$\langle x, y\rangle = \int_0^1 x(t)y(t)dt.$$

  This is not a Hilbert space since it is not complete.

Differentiation

- The results for derivatives of vector-valued functions carry over to functions taking values in a general Euclidean space.

Week 7

Monday, February 29, 2016

Convergence Rate of Newton's Method

- Suppose f is twice continuously differentiable, that $f(\alpha) = 0$, and $f'(\alpha) \neq 0$.
- Newton's method for finding α starts with an initial guess x_0 and computes

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

- By Taylor's theorem there is a z_n between α and x_n such that

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2} f''(z_n)(\alpha - x_n)^2.$$

- This can be rearranged as

$$-\frac{1}{2}\frac{f''(z_n)}{f'(x_n)}(\alpha - x_n)^2 = \frac{f(x_n)}{f'(x_n)} + \alpha - x_n = \alpha - x_{n+1}.$$

- If it can be shown that $x_n \to \alpha$, then $z_n \to \alpha$ and

$$\frac{|\alpha - x_{n+1}|}{(\alpha - x_n)^2} \to \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right|.$$

- This is called *quadratic convergence*.

The Riemann Integral

Basic Definitions

- A *partition* of a bounded interval $[a,b]$ is a set of points $x_1,\ldots,x_n$ with [R Def 6.1]

$$x_0 = a \le x_1 \le x_2 \le \cdots \le x_n = b.$$

- Let $\Delta x_i = x_i - x_{i-1}$ for $i = 1,\ldots,n$.
- For a bounded function $f : [a,b] \to \mathbb{R}$ define for $i = 1,\ldots,n$

$$M_i = \sup_{x_{i-1} \le x \le x_i} f(x)$$
$$m_i = \inf_{x_{i-1} \le x \le x_i} f(x).$$

- Let

$$U(P,f) = \sum_{i=1}^{n} M_i \Delta x_i$$
$$L(P,f) = \sum_{i=1}^{n} m_i \Delta x_i$$

- Define the *upper (Riemann) integral* and the *lower (Riemann) integral* as

$$\overline{\int_a^b} f(x)dx = \inf_P U(P,f)$$
$$\underline{\int_a^b} f(x)dx = \sup_P L(P,f)$$

- If the upper and lower integrals are equal, the function is said to be *(Riemann) integrable* on $[a,b]$, and the common value is the (Riemann) integral

$$\int_a^b f(x)dx.$$

Examples

- Suppose $f(x) = c$ for all $x \in [a,b]$. Then f is integrable with

$$\int_a^b f(x)dx = c(b-a).$$

- Suppose $f(x) = x$. Then

$$U(P,f) - L(P,f) = \sum_{i=1}^{n} (M_i - m_i)\Delta x_i \leq (b-a) \max_{1 \leq i \leq n} \Delta x_i,$$

so upper and lower integrals are equal, and equal to

$$\sum_{i=1}^{n} \frac{x_{i-1} + x_i}{2} \Delta x_i = \frac{1}{2} \sum_{i=1}^{n} (x_i^2 - x_{i-1}^2) = \frac{b^2 - a^2}{2}.$$

- Suppose

$$f(x) = \begin{cases} 1 & \text{if } x \in [a,b] \text{ is rational} \\ 0 & \text{if } x \in [a,b] \text{ is irrational.} \end{cases}$$

Then for every partition P

$$U(P,f) = b - a \qquad\qquad L(P,f) = 0.$$

So f is not Riemann integrable.

Simple Bounds

- Since f is bounded, there are values $m \le M$ such that for all $x \in [a,b]$

$$m \le f(x) \le M.$$

- For any partition P

$$m(b-a) \le L(P,f) \le U(P,f) \le M(b-a).$$

- Therefore upper and lower integrals are finite, and

$$\underline{\int_a^b} f(x)dx \ge m(a-b)$$
$$\overline{\int_a^b} f(x)dx \le M(b-a).$$

The Riemann-Stieltjes Integral

Definitions

- Let $\alpha : [a,b] \to \mathbb{R}$ be monotonically increasing. [R Def 6.2]
- For a partition P of $[a,b]$ let

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}).$$

- For a bounded function $f : [a,b] \to \mathbb{R}$ let

$$U(P,f,\alpha) = \sum_{i=1}^{n} M_i \Delta\alpha_i$$
$$L(P,f,\alpha) = \sum_{i=1}^{n} m_i \Delta\alpha_i$$

and define upper and lower *Riemann-Stieltjes integrals* as

$$\overline{\int_a^b} f(x)d\alpha(x) = \inf_P U(P,f,\alpha)$$
$$\underline{\int_a^b} f(x)d\alpha(x) = \sup_P L(P,f,\alpha)$$

- If upper and lower integrals are identical then f is *Riemann-Stieltjes* integrable with respect to α on $[a,b]$ with the common value

$$\int_a^b f d\alpha = \int_a^b f(x)d\alpha(x)$$

called the Riemann-Stieltjes integral, or just Stieltjes integral.

Examples

- Setting $\alpha(x) = x$ produces the Riemann integral.
- If α is differentiable we fill see later that

$$\int_a^b f(x)d\alpha(x) = \int_a^b f(x)\alpha'(x)dx.$$

- If f is continuous and α is a step function increasing by c_i at points $y_1, \ldots, y_k \in [a,b]$, then

$$\int_a^b f d\alpha = \sum_{i=1}^{k} f(y_i)c_i.$$

- If $[a,b] = [0,1]$,

$$f(x) = \begin{cases} 1 & \text{if } x \geq \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}$$

  and $\alpha = f$, then the integral does not exist:

  - $\Delta\alpha_i$ is only positive if $x_{i-1} < \frac{1}{2} \leq x_i$.
  - On such an interval

$$M_i = 1$$
$$m_i = 0.$$

  - So for all partitions P

$$U(P,f,\alpha) = 1$$
$$L(P,f,\alpha) = 0.$$

Wednesday, March 2, 2016

Properties of the Riemann-Stieltjes Integral

- A partition P^* is a *refinement* of P if $P \subset P^*$, i.e. every point is P is also a point in P^*. [R Def 6.3]
- The *common refinement* of two partitions P_1, P_2 is the partition

$$P^* = P_1 \cup P_2.$$

- If P^* is a refinement of P then [R Thm 6.4]

$$L(P, f, \alpha) \le L(P^*, f, \alpha)$$
$$U(P^*, f, \alpha) \le U(P, f, \alpha).$$

  - Think about adding a point between two partition points.

- Let P_1, P_2 be two partitions and P^* their common refinement. Then [R Thm 6.5]

$$L(P_1, f, \alpha) \le L(P^*, f, \alpha) \le U(P^*, f, \alpha) \le U(P_2, f, \alpha).$$

- So for any partition P_1

$$L(P_1, f, \alpha) \le \inf_{P_2} U(P_2, f, \alpha) = \overline{\int_a^b} f d\alpha.$$

- Taking the supremum over P_1 then gives

$$\underline{\int_a^b} f d\alpha = \sup_{P_1} L(P_1, f, \alpha) \le \overline{\int_a^b} f d\alpha.$$

- f is Riemann-Stieltjes integrable with respect to α, $f \in \mathscr{R}(\alpha)$, if

$$\underline{\int_a^b} f d\alpha = \overline{\int_a^b} f d\alpha.$$

- $f \in \mathscr{R}(\alpha)$ if and only if for every $\varepsilon > 0$ there is a partition P such that [R Thm 6.6]

$$U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon,$$

  - If: Given $\varepsilon > 0$ select a partition P such that

$$U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon.$$

    Then

$$\overline{\int_a^b} f d\alpha - \underline{\int_a^b} f d\alpha \le U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon.$$

    So

$$\overline{\int_a^b} f d\alpha - \underline{\int_a^b} f d\alpha = 0$$

    and $f \in \mathscr{R}(\alpha)$.

  - Only if: Suppose $f \in \mathscr{R}(\alpha)$ and let $\varepsilon > 0$ be given. Then there exist partitions P_1, P_2 such that

$$U(P_1,f,\alpha) - \int_a^b f d\alpha \le \frac{\varepsilon}{2}$$
$$\int_a^b f d\alpha - L(P_2,f,\alpha) \le \frac{\varepsilon}{2}.$$

    Let P be the common refinement of P_1, P_2. Then

$$\begin{aligned} U(P,f,\alpha) \le U(P_1,f,\alpha) &< \int_a^b f d\alpha + \frac{\varepsilon}{2} \\ &< L(P_2,f,\alpha) + \varepsilon \le L(P,f,\alpha) + \varepsilon \end{aligned}$$

    and therefore

$$U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon.$$

- Suppose $\varepsilon > 0$, $P = \{x_0, \ldots, x_n\}$ is a partition such that [R Thm 6.7]

$$U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon,$$

and $s_i, t_i \in [x_{i-1}, x_i]$ for $i = 1, \ldots, n$. Then

  - $\sum_{i=1}^{n} |f(s_i) - f(t_i)| \Delta\alpha_i < \varepsilon$;
  - if $f \in \mathscr{R}(\alpha)$, then

$$\left| \sum_{i=1}^{n} f(t_i) \Delta\alpha_i - \int_a^b f d\alpha \right| < \varepsilon.$$

- Proof outline:
  - $f(s_i), f(t_i) \in [m_i, M_i]$, so $|f(s_i) - f(t_i)| \leq M_i - m_i$ and

$$\sum_{i=1}^{n} |f(s_i) - f(t_i)| \Delta\alpha_i \leq U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon;$$

  - the second result follows since

$$L(P, f, \alpha) \leq \sum_{i=1}^{n} f(t_i) \Delta\alpha_i \leq U(P, f, \alpha)$$

$$L(P, f, \alpha) \leq \int_a^b f d\alpha \leq U(P, f, \alpha).$$

Some Existence Results

- If f is continuous on $[a,b]$, then $f \in \mathscr{R}(\alpha)$. [R Thm 6.8]
- Proof outline:
  - Fix $\varepsilon > 0$.
  - Choose $\eta > 0$ such that $(\alpha(b) - \alpha(a))\eta < \varepsilon$.
  - f is uniformly continuous, so there exists a $\delta > 0$ such that

    $$|f(x) - f(y)| < \eta$$

    for any $x, y \in [a,b]$ with $|x - y| < \delta$.
  - Choose a partition P with $\Delta x_i < \delta$ for all i.
  - Then $M_i - m_i \leq \eta$ for $i = 1, \ldots, n$, and therefore

$$\begin{aligned} U(P,f,\alpha) - L(P,f,\alpha) &= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\ &\leq \sum_{i=1}^{n} \eta\Delta\alpha_i = \eta(\alpha(b) - \alpha(a)) < \varepsilon. \end{aligned}$$

- If f is monotone on $[a,b]$ and α is continuous and monotone on $[a,b]$, then $f \in \mathscr{R}(\alpha)$ on $[a,b]$. [R Thm 6.9]
- Outline of proof:
  - Suppose f is increasing; the decreasing case is analogous.
  - Fix $\varepsilon > 0$.
  - By continuity of α we can choose a partition P such that

$$\Delta\alpha_i = \frac{\alpha(b) - \alpha(a)}{n}$$

  - Then $M_i = f(x_i), m_i = f(x_{i-1})$, and

$$\begin{aligned} U(P,f,\alpha) - L(P,f,\alpha) &= \sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))\Delta\alpha_i \\ &= \frac{\alpha(b) - \alpha(a)}{n} \sum_{i=1}^{n} (f(x_i) - f(x_{i-1})) \\ &= \frac{\alpha(b) - \alpha(a)}{n} (f(b) - f(a)). \end{aligned}$$

  - This is less than ε for n large enough.

- Suppose [R Thm 6.10]
  - f is bounded on $[a,b]$;
  - f has at most finitely many discontinuities on $[a,b]$;
  - α is continuous at every discontinuity of f.

  Then $f \in \mathscr{R}(\alpha)$.

- Outline of proof for one discontinuity at $c \in (a,b)$:
  - Let $M = \sup|f(x)|$ and fix $\varepsilon > 0$.
  - Choose $u < c < v$ such that $u, v \in (a,b)$ and

$$2M(\alpha(v) - \alpha(u)) < \frac{\varepsilon}{3}$$

  - Then for any partition of $[u,v]$

$$U(P[u,v], f, \alpha) - L(P[u,v], f, \alpha) \le 2M(\alpha(v) - \alpha(u)) < \frac{\varepsilon}{3}.$$

  - Since f is continuous on $[a,u]$ and $[v,b]$ we can chose partitions of these sets with

$$U(P[a,u], f, \alpha) - L(P[a,u], f, \alpha) < \frac{\varepsilon}{3}$$
$$U(P[v,b], f, \alpha) - L(P[v,b], f, \alpha) < \frac{\varepsilon}{3}.$$

  - Therefore for the combined partition

$$P = P[a,u] \cup P[u,v] \cup P[v,b]$$

    of $[a,b]$ we have

$$U(P, f, \alpha) - L(P, f, \alpha) < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

More Basic Properties of the Integral

- Suppose [R Thm 6.11]
  - $f \in \mathscr{R}(\alpha)$ on $[a,b]$;
  - $m \le f \le M$;
  - ϕ is continuous on $[m,M]$;
  - $h(x) = \phi(f(x))$.

  Then $h \in \mathscr{R}(\alpha)$ on $[a,b]$.
- Suppose $f \in \mathscr{R}(\alpha)$ on $[a,b]$. Then
  - $cf \in \mathscr{R}(\alpha)$ on $[a,b]$ for any $c \in \mathbb{R}$ and [R Thm 6.12a]

$$\int_a^b cf\,d\alpha = c\int_a^b f\,d\alpha;$$

  - $|f| \in \mathscr{R}(\alpha)$ on $[a,b]$, and [R Thm 6.13b]

$$\left|\int_a^b f\,d\alpha\right| \le \int_a^b |f|\,d\alpha.$$

  - $f \in \mathscr{R}(c\alpha)$ on $[a,b]$ $c > 0$, and [R Thm 6.12e]

$$\int_a^b f\,d(c\alpha) = c\int_a^b f\,d\alpha.$$

- Suppose $f, g \in \mathscr{R}(\alpha)$ on $[a,b]$. Then
  - $f+g \in \mathscr{R}(\alpha)$ on $[a,b]$, and [R Thm 6.12a]
  $$\int_a^b (f+g)d\alpha = \int_a^b f d\alpha + \int_a^b g d\alpha;$$
  - $fg \in \mathscr{R}(\alpha)$ on $[a,b]$; [R Thm 6.13a]
  - if $f(x) \le g(x)$ on $[a,b]$, then [R Thm 6.12b]
  $$\int_a^b f d\alpha \le \int_a^b g d\alpha;$$
- If $f \in \mathscr{R}(\alpha_1)$ on $[a,b]$ and $f \in \mathscr{R}(\alpha_2)$ on $[a,b]$, then $f \in \mathscr{R}(\alpha_1+\alpha_2)$ on $[a,b]$, and [R Thm 6.12e]
$$\int_a^b f d(\alpha_1+\alpha_2) = \int_a^b f d\alpha_1 + \int_a^b f d\alpha_2.$$
- Suppose $a \le c \le b$.
  - If $f \in \mathscr{R}(\alpha)$ on $[a,b]$, then $f \in \mathscr{R}(\alpha)$ on $[a,c]$ and $f \in \mathscr{R}(\alpha)$ on $[c,b]$, and [R Thm 6.12c]
  $$\int_a^b f d\alpha = \int_a^c f d\alpha + \int_c^b f d\alpha.$$
  - If $f \in \mathscr{R}(\alpha)$ on $[a,c]$ and $f \in \mathscr{R}(\alpha)$ on $[c,b]$, then $f \in \mathscr{R}(\alpha)$ on $[a,b]$, and
  $$\int_a^b f d\alpha = \int_a^c f d\alpha + \int_c^b f d\alpha.$$

Two Important Special Cases

- Suppose f is bounded, α is differentiable on $[a,b]$ and $\alpha' \in \mathscr{R}$ on $[a,b]$. [R Thm 6.17]
  - Then $f \in \mathscr{R}(\alpha)$ if and only if $f\alpha' \in \mathscr{R}$.
  - In this case
$$\int_a^b f d\alpha = \int_a^b f(x)\alpha'(x)dx.$$
- Suppose [R Thm 6.16]
  - f is continuous on $[a,b]$;
  - $c_n \geq 0$ and $\sum c_n$ converges;
  - $\{y_n\}$ is a sequence of distinct points in (a,b);
  - $\alpha(x) = \sum_{n=1}^{\infty} c_n 1_{[a,x]}(y_n)$.

  Then $f \in \mathscr{R}(\alpha)$ on $[a,b]$, and
$$\int_a^b f d\alpha = \sum_{n=1}^{\infty} c_n f(y_n).$$

Friday, March 4, 2016

Integration and Differentiation

- Suppose $f \in \mathscr{R}$ on $[a,b]$. For $x \in [a,b]$ define [R Thm 6.20]

$$F(x) = \int_a^x f(t)dt.$$

  Then

  - F is continuous on $[a,b]$;
  - if f is continuous at $x_0 \in [a,b]$, then F is differentiable at x_0 and $F'(x_0) = f(x_0)$.

- Continuity:

  - Since $f \in \mathscr{R}$

$$M = \sup_{a \le x \le b} |f(x)| < \infty.$$

  - So for $x < y$

$$\begin{aligned} |F(y) - F(x)| &= \left| \int_a^y f(t)dt - \int_a^x f(t)dt \right| \\ &= \left| \int_x^y f(t)dt \right| \\ &\le \int_x^y |f(t)|dt \\ &\le M(y-x). \end{aligned}$$

- Differentiability:
  - Fix $\varepsilon > 0$.
  - Since f is continuous at x_0 there is a $\delta > 0$ such that

$$|f(t) - f(x_0)| < \varepsilon$$

    when $|t - x_0| < \delta$.
  - For $u, v \in [a,b]$ with $u < v$ and

$$x_0 - \delta \le u \le x_0 \le v \le x_0 + \delta$$

    we have

$$\begin{aligned}
\left| \frac{F(v) - F(u)}{v-u} - f(x_o) \right| &= \frac{1}{v-u} |F(v) - F(u) - f(x_0)(v-u)| \\
&= \frac{1}{v-u} \left| \int_u^v f(t)dt - \int_u^v f(x_0)dt \right| \\
&= \frac{1}{v-u} \left| \int_u^v [f(t) - f(x_0)]dt \right| \\
&\le \frac{1}{v-u} \int_u^v |f(t) - f(x_0)|dt \\
&< \varepsilon.
\end{aligned}$$

Fundamental Theorem of Calculus

- If F is differentiable on $[a,b]$, $F' = f$, and $f \in \mathscr{R}$, then [R Thm 6.21]

$$\int_a^b f(x)dx = F(b) - F(a).$$

- Proof outline:
  - Fix $\varepsilon > 0$.
  - Choose a partition such that $U(P,f) - L(P,f) < \varepsilon$.
  - By the mean value theorem there are points $t_i \in [x_{i-1}, x_i]$ such that

$$F(x_i) - F(x_{i-1}) = f(t_i)\Delta x_i$$

    for $i = 1, \ldots, n$.
  - Then

$$\sum_{i=1}^{n} f(t_i)\Delta x_i = F(b) - F(a).$$

  - So

$$\left| F(b) - F(a) - \int_a^b f(x)dx \right| = \left| \sum_{i=1}^{n} f(t_i)\Delta x_i - \int_a^b f(x)dx \right| < \varepsilon.$$

  - Since this holds for any $\varepsilon > 0$ the result follows.

Integration By Parts

- Suppose F, G are differentiable on $[a,b]$ and suppose $f = F' \in \mathscr{R}$ and $g = G' \in \mathscr{R}$ on $[a,b]$. Then [R Thm 6.22]

$$\int_a^b F(x)g(x)dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)dx.$$

- Proof outline:
  - Let $H(x) = F(x)G(x)$.
  - Then $H'(x) = f(x)G(x) + F(x)g(x)$.
  - $H' \in \mathscr{R}$ since $f, g \in \mathscr{R}$ and F, G are continuous and therefore also $F, G \in \mathscr{R}$.
  - So by the fundamental theorem of calculus

$$\int_a^b H'(x)dx = H(b) - H(a),$$

and

$$\int_a^b H'(x)dx = \int_a^b F(x)g(x)dx + \int_a^b f(x)G(x)dx.$$

Uniform Convergence and Integration

- Suppose [R Thm 7.16]
  - α is monotonically increasing on $[a,b]$;
  - $f_n \in \mathscr{R}(\alpha)$ for $n = 1,2,\ldots$;
  - $f_n \to f$ uniformly on $[a,b]$.

  Then
  - $f \in \mathscr{R}(\alpha)$;
  - $\lim_{n\to\infty} \int_a^b f_n d\alpha = \int_a^b f d\alpha$.
- Proof outline:
  - Let $\varepsilon_n = \sup |f_n(x) - f(x)|$.
  - Then $f_n - \varepsilon_n \le f \le f_n + \varepsilon_n$ and therefore

$$\underline{\int_a^b} f d\alpha \ge \int_a^b (f_n - \varepsilon_n) d\alpha = \int_a^b f_n d\alpha - \varepsilon_n[\alpha(b) - \alpha(a)]$$
$$\overline{\int_a^b} f d\alpha \le \int_a^b (f_n + \varepsilon_n) d\alpha = \int_a^b f_n d\alpha + \varepsilon_n[\alpha(b) - \alpha(a)].$$

  - So

$$0 \le \overline{\int_a^b} f d\alpha - \underline{\int_a^b} f d\alpha \le 2\varepsilon_n[\alpha(b) - \alpha(a)] \to 0.$$

    and therefore $f \in \mathscr{R}(\alpha)$.
  - Since upper and lower integrals are identical and identical to the integral of f,

$$\left| \int_a^b f d\alpha - \int_a^b f_n d\alpha \right| \le \varepsilon_n[\alpha(b) - \alpha(a)] \to 0.$$

Improper Integrals

- If $f \in \mathscr{R}$ on $[c,b]$ for every $c \in (a,b]$ define [R Prob 6.7]

$$\int_a^b f(x)dx = \lim_{c \to a} \int_c^b f(x)dx$$

  if the limit exists and is finite.

- If $f \in \mathscr{R}$ then this coincides with the usual definition, since f is bounded and

$$\left| \int_a^c f(x)dx \right| \leq \sup |f(x)|(c-a) \to 0$$

  as $c \to a$.

- If $f \in \mathscr{R}$ on $[a,b]$ for every $b > a$ define

$$\int_a^\infty f(x)dx = \lim_{b \to \infty} \int_a^b f(x)dx$$

  if the limit exists and is finite.

  - If the limit exists then the integral is said to *converge*.
  - Ifthe limit exists with f replaced by $|f|$ then the integral is said to *converge absolutely*.

Integral Test For Series

- Suppose $f \geq 0$ and f is monotonically decreasing on $[1, \infty)$. Then

$$\int_1^\infty f(x)dx$$

converges if and only if

$$\sum_{k=1}^{\infty} f(k)$$

converges.

- Outline of proof:
  - Since $\int_1^b f(x)dx$ is increasing in b and $\sum_{k=1}^n f(k)$ is increasing in n is enough to show one it is bounded if and only if the other is.
  - Consider the partition of $[1,n]$ given by $P = \{1,2,\ldots,n\}$.
  - Since f is decreasing,

$$U(P,f) = \sum_{k=1}^{n-1} f(k)$$
$$L(P,f) = \sum_{k=2}^{n} f(k).$$

  - So

$$\sum_{k=2}^{n} f(k) \leq \int_1^n f(x)dx \leq \sum_{k=1}^{n-1} f(k).$$

  - So either both converge or both diverge.

Week 8

Monday, March 7, 2016

Weierstrass'Approximation Theorem

- Suppose $f : [a,b] \to \infty$ is continuous. Then there exists a sequence f_n of polynomials such that $f_n \to f$ uniformly on $[a,b]$. [R Thm 7.26]
- A proof using probability:
  - Assume, without loss of generality, that $[a,b] = [0,1]$.
  - Let $X \sim$ Bin(n, p), let $Y_n = \frac{X}{n}$, and let

$$f_n(p) = E_{n,p}[f(Y_n)] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} p^k (1-p)^{n-k}.$$

  - f_n is a polynomial of degree n in p.
  - f is continuous on the compact set $[0,1]$ and therefore uniformly continuous and bounded.
  - Let $M = \sup |f(x)|$.

