General Theorems

Squeeze Theorem

$$a_n \leq b_n \leq c_n$$

$$\lim a_n \leq \lim b_n \leq \lim c_n$$

Hölder's Inequality

Let p and q be positive real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$. For any two sequences of real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, Hölder's inequality states:

$$\sum_{i=1}^{n} |a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}}$$

In this inequality:

- p and q are positive real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$. These numbers are called conjugate exponents of each other.
- a_i and b_i are elements of the sequences $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, respectively.

Affine Function

Affine - linear map + constant C. $f : \mathbb{R} \to \mathbb{R}$

$$f = g + C$$

Taylor Formula

$$f(x+h) = f(x) + Df(x)(h) +$$

$$+\frac{1}{2} D^2 f(x)(h,h) + o(||h||^2) =$$

$$= f(x) + < \nabla f(x), h > + \frac{1}{2} h^T H f(x) h + o(||h||^2) =$$

$$= f(x) + \sum_{i=1}^{p} h_i \frac{\partial f}{x_i}(x) +$$

$$+\frac{1}{2} \sum_{i,j} h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(x) + o(||h||^2)$$

$f : \mathbb{R}^d \to \mathbb{R}, C^1$
$f(x+h) = f(x) + f'(x)h + o(|h|)$

$f : \mathbb{R} \to \mathbb{R}, C^2, h \in \mathbb{R}^d$
$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2} + o(h^2)$

Norm | Inner Product

Norm

A mapping $\| \cdot \| : \mathbb{R}^d \to \mathbb{R}$ is a norm on $\mathbb{R}^d$ if - For all $x \in \mathbb{R}^d$, $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$ - For all $x \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, $\|\lambda x\| = |\lambda| \|x\|$ - For all $x, y \in \mathbb{R}^d$, $\|x + y\| \leq \|x\| + \|y\|$ (Triangle inequality)

For $\mathbb{R}^1$ we have $(\mathbb{R}, |\cdot|)$ normed space.
For $\mathbb{R}^2$ we have: 1. Euclidian norm: $x = (x_1, x_2)$: $||x||_2 = \sqrt{x_1^2 + x_2^2}$ 2. $||x||_1 = |x_1| + |x_2|$ 3. $||x||_\infty = max(|x_1|, |x_2|)$

$(\mathbb{R}, ||\cdot||_1)$ and $(\mathbb{R}, ||\cdot||_2)$ are both normed spaces, but defined by different normes, thus different.

For $\mathbb{R}^d$ we have: $x = (x_1, x_2)$: 1. $||x||_1 = |x_1| + ... + |x_d|$ 2. $||x||_2 = \sqrt{x_1^2 + ... + x_d^2}$ 3. $||x||_\infty = max(|x_1|, ..., |x_d|)$ 4. $||x||_p = (\sum_{i=1}^{d} x_i^p)^{\frac{1}{p}}, p \geq 1$

The pair $(\mathbb{R}^d, \|\cdot\|)$ is a normed vector space.
Change norm => normed space changes.

Inner Product

Let $\mathbb{R}^d$ be a vector space. A mapping $\langle \cdot, \cdot \rangle : (\mathbb{R}^d)^2 \to \mathbb{R}$ is an inner product on $\mathbb{R}^d$ if: - For all $x, y \in \mathbb{R}^d$, $\langle x, y \rangle = \langle y, x \rangle$ (symmetry). - For all $x, y, z \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ (left linearity). - For all $x \in \mathbb{R}^d$, $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$ (positive definiteness).

Symmetry + left linearity defines bilinearity.

Inner product:

$$\mathbb{R} : \langle x, y \rangle = xy$$

$$\mathbb{R}^d : \langle x, y \rangle = \sum_{i=1}^{d} x_i y_i$$

Cauchy-Schwarz Inequality

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle, \forall x, y \in \mathbb{R}^d$$

With **equality** $\iff$ x and y are **linearly independent**.

Canonical Norm

Let $\langle \cdot, \cdot \rangle$ be the usual inner product on $\mathbb{R}^d$. The mapping

$$\| \cdot \| : x \in \mathbb{R}^d \mapsto \sqrt{\langle x, x \rangle}$$

is a norm on $\mathbb{R}^d$, the **canonical** norm associated with $\langle \cdot, \cdot \rangle$.

$||x|| = \sqrt{\langle x, x \rangle}$
$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$
$|\langle x, y \rangle|^2 \leq ||x||^2 ||y||^2$
$|\langle x, y \rangle| \leq ||x|| ||y||$

Topology

Open Ball

Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$.

$$\forall a \in \mathbb{R}^d, r > 0, B_{\|\cdot\|}(a, r) = \{x \in \mathbb{R}^d : \|x - a\| < r\}$$

$$B_{|\cdot|}(a, r) = (a - r, a + r)$$

is the open ball of $(\mathbb{R}^d, \| \cdot \|)$ with center a and radius r. From the definition we see that balls depend on the defined norm. In other words, distance from any x to center is less than r.

