Stochastic Processes: Martingales

Math 606, Spring 2024

Luc Rey-Bellet
University of Massachusetts Amherst

2025-05-01

Table of contents

1 Conditional Expectation

1.1 Motivation and definition

- Martingales are models of fair games and to understand then we need to understand first conditional expectations. Conditional expectations is a very useful concept to understand how information obtained from measurement can be incorporated to make predictions.

- Suppose we are given a random variable Y. If we know nothing about the outcome of experiment generating Y then our best guess for the value of Y would be the expectation $E[Y]$. On the contrary if we measure Y itself then our prediction would Y itself! Conditional expectations deals with making best guesses on the possible value of Y when we have some partial information which is described by some collection of other random variables $X_1, X_2, \cdots, X_n$.

- Example: discrete RV: Suppose X and Y are discrete random variables with joint density and marginals

$$\text{joint pdf} \quad p(x,y) = P(X=x, Y=y) \quad \text{marginals} \quad p(x) = \sum_y p(x,y)\,, \quad p(y) = \sum_x p(x,y)$$

To define the conditional expectation $E[Y|X]$ we need to give the best guess for Y given that we have observed $X = x$ which is

$$E[Y|X=x] = \sum_y y P(Y=y|X=x) = \sum_y y \, \frac{P(Y=y, X=x)}{P(X=x)} = \sum_y y \, \frac{p(x,y)}{p(x)}$$

which is well defined for those x with $p(x) > 0$.

- The function $E[Y|X=x]$ defines a function of the random variable X which we define to be $E[Y|X]$. For example if we roll two independent dice and X is value of the first roll and Y the sum of the two rolls
  then we have

$$f(x,y) = \frac{1}{36}, \quad x = 1, 2, \cdots, 6 \quad y = x+1, \cdots, x+6.$$

  and this

$$E[Y|X=x] = x + \frac{7}{2}$$

  so that $E[Y|X] = X + \frac{7}{2}$.

- In a similar way we can define $E[Y|X_1, \cdots, X_n]$ for discrete RV with joint pdf $p(x_1, \cdots, x_n, y)$.

Example: continuous RV: In a similar way, if $Y, X_1, \cdots X_n$ are continuous RV with a joint pdf $f(x_1, \cdots, x_n, y)$ with marginal $f(x_1, \cdots, x_n) = \int f(x_1, \cdots, x_n, y)dy$ then the function

$$y \mapsto \frac{f(x_1, \cdots, x_n, y)}{f(x_1, \cdots, x_n)}$$

defines a probability density function provided $(x_1, \cdots, x_n)$ is such that $f(x_1, \cdots, x_n) \neq 0$. Then the expectation

$$E[Y|X_1 = x_1 \cdots X_n = x_n] = \int y \frac{f(x_1, \cdots, x_n, y)}{f(x_1, \cdots, x_n)} \, dy$$

defines a function of $(x_1, \cdots, x_n)$. We leave this function undefined whenever $f(x_1, \cdots, x_n) = 0$ (or set it to 0 if you prefer).

Any function of $h(x_1, \cdots, x_n)$ can be used to define a RV $h(X_1, \cdots, X_n)$. Note that, as a RV, this does not depend on how the function is defined when $f(x_1, \cdots, x_n) = 0$ since such x have probability 0.

We call the corresponding random variable $E[Y|X_1, \cdots, X_n]$ and call it the conditional expectation of Y given $X_1, \cdots, X_n$.

The conditional expectation has the following properties

1. $E[Y|X_1 \cdots X_n]$ depend only on $X_1, \cdots, X_n$ in the sense that it is function $h(X_1, \cdots, X_n)$. In the language of measure theory $E[X|Y_1, \cdots, Y_n]$ is a measurable function with respect to $X_1, \cdots, X_n$, or better with respect to the σ-algebra generated by $X_1, \cdots, X_n$.

2. Suppose that A is an event which depends only on $X_1, \cdots, X_n$, for example the rectangle

$$A = \{a_i \leq X_i \leq b_i\,,\ i = 1, \cdots, n\}$$

and let 1_A be the corresponding indicator function. Then

$$E[Y1_A] = E[E[Y|X_1, \cdots, X_n]1_A]$$

To prove the second property note that

$$\begin{aligned}
E[E[Y|X_1, \cdots, X_n]1_A] &= \int E[Y|X_1 = x_1, \cdots, X_n = x_n]1_A(x_1, \cdots, x_n) f(x_1, \cdots, x_n) dx_1 \cdots dx_n \\
&= \int_A \left(\int y \frac{f(x_1, \cdots, x_n, y)}{f(x_1, \cdots, x_n)} dy \right) f(x_1, \cdots, x_n, y) dx_1 \cdots dx_n \\
&= \int_A \int y f(x_1, \cdots, x_n, y) dy dx_1 \cdots dx_n \\
&= E[Y1_A]
\end{aligned}$$

1.2 General defintion of the conditional expectation

What we have demonstrated with examples are instance of a general theorem. We use the notation $\mathcal{F}_n$ to denote all the information contained in the random variables $X_1, \cdots, X_n$. We say that a random variable Z is [measurable with respect to $\mathcal{F}_n$] if $Z = h(X_1, \cdots, X_n)$ can be expressed as a function of $(X_1, \cdots, X_n)$. We say that a set A is measurable with respect to $\mathcal{F}_n$ if the function 1_A is measurable with respect to $\mathcal{F}_n$. This simply means that A should be specifiued using the random variable X.

Theorem 1.1 Let Y and $X_1, \cdots, X_n$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and assume that $E[|Y|] < \infty$. Let $\mathcal{F}_n$ be the σ-algebra generated by $X_1, \cdots, X_n$. Then there exists a unique random variable $E[Y|X_1, \cdots, X_n]$ such that

1. $E[Y|X_1, \cdots, X_n]$ is measurable with respect to $\mathcal{F}_n$.
2. For any A measurable with respect to $\mathcal{F}_n$ we have

$$E[Y1_A] = E[E[Y|X_1, \cdots, X_n]1_A] \quad \text{for all } A \in \mathcal{F}_n\,.$$

A more geometric way to understand conditional expectation as an (orthogonal) projection is explored in the homework.

1.3 Properties of Conditional Expectations

From now we use the abbreviated notation for the conditional expectation

$$E[Y|\mathcal{F}_n] = E[Y|X_1, \cdots, X_n]$$

The conditional expectation has the following properties

Theorem 1.2 The conditional expectation has the following properties

1. Linearity: $E[a_1Y_1 + a_2Y_2|\mathcal{F}_n] = a_1E[Y_1|\mathcal{F}_n] + a_2E[Y_2|\mathcal{F}_n]$
2. If $Y = g(X_1, \cdots, X_n)$ then $E[Y|\mathcal{F}_n] = Y$
3. If Y is independent of $X_1, \cdots, X_n$ then $E[Y|\mathcal{F}_n] = E[Y]$
4. If $m < n$ then $E[E[Y|\mathcal{F}_n]|\mathcal{F}_m] = E[Y|\mathcal{F}_m]$.
5. If $Z = g(X_1, \cdots, X_n)$ then $E[YZ|\mathcal{F}_n] = ZE[Y|\mathcal{F}_n]$.

Proof. The idea is to use the uniqueness statement in Theorem 1.1.

For part 1. $E[Y_i|\mathcal{F}_n]$ are the unique $\mathcal{F}_n$ measurable random variables such that $E[Y_i 1_A] = E[E[Y_i|\mathcal{F}_n]1_A]$ and so by linearity

$$a_1 E[E[Y_1|\mathcal{F}_n]1_A] + a_2 E[E[Y_1|\mathcal{F}_n]1_A] = a_1 E[Y_1 1_A] + a_2 E[Y_2 1_A] = E[(a_1 Y_1 + a_2 Y_2)1_A]$$

and by uniqueness we must have $E[a_1 Y_1 + a_2 Y_2|\mathcal{F}_n] = a_1 E[Y_1|\mathcal{F}_n] + a_2 E[Y_2|\mathcal{F}_n]$.

For part 2. if $Y = g(X_1, \cdots, X_n)$ then Y itself satisfies the definition.

For part 3. if Y is independent of $X_1, \cdots, X_n$ then by independence and linearity

$$E[Y 1_A] = E[Y]E[1_A] = E[E[Y]1_A]$$

which, by uniqueness, proves the statement.

For part 4. note that $E[E[Y|\mathcal{F}_n]|\mathcal{F}_m]$ and $E[Y|\mathcal{F}_m]$ both depend only on $X_1, \cdots X_m$. Moreover if A is $\mathcal{F}_m$ measurable and $m \leq n$ then it is also $\mathcal{F}_n$ measurable. So we have

$$E[E[Y|\mathcal{F}_m]1_A] = E[Y 1_A] = E[E[Y|\mathcal{F}_n]1_A] = E[E[E[Y|\mathcal{F}_n]|\mathcal{F}_m]1_A]$$

which, by uniqueness, proves the statement.

Finally for 5. if $Z = 1_B$ and B is $\mathcal{F}_n$ measurable then

$$E[E[Y 1_B|\mathcal{F}_n]1_A] = E[Y 1_B 1_A] = E[Y 1_{A\cap B}] = E[E[Y|\mathcal{F}_n]]1_B 1_A]$$

which proves the statement. For general Z one use an approximation argument by simple functions.

1.4 Examples

Example Suppose X_i are IID random variables with $\mu = E[X_i]$ and let $S_n = X_1 + \cdots + X_n$. If we take $m < n$ then we have

$$\begin{aligned} E[S_n|\mathcal{F}_m] &= E[X_1 + \cdots + X_m|\mathcal{F}_m] + E[X_{m+1} + \cdots + X_n|\mathcal{F}_m] \\ &= X_1 + \cdots + X_m + E[X_{m+1} + \cdots + X_n] = S_m + (n-m)\mu \end{aligned}$$

since $X_1 + \cdots + X_m$ is $\mathcal{F}_m$ measurable and $X_{m+1} + \cdots + X_n$ is independent of $X_1, X_2, \cdots, X_m$ (see Theorem 1.2, properties 2 and 3.)

Example Let S_n as in the previous example and assume that $\mu = 0$ and let $\sigma^2 = V(X_i)$ be the variance. If we take $m < n$ we find, using properties 2. 3. and 5. of Theorem 1.2.

$$\begin{aligned} E[S_n^2|\mathcal{F}_m] &= E[(S_m + (S_n - S_m))^2|\mathcal{F}_m] \\ &= E[S_m^2|\mathcal{F}_m] + 2E[(S_n - S_m)S_m|\mathcal{F}_m] + E[(S_n - S_m)^2|\mathcal{F}_m] \\ &= S_m^2 + 2S_m E[(S_n - S_m)|\mathcal{F}_m] + E[(S_n - S_m)^2] \\ &+ S_m^2 + 2S_m E[S_n - S_m] + E[(S_n - S_m)^2] \\ &= S_m^2 + V(S_n - S_m) = S_m^2 + (n-m)\sigma^2 \end{aligned}$$

Example If X_i are Bernoulli RV and $m < n$ let us compute $E[S_m|S_n]$. Note first that

$$P(X_1 = 1|S_n = k) = \frac{p\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{k}{n} \implies E[X_1|S_n] = \frac{S_n}{n}$$

and thus

$$E[S_m|S_n] = \frac{m}{n}S_n\,.$$

1.5 Exercise

Exercise 1.1 (Conditional expectation as a projection)

1. Show the following: if $E[|Y|^2] < \infty$ (which ensures that all expectations exists) then $E[Y|\mathcal{F}_n]$ is is the random variable measurable with respect to $\mathcal{F}_n$ which minimizes the mean square error, that is it solves

$$\min\{E[(Y-Z)^2] \ : \ Z \ \mathcal{F}_n - \text{measurable}\}$$

   Hint: Write Z as $Z = E[Y|\mathcal{F}_n] + W$ and expand the square.

