Conditional Expectation and Martingales

Palash Sarkar

Applied Statistics Unit
Indian Statistical Institute
203, B.T. Road, Kolkata
INDIA 700 108
e-mail: palash@isical.ac.in

1 Introduction

This is a set of very hurriedly compiled notes. It has not been proof-checked and is likely to have some errors. These, though, should be minor errors and maybe cleared up by referring to the relevant texts. The texts that have been used in the preparation of the notes are Feller; Grimmett and Stirzaker; Goswami and Rao; David Williams; and Mitzenmacher and Upfal.

The purpose behind the notes is to provide an idea of conditional expectations and martingales while keeping within the ambit of discrete probability. In the future, these notes may be expanded into a more refined set of notes which will also include the more general definitions of conditional expectations and martingales.

These notes have been prepared to supplement two one-and-half hour lectures on discrete probability that had been delivered at a workshop on combinatorics at the Applied Statistics Unit of the Indian Statistical Institute on 17 and 18 February, 2011.

2 A Recapitulation of Discrete Probability

The sample space is a countable set $\{x_1, x_2, x_3, \ldots\}$ with a non-negative number p_i associated to x_i such that

$$p_1 + p_2 + p_3 + \cdots = 1.$$

This will also be written as $p_i = p(x_i)$. The x_i's are called points and subsets of the sample space are called events. If A is an event, then its probability is defined as

$$\mathsf{Pr}[A] = \sum_{x \in A} p(x).$$

If A_1 and A_2 are disjoint events, i.e., $A_1 \cap A_2 = \emptyset$, then it is easy to see that

$$\mathsf{Pr}[A_1 \cup A_2] = \mathsf{Pr}[A_1] + \mathsf{Pr}[A_2].$$

For two arbitrary events (i.e., not necessarily disjoint) A_1 and A_2, the following result is also quite easy to verify.

$$\begin{aligned} \mathsf{Pr}[A_1 \cup A_2] &= \mathsf{Pr}[A_1] + \mathsf{Pr}[A_2] - \mathsf{Pr}[A_1 \cap A_2] \\ &\leq \mathsf{Pr}[A_1] + \mathsf{Pr}[A_2]. \end{aligned}$$

The first row (i.e., the equality) extends using the principle of inclusion and exclusion. On the other hand, the second row (i.e., the inequality) extends directly by induction to give the following bound which is called the union bound.

$$\mathsf{Pr}[A_1 \cup A_2 \cup \cdots \cup A_k] \leq \mathsf{Pr}[A_1] + \mathsf{Pr}[A_2] + \cdots + \mathsf{Pr}[A_k].$$

Two events A and B are said to be independent if

$$\mathsf{Pr}[A \cap B] = \mathsf{Pr}[A] \times \mathsf{Pr}[B].$$

Given n events $A_1, A_2, \ldots, A_n$, these are said to be k-wise independent, if for any $1 \leq j \leq k$ and $\{i_1, \ldots, i_j\} \subseteq \{1, \ldots, n\}$,

$$\mathsf{Pr}[A_{i_1} \cap \cdots \cap A_{i_j}] = \mathsf{Pr}[A_{i_1}] \times \mathsf{Pr}[A_{i_j}].$$

If $k = 2$, then the events are said to be pairwise independent, while if $k = n$, then the events are said to be mutually independent. Intuitively, as k increases from 2 to n, then the amount of randomness in the sequence $A_1, \ldots, A_n$ increases. In many cases, it is of interest to investigate how much can be achieved with pairwise independence.

The notion of independence is of great importance. In fact, it is this multiplicative feature which combines with disjoint additivity to give probability theory much of its surprising results.

Let A be an event with $\mathsf{Pr}[A] > 0$. Let B be another event. The probability of B conditioned on A (also stated as the conditional probability of B given A) is denoted by $\mathsf{Pr}[B|A]$ and is defined in the following manner.

$$\mathsf{Pr}[B|A] = \frac{\mathsf{Pr}[A \cap B]}{\mathsf{Pr}[A]}.$$

For the sake of convenience, $A \cap B$ is also written as AB. If A and B are independent events, then $\mathsf{Pr}[B|A] = \mathsf{Pr}[B]$. (Note that $\mathsf{Pr}[A]$ has to be positive for $\mathsf{Pr}[B|A]$ to be defined.)

The chain rule for conditional probabilities is the following relation between two events A and B.

$$\mathsf{Pr}[A \cap B] = \mathsf{Pr}[B|A] \times \mathsf{Pr}[A].$$

If $\mathsf{Pr}[A] > 0$, then this follows from the definition of conditional probability. If, on the other hand, $\mathsf{Pr}[A] = 0$, then A is a null event (for the discrete case $A = \emptyset$, but, for the more general definition of probability, A may be non-empty but still have $\mathsf{Pr}[A] = 0$) and $A \cap B$ is a subset of A which is also a null event, so that $\mathsf{Pr}[A \cap B] = 0$. So, again both sides are equal. This relation easily generalises to more than two events.

$$\begin{aligned}\mathsf{Pr}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathsf{Pr}[A_1] \times \mathsf{Pr}[A_2|A_1] \times \mathsf{Pr}[A_3|A_2 \cap A_1] \\ \times \cdots \times \mathsf{Pr}[A_n|A_{n-1} \cap \cdots \cap A_1].\end{aligned}$$

A simple relation which is used very often is the following. Let A be an event and B be an event with positive probability.

$$\begin{aligned}\mathsf{Pr}[A] &= \mathsf{Pr}[A \cap (B \cup \overline{B})] \\ &= \mathsf{Pr}[(A \cap B) \cup (A \cap \overline{B})] \\ &= \mathsf{Pr}[A \cap B] + \mathsf{Pr}[A \cap \overline{B}] \\ &= \mathsf{Pr}[A|B]\mathsf{Pr}[B] + \mathsf{Pr}[A|\overline{B}]\mathsf{Pr}[\overline{B}].\end{aligned}$$

Using the fact that values of probabilities are between 0 and 1, the above relation is often used to obtain bounds on $\mathsf{Pr}[A]$.

$$\mathsf{Pr}[A|B]\mathsf{Pr}[B] \leq \mathsf{Pr}[A] \leq \mathsf{Pr}[A|B] + \mathsf{Pr}[\overline{B}].$$

The crux in using this relation is to choose the event B suitably.

A random variable X is a function from a sample space to the reals. Given a real a, the probability that X takes the value a is defined to be the probability of the event $\{x : X(x) = a\}$. This is written as

$$\mathsf{Pr}[X = a] = \mathsf{Pr}[\{x : X(x) = a\}].$$

Think of "$X = a$" as the event $\{x : X(x) = a\}$ which is a subset of the sample space. This will help in understanding much of the description of random variables. In the following, when we talk of two (or more) random variables, then all of these would be defined over the same sample space.