– Fix $\varepsilon > 0$ and choose $\delta > 0$ such that

$$|f(x) - f(y)| < \frac{\varepsilon}{2}$$

when $|x - y| < \delta$.

– Then

$$\begin{aligned} |f(p) - f_n(p)| &= |E_{n,p}[f(p) - f(Y_n)]| \\ &\le E_{n,p}[|f(p) - f(Y_n)|]. \end{aligned}$$

– Now

$$\begin{aligned} E_{n,p}[|f(p) - f(Y_n)|] =& E_{n,p}[|f(p) - f(Y_n)| 1_{\{|Y_n - p| < \delta\}}] \\ &+ E_{n,p}[|f(p) - f(Y_n)| 1_{\{|Y_n - p| \ge \delta\}}] \\ \le& \frac{\varepsilon}{2} + 2M P_{n,p}(|Y_n - p| \ge \delta). \end{aligned}$$

– By Chebychev's inequality,

$$\begin{aligned} P_{n,p}(|Y_n - p| \ge \delta) &\le \frac{\mathrm{Var}(Y_n)}{\delta^2} \\ &= \frac{p(1-p)}{n\delta^2} \\ &\le \frac{1}{4n\delta^2}. \end{aligned}$$

– So for $n > \frac{M}{\delta^2 \varepsilon}$

$$2M P_{n,p}(|Y_n - p| \ge \delta) \le 2M \frac{\delta^2 \varepsilon}{4\delta^2 M} = \frac{\varepsilon}{2}.$$

and therefore

$$|f(p) - f_n(p)| \le \varepsilon$$

for all $p \in [0,1]$.

Convergence of Descent Algorithms

- Many computational problems can be formulated in terms of
  - an *objective function* $f : X \to [0, \infty]$ on a metric space X;
  - a *solution set* $S = \{x \in X : f(x) = 0\}$;
  - the goal of finding a point in or close to the solution set.
- A general iterative algorithm is described by a function $A : X \to X$ that is used to produce a sequence of approximate solutions
  - starting with an initial value $x_0 \in X$;
  - then computing $x_{n+1} = A(x_n)$.
- An algorithm A is a *descent algorithm* if

$$f(A(x)) \leq f(x)$$

  for all $x \in X$, and

$$f(A(x)) < f(x)$$

  for all $x \notin S$.
- Suppose f and A are continuous.
  - Then S contains all limit points of x_n.
  - If S consists of a single point x^* and there is a compact set $K \supset S$ such that $x_n \in K$ for all n, then $x_n \to x^*$.

- Proof outline:
  - $f(x_{n+1}) = f(A(x_n)) \le f(x_n)$, so
    - $f(x_n)$ is a decreasing sequence;
    - for any subsequence x_{n_k} the values $f(x_{n_k})$ are decreasing as well.
  - Suppose x^* is a limit point of x_n and x_{n_k} is a subsequence converging to x^*.
  - Then $f(x_{n_k}) \downarrow f(x^*)$, and, in particular, $f(x_{n_k}) \ge f(x^*)$ for all k.
  - Suppose $x^* \notin S$.
  - Then $f(A(x^*)) < f(x^*)$, and we can choose a value z such that

$$f(A(x^*)) < z < f(x^*).$$

  - Since $f(A(x))$ is continuous, there is a $\delta > 0$ such that $f(A(x)) < z$ when $d(x, x^*) < \delta$.
  - Since $x_{n_k} \to x^*$, for all large k we have $|x_{n_k} - x^*| < \delta$ and therefore

$$f(x_{n_{k+1}}) \le f(x_{n_k+1}) = f(A(x_{n_k})) < z < f(x^*),$$

    which is a contradiction.
  - So $x^* \in S$.
  - If x_n is contained in a compact set then every subsequence has a further convergent subsequence; if $S = \{x^*\}$ then these all converge to x^*, and therefore $x_n \to x^*$.
- There are more general variants in which
  - A returns a set of values;
  - x_{n+1} is chosen as an element of $A(x_n)$.
- Many algorithms used in practice are covered by these formulations.

Smooth Interpolation

- Suppose $a = x_0 < x_1 < x_2 < \cdots < x_n < b = x + n + 1$ and $y_1, y_2, \ldots y_n$ are real numbers.
- We would like to find a function g such that
  - $g(x_i) = y_i$ for all $i = 1, \ldots, n$;
  - g is as smooth as possible.
- g is said to interpolate $(x_1, y_1), \ldots, (x_n, y_n)$.
- A specific smoothness requirement:
  - g, g', g'' should be continuous on $[a, b]$;
  - The *roughness* measured by

$$R(g) = \int_a^b (g''(x))^2 dx$$

    should be as small as possible.
- Let $\mathscr{S}$ be the set of all interpolating functions f such that f, f', f'' are continuous.
- Suppose $g \in \mathscr{S}$ minimizes $R(g)$ and f is another function in $\mathscr{S}$.
- Let $h = f - g$ and $f_u(x) = g(x) + uh(x)$ for some $u \in \mathbb{R}$.
- Then $f_u \in \mathscr{S}$, and for all u, and

$$R(f_u) = R(g) + u^2 R(h) + 2u \int_a^b g''(x)h''(x)dx \geq R(g).$$

- This has to be minimized at $u = 0$, and therefore

$$\int_a^b g''(x)h''(x)dx = 0.$$

- Furthermore,

$$R(f) = R(g) + R(f - g) \geq R(g);$$

  for $n \geq 2$ equality holds if and only if $f = g$.

- Suppose that g'' is differentiable on each of the intervals $[x_{i-1}, x_i]$.
- Using integration by parts,

$$\begin{aligned}\int_a^b g''(x)h''(x)dx &= \sum [g''(x_i)h'(x_i) - g''(x_{i-1})h'(x_{i-1})] - \sum \int_{x_{i-1}}^{x_i} g'''(x)h'(x)dx \\ &= g''(b)h'(b) - g''(a)h'(a) - \int_a^b g'''(x)h'(x)dx\end{aligned}$$

- Suppose g''' is continuously differentiable on each of the intervals $[x_{i-1}, x_i]$.
- Using integration by parts on these intervals, and the fact that $h(x_k) = 0$ for $k = 1, \ldots, n$,

$$\begin{aligned}\int_a^b g'''(x)h'(x)dx &= \sum [g'''(x_i-)h(x_i) - g'''(x_{i-1}+)h(x_{i-1})] - \sum \int_{x_{i-1}}^{x_i} g^{(4)}(x)h(x)dx \\ &= g'''(b-)h(b) - g'''(a+)h(a) - \int_a^b g^{(4)}(x)h(x)dx\end{aligned}$$

- These terms will be zero if
  - $g^{(4)}$ is zero on each interval;
  - g'' and g''' are zero on $[a, x_1)$ and $(x_n, b]$
- This means
  - g is linear on $[a, x_1)$ and $(x_n, b]$;
  - g is cubic on $[x_{i-1}, x_i]$ for $i = 2, \ldots, n$;
  - g, g', g'' are continuous;
  - $g(x_i) = y_i$.
- A function with these properties exists and is called a *natural cubic spline*.

- Efficient algorithms for computing the interpolating spline are available.
- The solution to finding the twice continuously differentiable function g on $[a,b]$ that minimizes

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda R(g)$$

  for $\lambda \geq 0$ is also a natural cubic spline, called a *smoothing spline*.
- λ is called a *smoothing* or *regularization* parameter.
- A drafting spline shows the motivation for the term *natural spline*.

Wednesday, March 7, 2016

Lebesgue Measure

Intervals

- Lebesgue measure generalizes the notion of interval length or rectangle area to more complicated sets.
- An interval in $\mathbb{R}^p$ is a set of points $x = (x_1, \ldots, x_p)$ such that

$$a_i \le x_i \le b_i$$

  for all $i = 1, \ldots, p$; any of the inequalities may also be strict.

- The measure of an interval is

$$m(I) = (b_1 - a_1) \times \cdots \times (b_p - a_p).$$

- The measure m is *additive* in the sense that if $I = I_1 \cup I_2$ and if I_1, I_2 are disjoint, then

$$m(I) = m(I_1) + m(I_2).$$

- By induction this implies that if $I = I_1 \cup \cdots \cup I_n$ and $I_k \cap I_l = \emptyset$ for $k \neq l$ then

$$m(I) = \sum_{k=1}^{n} m(I_k).$$

Elementary Sets

- A union of a finite number of intervals is an *elementary set*.
- Some properties of the collection $\mathscr{E}$ of elementary sets:
  - if $A, B \in \mathscr{E}$ then $A \cup B \in \mathscr{E}$;
  - if $A, B \in \mathscr{E}$ then $A \Delta B \in \mathscr{E}$, where

$$A \Delta B = (A \backslash B) \cup (B \backslash A);$$

- These mean $\mathscr{E}$ is a *ring* of subsets of $\mathbb{R}^p$.
- Suppose $A = I_1 \cup \cdots \cup I_n$ is an elementary set and the intervals are pairwise disjoint. Define the measure of A as

$$m(A) = \sum_{k=1}^{n} m(I_k).$$

- This is well-defined: if

$$A = \bigcup_k I_k = \bigcup_l J_l$$

with

$$I_i \cap I_j = \emptyset, \qquad J_i \cap J_j = \emptyset, \qquad i \neq j$$

then, since the intersection of two rectangles is a rectangle,

$$m(A) = \sum_k m(I_k) = \sum_k \sum_l m(I_k \cap J_l) = \sum_l m(J_l)$$

by the additivity of m for rectangles.

- If $A_1, \ldots, A_n \in \mathscr{E}$ are pairwise disjoint, then $\bigcup_{k=1}^{n} A_k \in \mathscr{E}$, and

$$m\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} m(A_k);$$

i.e. m is additive on $\mathscr{E}$.

Outer Measure

- For any subset B of $\mathbb{R}^p$ define the *outer measure* of B as

$$m^*(B) = \inf \sum_{n=1}^{\infty} m(A_n),$$

where the infimum is taken over all collections $A_1, A_2, \ldots$ of open elementary sets A_n such that

$$B \subset \bigcup_{n=1}^{\infty} A_n.$$

- Some properties of outer measure:
  - $m^*(A) = m(A)$ for every $A \in \mathscr{E}$.
  - $m^*(\mathbb{R}^p) = \infty$.
  - If $B_1 \subset B_2$ then $m^*(B_1) \leq m^*(B_2)$.
  - If $B \subset \bigcup_{n=1}^{\infty} B_n$, then

$$m^*(B) \leq \sum_{n=1}^{\infty} m^*(B_n).$$

This property is called *countable subadditivity*.

Measurable Sets

- A set A is *finitely measurable*, $A \in \mathscr{M}_F$, if there exists a sequence of elementary sets A_n such that

$$m^*(A \Delta A_n) \to 0.$$

- A set A is *measurable*, $A \in \mathscr{M}$, if it is a union of a countable collection of finitely measurable sets.

- The collection $\mathscr{M}$ of measurable sets is a *sigma-algebra* of subsets of $\mathbb{R}^p$:
  - $\mathbb{R}^p \in \mathscr{M}$;
  - if $A \in \mathscr{M}$ then $A^c \in \mathscr{M}$;
  - if $A_1, A_2, \cdots \in \mathscr{M}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{M}$.

- Outer measure is *countably additive* on $\mathscr{M}$: if $A_1, A_2, \cdots \in \mathscr{M}$ are pairwise disjoint, then

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

Lebesgue Measure

- A non-negative countably additive set function μ defined on a sigma-algebra such that $\mu(\emptyset) = 0$ is called a *measure*.
- The restriction of m^* to $\mathscr{M}$ is a measure, and is called *Lebesgue measure*.
- The sets in $\mathscr{M}$ are called *Lebesgue measurable*.
- All open sets are measurable.
- All closed sets are measurable.
- All finite or countable sets are measurable.
- A *Borel set* is any set that can be obtained, by a countable number of operations starting with open sets, with each operation consisting of taking unions, intersections, or complements.
- The collection of Borel sets is a sigma-algebra.
- The Borel sigma-algebra is the smallest sigma algebra containing all open sets.
- All Borel sets are measurable.
- For every $A \in \mathscr{M}$ there exists a Borel set B such that

$$m^*(A \Delta B) = 0.$$

- If $A \subset \mathbb{R}^p$ has $m^*(A) = 0$ then $A \in \mathscr{M}$.
- Non-Borel sets do exist; these can still be measurable.
- Non measurable sets do exist.

Friday, March 11, 2016

Measurable Functions

Basic Definitions and Properties

- Let X be a set, and $\mathscr{M}$ a sigma-algebra of subsets of X.
- A subset $A \subset X$ is called *measurable*, or *measurable with respect to* $\mathscr{M}$, if $A \in \mathscr{M}$.
- A function $f : X \to \overline{\mathbb{R}}$ is *measurable*, or *measurable with respect to* $\mathscr{M}$, or $\mathscr{M}$-measurable, if
$$\{x : f(x) > a\}$$
is measurable for all $a \in \mathbb{R}$.
- Each of the following is equivalent:
  - f is measurable;
  - $\{x : f(x) \geq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$;
  - $\{x : f(x) < a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$;
  - $\{x : f(x) \leq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$;
  - $f^{-1}(G) \in \mathscr{M}$ for all open sets G;
  - $f^{-1}(F) \in \mathscr{M}$ for all closed sets F;
  - $f^{-1}(B) \in \mathscr{M}$ for all Borel sets B.

Some Details

- If f is measurable, then $\{x : f(x) \geq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$:

$$\{x : f(x) \geq a\} = \bigcap_n \left\{ x : f(x) > a - \frac{1}{n} \right\}.$$

- If $\{x : f(x) \geq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$, then $\{x : f(x) < a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$:

$$\{x : f(x) < a\} = \{x : f(x) \geq a\}^c$$

- If $\{x : f(x) < a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$, then $\{x : f(x) \leq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$:

$$\{x : f(x) \leq a\} = \bigcap_n \left\{ x : f(x) < a + \frac{1}{n} \right\}$$

- If $\{x : f(x) \leq a\} \in \mathscr{M}$ for all $a \in \mathbb{R}$, then f is measurable:

$$\{x : f(x) > a\} = \{x : f(x) \leq a\}^c.$$

- If f is measurable, then $f^{-1}(G) \in \mathscr{M}$ for every open subset of $\overline{\mathbb{R}}$:
  - Every open subset is a countable union of bounded open intervals or open half lines;
  - every bounded open interval is an intersection of two open half lines;
  - measurability implies inverse images of open half lines are in $\mathscr{M}$;
  - since $\mathscr{M}$ is a sigma-algebra, intersections and unions of these inverse images are in $\mathscr{M}$ as well.
- Closed sets: take complements.
- If f is measurable, then $f^{-1}(A) \in \mathscr{F}$ for all Borel sets A:
  - The collection $\mathscr{B}$ of all sets B such that $f^{-1}(B) \in \mathscr{M}$ is a sigma-algebra.
  - $\mathscr{B}$ contains all open sets.
  - The Borel sigma-algebra is the smallest sigma-algebra that contains all open sets.
  - So $\mathscr{B}$ must contain all Borel sets.
- If $f^{-1}(A) \in \mathscr{M}$ for all Borel sets A, then f is measurable:
  - The half lines $(a, \infty]$ are Borel subsets of $\overline{\mathbb{R}}$.

More Properties

- Suppose $F : \overline{\mathbb{R}}^p \to \overline{\mathbb{R}}$ is continuous and $f_1, \ldots, f_p$ are measurable. Then

$$h(x) = F(f_1(x), \ldots, f_p(x))$$

  is measurable.

  - Since F is continuous, $F^{-1}(G)$ is open for any open set $G \subset \overline{\mathbb{R}}$.
  - Every open set in $\overline{\mathbb{R}}^p$ is a countable union of open sets of the form $G_1 \times \ldots G_p$ with G_i an open subset of $\overline{\mathbb{R}}$.
  - $$(f_1, \ldots, f_p)^{-1}(G_1 \times \ldots G_p) = \bigcap_{i=1}^{p} f_i^{-1}(G_i).$$
  - Since f_i is measurable, $f_i^{-1}(G_i)$ is measurable.

- Suppose f, g are measurable. Then
  - $|f|$ is measurable;
  - $f^+ = \max(f, 0), f^- = -\min(f, 0)$ are measurable;
  - $f + g, fg$ are measurable.

- Suppose $f_1, f_2, \ldots$ are measurable. Then

$$g(x) = \sup_n f_n(x)$$
$$h(x) = \limsup_n f_n(x)$$

are measurable. Analogous results hold for inf and liminf.

  - For any $a \in \mathbb{R}$

$$\{x : g(x) > a\} = \bigcup_{n=1}^{\infty} \{x : f_n(x) > a\}.$$

  - Let

$$g_m(x) = \sup_{n \geq m} f_n(x).$$

    Then

$$h(x) = \inf_m g_m(x).$$

- If $f_1, f_2, \ldots$ are measurable and

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

exists for all x, then f is measurable.

Functions on the Real Line

- Let $\overline{\mathscr{R}}$ be the Borel sigma-algebra on $\overline{\mathbb{R}}$.
- Let $\overline{\mathscr{L}}$ be the sigma algebra of Lebesgue-measurable sets.
  - Then $\overline{\mathscr{R}} \subset \overline{\mathscr{L}}$.
- A function $f : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is Borel measurable if it is measurable with respect to $\overline{\mathscr{R}}$.
- A function $f : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is Lebesgue measurable if it is measurable with respect to $\overline{\mathscr{L}}$.
- If f is continuous then f is Borel measurable
  - $\{x : f(x) > a\}$ is an open set, hence a Borel set.
- If f is monotone then f is Borel measurable.
  - For an increasing function and any a the set $\{x : f(x) > a\}$ is of the form $\{x : x > b\}$ or $\{x : x \geq b\}$ fof some b; these are Borel sets.
- If f is
  - the difference of two monotone functions, or
  - a function with finitely many discontinuities

  then f is Borel measurable.

- Suppose $g : X \to \overline{\mathbb{R}}$ is measurable and $f : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is Borel measurable. Then
$$h(x) = f(g(x)) : X \to \overline{\mathbb{R}}$$
is measurable.
  - Since f is Borel measurable, $B = f^{-1}(A)$ is a Borel set for any Borel set A.
  - Since B is a Borel set and g is measurable, $g^{-1}(B)$ is measurable.
  - $h^{-1}(A) = g^{-1}(f^{-1}(A)) = g^{-1}(B)$.
- This conclusion may fail if f is only Lebesgue measurable.
  - There may exist a Borel set A such that the inverse image $f^{-1}(A)$ is Lebesgue measurable but not a Borel set.
  - In this case h may fail to be measurable.
  - Rudin gives a reference.
- When working with functions on the real line we almost always assume these are Borel measurable.

Simple Functions

- A function $s : X \to \mathbb{R}$ is a *simple functions* if its range is finite, i.e. is takes on only finitely many values.
- For a set $A \subset X$ define the *indicator function* of A as

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

  - Some authors, including Rudin, call these *characteristic functions*.
  - Some use the notation χ_A.
  - In probability and statistics these are used elsewhere.

- If s is a simple function with range $\{c_1, \ldots, c_n\}$ and

$$A_i = \{s : s(x) = c_i\}$$

  then

$$s(x) = \sum c_i 1_{A_i}(x)$$

- If one of the possible values is zero we can discard it and use only the non-zero c_i.

- For any $f : X \to \overline{\mathbb{R}}$ there exists a sequence s_n of simple functions such that $s_n(x) \to f(x)$ for all x.
  - If f is measurable then the s_n can be taken to be measurable.
  - If $f \geq 0$ then the s_n can be chosen such that $s_n(x) \uparrow f(x)$.
- For $f \geq 0$ let

$$A_{ni} = \left\{ x : \frac{i-1}{n} \leq f(x) < \frac{i}{n} \right\}$$
$$B_n = \{x : f(x) \geq n\}$$

  for $n = 1, 2, \ldots$ and $i = 1, 2, \ldots, n^2$, and let

$$s_n = n 1_{B_n} + \sum_{i-1}^{n^2} \frac{i-1}{n} 1_{A_{ni}}.$$

- For general f write

$$f = f^+ - f^-$$

  and apply this construction to f^+ and f^-.
- If f is bounded, then s_n converges uniformly to f.

Week 9

Monday, March 21, 2016

Integration

Definitions

- Suppose
  - X is a set
  - $\mathscr{M}$ is a sigma-algebra on X
  - μ is a measure on $\mathscr{M}$
- For a measurable simple function

$$s(x) = \sum_{k=1}^{n} c_k 1_{A_k}(x)$$

with $c_k \neq 0$ and a set $A \in \mathscr{M}$ define

$$I_A(s) = \sum_{k=1}^{n} c_k \mu(A \cap A_k).$$

- Some basic properties:
  - If s_1, s_2 are simple functions, then $s_1 + s_2$ is also simple, and
$$I_A(s_1 + s_2) = I_A(s_1) + I_A(s_2).$$
  - If s is a simple function and $c \in \mathbb{R}$, then cs is simple, and
$$I_A(cs) = cI_A(s).$$
- If s is a simple function, for any $A \in \mathscr{M}$ define
$$\phi(A) = I_A(s).$$
Then $\phi(\emptyset) = 0$ and ϕ is countably additive on $\mathscr{M}$.
  - Suppose
$$s = \sum_{k=1}^{m} c_k 1_{A_k}.$$
  - Suppose $B_1, B_2, \cdots \in \mathscr{M}$ are pairwise disjoint, and
$$B = \bigcup_{n=1}^{\infty} B_n,$$
  - Then
$$\begin{aligned}
\phi(B) = I_B(s) &= \sum_{k=1}^{m} c_k \mu(B \cap A_k) \\
&= \sum_{k=1}^{m} c_k \left(\sum_{n=1}^{\infty} \mu(B_n \cap A_k) \right) \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{m} c_k \mu(B_n \cap A_k) \\
&= \sum_{n=1}^{\infty} I_{B_n}(s) = \sum_{n=1}^{\infty} \phi(B_n).
\end{aligned}$$

- For a non-negative measurable function f define

$$\int_A f d\mu = \sup I_A(s)$$

  with the supremum taken over all measurable simple functions s such that $0 \le s \le f$.

- This is the *Lebesgue integral* of f, with respect to μ, over A.
- This integral may be infinite.
- Suppose f is measurable and consider the integrals

$$\int_A f^+ d\mu \qquad\qquad \int_A f^- d\mu$$

- If at least one is finite, define

$$\int_A f d\mu = \int_A f^+ d\mu - \int_A f^- d\mu.$$

- If both are finite, f is said to be *integrable with respect to* μ.
- f is integrable if the integral exists and is finite.
- The set of functions integrable with respect to μ is $\mathscr{L}(\mu)$.
- The set of integrable functions with respect to Lebesgue measure is $\mathscr{L}$.

Some Elementary Properties

- If f is measurable and bounded on A and if $\mu(A) < \infty$, then $f \in \mathscr{L}$ on A
- If $a \le f(x) \le b$ for $x \in A$ and $\mu(A) < \infty$, then

$$a\mu(A) \le \int_A f d\mu \le b\mu(A).$$

- If $f, g \in \mathscr{L}(\mu)$ on A and $f(x) \le g(x)$ for $x \in A$, then

$$\int_A f d\mu \le \int_A g d\mu.$$

- If $f \in \mathscr{L}(\mu)$ on A and $c \in \mathbb{R}$, then $cf \in \mathscr{L}(\mu)$ and

$$\int_A cf d\mu = c \int_A d d\mu.$$

- If $\mu(A) = 0$ and f is measurable, then

$$\int_A f d\mu = 0.$$

- If $f \in \mathscr{L}(\mu)$ on A, $B \in \mathscr{M}$, and $B \subset A$, then $f \in \mathscr{L}(\mu)$ on B.

Wednesday, March 23, 2016

Defining a Measure as an Integral

- Suppose f is measurable and non-negative. For $A \in \mathscr{M}$ define

$$\phi(A) = \int_A f d\mu.$$

Then ϕ is countably additive on $\mathscr{M}$.