2. Suppose $E[Y^2] < \infty$ and $m \leq n$. Show that

$$E[(Y - E[Y|\mathcal{F}_n])^2] + E[(E[Y|\mathcal{F}_n] - E[Y|\mathcal{F}_m])^2] = E[(Y - E[Y|\mathcal{F}_m])^2]$$

   in particular $E[(Y - E[Y|\mathcal{F}_n])^2]$ is a decreasing function of n.

3. Suppose $E[Y^2] < \infty$ and define the conditional variance to be the random variable $V(Y|\mathcal{F}_n) = E[Y^2|\mathcal{F}_n] - E[Y|\mathcal{F}_n]^2$. Show that

$$V(Y) = E[V(Y|\mathcal{F}_n)] + V(E[Y|\mathcal{F}_n])$$

2 Martingales

2.1 Definition and simple examples

Definition 2.1 Consider a collection of random variables $X_1, X_2, \cdots$. A sequence of random variables $M_0, M_1, M_2, \cdots$ is called a Martingale with respect to the filtration $\mathcal{F}_n$ if

1. $E[|M_n|] < \infty$ for all n.
2. M_n is measurable with respect to $\mathcal{F}_n$
3. For each $m < n$ we have $E[M_n|\mathcal{F}_m] = M_m$

Remark To verify the martingale porperty 3. in Definition 2.1 it is enough to check that

$$E[M_{n+1}|\mathcal{F}_n] = M_n \text{ for all } n$$

since this property implies that, by item 4. in Theorem 1.1

$$E[M_{n+2}|\mathcal{F}_n] = E[E[M_{n+2}|\mathcal{F}_{n+1}]|\mathcal{F}_n] = E[M_{n+1}|\mathcal{F}_n] = M_n$$

and so on.

Example One of the prototype of a martingale is given by sum of IID random variables. Let

$$S_0 = 0\,, \quad S_n = X_1 + \cdots X_n$$

As we have seen before, if $m < n$,

$$E[S_n|\mathcal{F}_m] = S_m + (n-m)\mu$$

This implies that $M_n = S_n - n\mu$ is a martingale.

Example Suppose Y is a random variable with $E[|Y|] < \infty$. Then we can build a martingale with respect to $\mathcal{F}_n$ by setting

$$M_n = E[Y|\mathcal{F}_n]$$

Indeed we have, for $m \leq n$

$$E[M_n|\mathcal{F}_m] = E[E[Y|\mathcal{F}_n]\mathcal{F}_m] = E[Y|\mathcal{F}_m] = M_m$$

In that case we say that the martingale is closed by the random variable Y and we can think of M_n as succesively "better" approximation of Y as we incorporate more and more information.

2.2 Martingale and fair game

Suppose we are playing a sequence of independent fair game with two outcomes (e.g betting on the flip a fair coin.) We describe this by RV X_i such that

$$P(X_i = +1) = P(X_i = -1) = \frac{1}{2}$$

The RV describe the winning obtain by betting an amount of 1 on the i^{th} game and the game is fair since $E[X_i] = 0$.

A betting sequence is a sequence of RV B_n such that

1. B_n is the amount of money bet on the n^{th} game.
2. B_n is measurable with respect to $\mathcal{F}_{n-1}$.
3. $E[|B_n|] < \infty$.

The second property means that the way you bet on the n^{th} game is guided by the past outcomes of the $n-1$ previous bets. No peeking into the future allowed! The winnings after n games is given by

$$W_n = \sum_{k=1}^{n} B_k X_k \quad \text{with} W_0 = 0$$

and we show it is a martingale.

Clearly W_n is $\mathcal{F}_n$ measurable and $E[|W_n|] < \infty$. Moreover we have

$$E[W_{n+1}|\mathcal{F}_n] = E[B_{n+1}X_{n+1}|\mathcal{F}_n] + E[W_n|\mathcal{F}_n] = B_{n+1}E[X_{n+1}|\mathcal{F}_n] + W_n = W_n$$

where we have used that B_{n+1} is $\mathcal{F}_n$ measurable and $E[X_n] = 0$.

The martingale property implies that $E[W_n]$ is constant: we have

$$E[W_n] = E[E[W_n|\mathcal{F}_{n-1}]] = E[W_{n-1}]$$

which means that your expected winning in fair game is zero.

But this is not the end of the story. In a betting strategy you will do a sequence of bet and decide of a good moment when you actually stop betting. Consider for example the following well known strategy (often called the martingale strategy): double your bet until you win. If you win the first game you stop and take you gain of $W_1 = 1$. If you lose the first game you bet 2 on the second game. If you win the second game you winning is $W_2 = -1 + 2 = 1$ and then stop. If you lose the first two games you know bet 4 on the third game abnd if you win the third game you winning is $W_3 = -1 - 2 + 4 = 1$, and so on... We have then the transition probabilities $P(W_{n+1} = 1|W_n = 1) = 1$ and

$$P(W_{n+1} = 1|W_n = -(2^n - 1)) = \frac{1}{2} \qquad P(W_{n+1} = -(2^{n+1} - 1)|W_n = -(2^n - 1)) = \frac{1}{2}$$

and this is a martingale. It is true that $E[W_n] = 0$ but however when you stop, which happens at a random time T, you always win 1! The time T at which you first win occurs has here a geometric distribution. We will consider stopping time in the next section.

2.3 Polya Urn

Consider an urn with balls of two colors, say red and green. Assume that initially there is one ball of each color in the urn. At each time step, a ball is chosen at random from the urn. If a red ball is chosen, it is returned and in addition another red ball is added to the urn. Similarly, if a green ball si chosen, is is returned together with another green ball.

Let X_n denote the number of red balls in the urn after n draws. Then $X_0 = 1$ and X_n a (time-inhomogeneous) Markov chain with transitions

$$P(X_{n+1} = k+1 | X_n = k) = \frac{k}{n+2} \qquad P(X_{n+1} = k | X_n = k) = \frac{n+2-k}{n+2}$$

We now define

$$M_n = \frac{X_n}{n+2} \quad \text{fraction of red balls after } n \text{ draws}$$

Then M_n is a martingale since

$$E[X_{n+1}|X_n = k] = (k+1)\frac{k}{n+2} + k\frac{n+2-k}{n+2} = \frac{k}{n+2} + k \implies E[X_{n+1}|X_n] = X_n + \frac{X_n}{n+2}.$$

Since this is a Markov chain,

$$E[M_{n+1}|\mathcal{F}_n] = E[M_{n+1}|X_n] = E\left[\frac{X_{n+1}}{n+3}|X_n\right] = \frac{1}{n+3}\left(X_n + \frac{X_n}{n+2}\right) = \frac{X_n}{n+2} = M_n$$

2.4 Martingale and Markov chains

There is a natural connection between Markov chain and martingale. We explain this in the context of Markov chains but this is just an example. Consider a function $f : S \to \mathbb{R}$ and let us derive an equation for $E[f(X_t)]$. Using Kolmogorov equation we have

$$\frac{d}{dt}E[f(X_t)] = \frac{d}{dt}\sum_i f(i)p_t(i) = \sum_i f(i)(p_tA)(i) = \sum_i f(i)\sum_j p_t(j)A(j,i) = \sum_j\sum_i A(j,i)f(i)p_t(j)$$

and thus

$$\frac{d}{dt}E[f(X_t)] = E[g(X_t)] \quad \text{where } g = Af$$

Integrating the previous equation we find

$$E[f(X_t)] - E[f(X_0)] = \int_0^t E[Af(X_s)]ds$$

There is a martingale hidden in this equation. Indeed consider the random variables

$$Y_t = f(X_t) - f(X_0) - \int_0^t Af(X_s)\,ds$$

Our previous calculation shows that $E[Y_t] = 0$, moreover by the Markov property we have

2.5 Exercises

Exercise 2.1 (Martingales for IID random variables)

1. Suppose $X_1, X_2, \cdots$ are IID random variable with $E[X_i] = 1$. Show that $M_n = X_1 X_2 \cdots X_n$ is a martingale.
2. Suppose $X_1, X_2, \cdots$ are IID random variable with a moment generating function $\Lambda(t) = E[e^{tX_i}]$ and let $S_n = X_1 + \cdots + X_n$. Show that $M_n = \frac{e^{tS_n}}{\Lambda(t)^n}$ is a martingale.
3. Suppose $M_n(\alpha)$ is a martingale then under reasonable conditions on the derivatives, $\frac{d^k}{\mathrm{d}alpha^k} M_n(\alpha)$ is also a martingale.
4. Use the method of part 3. and the martingale the $M_n = M_n(t)$ in part 2. to derive the martingales associated to the first three derivative of the martingale (at $t = 0$).

Exercise 2.2 (Likelihood ratio martingale) Suppose the RV X has pdf $f(x)$ and the RV Y has PDF $g(x)$. Show that $M_n = \prod_{j=1}^n \frac{g(X_i)}{f(X_i)}$ is a martingale with respect to $X_1, X_2, \cdots$ where X_i are IID with pdf $f(x)$. This martingale is called the likelihood ratio martingale.

Exercise 2.3 (Martingale associated to the Poisson process) Suppose N_t is a Poisson process with rate λ. Here we denote $\mathcal{F}_t$ the σ-algebra generated by X_s for $0 \leq s \leq t$.

- Show that $N_t - \lambda t$ is a martingale with respect to $\mathcal{F}_t$.
- Show that $N_t^2 - \lambda t$ is martingale with respect to $\mathcal{F}_t$.
- Show that $e^{N_t - \lambda_t}$ is a martingale with respect to $\mathcal{F}_t$.

Exercise 2.4 Consider a branching process X_n with mean offspring number μ and extinction probability a. Show that

- $M_n = X_n \mu^{-n}$ is a martingale with respect to $X_0, \cdots, X_n$.
- $M_n = a^{X_n}$ is a martingale with respect to $X_0, \cdots, X_n$.

Exercise 2.5 Show (by induction) that for the polya's urn we have

$$P(X_n = k+1) = \frac{1}{n+1} \text{ for } k = 1, 2, \cdots$$

Show that M_n converges to M_∞ in distribution. Find M_∞.

3 Optional Sampling Theorem

3.1 Stopping times

A stopping time T with respect to a sequence of random variables $X_0, X_1, \cdots$ should be such that the random time at which you decide to stop depends only on the information you have accumulated so far. If you decide to stop at time n then it should depend only on $X_0, \cdots, X_n$ and you are not allowed to peek into the future.