Since we are working with discrete sample spaces, the possible values that a random variable may assume is also countable. So, suppose that X takes the values $a_1, a_2, a_3, \ldots$. Then

$$\sum_{i \geq 1} \mathsf{Pr}[X = a_i] = 1.$$

Let X and Y be two random variables taking values $a_1, a_2, a_3, \ldots$ and $b_1, b_2, b_3, \ldots$ respectively. Then the joint distribution of X and Y is defined as follows.

$$\mathsf{Pr}[X = a_i, Y = b_j] = \mathsf{Pr}[\{x : X(x) = a_i \text{ and } Y(x) = b_j\}].$$

Suppose $p(a_i, b_j)$ is the joint distribution of random variables X and Y. Then the marginal distributions of X and Y are obtained as follows.

$$f(a_i) \stackrel{\Delta}{=} \mathsf{Pr}[X = a_i] = p(a_i, b_1) + p(a_i, b_2) + p(a_i, b_3) + \cdots$$
$$g(a_i) \stackrel{\Delta}{=} \mathsf{Pr}[Y = b_j] = p(a_1, b_j) + p(a_2, b_j) + p(a_3, b_j) + \cdots.$$

Let X and Y be random variables. They are said to be independent if

$$\mathsf{Pr}[X = a, Y = b] = \mathsf{Pr}[X = a] \times \mathsf{Pr}[Y = b]$$

for every possible value a and b that X and Y can take. Extension to more than two variables is straightforward. Random variables $X_1, X_2, \ldots, X_n$ are said to be k-wise independent if for every choice of $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$,

$$\mathsf{Pr}[X_{i_1} = a_1, \ldots, X_{i_k} = a_k] = \mathsf{Pr}[X_{i_1} = a_1] \times \mathsf{Pr}[X_{i_k} = a_k]$$

for every choice of $a_1, \ldots, a_k$ that the random variables $X_{i_1}, \ldots, X_{i_k}$ can take. As in the case of events, if $k = 2$, then the random variables are said to be pairwise independent, where as if $k = n$, then the variables are said to be mutually independent.

If X is a random variable and g is a function from reals to reals, then $g(X)$ is a random variable. The probability that $g(X)$ takes a value b is

$$\mathsf{Pr}[g(X) = b] = \mathsf{Pr}[\{x : g(X(x)) = b\}].$$

If X is a random variable over a discrete sample space, taking the values $a_1, a_2, \ldots$, then the expectation of X is denoted as $E[X]$ and is defined as

$$E[X] = \sum_{i \geq 1} a_i \mathsf{Pr}[X = a_i]$$

provided the series converges absolutely.

The following basic result is easy to prove.

Lemma 1. *If X is a random variable taking values $a_1, a_2, a_3, \ldots$ and g is a function from reals to reals, then*

$$E[g(X)] = \sum_{i \geq 1} g(a_i) \mathsf{Pr}[X = a_i].$$

Setting $g(x) = ax$ for a constant a, the above result shows that $E[aX] = aE[X]$. Similarly, setting $g(x)$ to a constant b, we get

$$\begin{aligned} E[b] = E[g(X)] &= \sum_{i\geq 1} g(a_i)\mathsf{Pr}[X = a_i] \\ &= b\sum_{i\geq 1} \mathsf{Pr}[X = a_i] \\ &= b. \end{aligned}$$

Expectation distributes over a sum of random variables a feature which is called the linearity of expectation. This is stated in the following result.

Theorem 1. *Let $X_1, X_2, \ldots, X_k$ be random variables with finite expectations. Then*

$$E[X_1 + X_2 + \cdots + X_k] = E[X_1] + E[X_2] + \cdots + E[X_k].$$

Proof: It is sufficient to prove the result for $k = 2$. Let X and Y be two random variables having $p(a_i, b_j)$ as the joint distribution and $f(a_i)$ and $g(b_j)$ as the respective marginal distributions. Then

$$E[X] = \sum_{i\geq 1} a_i f(a_i) = \sum_{i\geq 1} a_i \sum_{j\geq 1} p(a_i, b_j) = \sum_{i\geq 1, j\geq 1} a_i p(a_i, b_j).$$

Similarly, $E[Y] = \sum_{i\geq 1, j\geq 1} b_j p(a_i, b_j)$ and so

$$\begin{aligned} E[X] + E[Y] &= \sum_{i\geq 1, j\geq 1} (a_i + b_j)p(a_i, b_j) \\ &= E[X + Y]. \end{aligned}$$

The rearrangements of the sums are possible as the series are absolutely convergent. □

It would have been nice if expectation also disbributed over the product of two random variables. That, however, is not true and the distributive result holds only if the two random variables are independent. (The distributive relation over arbitrary random variables would perhaps be too nice and would probably have made the theory uninteresting.)

Theorem 2. *Let X and Y be independent random variables. Then*

$$E[XY] = E[X] \times E[Y].$$

Proof: As in the above result, let $p(a_i, b_j)$ be the joint distribution and $f(a_i)$ and $g(b_j)$ be the respective marginal distributions of X and Y. By the independence of X and Y, we have $p(a_i, b_j) = f(a_i)g(b_j)$.

$$\begin{aligned} E[XY] &= \sum_{i\geq 1}\sum_{j\geq 1} a_i b_j p(a_i, b_j) \\ &= \sum_{i\geq 1}\sum_{j\geq 1} a_i b_j f(a_i) g(b_j) \\ &= \left(\sum_{i\geq 1} a_i f(a_i)\right) \times \left(\sum_{j\geq 1} b_j g(b_j)\right) \\ &= E[X] \times E[Y]. \end{aligned}$$

As before, the rearrangement of terms is allowed due to the absolute convergence of the series. □

The r-th moment of a random variable X is defined to be the expectation of X^r, if it exists, i.e., the r-th moment is

$$E[X^r] = \sum_{i \geq 1} a_i^r f(a_i)$$

provided the series converges absolutely.

Since $|a|^{r-1} \leq |a|^r + 1$, if $E[X^r]$ exists, then so does $E[X^{r-1}]$. Let $\mu = E[X]$. Then the r-th central moment of X is defined to be $E[(X - \mu)^r]$.

Observe that $(a - \mu)^2 \leq 2(a^2 + \mu^2)$ and so the second central moment of X exists whenever the second moment of X exists.

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

The second central moment of X is called its variance $\mathsf{Var}(X)$ and the positive square root of the variance is called its standard deviation. The variance is a measure of spread of X. It is a sum over i of terms $(a_i - \mu)^2 f(a_i)$. Each of these terms are positive. So, if the variance is small, then each of these terms should be small. But, then either $(a_i - \mu)^2$ is small, implying $|a_i - \mu|$ is small or, if it is large, then $f(a_i)$ is small. In other words, the probability that X takes values away from its mean is small.

For two random variables X and Y with expectations μ_x and μ_y, the co-variance $\mathsf{Cov}(X, Y)$ is defined to be the expectation of $(X - \mu_x)(Y - \mu_y)$, whenever it exists. Note that $|a_i b_j| \leq (a_i^2 + b_j^2)/2$ and so $E[XY]$ exists whenever $E[X^2]$ and $E[Y^2]$ exists. The following relation is easy to obtain using the linearity of expectation.

$$E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x \mu_y.$$

As a result, $\mathsf{Cov}(X, Y)$ equals 0 if X and Y are independent. The converse, however, is not true, i.e., it is possible that $\mathsf{Cov}(X, Y) = 0$ but X and Y are not independent.