- Suppose $A_1, A_2, \cdots \in \mathscr{M}$ are pairwise disjoint and

$$A = \bigcup_{n=1}^{\infty} A_n.$$

- Suppose s is simple and $0 \le s \le f$. Then

$$I_A(s) = \sum_{n=1}^{\infty} I_{A_n}(s) \le \sum_{n=1}^{\infty} \phi(A_n).$$

Taking the supremum over s then gives

$$\int_A f d\mu = \phi(A) \le \sum_{n=1}^{\infty} \phi(A_n).$$

– If $\phi(A_n) = \infty$ for some n, then $\phi(A) = \infty$, and

$$\phi(A) \geq \sum_{n=1}^{\infty} \phi(A_n).$$

– Suppose $\phi(A_n) < \infty$ for all n.

– Given $\varepsilon > 0$ we can chose a simple function s with $0 \leq s \leq f$ such that

$$\int_{A_1} s d\mu \geq \phi(A_1) - \varepsilon/2 \qquad \int_{A_2} s d\mu \geq \phi(A_2) - \varepsilon/2.$$

– Then

$$\begin{aligned} \phi(A_1 \cup A_2) &\geq \int_{A_1 \cup A_2} s d\mu \\ &= \int_{A_1} s d\mu + \int_{A_2} s d\mu \\ &\geq \phi(A_1) + \phi(A_2) - \varepsilon. \end{aligned}$$

So $\phi(A) \geq \phi(A_1 \cup A_2) \geq \phi(A_1) + \phi(A_2)$.

– Therefore, for all n

$$\phi(A) \geq \phi(A_1) + \cdots + \phi(A_n).$$

‘

– Taking $n \to \infty$ completes the proof.

- The result remains true if $f \in \mathscr{L}(\mu)$ and takes on positive and negative values.
- ϕ is then called a *signed measure*.
- If $\mu(A) = 0$ then $\phi(A) = 0$.
- This property is called *absolute continuity*; ϕ is said to be *absolutely continuous with respect to* μ.
- Common notation for absolute continuity: $\phi \ll \mu$.
- A measure μ on a space X is *finite* if $\mu(X) < \infty$.
- A measure μ on a space X is *sigma-finite* if there exists a countable collection of sets $A_1, A_2, \ldots$ such that
  - $\mu(A_n) < \infty$ for all n;
  - $X = \bigcup_{n=1}^{\infty} A_n$.
- All probabilities are finite and hence sigma-finite.
- Lebesgue measure on $\mathbb{R}^n$ is sigma-finite but not finite.

Radon-Nikodym Theorem

- Existence: If ϕ, μ are sigma-finite and $\phi \ll \mu$, then there exists a measurable function f such that

$$\phi(A) = \int_A f d\mu$$

for all measurable sets A.

- Uniqueness: if the function g also satisfies

$$\phi(A) = \int_A f d\mu$$

for all measurable sets A, then $f = g$ almost everywhere.

- The function f is called the (Radon-Nykodym) derivative of ϕ with respect to μ.

- Common notation:

$$f = \frac{d\phi}{d\mu}.$$

- The Radon-Nikodym derivative is also called the (Radon-Nikodym) density.

- Some other notations:

$$\phi(A) = \int_A f(x)\mu(dx)$$
$$f(x) = \frac{\phi(dx)}{\mu(dx)}.$$

Examples

- Suppose
  - $X = \{x_1, x_2, \dots\}$ is a countable set.
  - $\mathscr{M}$ consists of all subsets of x.
  - μ is *counting measure*,

$$\mu(A) = \text{number of elements in } A.$$

  - P is a probability on X.

  Then

$$f(x) = \frac{dP}{d\mu} = \frac{P(dx)}{\mu(dx)} = P(\{x\}).$$

- Suppose
  - $X = \mathbb{R}^p$.
  - $\mathscr{M}$ consists of the Borel subsets $\mathscr{R}^p$ of $\mathbb{R}^p$.
  - μ is Lebesgue measure on $\mathscr{R}^p$.
  - P is a probability on $\mathscr{R}^p$ and $P \ll \mu$.

  Then P has a density f with respect to μ, and

$$P(A) = \int_A f d\mu$$

  for all $A \in \mathscr{R}^p$.

Random Variables and Expected Values

- Let $(S, \mathscr{B}, P)$ be a probability space.
- A random variable is a measurable function $X : S \to \overline{\mathbb{R}}$.
- The *expected value* of X is

$$E[X] = \int X dP$$

  if the integral exists.

Conditional Expectation

- Suppose $E[X]$ exists and let $\mathscr{C} \subset \mathscr{B}$ be another, smaller, sigma-algebra.
- For example, suppose Y is another random variable and

$$\mathscr{C} = \sigma(Y) = \{Y^{-1}(A) : A \in \overline{\mathscr{R}}\}.$$

  - This is called the *sigma algebra generated by* Y.
  - It is the smallest sigma algebra with respect to which Y is measurable.

- Define $\phi(A) = \int_A X dP$ for $A \in \mathscr{C}$.
- Let $P_{\mathscr{C}}$ be the restriction of P to $\mathscr{C}$.
- Then $\phi \ll P_{\mathscr{C}}$.
- The *conditional expectation of* X *given* $\mathscr{C}$ is

$$E[X|\mathscr{C}] = \frac{d\phi}{dP_{\mathscr{C}}}.$$

- Two key properties of $E[X|\mathscr{C}]$:
  - $E[X|\mathscr{C}]$ is $\mathscr{C}$-measurable.
  - For any $A \in \mathscr{C}$

$$E[X1_A] = E[E[X|\mathscr{C}]1_A].$$

- The conditional expectation of X given Y is defined as

$$E[X|Y] = E[X|\sigma(Y)].$$

Monotone Convergence Theorem

- Suppose $f_1, f_2, \ldots$ are measurable functions and $0 \le f_1(x) \le f_2(x) \le \ldots$ for $x \in A$, and let
$$f(x) = \lim_{n\to\infty} f_n(x).$$
Then
$$\int_A f d\mu = \lim_{n\to\infty} \int_A f_n d\mu.$$
  - It is clear that
$$\alpha = \lim_{n\to\infty} \int_A f_n d\mu$$
exists and
$$\alpha \le \int_A f d\mu.$$
  - Choose a constant c with $0 < c < 1$ and a simple function s with $0 \le s \le f$, and let
$$A_n = \{x \in A : f_n(x) \ge cs(x)\}.$$
  - Then
$$A_1 \subset A_2 \subset \ldots$$
and
$$A = \bigcup_{n=1}^{\infty} A_n.$$
  - Also
$$\alpha \ge \int_A f_n d\mu \ge \int_{A_n} f_n d\mu \ge c \int_{A_n} s d\mu.$$
  - Since $\int_B s d\mu$ is countably additive in B taking the limit gives
$$\alpha \ge c \int_A s d\mu.$$
  - Taking she sup over s, c gives
$$\alpha \ge \int f d\mu.$$

Friday, March 25, 2016

More Basic Properties

- If $f \in \mathscr{L}(\mu)$ on A then $|f| \in \mathscr{L}(\mu)$ on A and
$$\left|\int_A f d\mu\right| \le \int_A |f| d\mu.$$
  - Let $B = \{x \in A : f(x) \ge 0\}$.
  - Then $|f| \in \mathscr{L}(\mu)$ in A since
$$\int_A |f| d\mu = \int_B |f| d\mu + \int_{A\setminus B} |f| d\mu = \int_B f^+ d\mu + \int_{A\setminus B} f^- d\mu < \infty.$$
  - Since $f \le |f|$ and $-f \le |f|$ we have
$$\int_A f d\mu \le \int_A |f| d\mu \qquad -\int_A f d\mu \le \int_A |f| d\mu$$
  and the result follows.
- If f_1, f_2 are non-negative and $f = f_1 + f_2$, then
$$\int_A f d\mu = \int_A f_1 d\mu + \int_A f_2 d\mu$$
  - Suppose $s_{n,1}, s_{n,2}$ are sequences of non-negative simple functions increasing monotonically to f_1, f_2.
  - Then $s_n = s_{n,1} + s_{n,2}$ is a sequence of non-negative simple converging monotonically to f.
  - So by the monotone convergence theorem
$$\begin{aligned}\int_A f d\mu &= \lim_{n\to\infty} \int_A s_n d\mu = \lim_{n\to\infty} I_A(s_n) \\ &= \lim_{n\to\infty} I_A(s_{n,1} + s_{n.2}) = \lim_{n\to\infty} (I_A(s_{n.1}) + I_A(s_{n,2})) \\ &= \lim_{n\to\infty} I_A(s_{n,1}) + \lim_{n\to\infty} I_A(s_{n.2}) \\ &= \lim_{n\to\infty} \int_A s_{n,1} d\mu + \lim_{n\to\infty} \int_A s_{n,3} d\mu = \int_A f_1 d\mu + \int_A f_2 d\mu.\end{aligned}$$
- If $f_1, f_2 \in \mathscr{L}(\mu)$ on A, then $f = f_1 + f_2 \in \mathscr{L}(\mu)$ and
$$\int_A f d\mu = \int_A f_1 d\mu + \int_A f_2 d\mu.$$

Fatou's Lemma

- Suppose $A \in \mathscr{M}$, $f_1, f_2, \ldots$ are non-negative measurable functions, and

$$f(x) = \liminf_{n\to\infty} f_n(x).$$

  Then

$$\int_A f d\mu \le \liminf_{n\to\infty} \int_A f_n d\mu.$$

- Proof outline:
  - Let
$$g_n(x) = \inf_{m\ge n} f_m(x).$$
  - Then $0 \le g_1 \le g_2(x) \le \ldots$.
  - Also
$$g_n(x) \le f_n(x)$$
$$g_n(x) \to f(x).$$
  - By the monotone convergence theorem
$$\lim_{n\to\infty} \int_A g_n d\mu = \int_A f d\mu.$$
  - So
$$\int_A f d\mu = \lim_{n\to\infty} \int_A g_n d\mu \le \liminf_{n\to\infty} \int_A f_n d\mu.$$

Dominated Convergence Theorem

- Suppose $A \in \mathscr{M}$. Let f_n be a sequence of measurable functions such that for $x \in A$

$$f_n(x) \to f(x).$$

Suppose there exists a function $g \in \mathscr{L}(\mu)$ such that for all $x \in A$ and all n

$$|f_n(x)| \le g(x).$$

Then

$$\lim_{n\to\infty} \int_A f_n d\mu = \int_A f d\mu.$$

- Proof outline:
  - $g + f_n \ge 0$, so by Fatou's Lemma

$$\int_A (g+f) d\mu \le \liminf_{n\to\infty} \int_A (g+f_n) d\mu.$$

  - Therefore,

$$\int_A f d\mu \le \liminf_{n\to\infty} \int_A f_n d\mu.$$

  - Similarly, $g - f_n \ge 0$, and

$$\int_A (g-f) d\mu \le \liminf_{n\to\infty} \int_A (g-f_n) d\mu.$$

  - This implies

$$-\int_A f d\mu \le -\liminf_{n\to\infty} \int_A f_n d\mu,$$

or

$$\int_A f d\mu \ge \limsup_{n\to\infty} \int_A f_n d\mu.$$

- The result remains true if the sets

$$B_n = \{x \in A : g(x) < f_n(x)\}$$

are non-empty as long as $\mu(B_n) = 0$.

- Thus the result is true as long as $|f_n| \le g$ *almost everywhere* on A.

- The result also remains true if $f_n(x) \to f(x)$ almost everywhere on A.

Connections to Riemann Integrals

- Suppose f is Riemann integrable on $[a,b]$. Then $f \in \mathscr{L}$ on $[a,b]$ and the two integrals are identical.
- If f is bounded on $[a,b]$ then f is Riemann integrable on $[a,b]$ if and only if it is continuous almost everywhere.
- If $f \in \mathscr{L}$ on $[a,b]$ and for $x \in [a,b]$

$$F(x) = \int_{[a,x]} f dm,$$

then

  - F is differentiable almost everywhere;
  - $F'(x) = f(x)$ almost everywhere.

Completeness for a Uniform Distribution

- Suppose $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0, \theta]$ for $\theta > 0$.
- The minimal sufficient statistic is

  $$T = X_{(n)} = \max(X_1, \ldots, X_n)$$

  with denisty

  $$f(t|\theta) = n\frac{t^{n-1}}{\theta^n} 1_{[0,\theta]}(t).$$

- Suppose h is a measurable function such that $E[h(T)|\theta] = 0$ for all θ.
- The expected value is

  $$E[h(T)|\theta] = \int_{[0,\theta]} h(t) n \frac{t^{n-1}}{\theta^n} dt = \frac{n}{\theta^n} \int_{[0,\theta]} h(t) t^{n-1} dt.$$

- Thus $E[h(T)|\theta] = 0$ for all $\theta > 0$ implies

  $$H(\theta) = \int_{[0,\theta]} h(t) t^{n-1} dt = 0$$

  for all $\theta > 0$.
- So $H'(\theta) = 0$ for all $\theta > 0$, and

  $$H'(\theta) = 0 = h(\theta)\theta^{n-1}$$

  for almost all θ.
- Thus $h(\theta) = 0$ for almost all θ, and therefore $h(T) = 0$ almost surely.
- So the family of distributions of T is complete.

Moment Generating Functions

- The moment generating function of a real-valued random variable X is

$$M(t) = E[e^{tX}].$$

  - $M(t)$ is defined for all t;
  - $M(t)$ can be infinite;
  - $M(0) = 1$.

- Let $B = \{t : M(t) < \infty\}$. Then B is an interval containing the origin and M is convex on B.

  - e^y is convex, so for any t_1, t_2 and any $\alpha \in [0,1]$

$$\begin{aligned} M(\alpha t_1 + (1-\alpha)t_2) &= E\left[e^{\alpha t_1 X + (1-\alpha)t_2 X}\right] \\ &\le E\left[\alpha e^{t_1 X} + (1-\alpha)e^{t_2 X}\right] \\ &= \alpha M(t_1) + (1-\alpha)M(t_2). \end{aligned}$$

  - So M is convex, and if $t_1, t_2 \in B$ then $M(t)$ is finite for any t between t_1 and t_2.
  - This means B is an interval, and since $M(0) = 1$ the set B contains the origin.
  - B may be the degenerate interval $\{0\}$.

- Suppose $M(t)$ is finite for all t in an open set G containing the the origin.
- Then for all positive integers n

$$E[|X|^n] < \infty,$$

and for all $t \in G$

$$E[|X|^n e^{tX}] < \infty.$$

  - Let $A = \{X \geq 0\}$ and write

$$M(t) = \int_A e^{tX} dP + \int_{A^c} e^{tX} dP.$$

  - Both integrals are finite for all $t \in G$.
  - For $h > 0$ let

$$K_n(h) = \max\{y^n e^{-yh} : y \geq 0\} < \infty.$$

  - Fix $t \in G$.
  - Since G is open and contains the origin there is a $u \in G$ with $t < u$ and $u > 0$.
  - Then on the set $A = \{X \geq 0\}$

$$X^n = (X^n e^{-uX}) e^{uX} \leq K_n(u) e^{uX}$$
$$X^n e^{tX} = (X^n e^{-(u-t)}) e^{uX} \leq K_n(u-t) e^{uX}$$

    and therefore

$$E[(X^+)^n] = \int_A X^n dP \leq K_n(u) \int_A e^{uX} dP \leq K_n(u) M(u) < \infty$$
$$\int_A |X|^n e^{tX} dP = \int_A X^n e^{tX} dP \leq K_n(u-t) \int_A e^{uX} dP \leq K_n(u-t) M(u) < \infty.$$

  - An analogous argument shows that

$$E[(X^-)^n] < \infty$$
$$\int_{A^c} |X|^n e^{tX} dP < \infty.$$

  - So $E[|X|^n e^{tX}] < \infty$ for all $t \in G$ and

$$E[|X|^n] = E[(X^+)^n] + E[(X^-)^n] < \infty.$$

Week 10

Monday, March 28, 2016

Moment Generating Functions

- Suppose $M(t)$ is finite on an open set G containing the origin.
- For $t \in G$ and non-negative integers n define

$$\varphi_n(t) = E[X^n e^{tX}].$$

Then φ_n is differentiable on G and

$$\varphi_n'(t) = \varphi_{n+1}(t).$$

Since G contains the origin, then this implies that

$$E[X^n] = \frac{d^n}{dt^n} M(t)\bigg|_{t=0}.$$

- Proof outline:
  - Again let $A = \{X \geq 0\}$ and write

$$\varphi_n(t) = \int_A X^n e^{tX} dP + \int_{A^c} X^n e^{tX} dP = \varphi_{n,1}(t) + \varphi_{n,2}(t).$$

  - Suppose $t \in G$. Since G is open there exist $v, u \in G$ such that

$$t < v < u.$$

  - By the mean value theorem, for any s, t and any $y \geq 0$ there exists a ξ between t and s such that

$$\frac{e^{ty} - e^{sy}}{t - s} = y e^{\xi y}.$$

  - If $s < v$ and $t < v$ then

$$0 \leq \frac{e^{ty} - e^{sy}}{t - s} \leq y e^{vy},$$

    and

$$0 \leq y^n \left(\frac{e^{ty} - e^{sy}}{t - s} \right) \leq y^{n+1} e^{vy} \leq K_{n+1}(u - v) e^{uy}.$$

– Therefore, if $s < v$ and $t < v$ the integrands in

$$\frac{\varphi_{n,1}(t) - \varphi_{n,1}(s)}{t-s} = \int_A X^n \frac{e^{tX} - e^{sX}}{t-s} dP$$

are dominated on $A = \{X \geq 0\}$ by

$$\left| X^n \frac{e^{tX} - e^{sX}}{t-s} \right| \leq K_{n+1}(u-v)e^{uX},$$

and

$$\int_A e^{uX} dP \leq M(u) < \infty.$$

– So by the dominated convergence theorem

$$\begin{aligned} \lim_{s \to t} \frac{\varphi_{n,1}(t) - \varphi_{n,1}(s)}{t-s} &= \int_A X^n \lim_{s \to t} \frac{e^{tX} - e^{sX}}{t-s} dP \\ &= \int_A X^{n+1} e^{tX} dP \\ &= \varphi_{n+1,1}(t). \end{aligned}$$

– A similar argument shows that $\varphi'_{n,2}(t) = \varphi_{n+1,2}(t)$.

An Inversion Theorem for Characteristic Functions

- The characteristic function of a real-valued random variable X is

$$\varphi_X(t) = E[e^{itX}] = E[\cos(tX)] + iE[\sin(tX)].$$

- Some properties:
  - The characteristic function exist for all t.
  - $\varphi_X(0) = 1$.
  - $|\varphi_X(t)| \leq 1$ for all t.
  - If $Z \sim \mathrm{N}(0,1)$ then $\varphi_X(t) = e^{-t^2/2}$.
- Suppose

$$\int_{-\infty}^{\infty} |\varphi_X(t)| dt < \infty.$$

  Then X is absolutely continuous and has a bounded density given by

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

- Suppose X is a random variable with characteristic function φ_X, $Z \sim N(0,1)$, $\sigma > 0$, and X, Z are independent. Then

$$\int_{-\infty}^{\infty} |\varphi_{X+\sigma Z}(t)| dt < \infty.$$

  - The characteristic function of $X + \sigma Z$ is

$$\varphi_{X+\sigma Z}(t) = \varphi_X(t) e^{-\sigma^2 t^2/2},$$

    so

$$\int_{-\infty}^{\infty} |\varphi_{X+\sigma Z}(t)| dt = \int_{-\infty}^{\infty} |\varphi_X(t) e^{-\sigma^2 t^2/2}| dt \leq \int_{-\infty}^{\infty} e^{-\sigma^2 t^2/2} dt < \infty$$

Uniqueness of Characteristic Functions

- Suppose X and Y have characteristic functions φ_X and φ_Y.
- If $\varphi_X(t) = \varphi_Y(t)$ for all t then X and Y have the same distribution.
  - Let X be independent of X and Y.
  - Then $X + \sigma Z$ and $Y + \sigma Z$ have the same distribution for all $\sigma > 0$.
  - now let $\sigma \to 0$.

Uniqueness of Moment Generating Functions

- Suppose X and Y have moment generating functions M_X and M_Y.
- If M_X and M_Y are finite and equal on a neighborhood of the origin, then X and Y have the same distribution.
  - Let B be a neighborhood of the origin where $M_X(t)$ and $M_Y(t)$ are finite and equal.
  - The complex generating functions

$$G_X(z) = E[e^{zX}]$$
$$G_Y(z) = E[e^{zY}]$$

    are defined and differentiable on the open set D of all complex numbers z with $\Re(z) \in B$.
  - Complex functions differentiable on an open subset of $\mathbb{C}$ are *analytic*, which implies
    * infinitely differentiable
    * they have power series expansions near every point of the open set
  - If a function f is analytic on an ope set D, and there is an infinite subset $E \subset D$ such that
    * $f(z) = 0$ for all $z \in E$;
    * E has a limit point in D;

    then $f(z) = 0$ for all $z \in D$.
  - Since G_X and G_Y are analytic on D and

$$G_X(t+0i) = M_X(t) = M_Y(t) = G_Y(t+0i)$$

    for all $t \in B$ this means that $G_X(z) = G_Y(z)$ for all $z \in D$.
  - Since D includes the imaginary axis, this means that X and Y have identical characteristic functions.
  - So X and Y have the same distribution.

Wednesday, March 30, 2016

Completeness in Exponential Families

- Let μ be a sigma-finite measure on the Borel subsets of $\mathbb{R}$.
- Let
$$\mathcal{N} = \left\{ \eta \in \mathbb{R} : \int_{-\infty}^{\infty} e^{\eta t} \mu(dt) < \infty \right\}$$
- Define the densities f_η with respect to μ
$$f_\eta(t) = \frac{e^{\eta t}}{\int_{-\infty}^{\infty} e^{\eta u} \mu(du)} = c(\eta) e^{\eta t}$$
- A family of densities
$$\{ f_\eta : \eta \in \mathcal{N} \}$$
is an exponential family.
- Suppose $D \subset \mathcal{N}$ is an open interval, g is measurable, and
$$\int_{-\infty}^{\infty} g(t) f_\eta(t) \mu(dt) = 0$$
for all $\eta \in D$. Then $g(t) = 0$ for μ-almost all t.

- Proof outline:
  - Under the hypotheses, for $\eta \in D$

$$\int_{-\infty}^{\infty} g^+(t)e^{\eta t}\mu(dt) = \int_{-\infty}^{\infty} g^-(t)e^{\eta t}\mu(dt)$$

  - Suppose for some $\eta_0 \in D$

$$d = \int_{-\infty}^{\infty} g^+(t)e^{\eta_0 t}\mu(dt) = \int_{-\infty}^{\infty} g^-(t)e^{\eta_0 t}\mu(dt) > 0.$$

  - Then $\frac{1}{d}g^+$ and $\frac{1}{d}g^-$ are probability density functions with respect to μ.
  - Taking $\eta = s + \eta_0$, their moment generating functions are

$$M^+(s) = \frac{1}{d}\int_{-\infty}^{\infty} g^+(t)e^{(s+\eta_0)t}\mu(dt)$$
$$M^-(s) = \frac{1}{d}\int_{-\infty}^{\infty} g^-(t)e^{(s+\eta_0)t}\mu(dt).$$

  - These are finite and equal on a neighborhood of the origin.
  - So $g^+ = g^-$ μ-almost everywhere, and $g = g^+ - g^- = 0$ μ-almost everywhere.
  - So the family of distributions $\{f_\eta : \eta \in D\}$ is complete.