Definition 3.1 A stopping time T with respect to the filtration $\mathcal{F}_n$ is a random variable taking values $\{0, 1, 2, \cdots, +\infty\}$ such that for any n the event $\{T = n\}$ is measurable with respect to $\mathcal{F}_n$.

Example: $T = k$ is a stopping time.

Example: The hitting time $T_A = \inf\{j, X_j \in A\}$, for some set A, is a stopping time.

Example: If T and S are stopiing time then $\min\{S, T\}$ is also a stopping time. In particular $T_n = \min\{T, n\}$ is a bounded stopping time since $T_n \leq n$ and we have $T_0 \leq T_1 \leq \cdots \leq T_n \leq T$.

3.2 The optional sampling theorem (version 1)

The optional sampling theorem says, roughly speaking, that "you cannot beat a fair game" which means that if M_n is a martingale and T is a stopping time then $E[M_T] = E[M_0]$. This is not true in general as we have seen when considering the martingale betting system where $1 = E[M_T] \neq E[M_0] = 0$. We start with the following result

Theorem 3.1 (Optional Sampling Theorem (T bounded)) Suppose M_n is a martingale and T is a bounded stopping time (i.e. $T \leq K$) then

$$E[M_T|\mathcal{F}_0] = M_0$$

and in particular $E[M_T] = M_0$.

Proof. Since $T \leq K$ we can write

$$M_T = \sum_{j=0}^{K} M_j 1_{\{T=j\}}$$

We compute next $E[M_T|\mathcal{F}_{K-1}]$. Note that since T is bounded by K we have $1_{\{T=K\}} = 1_{\{T>K-1\}}$ which is measurable with respect to $\mathcal{F}_{K-1}$. Therefore we find

$$E[M_T|\mathcal{F}_{K-1}] = E[M_K 1_{\{T>K-1\}}|\mathcal{F}_{K-1}] + \sum_{j=0}^{K-1} E[M_j 1_{\{T=j\}}|\mathcal{F}_{K-1}]$$

$$= 1_{\{T>K-1\}} E[M_K|\mathcal{F}_{K-1}] + \sum_{j=0}^{K-1} M_j 1_{\{T=j\}} = 1_{\{T>K-1\}} M_{K-1} + \sum_{j=0}^{K-1} M_j 1_{\{T=j\}}$$

$$= 1_{\{T>K-2\}} M_{K-1} + \sum_{j=0}^{K-2} M_j 1_{\{T=j\}}$$

where we used that $M_j 1_{\{T=j\}}$ is $\mathcal{F}_{K-1}$ measurable if $j \leq k-1$.

Using this we find

$$E[M_T|\mathcal{F}_{K-2}] = E[E[M_T|\mathcal{F}_{K-1}]|\mathcal{F}_{K-2}]$$

$$= E[1_{\{T>K-2\}} M_{K-1}|\mathcal{F}_{K-2}] + \sum_{j=0}^{K-2} E[M_j 1_{\{T=j\}}|\mathcal{F}_{K-2}]$$

$$= 1_{\{T>K-3\}} M_{K-2} + \sum_{j=0}^{K-3} M_j 1_{\{T=j\}}$$

and thus, inductively,

$$E[M_T|\mathcal{F}_0] = M_0 \quad \blacksquare$$

3.3 The optional sampling theorem (version 2)

To prove a more general version let us assume that $P(T < \infty) = 1$. Then $T_n = \min\{T, n\}$ converges to T and we can write

$$M_T = M_{T_n} 1_{\{T \leq n\}} + M_T 1_{\{T > n\}} = M_{T_n} + M_T 1_{\{T > n\}} - M_n 1_{\{T > n\}}$$

Since T_n is bounded by the optional sampling theorem we have $E[M_{T_n}] = M_0$. But we need then to control the remaining two terms.

If we assume that M_T is integrable, $E[|M_T|] < \infty$, the assumption $P(T < \infty) = 1$ means that $1_{\{T>n\}}$ converges to 0 and thus by the dominated convergence theorem we have $\lim_{n\to\infty} E[M_T 1_{\{T>n\}}] = 0$.

The third term is more troublesome. Indeed for the martingale betting systems, if $T > n$ it means we lost n bets in a row which happens with a probability of $\frac{1}{2^n}$ for a total loss of $-1 - 2 - \cdots - 2^{n-1} = -(2^n - 1)$. Therefore

$$E[M_n 1_{\{T>n\}}] = \frac{1}{2^n}(1 - 2^n) \to -1 \text{ as } n \to \infty$$

It does not converge to 0 but to 1 in accordance with the result $E[M_T] = 1$.

These considerations leads to the following

Theorem 3.2 (Optional Sampling Theorem (general version)) Suppose M_n is a martingale and T is a finite stopping time (i.e. $P(T < \infty) = 1$). If $E[|M_T| < \infty]$ and

$$\lim_{n\to\infty} E[|M_n|1_{\{T>n\}}] = 0$$

then $E[M_T] = M_0$.

Remark If the sequence M_n is uniformly bounded, i.e. $|M_n| \leq C$ then the optional sampling theorem holds since $E[|M_n|1_{\{T>n\}}] \leq CP(T > n)$.

Remark Another condition which gurantees the optional sampling theorem is if $C = \sup_n E[M_n^2] < \infty$. Indeed if this holds given $\epsilon > 0$ we have

$$\begin{aligned} E[|M_n|1_{T>n}] &= E[|M_n|1_{|M_n|>C/\epsilon}1_{T>n}] + E[|M_n|1_{|M_n|\leq C/\epsilon}1_{T>n}] \\ &\leq \frac{\epsilon}{C}E[|M_n|^2 1_{\{T>n\}}] + \frac{C}{\epsilon}P(T > n) \leq \epsilon + \frac{C}{\epsilon}P(T > n) \end{aligned}$$

Taking $n \to \infty$ shows that $\lim_n E[|M_n|1_{\{T>n\}}] \leq \epsilon$.

3.4 Applications of the optional sampling theorem.

Example: Gambler's ruin probability. Define $S_0 = a$ and $S_n = a + X_1 + \cdots + X_n$ where X_i arre IID fair bets $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$.
Then S_n is a martingale and consider the stopping time

$$T = \min\{n : S_n = 0 \text{ or } s_n = N\}$$

which describe the time at which a gambler starting with a fortune a either go bankrupt or reach a fortune of N. Note that if $T > n$ then $S_n \leq N$ and thus $E[|S_n|1_{\{T>n\}}] \leq NP(T > n) \to 0$ as $n \to \infty$.

$$E[S_T] = E[S_0] = a.$$

But we get then

$$a = E[S_T] = NP(S_T = N) \quad \Longrightarrow \quad P(S_T = N) = \frac{a}{N}$$

This gives another (computation free) derivation of the gambler's ruin formula! The case $p \neq \frac{1}{2}$ can also be treated using a (different) martingale and will be considered in the homework.

Example: Gambler's ruin playing time. Suppose S_n is like in the previous example. Then $M_n = S_n^2 - n$ is also martingale since

$$E[S_{n+1}^2 - n + 1|\mathcal{F}_n] = E[S_n^2 + 2X_{n+1}S_n + X_{n+1}^2 - n + 1|\mathcal{F}_n] = S_n^2 + 1 - (n+1) = S_n^2 - n$$

To apply the optional sampling theorem we note that $P(T > n) \leq C\rho^n$ since T is a hiting time for a finite state Markov chain.
Therefore we have $E[|M_n|1_{\{T>n\}}] \leq (N^2 + n^2)C\rho^n \to 0$ and we can apply the optional sampling theorem and

$$E[M_T] = E[M_0] = a^2$$

But, using the previous example we find

$$a^2 = E[M_T] = E[S_T^2] - E[T] = N^2 P(S_T = N) - E[T] = aN - E[T]$$

and thus

$$E[T] = a(N - a)$$

which gives the expected length of play.

3.5 The Voter Model

We can use martingale like for the gambler's ruin to analyze much more complicated models. The voter model is a simple opinion dynamics model. Here are the ingredients.

- A graph $G = (V, E)$ is given. At each vertex of the graph is an agent who can be either in state 0 or in state 1. We view this as two different possible opinions for the agent. We describe the state of the system by a vector

$$\sigma = (\sigma(v))_{v \in V} \qquad \text{with } \sigma(v) \in \{0, 1\}$$

  and the state space is $S = \{0, 1\}^{|V|}$.

- We think of G as a weighted directed graph. To every directed edge we associate a weight function $c(v, w) > 0$ and define $c(v) = \sum_w c(v, w)$. We do not need assume that $c(v, w) = c(w, v)$. But we assume that the graph is connected: there is path along directed edges between any pair of two vertices. To this weight we can associate transition probabilities

$$p(v, w) = \frac{c(v, w)}{c(v)}$$

  Let Y_n be the Markov chain on the state state space V with transtion probaility p_{vw}. It is irreducible and has a stationary distribution $\pi(v)$. If $c(v, w) = c(w, v)$ then the Markov chain Y_n is irreducible and we know that the stationary distribution is $\pi(v) = \frac{c(v)}{c_G}$ where $c_G = \sum_v c(v)$. In general π might be difficult to compute explicitly.

In the voter model, at each unit time you pick a voter, say voter at vertex v. The voter picks one of his neighbor w with probability p_{vw} and, if his neighbor at vertex w has a different opinion than his opinion, the voter at v adopts the opinion of w. This is admittedly a pretty simplistic model but let us analyze it nonetheless.

Let denote by X_n the corresponding Markov chain on S. The transition probabilities are given by

$$P(\sigma, \sigma + \mathbf{e}_v) = \frac{1}{|V|} 1_{\{\sigma(v)=0\}} \sum_{w:\sigma(w)=1} p(v,w)$$

$$P(\sigma, \sigma - \mathbf{e}_v) = \frac{1}{|V|} 1_{\{\sigma(v)=1\}} \sum_{w:\sigma(w)=0} p(v,w)$$

where $\mathbf{e}_v$ is the state with $\mathbf{e}_v(v) = 1$ and $\mathbf{e}_v(w) = 0$ if $w \neq v$.

The key insight is the following

Theorem 3.3 For the voter model

$$M_n = \pi X_n = \sum_v \pi(v) X_n(v)$$

is a martingale.

Proof. By the Markov property

$$E[M_{n+1} - M_n | X_0, \cdots, X_n] = E[M_{n+1} - M_n | X_n]$$

so it enough to show that $E[M_{n+1} - M_n | X_n = \sigma] = 0$.