Theorem 3. *If $X_1, X_2, \ldots, X_n$ are random variables, then*

$$\mathsf{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \mathsf{Var}(X_i) + \sum_{1 \leq i < j \leq n} 2\mathsf{Cov}(X_i, X_j).$$

Consequently, if $X_1, X_2, \ldots, X_n$ are pairwise independent random variables, then

$$\mathsf{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \mathsf{Var}(X_i).$$

Proof: Let $\mu_i = E[X_i]$.

$$\begin{aligned} \mathsf{Var}(X_1 + X_2 + \cdots + X_n) &= E[(X_1 + X_2 + \cdots + X_n - (\mu_1 + \mu_2 + \cdots + \mu_n))^2] \\ &= E[((X_1 - \mu_1) + (X_2 - \mu_2) + \cdots + (X_n - \mu_n))^2] \\ &= E[\sum_{i=1}^{n} (X_i - \mu_i)^2 + 2 \sum_{1 \leq i < j \leq n} (X_i - \mu_i)(X_j - \mu_j)]. \end{aligned}$$

Using the linearity of expectation, we get the first result. The second statement follows from the fact that pairwise independence ensures that the covariances are zeros. □

The collection $(\mathsf{Pr}[X = a_1], \mathsf{Pr}[X = a_2], \mathsf{Pr}[X = a_3], \ldots)$ is called the probability distribution of the random variable X. Next we define some useful distributions of random variables.

Bernoulli. Here X can take two values 0 and 1 (which is interpreted as failure F and success S) with probabilities q and p respectively. Then $p+q=1$. It is easy to verify that $E[X]=p$ and $\mathsf{Var}(X)=pq$.

Binomial. Let $X_1,\ldots,X_n$ be independent Bernoulli distributed random variables with probability of success p (and probability of failure q) and define $X=X_1+X_2+\cdots+X_n$. Then X is said to follow the binomial distribution. X can take values in the set $\{0,\ldots,n\}$ and

$$\Pr[X=i]=\binom{n}{i}p^iq^{n-i}.$$

Using linearity of expectation, it can be shown that $E[X]=np$ and that $\mathsf{Var}(X)=npq$.

Poisson. Let X be a random variable which can take any non-negative integer value and

$$\mathsf{Pr}[X=i]=e^{-\lambda}\frac{\lambda^i}{i!}.$$

Then X is said to follow the Poisson distribution with parameter λ. It is not too difficult to directly work out that $E[X]=\mathsf{Var}(X)=\lambda$.

Geometric. Let X be a random variable which can take and non-negative integer values and let p and q be such that $p+q=1$. Suppose $\mathsf{Pr}[X=i]=q^ip$. Then X is said to follow the geometric distribution. This can be interpreted as an unlimited sequence of independent Bernoulli trials and X denotes the number of failures before the first success. It can be shown that $E[X]=q/p$.

Let A be an event and I_A be a random variable taking values 0 and 1 such that $\mathsf{Pr}[I_A=1]=\mathsf{Pr}[A]$. Then I_A is called an indicator random variable (of the event A). Clearly, $I_A=1-I_{\overline{A}}$ and for two events A and B, $I_{AB}=I_AI_B$.

$$\begin{aligned}I_{A\cup B}=1-I_{\overline{A\cup B}}=1-I_{\overline{AB}}&=1-I_{\overline{A}}I_{\overline{B}}\\&=1-(1-I_A)(1-I_B)\\&=I_A+I_B-I_AI_B\\&=I_A+I_B-I_{AB}.\end{aligned}$$

Since an indicator variable can take only the two values 0 and 1, its expectation is equal to the probability that it takes the value 1.

$$\begin{aligned}\mathsf{Pr}[A\cup B]=\mathsf{Pr}[I_{A\cup B}]&=E[I_{A\cup B}]\\&=E[I_A+I_B-I_{AB}]\\&=E[I_A]+E[I_B]-E[I_{AB}]\\&=\mathsf{Pr}[A]+\mathsf{Pr}[B]-\mathsf{Pr}[AB].\end{aligned}$$

This technique extends to more than two variables. If $A_1,\ldots,A_n$ are events, then

$$\begin{aligned}I_{A_1\cup\cdots\cup A_n}=1-I_{\overline{A_1\cup\cdots\cup A_n}}&=1-I_{\overline{A_1}\cdots\overline{A_n}}\\&=1-I_{\overline{A_1}}\times\cdots\times I_{\overline{A_n}}\\&=1-(1-I_{A_1})(1-I_{A_2})\cdots(1-I_{A_n}).\end{aligned}$$

Now multiply out; use $I_{A_{i_1}\cdots A_{i_r}}=I_{A_{i_1}}\cdots I_{A_{i_r}}$, take expectations and use $P[A]=E[I_A]$ to obtain the principle of inclusion and exclusion.

3 Conditional Expectation

Let X and Y be random variables such that $E[X]$ is finite. Then

$$E[X|Y=y] = \sum x\mathsf{Pr}[X=x|Y=y] \stackrel{\Delta}{=} \phi(y).$$

In other words, the quantity $E[X|Y=y]$ is a function $\phi(y)$ of y. The conditional expectation of X given Y is defined to be $\phi(Y)$ and is written as $\phi(Y) \stackrel{\Delta}{=} E[X|Y]$. So, the conditional expectation of X given Y is a random variable which is function of the random variable Y.

Proposition 1. $E[E[X|Y]] = E[X]$.

Proof:

$$\begin{aligned}
E[E[X|Y]] = E[\phi(Y)] &= \sum_y \phi(y)\mathsf{Pr}[Y=y] \\
&= \sum_y \sum_x x\mathsf{Pr}[X=x|Y=y]\mathsf{Pr}[Y=y] \\
&= \sum_x x \sum_y \mathsf{Pr}[X=x|Y=y]\mathsf{Pr}[Y=y] \\
&= \sum_x x \sum_y \mathsf{Pr}[X=x, Y=y] \\
&= \sum_x x\mathsf{Pr}[X=x] \\
&= E[X].
\end{aligned}$$

□

Proposition 2. *If X has finite expectation and if g is a function such that $Xg(Y)$ also has finite expectation, then $E[Xg(Y)|Y] = E[X|Y]g(Y)$.*

Proof: Let $\phi_1(Y) = E[Xg(Y)|Y]$ and $\phi_2(Y) = E[X|Y]g(Y)$. We have to show that for each y, $\phi_1(y) = \phi_2(y)$.

$$\begin{aligned}
\phi_1(y) &= E[Xg(Y)|Y=y] \\
&= \sum_x xg(y)\mathsf{Pr}[X=x|Y=y] \\
&= g(y)\sum_x x\mathsf{Pr}[X=x|Y=y] \\
&= g(y)E[X|Y=y] \\
&= \phi_2(y).
\end{aligned}$$

If Y is a constant, i.e., $\mathsf{Pr}[Y=a]=1$ for some a, then $E[X|Y=a] = \sum_x x\mathsf{Pr}[X=x|Y=a] = \sum_x x\mathsf{Pr}[X=x] = E[X]$ and so, in this case, $E[X|Y] = E[X]$.