Some Function Spaces

- For $1 \le p < \infty$ a measurable function f is in $\mathscr{L}^p(\mu)$ if

$$\int |f|^p d\mu < \infty$$

- A measurable function f is in $\mathscr{L}^\infty(\mu)$ if

$$\text{esssup}|f| < \infty.$$

- For $f \in \mathscr{L}(\mu)$ Define

$$\|f\|_p = \sqrt[p]{\int |f|^p d\mu}$$
$$\|f\|_\infty = \text{esssup}|f|$$

- Some properties:
  - $\|f\|_p$ is a norm for $1 \le p \le \infty$ if functions that are μ-almost everywhere equal are considered equivalent.
  - If $\|f\|_r < \infty$ for some $r < \infty$, then

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty.$$

  - If $1 \le p \le \infty$, $1/p + 1/q = 1$, $f \in \mathscr{L}^p(\mu)$ and $g \in \mathscr{L}^q(\mu)$, then $fg \in \mathscr{L}^1(\mu)$ and

$$\|fg\|_1 \le \|f\|_p \|g\|_q.$$

- $\mathscr{L}^p(\mu)$ is a complete metric space for $0 \le p \le \infty$.
- Proof outline for $p < \infty$:
  - Let f_n be a Cauchy sequence.
  - There exists a subsequence f_{n_k} such that

$$\|f_{n_{i+1}} - f_{n_i}\|_p < \frac{1}{2^i}.$$

  - Let

$$g_k = \sum_{i=1}^{k} |f_{n_{i+1}} - f_{n_i}|$$
$$g = \sum_{i=1}^{\infty} |f_{n_{i+1}} - f_{n_i}|.$$

  - Then

$$\|g_k\|_p \le \sum_{i=1}^{k} \|f_{n_{i+1}} - f_{n_i}\|_p \le \sum_{i=1}^{k} \frac{1}{2^i} < 1.$$

  - By Fatou's lemma, $\|g\|_p \le 1$.
  - So the series

$$f_{n_1} + \sum_{i=1}^{\infty} (f_{n_{i+1}} - f_{n_i})$$

    converges on a set A with $\mu(A^c) = 0$.
  - Define

$$f(x) = \begin{cases} f_{n_1}(x) + \sum_{i=1}^{\infty} (f_{n_{i+1}}(x) - f_{n_i}(x)) & x \in A \\ 0 & x \notin A. \end{cases}$$

  - Then for almost all x

$$f_{n_k} = f_{n_1}(x) + \sum_{i=1}^{k-1} (f_{n_{i+1}}(x) - f_{n_i}(x)) \to f(x).$$

  - We now need to show that $f \in \mathscr{L}^p(\mu)$ and $\|f - f_n\|_p \to 0$.
  - Fix $\varepsilon > 0$ and find N such that $n, m \ge N$ imply $\|f_m - f_n\|)p < \varepsilon$.

  - Then by Fatou's lemma

$$\int |f - f_n|^p d\mu \le \liminf_{i\to\infty} \int |f_{n_i} - f_n|^p d\mu \le \varepsilon^p.$$

  - So

$$\|f\|_p \le \|f_n\|_p + \|f - f_n\|_p < \infty$$

    and $\|f - f_n\|_p \le \varepsilon$ for all $n \ge N$.

- Proof outline for $p = \infty$:
  - Let A be the set of all x where

$$|f_n(x)| \le \|f_n\|_\infty$$
$$|f_n(x) - f_m(x)| \le \|f_n - f_m\|_\infty$$

    for all n, m.
  - Then $\mu(A^c) = 0$ and on A the sequence f_n converges uniformly to a function f.
  - Set $f(x) = 0$ for $x \notin A$.
  - Then $\|f - f_n\|_\infty \to 0$.

Friday, April 1, 2016

Hilbert Spaces

- For $\mathscr{L}^2(\mu)$ define the inner product

$$\langle f, g \rangle = \int f g d\mu.$$

- Then $\|f\|_2 = \sqrt{\langle f, f \rangle}$.
- Since $\mathscr{L}^2(\mu)$ is a complete inner product space, it is a Hilbert space.
- Let L be a closed subspace of a Hilbert space and x and arbitrary point.
- Then there exists a unique $x_0 \in L$ such that

$$\|x - x_0\| = \inf\{\|x - y\| : y \in L\}.$$

- x_0 is the unique vector in L such that $\langle x - x_0, y \rangle = 0$ for all $y \in L$.
- x_0 is the *orthogonal projection* of x on L.
  - A normed space is Euclidean if and only if its norm satisfies the *parallelogram identity* for all u, v:

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

  - Let $D = \inf\{\|x - y\| : y \in L\}$.
  - Suppose $y_n \in L$ are such that $\|x - y_n\| \to D$.
  - Then applying the parallelogram identity to

$$u = \frac{x - y_n}{2} \qquad\qquad v = \frac{x - y_m}{2}$$

gives

$$\frac{1}{4}\|y_n - y_m\|^2 = \frac{1}{2}\|x - y_n\|^2 + \frac{1}{2}\|x - y_m\|^2 - \|x - \frac{y_n + y_m}{2}\|^2$$

- Since $(y_n + y_m)/2 \in L$,

$$\|x - \frac{y_n + y_m}{2}\| \geq D.$$

- Since $\|x - y_n\| \to D$ this shows that y_n is a Cauchy sequence which converges to a limit y by completeness.
- By continuity of the norm, $\|x - y\| = D$.
- If $y' \in L$ also satisfies $\|x - y'\| = D$ then the same use of the parallelogram identity implies $\|y - y'\| = 0$.
- If $y \in L$, then for any scalar t

$$\|x - x_0 + ty\|^2 = \|x - x_0\|^2 + t^2\|y\|^2 + 2t\langle x - x_0, y\rangle.$$

- This is must be minimized at $t = 0$, so $\langle x - x_0, y\rangle = 0$.

- If $x_1 \in L$ satisfies $\langle x - x_1, y\rangle = 0$ for all $y \in L$, then

$$\begin{aligned}\|x - x_0\|^2 &= \|x - x_1\|^2 + \|x_1 - x_0\|^2 + 2\langle x - x_1, x_1 - x_0\rangle \\ &\geq \|x - x_1\|^2\end{aligned}$$

So $x_1 = x_0$.

Conditional Expectation as Projection

- Suppose
  - $(S, \mathscr{B}, P)$ is a probability space;
  - $\mathscr{C} \subset \mathscr{B}$ is a sigma-algebra;
  - $X \in \mathscr{L}^2(P)$.
- Let L he the set of all $\mathscr{C}$-measurable random variables with finite second moments.
- L is a closed subspace.
- There exists a unique random variable $X_0 \in L$ such that

$$E[XY] = E[X_0 Y]$$

  for all $Y \in L$.
- X_0 is the conditional expectation of X given $\mathscr{C}$.
- $E[(X - E[X|\mathscr{C}])^2] \leq E[(X - Y)^2]$ for all $Y \in L$.
- This definition can be extended to $X \in \mathscr{L}^1(P)$ and non-negative X by taking limits.

Linear Functionals on a Hilbert space

- A linear functional on a vector space is a real-valued function f such that
  - $f(ax) = af(x)$ for any vector x and scalar a;
  - $f(x+y) = f(x) + f(y)$ for any vectors x, y.
- If f is a continuous linear functional on a Hilbert space, then there exists a vector y such that
$$f(x) = \langle y, x \rangle$$
for all x.
- This is called the *Riesz representation theorem*. There are several other related results known by the same name.
- In finite dimensional Euclidean spaces all linear functionals are continuous, and this result is easy to establish by considering an orthogonal basis.
- In infinite dimensional spaces not all linear functionals are continuous.
- For example, in $C^2_{[a,b]}$, the space of continuous functions on $[a,b]$ with the norm
$$\|x\| = \sqrt{\int_a^b x(t)^2 dt}$$
the *evaluation functional*
$$f_t(x) = x(t)$$
is linear but is not continuous.
- One possible proof for the general case is based on projections.

- Outline of proof:
  - If $f(x) = 0$ for all x use $y = 0$.
  - Suppose there is some z such that $f(z) \neq 0$.
  - Let $L = \{x : f(x) = 0\}$.
  - L is a subspace since f is linear; L is the *null space* of f.
  - L is closed since f is continuous.
  - Let z_0 be the projection of z onto L, and let $v = z - z_0$.
  - Then $\langle v, w \rangle = 0$ for all $w \in L$, and $f(v) = f(z) \neq 0$.
  - Let
$$y = \frac{f(v)}{\|v\|^2} v$$
  - So $\langle y, w \rangle = 0$ for all $w \in L$, and
$$f(y) = \frac{f(v)^2}{\|v\|^2} = \|y\|^2 = \langle y, y \rangle.$$
  - For any x the vector
$$x - \frac{f(x)}{f(y)} y$$
    is in L since
$$f\left(x - \frac{f(x)}{f(y)} y\right) = f(x) - \frac{f(x)}{f(y)} f(y) = f(x) - f(x) = 0.$$
  - Therefore
$$\langle y, x - \frac{f(x)}{f(y)} y \rangle = 0,$$
    and
$$\langle y, x \rangle = \langle y, x - \frac{f(x)}{f(y)} y \rangle + \langle y, \frac{f(x)}{f(y)} y \rangle = \frac{f(x)}{f(y)} \langle y, y \rangle = f(x).$$

Reproducing Kernel Hilbert Spaces

- Suppose the elements of a Hilbert space are real-valued functions on some domain D.
- The *evaluation functional*

$$f_t(x) = x(t)$$

  is a linear functional.
- If these functionals are continuous for all $t \in D$ then the Hilbert space is said to be a *reproducing kernel Hilbert space* (RKHS).
- By the Riesz representation theorem, for each t there exists a function y_t such that

$$f_t(x) = \langle y_t, x \rangle.$$

- The *reproducing kernel* is

$$R(s,t) = y_s(t) = \langle y_s, y_t \rangle$$

- Reproducing kernel Hilbert spaces are used heavily to study a wide range of regression and machine learning methods.

Week 11

Monday, April 4, 2016

Orthogonal Systems and Bases

- A collection $\{x_\alpha\}$ of points in a Hilbert space is an *orthogonal system* if $\|x_\alpha\| > 0$ for all α and
$$\langle x_{\alpha_1}, x_{\alpha_2} \rangle = 0$$
for all $\alpha_1 \neq \alpha_2$.
- An orthogonal system is an orthonormal system if $\|x_\alpha\| = 1$ for all α.
- If a Hilbert space is separable then any orthogonal system is at most countable.
- An orthogonal system is complete, or an orthogonal basis, if the smallest closed subspace containing the system is the space itself.
- Every separable Hilbert space has a countable orthonormal base.

- Suppose $\phi_1, \phi_2, \ldots$ is an orthonormal system and x is an element of a Hilbert space.
  - For a fixed n the expression

$$\left\| x - \sum_{i=1}^{n} a_i \phi_i \right\|^2$$

    is minimized by

$$a_i = c_i = \langle x, \phi_i \rangle.$$

  - The minimal value is

$$\|x\|^2 - \sum_{i=1}^{n} c_i^2.$$

  - This implies *Bessel's inequality*

$$\sum_{i=1}^{\infty} c_i^2 \leq \|x\|^2$$

- Proof outline:
  - The space L spanned by $\phi_1, \ldots, \phi_n$ is a closed linear subspace.
  - By the general projection result there is a unique closest point $x_0 \in L$ to x, and $x - x_0$ is orthogonal to all elements of L.
  - So if
    $$x_0 = \sum_{i=1}^{n} b_i \phi_i$$
    then for $k = 1, \ldots, n$
    $$0 = \langle x - x_0, \phi_k \rangle = \langle x, \phi_k \rangle - \left\langle \sum_{i=1}^{n} b_i \phi_i, \phi_k \right\rangle = c_k - b_k.$$
  - So $b_k = c_k = \langle x, \phi_k \rangle$.
  - Since $x_0 \in L$ and thus $\langle x - x_0, x \rangle = 0$, it follows that
    $$\|x\|^2 = \|x - x_0\|^2 + \|x_0\|^2.$$
  - So
    $$\|x - x_0\|^2 = \|x\|^2 - \|x_0\|^2,$$
    and
    $$\|x_0\|^2 = \langle x_0, x_0 \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \phi_i, \phi_j \rangle = \sum_{i=1}^{n} c_i^2.$$

- If $\phi_1, \phi_2, \phi_3, \ldots$ is an orthonormal basis, then

$$\lim_{n\to\infty} \left\| x - \sum_{i=1}^{n} c_i \phi_i \right\| = 0$$

and

$$\|x\|^2 = \sum_{i=1}^{\infty} c_i^2.$$

- This identity is known as *Parseval's Theorem*.
- The coefficients $c_i = \langle x, \phi_i \rangle$ are called *Fourier coefficients*
- The series

$$\sum_{i=1}^{\infty} c_i \phi_i$$

is called the *Fourier series*.

- A basis for $\mathscr{L}^2$ on $[-\pi, \pi]$ is given by

$$1, \quad \cos t, \quad \sin t, \quad \cos 2t, \quad \sin 2t, \quad \ldots \quad .$$

Orthogonal Complements and Direct Sums

- Suppose L is a closed linear subspace of a Hilbert space H.
- Let
$$L^\perp = \{x \in H : \langle x, y\rangle = 0 \text{ for all } y \in L\}$$
- $L^\perp$ is called the *orthogonal complement* of L.
- $L^\perp$ is a closed linear subspace of H.
- Every vector $x \in H$ can be written as
$$x = x_1 + x_0$$
with $x_0 \in L$ and $x_1 \in L^\perp$. This decomposition is unique.
- H is said to be the *direct sum* of L and $L^\perp$:
$$H = L \oplus L^\perp.$$
- Direct sums are also called *tensor sums*.
- The orthogonal complement is also sometimes denoted
$$L^\perp = H \ominus L.$$
- Suppose
  - $L_1, \ldots, L_K$ are closed linear subspaces of L;
  - $\langle x_i, x_j\rangle = 0$ for any $x_i \in L_i, x_j \in L_i$ with $i \neq j$;
  - every $x \in L$ can be written as $x = x_1 + \cdots + x_K$ with $x_i \in L_i$.

  Then L is the direct sum of $L_1, \ldots, L_K$,
$$L = L_1 \oplus \cdots \oplus L_K.$$
- This is often used in hypothesis testing for linear models, with
  - L representing a full model;
  - $L_1 \subset L$ a reduced model;
  - L_2 the orthogonal complement of L_1 within L,
$$L_2 = L \ominus L_1.$$

Projection Operators

- Given a closed subspace L or a Hilbert space H the *projection operator* P_L is the function that returns the projection of its argument on L:

$$x_0 = P_L x.$$

- The projection operator P_L
  - is a linear operator;
  - is continuous.
- If $L_1, \ldots, L_K$ are closed subspaces of L and

$$L = L_1 \oplus L_2 \oplus \cdots \oplus L_K,$$

then

$$P_L = P_{L_1} + P_{L_2} + \cdots + P_{L_K}.$$

- If $y \in L$ and $x \perp y$, i.e. $\langle x, y \rangle = 0$, then $P_L x \perp y$ as well:
  - $\langle x, y \rangle = \langle x - P_L x, y \rangle + \langle P_L x, y \rangle$.
  - $\langle x - P_L x, y \rangle = 0$ since $y \in L$.

Wednesday, April 6, 2016

RKHS and Model Fitting

- Suppose we have data of the form $(x_1, y_1), \ldots, (x_n, y_n)$, with
  - $x_i \in D$ for some domain D;
  - $y_i \in \mathbb{R}$.
- We eould like to model the relationship between the predictors, or features, x_i and the responses y_i.
- A parametric model might consider functions f of the form

$$f(x) = \sum_{j=1}^{K} d_j g_j(x)$$

  for some basis functions $g_1, \ldots, g_K$.
- The coefficients $d_1, \ldots, d_K$ could be estimated by minimizing a loss function representing lack of fit, minimizing

$$\sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{K} d_j g_j(x_i)\right)$$

- The loss function could be
  - squared error $L(y, \theta) = (y - \theta)^2$;
  - a negative log likelihood term, e.g. for a generalized linear model

- A more flexible approach allows f to vary in infinite dimensions.
- Minimizing a loss function over too many parameters results in *over-fitting*.
- To avoid over-fitting requires *regularization*, or a *penalty*.
- Penalties usually are functions $J(f)$ that are larger for functions behaving less nicely, e.g. larger for functions with less smoothness.
- Regularized, or penalized, function estimates are obtained by minimizing

$$\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(f)$$

  where $\lambda > 0$ is a *regularization* parameter.

- An intermediate version minimizes

$$\sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{K} d_j g_j(x_i) + f(x_i)\right) + \lambda J(f)$$

  over f and $d_1, \ldots, d_K$.

- The estimates of $d_1, \ldots, d_K$ are not penalized, or regularized.

- Many methods of this type
  - take the set of admissible functions f to be a RKHS H with kernel $R(s,t)$;
  - take $J(f) = \|f\|^2$.
- A function f can be written as

$$f(x) = \sum_{j=1}^{n} c_j R(x_j, x) + \rho(x)$$

  where $\rho(\cdot) \perp R(x_j, \cdot)$ for $j = 1, \ldots, n$.
- As a result,

$$\rho(x_i) = \langle R(x_i, \cdot), \rho \rangle = 0,$$

  and

$$\begin{aligned} \sum_{i=1}^{n} L(y_i, f(x_i)) &= \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} c_j R(x_j, x_i) + \rho(x_i)\right) \\ &= \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} c_j R(x_j, x_i)\right). \end{aligned}$$

- Furthermore,

$$J(f) = \|f\|^2 = \left\| \sum_{j=1}^{n} c_j R(x_j, \cdot) \right\|^2 + \|\rho\|^2.$$

- Thus the penalized loss function is

$$\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(f)$$
$$= \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} c_j R(x_j, x_i)\right) + \lambda \left\| \sum_{j=1}^{n} c_j R(x_j, \cdot) \right\|^2 + \lambda \|\rho\|^2.$$

- ρ enters the objective function only in the term $\|\rho\|^2$, which is minimized by $\rho = 0$.

- So the problem of minimizing the penalized loss over the infinite dimensional space H reduces to solving the finite dimensional problem of minimizing

$$\sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} c_j R(x_j, x_i)\right) + \lambda \left\| \sum_{j=1}^{n} c_j R(x_j, \cdot) \right\|^2.$$

  over $c_1, \ldots, c_n$.

- A common variation takes

$$J(f) = \|P_L f\|^2$$

  where L is a closed subspace.

- This approach considers representations of f of the form

$$f(x) = \sum_{j=1}^{K} d_j g_j(x) + \sum_{j=1}^{n} c_j R_L(x_j, x) + \rho(x)$$

  where

  - $g_1, \ldots, g_K$ is a basis for $L^\perp$;
  - R_L is the reproducing kernel corresponding to L;
  - $R_L(x, \cdot) \in L$ for all x;
  - $\rho \in L$ and $\rho \perp R_L(x_i, \cdot)$.

- Minimizing

$$\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(f)$$

  over f is then equivalent to the finite dimensional problem of minimizing

$$\sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{K} d_j g_j(x) + \sum_{j=1}^{n} c_j R_L(x_j, x)\right) + \lambda \left\| \sum_{j=1}^{n} c_j R_L(x_j, x) \right\|^2$$

  over $d_1, \ldots, d_K$ and $c_1, \ldots, c_n$.

- The coefficients $d_1, \ldots, d_K$ are not affected by the penalty term.

- For $D = [0,1]$ and m a positive integer
  - let H be the set of all m times differentiable functions on D
  - with the inner product

$$\langle f, g \rangle = \sum_{j=0}^{m-1} f^{(j)}(0) g^{(j)}(0) + \int_0^1 f^{(m)}(t) g^{(m)}(t) dt$$

- The unpenalized basis functions $g_0, \ldots, g_{m-1}$ are

$$g_i(x) = \frac{x^i}{i!}$$

- The subspace L is the orthogonal complement of the span of $g_0, \ldots, g_{m-1}$.
- The penalty function is

$$\|P_L f\| = J(f) = \int_0^1 (f^{(m)}(t))^2 dt$$

and

$$R_L(x,y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du,$$

where $(x)_+ = \max(x, 0)$.

- $m = 1$ corresponds to fitting piecewise linear continuous functions, with

$$R_L(s,t) = \min(s,t).$$

- $m = 2$ corresponds to cubic smoothing splines, with

$$R_L(s,t) = \frac{max(s,t) \min(s,t)^2}{2} - \frac{\min(s,t)^3}{6}.$$

- *Support vector machines* are another fitting approach that builds on RKHS methodology.

Multiple Integration

- Suppose
  - $X = \mathbb{R}^p$;
  - $\mathscr{M} = \mathscr{R}^p$, the Borel sets in $\mathbb{R}^p$;
  - $\mu = m_p$ is Lebesgue measure restricted to the Borel sets.
- For a Borel-measurable function $f : \mathbb{R}^p \to \mathbb{R}$ the Lebesgue integral

$$\int f d\mu = \int f dm_p$$

  is defined if f is non-negative or $f \in \mathscr{L}(m_p)$.
- To compute the integral for, say, $p = 2$ we would like to have

$$\begin{aligned} \int f dm_2 &= \int \left(\int f(x,y) m_1(dx) \right) m_1(dy) \\ &= \int \left(\int f(x,y) m_1(dy) \right) m_1(dx) \end{aligned}$$

- This is the subsect of Fubini's theorem.

- Suppose $(X, \mathscr{M}, \mu), (Y, \mathscr{N}, \nu)$ are measure spaces with sigma-finite measures.
- Let $X \times Y = \{(x,y) : x \in X, y \in Y\}$.
- A subset $A \times B \subset X \times Y$ is called a *measurable rectangle* if $A \in \mathscr{M}$ and $B \in \mathscr{N}$.
- The *product sigma-algebra* $\mathscr{M} \times \mathscr{N}$ is the smallest sigma-algebra containing all measurable rectangles.
- If $X = Y = \mathbb{R}$ and $\mathscr{M} = \mathscr{N} = \mathscr{R}$, then $\mathscr{R} \times \mathscr{R} = \mathscr{R}^2$.
- For $E \subset X \times Y$, $x \in X$, and $y \in Y$ define

$$E_x = \{y : (x,y) \in E\}$$
$$E^y = \{x : (x,y) \in E\}$$

  These are called the x-section and y-section of E.

- If $E \in \mathscr{M} \times \mathscr{N}$ then $E_x \in \mathscr{N}$ and $E^y \in \mathscr{M}$ for all $x \in X$ and $y \in Y$:
  - Let Ω be the class of all $E \in \mathscr{M} \times \mathscr{N}$ such that $E_x \in \mathscr{N}$ for all $x \in X$.
  - If $E = A \times B$ then
$$E_x = \begin{cases} B & \text{if } x \in A \\ \emptyset & \text{if } x \notin A \end{cases}$$
    So Ω contains all measurable rectangles.
  - $(X \times Y)_x = Y \in \mathscr{N}$.
  - If $E \in \Omega$ then $(E^c)_x = (E_x)^c \in \mathscr{N}$, so $E^c \in \Omega$.
  - If $E_1, E_2, \cdots \in \Omega$ and $E = \bigcup E_i$ then
$$E_x = \left(\bigcup E_i\right)_x = \bigcup (E_i)_x \in \mathscr{N},$$
    so $E \in \Omega$. So $\Omega \subset \mathscr{M} \times \mathscr{N}$ is a sigma-algebra containing all open rectangles.
  - Therefore $\Omega = \mathscr{M} \times \mathscr{N}$.
  - The proof for y-sections is analogous.

Friday, April 8, 2016

Multiple Integration

The $\pi - \lambda$ Theorem

- A common line of argument for proving results will be:

  (i) show the result holds for measurable rectangles;

  (ii) show the result holds for A^c if it holds for A;

  (iii) show the result holds for $A = \bigcup_{i=1}^{\infty} A_i$ if it holds for $A_1, A_2, \ldots$.

  This then implies that the result holds for the product sigma algebra.

- Sometimes it is easy to show (iii) only of the A_i are assumed disjoint.

- The $\pi - \lambda$ theorem bridges the gap.

- The collection of measurable rectangles is a π-system.

- A collection $\mathscr{P}$ of subsets of a set X is a π-*system* if $A, B \in \mathscr{P}$ implies $A \cap B \in \mathscr{P}$.

- A collection $\mathscr{L}$ of subset of a set X is a λ *system* if

  - $X \in \mathscr{L}$;
  - if $A \in \mathscr{A}$ then $A^c \in \mathscr{A}$;
  - if $A_1, A_2, \cdots \in \mathscr{A}$ are pairwise disjoint, then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{L}$.

- If a λ system is also a π-system then it is a sigma algebra.

- The $\pi - \lambda$ theorem: If $\mathscr{P}$ is a π, $\mathscr{L}$ is a λ system, and $\mathscr{P} \subset \mathscr{L}$, then $\sigma(\mathscr{P}) \subset \mathscr{L}$.

