Mathematical Methods Concise Notes Pavel Bezkovich

Course structure:

1. Probability methods
   - 1.1. Probability generating functions
   - 1.2. Inequalities and limit theorems
   - 1.3. Stohastic processes
2. Fourier & related methods
   - 2.1. Fourier representations
   - 2.2. Discrete Fourier Methods
   - 2.3. Wavelets

1. Probability methods

(1.1.) Probability generating functions (PGFs)

Given a random variable $X \in \mathbb{N}$, s.t. $\mathbb{P}(X=k) = p_k$ $(k \in \mathbb{N})$, define the probability generating function as:

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k .$$

for all such z, for which the sum converges.

This function "encodes" all p_k's and so defines the RV completely.

Properties:

1.) $G_X(0) = p_0$

$G_X(1) = \sum_{k=0}^{\infty} p_k = 1$

2.) $G_X(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} \mathbb{P}(X=k) \cdot z^X \big|_{X=k} = \mathbb{E}(z^X)$.

3.) PGF is defined for all $|z| \leq 1$,
since $\sum_{k=0}^{\infty} |p_k z^k| \leq \sum_{k=0}^{\infty} p_k = 1$ and abs. convergence $\Rightarrow$ convergence

4.) PGF characterizes distribution of RV, namely

$G_X(z) = G_Y(z) \iff \forall k \in \{0,1,2,\dots\}.\ \mathbb{P}(X=k) = \mathbb{P}(Y=k)$.

Example: For binomial distribution

$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, so PGF:

$$G_X(z) = \sum_{k=0}^{n} p^k (1-p)^{n-k} \binom{n}{k} \cdot z^k = \sum_{k=0}^{n} \binom{n}{k} (pz)^k (1-p)^{n-k} = (pz+1-p)^n .$$

Uses of PGFs: 1.) Working with sums of IID RVs (such as sum of a pair of dice)

2.) Finding moments (mean, variance etc) using differentiation

Theorem

$$\frac{d^k}{dz^k} G_X(1) = \mathbb{E}[X(X-1)\dots(X-k+1)]$$

(proven by induction)

It follows that:

1.) $\mathbb{E}(X) = G'_X(1)$.

2.) $\underline{Var(X)} = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2 - 2X\mathbb{E}(X) + \mathbb{E}^2(X)] =$
$= \mathbb{E}(X^2) - \mathbb{E}^2(X) = \mathbb{E}[X(X-1)] + \mathbb{E}(X) - \mathbb{E}^2(X) =$
$= \underline{G''(1) + G'(1) - (G'(1))^2}$.

Theorem

$G_{X+Y}(z) = G_X(z) \cdot G_Y(z)$ for independent X and Y

Proof: $\underline{G_X(z) \cdot G_Y(z)} = \mathbb{E}(z^X) \cdot \mathbb{E}(z^Y) = \mathbb{E}(z^{X+Y}) = \underline{G_{X+Y}(z)}$.

Hence,

$\mathbb{E}(X+Y) = G'_X(1) \cdot G_Y(1) + G_X(1) \cdot G'_Y(1) = \mathbb{E}(X) + \mathbb{E}(Y)$.

$Var(X+Y) = \dots$

(1.2) Limits and inequalities

Topic structure:

1.) Moment generating functions
2.) Probabilistic inequalities
3.) Limit theorems

1.2.1 Moment generating functions

Def For a RV X define the moment generating function as:

$$M_X(t) = \mathbb{E}(e^{tX})$$

for those $t \in \mathbb{R}$ for which the expectation exists.

Properties: 1.) $M_X(0) = 1$.

2.) If X has MGF $M_X(t)$ and $Y = aX + b$, then

$$M_Y(t) = \mathbb{E}(e^{t(aX+b)}) = e^{bt} \cdot M_X(ta).$$

3.) If X and Y are independent RVs then

$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX} \cdot e^{tY}) = \mathbb{E}(e^{tX}) \cdot \mathbb{E}(e^{tY}) = M_X(t) \cdot M_Y(t).$$

4.) Since $M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left(1 + tX + \frac{(tX)^2}{2!} + \dots\right) = 1 + t \cdot \mathbb{E}(X) + \frac{t^2}{2!}\mathbb{E}(X^2) + \dots,$

$$\left(\frac{d^n}{dt^n} M_X(t)\right)\Big|_{t=0} = \mathbb{E}(X^n)$$

(n-th moment of X).