Proposition 3. *$E[(X-g(Y))^2] \geq E[(X-E[X|Y])^2]$ for any pair of random variables X and Y such that X^2 and $g(Y)^2$ havve finite expectations.*

Proof: Let $\phi(Y) = E[X|Y]$.

$$(X - g(Y))^2 = (X - \phi(Y) + \phi(Y) - g(Y))^2 = (X - \phi(Y))^2 + (\phi(Y) - g(Y))^2 \\ +2(X - \phi(Y))(\phi(Y) - g(Y)).$$

Now,

$$\begin{aligned} E[(X - \phi(Y))(\phi(Y) - g(Y))] &= \sum_{x,y}(x - \phi(y))(\phi(y) - g(y))\mathsf{Pr}[X = x, Y = y] \\ &= \sum_{y}(\phi(y) - g(y)) \sum_{x}(x - \phi(y))\mathsf{Pr}[X = x, Y = y] \\ &= \sum_{y}(\phi(y) - g(y))\mathsf{Pr}[Y = y] \sum_{x}(x - \phi(y))\mathsf{Pr}[X = x|Y = y]. \end{aligned}$$

The last term can be simplified as follows.

$$\begin{aligned} \sum_{x}(x - \phi(y))\mathsf{Pr}[X = x|Y = y] &= \sum_{x} x\mathsf{Pr}[X = x|Y = y] - \phi(y) \sum_{x} \mathsf{Pr}[X = x|Y = y] \\ &= \phi(y) - \phi(y) \\ &= 0. \end{aligned}$$

So,

$$\begin{aligned} E[(X - g(Y))^2] &= E[(X - \phi(Y))^2] + E[(\phi(Y) - g(Y))^2] \\ &\geq E[(X - \phi(Y))^2]. \end{aligned}$$

□

If Y is a constant such that $g(Y) = b$, then $E[(X - b)^2] \geq E[(X - E[X])^2] = \mathsf{Var}(X)$. This gives us the result that the mean squared error is minimum about the expectation.

An interpretation of the above result is the following. Suppose we observe the random variable Y. From this observation, we would like to form an opinion of the random variable X. This opinion is in the form of a prediction function of Y. The above proposition tells us that the best predictor for X from Y is the conditional expectation of X given Y. In fact, this can itself be used to obtain a definition of the conditional expectation and proves to be useful in other parts of the theory.

Proposition 4. *For any function g, such that $g(X)$ has finite expectation,*

$$E[g(X)|Y = y] = \sum_{x} g(x)\mathsf{Pr}[X = x|Y = y].$$

Proof: Let $Z = g(X)$. Then

$$\begin{aligned} E[Z|Y = y] &= \sum_{z} z\mathsf{Pr}[Z = z|Y = y] \\ &= \sum_{z} z \sum_{x:g(x)=z} \mathsf{Pr}[X = x|Y = y] \\ &= \sum_{x} g(x)\mathsf{Pr}[X = x|Y = y]. \end{aligned}$$

□

Proposition 5. $|E[X|Y]| \leq E[|X||Y]$.

Proof: We will use $g(x) = |x|$ in the previous proposition.

$$\begin{aligned} |E[X|Y=y]| &= \left|\sum_x x\mathsf{Pr}[X=x|Y=y]\right| \\ &\leq \sum_x |x|\mathsf{Pr}[X=x|Y=y] \\ &= E[|X||Y-y]. \end{aligned}$$

□

Proposition 6. $E[E[X|Y,Z]|Y] = E[X|Y]$.

Proof: Let $\phi(Y,Z) = E[X|Y,Z]$. Then

$$\begin{aligned} E[\phi(Y,Z)|Y=y] &= \sum_z \phi(y,z)\mathsf{Pr}[Z=z|Y=y] \\ &= \sum_z \left(\sum_x x\mathsf{Pr}[X=x|Y=y,Z=z]\right)\mathsf{Pr}[Z=z|Y=y] \\ &= \sum_z \sum_x x\frac{\mathsf{Pr}[X=x,Y=y,Z=z]}{\mathsf{Pr}[Y=y,Z=z]} \times \frac{\mathsf{Pr}[Y=y,Z=z]}{\mathsf{Pr}[Y=y]} \\ &= \sum_z \sum_x x\mathsf{Pr}[X=x,Z=z|Y=y] \\ &= \sum_x x\sum_z \mathsf{Pr}[X=x,Z=z|Y=y] \\ &= \sum_x x\mathsf{Pr}[X=x|Y=y] \\ &= E[X|Y=y]. \end{aligned}$$

From this the result follows. □

Let $Z = X_n$ and $Y = (X_1, \ldots, X_{n-1})$. Then

$$E[E[X|X_1,\ldots,X_n]|X_1,\ldots,X_{n-1}] = E[X|X_1,\ldots,X_{n-1}].$$

Proposition 7. $E[E[g(X,Y)|Z,W]|Z] = E[g(X,Y)|Z]$.

Proof: Let $U = g(X,Y)$ and apply the previous result. □

More generally, if $U = g(X_1,\ldots,X_m)$, then

$$E[E[U|Y_1,\ldots,Y_n]|Y_1,\ldots,Y_{n-1}] = E[U|Y_1,\ldots,Y_{n-1}].$$

Next, we consider a situation, where conditional expectation plays a major role. Suppose a fair coin is tossed a Poisson number of times. Then we would like to find the conditional expectation of the time of occurrence of the first head, given the total number of heads.

Let $N \sim \mathsf{Poisson}(\lambda)$. Suppose a coin is tossed N times and let X be the number of heads and T be the time of occurrence of the first head. (In case there are no heads, T is defined to be the number of tosses plus one, i.e., $T = N+1$, if $X = 0$.) If $N = 0$, then $X = 0$ and so $T = 1$. We wish to find $E[T|X=x]$ for each $x \geq 0$.

The plan is to first compute $E[T|N,X]$ and then to compute $E[T|X]$ as $E[T|X] = E[E[T|N,X]|X]$.

For $0 \le x \le n$, let $f(n,x) = E[T|N=n, X=x]$. Then $f(n,x)$ is the expected waiting time for the first head given that the total number of tosses is n and the number of heads is x. A recurrence for $f(n,x)$ is obtained as follows.

For $1 \le x \le n$, $f(n,x) = \alpha + \beta$, where

1. $\alpha = E[T|x \text{ heads in } n \text{ tosses, first is head}] \times \mathsf{Pr}[\text{first head}|N=n, X=x]$.
2. $\beta = E[T|x \text{ heads in } n \text{ tosses, first is tail}] \times \mathsf{Pr}[\text{first tail}|N=n, X=x]$.