- Thus to show that a property holds for all sets in $\mathscr{M} \times \mathscr{N}$ it will suffice to show

  - that is holds for all measurable rectangles;
  - that the collection of sets in $\mathscr{M} \times \mathscr{N}$ for which it holds is a λ-system.

- For a function f on $X \times Y$ and each $x \in X$ define

$$f_x(y) = f(x,y).$$

- Similarly, for $y \in Y$ define

$$f^y(x) = f(x,y).$$

- If f is $\mathscr{M} \times \mathscr{N}$-measurable, then f_x is $\mathscr{N}$-measurable for each $x \in X$, and f^y is $\mathscr{M}$-measurable for each $y \in Y$:
  - If $G \subset \mathbb{R}$ is open, then $E = f^{-1}(G) \in \mathscr{M} \times \mathscr{N}$ since f is measurable.
  - $f_x^{-1}(G) = E_x \in \mathscr{N}$.
- For $E \in \mathscr{M} \times \mathscr{N}$ let

$$\varphi_E(x) = \nu(E_x)$$
$$\psi_E(y) = \mu(E^y).$$

- Then φ_E is $\mathscr{M}$-measurable, ψ_E is $\mathscr{N}$-measurable, and

$$\int_X \varphi_E d\mu = \int_Y \psi_E d\nu.$$

- Proof for the case where $\mu(X)$ and $\nu(Y)$ are finite:
  - Let Ω be the set of all $E \in \mathscr{M} \times \mathscr{N}$ for which the conclusion holds.
  - If $E = A \times B$ is a measurable rectangle,

$$\varphi_{A\times B}(x) = \nu(B)1_A(x) \qquad \psi_{A\times B}(y) = \mu(A)1_B(y)$$

    and thus

$$\int_X \varphi_{A\times B} d\mu = \mu(A)\nu(B) = \int_Y \psi_{A\times B} d\nu,$$

    and $E \in \Omega$.
  - $X \times Y \in \Omega$ since it is a measurable rectangle.
  - If $E \in \Omega$ then

$$\begin{aligned}\varphi_{E^c}(x) &= \nu((E^c)_x) = \nu((E_x)^c) = \nu(Y) - \nu(E_x) = \nu(Y) - \varphi_E(x)\\ \psi_{E^c}(y) &= \mu((E^c)^y) = \mu((E^y)^c) = \mu(Y) - \mu(E^y) = \mu(X) - \psi_E(y).\end{aligned}$$

    So φ_E and ψ_E are measurable. Furthermore,

$$\begin{aligned}\int_X \varphi_{E^c} s\mu &= \int_X \nu(Y) - \varphi_E d\mu = \mu(X)\nu(Y) - \int_X \varphi_E d\mu\\ &= \mu(X)\nu(Y) - \int_Y \psi_E d\nu = \int_Y \mu(X) - \psi_E d\nu = \int_Y \psi_{E^c} d\nu\end{aligned}$$

    So $E^c \in \Omega$.
  - A similar argument shows that if $E_1, E_2, \cdots \in \Omega$ and are pairwise disjoint, then $\bigcup E_i \in \Omega$ as well.
  - So $\Omega \subset \mathscr{M} \times \mathscr{N}$ is a λ-system containing the π-system of measurable rectangles.
  - Therefore $\Omega = \mathscr{M} \times \mathscr{N}$.

- The product measure $\mu \times \nu$ on $\mathscr{M} \times \mathscr{N}$ is defined as

$$(\mu \times \nu)(E) = \int_X \nu(E_x) d\mu = \int_Y \mu(E^y) d\nu$$

  for $E \in \mathscr{M} \times \nu$.

- If m_p is Lebesgue measure restricted to the Borel subsets of $\mathbb{R}^p$, then $m_1 \times m_1 = m_2$.

- $\mu \times \nu$ is the unique measure λ on $\mathscr{M} \times \mathscr{N}$ such that for all measurable rectangles $A \times B$

$$\lambda(A \times B) = \mu(A)\nu(B).$$

- If μ and ν are probabilities and

$$V(x,y) = x$$
$$W(x,y) = y$$

  then $\mu \times \nu$ is a probability under which V and W are independent.

- Suppose f is non-negative and $\mathscr{M} \times \mathscr{N}$-measurable.
- Let

$$\varphi_f(x) = \int_Y f_x d\nu \qquad\qquad \psi_f(y) = \int_X f^y d\mu.$$

- Then φ_f is $\mathscr{M}$-measurable, ψ_f is $\mathscr{N}$-measurable, and

$$\int_{X\times Y} f d(\mu \times \nu) = \int_X \varphi_f d\mu = \int_Y \psi_f d\nu.$$

- If $f \in \mathscr{L}^1(\mu \times \nu)$
  - $f_x \in \mathscr{L}^1(\nu)$ for μ-almost all x;
  - $f^y \in \mathscr{L}^1(\mu)$ for ν-almost all y;
  - $\varphi_f(x)$ is defined for μ-almost all x and is in $\mathscr{L}^1(\mu)$;
  - $\psi_f(y)$ is defined for ν-almost all y and is in $\mathscr{L}^1(\nu)$;
  - The integrals are defined, finite, and equal.
- These results are known as *Fubini's theorem*.
- The results for non-negative f are sometimes called *Tonelli's theorem*.
- The proof
  - uses the results already obtained to handle simple functions;
  - monotone convergence extends to non-negative functions;
  - positive and negative parts handled separately for integrable functions.

- Suppose μ is a sigma-finite measure on the Borel subsets of $[0,\infty)$.
- Then

$$\begin{aligned}\int_{[0,\infty)} x d\mu &= \int_{[0,\infty)} \int_{[0,\infty)} 1_{[0,x)}(y) dy d\mu \\ &= \int_{[0,\infty)} \int_{[0,\infty)} 1_{(y,\infty)}(x) dy d\mu \\ &= \int_{[0,\infty)} \int_{[0,\infty)} 1_{(y,\infty)}(x) d\mu dy \\ &= \int_{[0,\infty)} \mu((y,\infty)) dy.\end{aligned}$$

- If μ is a probability measure with CDF $F(y) = \mu([0,y])$, then

$$E[X] = \int x d\mu = \int_0^\infty (1 - F(y)) dy.$$

- The result is true
  - even if $E[X] = \infty$
  - for any distribution on $[0,\infty)$, whether continuous, discrete, ...

Riemann Integrals and Sequences of Partitions

.

- Suppose f is a bounded function on $[a,b]$ and P_n and Q_n are sequences of partitions such that

$$\lim_{n\to\infty} L(P_n,f) = \lim_{n\to\infty} U(Q_n,f).$$

  Then $f \in \mathscr{R}$ on $[a,b]$:

$$\lim_{n\to\infty} L(P_n,f) \le \sup_P L(P,f) \le \inf_P U(P,f) \le \lim_{n\to\infty} U(Q_n,f).$$

- For a converse, define the *mesh* of a partition P with points $a = x_0 \le x_1 \le x_2 \le \cdots \le x_n = b$ as

$$|P| = \max_{1\le i\le n} \Delta x_i = \max_{1\le i\le n} |x_i - x_{i-1}|.$$

- If $f \in \mathscr{R}$ on $[a,b]$ and $\varepsilon > 0$, then there exists a $\delta > 0$ such that

$$\int_a^b f dx - \varepsilon < L(P,f) \le U(P,f) < \int_a^b f dx + \varepsilon$$

  for all partitions P with $|P| < \delta$:

  - Since $f \in \mathscr{R}$ there exists an M such that $|f| \le M$ and a partition Q with points $a = y_0 \le y_1 \le y_2 \le \cdots \le y_m = b$ such that

$$\int_a^b f dx - \frac{\varepsilon}{2} < L(Q,f) \le U(Q,f) < \int_a^b f dx + \frac{\varepsilon}{2}.$$

  - Given a partition P, let P' be the common refinement of P and Q.
  - Then

$$U(P',f) \le U(Q,f) < \int_a^b f dx + \frac{\varepsilon}{2}.$$

- Now in the difference of sums $U(P,f) - U(P',f)$
  - the contributions from all intervals that do not contain a point from Q are equal and cancel;
  - at most $m-1$ intervals contain a point in Q;
  - these each contribute at most $M|P|$ to each sum and hence at most $2M|P|$ to the difference;
  - so the difference is bounded by

$$U(P,f) - U(P',f) \le 2M(m-1)|P|.$$

- Let

$$\delta = \frac{\varepsilon}{4M(m-1)}.$$

- Then for a partition P with $|P| < \delta$

$$U(P,f) - U(P',f) \le 2M(m-1)|P| < \frac{\varepsilon}{2},$$

  and

$$U(P,f) = U(P',f) + U(P,f) - U(P',f) < \int_a^b f\,dx + \varepsilon.$$

- The lower bound is analogous.

Week 12

Monday, April 11, 2016

Weak Convergence of Probability Measures

Basic Definitions

- Suppose
  - X is a complete, separable metric space.
  - $\mathscr{X}$ is the Borel sigma-algebra on X.
  - $\mu, \mu_1, \mu_2, \ldots$ are probability measures on $\mathscr{X}$.
- μ_n is said to *converge weakly* to μ if

$$\int f d\mu_n \to \int f d\mu$$

  for all bounded, continuous $f : X \to \mathbb{R}$.
- A common notation is $\mu_n \Rightarrow \mu$.
- A set $A \in \mathscr{X}$ is a *μ-continuity set* if $\mu(\partial A) = \mu(\overline{A} \backslash A^\circ)$.
- Suppose $Y, Y_1, Y_2, \ldots$ are X-valued random variables with distributions μ, $\mu_1, \mu_2, \ldots$. Then Y_n *converges in distribution* to Y if $\mu_n \Rightarrow \mu$.

Equivalent Characterizations

- The following are equivalent:

  (a) $\mu_n \Rightarrow \mu$.

  (b) $\limsup_n \mu_n(F) \leq \mu(F)$ for all closed sets F.

  (c) $\liminf_n \mu_n(G) \geq \mu(G)$ for all open sets G.

  (d) $\lim_n \mu_n(A) = \mu(A)$ for all μ-continuity sets A.

- (a) implies (b):
  - Fix $\varepsilon > 0$.
  - For $y > 0$ define
$$F^y = \{x : d(x,F) < y\}$$
  - Since F is closed, $F = \bigcap_y F^y$, so we can find $\delta > 0$ such that
$$\mu(F^\delta) < \mu(F) + \varepsilon.$$
  - Let
$$\phi(t) = \begin{cases} 1 & \text{if } t \le 0 \\ 1-t & \text{if } 0 \le t \le 1 \\ 0 & \text{if } t \ge 1, \end{cases}$$
    and set
$$f(x) = \phi\left(\frac{1}{\delta} d(x,F)\right).$$
  - f is continuous and bounded by 1.
  - $f = 1$ on F and $f = 0$ on F^δ.
  - So
$$\mu_n(F) = \int_F f d\mu_n \le \int f d\mu_n$$
$$\int f d\mu = \int_{F^\delta} f d\mu \le \mu(F^\delta) \le \mu(F) + \varepsilon$$
  - Therefore
$$\limsup_n \mu_n(F) \le \lim_n \int f d\mu_n = \int f d\mu \le \mu(F) + \varepsilon.$$
  - Since this holds for all $\varepsilon > 0$ (b) follows.
- (b) implies (c): take complements.
- (b), (c) imply (d):
  - For any $A \in \mathscr{X}$
$$\mu(A^\circ) \le \liminf_n \mu_n(A^\circ) \le \limsup_n \mu_n(\overline{A}) \le \mu(\overline{A}).$$
  - If A is a μ-continuity set then $\mu(A^\circ) = \mu(\overline{A})$.

- (d) implies (a):
  - Let $f : X \to \mathbb{R}$ be a bounded continuous function.
  - Fix $\varepsilon > 0$.
  - Since f is bounded we can find $a, b \in \mathbb{R}$ such that $a < f(x) < b$ for all x.
  - There are at most countably many $y \in (a,b)$ such that

$$\mu(\{x : f(x) = y\} > 0.$$

  - So we can choose $y_0 = a < y_1 < y_2 < \ldots y_k = b$ such that
    - $y_i - y_{i-1} < \varepsilon$ for $i = 1, \ldots, k$;
    - $\mu(\{x : f(x) = y_i\} = 0$ for $i = 0, 1, \ldots, k$.
  - Let

$$\begin{aligned} A_i &= \{x : y_{i-1} < f(x) \le y_i\} \\ B_i &= \{x : y_{i-1} < f(x) < y_i\} \\ C_i &= \{x : y_{i-1} \le f(x) \le y_i\} \end{aligned}$$

  - Since f is continuous,
    - B_i is open and $B_i \subset A_i^\circ$;
    - C_i is closed and $\overline{A}_i \subset C_i$.
  - $\mu(B_i) = \mu(C_i)$, so

$$\mu(\partial A_i) \le \mu(C_i \backslash B_i) = \mu(C_i) - \mu(B_i) = 0$$

    and A_i is a μ-continuity set.

– Let

$$s(x) = \sum_{i=1}^{k} y_{i-1} 1_{A_i}(x)$$

– Then $s \le f \le s + \varepsilon$.

– So

$$\int s d\mu_n \le \int f d\mu_n \le \int s d\mu_n + \varepsilon$$
$$\int s d\mu \le \int f d\mu \le \int s d\mu + \varepsilon$$

– Since the A_i are μ-continuity sets

$$\int s d\mu_n = \sum_{i=1}^{k} y_{i-1} \mu_n(A_i) \to \sum_{i=1}^{k} y_{i-1} \mu(A_i) = \int s d\mu.$$

– So for any $\varepsilon > 0$

$$\limsup_{n\to\infty} \int f d\mu_n \le \int f d\mu + \varepsilon$$
$$\liminf_{n\to\infty} \int f d\mu_n \ge \int f d\mu - \varepsilon.$$

Some Properties

- Suppose every subsequence μ_{n_k} of μ_n contains a further subsequence $\mu_{n_{k_j}}$ such that $\mu_{n_{k_j}} \Rightarrow \mu$. Then $\mu_n \Rightarrow \mu$:
  - Suppose $\mu_n \not\Rightarrow \mu$.
  - Then for some bounded, continuous f and some $\varepsilon > 0$ there is a subsequence μ_{n_k} such that

$$\left| \int f d\mu_{n_k} - \int f d\mu \right| \geq \varepsilon.$$

  - No subsequence of μ_{n_k} can converge weakly to μ.
- There exists a metric ρ on the set of all probability distributions on $(X, \mathscr{X})$ such that $\mu_n \Rightarrow \mu$ if and only if $\rho(\mu_n, \mu) \to 0$.
- This result may fail if X is not separable.

Convergence in Distribution on the Real Line

- Suppose $X = \mathbb{R}$ and let

$$F_n(x) = \mu_n((-\infty, x])$$
$$F(x) = \mu((-\infty, x])$$

- Then $\mu_n \Rightarrow \mu$ if and only if $\lim_{n\to\infty} F_n(x) = F(x)$ for all continuity points x of F.
- Necessity:
  - The boundary of $(-\infty, x]$ is $\{x\}$
  - If F is continuous at x then $(-\infty, x]$ is a μ-continuity set.
  - So if $\mu_n \Rightarrow \mu$ then $F_n(x) \to F(x)$.

- Sufficiency:
  - Let $f : X \to \mathbb{R}$ be a bounded continuous function.
  - Fix $\varepsilon > 0$.
  - Let M be an upper bound on $|f(x)|$.
  - Select continuity points $a < b$ of F such that

$$F(a) < \varepsilon \qquad\qquad 1 - F(b) < \varepsilon.$$

  - Since f is continuous we can select $x_0 = a < x_1 < \cdots < x_k = b$ such that
    - $x_0, x_1, \ldots, x_k$ are continuity points of F;
    - $|f(x) - f(x_i)| < \varepsilon$ for all $x \in [x_{i-1}, x_i]$.
  - Let

$$S_k = \sum_{i=1}^{k} f(x_i)(F(x_i) - F(x_{i-1}))$$
$$S_{k,n} = \sum_{i=1}^{k} f(x_i)(F_n(x_i) - F_n(x_{i-1})).$$

  - Then $\lim_{n\to\infty} S_{k,n} = S_k$, and

$$\left| \int f d\mu - S_k \right| \le \varepsilon + MF(a) + M(1 - F(b)) \le (2M+1)\varepsilon$$
$$\left| \int f d\mu_n - S_{k,n} \right| \le \varepsilon + MF_n(a) + M(1 - F_n(b)) \le (2M+1)\varepsilon.$$

  - Therefore, for all $\varepsilon > 0$,

$$\liminf_{n\to\infty} \int f d\mu_n \ge S_k - (2M+1)\varepsilon \ge \int f d\mu - 2(2M+1)\varepsilon$$
$$\limsup_{n\to\infty} \int f d\mu_n \le S_k + (2M+1)\varepsilon \le \int f d\mu + 2(2M+1)\varepsilon.$$

  - So

$$\lim_{n\to\infty} \int f d\mu_n = \int f d\mu.$$

Wednesday, April 13, 2016

Skorokhod's Representation Theorem

- Suppose $(S, \mathscr{B}, P)$ is a probability space.
- $V : S \to X$ is a random element if $V^{-1}(A) \in \mathscr{B}$ for all $A \in \mathscr{X}$.
- The probability distribution of a random element V is the probability measure

$$\mu_V(A) = P(V^{-1}(A)) = P(\{s \in S : V(s) \in A\}).$$

- If V has probability distribution μ we write $V \sim \mu$.
- If $V, V_1, V_2, \ldots$ are random elements, we say $V_n \Rightarrow V$, or V_n converges to V in distribution, if

$$\mu_{V_n} \Rightarrow \mu_V.$$

- Suppose $\mu_n \Rightarrow \mu$. Then there exists a probability space $(S, \mathscr{B}, P)$ and random elements $V, V_1, V_2, \ldots$ such that $V \sim \mu$, $V_n \sim \mu_n$, and for all $s \in S$

$$\lim_{n\to\infty} V_n(s) = V(s).$$

- The probability space can always be taken to be $([0,1], \mathscr{R}, m)$ with $\mathscr{R}$ the Borel subsets of $[0,1]$ and m Lebesgue measure.
- This is known as *Skorokhod's representation theorem*.
- For the case $X = \mathbb{R}$, use the left-continuous inverse CDFs,

$$V(s) = F^{-1}(s)$$
$$V_n(s) = F_n^{-1}(s)$$

at points s where F^{-1} is continuous and set $V(s) = V_n(s) = 0$ otherwise.

Helly's Selection Principle

- Suppose F_n is a sequence of non-decreasing functions on $\mathbb{R}$ with $a \le F_n \le b$.
- Then there exists a subsequence F_{n_k} and a non-decreasing, right continuous function F such that

$$\lim_{k\to\infty} F_{n_k}(x) = F(x)$$

  for all continuity points x of F.

- Proof outline:
  - Let $r_1, r_2, \ldots$ be an enumeration of the rationals.
  - Since $[a,b]$ is compact, the sequence $F_n(r_1)$ has a subsequence $F_{n_{1,k}}(r_1)$ that converges to a limit $F_0(r_1)$.
  - From the corresponding subsequence of functions extract a second subsequence $F_{n_{2,1}}, F_{n_{2,2}}, \ldots$ such that $F_{n_{2,k}}(r_2)$ converges to a limit $F_0(r_2)$, and so on for $r_3, r_4, \ldots$.
  - Let $F_{n_k} = F_{n_{k,k}}$. Then $F_{n_k}(r) \to F_0(r)$ for every rational r.
  - F_0 is non-decreasing.
  - Define F as

$$F(x) = \inf\{F_0(r) : r \in \mathbb{Q} \text{ and } r \ge x\}.$$

  - Then F is non-decreasing and right continuous, and for and $x \in \mathbb{R}$ and any $s,t \in \mathbb{Q}$ with $s < x < t$

$$F_0(s) \le F(x) \le F_0(t).$$

    [If $x \in \mathbb{Q}$ we could have $F_0(x) \ne F(x)$ as F_0 need not be right continuous.]
  - Suppose x is a continuity point of F and fix $\varepsilon > 0$.
  - Since x is a continuity point of F, there exist a $\delta > 0$ such that

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

    when $|y - x| < \delta$.

- Choose $y, z \in \mathbb{R}$ and $s, t \in \mathbb{Q}$ such that $x - \delta < y < s < x$ and $x < t < z < x + \delta$.
- Then

$$\limsup_{k\to\infty} F_{n_k}(x) \le \lim_{k\to\infty} F_{n_k}(t) = F_0(t) \le F(z) \le F(x) + \varepsilon$$

  and

$$\liminf_{k\to\infty} F_{n_k}(x) \ge \lim_{\to\infty} F_{n_k}(s) = F_0(s) \ge F(y) \ge F(x) - \varepsilon$$

- So $F_{n_k}(x)$ converges to $F(x)$.

Tightness and Prokhorov's Theorem

- Even if

$$\lim_{x\to-\infty} F_n(x) = a \qquad\qquad \lim_{x\to\infty} F_n(x) = b$$

  for all n, all we ca say for F is

$$a \le \lim_{x\to-\infty} F(x) \qquad\qquad \lim_{x\to\infty} F_n(x) \le b.$$

- For distribution functions F_n this means that probability mass can "escape" to $\pm\infty$.

- To ensure this does not happen we need something like: for every $\varepsilon > 0$ there exists a positive number M such that

$$F_n(-M) \le \varepsilon \qquad\qquad F_n(M) \ge 1-\varepsilon.$$

- By increasing M if necessary we can take $\pm M$ to be continuity points of F.

- Then

$$F(-M) = \lim_{k\to\infty} F_{n_k}(-M) \le \varepsilon \qquad F(M) = \lim_{k\to\infty} F_{n_k}(M) \ge 1-\varepsilon.$$

- So F is a proper distribution function.

- A family $\mathscr{P}$ of probability distributions on a metric space is said to he *tight* if for every $\varepsilon > 0$ there exists a compact set K such that

$$\mu(K^c) < \varepsilon$$

  for all $\mu \in \mathscr{P}$

- Prokhorov's theorem: If the metric space is separable and $\mathscr{P}$ is tight, then there exists a sequence of probabilities $\mu_n \in \mathscr{P}$ and a probability μ such that $\mu_n \Rightarrow \mu$.

- If $\mu_n \Rightarrow \mu$ then the collection $\{\mu_n\}$ is tight.
- This is easy to verify if $X = \mathbb{R}^p$: take

  $$K = [-A, A]^p$$

  for A sufficiently large.
- More generally, if closed balls are compact, let

  $$K = \{x : d(x, x_0) \leq A\}$$

  for some x_0 and A sufficiently large.
- To handle more general separable metric spaces we need to use the fact a subset of a metric space is compact if and only if it is closed and totally bounded.

- Proof outline:
  - Fix $\varepsilon > 0$.
  - Let $x_1, x_2, \ldots$ be a countably dense subset of X.
  - For each k let
$$A_{i,k} = \{x : d(x, x_i) < 1/k\}.$$
  - Then $\bigcup_i A_{i,k} = X$.
  - Choose m_k such that
$$\mu\left(\bigcup_{i=1}^{m_k} A_{i,k}\right) \geq 1 - \frac{\varepsilon}{2^{k+1}}.$$
  - Then, since the union is an open set and $\mu_n \Rightarrow \mu$,
$$\liminf_{n\to\infty} \mu_n\left(\bigcup_{i=1}^{m_k} A_{i,k}\right) \geq 1 - \frac{\varepsilon}{2^{k+1}}$$
    and there is an integer N_k such that
$$\mu_n\left(\bigcup_{i=1}^{m_k} A_{i,k}\right) \geq 1 - \frac{\varepsilon}{2^k}$$
    for all $n > N_k$.
  - Increase m_k if necessary so
$$\mu_n\left(\bigcup_{i=1}^{m_k} A_{i,k}\right) \geq 1 - \frac{\varepsilon}{2^k}$$
    for $n = 1, \ldots, N_K$ as well.

– Let

$$B_k = \bigcup_{i=1}^{m_k} A_{i,k}.$$

and set $B = \bigcap_{k=1}^{\infty} B_k$.

– Then $x_1, \ldots, x_{m_k}$ is a $\frac{1}{k}$-net for B for every k, so B is totally bounded.