5.) If X is discrete RV $\in \{0, 1, 2 \dots\}$ with PGF $G_X(t)$, then

$$M_X(t) = \mathbb{E}(e^{tX}) = G_X(e^t).$$

6.) Uniqueness: same MGF $\Leftrightarrow$ same distribution function

7.) Continuity: $\left[\lim_{n\to\infty} F_n(X) = F(X)\right] \Leftrightarrow \left[\lim_{n\to\infty} M_n(X) = M(X)\right]$

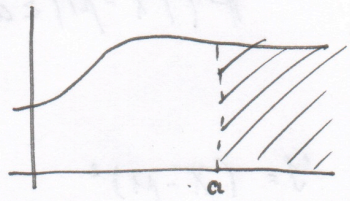
where $F_n(X)$ – probability distrib. functions

$M_n(X)$ – corresponding MGFs

1.2.2 Probabilistic inequalities

It is generally useful to have upper bounds on probabilities of the form:

$$\mathbb{P}(X \geq a)$$



We'll consider 3 such inequalities:

- Markov's inequality
- Chebyshev's inequality
- Chernoff's inequality

Useful concept:

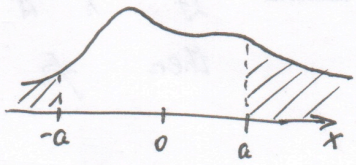
$I(A) = \begin{cases} 1, & \text{if } A \text{ happens} \\ 0, & \text{otherwise} \end{cases}$ (indicator)

Note: $\mathbb{E}(I(A)) = P(A) \cdot 1 + P(\bar{A}) \cdot 0 = P(A)$

1.) Markov's inequality

Theorem If $\mathbb{E}(X) < \infty$, then for any $a > 0$

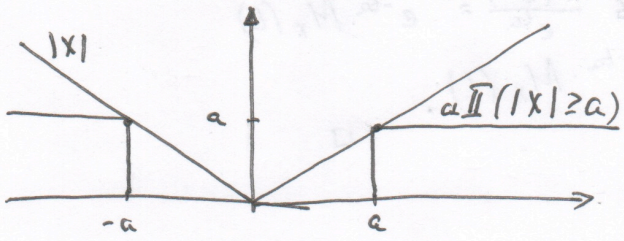
$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}$$



Proof

$$\mathbb{I}(|X| \geq a) = \begin{cases} 1, & \text{if } |X| \geq a \\ 0, & \text{if } -a < x < a \end{cases}$$

Consider the graphs of $|X|$ and $a\mathbb{I}(|X| \geq a)$.



From the drawing, it follows that

$$|X| \geq a \cdot \mathbb{I}(|X| \geq a)$$

Taking expectations of both sides:

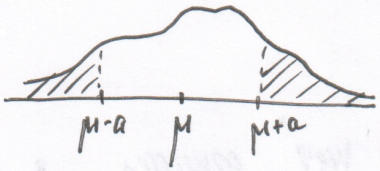
$$\mathbb{E}(|X|) \geq a \cdot \mathbb{P}(|X| \geq a)$$

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

2.) Chebyshev's inequality

Theorem If X is a RV with mean μ and finite variance σ^2, then for all $a>0$

$$\mathbb{P}(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}$$



Proof Let $Y = (X-\mu)^2$

Then $\mathbb{E}(Y) = \mathbb{E}((X-\mu)^2) = Var(X) = \sigma^2$

By Markov: $\mathbb{P}(|Y| \geq b) \leq \frac{\mathbb{E}(|Y|)}{b}$

$$\mathbb{P}((X-\mu)^2 \geq b) \leq \frac{\sigma^2}{b}$$

Now let $b = a^2$, $a>0$

$$\mathbb{P}((X-\mu)^2 \geq a^2) = \mathbb{P}(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}$$ □

3.) Chernoff's inequality

Theorem If X is defined by MGF $M_X(t)$ and $a \in \mathbb{R}$, then for all $t>0$

$$\mathbb{P}(X \geq a) \leq e^{-ta} \cdot M_X(t).$$

Note: we can choose $t>0$ to make the bound as tight as possible

Proof $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta})$

By Markov: $\mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}} = e^{-ta} \cdot M_X(t)$

So $\mathbb{P}(X \geq a) \leq e^{-ta} \cdot M_X(t)$.

□

1.3.3 Limit theorems

Topic structure

1.) convergence

2.) Weak Law of Large Numbers (WLLN)

3.) Central Limit Theorem (CLT)

- Application: calculation of confidence intervals

1.) Convergence

For a sequence of random variables $(X_n)_{n\geq 1}$ we shall define two notions of convergence* to some RV X as $n\to\infty$:

a.) convergence in distribution ($X_n \xrightarrow{D} X$)

Def: $X_n \xrightarrow{D} X \iff F_{X_n} \to F_X$ as $n\to\infty$.

distributions of X_n become increasingly close to that of X

b.) convergence in probability ($X_n \xrightarrow{P} X$)

Def: $X_n \xrightarrow{P} X \iff \mathbb{P}(|X_n - X| < \varepsilon) \to 0$ as $n\to\infty$ for all $\varepsilon > 0$.

Theorem

convergence in probability implies convergence in distribution

(convergence in probability is a stronger statement)

Proof $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$.