To see this we note that

$$X_{n+1} = X_n \pm \mathbf{e}_v \implies M_{n+1} - M_n = \pm \pi(v)$$

Therefore

$$E[M_{n+1} - M_n | X_n = \sigma] = \frac{1}{|V|} \sum_v \sum_w \pi(v) p(v,w) \left(1_{\{\sigma(v)=0\}} 1_{\{\sigma(w)=1\}} - 1_{\{\sigma(v)=1\}} 1_{\{\sigma(w)=0\}} \right)$$

Now we rewrite

$$\begin{aligned} 1_{\{\sigma(v)=0\}} 1_{\{\sigma(w)=1\}} - 1_{\{\sigma(v)=1\}} 1_{\{\sigma(w)=0\}} &= 1_{\{\sigma(v)=0\}} (1 - 1_{\{\sigma(w)=0\}}) - 1_{\{\sigma(v)=1\}} 1_{\{\sigma(w)=0\}} \\ &= 1_{\{\sigma(v)=0\}} - 1_{\{\sigma(w)=0\}} \end{aligned}$$

and thus

$$E[M_{n+1} - M_n | X_n = \sigma] = \frac{1}{|V|} \sum_{v:\sigma(v)=0} \sum_w \pi(v) p(v,w) - \sum_{w:\sigma(w)=0} \sum_v \pi(v) p(v,w)$$

Using the fact that π is the stationary distribution we have the balance equation

$$\sum_v \pi(v)p(v,w) = \sum_v \pi(w)p(w,v)$$

Using this and then exchanging the indices v and w we find

$$\sum_{w:\sigma(w)=0} \sum_v \pi(v)p(v,w) = \sum_{w:\sigma(w)=0} \sum_v \pi(w)p(w,v) = \sum_{v:\sigma(v)=0} \sum_w \pi(v)p(v,w)$$

which implies that

$$E[M_{n+1} - M_n | X_n = \sigma] = 0$$

and therefore M_n is a martingale. ■

We can now apply Theorem 3.1 where we take the stopping time to be the consensus time at which everyone is agreement.

$$T = \inf\{n, X_n(v) = 0 \text{ for all } v \text{ or } X_n(v) = 1 \text{ for all } v\}$$

Note that these are absorbing states and since the Markov chain has finite state space $P(T < \infty) = 1$ and M_n is bounded (in fact $0 \leq M_n \leq 1$). We have, similarly to the gambler's ruin,

$$P(X_T = (1, \cdots, 1)) = E[M_T] = E[M_0] = \pi X(0)$$

which is the absorbtion probabilties.

Example If the underlying model is a random walk on the graph $c(v,w) = c(w,v) = 1$ then $\pi(v) = \frac{\deg(v)}{2|E|}$ and we can easily compute
πX_0. For example if we consider the complete graph on N vertices or any regular graph such that all vertices have the same degree then $P(X_T = (1, \cdots, 1))$ is simply equal to the proportion of vertices with opinion 1.

3.6 Exercises

Exercise 3.1 Here is another version of the optional sampling theorem. Suppose that M_n is martingale. The *increments* B_n of the martingales are given by $B_n = M_n - M_{n-1}$.

Show that if the stopping time T hase finite expectation $E[T] < \infty$ and if the martingale M_n has uniformly bounded increments, i.e.

$$\sup_n E[|B_n|] \leq C$$

then $E[M_T] = E[M_0]$.
Hint: Bound $E[|M_{T_n} - M_0|]$ and then use the dominated convergence theorem to take $n \to \infty$.

Exercise 3.2 (Wald's identity) As we have seen before, if (X_i) are IID copies of a random variable X and N is integer value RV which is independent of the $X_i's$ then $E[\sum_{k=1}^{N} X_i] = E[X]E[N]$.
Prove the folliwing generalization: If T is a stopping time with $E[T]$ finite then $E[\sum_{k=1}^{T} X_i] = E[X]E[N]$. *Hint*: Consider a suitable martingale and use Exercise 3.1.

Exercise 3.3 (Gambler's ruin probabilities and expected playing time) Suppose $P(X_i = 1) = p$, $P(X_i = -1) = q = 1 - p$ and let $S_n = a + X_0 + \cdots + X_n$. And let $T = \min\{n : S_n = 0 \text{ or } s_n = N\}$

1. Show that $M_n = \left(\frac{q}{p}\right)^{S_n}$ is a martingale with respect to $X_1, X_2, \cdots$.

2. Use part 1. and the optional sampling theorem to compute $P\{S_T = 0\}$.

3. Use the martingale $Z_n = S_n + (1 - 2p)n$ and part 2. to compute $E[T]$.

4 Martingale Convergence Theorem

4.1 The convergence theorem

The martingale convergence theorem asserts that, under quite general circumstances, a martingale M_n converges to a limiting random variabel M_∞.

Theorem 4.1 (Martingale convergence theorem (version 1)) If M_n is a martingale such that $E[|M_n|] \leq c$ for all n then there exists a random variable M_∞ such that M_n converge to M_∞ almost surely.

Proof. Pick two (arbitrary) numbers $a < b$. The idea of the proof is to show that the probability that the martingale fluctuate infinitely often between a and b is zero. Since this will be true for any a, b this shows almost sure convergence.

Consider the following betting strategy. Think of M_n as the cumulative gain from a sequence of fair games and thus $M_{n+1} - M_n$ is the gain from the $n+1$st game. Take make the following sequence of bets

- If $M_n < a$ makes bets $B_n = 1$ until the martingale M_n reaches or exceeds the value b.
- Once M_n reaches b, stop betting (that is $B_n = 0$) until M_n comes back to less than a.

Continue this process. If the martingale fluctuates infinitely often between a and b then the betting strategy will provide a long term gain and we show that this cannot happen because of the martingale property.

The gain from this strategy after n games is given by

$$W_n = \sum_j B_j (M_j - M_{j-1})$$

where B_j is either 0 or 1 depending on the position of the martingale. Since M_j is a martingale then W_n is a martingale with respect M_0, M_1, M_2. Indeed we we have $E[M_{n+1} - M_n | \mathcal{F}_n] = 0$.

$$E[W_{n+1}|\mathcal{F}_n] = E[B_{n+1}(M_{n+1} - M_n)|\mathcal{F}_n] + E[W_n|\mathcal{F}_n] = W_n.$$

Let U_n the number of *upcrossing* up to step n that is the number of times the martingales goes from below a to above b. Then from the structure of the betting

$$W_n \geq (b-a)U_n - (W_n - a)_-$$

since $(W_n - a)_-$ overestimate the loss during the last interval of play (if $B_n = 1$). Since $E[W_n] = E[W_0]$ we have

$$E[W_0] = E[W_n] \geq (b-a)E[U_n] - E[(W_n - a)_-] \implies E[U_n] \leq \frac{E[(W_n - a)_-]}{b-a} \leq \frac{c+a}{b-a}$$

Since the right hand side is independent of n, the number of upcrossing up to infinity U_∞ has finite expection and thus U_∞ is finite almost surely. ■

4.2 Uniform integrability

- From the convergence theorem we have $M_n \to M_\infty$ almost surely. In general it does *not* imply that $\lim_n E[M_n] = E[M_\infty]$ without further assumption on M_n. For example, for the martingale betting system we have $E[M_n] = 0$ for all n. However if we stop betting at time T (first win) then our gain after that stay at 1 and thus we have $\lim_n W_n = 1$ almsot surely so $M_\infty = 1$ and clearly $E[M_n]$ does not converge to $E[M_\infty]$!

- In order to obtain convergence we need a stronger condition on M_n than our assumption $\sup_n E[|M_n|] \leq C < \infty$. The proper mathemtical asuumption is that the sequence M_n should be *uniformly integrable* (see Math 605 for more details). The sequence $\{X_n\}$ is uniformaly integrable if

$$\sup_n E[|X_n| 1_{|X_n| \geq R}] \to 0 \text{ as } R \to \infty$$

  This means that the tail behavior of X_n is controlled uniformly. We have the following

Theorem 4.2 If X_n converges almost surely to X and $\{X_n\}$ is uniformly integrable then $\lim_n E[X_n] = E[X]$.

For example if $\sup_n E[||X_n^2] \leq c < \infty$ then the $\{X_n\}$ are uniformly integrable. The argument is the same as the one used after Theorem 3.2

Theorem 4.3 (Martingale convergence theorem (version 2)) If M_n is a martingale and $\{M_n\}$ are uniformly integrable then there is a random variable M_∞ such that $M_n \to M_\infty$ almost surely and $\lim_{n\to\infty} E[M_n] = E[M_\infty]$

4.3 Random Harmonic series

Random harmonic series It is well-known that the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ diverges while the alternating harmonic series $1 - \frac{1}{2} + \frac{1}{3} - \cdots$ converges.

What does happen if we chose the sign in the series randomly? Let X_i be IID random variables with $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$. Let $M_0 = 0$ and for $n > 0$ define

$$M_n = \sum_{j=1}^{n} \frac{1}{j} X_j \, .$$

Since M_n is a sum of independent random variable with mean 0, it is a martingale and we have $E[M_n] = 0$ for all n.

By the martingale convergence theorem we have $M_n \to M_\infty$ converges almost surely. Therefore the random harmonic series $\sum_{j=1}^{\infty} \frac{1}{j} X_j$ converges almost surely.

Note that by independence $E[M_n^2] = \sum_{j=1}^{n} \frac{1}{j^2} < \infty$ and thus M_n is uniformly integrable and $E[\sum_{j=1}^{\infty} \frac{1}{j} X_j] = 0$.

4.4 Branching process

Suppose X_n is a branching process with offspring distribution given by a random variable Z with mean $\mu = E[Z]$.

First we prove that $M_n = \mu^{-n} X_n$ is a martingale. We have, using the Markov property, that $E[X_{n+1}|\mathcal{F}_n] = E[X_{n+1}|X_n]$ and

$$E[X_{n+1}|X_n = k] = E[Z_1^{(n)} + \cdots + Z_k^{(n)}] = \mu k$$

and thus $E[X_{n+1}|X_n] = \mu X_n$. This proves that $\mu^{-n} X_n$ is a Martingale.

If $\mu \leq 1$ we have proved that the probability of extinction is 1 and thus $X_n = 0$ for all n sufficiently large and thus M_n converges to 0 almost surely.

If $\mu > 1$ and then $M_n = X_n/\mu^n$. By the Martingale convergence theorem we have $M_n \to M_\infty$. We show that M_∞ is a non-trivial random variable by showing that $E[M_\infty] = \lim_n E[M_n] = 1$ (if we start with single individual).

To do this we need to show uniform integrability. Suppose $\sigma^2 = V(Z)$. We have the formula

$$E[(M_{n+1} - M_n)^2|\mathcal{F}_n] = E[M_{n+1}^2|\mathcal{F}_n] - 2E[M_{n+1}M_n|\mathcal{F}_n] + E[M_n^2|\mathcal{F}_n] = E[M_{n+1}^2|\mathcal{F}_n] - M_n^2$$

and so

$$E[M_{n+1}^2|\mathcal{F}_n] = M_n^2 + E[(M_{n+1} - M_n)^2|\mathcal{F}_n]$$

To compute the second term note that

$$E[(M_{n+1} - M_n)^2|\mathcal{F}_n] = E[(\mu^{-(n+1)} X_{n+1} - \mu^{-n} X_n)^2|\mathcal{F}_n] = \mu^{-2(n+1)} E[(X_{n+1} - \mu X_n)^2|\mathcal{F}_n]$$

Now

$$E[(X_{n+1} - \mu X_n)^2 | X_n = k] = E[(Z_1^{(n)} + \cdots + Z_k^{(n)} - \mu k)^2] = k\sigma^2$$

and so $E[(X_{n+1} - \mu X_n)^2 | X_n] = X_n \sigma^2$.