– $K = \overline{B}$ is compact,

$$K^c \subset B^c = \bigcup_{k=1}^{\infty} B_k^c,$$

and

$$\mu_n(K^c) \leq \mu_n(B^c) \leq \sum_{k=1}^{\infty} \mu_n(B_k^c) \leq \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon.$$

for every n.

Characteristic Functions and Weak Convergence

- Suppose μ, μ_n have characteristic functions φ, φ_n.
- If $\mu_n \Rightarrow \mu$ then $\varphi_n(t) \to \varphi(t)$ for all t.
  - $f(x) = \sin(tx)$ and $g(x) = \cos(tx)$ are bounded and continuous for all t.
- Suppose μ_n is tight and $\varphi_n(t) \to \varphi(t)$ for all t.
  - Let μ_{n_k} be a subsequence of μ_n.
  - This subsequence is tight as well.
  - So by Prokhorov's theorem there is a further subsequence $\mu_{n_{k_i}}$ that converges weakly to some probability μ' with characteristic function φ'.
  - Then $\varphi_{n_{k_i}}(t) \to \varphi'(t)$ for all t.
  - But we also have $\varphi_{n_{k_i}}(t) \to \varphi(t)$ for all t.
  - So $\varphi'(t) = \varphi(t)$ for all t.
  - By uniqueness theorem for characteristic functions this means $\mu' = \mu$.
  - So every subsequence of μ_n has a further subsequence that converges weakly to μ.
  - Therefore $\mu_n \Rightarrow \mu$.
- It turns out that the fact that φ is continuous at the origin is enough to ensure tightness.

Characteristic Functions and Tail Probabilities

- Suppose the probability μ on $\mathbb{R}$ has characteristic function φ.
- The behavior of φ near the origin can be used to bound the probability of $[-M,M]^c$ for large M:
  - $$\begin{aligned}\frac{1}{a}\int_{-a}^{a} 1-\Re\varphi(t)dt &= \frac{1}{a}\int_{-a}^{a}\int 1-\cos(tx)\mu(dx)dt\\ &= \frac{1}{a}\int\int_{-a}^{a} 1-\cos(tx)dt\mu(dx)\\ &= 2\frac{1}{a}\int a-\frac{\sin(ax)}{x}\mu(dx)\\ &= 2\int 1-\frac{\sin(|ax|)}{|ax|}\mu(dx)\end{aligned}$$

    with $\sin(0)/0 = 1$.
  - For $u \geq 1$
    $$u\sin(1) \geq \sin(u)$$
    and therefore
    $$1-\frac{\sin(u)}{u} \geq 1-\sin(1)$$
  - So
    $$\begin{aligned}2\int 1-\frac{\sin(|ax|)}{|ax|}\mu(dx) &\geq 2\int 1_{[1,\infty)}(|ax|)\left(1-\frac{\sin(|ax|)}{|ax|}\right)\mu(dx)\\ &\geq 2(1-\sin(1))\int 1_{(1,\infty)}(|ax|)\mu(dx)\\ &= 2(1-\sin(1))\mu([-1/a,1/a]^c).\end{aligned}$$
  - So for any $M > 0$
    $$\mu([-M,M]^c) \leq \frac{M}{2(1-\sin(1))}\int_{-1/M}^{1/M} 1-\Re\varphi(t)dt$$

Friday, April 15, 2016

Class Canceled

Week 13

Monday, April 18, 2016

Weak Convergence of Probability Measures

Characteristic Functions and Weak Convergence

- Suppose μ_n is a sequence of probabilities on $\mathbb{R}$ with characteristic functions φ_n.
- Suppose $\varphi_n(t) \to \varphi(t)$ for all t and φ is continuous at the origin. Then
  - φ is the characteristic function of a probability μ;
  - $\mu_n \Rightarrow \mu$.
- Proof outline:
  - Continuity at the origin of φ implies that that the sequence μ_n is tight.
  - By Prokhorov's theorem there exists a subsequence μ_{n_k} and a probability μ such that $\mu_{n_k} \Rightarrow \mu$.
  - $\varphi = \lim \varphi_{n_k}$ is therefore the characteristic function of μ.
  - Every subsequence has a further subsequence that converges.
  - The limit of the further subsequence has characteristic function φ, so it is μ.
  - So $\mu_n \Rightarrow \mu$.

- To show tightness:
  - $\varphi_n(0) = 1$ for all t so $\varphi(0) = 1$.
  - Fix $\varepsilon > 0$. Choose M such that

$$|1 - \Re\varphi(t)| < \frac{\varepsilon}{2}(1 - \sin(1))$$

    for $|t| < 1/M$.
  - Then by the dominated convergence theorem

$$\begin{aligned}\limsup_{n\to\infty} \mu_n([-M,M]^c) &\le \lim_{n\to\infty} \frac{M}{2(1-\sin(1))} \int_{-1/M}^{1/M} 1 - \Re\varphi_n(t)dt \\ &= \frac{M}{2(1-\sin(1))} \int_{-1/M}^{1/M} 1 - \Re\varphi(t)dt \\ &\le \frac{\varepsilon}{2}.\end{aligned}$$

  - So there exists an N such that $\mu_n([-M,M]^c) < \varepsilon$ for all $n \ge N$.
  - Increase M if necessary so that $\mu_n([-M,M]^c) < \varepsilon$ for $n = 1,\ldots,N$ as well.

Multivariate Characteristic Functions

- If Y takes values in $\mathbb{R}^p$ and $t \in \mathbb{R}^p$ then the multivariate characteristic function of Y is

$$\varphi(t) = E[\exp\{it^T Y\}] = E[\exp\{i \sum_{k=1}^{p} t_k Y_k\}].$$

- Uniqueness and convergence results carry over to p dimensions.
- A consequence is the *Cramer-Wold* device:
  - Y and Z have the same distribution if and only it $a^T Y$ and $a^T Z$ have the same distribution for all $a \in \mathbb{R}^p$.
  - $Y_n \Rightarrow Y$ if and only if $a^T Y_n \Rightarrow a^T Y$ for all $a \in \mathbb{R}^p$.
- These follow from the observation that

$$\varphi_{a^T Y}(s) = \varphi_Y(sa).$$

Convergence of Moment Generating Functions

- In general weak convergence does not imply convergence of moment generating functions.
- A uniform boundedness condition is needed to obtain this.
- Suppose $\mathscr{P}$ is a collection of probabilities on $(\mathbb{R}, \mathscr{R})$ and that for some $t > 0$ and some $K < \infty$
$$\int e^{sx} \nu(dx) \le K$$
for all $\nu \in \mathscr{P}$ and all $s \in [-t, t]$.
- Then $\mathscr{P}$ contains a sequence μ_n with moment generating functions $M_n(t)$ that converges weakly to a probability μ with moment generating function M, and
$$\lim_{n\to\infty} M_n(s) = M(s)$$
for all $s \in (-t, t)$.
- In particular, if μ_n have moment generating functions M_n such that
$$M_n(s) \le K < \infty$$
for some $t > 0$ and all $s \in [-t, t]$, and if $\mu_n \Rightarrow \mu$, then the MGF M of μ is finite on $(-t, t)$ and
$$\lim_{n\to\infty} M_n(s) = M(s)$$
for $s \in (-t, t)$.
- Key points in the proof:
  - Uniform boundedness of the moment generating functions provides a uniform bound on tail probabilities.
  - This establishes tightness.
  - The bound is strong enough to uniformly bound tails of the MGF integrals.

- Proof outline:
  - For any $\lambda \in \mathscr{P}$ and any $A > 0$ by Markov's inequality

$$\lambda((A,\infty)) = \lambda(\{x : e^{tx} > e^{tA}\}) \le e^{-tA} \int e^{tx} \lambda(dx) \le e^{-tA} K.$$

    and

$$\lambda((-\infty,-A)) = \lambda(\{x : e^{-tx} > e^{tA}\}) \le e^{-tA} \int e^{-tx} \lambda(dx) \le e^{-tA} K.$$

  - This implies that $\mathscr{P}$ is tight, so by Prokhorov's theorem there exists a sequence $\mu_n \in \mathscr{P}$ and a probability μ such that $\mu_n \Rightarrow \mu$.
  - Since the sets (A,∞) and $(-\infty,-A)$ are open and $\mu_n \Rightarrow \mu$ we have

$$\mu((A,\infty)) \le \liminf_{n\to\infty} \mu_n((A,\infty)) \le e^{-tA} K$$
$$\mu((\infty,-A)) \le \liminf_{n\to\infty} \mu_n((\infty,-A)) \le e^{-tA} K$$

  - Suppose $s \in [0,t)$. For any probability λ and any A

$$\int e^{sx} \lambda(dx) = \int e^{s\min(x,A)} \lambda(dx) + \int (e^{sx} - e^{sA}) 1_{(A,\infty)}(x) \lambda(dx).$$

  - Since the integrand is non-negative, for any probability λ

$$\begin{aligned}
\int (e^{sx} - e^{sA}) 1_{(A,\infty)}(x) \lambda(dx) &= \int_0^\infty \lambda(\{x : e^{sx} - e^{sA} > u\}) du \\
&= \int_0^\infty \lambda(\{x : e^{sx} > u + e^{sA}\}) du \\
&= \int_{e^{sA}}^\infty \lambda(\{x : e^{sx} > v\}) dv \\
&= \int_{e^{sA}}^\infty \lambda((\log(v)/s,\infty)) dv
\end{aligned}$$

– If λ satisfies the inequality

$$\lambda((A,\infty)) = e^{-tA}K.$$

for all A, and the μ_n and μ do, then

$$\begin{aligned}\int_{e^{sA}}^{\infty} \lambda((\log(v)/s,\infty))dv &\le K\int_{e^{sA}}^{\infty} v^{-t/s}dv \\ &= \frac{K}{t/s-1}e^{(1-t/s)sA} = \frac{sK}{t-s}e^{(s-t)A}.\end{aligned}$$

– This shows that the moment generating function of μ is finite at s.

– Fix $\varepsilon > 0$ and chose A such that

$$\frac{sK}{t-s}e^{(s-t)A} < \frac{\varepsilon}{2}.$$

– The function $e^{s\min(x,A)}$ is bounded and continuous, so

$$\lim_{n\to\infty}\int e^{s\min(x,A)}\mu_n(dx) = \int e^{s\min(x,A)}\mu(dx).$$

– Therefore

$$\limsup_{n\to\infty} |M_n(s) - M(s)| \le \varepsilon$$

– Since this holds for all $\varepsilon > 0$ we have

$$\lim_{n\to\infty} M_n(s) = M(s)$$

for all $s \in [0,t)$.

– The case $s \in (-1,0]$ is analogous.

- Suppose μ_n are probabilities with moment generating functions M_n, μ is a probability with moment generating function M, and for some $t > 0$

$$\lim_{n\to\infty} M_n(s) = M(s) < \infty$$

for all $s \in [-t,t]$. Then $\mu_n \Rightarrow \mu$.

  - Since $M_n(t)$ and $M_n(-t)$ converge to finite limits there is a constant $K < \infty$ such that

$$M_n(t) \le K \qquad\qquad M_n(-t) \le K$$

for all n.

  - Since moment generating functions are convex, this means $M_n(s) \le K$ for all n and all $s \in [-t,t]$.
  - Let μ_{n_k} be a subsequence of μ_n. Then there exists a further subsequence $\mu_{n_{k_j}}$ that converges weakly to a limit μ' and for which the moment generating functions converge to the MGF of μ'.
  - Then μ' and μ have identical finite moment generating functions on a neighborhood of the origin, and therefore $\mu' = \mu$.
  - Thus every subsequence of μ_n has a further subsequence that converges weakly to μ, so $\mu_n \Rightarrow \mu$.

Wednesday, April 20, 2016

Weak Convergence of Probability Measures

Product Spaces

- Suppose $(X',d'),(X'',d'')$ are separable metric spaces with Borel sigma algebras $\mathscr{X}',\mathscr{X}''$.
- Let $X = X' \times X''$ and

$$d((x',x''),(y',y'')) = \max(d'(x',y'),d''(x'',y'')).$$

- Then (X,d) is separable with Borel sigma algebra $\mathscr{X}$, and

$$\mathscr{X} = \mathscr{X}' \times \mathscr{X}''.$$

- Suppose μ,μ_n are probabilities on $\mathscr{X}'$ and ν,ν_n are probabilities on $\mathscr{X}''$.
- Then $\mu_n \times \nu_n \Rightarrow \mu \times \nu$ if and only if $\mu_n \Rightarrow \mu$ and $\nu_n \Rightarrow \nu$.
- For random elements, if U,V are independent and U_n,V_n are independent for each n, then $(U_n,V_n) \xrightarrow{\mathscr{D}} (U,V)$ if and only if $U_n \xrightarrow{\mathscr{D}} U$ and $V_n \xrightarrow{\mathscr{D}} V$.

Continuous Mappings

- Let Y be a separable metric space and suppose $h : X \to Y$ is measurable.
- For a measure μ on the Borel sets of X define
$$\mu h^{-1}(A) = \mu(h^{-1}(A)) = \mu(\{x : h(x) \in A\})$$
for all Borel sets of Y.
- If $f : Y \to \mathbb{R}$ is measurable and $\int f d(\mu h^{-1})$ exists, then
$$\int f d(\mu h^{-1}) = \int f(y)\mu h^{-1}(dy) = \int f(h(x))\mu(dx)$$
- Suppose μ, μ_n are probabilities on the Borel sets of X and $\mu_n \Rightarrow \mu$.
- If h is continuous, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$:
  - If f is a bounded continuous function on Y then
$$\begin{aligned}\int f(y)\mu_n h^{-1}(dy) &= \int f(h(x))\mu_n(dx) \\ &\to \int f(h(x))\mu(dx) = \int f(y)\mu h^{-1}(dy)\end{aligned}$$
since $g(x) = f(h(x))$ is a bounded, continuous function on X.
  - For random elements $U_n \xrightarrow{\mathscr{D}} U$ and $V_n = h(U_n), V = h(U)$,
$$\begin{aligned}E[f(V_n)] &= E[f(h(U_n))] = E[g(U_n)] \\ &\to E[g(U)] = E[f(h(U))] = E[f(V)].\end{aligned}$$
- More generally, if D_h is the set of points in X where h is not continuous and $\mu(D_h) = 0$, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$:
  - If F is a closed subset of X, then
$$\overline{(h^{-1}(F))} \subset D_h \cup h^{-1}(F).$$
  - Therefore
$$\begin{aligned}\limsup_{n\to\infty} \mu_n h^{-1}(F) &= \limsup_{n\to\infty} \mu_n(h^{-1}(F)) \\ &\le \limsup_{n\to\infty} \mu_n(\overline{(h^{-1}(F))}) \\ &\le \mu(\overline{(h^{-1}(F))}) \le \mu(D_h) + \mu(h^{-1}(F)) \\ &= \mu h^{-1}(F).\end{aligned}$$

Simple Random Walks

- Suppose $Y_1, Y_2, \ldots$ are independent with

$$P(Y_i = 1) = P(Y_i = -1) = \frac{1}{2}.$$

- Then $E[Y_i] = 0$ and $\text{Var}(Y_i) = 1$.
- Let $S_0 = 0$ and for $n = 1, 2, \ldots$

$$S_n = \sum_{i=1}^{n} Y_i.$$

- S_n is called a *random walk* on the line.
- Some simple simulations:

```
rw <- function(n) {
    y <- 2 * rbinom(n, 1, 0.5) - 1
    c(0, cumsum(y))
}
plot(rw(10), type = "l")
plot(rw(100), type = "l")
```

- By the central limit theorem,

$$\frac{1}{\sqrt{n}} S_n \Rightarrow Z \sim \text{N}(0,1)$$

- Plotting $\frac{1}{\sqrt{n}} S_k$ against $\frac{k}{n}$ for $k = 0, 1, \ldots, n$ stabilizes the picture:

```
plot((0 : 10) / 10, rw(10) / sqrt(10), type = "l", ylim = c(-2.5,2.5))
lines((0 : 100) / 100, rw(100) / sqrt(100), col = "red")
lines((0 : 10000) / 10000, rw(10000) / sqrt(10000), col = "blue")
```

- The linear interpolation in the plots corresponds to

$$X_n(t) = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} + \frac{(nt - \lfloor nt \rfloor)}{\sqrt{n}} Y_{\lfloor nt \rfloor + 1}$$

- A useful variation is

$$V_k = S_k - \frac{k}{n} S_n$$

- Then $V_0 = V_n = 0$.

```
rwb <- function(n) {
  s <- rw(n)
  s - s[n + 1] * ((0 : n) / n)
}
plot((0 : 10) / 10, rwb(10) / sqrt(10), type = "l", ylim = c(-2.5,2.5))
lines((0 : 100) / 100, rwb(100) / sqrt(100), col = "red")
lines((0 : 10000) / 10000, rwb(10000) / sqrt(10000), col = "blue")
```

- The linear interpolation in the plots corresponds to

$$V_n(t) = X_n(t) - tX_n(1).$$

Empirical Distribution Functions

- Suppose $Y_1, Y_2, \ldots, Y_n$ are independent with continuous CDF F.
- The empirical CDF is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(Y_i)$$

- A useful measure of how well a distribution F fits a sample is

$$D_n = \sup_x |F_n(x) - F(x)|.$$

- If F is the correct distribution, then the distribution of D_n does not depend on F, as long as F is continuous.
- So just consider the case where F corresponds to the uniform distribution on $[0,1]$.
- For $Y_i \sim \text{Uniform}[0,1]$ and for large n the sample paths of

$$\sqrt{n}(F_n(t) - F(t))$$

behave like the sample paths $V_n(t)$.

Distributions on $C_{[0,1]}$

- The linearly interpolated and scaled random walks are random elements in $C_{[0,1]}$.
- $C_{[0,1]}$ with the distance

$$d(x,y) = \max_t |x(t) - y(t)|$$

  is a complete separable metric space.
- To show that the probability distributions μ_n of X_n converge weakly to some distribution μ we need some preliminary results.
- A mapping $h : C_{[0,1]} \to \mathbb{R}^k$ defined by

$$h(x) = h_{t_1,\ldots,t_k}(x) = (x(t_1), \ldots, x(t_k))$$

  for some $t_1, \ldots, t_k$ is called a *finite dimensional projection*.
- The probability μh^{-1} on the Borel sets of $\mathbb{R}^k$ is a *finite dimensional distribution*.
- Sets of the form

$$h^{-1}(A) = \{x \in C_{[0,1]} : (x(t_1), \ldots, x(t_k)) \in A\}$$

  for A a Borel set in $\mathbb{R}^k$ are called *cylinder sets*.
- The finite dimensional projections are continuous.
- So if $\mu_n \Rightarrow \mu$ then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ for any finite dimensional projection h; i.e. the finite dimensional distributions converge.

- The collection of cylinder sets is a π-system.
- The Borel sigma algebra of $C_{[0,1]}$ is generated by the cylinder sets:
  - Every closed sphere in $C_{[0,1]}$ can be expressed as

$$\{x : d(x,x_0) \le \varepsilon\} = \bigcap_{r \in \mathbb{Q}} \{x : |x(r) - x_0(r)| \le \varepsilon\}$$

  - Every open set can be expressed as a countable union of closed spheres.
- If two probabilities μ and μ' agree on all cylinder sets, i.e. have the same finite dimensional distributions, then they agree on all Borel sets of $C_{[0,1]}$.
- Suppose μ_n is tight and that for every finite dimensional projection $h = h_{t_1,\ldots,t_k}$ there is a distribution ν_h on the Borel sets of $\mathbb{R}^k$ such that $\mu h^{-1} \Rightarrow \nu_h$.
- Then there exists a probability μ such that
  - $\mu h^{-1} = \nu_h$;
  - $\mu_n \Rightarrow \mu$.

Finite Dimensional Limits for the Random Walk

- For any $t \in [0,1]$

$$\left| X_n(t) - X_n\left(\frac{\lfloor nt \rfloor}{n}\right) \right| = \frac{(nt - \lfloor nt \rfloor)}{\sqrt{n}} |Y_{\lfloor nt \rfloor + 1}| \le \frac{1}{\sqrt{n}} \to 0.$$

- Furthermore, by the central limit theorem,

$$X_n\left(\frac{\lfloor nt \rfloor}{n}\right) = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} = \sqrt{\frac{\lfloor nt \rfloor}{n}} \frac{1}{\sqrt{\lfloor nt \rfloor}} S_{\lfloor nt \rfloor} \xrightarrow{\mathscr{D}} Z\sqrt{t}$$

  where $Z \sim \mathrm{N}(0,1)$.

- For any s,t with $t \in [0,1]$ and $t+s \in [0,1]$ the random variables $S_{\lfloor nt \rfloor}$ and $S_{\lfloor n(t+s) \rfloor} - S_{\lfloor nt \rfloor}$ are independent, and

$$\frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} \xrightarrow{\mathscr{D}} Z_1\sqrt{t}$$
$$\frac{1}{\sqrt{n}} \left(S_{\lfloor n(t+s) \rfloor} - S_{\lfloor nt \rfloor} \right) \xrightarrow{\mathscr{D}} Z_2\sqrt{s}$$

  with $Z_1, Z_2 \sim \mathrm{N}(0,1)$.

- So

$$(X_n(t), X_n(t+s) - X_n(t)) \xrightarrow{\mathscr{D}} (Z_1\sqrt{t}, Z_2\sqrt{s})$$

  with Z_1, Z_2 independent.

- Therefore

$$(X_n(t), X_n(t+s)) \xrightarrow{\mathscr{D}} (Z_1\sqrt{t}, Z_1\sqrt{t} + Z_2\sqrt{s}),$$

  and

$$\mathrm{Cov}(Z_1\sqrt{t}, Z_1\sqrt{t} + Z_2\sqrt{s}) = t\mathrm{Var}(Z_1) = t.$$

- Put another way, for $t_1, t_2 \in [0,1]$

$$(X_n(t_1), X_n(t_2)) \xrightarrow{\mathscr{D}} (B(t_1), B(t_2))$$

where $(B(t_1), B(t_2))$ has a bivariate normal distribution with

$$E[B(t_i)] = 0 \qquad \text{Var}(B(t_i)) = t_i$$
$$\text{Cov}(B(t_1), B(t_2)) = \min(t_1, t_2).$$

- This extends to all finite dimensional distributions.

Friday, April 22, 2016

Weak Convergence of Probability Measures

Tightness in $C[0,1]$

- To show that
  - that there exists a probability μ that is the distribution of a random element B with these finite dimensional distributions;
  - that $X_n \xrightarrow{\mathscr{D}} B$

  we need to show that the distributions of X_n are tight.
- A set in $C_{[0,1]}$ is compact if and only if and only if it is
  - closed
  - equicontinuous.
  - uniformly bounded
- A slight variation: A set in $C_{[0,1]}$ is compact if it
  - it is closed;
  - it is equicontinuous;
  - the set of values at one point in $[0,1]$ is bounded.
- The *modulus of continuity* of a function $x \in C_{[0,1]}$ is defined for $0 < \delta \leq 1$ as
$$w_x(\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)|.$$
- A set $A \in C_{[0,1]}$ is equicontinuous if for every $\varepsilon > 0$ there exists a $\delta \in (0,1]$ such that
$$w_x(\delta) < \varepsilon$$
for all $x \in A$.

- A collection $\mathscr{P}$ of probabilities on $C_{[0,1]}$ is tight if and only if these two conditions hold:

  (i) For every $\eta > 0$ there exists a number a such that

  $$\mu(\{x : |x(0)| > a\}) < \eta$$

  for all $\mu \in \mathscr{P}$.

  (ii) For every $\varepsilon > 0$ and $\eta > 0$ there exists a $\delta \in (0,1)$ such that

  $$\mu(\{x : w_x(\delta) \geq \varepsilon\}) \leq \eta$$

  for all $\mu \in \mathscr{P}$.

- Suppose $\mathscr{P}$ is tight and fix $\eta > 0$ and $\varepsilon > 0$.
  - Then there is a compact $K \subset C_{[0,1]}$ such that $\mu(K) > 1 - \eta$ for all $u \in \mathscr{P}$.
  - K is uniformly bounded and equicontinuous, so there exists and a and a δ such that conditions (i) and (ii) hold for all $x \in K$.