Closed Ball

Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$.

$$\forall a \in \mathbb{R}^d, r > 0, \overline{B}_{\|\cdot\|}(a, r) = \{x \in \mathbb{R}^d : \|x - a\| \leq r\}$$

is the closed ball of $(\mathbb{R}^d, \| \cdot \|)$ with center a and radius r.

$$\overline{B}_{\|\cdot\|}(a, r) = [a - r, a + r]$$

Neighbourhood of a Point

Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$. A subset $V \subset \mathbb{R}^d$ is a neighborhood of $a \in \mathbb{R}^d$ if

$$\forall r > 0 : B_{\|\cdot\|}(a, r) \subset V$$

Open Set

Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$.

A subset $O \subset \mathbb{R}^d$ is an open set of $(\mathbb{R}^d, \| \cdot \|)$ if

$$\forall a \in O, \exists r > 0 : B_{\|\cdot\|}(a, r) \subset O$$

We can find any open ball with small r s.t. for any point in O open ball will be a subset of O. **Frontier not included.**

Closed Set

A subset $F \subset \mathbb{R}^d$ is a closed set of $(\mathbb{R}^d, \| \cdot \|)$ if its complement ($F^c := \mathbb{R}^d \setminus F$) is an open set of $(\mathbb{R}^d, \| \cdot \|)$.

If we take a ball, where center is on the frontier, we will notice that some part of the ball is not in the set ($r > 0$) => we have closed set.

Open & Closed Set Properties

- An open set is a set which is a neighborhood of all its points
- Any open ball is open
- Any closed ball is closed
- In $\mathbb{R}^d$ a **compact** set is a set which is **closed** and **bounded**.

Remark: If F is closed => F^c is open.

Interior of a Set

Let $\| \cdot \|_{\mathbb{R}^d}$ be a norm on $\mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$. The interior of A is the largest open subset of $\mathbb{R}^d$ contained in A (it exists) and is denoted by $\overset{\circ}{A}$

Remarks: $\overset{\circ}{A}$ is open and $\overset{\circ}{A} \subset A$

Example: $A = [1, 3]$, $\overset{\circ}{A} = (1, 3)$, $B = (1, 3)$, $\overset{\circ}{B} = (1, 3)$

Closure of a Set

Let $\| \cdot \|_{\mathbb{R}^d}$ be a norm on $\mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$. The closure of A is the smallest closed subset of $\mathbb{R}^d$ containing A (it exists) and is denoted by $\overline{A}$

Remarks: $\overline{A}$ is closed and $A \subset \overline{A}$

Example: $A = (1, 3]$, $\overline{A} = [1, 3]$

Frontier of a Set

Let $\|\cdot\|_{\mathbb{R}^d}$ be a norm on $\mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$. The frontier of A is $\delta A = \overline{A} \setminus \overset{\circ}{A}$

Example: $A = [1, 3)$, $\delta A = \{1, 3\}$ (points 1 and 3)

Family of Open Sets

Let $(O_\alpha)_{\alpha \in I}$ be a family of open sets, then

$$\bigcup_{\alpha \in I} O_\alpha$$

is also open (infinite sets).

Let $(A_\delta)_{\delta \in \Delta}$ be a family of open sets s.t. $|\Delta| < +\infty$, then

$$\bigcap_{\delta \in \Delta} A_\delta$$

is also open (finite sets).

Sequential Characterization of Closed Set

Let $(\mathbb{R}^d, ||\cdot||)$ be a normed space. $F \subseteq \mathbb{R}^d$ is closed $\iff \forall (x_n)_{n \geq 1} \subseteq F$ s.t. $x_n \to l$ then $l \in F$

Sequences and Mappings

Convergent Sequence

Let $\|\cdot\|_{\mathbb{R}^d}$ be a norm on $\mathbb{R}^d$. A sequence $(x_n)_{n \in \mathbb{N}}$ of elements of $\mathbb{R}^d$ converges if there exists $x \in \mathbb{R}^d$ such that

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} : \forall n \in \mathbb{N}, n \geq N, \|x_n - x\| < \varepsilon$$

Limit of Convergent Sequence

Let $\|\cdot\|_{\mathbb{R}^d}$ be a norm on $\mathbb{R}^d$, and let $(x_n)_{n \in \mathbb{N}}$ be a sequence of elements of $\mathbb{R}^d$

If $(x_n)_{n \in \mathbb{N}}$ converges, it has a unique limit denoted $\lim_{n \to \infty} x_n$

Furthermore,

$$x = \lim_{n \to \infty} x_n \iff \lim_{n \to \infty} \|x_n - x\| = 0$$

Proposition

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, $\|\cdot\|'$ be a norm on $\mathbb{R}^k$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ be a function. The function has the limit l in $\mathbb{R}^k$ at $a \in \mathbb{R}^d$ if

$$\forall \varepsilon > 0, \exists \delta > 0 : \forall x \in \mathbb{R}^d$$

$$\|x - a\| < \delta \Rightarrow \|f(x) - l\|' < \varepsilon$$

If it exists, the limit l is unique and is denoted $\lim_{x \to a} f(x)$

Proposition

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, $\|\cdot\|'$ be a norm on $\mathbb{R}^k$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ be a function. The function has the limit l in $\mathbb{R}^k$ at $a \in \mathbb{R}^d \iff$ for any sequence $(x_n)_{n \in \mathbb{N}}$ of elements of $\mathbb{R}^d$,

$$\lim_{n \to \infty} \|x_n - a\| = 0 \iff ||f(x) - l||' = 0$$

Continious Function

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, $\|\cdot\|'$ be a norm on $\mathbb{R}^k$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ be a function. The function f is continuous at a point a in $\mathbb{R}^d$ if

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

Let O be an open subset of $\left(\mathbb{R}^d, \|\cdot\|\right)$. The function f is continuous on O if f is continuous at any point of O.