Consider the cumulative distribution function of X_n

$$F_{X_n}(a) \equiv \mathbb{P}(X_n \leq a)$$

Now

$$F_{X_n}(a) = \mathbb{P}(X_n \leq a) = \mathbb{P}(X_n \leq a \wedge X > a+\varepsilon) + \mathbb{P}(X_n \leq a \wedge X \leq a+\varepsilon) \quad \text{(I)}$$

for an arbitrary fixed $\varepsilon > 0$.

If $\begin{cases} X_n \leq a \\ X > a+\varepsilon \end{cases} \iff \begin{cases} -X_n \geq -a \\ X > a+\varepsilon \end{cases}$, then $X - X_n > \varepsilon \Rightarrow |X - X_n| > \varepsilon$. (II)

Also: $X_n \leq a \wedge X \leq a+\varepsilon \Rightarrow X \leq a+\varepsilon$, so $\mathbb{P}(X_n \leq a \wedge X \leq a+\varepsilon) \leq \mathbb{P}(X \leq a+\varepsilon)$ (III)

Putting (II) and (III) into (I), we get

$$F_{X_n}(a) \leq \mathbb{P}(|X - X_n| > \varepsilon) + \mathbb{P}(X \leq a+\varepsilon) = \mathbb{P}(|X - X_n| > \varepsilon) + F_X(a+\varepsilon).$$

Similarly, $F_{X_n}(a) \geq \mathbb{P}(X \leq a-\varepsilon) - \mathbb{P}(|X - X_n| > \varepsilon)$

Similarly, $F_{X_n}(a-\varepsilon) = \mathbb{P}(X \le a-\varepsilon \wedge X_n > a) + \mathbb{P}(X \le a-\varepsilon \wedge X_n \le a) \le$

$\le \mathbb{P}(|X_n - X| > \varepsilon) + F_{X_n}(a)$

Using the two inequalities, we can sandwich $F_{X_n}(a)$:

$$F_X(a-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \le F_{X_n} \le \mathbb{P}(|X_n - X| > \varepsilon) + F_X(a+\varepsilon)$$

Since $X_n \xrightarrow{P} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ (by definition),

so

$$F_X(a-\varepsilon) \le F_{X_n}(a) \le F_X(a+\varepsilon) \quad \text{for an arbitrary small } \varepsilon$$

So $F_{X_n} \to F_X$, which is the definition of <u>convergence in dist.</u>

□

<u>Limit theorems</u>

<u>Idea:</u> Given a sequence of RVs $(X_n)_{n \ge 1}$, let

$$S_n = X_1 + X_2 + X_3 + \dots + X_n \quad \text{and} \quad \bar{X}_n = \frac{S_n}{n}$$

What happens to the <u>sample average</u> $\bar{X}_n$ for a large n?

a.) The Weak Law of Large Numbers (WLLN)

<u>Theorem</u> If $(X_n)_{n \ge 1}$ is a sequence of IID RVs with finite mean μ, then $\bar{X}_n \xrightarrow{P} \mu$.

<u>Proof</u>

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \mathrm{Var}\left(\frac{X_1}{n}\right) + \mathrm{Var}\left(\frac{X_2}{n}\right) + \dots + \mathrm{Var}\left(\frac{X_n}{n}\right) =$$

$$= \underbrace{\frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \dots + \frac{\sigma^2}{n^2}}_{n \text{ terms}} = \frac{\sigma^2}{n}$$

By Chebyshev's inequality:

$$0 \le \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n \cdot \varepsilon^2}, \quad \text{for an arbitrary small } \varepsilon$$

Hence, $\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) = 0$, so $\bar{X}_n \xrightarrow{P} \mu$ for large n.

□

3.) Central Limit Theorem (CLT)

Theorem For a sequence of IID RVs $(X_n)_{n\geq 1}$ with mean μ and variance σ^2, whose MGF converges in some interval $-a<t<a$ $(a>0)$,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)$$

What it means: Given that some conditions hold, the sample average $(\bar{X}_n)$ is distributed roughly normally (in a bell curve). That is, if we compute the averages of many samples and then plot them, we'll get the familiar bell curve.

Proof Let $Y_i = \frac{X_i - \mu}{\sigma}$, then

$$\mathbb{E}(Y_i) = \mathbb{E}\left(\frac{X_i-\mu}{\sigma}\right) = \frac{\mathbb{E}(X_i)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0.$$

$$\mathbb{E}(Y_i^2) = \mathbb{E}\left[(Y_i - \cancelto{0}{\mathbb{E}(Y_i)})^2\right] = \mathrm{Var}(Y_i) = 1$$

Hence, the MGF:

$$M_{Y_i}(t) = 1 + \frac{t^2}{2} + \underbrace{o(t^2)}_{\text{higher order terms}}$$

Now,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{n} - \frac{\mu}{n}\right)}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = \frac{1}{\sqrt{n}}\cdot\sum_{i=1}^{n} Y_i$$

Combining properties (2) and (3) of MGFs:

$$M_{Z_n}(t) = \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n$$

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \cancelto{0}{o\left(\frac{t^2}{n}\right)}\right)^n = e^{\frac{t^2}{2}}$$

which is exactly the MGF of $N(0,1)$.