We have

$$\alpha = \frac{\binom{n-1}{x-1}}{\binom{n}{x}} = \frac{x}{n}$$

and

$$\begin{aligned}\beta &= (1 + f(n-1,x)) \times \frac{\binom{n-1}{x}}{\binom{n}{x}} \\ &= \frac{(n-x)}{n}\left(1 + f(n-1,x)\right).\end{aligned}$$

This gives

$$f(n,x) = 1 + \frac{n-x}{n} f(n-1,x).$$

Since $f(x,x) = 1$, we obtain by induction on n that for $n \ge x$, $f(n,x) = (n+1)/(x+1)$. So, $E[T|X,N] = (N+1)/(X+1)$. We need to compute the conditional expectation of this given $X = x$ and for that we need the conditional distribution of N given $X = x$.

$$\mathsf{Pr}[N=n, X=x] = e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{x}\frac{1}{2^n}.$$

$$\begin{aligned}\mathsf{Pr}[X=x] &= \sum_{n \ge x} e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{x}\frac{1}{2^n} \\ &= \frac{e^{-\lambda}}{x!}\sum_{m \ge 0}\left(\frac{\lambda}{2}\right)^{m+x}\frac{1}{m!} \\ &= \frac{e^{-\lambda}}{x!}\left(\frac{\lambda}{2}\right)^x \sum_{m \ge 0}\left(\frac{\lambda}{2}\right)^m \frac{1}{m!} \\ &= \frac{e^{-\lambda/2}}{x!} \times \left(\frac{\lambda}{2}\right)^x.\end{aligned}$$

Now $\mathsf{Pr}[N=n|X=x] = \mathsf{Pr}[N=n, X=x]/\mathsf{Pr}[X=x]$ and simplification gives

$$\mathsf{Pr}[N=n|X=x] = e^{-\lambda/2}\left(\frac{\lambda}{2}\right)^{n-x}\frac{1}{(n-x)!}.$$

So for $x \ge 1$,

$$\begin{aligned}E[T|X=x] &= E\left[\frac{N+1}{X+1}|X=x\right] \\ &= \frac{x+\lambda/2+1}{x+1} \\ &= 1 + \frac{\lambda}{2(x+1)}.\end{aligned}$$

So, $E[T|X] = 1 + \lambda/(2(X+1))$.

Given $X = 0$, $T = 1 + N$ and $E[N] = \lambda$, but, $E[T|X = 0] = 1 + \lambda/2$.

4 Martingales

A sequence of random variables $S_1, S_2, \ldots$ is a martingale with respect to another sequence of random variables $X_1, X_2, \ldots$ if for all $n \geq 1$ the following two conditions hold.

1. $E[|S_n|] < \infty$.
2. $E[S_{n+1}|X_1, \ldots, X_n] = S_n$.

If $S_n = X_n$ for $n \geq 1$, then the sequence is a martingale with respect to itself.

$E[S_{n+1}|X_1, \ldots, X_n]$ is a function $\psi(X_1, \ldots, X_n)$ and so the relation $E[S_{n+1}|X_1, \ldots, X_n] = S_n$ is meaningless unless S_n itself is a function of $X_1, \ldots, X_n$. A specified sequence itself may not be a martingale. But, it is often possible to find ϕ such that $S_n = \phi(X_n)$ is a martingale.

The Martingale. The following gambling strategy is called a martingale. A gambler lays Rs. 1 bet on the first game. Everytime he loses, he doubles his earlier stake. If he wins on the T-th bet, then he leaves with a profit of $2^T - (1 + 2 + \cdots + 2^{T-1}) = 1$.

Let Y_n be the accumulated gain after the n-th play (losses are negative). We have $Y_0 = 0$ and $|Y_n| \leq 1 + 2 + \cdots + 2^{n-1} = 2^n - 1$. Also, $Y_{n+1} = Y_n$ if the gambler has stopped by time $n + 1$. Otherwise, $Y_{n+1} = Y_n - 2^n$ with probability $1/2$; or $Y_{n+1} = Y_n + 2^n$ with probability $1/2$. So $E[Y_{n+1}|Y_1, \ldots, Y_n] = Y_n$ which shows that $\{Y_n\}$ is a martingale with respect to itself.

Example 1. Let $X_1, X_2, \ldots$ be a sequence of integer valued random variables and $S_0 = 0$, $S_1 = S_0 + X_1$, $\ldots$, $S_n = X_n + S_{n-1}$ be the partial sums.

$$\begin{aligned}
E[S_{n+1}|S_1 = s_1, \ldots, S_n = s_n] &= \sum_{s_{n+1}} s_{n+1}\mathsf{Pr}[S_{n+1} = s_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \\
&= \sum_{x_{n+1}} (s_n + x_{n+1})\mathsf{Pr}[S_n = s_n, X_{n+1} = x_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \\
&= \sum_{x_{n+1}} (s_n + x_{n+1})\mathsf{Pr}[X_{n+1} = x_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \\
&= s_n + \sum_{x_{n+1}} x_{n+1}\mathsf{Pr}[X_{n+1} = x_{n+1}|S_1 = s_1, \ldots, S_n = s_n].
\end{aligned}$$

So, $E[S_{n+1}|S_1, \ldots, S_n] = S_n + E[X_{n+1}|S_1, \ldots, S_n]$.

1. If $E[X_{n+1}|S_1, \ldots, S_n] = 0$, then $\{S_n\}$ is a martingale.
2. If $X_n = \varepsilon_n Y_n$, where ε_n is a random variable taking values ± 1 with probability $1/2$ each and ε_n is independent of all other random variables, then

$$\begin{aligned}
E[X_{n+1}|S_1 = s_1, \ldots, S_n = s_n] &= \sum_{x_{n+1}} x_{n+1}\mathsf{Pr}[X_{n+1} = x_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \\
&= \sum_{y_{n+1}} \Big(\frac{y_{n+1}}{2}\mathsf{Pr}[Y_{n+1} = y_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \\
&\quad - \frac{y_{n+1}}{2}\mathsf{Pr}[Y_{n+1} = y_{n+1}|S_1 = s_1, \ldots, S_n = s_n] \Big) \\
&= 0.
\end{aligned}$$

This captures the idea of a gambler's gain in a fair game.

Example 2. Let $X_1, X_2, \ldots$ be independent random variables with zero means and let $S_n = X_1 + \cdots + X_n$.

$$\begin{aligned} E[S_{n+1}|X_1, \ldots, X_n] &= E[S_n + X_{n+1}|X_1, \ldots, X_n] \\ &= E[S_n|X_1, \ldots, X_n] + E[X_{n+1}|X_1, \ldots, X_n] \\ &= S_n + 0. \end{aligned}$$

The last equality follows from a simple calculation.

Example 3. Let $X_0, X_1, \ldots$ be a discrete time Markov chain with transition matrix $P = [p_{i,j}]$ and the number of states in the state space is countable. Suppose $\psi : S \to \mathbb{R}$ is a bounded function which satisfies the following. For all $i \in S$,

$$\sum_{j \in S} p_{i,j}\psi(j) = \psi(i).$$

Let $S_n = \psi(X_n)$. Then

$$\begin{aligned} E[S_{n+1}|X_1, \ldots, X_n] &= E[\psi(X_{n+1})|X_1, \ldots, X_n] \\ &= E[\psi(X_{n+1})|X_n] \\ &= \sum_{j \in S} p_{X_n,j}\psi(j) \\ &= \psi(X_n) \\ &= S_n. \end{aligned}$$

Example 4. Let $X_1, X_2, \ldots$ be independent variables with zero means and finite variances. Let $S_n = \sum_{i=1}^n X_i$. Define $T_n = S_n^2$.

$$\begin{aligned} E[T_{n+1}|X_1, \ldots, X_n] &= E[S_n^2 + 2S_nX_{n+1} + X_{n+1}^2|X_1, \ldots, X_n] \\ &= T_n + 2E[X_{n+1}]E[S_n|X_1, \ldots, X_n] + E[X_{n+1}^2] \\ &= T_n + E[X_{n+1}^2] \geq T_n. \end{aligned}$$

So $\{T_n\}$ is not a martingale. It is a sub-martingale. If the inequality had been "$\leq$", then it would have been a super-martingale.