To Compute the Limit of f at the Given Point x_0

$$\lim_{x \to x_0} \frac{g(x) - g(x_0)}{x} = g'(x_0)$$

Differential Calculus

$f : (\mathbb{R}^d, ||\cdot||) \to (\mathbb{R}^k, ||\cdot||')$

f is differentiable on $a \in \mathbb{R}^d$ if $\exists L \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$ s.t.

$$\lim_{h \to 0} \frac{||f(a+h) - f(a) - L(h)||'}{||h||} = 0$$

We say L is differential of $f : L = Df(a)$

$f : \mathbb{R}^d \to \mathbb{R}^k$ Linear Map

If f is a linear map, then f is differentiable on $\mathbb{R}^d$ and then

$$Df(a)(h) = f(h) \quad \forall a \in \mathbb{R}^d, \forall h \in \mathbb{R}^d$$

$f : \mathbb{R} \to \mathbb{R}$

f is derivable on a if and only if f is differentiable on a and

$$Df(a)(h) = f'(a)h$$

Chain Rule: $f : \mathbb{R} \to \mathbb{R}, g : \mathbb{R} \to \mathbb{R}$

If f is derivable on a and g is derivable on $f(a)$, then $g \circ f$ is derivable on a and:

$$(g \circ f)'(a) = g'(f(a))f'(a)$$

$$D(g \circ f)(a)(h) = [Dg(f(a))](Df(a)(h))$$

First Order Directional Derivative

A mapping $f : \mathcal{O} \to \mathbb{R}^k$ is differentiable at $a \in \mathcal{O}$ along a given direction $h \in \mathbb{R}^d$ if the limit

$$D_h f(a) = \lim_{t \to 0} \frac{f(a + th) - f(a)}{t}$$

exists and is finite.

If a mapping $f : \mathcal{O} \to \mathbb{R}^k$ is differentiable at $a \in \mathcal{O}$, then f is differentiable at a along all directions, and

$$D_h f(a) = Df(a)(h); \quad \forall h \in \mathbb{R}^d.$$

Note: if f is differentiable at a, then f is differentiable along all directions at a, but if f is differentiable along all directions at a != f is differentiable at a

$f : \mathbb{R}^d \to \mathbb{R}, \nabla$

If all partial derivatives $\frac{\partial}{\partial x_i} f, \forall 1 \leq i \leq d$ exist and continious, then f is differentiable and

$$Df(a)(h) =< \nabla f(a), h >$$

$$Df(a) = \begin{bmatrix} \frac{\partial}{\partial x_i} f(a) \\ \vdots \\ \frac{\partial}{\partial x_d} f(a) \end{bmatrix}$$

$f : \mathbb{R} \to \mathbb{R}^d$

$$f(x) = f(x * 1) = xf(1)$$

$$f(x) = cx \quad \forall x \in R, c \in R^d$$

$$Df(x)(h) = Df(x)(1)h$$

$f : \mathbb{R}^d \to \mathbb{R}^k, J$

$x \to f(x) = (f_1(x), ..., f_k(x))$
If all partial derivatives exist and continious:

$$\frac{\partial}{\partial x_j} f_i, \quad \forall 1 \leq i \leq k, \forall 1 \leq j \leq d$$

Then f is differentiable and

$$Df(a)(h) = J_f(a)(h), \quad J_f(a) = (k \times d), h = (d \times 1)$$

$$J_f(a) = \left(\frac{\partial}{\partial x_j} f_i(a) \right)$$

$$= \begin{bmatrix} 1 \leq i \leq k \\ 1 \leq j \leq d \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(a) & \dots & \frac{\partial}{\partial x_d} f_1(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_k(a) & \dots & \frac{\partial}{\partial x_d} f_k(a) \end{bmatrix}$$

$$Df(a) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$$

$$Df(a)(h) \to \mathbb{R}^d, \forall h \in \mathbb{R}^d$$

Class C^1

Let O an open subset of $\mathbb{R}^d$. A function $f : O \to \mathbb{R}^k$ is of class C^1 on O if all partial derivatives exist and are continious.

Let O an open subset of $\mathbb{R}^d$ and a function $f : O \to \mathbb{R}^k$ is of class C^1. Then f is differentiable on O.

A function $f : \mathbb{R}^d \to \mathbb{R}$ of class C^1. Then f is differentiable on $\mathbb{R}^d$ and

$$Df(a)(h) =< \nabla f(a), h >$$

Twice Differentiable Map

If all 2nd order partial derivatives of f on $a, \forall a \in R^d$ continious, then f is twice differentiable.

Second Order Directional Derivative

A map $f : O \to R^k$ is twice differentiable at $a \in O$ in the direction $h \in R^d$, then in the direction $n \in R^d$ if

$$D_{h,n}f(a) = D_n D_h f(a)$$

We first differentiate along the direction h and then along the direction n

Schwarz Theorem

If map $f : O \to R^k$ has continious second order partial derivatives on O then for all $a \in O$:

$$\partial^2_{jl} f(a) = \partial^2_{lj} f(a), \quad \forall (l, j) \in \{1, ..., d\}^2$$

In other words, we can switch the order when we differentiate and get the same result.