□

Applications of CLT: Confidence Intervals

Confidence interval — interval estimate of some unknown value, accurate with some known, parametrically controllable probability.

Theory

By CLT

$$Z_n = \frac{\bar{X}_n - M}{\sigma/\sqrt{n}} \qquad (I)$$

is asymptotically distributed as $N(0,1)$.

In real life, we can't know the true value of the variance σ^2, so instead we use the so-called sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

We can do it because we can show that $\mathbb{E}(S^2) = \sigma^2$.

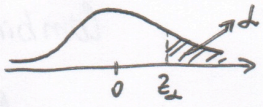
So, instead of (I), we use

$$Z_n = \frac{\bar{X}_n - M}{S/\sqrt{n}}$$

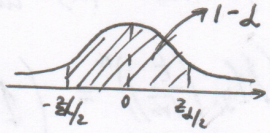
which is also approximately distributed as $N(0,1)$. (II)

Now let's define a z_α, so that $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim N(0,1)$

(values of z_α can usually be found in a table)



Then $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$



From (II), we can substitute Z_n since it is approximately distributed as $N(0,1)$:

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1-\alpha$$

So the true mean μ lies within the interval $\bar{X}_n \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$ with probability $1-\alpha$. We can vary α to make interval tighter or increase coverage probability.

(1.3) Stohastic processes

Topic structure:

1.) Random walks
   a. Intro and basic concepts
   b. Returning to starting state
   c. Gambler's ruin problem
   d. [redacted]
2.) Markov chains
   a. Basic definitions, concepts and theorems
   b. [redacted]
   b. Hidden Markov Models (HMMs) and associated problems

1.3.1. Random Walks

a.) Introduction and basic concepts

Consider a sequence $Y_1, Y_2 \ldots$ of IID RVs such that:
$\mathbb{P}(Y_i = 1) = p$, $\mathbb{P}(Y_i = -1) = 1-p$ for $p \in [0;1]$.
(so Y_i only takes values $+1$ or -1 with probabilities p and $1-p$).

Definition A simple random walk is a sequence of RVs $\{X_n \mid n \in \mathbb{Z}^+\}$ defined by

$$X_n = X_0 + \sum_{i=1}^{n} Y_i$$

where X_0 is the starting value

(so $\sum_{i=1}^{n} Y_i$ is the sum of several ± 1 steps)

Definition A simple symmetric random walk is that where Y_i is equally likely to be $+1$ and -1 (so $p = \frac{1}{2}$).

b.) Returning to the starting state

What is the probability of returning after exactly n steps?

$$\mathbb{P}(X_n = X_0) = \begin{cases} 0, & n \text{ is odd} \\ \binom{2m}{m} p^m (1-p)^m \end{cases}$$

Need the same number (m) of up- and down-steps.

What is the probability of ever returning to the starting state?

A simple random walk is called recurrent if it is certain to come back to its starting state at some time in the future. Otherwise, it is called transient.

Theorem

For a simple random walk starting at $X_0 = 0$, the probability of ever revisiting the starting state is:

$$\mathbb{P}(X_n = 0 \text{ for some } n \in \mathbb{Z}^+) = 1 - |\mathbb{E}(Y_i)| = 1 - |2p - 1|.$$

Interesting fact

$\mathbb{E}(Y_i) = 1 \cdot p - 1(1-p) = 2p - 1$

if $p > \frac{1}{2}$, $\mathbb{E}(Y_i) > 0 \rightarrow$ net drift upwards

if $p < \frac{1}{2}$, $\mathbb{E}(Y_i) < 0 \rightarrow$ net drift downwards

if $p = \frac{1}{2}$, $\mathbb{E}(Y_i) = 0 \rightarrow$ no net drift

Note: this implies that a simple random walk can only be recurrent if it is symmetric ($p = \frac{1}{2}$), and it is transient if $p \neq \frac{1}{2}$.

Proof:

Without loss of generality, assume that $X_0 = 0$. Define event $R_n = \{X_n = 0\}$, that is that the simple random walk returns to X_0 at time n.

Now consider event:

$$F_n = \{X_n = 0, X_m \neq 0 \ \forall m. m \in \mathbb{Z}^+, m < n\}$$

which is that the random walk *first* revisits its starting state at time n.