Example 5. Suppose $Y_n = S_n^2 - \sum_{i=1}^n \sigma_i^2$, where σ_i is the standard deviation of X_i. Then $\{Y_n\}$ is a martingale with respect to $\{X_n\}$. If each X_i takes the values ± 1, then $Y_n = S_n^2 - n$ is a martingale.

Example 6. Polya's urn scheme. A urn contains b black and r red balls. A ball is drawn at random, its colour noted, it is replaced and along with it an additional ball of the same colour is also put into the urn. This process is repeated. Note that, at each stage, the number of balls increases by one and after n trials, the urn will have $b + r + n$ balls.

Let X_n be the proportion of red balls after n trials with $X_0 = r/(b + r)$. A computation shows that $E[X_n|X_0 = x_0, \ldots, X_{n-1} = x_{n-1}] = x_{n-1}$ and so $E[X_n|X_0, \ldots, X_{n-1}] = X_{n-1}$.

Example 7. Let $\{Y_n\}$ be an independent identically distributed sequence taking values ± 1 with probability $1/2$ each. Set $S_n = \sum_{i=1}^n Y_i$. For any $\theta \in (0, 1)$, the sequence $\{X_n\}$ defined as $X_0 = 1$ and $X_n = 2^n\theta^{(n+S_n)/2}(1 - \theta)^{(n-S_n)/2}$ defines a martingale.

$$\begin{aligned} E[X_n|X_0, \ldots, X_{n-1}] &= 2^n E[\theta^{(n+S_n)/2}(1 - \theta)^{(n-S_n)/2}|X_0, \ldots, X_{n-1}] \\ &= 2^n\theta^{n/2}(1 - \theta)^{n/2}E[\theta^{(S_{n-1}+X_n)/2}(1 - \theta)^{(-S_{n-1}-X_n)/2}|X_0, \ldots, X_{n-1}]. \end{aligned}$$

$$E[\theta^{(S_{n-1}+X_n)/2}(1-\theta)^{(-S_{n-1}-X_n)/2}|X_0,\ldots,X_{n-1}] = E[\theta^{X_n/2}(1-\theta)^{-X_n/2}] \times E[\theta^{S_{n-1}/2}(1-\theta)^{-S_{n-1}/2}|X_0,\ldots,X_{n-1}] = E[\theta^{X_n/2}(1-\theta)^{-X_n/2}]E[\theta^{S_{n-1}/2}(1-\theta)^{-S_{n-1}/2}] = \cdots$$

Note that

$$E[\psi(S_{n-1})|X_1 = x_1,\ldots,X_{n-1} = x_{n-1}] = \sum_{s_{n-1}} \psi(s_{n-1})P[S_{n-1} = s_{n-1}|X_1 = x_1,\ldots,X_{n-1} = x_{n-1}] = \psi(x_1 + \cdots + x_{n-1})$$

and so $E[\psi(S_{n-1})|X_1,\ldots,X_{n-1}] = \psi(S_{n-1})$.

Doob Martingale. Let $X_0, X_1, \ldots, X_n$ be a sequence of random variables and Y is a random variable with $E[|Y|] < \infty$. Let $Z_i = E[Y|X_0,\ldots,X_i]$ for $i = 0,\ldots,n$. Then

$$E[Z_{i+1}|X_0,\ldots,X_i] = E[E[Y|X_0,\ldots,X_{i+1}]|X_0,\ldots,X_i] = E[Y|X_0,\ldots,X_i] = Z_i.$$

So, $\{Z_n\}$ is a martingale with respect to $\{X_n\}$. In most applications, we start the Doob martingale with $Z_0 = E[Y]$, which corresponds to Z_0 being a trivial random variable which is independent of Y.

The interpretation of the Doob martingale is the following. We want to estimate Y which is a function of $X_1,\ldots,X_n$. The Z_i's are refined estimates giving more information gradually. If Y is fully determined by $X_1,\ldots,X_n$, then $Z_n = Y$.

4.1 A Branching Process Example

Let X be a random variable taking non-negative values and assume that $\mathsf{Pr}[X = 0] > 0$. The probability generating function for X is defined to be

$$f(\theta) = E[\theta^X] = \sum_{k\geq 0} \theta^k \mathsf{Pr}[X = k].$$

Taking derivatives

$$f'(\theta) = E[X\theta^{X-1}] = \sum k\theta^{k-1}\mathsf{Pr}[X = k]$$

and $\mu = E[X] = f'(1) = \sum k\mathsf{Pr}[X = k] \leq \infty$.

Suppose $X_r^{(m)}$ be a doubly infinite sequence of independent random variables each of which is distributed according to the distribution of X. The idea is that $X_r^{(n+1)}$ represents the number of children (who will be in the n-th generation) of the r-th animal (if there is one) in the n-th generation. Let $Z_0 = 1$ and

$$Z_{n+1} = X_1^{(n+1)} + \cdots + X_{Z_n}^{(n+1)}.$$

Then Z_n is the size of the n-th generation. Let $f_n(\theta) = E[\theta^{Z_n}]$ be the probability generating function for Z_n.

Proposition 8. $f_{n+1}(\theta) = f_n(f(\theta))$.

Consequently, f_n is the n-fold composition of f.

Proof: Let $U = \theta^{Z_{n+1}}$ and $V = Z_n$. We compute $E[U]$ as $E[E[U|V]]$, i.e., we use the basic tower property of conditional expectation.

$$E[\theta^{Z_{n+1}}|Z_n = k] = E[\theta^{X_1^{(n+1)}+\cdots+X_k^{(n+1)}}|Z_n = k].$$

But, Z_n is independent of the variables $X_1^{(n+1)}, \ldots, X_k^{(n+1)}$ and so the conditional expectation is equal to the absolute expectation.

$$\begin{aligned} E[\theta^{Z_{n+1}}|Z_n = k] &= E[\theta^{X_1^{(n+1)}} \cdots \theta^{X_k^{(n+1)}}] \\ &= E[\theta^{X_1^{(n+1)}}] \times \cdots \times E[\theta^{X_k^{(n+1)}}] \\ &= f(\theta)^k. \end{aligned}$$

The last but one equality follows due to the multiplicative property of expectation for independent variables and the last equality follows from the fact that all the X's have the same distribution as that of X. This shows that $E[\theta^{Z_{n+1}}|Z_n] = f(\theta)^{Z_n}$. Now,

$$E[\theta^{Z_{n+1}}] = E[E[\theta^{Z_{n+1}}|Z_n]] = E[f(\theta)^{Z_n}].$$

By definition, $E[\alpha^{Z_n}] = f_n(\alpha)$ and so, $E[f(\theta)^{Z_n}] = f_n(f(\theta))$. □