Combining all this gives (using again that $E[\mu^{-n} X_n] = E[M_n] = 1$)

$$E[M_{n+1}^2] = E[M_n^2] + \mu^{-2(n+1)} \sigma^2 E[X_n] = E[M_n^2] + \frac{\sigma^2}{\mu^{n+2}}$$

Using that $E[M_0^2] = 1$ we find $E[M_1] = 1 + \frac{\sigma^2}{\mu^2}$ and by induction

$$E[M_n^2] = 1 + \sigma^2 \sum_{k=2}^{n+1} \frac{1}{\mu^k}$$

This proves that $\sup_n E[|M_n|^2] < \infty$ and thus M_n is uniformly integrable.

Thus M_∞ is non-trivial.

4.5 Estimating the mean in Bayesian statistics

Estimation problem Suppose $X_1, X_2, \cdots$ are IID random variables whose mean $E[X_i] = \theta^*$ is unknown. In a Bayesian spirit we equip θ with a probability distribution (called the prior distribution) and called this random variable Θ.

Example: The simplest model is when X_i are Bernoulli random variables with a unknown probability of success θ. A natural prior distribution for Θ is the uniform distribution on $[0, 1]$. In this case we would have the joint distribution

$$f(x_1, \cdots, x_n, \theta) = f(x_1, \cdots, X_n|\theta) f(\theta) = \binom{n}{k} \theta^{x_1+\cdots+x_n} (1-\theta)^{n-x_1+\cdots+x_n}$$

There is a natural martingale associated, namely

$$M_0 = E[\Theta], \qquad M_n = E[\Theta|X_1, \cdots, X_n]$$

This means that M_0 is the expectation of θ under the prior distribution $f(\theta)$ and M_n is the expectation of θ inder the *posterior distribution* $f(\theta|x_1, \cdots, x_n)$.

By the martigale convergence theorem (under suitable assumptions on θ to assure uniform integrability of M_n, for example if Θ is bounded) we have

$$\lim_{n\to\infty} M_n = M_\infty$$

where M_∞ is a random variable which depends on the infnite sequence $X_1, X_2, \cdots$.

Since for $m > n$ we have $M_n = E[M_m|X_1, \cdots, X_n]$ taking $m \to \infty$ shows that $M_n = E[M_\infty|X_1, \cdots, X_n]$. Since θ is the mean, by the LLN, for fixed θ we have

$$\theta = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n}$$

and thus θ is a function of $X_1, X_2, \cdots$. This shows that $M_\infty = \theta$ and thus

$$\lim_{n\to\infty} E[\Theta | X_1, \cdots, X_n] = \theta$$

In our simple example we can compute the posterior distribution by Bayes rule (after dusting up our knowledge of the Beta random variables)

$$f(\theta|x_1, \cdots, x_n) = \frac{\binom{n}{k}\theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1+\cdots+x_n}}{\int_0^1 \binom{n}{k}\theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1+\cdots+x_n} d\theta} = \frac{(n+1)!}{k!(n-k)!}\theta^{x_1+\cdots+x_n}(1-\theta)^{n-x_1+\cdots+x_n}$$

which a beta random variable with parameter $\alpha = k+1$ and $\beta = n+1-k$. The mean of beta random variables with parameters α and β is $\frac{\alpha}{\alpha+\beta}$

$$E[\Theta | X_1, \cdots, X_n] = \frac{1 + X_1 + \cdots + X_n}{n+2}$$

This is related to the Polya's urn. To see this let us compute $P(X_1 + \cdots + X_{n+1} = k+1 | X_1 + \cdots + X_n = k)$ under that model. By conditioning we find

$$\begin{aligned} &P(X_1 + \cdots + X_{n+1} = k+1 | X_1 + \cdots + X_n = k) \\ &= \int_0^1 P(X_{n+1} = 1 | \Theta = \theta) f(\theta | X_1 + \cdots X_n = k) d\theta = E[\theta | X_1 + \cdots X_n = k] = \frac{k+1}{n+2} \end{aligned}$$

which is the same transition as Polya's urn.

4.6 Polya's urn

Let us consider the general Polya's urn starting with r red balls and g green balls. At each time a ball is drawn at random, replaced in the urn together with c extra balls of the same color.

Theorem 4.4 (Polya's urn is a martingale) Let X_n be the number of green balls in the Polya's urn at time n. Then $M_n = \frac{X_n}{r+g+nc}$, that is the fraction of green balls at time n is a martingale with respect to $X_0, X_1, \cdots$.

Proof. At time n there is a total of $r + g + nc$ balls in the urn. The sequence X_n form a time-inhomogenous Markov chain with transition probabilities

$$P(X_{n+1} = j + c | X_n = j) = \frac{j}{r+g+nc} \qquad P(X_{n+1} = j | X_n = j) = \frac{r+g+nc-j}{r+g+nc}$$

By the Markov property $E[M_{n+1}|X_0, \cdots, X_n] = E[M_{n+1}|X_n]$ and we have

$$\begin{aligned} E[M_{n+1}|X_n = j] &= \frac{j+c}{r+g+(n+1)c}\frac{j}{r+g+nc} + \frac{j}{r+g+(n+1)c}\frac{r+g+nc-j}{r+g+nc} \\ &= \frac{j(r+g+(n+1)c)}{(r+g+(n+1)c)(r+g+nc)} = \frac{j}{r+g+nc} \end{aligned}$$

and thus $E[M_n + 1|X_n] = \frac{X_n}{r+g+nc} = M_n$. ■

By the martingale convergence theorem we know that $M_n \to M_\infty$ almost surely and we know identify the distribution of M_∞ by computing the distribution of M_n.

The basic observation is that the probability to get first m green balls and then $n-m$ red balls is given by

$$\frac{g}{g+r}\frac{g+c}{g+r+c}\cdots\frac{g+(m-1)c}{g+r+(m-1)c}\frac{r}{g+r+mc}\cdots\frac{r+(n-m-1)c}{g+r+(n-1)c} \tag{4.1}$$

Note also that if we pick m green balls and $n-m$ red balls *in any order* then the probability of that event will have the same probability. Indeed the denominator in Equation 4.1 will be the same and the terms in the numnerators with also be the same but will permuted and appear in a different order.

As a warm up let us consider the special case $r=g=c=1$ we obtain from Equation 4.1

$$P(X_n = m+1) = \binom{n}{m}\frac{m!(n-m)!}{(n+1)!} = \frac{1}{n+1} \quad m = 0,1,2,\cdots,n$$

Therefore M_n is uniformly distributed on $\frac{1}{n+2}, \frac{1}{n+2}, \cdots, \frac{n+1}{n+2}$.

This shows M_n converges in dsitribution to the uniform distribution on $[0,1]$. Indeed for any bounded continuous function $h:[0,1]\to\mathbb{R}$ we have by a Riemann sum argument

$$E[h(M_n)] = \sum_{k=1}^{n+1} h\left(\frac{k}{n+2}\right)\frac{1}{n+1} \to \int h(x)dx$$

In general we have

Theorem 4.5 (Asymptotic distribution in the Polya's urn) The proportion of green balls in the the Polya's urn with paramter r, g, c converges to a beta distribution with paramter $\frac{g}{c}$ and $\frac{r}{c}$

Proof. We prove it only for $c = 1$ and leave the general case $c > 1$ to the reader. Starting from Equation 4.1 we find for the number of green balls

$$\begin{aligned} P(X_n = m + g) &= \binom{n}{m} \frac{g(g+1)\cdots(g+m-1)r(r+1)\cdots(r+n-m-1)}{(g+r)(g+r+1)\cdots(g+r+(n-1))} \\ &= \frac{(g+r-1)!}{(r-1)!(g-1)!} \frac{n!}{m!(n-m)!} \frac{(g+m-1)!(r+n-m-1)!}{(g+r+(n-1))!} \\ &= \frac{(g+r-1)!}{(r-1)!(g-1)!} \frac{m^{g-1}(n-m)^{r-1}}{n^{r+g-1}} \frac{\frac{(g+m-1)!}{m!m^{g-1}} \frac{(r+n-m-1)!}{(n-m)!(n-m)^{r-1}}}{\frac{(g+r+(n-1))!}{n!n^{r+g-1}}} \end{aligned}$$

We now take the limit $n \to \infty$ and $m \to \infty$ such that $\frac{m}{n} \to x$ and $\frac{n-m}{n} \to 1 - x$ Note that

$$\frac{g + (m-1)!}{m!m^{g-1}} = \frac{(m+1)}{m} \frac{m+2}{m} \frac{m+g-1}{m} \to 1 \quad \text{as } m \to \infty$$

The other terms in the last fraction likewise converge to 1.

Therefore as $n \to \infty$

$$P\left(M_n = \frac{g+r+m}{g+r+n}\right) \approx \frac{(g+r-1)!}{(r-1)!(g-1)!}\left(\frac{m}{n}\right)^{r-1}\left(\frac{n-m}{n}\right)^{g-1}\frac{1}{n}$$

By a Riemann sum argument this shows that the distribution of M_n converges to beta random variable with parameters g and r. ■

The case when $c > 1$ is handled in a similarly manner using the Gamma function and details are left to the interested reader.

4.7 Exercises

Exercise 4.1 In this problem we consider a Branching process with a geometric distribution of the offsprings $P(Z = k) = pq^k$, Note that Z has moment generating function $E[\theta^Z] = \frac{1}{1-q\theta}$.

1. We will need to compute the n-fold composition $f^n = f f $. To do this we will use fact from geometry about fractional linear transformation.
2.
3. To identify the distribution of M_∞ we compute for $\lambda \geq 0$

$$E[e^{-\lambda} M_\infty] = \lim_{n\to\infty}$$

$

5 Martingale concentration inequalities

5.1 Conditional Jensen inequality

Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function. There are two equivalent ways to characterize convexity.

1. ϕ is conevx if and only if, for all x, y and $0 < \alpha < 1$ we have

$$\phi(\alpha x + (1-\alpha)y) \leq \alpha\phi(x) + (1-\alpha)\phi(y)$$

   that is, the line segment between $(x, \phi(x))$ and $(y, \phi(y))$ lies *above* the graph of $\phi(t)$ for t between x and y.

2. ϕ is convex if and only if for any x_0 there exists a such that

$$\phi(x) \geq \phi(x_0) + a(x - x_0)$$

   that is there exists a line which lies below the graph of ϕ and interesects the graph at $x = x_0$. If ϕ is differentiable then $a = \phi'(x_0)$ and ϕ lies above its linear approximation at any point x_0.

Both characterizations of convex function will come handy. An fairly immediate condition of convexity is Jensen inequality which we present here in its conditional version.

Theorem 5.1 (Conditional Jensen inequality) Suppose X is a random variable with $E[|X|] < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function. Then we have

$$\phi(E[X|Y]) \leq E[\phi(X)|Y] \quad \text{conditional Jensen inequality}$$

In particular we have $\phi(E[X]) \leq E[\phi(X)]$ (Jensen inequality).

Proof. Choosing $x_0 = E[X|Y]$ we have

$$\phi(X) \geq \phi(E[X|Y]) + a(X - E[X|Y])$$

where $a = a(Y)$ is a random variable which is measurable with respect to Y.