$f : \mathbb{R}^d \to \mathbb{R}, H$

If all second partial derivatives exist and continious:

$$\frac{\partial^2}{\partial x_i \partial x_j} f, \quad \forall 1 \leq i, j \leq d$$

Then:

$$\frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2}{\partial x_j \partial x_i} f$$

and f is twice differentiable and

$$D^2 f(a)(h, k) = k^T H_f(a) h$$

$$D^2 f(a) \in \mathcal{L}_2(\mathbb{R}^d, \mathbb{R}^k)$$

$$H_f(a) = \left(\frac{\partial^2}{\partial x_i \partial x_j} f(a) \right) =$$

$$\begin{bmatrix} \frac{\partial^2}{\partial^2 x_{11}} f(a) & \dots & \frac{\partial^2}{\partial x_1 \partial x_d} f(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} f(a) & \dots & \frac{\partial^2}{\partial^2 x_{dd}} f(a) \end{bmatrix}$$

Convex Sets and Functions

Convex Set

A subset $C \subset \mathbb{R}^d$ is convex if

$$\forall x, y \in C, \forall t \in [0, 1], (1 - t)x + ty \in C$$

In simpler terms, C is convex $\iff$ any time we pick and two points, their segment stays in C

If C is convex:

$$\forall n \in \mathbb{N}, x_1, ..., x_n \in C, t_1, ..., t_n \in \mathbb{R}_+, t_1 + ... + t_n = 1$$

$$\sum_{k=1}^{n} t_k x_k \in C$$

$\sum_{k=1}^{n} t_k x_k \in C$ is called **convex combination** ($\sum_{k=1}^{n} t_k x_k \in C$ is a linear combination, and when $t_k \geq 0$ then it is convex combination).

Particular Convex Sets

Following sets are convex:

- Vector subspaces of $\mathbb{R}^d$
- Intersection of two convex sets of $\mathbb{R}^d$
- Translation of convex set is also a convex set
- The open $B_{||\cdot||}(a, r) = \{x \in \mathbb{R}^d : ||x - a|| < r\}$ and closed balls $\overline{B}_{||\cdot||}(a, r) = \{x \in \mathbb{R}^d : ||x - a|| \leq r\}$ of $(\mathbb{R}^d, ||\cdot||)$ are convex

Proof for closed ball:
$\overline{B}_{||\cdot||}(a, r) = a + \overline{B}_{||\cdot||}(0, r)$ - translation by a, still convex. Now prove that $\overline{B}_{||\cdot||}(0, r)$ is convex.
Let $x, y \in \overline{B}_{||\cdot||}(0, r)$ and let $t \in [0, 1]$
$(1 - t)x + ty \in \overline{B}_{||\cdot||}(0, r)$
$||(1 - t)x + ty - 0|| \leq r$
$||(1-t)x+ty-0|| \leq ||(1-t)x|| + ||ty|| = |(1-t)|||x|| + |t|||y|| = (1 - t)||x|| + t||y||$
Since $x, y \in \overline{B}_{||\cdot||}(0, r) \Rightarrow ||x|| \leq r, ||y|| \leq r, ||x - 0|| \leq r, ||y - 0|| \leq r$
$\Rightarrow ||(1 - t)x + ty - 0|| \leq (1 - t)r + tr = r$
$\Rightarrow (1 - t)x + ty \in \overline{B}_{||\cdot||}(0, r)$

Convex Function

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the following set is convex:

$$\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq y\}$$

We also have:

$$\forall x, y \in \mathbb{R}^d, t \in [0, 1]$$

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

Example:
$f(x) = x^2$
$\forall x, y \in \mathbb{R}, t \in [0, 1]$
$((1-t)x + ty)^2 \leq (1-t)x^2 + ty^2$
$\alpha + \beta = 1$
$(\alpha x + \beta y)^2 \leq \alpha x^2 + \beta y^2$

Convexity and Derivative

We have $(\mathbb{R}^d, ||\cdot||)$ and O open and convex, $f \in C^1(O, \mathbb{R})$ 1. f is convex if and only if

$$f(y) \geq f(x) + < \nabla f(x), y - x >; \forall x, y \in O$$

2. The map ∇f is monotone:

$$< \nabla f(y) - \nabla f(x), y - x > \geq 0; \forall x, y \in O$$

Examples:

$f : \mathbb{R} \to \mathbb{R}$
$f'(x)$ is non-decreasing (respectively increasing) => f is convex (respectively strictly convex).

$f(x) = x^2$
$f'(x) = 2x \uparrow => f(x)$ is convex
$f'(x) \uparrow \iff \forall x < y => f'(x) \leq f'(y) \iff \forall x, y \in \mathbb{R} : (x - y)|f'(x) - f'(y)| \geq 0$

$f : \mathbb{R} \to \mathbb{R}, C^2$
if $f''(x) \geq 0, \forall x \in \mathbb{R}$
=> $f'(x)$ is non-decreasing
=> $f(x)$ is convex

$f(x) = x^2, f'(x) = 2x, f''(x) = 2$
=> $f(x)$ is convex.