Then:

$$\mathbb{P}(R_n) = \sum_{m=1}^{n} \mathbb{P}(R_n \cap F_m)$$

but

$$\mathbb{P}(R_n \cap F_m) = \mathbb{P}(F_m) \cdot \mathbb{P}(R_{n-m})$$

So $\mathbb{P}(R_n) = \sum_{m=1}^{n} \mathbb{P}(F_m) \cdot \mathbb{P}(R_{n-m})$, or, more briefly:

$$r_n = \sum_{m=1}^{n} f_m \cdot r_{n-m} \quad (I)$$

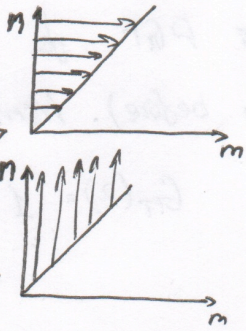
Now let's consider the probability generating functions for r_n and f_n:

$$R(z) = \sum_{n=0}^{\infty} r_n z^n \qquad F(z) = \sum_{n=0}^{\infty} f_n z^n \qquad |z| < 1$$

$r_0 = 1$ (we start at X_0) $\qquad f_0 = 0$ $\qquad \downarrow$ to make the series converge

Now let's derive an expression for $F(z)$ in terms of $R(z)$:

$$R(z) - \cancel{r_0 z^0}^{1} = \sum_{n=1}^{\infty} r_n z^n \overset{\text{by (I)}}{=} \underbrace{\sum_{n=1}^{\infty} \sum_{m=1}^{n}}_{\text{summing over the same range of m's and n's}} f_m r_{n-m} z^n =$$



$$= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} f_m z^m r_{n-m} \cdot z^{n-m} =$$

$$= \left| \begin{array}{l} \text{substitution} \\ k = n-m \end{array} \right| \sum_{m=1}^{\infty} f_m z^m \sum_{k=0}^{\infty} r_k z^k = \left(F(z) - \cancel{f_0 z^0}^{0}\right) \cdot R(z).$$

So $\quad R(z) - 1 = F(z) \cdot R(z)$

$$F(z) = 1 - R^{-1}(z). \quad (\text{II}).$$

Now let's explicitly calculate $R(z)$:

$$R(z) = \sum_{n=0}^{\infty} r_n \cdot z^n = \sum_{m=0}^{\infty} r_{2m} z^{2m} = \quad (\text{since } r_{2m+1} = 0 \ \ \forall m \in \mathbb{Z}^{\geq 0}).$$

$$= \sum_{m=0}^{\infty} \binom{2m}{m} p^m (1-p)^m z^{2m} =$$

$$= \sum_{m=0}^{\infty} \binom{2m}{m} \left(p(1-p)z^2\right)^m =$$

$$= \left(1 - 4p(1-p)z^2\right)^{-\frac{1}{2}}$$

(can expand it to see it gives same series as previous line)

By (II):

$$F = 1 - \left(1 - 4p(1-p)z^2\right)^{\frac{1}{2}}.$$

Finally:

$$\mathbb{P}(X_n = 0, \text{ for some } n \in \mathbb{Z}^+) = \mathbb{P}(\mathcal{F}_1 \vee \mathcal{F}_2 \vee \mathcal{F}_3 \dots) = f_1 + f_2 + f_3 + \dots =$$

$$= \lim_{z \to 1^-} \sum_{n=1}^{\infty} f_n z^n = F(1) = 1 - \left((2p-1)^2\right)^{\frac{1}{2}} = 1 - |2p-1|$$

□.

Mean return time

Suppose $p=\frac{1}{2}$, so the simple random walk is recurrent.

Let's say:

$$T = \min\{n \geq 1 \mid X_n = 0\}$$ (time of first return)

Then

$$\mathbb{P}(T=n) = f_n$$

So the PGF for T is the same as that for F (in theorem before). Namely:

$$G_T(z) = 1 - (1-4p(1-p)z^2)^{\frac{1}{2}} = 1 - (1-z^2)^{\frac{1}{2}} \quad |z|<1$$

(since $p=\frac{1}{2}$, we are considering the symmetric case)

By definition of PGF, the mean of return time T:

$$\mathbb{E}(T) = \frac{d}{dz}G_T(z)\Big|_{z=1} = \lim_{z\to 1^-} z(1-z^2)^{-\frac{1}{2}} = \infty$$

→ since we said $|z|<1$.

So the symmetric simple random walk is certain to return to starting state, but the mean return time is infinite. Such random walks are called null-recurrent.

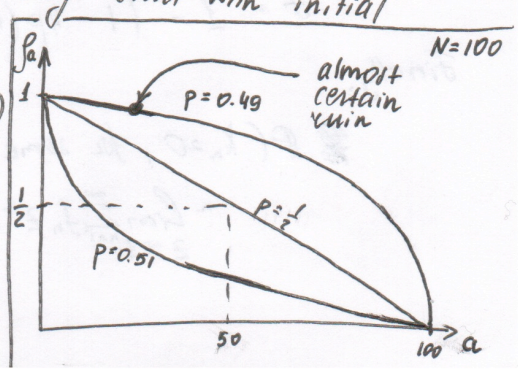
c.) The Gambler's ruin problem

Problem: Two players (A and B) with joint capital between them £N. Player A starts with £a ($0 \leq a \leq N$) of capital and at each time step either receives £1 from B with probability p, or gives £1 to B with probability $q=1-p$. What is the probability of ruin (losing the whole capital) for Player A?