Let $\pi_n = \mathsf{Pr}[Z_n = 0]$, i.e., π_n is the probability that the population vanishes at the n-th stage. Then $\pi_n = f_n(0)$ and so

$$\pi_{n+1} = f_{n+1}(0) = f(f_n(0)) = f(\pi_n).$$

Let π be the limit of π_n as n goes to infinity. Then

$$\pi = f(\pi).$$

Theorem 4. *If $E[X] > 1$, then the extinction probability π is the unique root of the equation $\pi = f(\pi)$ which lies strictly between 0 and 1. If $E[X] \leq 1$, then $\pi = 1$.*

Recall that $Z_{n+1} = X_1^{(n+1)} + \cdots + X_{Z_n}^{(n+1)}$ where $X_i^{(n+1)}$ are independent of the values $Z_1, Z_2, \ldots, Z_n$. Then, it is clear that

$$\mathsf{Pr}[Z_{n+1} = j|Z_0 = i_0, \ldots, Z_n = i_n] = \mathsf{Pr}[Z_{n+1} = j|Z_n = i_n].$$

This shows that the sequence $\{Z_n\}$ forms a Markov chain.

$$E[Z_{n+1}|Z_0 = i_0, \ldots, Z_n = i_n] = \sum_j j\mathsf{Pr}[Z_{n+1} = j|Z_n = i_n] = E[Z_{n+1}|Z_n = i_n].$$

This in turn says that $E[Z_{n+1}|Z_0, \ldots, Z_n] = E[Z_{n+1}|Z_n]$.

Since, each of the animals in the n-th generation gives rise to μ children it is intuitively obvious that $E[Z_{n+1}|Z_n] = \mu Z_n$. This is confirmed by differentiating both sides of $E[\theta^{Z_{n+1}}|Z_n] = f(\theta)^{Z_n}$ with respect to θ and setting $\theta = 1$.

Define $M_n = Z_n/\mu^n$, $n \geq 0$. Then

$$E[M_{n+1}|Z_0, Z_1, \ldots, Z_n] = M_n.$$

So, $\{M_n\}$ is a martingale with respect to the sequence $\{Z_n\}$. In other words, this says that given the history of Z up to stage n, the next value M_{n+1} of M is on average what it is now. The notion of constant on average conveys much more information than the correct but, less informative statement $E[M_n] = 1$.

5 Stopping Times

Suppose that $Z_1, \ldots, Z_n$ is a (finite) martingale with respect to $X_0, X_1, \ldots, X_n$. Assume that the Z_i's are the winnings of a gambler in a fair game and that the gambler had decided (before the start of the game) to quit after n games. The following result tells us that the expected win of a gambler does not change.

Lemma 2. *If the sequence $Z_1, \ldots, Z_n$ is a martingale with respect to $X_0, X_1, \ldots, X_n$, then $E[Z_n] = E[Z_0]$.*

Proof: From the martingale property, we have $Z_i = E[Z_{i+1}|X_0, \ldots, X_i]$. Taking expectation on both sides, we get

$$E[Z_i] = E[E[Z_{i+1}|X_0, \ldots, X_i]] = E[Z_{i+1}].$$

Repeating this argument gives the result. □

Suppose now that the number of games that the gambler decides to play is not fixed at the beginning. Instead, the gambler decides to quit mid-way depending on the outcomes of the games played so far. Then the time that the gambler chooses to quit is called the stopping time.

Formally, a non-negative, integer-valued random variable T is a stopping time for the sequence $\{Z_n\}$ if the event $T = n$ depends only on the value of the random variables $Z_0, \ldots, Z_n$. If $Z_1, Z_2, \ldots$ are independent, then T is a stopping time if the event $T = n$ is independent of $Z_{n+1}, Z_{n+2}, \ldots$

Examples of stopping time would be if the gambler decides to stop after winning three times in succession, or after winning a certain amount of money, etcetera. On the other hand, let T be the *last* time that the gambler wins five times in a row. Then T would not be a stopping time, since the *last* time cannot be determined without reference to the future.

Theorem 5. *If $Z_0, Z_1, \ldots$ is a martingale with respect to $X_1, X_2, \ldots$ and if T is a stopping time for $X_1, X_2, \ldots$, then*

$$E[Z_T] = E[Z_0]$$

whenever one of the following conditions hold.

1. *The Z_i are bounded, i.e., $|Z_i| \leq c$ for some constant c and for all $i \geq 0$.*
2. *T is bounded.*
3. *$E[T] < \infty$ and there is a constant c such that $E[|Z_{i+1} - Z_i||X_1, \ldots, X_i] < c$.*

Consider a sequence of independent, fair games. In each round, a player wins a dollar with probability $1/2$ or loses a dollar with probability $1/2$. Let $Z_0 = 0$ and let X_i be the amount won on the i-th game. Also, let Z_i be the amount of winnings after i games. Suppose that the player quits the game when she either loses l_1 dollars or wins l_2 dollars. What is the probability that the player wins l_2 dollars before losing l_1 dollars?

Let T be the first time the player has either won l_2 or lost l_1. Then T is a stopping time for $X_1, X_2, \ldots$. The sequence $Z_0, Z_1, \ldots$ is a martingale and since the values of the Z_i are clearly bounded, we can apply the martingale stopping theorem. So, $E[Z_T] = 0$. Let q be the probability that the gambler quits playing after winning l_2 dollars. Then

$$E[Z_T] = l_2 q - l_1(1 - q) = 0.$$

This shows that $q = l_1/(l_1 + l_2)$. So, the probability q is obtained using the martingale stopping theorem.

Theorem 6 (Wald's equation). *Let $X_1, X_2, \ldots$ be nonnegative, independent, identically distributed random variables with distribution X. Let T be a stopping time for this sequence. If T and X have bounded expectations, then*

$$E\left[\sum_{i=1}^{T} X_i\right] = E[T] \times E[X].$$

There are different proofs of the equality that do not require the random variables $X_1, X_2, \ldots$ to be nonnegative.

Proof: For $i \geq 1$, let $Z_i = \sum_{j=1}^{i}(X_i - E[X])$. The sequence $Z_1, Z_2, \ldots$ is a martingale with respect to $X_1, X_2, \ldots$ and $E[Z_1] = 0$. It is given that T has bounded expectation and

$$E[|Z_{i+1} - Z_i||X_1, \ldots, X_i] = E[|X_{i+1} - E[X]|] \leq 2E[X].$$

So, applying the martingale stopping theorem, we get $E[Z_T] = E[Z_1] = 0$ and so

$$\begin{aligned} 0 = E[Z_T] &= E\left[\sum_{j=1}^{T}(X_j - E[X])\right] \\ &= E\left[\left(\sum_{j=1}^{T} X_j\right) - TE[X])\right] \\ &= E\left[\sum_{j=1}^{T} X_j\right] - E[T]E[X]). \end{aligned}$$

□

Consider a gambling game, in which a player first rolls a die. If the outcome is X, then X more dice are rolled and the gain Z is the sum of the outcomes of the X dice. What is the expected gain of the gambler?