- Suppose conditions (i) and (ii) hold and fix $\eta > 0$.
  - Choose a such that

  $$\mu(\{x : |x(0)| \leq a\}) = \mu(A_0) \geq 1 - \frac{\eta}{2}$$

  for all $\mu \in \mathscr{P}$.
  - For each $k = 1, 2, \ldots$ chose δ_k such that

  $$\mu\left(\left\{x : w_x(\delta_k) < \frac{1}{k}\right\}\right) = \mu(A_k) \geq 1 - \frac{\eta}{2^{k+1}}.$$

  for all $\mu \in \mathscr{P}$.
  - Let

  $$A = \bigcap_{i=0}^{\infty} A_i,$$

  and let $K = \overline{A}$.
  - Then K is compact, and for all $\mu \in \mathscr{P}$

  $$\mu(K) \geq \mu(A) = 1 - \mu(A^c) \geq 1 - \sum_{i=0}^{\infty} \mu(A_i^c) \geq 1 - \eta \sum_{i=0}^{\infty} \frac{1}{2^{i+1}} = 1 - \eta.$$

- For a sequence μ_n of probabilities, (ii) can be replaced by

  (ii') For every $\varepsilon > 0$ and $\eta > 0$ there exists a $\delta \in (0,1)$ and an integer N such that
  $$\mu_n(\{x : w_x(\delta) \geq \varepsilon\}) \leq \eta$$
  for all $n \geq N$.

- This follows since
  - every finite set of probabilities is tight;
  - therefore, once δ and N have been found, δ can be increased until
    $$\mu_n(\{x : w_x(\delta) \geq \varepsilon\}) \leq \eta$$
    for $n = 1, \ldots, N-1$ as well.

- This can be useful when the bound is easier to establish for large n.

- Suppose $\delta \in (0,1)$ and $t_0 = 0 < t_1 < \ldots t_m = 1$ is a partition of $[0,1]$ with
  - $t_i - t_{i-1} \geq \delta$ for $i = 1, \ldots, m-1$.
  - The last partition interval might be smaller that δ
- If $0 \leq u \leq v \leq 1$ and $v - u < \delta$, then either
  - $t_{i-1} \leq u < v \leq t_i$ for some $i = 1, \ldots, m$, or
  - $t_{i-1} \leq u \leq t_i \leq v \leq t_{i+1}$ for some $i = 1, \ldots, m-1$.
- For $x \in C_{[0,1]}$ let

$$M_x(\delta) = \max_{i=0,\ldots,m-1} \sup_{t_i \leq s \leq t_{i+1}} |x(s) - x(t_i)|.$$

- If u, v are in the same partition interval $[t_i, t_{i+1}]$, then

$$\begin{aligned} |x(u) - x(v)| &\leq |x(u) - x(t_i)| + |x(v) - x(t_i)| \\ &\leq 2M_x(\delta). \end{aligned}$$

- If $u \in [t_i, t_{i+1}]$ and $v \in [t_{i+1}, t_{i+2}]$, then

$$\begin{aligned} |x(u) - x(v)| &\leq |x(u) - x(t_i)| + |x(t_i) - x(t_{i+1})| + |x(v) - x(t_{i+1})| \\ &\leq 3M_x(\delta). \end{aligned}$$

- Therefore $w_x(\delta) \leq 3M_x(\delta)$.
- This implies that

$$\begin{aligned} \mu(\{x : w_x(\delta) \geq a\}) &\leq \mu\left(\left\{x : M_x(\delta) \geq \frac{a}{3}\right\}\right) \\ &\leq \sum_{i=0}^{m-1} \mu\left(\left\{x : \sup_{t_i \leq s \leq t_{i+1}} |x(s) - x(t_i)| \geq \frac{a}{3}\right\}\right) \end{aligned}$$

Tightness for the Random Walk

- For the random walk, take μ_n to the distribution of X_n and fix $\eta > 0$ and $\varepsilon > 0$.
- Condition (i) is satisfied since $X_n(0) = 0$ for all n.
- For a $\delta \in (0,1)$ use a partition with $t_i = \frac{i}{n}\lceil n\delta \rceil$ and

$$m = \lceil \frac{1}{\delta} \rceil \le \frac{1}{\delta} + 1 \le \frac{2}{\delta}.$$

- Then, because of the piece-wise linear form of X_n, for $i \le m-2$

$$P\left(\sup_{t_i \le s \le t_{i+1}} |X_n(s) - X_n(t_i)| \ge \frac{\varepsilon}{3}\right) = P\left(\frac{1}{\sqrt{n}} \max_{j \le \lceil n\delta \rceil} |S_j| \ge \frac{\varepsilon}{3}\right)$$

- For the possibly shorter final interval this becomes an inequality,

$$P\left(\sup_{t_{m-1} \le s \le t_m} |X_n(s) - X_n(t_i)| \ge \frac{\varepsilon}{3}\right) \le P\left(\frac{1}{\sqrt{n}} \max_{j \le \lceil n\delta \rceil} |S_j| \ge \frac{\varepsilon}{3}\right)$$

- Thus

$$P(w_{X_n}(\delta) \ge \varepsilon) \le mP\left(\frac{1}{\sqrt{n}} \max_{j \le \lceil n\delta \rceil} |S_j| \ge \frac{\varepsilon}{3}\right) \le \frac{2}{\delta} P\left(\max_{j \le \lceil n\delta \rceil} |S_j| \ge \sqrt{n}\frac{\varepsilon}{3}\right)$$

- Because of the symmetry of the random walk,

$$\begin{aligned} P\left(\max_{j\le k}|S_j|\ge a\right) &\le P\left(\max_{j\le k}S_j\ge a\right)+P\left(\min_{j\le k}S_j\le -a\right)\\ &= 2P\left(\max_{j\le k}S_j\ge a\right)\end{aligned}$$

- Using the *reflection principle*,

$$\begin{aligned} P\left(\max_{i\le k}S_j\ge a\right) &= P(S_k=a)+2P(S_k>a)\\ &\le 2P(S_k\ge a).\end{aligned}$$

- Using Markov's inequality, for $a>0$

$$\begin{aligned} 2P(S_k\ge a) = P(|S_k|\ge a) &\le \frac{E[S_k^4]}{a^4}\\ &= \frac{k(3k-2)}{a^4}\le\frac{3k^2}{a^4}\end{aligned}$$

- So

$$P\left(\max_{j\le k}|S_j|\ge a\right)\le\frac{6k^2}{a^4}.$$

- And therefore

$$P\left(w_{X_n}(\delta)\ge\varepsilon\right)\le\frac{2}{\delta}\frac{6\lceil n\delta\rceil^2}{(\sqrt{n}\varepsilon/3)^4}=\frac{972\lceil n\delta\rceil^2}{\delta n^2\varepsilon^4}$$

- For a fixed δ,

$$\lim_{n\to\infty}\frac{972\lceil n\delta\rceil^2}{\delta n^2\varepsilon^4}=\frac{972\delta}{\varepsilon^4}.$$

- So for the given $\eta>0$ and $\varepsilon>0$ chose δ such that

$$\frac{972\delta}{\varepsilon^4}=\frac{\eta}{2}$$

and N such that

$$\frac{972\lceil n\delta\rceil^2}{\delta n^2\varepsilon^4}<\eta$$

for all $n\ge N$.

- So the sequence μ_n of distributions of X_n is tight.

Week 14

Monday, April 25, 2016

Weak Convergence of Probability Measures

Brownian Motion

- The probability μ on $C_{[0,1]}$ that is the limit of the distributions μ_n is called *Wiener measure*.
- A random element B with distribution μ is called *Brownian Motion*.
- Brownian motion
  - has $B(0) = 0$;
  - has continuous sample paths;
  - has finite dimensional distributions that are normal with mean zero and
    $$\mathrm{Cov}(B(t_1), B(t_2)) = \min(t_1, t_2).$$
- This implies that for $t_1 < t_2 < \cdots < t_k$ the *increments*
  $$B(t_1), B(t_2) - B(t_1), \ldots, B(t_k) - B(t_{k-1})$$
  are independent.

- With probability one Brownian motion paths
  - are nowhere differentiable
  - have infinite length
- If $0 \le a < b \le 1$ and

$$B_1(t) = \frac{B(tb+(1-t)a) - B(a)}{\sqrt{b-a}}$$

  the B_1 has the same distribution as B. This is sometimes called *self-similarity*.
- The process

$$B_0(t) = B(t) - tB(1)$$

  is called a *Brownian bridge* or *tied-down Brownian motion*.
- The Brownian bridge can also be obtained as the limit, as $\varepsilon \to 0$, of Brownian motion conditioned on the event $B(1) \in [-\varepsilon, \varepsilon]$.
- Brownian motion can be extended to $t \in [0, \infty)$ by piecing together independent Brownian motions on $[0,1]$.
- A Brownian motion on $[0,\infty)$ can also be obtained from a Brownian bridge as

$$B(t) = (1+t)B_0\left(\frac{t}{1+t}\right).$$

Limiting Distribution of the Kolmogorov-Smirnov Statistic

- Suppose $Y_1, \ldots, Y_n$ are independent with common continuous CDF F.
- Let F_n be the empirical CDF for the sample.
- Then

$$D_n^+ = \sqrt{n} \sup_x (F_n(x) - F(x)) \xrightarrow{\mathscr{D}} \sup_{0 \le t \le 1} B_0(t) = D_0^+$$
$$D_n^- = \sqrt{n} \inf_x (F_n(x) - F(x)) \xrightarrow{\mathscr{D}} \inf_{0 \le t \le 1} B_0(t) = D_0^-$$
$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)| \xrightarrow{\mathscr{D}} \sup_{0 \le t \le 1} |B_0(t)| = D_0$$

- The distribution of D_0^+ is easy to obtain using the reflection principle.
  - For $a > 0$ and $D^+ = \sup_{0 \le t \le 1} B(t)$

$$\begin{aligned} P(D_0^+ > a) &= P(D^+ > a | B(1) = 0) \\ &= \lim_{\varepsilon \downarrow 0} P(D^+ > a | B(1) \in (-\varepsilon, \varepsilon)) \\ &= \lim_{\varepsilon \downarrow 0} \frac{P(D^+ > a, B(1) \in (-\varepsilon, \varepsilon))}{P(B(1) \in (-\varepsilon, \varepsilon))}. \end{aligned}$$

  - By the reflection principle, assuming $\varepsilon < a$,

$$\begin{aligned} P(D^+ > a, B(1) \in (-\varepsilon, \varepsilon)) &= P(D^+ > a, B(1) \in (2a - \varepsilon, 2a + \varepsilon)) \\ &= P(B(1) \in (2a - \varepsilon, 2a + \varepsilon)). \end{aligned}$$

  - Thus

$$\begin{aligned} P\left(D_0^+ > a\right) &= \lim_{\varepsilon \downarrow 0} \frac{P(B(1) \in (2a - \varepsilon, 2a + \varepsilon))}{P(B(1) \in (-\varepsilon, \varepsilon))} \\ &= \frac{\varphi(2a)}{\varphi(0)} \\ &= e^{-2a^2}. \end{aligned}$$

- For the two-sided case consider, for $a, b > 0$,

$$P(D^+ > a \text{ or } D^- < -b) = P(B_0 \text{ hits either } a \text{ or } -b).$$

- Let

$$A_1 = \{B_0 \text{ hits } a\} \qquad B_1 = \{B_0 \text{ hits } -b\}.$$

- Then

$$P(B_0 \text{ hits either } a \text{ or } -b) = P(A_1 \cup B_1) = P(A_1) + P(B_1) - P(A_1 \cap B_1)$$

- Let

$$A_2 = \{B_0 \text{ hits } a \text{ then } -b\} \qquad B_2 = \{B_0 \text{ hits } -b \text{ then } a\}.$$

- Then

$$P(A_1 \cap B_1) = P(A_2 \cup B_2) = P(A_2) + P(B_2) - P(A_2 \cap B_2)$$

- So if $P(A_m) \to 0$ and $P(B_m) \to 0$ then

$$P(B_0 \text{ hits either } a \text{ or } -b) = \sum_{m=1}^{\infty} (-1)^{m+1} (P(A_m) + P(B_m)).$$

- For A_2

$$\begin{aligned} P(A_2) &= P(B \text{ hits } a \text{ then } -b | B(1) = 0) \\ &= \lim_{\varepsilon \downarrow 0} \frac{P(B \text{ hits } a \text{ then } -b \text{ and } B(1) \in (-\varepsilon, \varepsilon))}{P(B(1) \in (-\varepsilon, \varepsilon))} \end{aligned}$$

- Using the reflection principle twice gives

$$\begin{aligned} &P(B \text{ hits } a \text{ then } -b \text{ and } B(1) \in (-\varepsilon, \varepsilon)) \\ &\quad = P(B \text{ hits } 2a+b \text{ and } B(1) \in (2a-\varepsilon, 2a+\varepsilon)) \\ &\quad = P(B(1) \in (2a+2b-\varepsilon, 2a+2b\varepsilon)) \\ &\quad \approx 2\varepsilon\varphi(2(a+b)) \end{aligned}$$

- So

$$P(A_2) = \frac{\varphi(2(a+b))}{\varphi(0)} = e^{-2(a+b)^2}.$$

- Using the reflection principle m times for $m \geq 2$ gives

$$P(A_m) = \frac{\varphi(m(a+b))}{\varphi(0)} = e^{-m^2(a+b)^2/2}.$$

- Since a, b appear symmetrically, $P(B_m) = P(A_m)$ for $m \geq 2$.
- Thus $P(A_m) \to 0$ and $P(B_m) \to 0$ and

$$P(B_0 \text{ hits either } a \text{ or } -b) = e^{-2a^2} + e^{-2b^2} + 2\sum_{m=2}^{\infty} (-1)^{m+1} e^{-m^2(a+b)^2/2}.$$

- Taking $b = a$ then gives

$$\begin{aligned} P(D \geq a) &= 2e^{-2a^2} + 2\sum_{m=2}^{\infty} (-1)^{m+1} e^{-2m^2a^2} \\ &= 2\sum_{m=1}^{\infty} (-1)^{m+1} e^{-2m^2a^2} \end{aligned}$$

- Except for small a this series converges very quickly.
- Some plotting and simulation code is available on line.

Wednesday, April 27, 2016

Linear Transformations

- Some basic notions:
  - vector space or linear space
  - (linear) span of a set of vectors
  - (linear) independence
  - dimension
  - basis
- A mapping A of a vector space X into a vector space Y is a *linear transformation* if [R Def 9.4]

$$A(x_1+x_2) = A(x_1)+A(x_2) \qquad A(cx) = cA(x)$$

  for all $x, x_1, x_2 \in X$ and all scalars c.
- $A(x)$ is often written as Ax.
- If $\{x_1, \ldots, x_n\}$ is a basis of X then the action of A is completely determined by $Ax_1, \ldots, Ax_n$:
  - Each $x \in X$ has a unique representation of the form

$$x = \sum_{i=1}^{n} c_i x_i.$$

  - In terms of the c_i,

$$Ax = \sum_{i=1}^{n} c_i A x_i.$$

- A linear transformation of X into X is called a *linear operator*.
- If A is a linear operator that

  (i) is one-to-one

  (ii) maps X onto X

  then A is *invertible* with inverse A^{-1}.
- If A is invertible, then A^{-1} is a linear operator.

- A linear operator A on a finite dimensional vector space is one-to-one if and only if the range of A is all of X. [R Thm 9.5]
- Let [R Ref 9.6]
  - $L(X,Y)$ be the set of all linear transformations from X into Y;
  - $L(X)$ be the set of all linear operators on X.
- $L(X,Y)$ and $L(X)$ are vector spaces.
- If X,Y,Z are vector spaces, $A \in L(X,Y), B \in L(Y,Z)$, define

$$(BA)(x) = B(A(x)).$$

- Then the *product* BA is in $L(X,Z)$.
- The norm of a linear transformation $A \in L(X,Y)$ is defined as

$$\|A\| = \sup_{\|x\| \le 1} \|Ax\|.$$

- If A has norm $\|A\|$, then

$$\|Ax\| \le \|A\|\|x\|$$

  for all $x \in X$.
- A linear transformation A is said to be *bounded* if $\|A\| < \infty$.
- If X is finite dimensional, then every linear transformation from X into Y is bounded.
- $\|A\|$ is a norm on the set of bounded linear operators, and $d(A,B) = \|A-B\|$ is a metric on this set.
- If $A \in L(X,Y)$ and $A \in L(Y,Z)$, then

$$\|BA\| \le \|B\|\|A\|.$$

- Let Ω be the set of all invertible linear operators on $\mathbb{R}^n$. [R Thm 9.8]
  - If $A \in \Omega$, $B \in L(\mathbb{R}^n)$, and
    $$\|B-A\|\|A^{-1}\| < 1$$
    then $B \in \Omega$; this implies that Ω is an open subset of $L(\mathbb{R}^n)$.
  - The mapping $A \to A^{-1}$ is continuous on Ω.
- Suppose $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$ are bases for X,Y.
  - Every $A \in L(X,Y)$ determines, and is determined by, numbers a_{ij} such that
    $$Ax_j = \sum_{i=1}^{m} a_{ij} y_i.$$
  - The matrix
    $$[A] = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$
    will differ for different bases.
- If
  - S is a metric space;
  - $a_{ij}(p) : S \to \mathbb{R}$ are continuous;
  - A_p is the linear transformation determined by the $a_{ij}(p)$;

  then the mapping $p \to A_p$ is continuous from S into $L(\mathbb{R}^n,\mathbb{R}^m)$.

Differentiation

- Suppose X, Y are complete normed linear spaces (i.e. Banach spaces).
- Suppose $E \subset X$ is open.
- A function $f : E \to Y$ is differentiable at $x \in E$ if there exists a bounded linear transformation A such that [R Def 9.11]

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0.$$

- If such a linear transformation exists then it is unique: [R Thm 9.12]
  - Suppose A_1 and A_2 both satisfy this definition.
  - Let $B = A_1 - A_2$. Then as $h \to 0$

$$\frac{\|Bh\|}{\|h\|} \le \frac{\|f(x+h) - f(x) - A_1 h\|}{\|h\|} + \frac{\|f(x+h) - f(x) - A_2 h\|}{\|h\|} \to 0.$$

  - For a fixed $v \neq 0$ and scalars t let $h = tv$.
  - Then

$$0 = \lim_{t \to 0} \frac{\|B(tv)\|}{\|tv\|} = \frac{\|Bv\|}{\|v\|}.$$

    So $Bv = 0$ for all v and thus $B = 0$ and $A_1 = A_2$.

- The bounded linear transformation A is the derivative, or differential,

$$f'(x) = A \qquad\qquad \delta f(x) = A.$$

- If X is finite dimensional then all linear transformations in $L(X, Y)$ are bounded.
- A differentiable function is also said to be *Fréchet differentiable*, in particular if X is infinite-dimensional.
- The differential/derivative is also called the Fréchet differential/derivative.
- The derivative definition can also be written as [R Rem 9.13]

$$f(x+h) = f(x) + Ah + o(h).$$

  as $h \to 0$.

- For a fixed x the derivative $f'(x)$ is a function mapping X to Y.
- If f is differentiable for all $x \in E$ then $f'(x)$ is a function mapping E into the bounded linear transformations in $L(X,Y)$.
- If $f(x) = Ax$ for some $A \in L(X,Y)$ and A is bounded, then f is differentiable with derivative $f'(x) = A$. [R Ex 9.14]
- **Chain rule:** Suppose [R Thm 9.15]
  - f maps a neighborhood of $x_0 \in X$ into Y and is differentiable at x_0 with $f'(x_0) = A$;
  - g maps a neighborhood if $y_0 = f(x_0)$ into Z and is differentiable at y_0 with $g'(y_0) = B$;
  - $F(x) = g(f(x))$.

  Then F is differentiable at x_0 and

$$F'(x_0) = g'(f(x_0))f'(x_0) = BA.$$

- Proof outline:
  - For sufficiently small $h \in X, k \in Y$

$$u(h) = f(x_0+h) - f(x_0) - Ah$$
$$v(k) = g(y_0+k) - g(y_0) - Bk.$$

  - Then

$$\|u(h)\| \le \varepsilon(h)\|h\| \qquad \|v(k)\| \le \eta(k)\|k\|$$

    with $\varepsilon(h) \to 0$ as $h \to 0$ and $\eta(k) \to 0$ as $k \to 0$.
  - Let $k = f(x_0+h) - f(x_0)$.
  - Then $\|k\| = \|Ah + u(h)\| \le [\|A\| + \varepsilon(h)]\|h\|$, and

$$\begin{aligned} F(x_0+h) - F(x_0) - BAh &= g(y_0+k) - g(y_0) - Bk + B(k - Ah) \\ &= v(k) + B(k - Ah) \\ &= v(k) + Bu(h) \end{aligned}$$

  - Therefore as $h \to 0$

$$\frac{\|F(x_0+h) - F(x_0) - BAh\|}{\|h\|} \le \eta(k)[\|A\| + \varepsilon(h)] + \|B\|\varepsilon(h) \to 0.$$

- For a vector $v \in X$ the *directional derivative* of f at x in the direction v is defined as [R p 217]

$$(D_v f)(x) \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t}$$

if the limit exists.

- The directional derivative is also called the *Gateaux derivative*, in particular in infinite dimensional settings.

- If f is differentiable at x, then $(D_v f)(x) = f'(x)v$.

- If is possible for f to be not differentiable at x but for the directional derivatives at x to exist for all v.

- If $X = \mathbb{R}^n$ and $\{e_1, \ldots, e_n\}$ is the standard basis for $\mathbb{R}^n$, then we can define

$$(D_j f)(x) = (D_{e_j} f)(x).$$

- If $v = \sum_{j=1}^n v_j e_j$ then

$$(D_v f)(x) = \sum_{j=1}^n (D_j f)(x) v_j$$

- If $Y = \mathbb{R}^m$ with standard basis $\{u_1 \ldots, u_m\}$ then the components $f_1, \ldots, f_m$ of f are defined by

$$f(x) = \sum_{i=1}^m f_i(x) u_i.$$

- The *partial derivatives* of f are

$$\left(\frac{\partial f_i}{\partial x_j}\right)(x) = (D_j f_i)(x) = \lim_{t \to 0} \frac{f_i(x+te_j) - f_i(x)}{t}.$$

- If f is defined on a neighborhood of x and is differentiable at x then the partial derivatives exist and

$$f'(x)e_j = \sum_{i=1}^m (D_j f_i)(x) u_i.$$

Friday, April 29, 2016

Differentiation

- Suppose
  - $E \subset X$ is an open, convex set;
  - f is differentiable at every point in E;
  - $\|f'(x)\| \le M < \infty$ for all $x \in E$.

  Then
  $$\|f(x) - f(y)\| \le M\|x - y\|$$
  for all $x, y \in E$.
- Proof outline:
  - Fix $x, y \in E$, and for $t \in [0,1]$ define
    $$\gamma(t) = (1-t)x + ty$$
    $$g(t) = f(\gamma(t)).$$
  - Then $g'(t) = f'(\gamma(t))\gamma'(t) = f'(\gamma(t))(x-y)$, and
    $$\|g'(t)\| \le \|f'(\gamma(t))\|\|x-y\| \le M\|x-y\|.$$
  - By Theorem 5.19 in Rudin, this implies that
    $$\|f(x) - f(y)\| = \|g(1) - g(0)\| \le M\|x-y\|.$$
  - Theorem 5.19 is valid for any Banach space Y.

- A mapping f from an open set $E \subset$X into Y is *continuously differentiable* on E if [R Def 9.20]
  - f is differentiable at every $x \in E$;
  - the mapping $f'(x) : E \to L(X,Y)$ is continuous.
- The collection of continuously differentiable mappings from E to $\mathbb{R}^m$ is denoted by $\mathscr{C}'(E)$.
- A mapping f from an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^m$ is in $\mathscr{C}'(E)$ if and only if the partial derivatives $D_j f_i$ exist and are continuous on E. [R Thm 9.21]

- Suppose $f \in \mathscr{C}'(E)$:
  - Then $(D_j f_i)(x) = \langle (D_j f)(x), u_i \rangle = \langle f'(x) e_j, u_i \rangle$.
  - So any $x, y \in E$

$$\begin{aligned} |(D_j f_i)(y) - (D_j f_i)(x)| &= |\langle (f'(x) - f'(y)) e_j, u_i \rangle| \\ &\leq \|(f'(x) - f'(y)) e_j\| \\ &\leq \|f'(x) - f'(y)\|. \end{aligned}$$

  - Thus $D_j f_i$ is continuous on E.