Convexity and Second Derivative (Hessian)

Let O open set of $(\mathbb{R}^d, ||\cdot||)$, $f \in C^2(O, \mathbb{R})$. 1. f is convex $\iff$ $Hf(x)$ is positive, meaning $\forall h \in \mathbb{R}^d : h^T Hf(x) H \geq 0$ or $sp[Hf(x)] \subset \mathbb{R}_+ \forall x \in O$ (all eigenvalues are real and nonnegative) 2. The function f is strictly convex $\iff$ $Hf(x)$ is positive definite, meaning $\forall h \in \mathbb{R}^d : h^T Hf(x) H > 0$ or $sp[Hf(x)] \subset \mathbb{R}_+^* \forall x \in O$ (all eigenvalues are real and strictly positive)

Hilbert Projection

Let $||\cdot||_2$ the Euclidean norm on $\mathbb{R}^d$. Let C non-empty closed convex subset of $\mathbb{R}^d$. 1. . There exists a unique element $x_0 \in C$ s.t.

$$||x - x_0||_2 = d(x, C) := \inf_{y \in C} ||x - y||_2$$

2. For all $y \in C$

$$< x - x_0, y - x_0 > \leq 0$$

Numerical Probabilities

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$

$$f_X = (F_X)' = \frac{d}{dx}F_X$$

Property	$X(\Omega)$ finite or countable set	$X(\Omega) \subset R$
Distribution $P_X(\{x\})$	$P_X(\{x\}) = P(X = x)$	$dP_X(x) = f_X(x)dx$
Cumulative Distribution Function $F_X(x)$	$F_X(x) = \sum_{k \in X(\Omega), k \leq x} P(X = k)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$
Expectation $E[X]$	$\sum_{k \in X(\Omega)} kP(X = k)$	$\int_{-\infty}^{\infty} x f_X(x)dx$
Transfer Theorem $E[g(X)]$	$E[g(X)] = \sum_{k \in X(\Omega)} g(k)P(X = k)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

$$Var(X) = E[X^2] - (E[X])^2 \text{ and } \sigma_X = \sqrt{Var(X)}$$

Distribution	$X(\Omega)$	Support + Definition	$E[X]$	$Var(X)$
Bernoulli	$X \sim B(p)$	$\begin{bmatrix} X(\Omega) = \{0,1\} \\ P(X=1) = p \\ P(X=0) = 1-p \end{bmatrix}$	p	$p(1-p)$
Binomial	$X \sim B(n,p)$	$\begin{bmatrix} X(\Omega) = \{0,\ldots,n\} \\ P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \end{bmatrix}$	np	$np(1-p)$
Geometric	$X \sim G(p)$	$\begin{bmatrix} X(\Omega) = \{N \geq 1\} \\ P(X=k) = p(1-p)^{k-1} \end{bmatrix}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$X \sim P(\lambda)$	$\begin{bmatrix} X(\Omega) = \mathbb{N}_0 \\ P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \end{bmatrix}$	λ	λ
Uniform	$X \sim U([a,b])$	$\begin{bmatrix} X(\Omega) = [a,b] \\ f_X(x) = \frac{1}{b-a} 1_{R^+}(x) \end{bmatrix}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$X \sim E(\lambda)$	$\begin{bmatrix} X(\Omega) = \mathbb{R}^+ \\ f_X(x) = \lambda e^{-\lambda x} 1_{R^+}(x) \end{bmatrix}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$X \sim N(\mu, \sigma^2)$	$\begin{bmatrix} X(\Omega) = \mathbb{R} \\ f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \end{bmatrix}$	μ	σ^2
Standard Normal	$X \sim N(0,1)$	$X(\Omega) = \mathbb{R}$	0	1

Simulation of Random Variables

The generation of "random numbers" is the problem of producing a deterministic sequence of values in $[0, 1]$ which imitates a sequence of i.i.d. uniform random variables of distribution $U([0, 1])$.

Pseudo-Random Variables

If $u_n = U_n(\omega)$, we call $(u_n)_n$ a sequence of pseudo-random numbers (because it is deterministic).

Distributions

Distributions can be characterized by:

- if $X(\Omega)$ - image of random variable - is finite and countable, then the law of X is characterized by the probability of singletons and we have CDF
- if $X(\Omega) \subset R$ absolutely continious w.r.t the Lebesque measure, then the law is characterized by PDF - probability density function - distributions that have a density w.r.t. the Lebesque measure.

Characteristic Function

$X : \Omega \to R, \quad dR_x = f_X(x)dx$

$$G_X(t) = E[e^{itX}] = \int_R e^{itx} f_X(x)dx$$

$$\phi(x) = e^{itx}$$

Generalized Inverse Function

Let F be a non-decreasing function on R. The generalized inverse of F, denoted by F^- is the function defined by

$$F^-(u) = \inf\{x \in R : F(x) \geq u\}$$

The function F^- is non-decreasing, left-continuous and satisfies

$$F^-(u) \leq x \iff u \leq F(x), \forall u \in (0, 1)$$

If F is **increasing and continious** on R, then F has an inverse function defined by F^{-1} s.t. $F * F^{-1} = Id_{(0,1)}$ and $F^{-1} * F = Id_R$

Proposition

Let $U \sim U((0, 1))$ and $F = \mu((-\infty, x])$ where μ is a probability distribution on $(R, B(R))$. Then $F^-(U) \sim \mu$.