Taking conditional expectation with respect to Y and using the property of conditional expectation we have

$$\begin{aligned} E[\phi(X)|Y] &\geq E[\phi(E[X|Y])|Y] + E[a(Y)(X - E[X|Y])|Y] \\ &= \phi(E[X|Y]) + a(Y)(E[X|Y] - E[X|Y]) = \phi(E[X|Y]) \end{aligned}$$

and this proves Jensen inequality. ■.

5.2 Hoeffding's bound

Recall that if X is a normal random variable then its moment generating function is given by $E[e^{tX}] = e^{\mu t + \frac{\sigma^2 t^2}{2}}$ which implies, by Chernov bound the concentration bound

$$P(X - E[X] \geq \epsilon) \leq \inf_{t \geq 0} \frac{E\left[e^{t(X - E[X])}\right]}{e^{t\epsilon}} = e^{-\sup_{t \geq 0}\left\{t\epsilon - \frac{\sigma^2 t^2}{2}\right\}} = e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Note that we also have a similar bound for the left tail, $P(X - E[X] \leq -\epsilon) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$ prove in the same way.

It turns out that bounded random variables also satisfies such Gaussian concentration bound. To do this we need the following fact which expresses the fact that among all random variable supported on the interval $[-A, B]$ with mean 0 the one which is most spread out is a random variable concentrated on the endpoints $-A$ and B.

Theorem 5.2 (Hoeffdings bound in conditional form) Suppose X is a random variable such that $E[X|Y] = 0$ and $-A \leq X \leq B$. Then for any convex function we have

$$E[\phi(X)|Y] \leq \phi(-A)\frac{B}{A+B} + \phi(B)\frac{A}{A+B}$$

In particular if $A = B$, we have

$$E[e^{tX}] \leq \cosh(At) \leq e^{\frac{A^2 t^2}{2}}$$

Proof. If $x \in [-A, B]$ let us write x as a convex combination of $-A$ and B, that is

$$x = \frac{B-x}{A+B}(-A) + \frac{A+x}{A+B}B\,.$$

The convexity of ϕ implies that

$$\phi(X) \leq \frac{B-X}{A+B}\phi(-A) + \frac{A+X}{A+B}\phi(B)\,.$$

and taking conditional expectation with respect to Y and uising that $E[X|Y] = 0$ gives the result.

Taking now $\phi(x) = e^{tx}$ and $A = B$ we find

$$E[e^{tX}] \leq \frac{1}{2}(e^{-At} + e^{At}) = \cosh(At) = \sum_{n=0}^{\infty} \frac{A^{2n}t^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{A^{2n}t^{2n}}{2^n n!} = e^{\frac{t^2A^2}{2}}$$

5.3 Azuma-Hoeffdings concentration inequality

The next theorem provides concentration inequality for martingales with bounded increments.

Theorem 5.3 (Azuma-Hoeffding's theorem) Suppose M_n is a martingale with $M_0 = 0$ and with bounded increments, i.e. $B_n = M_n - M_{n-1}$ satisfies the bound

$$|B_n| \leq \sigma_n \, .$$

Then we have the Gaussian concentration bound

$$P(M_n \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2\sum_{k=1}^n \sigma_k^2}}$$

Proof. Using the Martingale property and Theorem 5.2

$$E[e^{tM_n}] = E[E[e^{tM_n}|\mathcal{F}_{n-1}]] = E[E[e^{tM_{n-1}}e^{tB_n}|\mathcal{F}_{n-1}]] = E[e^{tM_{n-1}}E[e^{tB_n}|\mathcal{F}_{n-1}]] \leq e^{\frac{t^2\sigma_n^2}{2}}E[e^{tM_{n-1}}]$$

Iterating we find $E[e^{tM_n}] \leq e^{\frac{t^2\sum_{k=1}^n \sigma_k^2}{2}}$ and Chernov bound gives the result.

5.4 McDiarmid Theorem

McDiarmid theorem is an application of Azuma-Hoeffdings theorem and provides concentration bounds for (nonlinear) function of independent random variables $X_1, \cdots, X_n$, under certain conditions.

Definition 5.1 We say that $h(x_1, x_2, \cdots, x_n)$ satisifies the **bounded difference property** if there exist constants c_k such that for all x_k, x_k'

$$|h(x_1, \cdots, x_k, \cdots, x_n) - h(x_1, \cdots, x_k', \cdots, x_n)| \leq c_k$$

that is we control the change of h when changing only one coordinate at a time.

Theorem 5.4 (McDiarmid Theorem) Suppose $X_1, \cdots X_n$ are independent RVs and $h(x_1, \cdots, x_n)$ satisfies the bounded difference property (almost surely). Then we have

$$P\left(h(X_1, \cdots, X_n) - E[h(X_1, \cdots, X_n)] \geq \varepsilon\right) \leq e^{-\frac{\varepsilon^2}{2\sum_{k=1}^n c_k^2}}$$

Proof. **"Martingale trick"**: We construct a suitable martingale. Let us define random variable Y_k by $Y_0 = E[f(X_1, \cdots, X_n)]$ and, for $1 \leq k \leq n$,

$$Y_k = E[h(X_1, \cdots, X_n)|X_1, \cdots, X_k]$$

Note that $Y_n = h(X_1, \cdots, X_n)$ and by construction $E[Y_k|X_1, \cdots, X_{k-1}] = Y_{k-1}$ that is Y_n is a martingale and so is $M_n = Y_n - Y_0$. To use Theorem 5.3 we need to prove that the increment

$$Y_k - Y_{k-1} = E[h(X_1, \cdots, X_k, \cdots, X_n)|X_1, \cdots, X_k] - E[h(X_1, \cdots, X_k, \cdots, X_n)|X_1, \cdots, X_{k-1}]$$

is bounded.

Duplication trick: Let $\widehat{X}_k$ be an independent copy of the random variable X_k. Then by linearity of the expectation and the bounded difference property we have, almost surely,

$$\left|E[h(X_1, \cdots, X_k, \cdots, X_n)|X_1, \cdots, X_k] - E[h(X_1, \cdots, \widehat{X}_k, \cdots, X_n)|X_1, \cdots, X_k]\right| \leq c_k \quad (5.1)$$

Furthermore we have

$$\begin{aligned} E[h(X_1, \cdots, X_k, \cdots, X_n)|X_1, \cdots, X_{k-1}] &= E[h(X_1, \cdots, \widehat{X}_k, \cdots, X_n)|X_1, \cdots, X_{k-1}] \\ &= E[h(X_1, \cdots, \widehat{X}_k, \cdots, X_n)|X_1, \cdots, X_K] \end{aligned} \quad (5.2)$$

The first equality holds because X_k and $\widehat{X}_k$ are identically distributed and left-hand side is a function of $X_1, \cdots, X_{k-1}$. The second equality holds because X_k and $\widehat{X}_k$ are idependent. Combining Equation 5.1 and Equation 5.2 shows that $|Y_k - Y_{k-1}| \leq c_k$ almost surely. ■

5.5 Appplication to statistical estimators

Example: empirical mean Suppose $S_n = X_1 + \cdots + X_n$ is a sum of IID random variables such that $E[X_i] = \mu$ and $a \leq X_i \leq b$. Then the function $h(x_1, \cdots, x_n) = \frac{x_1 + \cdots + x_n}{n}$ satisfies the bounded difference property with $c_k = \frac{(b-a)}{n}$ and we recover the classical Hoeffding's theorem for bounded random variables

$$P\left(\frac{X_1 + \cdots + X_n}{n} - \mu \geq \varepsilon\right) \leq e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

Example: empirical variance Suppose we are interested in estimating the variance σ^2. Then using the unbiased variance estimator

$$V_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i - \frac{S_n}{n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{S_n^2}{(n-1)n}$$

with $E[V_n] = \sigma^2$. If we change X_1 into $\widehat{X}_1$ then $S_n(X_1, \cdots, X_n) - S_n(\widehat{X}_1, \cdots, X_n) = X_1 - \widehat{X}_1$ and so

$$V_n(X_1, \cdots, X_n) - V_n(\widehat{X}_1, \cdots, X_n) = \frac{X_1^2 - \widehat{X}_1^2}{n-1} - \frac{X_1 - \widehat{X}_1}{n-1}\left(\frac{S_n(X_1, \cdots, X_n)}{n} + \frac{S_n(\widehat{X}_1, \cdots, X_n)}{n}\right)$$

Let us assume $a \leq X_i \leq b$, since we can replace X_i by $X_i - (a + b/2)$ without changing V_n we can instead assume that $-\frac{(b-a)}{2} \leq X_i \leq \frac{(b-a)}{2}$ and then find the bounded difference bound

$$|V_n(X_1, \cdots, X_n) - V_n(\widehat{X}_1, \cdots, X_n)| \leq \frac{\frac{5}{4}(b-a)^2}{n-1} = c_k$$

and thus by McDiarmid we get

$$P(V_n - \sigma^2 \geq \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sum_{k=1}^n c_k^2}} \leq e^{-\frac{(n-1)^2}{n}\frac{8\varepsilon^2}{25(b-a)^4}}$$

and this decay exponentially in N again. This can be used for a non-asymptotic confidence interval for the variance.

5.6 Balls and Bins

Another classical example in probability is the so-called "Balls and Bins problem": suppose we have m balls that are thrown in n bins, the location of each ball is chosen at random independently of the other balls. It turns out this problem occur in many algorithmic optimization problems. We can ask many questions realated to this problem. For example what is the probability that one bin has more than two balls in it (this is a version of the famous birthday problem!), or we can ask what is the maximal number of balls in a bin or the number of empty bins, and so on…

Let us compute first the expected number of empty bins. We write

$$N = X_1 + \cdots X_n$$

where $X_i = 1$ is the i^{th} bin is empty and 0 otherwise. For the first urn to be empty, it must be missed by all balls and this occur with probbaility $\left(1 - \frac{1}{n}\right)^m$ and thus

$$E[N] = n\left(1 - \frac{1}{n}\right)^m$$

We prove a concentration bound around the mean using a martingale argument. We consider the sequence of random variable Y_i to be the bin in which the i^{th} ball falls and write $N = N(Y_1, \cdots, Y_m)$. To apply Theorem 5.4 we check the bounded difference equality. Consider how N changes when we change the location of the i^{th} ball. If the i^{th} balls lands in a bin of its own then changing Y_i may increase N by 1 or left it unchanged. If it lands in a bin with other balls, then changing Y_i may decrease N by 1. Then, by Theorem 5.3 we have

$$P(|N - E[N]| \geq \epsilon) \leq 2e^{-\frac{2\epsilon^2}{m}}$$

5.7 Empirical processes

As another application of Azuma-Hoeffding's theorem we consider *empirical processes* which are given by

$$Z(X_1, \cdots, X_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(X_i) - E[f(X_i)] \right|$$

where X_i are IID random variables and f belongs to a suitable class of function $\mathcal{F}$. We assume here that

$$\sup_x |f(x)| \leq B \text{ for all } f \in \mathcal{F}$$

i.e. the functions f are uniformly bounded. One of the simplest example of empirical process to consider the function $f_t(x) = 1_{x \leq t}$. By the LLN we have

$$\frac{1}{n} \sum_{k=1}^{n} 1_{\{X_i \leq t\}} \to P(X \leq t) = F_X(t) \quad \text{almost surely}$$

where F_X is the distribution of the RVs X_i. By *Glivenko–Cantelli Theorem* we have

$$\sup_{t \in \mathbb{R}} |\frac{1}{n} \sum_{k=1}^{n} 1_{\{X_i \leq t\}} - F_X(t)| \to 0 \quad \text{almost surely}$$

that is we have *uniform convergence* of the empirical distribution to the true distribution.