Theorem: The probability of ruin for A when they start with initial capital a is given by:

$$f_a = \begin{cases} \frac{\theta^a - \theta^N}{1-\theta^N} & \text{if } p \neq q \text{ (i.e. } p \neq \frac{1}{2}) \\ 1 - \frac{a}{N} & \text{if } p = q \text{ (i.e. } p = \frac{1}{2}) \end{cases}$$

where $\theta = \frac{q}{p}$.



Proof

$$f_a = P(y_1 = +1)\cdot f_{a+1} + P(y_1 = -1)\cdot f_{a-1} = p\cdot f_{a+1} + q\cdot f_{a-1}$$

The above is called a difference equation, and is solved by methods similar to those used for differential equations.

Try a solution of the form $f_a = \lambda^a$:

$$\lambda^a = p\cdot\lambda^{a+1} + q\lambda^{a-1}$$

$$p\lambda^2 - \lambda + q = 0.$$

$$\lambda = 1 \quad \text{or} \quad \lambda = \frac{q}{p}.$$

if $p \neq q$

general solution of the form $A + B\cdot\left(\frac{q}{p}\right)^a$

Applying boundary conditions

$f_0 = 1$ $\quad f_N = 0$

We get

$$f_a = \frac{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}$$

if $p = q$

general solution of the form $C + Da$

Applying boundary conditions

$f_0 = 1$ and $f_N = 0$

we get

$$f_a = 1 - \frac{a}{N}.$$

Mean duration time

How long (on average) will it take before one of the players gets ruined? Let μ_a be mean duration time for the case when A starts with $£a$. The, similarly to above:

$$\mu_a = 1 + p\cdot\mu_{a+1} + q\cdot\mu_{a-1}$$

Using methods similar to above it can be now shown that:

$$\mu_a = \begin{cases} \frac{1}{p-q}\left(N\frac{(q/p)^a - 1}{(q/p)^N - 1} - a\right) & \text{if } p \neq q \\ a(N-a) & \text{if } p = q \end{cases}$$

1.3.2. Markov Chains (MCs)

a.) Basic definitions, concepts and theorems

Definition

A Markov chain is a sequence of discrete RVs taking values in some countable state space S with the property:

$$P(X_n = a_n \mid X_0 = a_0, X_1 = a_1, \dots, X_{n-1} = a_{n-1}) = P(X_n = a_n \mid X_{n-1} = a_{n-1}).$$

→ this means that the next state is only determined by the current state and not the previous ones (MCs are memoryless)

Since S is countable, we usually map its elements to integers and "mark" the states with those.

The dynamics of a Markov chain is determined by transition probabilities of the form $P(X_n = i \mid X_{n-1} = j)$. Since we usually want to know the probabilities of transition between any 2 states in the MC, it is convenient to define the transition matrix P the elements of which are:

$$(P)_{ij} = p_{ij} = P(X_n = j \mid X_{n-1} = i)$$

where i, j are in S.

row – current state
column – next state

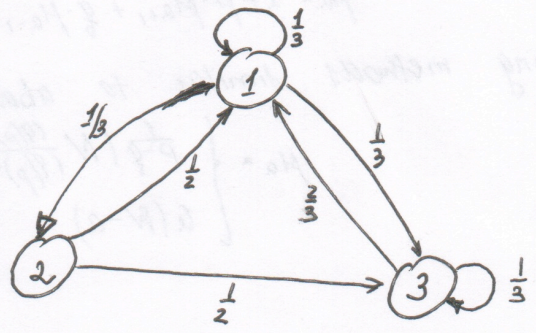
This matrix P completely defines the behaviour of the MC.

P is called a stohastic matrix (it contains probabilities), so the sum of each of its rows is 1 (total probability of going anywhere).

If the probabilities in the matrix do not depend on time, the MC is said to be time-homogeneous.

Example $S = \{1, 2, 3\}$

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



We also define the n-step transition matrix $P^{(n)} = (p_{ij}^{(n)})$ where

$$p_{ij}^{(n)} = P(X_n = j \mid X_0 = i)$$

so it consists of probabilities of moving from i to j in exactly n steps.

Theorem (Chapman-Kolmogorov)

For all states i, j and steps m, n:

$$p_{ij}^{(m+n)} = \underbrace{\sum_k p_{ik}^{(m)} \cdot p_{kj}}_{\text{matrix multiplication}}$$

Equivalently, $P^{(m+n)} = P^{(m)} \cdot P^{(n)}$

$P^{(n)} = P^n$ (we can multiply matrices together to obtain the n-step transition matrix)

Proof

$$p_{ij}^{(m+n)} = \mathbb{P}(X_{m+n} = j \mid X_0 = i) = \sum_k \mathbb{P}(X_{m+n} = j, X_m = k \mid X_0 = i) =$$