In other words, $F^-(U)$ follows the law of X.

Acceptance-Rejection

Let $f : R^d \to R_+$ s.t. there exists a (positive) probability density g (instrumental, easier simulationwise) and a positive real constant $c > 0$ s.t.

$$f(x) \leq cg(x)$$

Remember: we want the curve of g to be always above the curve of f no matter the x, thus we have the constant c. g is easy to sample from. Sampling is going to be more likely in places where g has a higher density and less likely where g has a lower density.

Problem: if c is very large (meaning we need to scale g my large c so that it is always above f), then $P(A)$ is very low and it will take a lot of time to simulate p.

Markov Chains

Discrete Stochastic Process

A discrete stochastic process is a collection of random variables indexed by time.

Distribution of a Process

Distribution of a process is a collection of all the laws and distributions of random variables.

$$\mathbb{P}_{(X_{k_1},\dots,X_{k_n})}, \forall n \in N, k_1, \dots, k_n \in N, k_1 < \dots < k_n$$

Example

We flip coin n times. $X_1, \dots, X_n$ are independent.
$P[\{0\}] = P[X_k = 0] = 1 - p, P[\{1\}] = P[X_k = 1] = p$
$=> P[X_k = x_k] = p^{x_k}(1 - p)^{1-x_k}$
$\mathbb{P}_{(X_{k_1},\dots,X_{k_n})}(\{x_1, \dots, x_n\}) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1 - p)^{1-x_i} = p^{\sum_{i-1}^n x_i}(1 - p)^{n-\sum_{i=1}^n x_i}$

Filtration

A sequence of $F = (F_n)_{n\geq 0}$ of sub sigma-algebras of A is a filtration of (Ω, A) if

$$\forall m, n \in N_0, n \leq m, F_n \subset F_m$$

The quadruplet (Ω, A, F, P) is a filtered probability space.
Remark: F_n includes all the information provided in the r.v.

Example

Let $(X_n)_n$ be a process and $F_n = \sigma(X_0, ..., X_n)$ is the smallest sigma-albegra making the application $\omega \in \Omega \to (X_0(\omega), \ldots, X_n(\omega))$ measurable. The familty $(F_n)_n$ is a called the natural filtration assossicated to X.

Adapted Process

A process $X = (X_n)_n$ is adapted to the filtration $F = (F_n)_n$ if X_n is F_n- measurable for all $n \in N_0$

Examples

1. $S_n = \sum_{k=1}^{n} X_k$. Is S_n F_n- measurable? Yes, for all n, then it is adapted.
2. $T_n = X_n + X_{n+1}$. T_n is not F_n- measurable and $(T_n)_n$ is not $(F_n)_n$ adapted.

Stopping Time

A random variable τ is a $F-$ stopping time if

$$\{\omega \in \Omega : \tau(\omega) = n\} = \{\tau = n\} \in F_n$$

In other words, τ_A is the first time the process X appears in A.

1. Constant applications from Ω to N_0 are $F-$ stopping times
2. if τ_1 and τ_2 are $F-$ stopping times, then $\min(\tau_1, \tau_2)$ and $\max(\tau_1, \tau_2)$ are $F-$ stopping times
3. If the process $(X_n)_n$ is $F-$ adapted, then $\tau_A = \min\{n \in N_0 : X_n \in A\}$ is a $F-$ stopping time.

Example

$\{\tau_A = 2\} = \{X_0 \notin A\} \cap \{X_1 \notin A\} \cap \{X_2 \in A\} \in F_2$
$F_2 = \sigma(X_0, X_1, X_2)$

Markov Chain

Probability of the future knowing all the past is a conditional probability of the future knowing jus the present.

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

The distribution of X_0 is the initial distribution of the Markov chain X.

Stochastic Matrix

A matrix $(P(x, y))_{(x,y)\in E}$ is stochastic if

1. $P(x, y) \geq 0 \quad \forall x, y \in E$
2. For all $x \in E, \sum_{y\in E} P(x, y) = 1$

Homogeneous Markov Chain

A Markov Chain X is homogeneous if there exists a stochastic matrix $(P(x, y))_{(x,y)\in E}$ s.t.

$$P(X_{n+1} = y | X_n = x) = P(x, y)$$

The matrix P is the transition matrix of the Markov chain X.

Knowing the present, predicting the future, does not depend on time n.

Example

$E = \{1, 2, 3\}-$ state space.
$$P = \begin{bmatrix} p(1,1) & p(1,2) & p(1,3) \\ p(2,1) & p(2,2) & p(2,3) \\ p(3,1) & p(3,2) & p(3,3) \end{bmatrix} = \begin{bmatrix} 1/2 & 1/4 & 1/4 & \sum = 1 \\ 1/3 & 0 & 2/3 & \sum = 1 \\ 0 & 0 & 1 & \sum = 1 \end{bmatrix}$$
P is stochastic transition matrix of Markov Chain X, X is homogeneous.