- Suppose $D_j f_i$ exist and are continuous on E for all i, j.
  - If $f_1, \ldots, f_m$ are continuously differentiable, then f is continuously differentiable, so it is enough to consider $m = 1$.
  - Fix $x \in E$ and $\varepsilon > 0$.
  - Choose $\delta > 0$ such that $\|y - x\| < \delta$ implies that $y \in E$ and for $j = 1, \ldots, n$
$$|(D_j f)(y) - (D_j f)(x)| < \frac{\varepsilon}{n}$$
  - Suppose $\|h\| < \delta$, $h = \sum_{j=1}^n h_j e_j$, set $v_0 = 0$, and for $k = 1, \ldots, n$ set
$$v_k = h_1 e_i + \cdots + h_k e_k.$$
  - Then
$$f(x+h) - f(x) \sum_{k=1}^n (f(x+v_k) - f(x+v_{k-1})).$$
  - Since $\{y : \|y - x\| < \delta\}$ is convex, every point
$$y_{t,k} = x + v_{k-1} + t h_k e_k$$
for $t \in [0,1]$ on the segment between $x + v_{k-1}$ and $x + v_k$ satisfies $\|y_{t,k} - x\| < \delta$.
  - By the mean value theorem there exist $\theta_j \in [0,1]$ such that
$$f(x+v_k) - f(x+v_{k-1}) = h_j (D_j f)(x + v_{k-1} + \theta_j h_j e_j). = h_j (D_j f)(y_{\theta_j, j}).$$
  - Therefore
$$\left| f(x+h) - f(x) - \sum_{j=1}^n h_j (D_j f)(x) \right| = \left| \sum_{j=1}^n h_j ((D_j f)(y_{\theta_j, j}) - (D_j f)(x)) \right|$$
$$\leq \sum_{j=1}^n |h_j| \frac{\varepsilon}{n} \leq \|h\| \varepsilon.$$

Examples

- Suppose $X = [0,\infty) \times [0,\infty)$ and

$$f(x,y) = \begin{cases} \frac{xy}{x+y} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0). \end{cases}$$

  - f is continuous at $(0,0)$ since

$$0 \leq f(x,y) \leq \min(x,y).$$

  - The partial derivatives are

$$(D_1 f)(x,y) = \begin{cases} \left(\frac{y}{x+y}\right)^2 & y \neq 0 \\ 0 & y = 0 \end{cases} \quad (D_2 f)(x,y) = \begin{cases} \left(\frac{x}{x+y}\right)^2 & x \neq 0 \\ 0 & x = 0. \end{cases}$$

  - These are not continuous at $(0,0)$ since for $y \neq 0$

$$\lim_{x \to 0} (D_1 f)(x,y) = 1.$$

  - The directional derivative in direction $v = (v_1, v_2)$ at $(0,0)$ in

$$(D_v f)(0,0) = \lim_{t \to 0} \frac{f(tv_1, tv_2)}{t} = \frac{v_1 v_2}{v_1 + v_2}.$$

  - This is not linear in v, to f is not differentiable at $(0,0)$.
  - Some plots:

```
x <- seq(0, 1, len = 101)
y <- x
v <- outer(x, y, function(x, y) x * y / (x + y))
persp(x, y, v)
contour(x, y, v)
contour(x, y, v, nlevels = 30)
```

- Suppose $X = C_{[0,1]}$ and

$$f(x) = \int_0^1 (x(s))^2 ds.$$

  - The directional derivative in direction v is

$$\begin{aligned}(D_v f)(x) &= \frac{d}{dt} f(x+tv)|_{t=0} \\ &= \frac{d}{dt} \int_0^1 (x(s)+tv(s))^2 ds \Big|_{t=0} \\ &= \int_0^1 \frac{d}{dt}(x(s)+tv(s))^2 ds \Big|_{t=0} \\ &= 2\int_0^1 x(s)v(s)ds\end{aligned}$$

  - The linear operator A defined by $Ah = 2\int_0^1 x(s)h(s)ds$ is bounded:

$$|Ah| = 2\left|\int_0^1 x(s)h(s)ds\right| \le 2\|x\|\|h\|.$$

  - By the mean value theorem, for each $s \in [0,1]$ there is a $u(s)$ between $x(s)$ and $x(s)+h(s)$ such that

$$(x(s)+h(s))^2 - (x(s))^2 = 2(x(s)+u(s))h(s).$$

  - So

$$\begin{aligned}f(x+h) - f(x) &= \int_0^1 (x(s)+h(s))^2 - (x(s))^2 ds \\ &= \int_0^1 2(x(s)+u(s))h(s)ds.\end{aligned}$$

  - Therefore

$$|f(x+h) - f(x) - Ah| = \left|2\int_0^1 u(s)h(s)ds\right| \le 2\|h\|^2.$$

  - So A is the Fréchet derivative of f.

Week 15

Monday, May 2, 2016

Inverse Function Theorem

- Suppose [R Thm 9.24]
  - f is a $\mathscr{C}'$ mapping of an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^n$;
  - $f'(a)$ is invertible for some $a \in E$, and let $b = f(a)$.

  Then

  (a) there exists open sets $U, V \subset \mathbb{R}^n$, $a \in U$, $b \in V$, f is one-to-one on U with $f(U) = V$;

  (b) if $g : V \to U$ is the inverse of f, then $g \in \mathscr{C}'(V)$, and for $y \in V$

$$g'(y) = (f'(g(y)))^{-1}.$$

- Proof outline for (a):
  - Let $A = f'(a)$ and
$$\lambda = \frac{1}{2\|A^{-1}\|}$$
  - Let $U \subset E$ be an open ball with center a such that for $x \in U$
$$\|f'(x) - A\| < \lambda.$$
  - For $y \in \mathbb{R}^n$ define $\varphi_y : E \to \mathbb{R}^n$ as
$$\varphi_y(x) = x + A^{-1}(y - f(x)).$$
  - Then $f(x) = y$ if and only if $\varphi_y(x) = x$.
  - φ_y is differentiable with
$$\varphi_y'(x) = I - A^{-1}f'(x) = A^{-1}(A - f'(x)),$$
    so
$$\|\varphi_y'(x)\| \le \|A^{-1}\|\|A - f'(x)\| < \frac{1}{2}.$$
  - Therefore if $x_1, x_2 \in U$ then
$$\|\varphi_y(x_1) - \varphi_y(x_2)\| \le \frac{1}{2}\|x_1 - x_2\|.$$
  - So φ_y can have at most one fixed point in U and thus f is one-to one on U.
  - Let $V = f(U)$. We need to show that V is open.
  - Pick $y_0 \in V$; then $y_0 = f(x_0)$ for a unique $x_0 \in U$.
  - Let B be an open ball with center x_0 and radius r small enough that $\overline{B} \subset U$.
  - Suppose y satisfies $\|y - y_0\| < \lambda r$.
  - Then
$$\|\varphi_y(x_0) - x_0\| = \|A^{-1}(y - f(x_0))\| \le \|A^{-1}\|\|y - y_0\| < \frac{r}{2}.$$

- Therefore if $x \in \overline{B}$ then

$$\|\varphi_y(x) - x_0\| \le \|\varphi_y(x) - \varphi_y(x_0)\| + \|\varphi_y(x_0) - x_0\| < \frac{1}{2}\|x - x_0\| + \frac{r}{2} \le r,$$

  and so $\varphi_y(x) \in \overline{B}$.
- So φ_y is a contraction of $\overline{B}$ into $\overline{B}$.
- This implies that φ_y has a fixed point $x \in \overline{B}$, and this fixed point satisfies $f(x) = y$.
- So $y \in f(\overline{B}) \subset f(U) = V$, and thus V is open.

- Proof outline for (b):
  - Fix $y \in V$ and suppose k satisfies $y+k \in V$.
  - Then there exist unique $x, x+h \in U$ such that

$$y = f(x) \qquad\qquad y+k = f(x+h).$$

  - Then

$$\begin{aligned}\varphi_y(x+h) &= x+h+A^{-1}(y-f(x+h)) \\ &= \varphi_y(x)+h+A^{-1}(f(x)-f(x+h)) \\ &= \varphi_y(x)+h-A^{-1}k.\end{aligned}$$

  - So

$$\|h-A^{-1}k\| = \|\varphi_y(x+h)-\varphi_y(x)\| \le \frac{1}{2}\|h\|,$$

    and hence $\|A^{-1}k\| \ge \frac{1}{2}\|h\|$ and

$$\|h\| \le 2\|A^{-1}\|\|k\| = \frac{1}{\lambda}\|k\|.$$

    So $h \to 0$ as $k \to 0$.
  - By the definition of λ and U, $f'(x)$ has an inverse, T, and

$$\begin{aligned}g(y+k)-g(y)-Tk &= h-Tk \\ &= Tf'(x)h-T(f(x+h)-f(x)) \\ &= -T[f(x+h)-f(x)-f'(x)h].\end{aligned}$$

  - This implies

$$\begin{aligned}\frac{\|g(y+k)-g(y)-Tk\|}{\|k\|} &\le \|T\|\frac{\|f(x+h)-f(x)-f'(x)h\|}{\|k\|} \\ &\le \frac{\|T\|}{\lambda}\frac{\|f(x+h)-f(x)-f'(x)h\|}{\|h\|} \\ &\to 0\end{aligned}$$

    as $k \to 0$.

- So g is differentiable as y and

$$g'(y) = T = (f'(x))^{-1} = (f'(g(y)))^{-1}.$$

- Finally,
  - g is continuous since it is differentiable;
  - f' is a continuous mapping of U into Ω, the set invertible linear operators;
  - computing the inverse operator is a continuous mapping from Ω onto Ω.

  So $g'(y)$ is continuous on V.

Method of Moments

- The method of moments assumes that moments can be expressed as nice functions of the parameters,

$$\begin{aligned}\mu_1 &= f_1(\theta_1, \ldots, \theta_k)\\ &\vdots\\ \mu_k &= f_k(\theta_1, \ldots, \theta_k),\end{aligned}$$

  or

$$\mu = f(\theta).$$

- Sample moments $m_1, \ldots, m_k$ are then computed, and parameters are estimated by solving

$$m = f(\theta)$$

- If g is the inverse of f, then $\theta = g(\mu)$, and

$$\widehat{\theta} = g(m).$$

- If f is continuously differentiable near θ, then g is continuously differentiable near μ, and for large samples

$$\widehat{\theta} = g(m) \approx g(\mu) + g'(\mu)(m - \mu).$$

- Limiting distributions are typically obtained using a central limit theorem for the sample moments m and approximations

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta) &= \sqrt{n}(\widehat{\theta} - g(\mu))\\ &\approx g'(\mu)[\sqrt{n}(m - \mu)]\\ &\approx g'(m)[\sqrt{n}(m - \mu)].\end{aligned}$$

Implicit Function Theorem

- Suppose $f(a,b) = 0$. The implicit function theorem addresses when

$$f(x,y) = 0$$

can be solved uniquely for x when y is near b.

- If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, write [R Not 9.26]

$$(x,y) = (x_1, \ldots, x_n, y_1, \ldots, y_m) \in \mathbb{R}^{n+m}.$$

- If $A \in L(\mathbb{R}^{n+m}, \mathbb{R}^n)$ define

$$A_x h = A(h,0) \qquad\qquad A_y k = A(0,k)$$

for $h \in \mathbb{R}^n, k \in \mathbb{R}^m$.

- Then $A_x \in L(\mathbb{R}^n)$, $A_y \in L(\mathbb{R}^m, \mathbb{R}^n)$, and

$$A(h,k) = A_x h + A_y k.$$

- Linear version: If $A \in L(\mathbb{R}^{m+n}, \mathbb{R}^n)$ and A_x is invertible, then for every [R Thm 9.27] $k \in \mathbb{R}^m$ there exists a unique h given by

$$h = -A_x^{-1} A_y k$$

such that $A(h.k) = 0$.

- General version: [R Thm 9.28]
  - Let f be a $\mathscr{C}'$ mapping of an open set $E \subset \mathbb{R}^{n+m}$ into $\mathbb{R}^n$ such that $f(a,b) = 0$ for some $(a,b) \in E$.
  - Let $A = f'(a,b)$ and assume A_x is invertible.
  - Then there exist open sets $U \subset \mathbb{R}^{n+m}$, $W \subset \mathbb{R}^m$, such that
    - $(a,b) \in U$ and $b \in W$;
    - for every $y \in W$ there exists a unique $x = g(y)$ such that

$$(x,y) \in U \qquad\qquad f(x,y) = f(g(y),y) = 0;$$

    - g is a $\mathscr{C}'$ mapping of W into $\mathbb{R}^n$;
    - $g(b) = a$ and

$$g'(b) = -A_x^{-1} A_y.$$

- This can be obtained from the inverse function theorem:
  - Let $F(x,y) = (f(x,y),y)$.
  - F is a $\mathscr{C}'$ mapping of E into $\mathbb{R}^{n+m}$ with

$$F'(a,b)(h,k) = (A(h,k),k),$$

    which is invertible.
  - So F has an inverse G near $(0,b)$ with

$$G(u,v) = (G_1(u,v), G_2(u,v)) = (G_1(u,v),v).$$

    Let $g(y) = G_1(0,y)$.

- By writing $y = f(x)$ as

$$f(x) - y = 0$$

  the inverse function theorem could also be obtained from the implicit function theorem.

Wednesday, May 4, 2016

Implicitly Defined Statistical Functionals

- Suppose $X_1, \ldots, X_n$ is a sample from a population with distribution F.
- A number of estimates are defined as the solution to equations of the form
$$0 = \int \psi(x,t) F_n(dx)$$
with F_n the empirical distribution.
- Examples include
  - maximum likelihood estimators;
  - M-estimators.
- The population characteristic estimated generally satisfies
$$0 = \int \psi(x,t) F(dx).$$
- Let
$$f(t,F) = \int \psi(x,t) F(dx).$$
- Then $T(F)$ is the implicit function defined by the equation
$$0 = f(t,F).$$
- The derivative of $T(F)$ with respect to F is
$$D_F T = -(D_t f)^{-1}(D_F f).$$

- For well-behaved ψ

$$D_t f = D_t \int \psi(x,t) F(dx) = \int D_t \psi(x,t) F(dx)$$

and for $H = G - F$

$$\begin{aligned}(D_F f)H &= \int \psi(x,t) H(dx) \\ &= \int \psi(x,t) G(dx) - \int \psi(x,t) F(dx) \\ &= \int \psi(x,t) G(dx).\end{aligned}$$

- So

$$(D_F T)(G-F) = -\left(\int D_t \psi(x, T(F)) F(dx)\right)^{-1} \int \psi(x, T(F)) G(dx)$$

- For $G = \delta_x$ this produces the *influence function*

$$IC_T(x,F) = -\left(\int D_t \psi(x, T(F)) F(dx)\right)^{-1} \psi(x, T(T))$$

- The sample mean corresponds to $\psi(t,x) = x - t$, so

$$IC_{\overline{X}_n}(x,F) = x - \mu_F.$$

This is unbounded.

- For the Huber M-estimator the influence function is bounded.

- For $G = F_n$,

$$\begin{aligned}\sqrt{n}(T(F_n) - T(F)) &\approx -\sqrt{n}\left(\int D_t\psi(x,T(F))F(dx)\right)^{-1}\int \psi(x,T(F))F_n(dx)\\ &= -\left(\int D_t\psi(x,T(F))F(dx)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,T(F))\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_T(X_i,F)\end{aligned}$$

- Viewing the value of ψ as a column vector, the central limit theorem gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i,T(F)) \xrightarrow{\mathscr{D}} Y \sim \mathrm{N}(0,\Sigma)$$

with

$$\Sigma = E[\psi(X,T(F))\psi(X,T(F))^T]$$

- This then gives

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathscr{D}} V \sim \mathrm{N}(0,A)$$

with

$$A = B\Sigma B^T$$

and

$$B = \left(\int D_t\psi(x,T(F))F(dx)\right)^{-1}.$$

- This is sometimes called the *sandwich formula*.

A Generalized Inverse Function Theorem

- Suppose X, Y are Banach spaces.
- Let f be a continuously differentiable mapping of an open set $E \subset X$ into Y.
- A point $x_0 \in E$ is a *regular point* if $f'(x_0)$ maps X onto Y.
- Suppose x_0 is a regular point of f. Then there exists an open set $V \subset Y$ and a constant K such that
  - $y_0 = f(x_0) \in V$;
  - for any $y \in V$ there exists an $x \in X$ such that $f(x) = y$ and

$$\|x - x_0\| \leq K\|y - y_0\|.$$

- This can be proved using some of the same ideas used to prove the regular inverse function theorem.

- Suppose
  - f is real-valued;
  - f and H are continuously differentiable on an open set E;
  - $x_0 \in E$ is a regular point of H;
  - $H(x_0) = 0$;
  - x_0 is a local extremum among all points $x \in E$ satisfying $H(x) = 0$.

  Then $f'(x_0)h = 0$ for all h satisfying $H'(x_0)h = 0$.
  - Suppose x_0 is a local minimum.
  - Let $F(x) = (f(x), H(x))$.
  - F is continuously differentiable.
  - If $f'(x_0)h \neq 0$ and $H'(x_0)h = 0$ for some h then $F'(x_0)$ is onto.
  - By the generalized inverse function theorem, given any $\varepsilon > 0$ there exists an x and a $\delta > 0$ with

    $$\|x - x_0\| < \varepsilon$$

    and

    $$F(x) = (f(x_0) - \delta, 0).$$

  - This contradicts the assumption that x_0 is a local minimum.

- Suppose $X = \mathbb{R}^n$ and $H : \mathbb{R}^n \to \mathbb{R}^m$.
- Suppose x_0 is a local extremum of f subject to $H(x) = 0$.
- View $H'(x_0)$ as an $m \times n$ matrix of partial derivatives.
- $f'(x_0)$ is orthogonal to the null space of $H'(x_0)$
- This implies that $f'(x_0)$ is in the row space of $H'(x_0)$.
- So there exists a vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ such that

$$f'(x_0) + \lambda^T H'(x_0) = 0.$$

- Together with $H(x_0) = 0$ this provides a set of $n+m$ equations in $n+m$ unknowns.
- The λ_i are called *Lagrange multipliers*.
- Versions of this result for more general spaces are also available.

Friday, May 4, 2016

Derivatives of Higher Order

- Suppose f is real valued defined on an open set $E \subset \mathbb{R}^n$.
- Suppose f has partial derivatives $D_1 f, \ldots, D_n f$ on E.
- If the partial derivatives are differentiable then the *second order partial derivatives* are defined as

$$D_{ij} f = D_i D_j f$$

  for $i, j = 1, \ldots . n$.

- If the $D_{ij} f$ are all defined and continuous on E then f is said to be $\mathscr{C}''$ on E.
- If f maps E into $\mathbb{R}^m$ then it is $\mathscr{C}''$ if each component is.
- For simplicity just consider $m = 1$ and $n = 2$.

A Mean Value Theorem

- Suppose $D_1 f$ and $D_{21} f$ exist at every point of E. [R Thm 9.40]
- Let $Q \subset E$ be a closed rectangle with
  - sides parallel to the axes
  - opposite corners (a,b) and $(a+h,b+k)$, for $h \neq 0$ and $k \neq 0$.
- Let

$$\Delta(f,Q) = f(a+h,b+k) - f(a+h,b) - f(a,b+k) + f(a,b).$$

- Then there is a point (x,y) in the interior of Q such that

$$\Delta(f,Q) = hk(D_{21} f)(x,y).$$

- Proof:
  - Let $u(t) = f(t,b+k) - f(t,b)$.
  - Then $\Delta(f,Q) = u(a+h) - u(a)$.
  - By the mean value theorem there is an x between a and $a+h$ such that

$$u(a+h) - u(a) = hu'(x) = h((D_1 f)(x,b+k) - (D_1 f)(x,b)).$$

  - Using the mean value theorem again, there is a y between b and $b+k$ such that

$$(D_1 f)(x,b+k) - (D_1 f)(x,b) = k(D_{21} f)(x,y)$$

Derivative Order

- Suppose
  - $D_1 f, D_{21} f, D_2 f$ exist at every point of E;
  - $D_{21} f$ is continuous at some $(a,b) \in E$.

  Then $D_{12} f$ exists at (a,b) and $(D_{12} f)(a,b) = (D_{21} f)(a,b)$.
- Proof:
  - Let $A = (D_{21} f)(a,b)$.
  - Fix $\varepsilon > 0$.
  - For h,k in the definition of Q sufficiently small we have $Q \subset E$.
  - Since $D_{21} f$ is continuous at (a,b), for all sufficiently small h,k and for all $(x,y) \in Q$
$$|A - (D_{21} f)(x,y)| < \varepsilon.$$
  - Thus for all sufficiently small non-zero h,k
$$\left| \frac{\Delta(f,Q)}{hk} - A \right| < \varepsilon.$$
  - Using the fact that $D_2 f$ exists everywhere in E, fix h and let $k \to 0$ to get
$$\left| \frac{(D_2 f)(a+h,b) - (D_2 f)(a,b)}{h} - A \right| \leq \varepsilon.$$
    for all sufficiently small $h \neq 0$.
  - Since $\varepsilon > 0$ was arbitrary, this shows that $D_{12} f$ exists at (a,b) and
$$(D_{12} f)(a,b) = A.$$
- If $f \in \mathscr{C}''(E)$ then $D_{12} f = D_{21} f$ on E.

Differentiation of Integrals

- Suppose $\varphi(x,t) : [a,b] \times [c,d] \to \mathbb{R}$.
- Suppose α is non-decreasing on $[a,b]$.
- A useful notation to view φ for a fixed t as a function of x:

$$\varphi^t(x) = \varphi(x,t).$$

- Suppose $\varphi^t \in \mathscr{R}(\alpha)$ on $[a,b]$ for every $t \in [c,d]$.
- Suppose $(D_2\varphi)(x,t)$ exists for all $x \in (a,b)$ and $t \in (c,d)$.
- For $t \in [c,d]$ define

$$f(t) = \int_a^b \varphi(x,t)d\alpha(x) = \int_a^b \varphi^t d\alpha.$$

- Suppose $s \in (c,d)$ and for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$|(D_2\varphi)(x,t) - (D_2\varphi)(x,s)| < \varepsilon$$

  for all $x \in [a,b]$ whenever $|t-s| < \delta$.

- Then
  - $(D_2\varphi)^s \in \mathscr{R}(\alpha)$ on $[a,b]$;
  - $f'(s)$ exists, and

$$f'(s) = \int_a^b (D_2\varphi)^s d\alpha = \int_a^b (D_2\varphi)(x,s)d\alpha(x).$$

- Proof outline:
  - Consider the difference quotiont

$$\psi(x,t) = \frac{\varphi(x,t) - \varphi(x,s)}{t-s}.$$

    for $t \neq s$ with $|t-s| < \delta$.
  - By the mean value theorem for each x,t there is a $u_{x,t}$ between s and t such that

$$\psi(x,t) = (D_2\varphi)(x,u_{x,t}).$$

  - Then for t with $|t-s| < \delta$ we have $|u_{x,t} - s| < \delta$ and therefore

$$|\psi(x,t) - (D_2\varphi)(x,s)| = |(D_2\varphi)(x,u_{x,t}) - (D_2\varphi)(x,s)| < \varepsilon.$$

  - Thus $\psi^t \to (D_2\varphi)^s$ uniformly on $[a,b]$ as $t \to s$, and therefore

$$\int_a^b \psi^t d\alpha \to \int_a^b (D_2\varphi)^s d\alpha.$$

  - Since

$$\frac{f(t) - f(s)}{t-s} = \int_a^b \psi^t d\alpha,$$

    this shows that $f'(s)$ exists and has the value claimed.

Final Notes

- We have gone over many mathematical results and methods.
- All are used within statistics and probability.
- Many will come up again in courses you take in the future.
- You have practiced developing and presenting careful proofs.
- This will help you
  - to convince yourself that your new method is working;
  - to convince others that your new method is working;
  - to make sure your computer programs are doing the right thing;
  - ...