(introducing a state in the middle)

$$= \sum_k \mathbb{P}(X_{m+n} = j \mid X_m = k \cancel{,} X_0 = i) \cdot \mathbb{P}(X_m = k \mid X_0 = i) =$$

$$= \sum_k \mathbb{P}(X_{m+n} = j \mid X_m = k) \cdot \mathbb{P}(X_m = k \mid X_0 = i) = \sum_k p_{ik}^{(n)} \cdot p_{kj}^{(m)}.$$

Related Lemma

If $\lambda^{(n)}$ is a row vector containing probabilities of being in each state after n steps, then

$$\lambda^{(m+n)} = \lambda^{(m)} \cdot P^{(n)}$$

Proof:

$$\lambda_j^{(m+n)} = \mathbb{P}(X_{m+n} = j) = \sum_i \mathbb{P}(X_{m+n} = j \mid X_m = i) \cdot \mathbb{P}(X_m = i) =$$

$$= \sum_i \mathbb{P}(X_n = j \mid X_0 = i) \cdot \mathbb{P}(X_m = i) = \sum_i p_{ij}^{(n)} \cdot \lambda_i^{(m)} = \left(\lambda^{(m)} \cdot P^{(n)}\right)_j.$$

Classification of states

Definition

1.) j is *accessible* from i $(i \rightsquigarrow j)$ iff $\exists n \geq 0.\ p_{ij}^{(n)} > 0$

2.) i and j *communicate* $(i \leftrightsquigarrow j)$ iff $(i \rightsquigarrow j) \wedge (j \rightsquigarrow i)$

The communicates relation $(\leftrightsquigarrow)$ is reflexive, symmetric and transitive, so it is an equivalence relation.

We can partition the state space into subsets in which every state communicates with all others. These subsets are called *communicating classes*.

Definitions

1.) A communicating class which, once entered, cannot be left is called closed.

2.) A closed communicating class containing one state is called absorbing.

3.) When the state space forms a single communicating class (can get to every state from every state), the Markov chain is called irreducible. Otherwise, it's reducible.

Recurrence and transience of MCs

Let's define the matrix $F^{(n)} = (f_{ij}^{(n)})$ where

$$f_{ij}^{(n)} = \mathbb{P}(X_1 \neq j, X_2 \neq j, \dots, X_n = j \mid X_0 = i)$$

so $f_{ij}^{(n)}$ is the probability of visiting state j for the first time at n^{th} step, after starting from state i.

Then we can say that

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

is the probability of ever visiting j starting from i.

If $f_{ii} < 1$, the state i is transient
(meaning that we are not guaranteed to even revisit it after visiting once)

If $f_{ii} = 1$, the state i is recurrent
(meaning that we are certain to revisit it)

Since MCs are memoryless, the probability of coming back to state i at least N times is $\underbrace{f_{ii} f_{ii} f_{ii} \cdots f_{ii}}_{N \text{ terms}} = f_{ii}^N$ (because each return is an independent event). Then the probability of coming back to i an infinite number of times is:

$$\mathbb{P}_\infty = \begin{cases} 1^\infty = 1, & \text{if } f_{ii} = 1 \\ f_{ii}^\infty = 0, & \text{if } f_{ii} < 0. \end{cases}$$

State i is transient $\iff \sum_{n=1}^{\infty} p_{ii}^{(n)}$ converges

State i is recurrent $\iff \sum_{n=1}^{\infty} p_{ii}^{(n)}$ diverges

(can be used as a criterion for recurrence/transience)

Theorem

If i and j belong to the same communicating class, then they are either both recurrent or both transient. (the solidarity property)

Proof — Slide #5 in the notes.

Mean recurrence time

Let's say

$$T_j = \min\{n \geq 1 : X_n = j\}$$

is the first time we visit j. $T_j = \infty$ if it never happens.

The mean recurrence time:

$$\mu_i = \mathbb{E}(T_i \mid X_0 = i) = \sum_{n=1}^{\infty} n \cdot \mathbb{P}(T_i = n \mid X_0 = i) = \begin{cases} \sum_n n \cdot f_{ii}^{(n)}, & \text{if } i \text{ is recurrent} \\ \infty \cdot \underbrace{\mathbb{P}(T_i = \infty \mid X_0 = i)}_{>0} + \ldots = \infty, & \text{if } i \text{ is transient} \end{cases}$$

NB: $\sum_n n \cdot f_{ii}^{(n)}$ may not converge, so the mean recurrence time may be infinite, even if i is recurrent.

To summarise, there are 3 possible cases. A state i can be:

1.) Transient — $f_{ii} < 1$, $\mu_i = \infty$

2.) Null-recurrent — $f_{ii} = 1$, $\mu_i = \infty$

3.) Positive recurrent — $f_{ii} = 1$, $\mu_i < \infty$ } recurrent

Periodicity

Let

$$d_i = \text{GCD}\{n : p_{ii}^{(n)} > 0\}$$

if $d_i = 1$, the state is aperiodic

else if $d_i > 1$, the state is periodic

It can be shown that all states in a communicating class have same period.