Initial Distribution

A process $X = (X_1, \ldots, X_n) = (X_n)_{n\geq 0}-$ is a homogeneous Markov chain of P_{X_0} initial distribution of and transition matrix P if and only if

$$P(X_0 = x_0, \ldots, X_n = x_n) = P_{X_0} \prod_{k=0}^{n-1} P(X_k, X_{k+1})$$

If we know the initial distribution and transition matrix, we know all the laws of Markov chain.

HMC is characterized by its initial distribution P_{X_0} and the transition matrix P.

This gives the expression of the joint distribution of $(X_1, \ldots, X_n)$.

$$P(X_{n+m} = y | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x_n) = (P^m)(x_n, y)$$

This gives the expression of the probability of X_{n+m} given $(X_1, \ldots, X_n)$.

Example

$P[X_{0+3} = y | X_0 = x] = P[X_{3+0} = y | X_0 = x] = P^3(x, y)$

Example (conitnued)

1. $P[X_2 = 3 | X_0 = 1] = P[X_{0+2} = 3 | X_0 = 1] = p^2(1, 3)$
2. $P[X_2 = 3 | X_0 = 1] = P[X_1 = 1 | X_0 = 1] P[X_2 = 3 | X_1 = 1] + P[X_1 = 2 | X_0 = 1] P[X_2 = 3 | X_1 = 2] + P[X_1 = 3 | X_0 = 1] P[X_2 = 3 | X_1 = 3] = 1/2 * 1/4 + 1/4 * 2/3 + 1/4 * 1 = 13/24$

$$p^2(1,3) = \begin{bmatrix} *1/2* & 1/4* & *1/4* \\ 1/3 & 0 & 2/3 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & 1/4 & **1/4** \\ 1/3 & 0 & *2/3* \\ 1 & 0 & *1* \end{bmatrix} =$$

$$= 1/2 * 1/4 + 1/4 * 2/3 + 1/4 * 1 = 13/24$$

Reaching States

$$\mathbb{P}_x(X_n = j) = \mathbb{P}(X_n = j | X_0 = x), \forall j \in E$$

probability to reach j at time n from x

$$N_x = \sum_{n=1}^{\infty} 1_{\{x\}}(X_n)$$

N_x - starting from x. N_x - number of times you leave state x and come back to it. It is random, can be infinity.

Types of States

1. The state x is **recurrent** for X if $E_x[N_x] = \infty$ (expectation). Recurrent means the process visits x infinity of times, N_x is unbounded $(1 \to 1, 1 \to 2, 2 \to 2, 2 \to 1)$. We say the **chain** X is **reccurent** if all **states** are **recurrent**.
2. The state x is **transient** for X if it is not recurrent, i.e., $E_x[N_x] < \infty$ (expectation is finite) $(1 \to 1p(1/2), 1 \to 2p(1/2), 2 \to 2p(1))$. Since we have probability to go to 2, it is not recurrent (no path from 2 to 1).
3. The state x is **absorbing** for X if $P(x, x) = 1$ (in previous example 2 is absorbing).
4. For X, the state y is **reachable from** x, which is denoted by $x \to y$, if there exists $n \in \mathbb{N}_0$ such that $P_x(X_n = y) > 0$ (in previous example 2 is rechable from 1). **The path does not need to be direct**.
5. For X, x and y **communicate** if $x \to y$ and $y \to x$ (in previous example, 1 and 2 do not communicate). **The path does not need to be direct**.
6. The state is **nonnull** if the expected time to return is finite.
7. **Steady** state: $P(X_{t+1} = s) = P(X_t = s)$. **Markov chain** is **steady** if all **states** are **steady**.

Closed Classes, Irreducible Classes

1. The class C is closed for X if
   - $\forall x, y \in E$, if $x \in C$ and $x \to y$, then $y \in C$ ($x \in C$, y is reachable from x, then C is closed)

$C = \{4, 5, 6\}$ closed, but $C = \{2, 4\}$ not closed, since 5 is reachable from 2, but $5 \notin C$

1. The class C is **irreducible** for X if all its states communicate (also have to check $x \to x$)
2. The Markov chain X is **irreducible** if E is an irreducible class of X. A Markov chain is considered **irreducible** if it is possible to reach any state from any other state with a positive probability in a finite number of steps

Irreducible: Recurrent and Transient States

Let X be a Homogeneous Markov chain.

1. An irreducible, finite and closed class consists of recurrent states.
2. An irreducible and non-closed class consists of transient states.

If the number T of transient states of X is finite then for any irreducible and closed class of recurrent states R and any $x \in T$,

$$\mathbb{P}_x(\tau_R < +\infty) = \sum_{y \in R} P(x,y) + \sum_{y \in T} P(x,y)\mathbb{P}_y(\tau_R < +\infty)$$

$$\tau_R = \min\{n \geq 1 : X_n \in R\}$$

Invariant Probability Measure | Steady State

Answers question: to which state does the Markov Chain converge?

$\mathcal{P}(E)-$ power set, $\mu-$ probability measure on $(E, \mathcal{P}(E))$. We say X is invariant with μ if $\mu \cdot P = \mu$.