Empirical processes satisfy the bounded difference property: indeed if we set

$$g(x_1, \cdots, x_n) = \left| \frac{1}{n} \sum_{k=1}^{n} f(x_i) - E[f(X)] \right|$$

Then

$$g(x_1, \cdots, \tilde{x}_k, \cdots, x_n) = \left| \frac{1}{n} \sum_{k=1}^{n} f(x_i) - E[f(X)] + \frac{1}{n}(f(\tilde{x}_k) - f(x_k)) \right| \leq g(x_1, \cdots, \tilde{x}_k, \cdots, x_n) + \frac{2B}{n}$$

from which we conclude that $Z(x_1, \cdots, \tilde{x}_k, \cdots, x_n) \leq Z(x_1, \cdots, x_k, \cdots, x_n) + \frac{2B}{n}$. Interchanging x_k and $\tilde{x}_k$ proves the bounded difference property with $c_k = \frac{2B}{n}$. Then Theorem 5.3 show that, for every $\epsilon > 0$.

$$P(|Z - E[Z]| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2B^2}}$$

This tells us that the behavior of Z is controlled by its mean $E[Z]$. For example if we set $\delta = e^{-\frac{n\epsilon^2}{2B^2}}$ we have

$$|Z - E[Z]| \leq B\sqrt{\frac{2}{n} \log \frac{1}{\delta}} \quad \text{with probability at least } 1 - 2\delta$$

for any $\delta > 0$.

To go further we need to control the mean $E[Z]$ and to do this we need two results.

Theorem 5.5 Suppose $Y_1, Y_2, \cdots, Y_N$ are RVs (no independence needed) with $E[e^{tY_i}] \leq e^{\frac{t^2\sigma^2}{2}}$ for $i = 1, \cdots, N$. Then

$$E[\max_i Y_i] \leq \sigma\sqrt{2\ln(N)}$$

and

$$E[\max_i |Y_i|] \leq \sigma\sqrt{2\ln(2N)}$$

Proof. By Jensen inequality we have

$$e^{tE[\max_i Y_i]} \leq E[e^{t\max_i Y_i}] \leq \sum_i E[e^{tY_i}] \leq Ne^{\frac{t^2\sigma^2}{2}}$$

Thus

$$E[\max_i Y_i] \leq \frac{\log N}{t} + \frac{t\sigma^2}{2}$$

and optimizing over t yields the result. Note that if Y_i satisfies the condition of the theorem so does $-Y_i$ and thus we also have

Theorem 5.6 Suppose ϵ_i are IID Rademacher random variable (i.e. equal to ± 1 with probability $\frac{1}{2}$). Then we have

$$E\left[\sup_f \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - E[f(X)]\right|\right] \leq 2E_X\left[E_\epsilon\left[\sup_f |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)|\right]\right]$$

Proof. We use a duplication trick and consider RVs $Y_1, \cdots Y_n$ which are indepndent copies of $X_1, \cdots X_n$. We have then

$$\begin{aligned}
E\left[\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - E[f(X_i)]|\right] &= E\left[\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - E[f(Y_i)]|\right] \\
&= E_X\left[\sup_f \left|E_Y[\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)]\right|\right] \\
&\leq E_X E_Y\left[\sup_f \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right] \\
&\leq E_X E_Y\left[\sup_f \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i) - f(Y_i))\right|\right]
\end{aligned}$$

where in the last line we have used that multiplying $(f(X_i) - f(Y_i))$ by a factor -1 is equivalent to exchanging X_i and Y_i and thus does not change the expectation. Taking now expectation over ϵ_i and using the triangle inequality yields the result.

This is a very useful since have gotten rid of $E[f(X_i)]$ which maybe unknown and we have now a quantity which depnds only on the data $X_1, \cdots, X_n$. How useful this can be is deonstrated next in the context of Glivenko-Catelli Theorem.

Theorem 5.7 We have

$$P\left(\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{k=1}^{n}1_{\{X_i\leq t\}}-F_X(t)\right|\geq 2\sqrt{2\frac{\log 2(n+1)}{n}}+\epsilon\right)\leq e^{-\frac{n\epsilon^2}{2}}$$

Proof. Using Theorem 5.6 we bound $E_\epsilon[\sup_{t\in\mathbb{R}}|\frac{1}{n}\sum_{k=1}^{n}\epsilon_i 1_{\{x_i\leq t\}}|]$ for some *fixed* values of the x_i. Given those values and since $f(x)=1_{\{x\leq t\}}$ is a characteristic function of a half interval, as t varies there are only $n+1$ possible values for the functions $(f(x_1),\cdots,f(x_n))$ (To see this order the x_i in increasing order). Therefore the supremum over t reduces to a supremum over $n+1$ values (which depend on $x_1,\cdots,x_n$)).

Since ϵ_i satisfies a Gaussian bound $E[e^{s\epsilon_i}]\leq e^{\frac{s^2}{2}}$ we have

$$E_\epsilon[e^{s\frac{1}{n}\sum_{k=1}^{n}\epsilon_k 1_{\{x_k\leq t\}}}]=\prod_{k=1}^{n}E_\epsilon[e^{\frac{s}{n}\epsilon_k 1_{\{x_k\leq t\}}}]\leq\prod_{k=1}^{n}e^{\frac{s^2}{2n^2}}=e^{\frac{s^2}{2n}}$$

a bound which is independent of the x_i's! Thus by Theorem 5.5 with $\sigma^2=\frac{1}{n}$ we find

$$E_\epsilon[\sup_{t\in\mathbb{R}}|\frac{1}{n}\sum_{k=1}^{n}\epsilon_i\frac{1}{n}1_{\{x_i\leq t\}}|]\leq\sqrt{2\ln(2(n+1))}\frac{1}{\sqrt{n}}$$

6 Martingale, Markov chain, and CLT

6.1 The Dynkin Martingale and Markov chains

Consider a Markov chain X_j with state space S and transition probabilities P and consider any (bounded) function $f : S \to \mathbb{R}$.

We want to construct a martingale associated to the sequence of random variables $Y_j = f(X_j), j = 0, 1, 2, \cdots$.

If the $f(X_j)$ were independent it would be enough to assume that they have mean 0. In general we use the following construction to build a martingale. If we are given sequence of integrable RV $Y_1, Y_2, \cdots$ then we set

$$D_m = Y_m - E[Y_m | Y_0, \cdots, Y_{m-1}]$$

and we have then $E[D_m | Y_0, \cdots, Y_{m-1}] = 0$. So we can build a martingale by using D_m as the increment of the martingale. We set

$$M_n = \sum_{j=1}^{n} D_j = \sum_{j=1}^{n} (Y_m - E[Y_m | Y_0, \cdots, Y_{m-1}]) .$$

Applying this idea to the sequence $Y_n = f(X_n)$ generated by a Markov chain with generator P we set

$$D_m = f(X_m) - E[f(X_m) | X_0, \cdots, X_{m-1}] = f(X_m) - E[f(X_m) | X_{m-1}] = f(X_m) - Pf(X_{m-1})$$

is a martingale increment with respect to the sequence $X_1, X_2, \cdots$.

We obtain therefore the martingale

$$\begin{aligned} M_n = \sum_{j=1}^{n} f(X_j) - Pf(X_{j-1}) &= f(X_n) - f(X_0) + \sum_{j=0}^{n-1} (f(X_j) - Pf(X_j)) \\ &= f(X_n) - f(X_0) - \sum_{j=0}^{n-1} Af(X_j) \end{aligned} \tag{6.1}$$

for f bounded and where we used the notation $A = P - I$. This martingale is called the *Dynkin's martingale*.

If we want to apply martingale theory to analyze the sum $\sum_{j=1}^{n} g(X_j)$ then we would like to use the martingale Equation 6.1. To do this we must a find a function f such that

$$Af = (P - I)f = -g$$

Then we would have a martingale

$$M_n = f(X_n) - f(X_0) + \sum_{j=0}^{n-1} g(X_j) \qquad \text{with } Af = -g$$

which, up to the correction term $f(X_n) - f(X_0)$ is equal to $\sum_{j=1}^{n} g(X_j)$.

6.2 Poisson equation

What we have uncovered to build a martingale from a Markov chain is an important equation

Definition 6.1 (Poisson equation) Let X_n be a Markov chain with transition matrix P. A function f satisfies the *Poisson equation* for the function g if

$$Af = -g \qquad \text{where} \qquad A = P - I$$

Note that Poisson equation needs not have a solution for arbitrary g. We investigate this in the context of finite state space Markov chain.

Theorem 6.1 Suppose X_n is an irreducible Markov chain with transition matrix P and stationary distribution π. Then the Poisson equation $Af = -g$ has a solution if and only if g has mean 0 with respect to the stationary distribution, that is $\pi g = \sum_i \pi(i) g(i) = 0$.

The function f is then given by

$$f = (\Pi - (P - I))^{-1} g$$

where Π is the matrix whose each row is the stationary distribution π.

Proof. By irreducibility the matrix P has a unique right eigenvector $h = (1, \cdots, 1)^T$ (up to a mutliplicative constant) and unique left eigenvector π for the eigenvalue 1 and the algebraic multiplicity of 1 is also equal to 1.

The matrix Π is a projection, $\Pi^2 = \Pi$ and we have $P\Pi = \Pi P = \Pi$, that is Π projects onto the eigenspace corresponding to the eigenvalue 1 and we have $\Pi f = (\pi f)h$. So if $\pi f = 0$ then $(P - \Pi)f = Pf$. As a consequence $(P - \Pi) = P(I - \Pi)$ has the same eigenvalues as P except the eigenvalue 1 which is replaced by the eigenvalue 0 and thus $I - (P - \Pi)$ is invertible.

If f solves the Poisson equation then $(P - I)f = -g$ and thus $\pi P f - \pi f = \pi f - \pi f = 0 = -\pi g$ which implies that $\pi g = 0$.

Conversely taking g with $\pi g = 0$ let us set $f = (\Pi - (P - I))^{-1} g$. This implies that

$$\Pi f - (P - I)f = \Pi f - Af = g$$

which proves the claim if we can show that $\Pi f = 0$. But since Π commutes with P we have

$$\Pi f = \Pi(\Pi - (P - I))^{-1} g = (\Pi - (P - I))^{-1} \Pi g = 0$$

provided that $\pi g = 0$. ■.

The object involved in solving the Poisson equation occur in various other contexts and deserve a name.