From here onwards, we only consider irreducible, aperiodic Markov chains (i.e. consisting of a single communicating class with aperiodic states).

Stationary distribution

Stationary (a.k.a. equilibrium) distribution is basically:

$$\pi = \lim_{n\to\infty} \lambda^{(n)} \qquad \text{(if the limit exists)}$$

(i.e. the the long-term probability distribution between states).

As such, it has the following properties:

1. $\pi_i \geq 0 \quad \forall i \in S$ (probabilities are not negative)
2. $\sum_{i\in S} \pi_i = 1$ (total prob. of being anywhere)
3. $\pi = \pi P$ (1, or any number, further transition doesn't make a difference)
   ↳ $\pi P^n = \pi, \ \forall n.$

Theorem (Erdős-Feller-Pollard)

For all states i, j in an irreducible, aperiodic MC:

1.) if the chain is transient, $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ (or $\lim_{n\to\infty} P^n = Z$ ↳ zero-matrix)

2.) if the chain is recurrent, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ (each row of P converges to π)

2.1.) if chain is null-recurrent, $\pi = 0$

2.2.) if chain is positive-recurrent, every $\pi_i > 0$, $\sum_{i\in S} \pi = 1$

π is the unique solution of $\pi P = \pi$

In case (2), mean time of return to i:

$$\mu_i = \frac{1}{\pi_i}. \quad \text{(note: } \mu_i = \infty \text{ if } \pi_i = 0\text{)}.$$

Proof: non-examinable.

Time-reversibility

We can define a reversed chain by:

$$Y_n = X_{-n} \quad \text{for} \quad -\infty < n < \infty$$

(Y_n) is also an MC with same stationary distribution.

Definition:

A MC is reversible if the transition matrices of (X_n) and (Y_n) are equal.

Theorem

A MC is reversible iff $\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i,j \in S$

Proof:

Suppose $Q = (q_{ij})$ is the transition matrix of (Y_n).

Then

$$q_{ij} = \mathbb{P}(Y_{n+1} = j \mid Y_n = i) = \mathbb{P}(X_{-n-1} = j \mid X_{-n} = i) = \text{(by Bayes' theorem)}$$

$$= \frac{\mathbb{P}(X_{-n} = i \mid X_{-n-1} = j) \cdot \mathbb{P}(X_{-n-1} = j)}{\mathbb{P}(X_{-n} = i)} = \frac{p_{ji} \cdot \pi_j}{\pi_i}$$

For MC to be reversible, we need $p_{ij} = q_{ij}$

$$p_{ij} = \frac{p_{ji}\pi_j}{\pi_i} \Longrightarrow \underline{p_{ij}\pi_i = p_{ji}\pi_j} \quad \square$$

Theorem

For an irreducible MC, if there exists a vector π such that:

1) $0 \le \pi_i \le 1$ and $\sum_{i \in S} \pi_i = 1$
2) $\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i,j \in S$ (local balance condition)

then the MC is reversible with stationary distribution π.

Proof:

$$(\pi P)_j = \sum_{i \in S} \pi_i p_{ij} = \sum_{i \in S} \pi_j p_{ji} = \pi_j \underbrace{\sum_{i \in S} p_{ji}}_{=1} = \pi_j$$

so $\pi P = \pi$ and distribution π is stationary. $\square$

Ergodic results

Let's consider the proportion of time we spend in a state i.

I.e. we visit state i $V_i(n)$ times before time n, then:

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{I}(\{X_k = i\}).$$

Then $\frac{V_i(n)}{n}$ is the proportion of time we spend in state i before time n.

Theorem (Ergodic theorem)

For an irreducible MC (X_n) with transition matrix P

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{V_i(n)}{n} = \frac{1}{\mu_i} = \pi_i\right) = 1.$$

where μ_i is expected return time to state i.

Proof: in the notes.

[redacted]

b.) Hidden Markov Models (HMMs)

Sometimes statistical model of observed data is constructed from an underlying but hidden MC. Can be useful in: OCR, Nat. language processing etc.

HMMs: We have an MC with transition matrix P, but the [redacted] current state of the MC is not directly observable. Instead we can observe a randomly chosen (distribution is given) token Y_n. For each state there is a distribution b_i, such that $(b_i)_t = \mathbb{P}(Y_n = t \mid X_n = i)$.

Three central problems:

Evaluation

Given an observed sequence of tokens, determine its probability given the HMM parameters.

$\mathbb{P}(Y_1 = a_1, Y_2 = a_2, Y_3 = a_3 \ldots \mid \text{HMM params})$

In practice, solved by forward algorithm.

Decoding

Given a sequence of observed tokens and HMM parameters, determine the best-fitting sequence of hidden underlying states.

In practice, solved by Viterbi algorithm.

Learning

Given a sequence of tokens, determine the HMM parameters to maximise its probability.

$\mathbb{P}(Y_1 = a_1, \ldots, Y_n = a_n \mid \text{HMM params})$.

Solved by the Baum-Welch iterative method.