If the Markov chain X is irreducible, then it admits a unique invariant probability measure μ. Moreover, for all $x \in E$,

$$\mathbb{E}_x[\tau_x] = \frac{1}{\mu(\{x\})}$$

$$\mu \cdot P = \mu \iff (\mu \cdot P)^T = \mu^T \iff P^T \mu^T = \mu^T$$

μ^T is an eigenvector (for the eigenvalue 1) of P^T. Then to compute invariant measure, it can be easier to see it as an eigenvector of P^T.

Example

Find if there exists the invariant measures of P^T.

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/3 & 0 & 2/3 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathcal{P}^T \mu^T = \mu^T \iff \begin{bmatrix} 1/2 & 1/3 & 0 \\ 1/4 & 0 & 0 \\ 1/4 & 2/3 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

$$\begin{cases} 1/2\mu_1 + 1/3\mu_2 = \mu_1 \\ 1/4\mu_1 = \mu_2 \\ 1/4\mu_1 + 2/3\mu_2 + \mu_3 = \mu_3 \\ \mu_1 + \mu_2 + \mu_3 = 1 \end{cases}$$

$$\iff \mu_1 = \mu_2 = 0, \mu_3 = 1$$

Thus, the only invariant measure of X is $\mu = (0,0,1)$

This means: if there is 0% chance of being at 1 in the current step, 0% chance of being at 2 in the current step and 100% chance of being at 3 in the current step, then what is the probability of being at $1,2,3$ in the next step? They are the same, since we have stationary distribution. The probability will not change ever again.

So, **to find invarient measures**:

1. Compute transition probability matrix P
2. Solve system $\mu \cdot P = \mu$

Corollary

Let $R_1, \ldots, R_k$ ($k \in \{1, \ldots, \text{card}(E)\}$) be the irreducible classes of X. For all $i \in \{1, \ldots, k\}$, let μ_i be the invariant probability measure of X with support R_i. The invariant measures of X are of the form

$$\sum_{i=1}^{k} \lambda_i \mu_i,$$

where $\lambda_1, \ldots, \lambda_k \in [0,1]$ such that $\sum_{i=1}^{k} \lambda_i = 1$.

Ergodic Theorem

If X is irreducible of invariant probability measure μ, then for all functions $f : E \to R$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow[n \to +\infty]{P} \sum_{x \in E} f(x)\mu(\{x\}) = E[f(Y)], Y \sim \mu$$

In statistics, we have $E[f(Y)]$, in LLN: $\frac{1}{n}\sum_{i=1}^{n} f(Y_i) \to E[f(Y)]$, **where** $\perp Y_i$

By Markov chain we can achieve: 1. $X = (X_n)_{n \geq 0}-$ HMC, irreducible 2. $\mu-$ invariant measure for X 3. $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow[n \to +\infty]{P} \sum_{x \in E} f(x)\mu(\{x\}) = E[f(Y)], Y \sim \mu$ (here is μ is now target distribution). We do not require independence of X_i !

Stationarity, Irreducibility

1. **Stationarity**: X is stationary of invariant (or stationary) distribution f is $X_0 \to f \cdot \lambda$ implies $X_n \to f \cdot \lambda$ for all n. In other words, if we have a distribution at some time step and we allow the Markov Chain to proceed, we want the distribution at each of the states to stay the same. This is important for sampling, that is the Markov Chain is steady.

2. **Irreducibility**: X is $(f \cdot \lambda)-$irreducible if for all $A \in \mathcal{E}$ s.t. $f \cdot \lambda(A) > 0, \mathbb{P}_x(\tau_A < \infty) > 0$ for all $x \in E$.

Markov Chains Monte Carlo (MCMC)

Issue with accept-reject method: if c is large, then we almost never accept the sample. When $f(x)$ is high, thus close to $cg(x)$, we want to sample in this region, since $f(x)$ is high (density in this region is high). MCMC learns from previous samples - pick samples based on what learned from previous sample. In MCMC next sample depends on the last sample (Markov chain).

Let $Y : \Omega \to \mathcal{X}, Y \sim \mu = \mathbb{P}_Y, Y$ is absolutely continous.

$E[h(Y)] = \int_{\mathcal{X}} h(x) f_Y(x) dx$

- By **LLN**:
  - $y_1, \ldots, y_n \sim \mathbb{P}_Y \perp$ (can use generalized inverse or acceptance rejection to generate $y_1, \ldots, y_n$).
  - $\frac{1}{n} \sum_{i=1}^n h(y_i) \to E[h(y)]$
- By Markov Chain

$E[h(Y)]-?$

1. Look for stochastic matrix P, s.t. $\mu \cdot P = \mu$, μ is invariant.
2. generate a HMC X irreducible (all states communicate) + P its transition matrix and μ is its invariant measure (using Metropolis-Hastings).
3. By Ergodic Theorem approximate $\frac{1}{n} \sum_{i=1}^n h(X_i) = E[h(Y)], Y \sim \mu$

Theorem

Let $X = (X_n)_{n \geq 0}$ be the chain produced by the Metropolis-Hastings algorithm. Assumje that X is $(f \cdot \lambda)-$ irreducible. If $h \in L^1(E, f \cdot \lambda)$, then

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^n h(X_i) = \int_E h(x) f(x) \lambda(dx)$$

Symmetric Law

$$q(x, y) = q(y, x)$$

$$P(1, 2) = P(2, 1)$$