Definition 6.2 (Fundamental matrix) For an irreducible finite state Markov chain X_n with transition matrix P and stationary distribution π, the *fundamental matrix* is given by

$$Z = (I - (P - \Pi))^{-1}$$

where Π is the matrix whose every rows is the stationary distribution π.

The solution to the Poisson equation $Af = -g$ is simply $f = Zg$.

Note that if X_n is irreducible and aperiodic we have proved that $P^n - \Pi$ converges exponentially fast to 0 and using that $(P - \Pi)^n = P^n - \Pi$ we have the convergent series

$$Zf = \sum_{n=0}^{\infty} P^n (f - \pi f)\,.$$

6.3 The Central Limit Theorem

We can now use this to prove a central limit theorem for Markov chain.

Theorem 6.2 (central limit theorem for Markov chain) For $g : S \to \mathbb{R}$ with $\pi g = 0$ and for any initial distribution of X_0, as $n \to \infty$,

$$\frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} g(X_j)$$

converges in distribution to a mean zero normal random variable with variance

$$\sigma^2(g) = \pi f^2 - \pi((Pf)^2)$$

where f is the solution of the Poisson equation $(P - I)f = -g$.

Proof. Consider the Dynkin martingale

$$M_n = f(X_n) - f(X_0) + \sum_{j=0}^{n-1} g(X_j) = \sum_{j=1}^{n} D_j$$

with increments are $D_j = f(X_j) - Pf(X_{j-1})$.

We claim that, for any n,

$$E\left[\frac{e^{i\theta M_n}}{\prod_{j=1}^n E[e^{i\theta D_j}|\mathcal{F}_{j-1}]}\right] = 1 \tag{6.2}$$

Indeed, using the properties of conditional expectations we have

$$E\left[\frac{e^{i\theta M_n}}{\prod_{j=1}^n E[e^{i\theta D_j}|\mathcal{F}_{j-1}]}\right] = E\left[E\left[\frac{e^{i\theta M_n}}{\prod_{j=1}^n E[e^{i\theta D_j}|\mathcal{F}_{j-1}]}|\mathcal{F}_{n-1}\right]\right] = E\left[\frac{E[e^{i\theta M_n}|\mathcal{F}_{n-1}]}{\prod_{j=1}^n E[e^{i\theta}D_j|\mathcal{F}_{j-1}]}\right]$$
$$= E\left[\frac{E[e^{i\theta M_{n-1}+D_n}|\mathcal{F}_{n-1}]}{\prod_{j=1}^n E[e^{i\theta}D_j|\mathcal{F}_{j-1}]}\right] = E\left[\frac{e^{i\theta M_{n-1}}E[e^{i\theta D_n}|\mathcal{F}_{n-1}]}{\prod_{j=1}^n E[e^{i\theta}D_j|\mathcal{F}_{j-1}]}\right] = E\left[\frac{e^{i\theta M_{n-1}}}{\prod_{j=1}^{n-1} E[e^{i\theta}D_j|\mathcal{F}_{j-1}]}\right]$$

Iterating proves the statement. We now use a Taylor expansion and $E[D_j|\mathcal{F}_{j-1}] = 0$ to find

$$\log E[e^{i\frac{\theta}{\sqrt{n}}D_j}|\mathcal{F}_{j-1}] = -\frac{\theta^2}{2n}E[D_j^2|\mathcal{F}_{n-1}] + \frac{1}{n^{3/2}}R_n$$

where R_n is a bounded random variable. Note that

$$E[D_j^2|\mathcal{F}_{n-1}] = E[(f(X_j) - Pf(X_{j-1}))^2|X_{j-1}] = Pf^2(X_{j-1}) - (Pf(X_{j-1}))^2$$

Using the the strong law of large numbers for Markov chain,

$$\lim_{n\to\infty}\sum_{j=1}^{n}\log E[e^{i\frac{\theta}{\sqrt{n}}D_j}|\mathcal{F}_{j-1}] = -\frac{\theta^2}{2}\pi(Pf^2-(Pf)^2) = -\frac{\theta^2}{2}\pi(f^2-(Pf)^2) \quad \text{almost surely}$$

Therefore, from Equation 6.2, we find that

$$\lim_{n\to\infty} E\left[e^{i\frac{\theta}{\sqrt{n}}M_n}\right] = e^{-\frac{\theta^2\sigma^2}{2}}$$

where

$$\sigma^2(g) = \pi(f^2-(Pf)^2)$$

.
Since

$$\frac{1}{\sqrt{n}}M_n = \frac{1}{\sqrt{n}}\left(f(X_n)-f(X_0)\right) + \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}g(X_j)$$

and the term $\frac{1}{\sqrt{n}}\left(f(X_n)-f(X_0)\right)$ is negligible as $n\to\infty$ we have shown that $\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}g(X_j)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(g)$. ■

We can rewrite the asymptotic variance in terms of g using the fundamental matrix $Z = (I - (P - \Pi))^{-1}$. Indeed since $f = Zg$ and $\pi f = \pi g = 0$

$$\begin{aligned} f^2 - (Pf)^2 &= (f - Pf)(f + Pf) = (f - (P - \Pi)f)(f + (P - \Pi)f) \\ &= (I - (P - \Pi))f(I + (P - \Pi))f = (I - (P - \Pi))f(2I - (I - (P - \Pi)))f \\ &= 2gZg - g^2 \end{aligned}$$

Thus the asymptotic variance is given by

$$\sigma^2(g) = \pi(2gZg - g^2)$$

If X_n is aperiodic then we have $Zg = \sum_{n=0}^{\infty}(P - \pi)^n g$ and using, the scalar product $\langle f, g \rangle_\pi = \sum_j \pi(j) f(j) g(j)$ we get

$$\sigma^2(g) = \langle g, g \rangle_\pi + 2 \sum_{n=1}^{\infty} \langle g, P^n g \rangle_\pi$$

which is often the form of the asymptotic variance found in the literature.

6.4 Peskun theorem

In the context of Monte-Carlo methods one can try and use the central limit theorem as way to compare different Monte-Carlo Markov chain methods used to sample the same distribution π. The idea is that the smaller the variance, the "better" the Monte-Carlo will perform since the fluctuations around the stationary values πg will be smaller. The following theorem is a result in this direction. It shows that the more a Monte-Carlo Markov chain "jumps", the smaller the asymptotic variance will be.

The proof will use result from matrix algebra. Suppose a vector space E is equipped with a scalar product $\langle \cdot, \cdot \rangle$ and A and B are self-adjoint. We say that $A \preceq B$ if

$$\langle x, Ax \rangle \leq \langle x, Bx \rangle \quad \text{for all } x \in \mathbb{R}^n$$

and we say that A is positive definite if $0 \prec A$, in which case A is invertible.

Theorem 6.3 Suppose A and B are positive definite with $A \preceq B$ then $B^{-1} \preceq A^{-1}$

Proof. Since B is positive definite then $B^{1/2}$ exists and is also invertible. Then $A \preceq B$ implies that $B^{-1/2} A B^{1/2} \preceq I$. Indeed since $B^{1/2}$ is invertible we can write $x = B^{-1/2} y$ and

$$\langle x, Ax \rangle \leq \langle x, Bx \rangle \implies \langle B^{-1/2}y, AB^{-1/2}y \rangle = \langle y, B^{-1/2}AB^{-1/2}y \rangle \leq \langle B^{-1/2}y, BB^{-1/2}y \rangle = \langle y\,, y \rangle$$

To conclude we need to show that if $C \preceq I$ then $I \preceq C^{-1}$. Since the eigenvectors e_i of C can be chosen to be orthonormal basis of E, $C \preceq I$ implies that the eigenvalues λ_i of C satisfy $0 < \lambda_i \leq 1$.

We have, with $x = \sum_i x_i e_i$,

$$\langle x, x\rangle = \sum_i |x_i|^2 \le \sum_i \frac{1}{\lambda_i}|x_i|^2 = \langle x, C^{-1}x\rangle$$

and thus $I \preceq C^{-1} = B^{1/2}A^{-1}B^{1/2}$. Arguing as above this implies that $B^{-1} \preceq A^{-1}$. ■.

Theorem 6.4 (Peskun Theorem) Suppose P_1 and P_2 are two transition probabilities such that P_1 and P_2 satisfy detailed balance with respect to the stationary distribution π. Assume that

$$P_1(i,j) \le P_2(i,j) \quad \text{for all } i \ne j$$

then for all g with $\pi g = 0$ we have

$$\sigma^2_{P_2}(g) \le \sigma^2_{P_1}(g)\,.$$

Proof. Suppose P satisfies detailed balance with stationary distribution π. Then P is self-adjoint with respect to the scalar product $\langle f, g\rangle_\pi$:

$$\langle f, Pg\rangle_\pi = \langle Pf, g\rangle_\pi\,.$$

Consider the so-called Dirichlet form associated to the Markov chain (with $A = P - I$)

$$\mathcal{E}(f,f) = \langle f, (-A)f\rangle_\pi = \sum_i \pi(i) f(i) \left(f(i) - \sum_j P(i,j) f(j) \right) = \sum_{i,j} \pi(i) f(i) P(i,j) (f(i) - f(j))$$

Using detailed balance, $\pi(i)P(i,j) = \pi(j)P(j,i)$, and then exchanging the indices we find

$$\mathcal{E}(f,f) = \sum_{i,j} \pi(j) f(i) P(j,i) (f(i) - f(j)) = -\sum_{i,j} \pi(i) f(j) P(i,j) (f(i) - f(j))$$

Therefore, averaging these two formulas, we find

$$\mathcal{E}(f,f) = \frac{1}{2} \sum_{i \neq j} \pi(i) (f(i) - f(j)) P(i,j) (f(i) - f(j))$$

Note that the Dirichlet form does not involve the diagonal terms $P(i,i)$. Therefore if $P_1(i,j) \leq P_2(i,j)$ for $i \neq j$ and both satisfy detailed balance then we have an inequality between the Dirichlet forms

$$\mathcal{E}_{P_1}(f,f) \leq \mathcal{E}_{P_2}(f,f)$$

Note also that if we assume $\pi f = 0$ we can write $\mathcal{E}_{P_i}(f,f) = \langle f, (I - (P_i - \Pi))f\rangle_\pi$. Since the since the asymptotic variances $\sigma^2_{P_i}(g)$ have the from $\sigma^2_{P_i}(g) = 2\langle g, (I - (P_i - \Pi))^{-1} g\rangle_\pi - \langle g, g\rangle_\pi$ the proof is complete by invoking Theorem 6.3. ■.

[Example: Metropolis vs Barker]. We can build a Markov chain whose stationary distribution is π by using the transition probabilities

$$P(i,j) = Q(i,j) \min \left\{ 1, \frac{\pi(j)Q(j,i)}{\pi(i)P(i,j)} \right\} \quad Metropolis - Hastingsalgorithm$$

or

$$P(i,j) = Q(i,j) \frac{\frac{\pi(j)Q(j,i)}{\pi(i)P(i,j)}}{1 + \frac{\pi(j)Q(j,i)}{\pi(i)P(i,j)}} \quad Barkeralgorithm$$

Since, for $x \geq 0\$$

$$\frac{x}{1+x} \leq \min\{1, x\}$$

the Metropolis algorithm as a smaller asymptotic variance.