For $1 \leq i \leq X$, let Y_i be the outcome of the i-th die. Then $E[Z] = E[\sum_{i=1}^{X} Y_i]$. By definition, X is a stopping time for the sequence $Y_1, Y_2, \ldots$ and so by Wald's equation,

$$E[Z] = E[X]E[Y_i] = (7/2)^2 = 49/4.$$

6 Martingale Tail Inequalities

Theorem 7 (Azuma-Hoeffding Inequality). *Let $X_0, \ldots, X_n$ be a martingale such that $|X_k - X_{k-1}| \leq c_k$. Then, for all $t \geq 0$ and any $\lambda > 0$,*

$$\mathsf{Pr}[|X_t - X_0| \geq \lambda] \leq 2\exp\left(-\left(\lambda^2/2\right) / \left(\sum_{k=1}^{n} c_k^2\right)\right).$$

Proof: Define the so-called martingale difference sequence $Y_i = X_i - X_{i-1}$, $i = 1, \ldots, t$. Note that $|Y_i| \leq c_i$ and since $X_0, X_1, \ldots$ is a martingale,

$$\begin{aligned} E[Y_i|X_0, X_1, \ldots, X_{i-1}] &= E[X_i - X_{i-1}|X_0, X_1, \ldots, X_{i-1}] \\ &= E[X_i|X_0, X_1, \ldots, X_{i-1}] - X_{i-1} \\ &= 0. \end{aligned}$$

Consider $E[e^{\alpha Y_i}|X_0, X_1, \ldots, X_{i-1}]$. Write

$$Y_i = -c_i\frac{1 - Y_i/c_i}{2} + c_i\frac{1 + Y_i/c_i}{2}.$$

From the convexity of $e^{\alpha Y_i}$, it follows that

$$\begin{aligned} e^{\alpha Y_i} &\leq \frac{1 - y_i/c_i}{2}e^{-\alpha c_i} + c_i\frac{1 + Y_i/c_i}{2}e^{\alpha c_i} \\ &= \frac{e^{\alpha c_i} + e^{-\alpha c_i}}{2} + \frac{Y_i}{2c_i}\left(e^{\alpha c_i} - e^{-\alpha c_i}\right). \end{aligned}$$

Since $E[Y_i|X_0, X_1, \ldots, X_{i-1}] = 0$, we have

$$\begin{aligned} E\left[e^{\alpha Y_i}|X_0, X_1, \ldots, X_{i-1}\right] &\leq E\left[\frac{e^{\alpha c_i} + e^{-\alpha c_i}}{2} + \frac{Y_i}{2c_i}\left(e^{\alpha c_i} - e^{-\alpha c_i}\right)|X_0, X_1, \ldots, X_{i-1}\right] \\ &= \frac{e^{\alpha c_i} + e^{-\alpha c_i}}{2} \\ &\leq e^{(\alpha c_i)^2/2}. \end{aligned}$$

The last inequality follows using the Taylor series expansion of e^x.

Using

$$E\left[e^{\alpha(X_t - X_0)}\right] = E\left[e^{\alpha(X_{t-1} - X_0)} e^{\alpha(X_t - X_{t-1})}\right]$$

we can write

$$\begin{aligned} E\left[e^{\alpha(X_t - X_0)}|X_0, \ldots, X_{t-1}\right] &= E\left[e^{\alpha(X_{t-1} - X_0)} e^{\alpha(X_t - X_{t-1})}|X_0, \ldots, X_{t-1}\right] \\ &= e^{\alpha(X_{t-1} - X_0)} E\left[e^{\alpha(X_t - X_{t-1})}|X_0, \ldots, X_{t-1}\right] \\ &= e^{\alpha(X_{t-1} - X_0)} E\left[e^{\alpha Y_t}|X_0, \ldots, X_{t-1}\right] \\ &\leq e^{\alpha(X_{t-1} - X_0)} e^{\alpha c_t^2/2}. \end{aligned}$$

Taking expectations and iterating, we obtain the following.

$$\begin{aligned} E[e^{\alpha(X_t - X_0)}] &= E[E[e^{\alpha(X_t - X_0)}]|X_0, \ldots, X_{t-1}] \\ &\leq E[e^{\alpha(X_{t-1} - X_0)}] e^{\alpha c_t^2/2} \\ &\cdots \\ &\leq \exp\left(\frac{\alpha^2}{2}\sum_{i=1}^{t} c_i^2\right). \end{aligned}$$

Hence,

$$\begin{aligned} \mathsf{Pr}[X_t - X_0 \geq \lambda] &= \mathsf{Pr}[e^{\alpha(X_t - X_0)} \geq e^{\alpha\lambda}] \\ &\leq \frac{E\left[e^{\alpha(X_t - X_0)}\right]}{e^{\alpha\lambda}} \\ &\leq \exp\left(\alpha^2 \sum_{k=1}^{t} \frac{c_i^2}{2} - \alpha\lambda\right) \\ &\leq \exp\left(\frac{-\lambda^2}{2\sum_{k=1}^{t} c_k^2}\right). \end{aligned}$$

The last inequality comes from choosing $\alpha = \lambda/(\sum_{k=1}^{t} c_k^2)$. A similar argument gives the bound for $\mathsf{Pr}[X_t - X_0 \leq -\lambda]$ and can be seen by replacing X_i with $-X_i$ everywhere. □

Corollary 1. *Let $X_0, X_1, \ldots$ be a martingale such that for all $k \geq 1$, $X_k - X_{k-1} \leq c$. Then for all $t \geq 1$ and $\lambda > 0$,*

$$\mathsf{Pr}\left[|X_t - X_0| \geq \lambda c\sqrt{t}\right] \leq 2e^{-\lambda^2/2}.$$

A more general form of the Azuma-Hoeffding inequality is given below and yields slightly tighter bounds in applications.

Theorem 8. *Let $X_0, X_1, \ldots, X_n$ be a martingale such that*

$$B_k \leq X_k - X_{k-1} \leq B_k + d_k$$

for some constants d_k and for some random variables B_k that may be functions of $X_0, X_1, \ldots, X_{k-1}$. Then, for all $t \geq 0$ and any $\lambda > 0$,

$$\mathsf{Pr}[|X_t - X_0| \geq \lambda] \leq 2\exp\left(\frac{-2\lambda^2}{\sum_{k=1}^{t} d_k^2}\right).$$

A general formalization. A function $f(x_1, \ldots, x_n)$ satisfies the Lipschitz condition if for any $x_1, \ldots, x_n$ and y_n,

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)| \leq c.$$

In other words, changing any component in the input changes the output by at most c.

Let $Z_0 = E[f(X_1, \ldots, X_n)]$ and $Z_k = E[f(X_1, \ldots, X_n)|X_1, \ldots, X_k]$. Then $\{Z_n\}$ is a Doob martingale. If the X_k are *independent random variables*, then it can be shown that there exist random variables B_k such that

$$B_k \leq Z_k - Z_{k-1} \leq B_k + c.$$

(For a proof of this see Mitzenmacher and Upfal.) It is necessary that $X_1, X_2, \ldots$ are independent. If they are not, then the relation may not hold. The advantage of this fact is that the gap between the lower and upper bounds on $Z_k - Z_{k-1}$ is at most c and so the Azuma-Hoeffding inequality applies.

7 Applications of the Azuma-Hoeffding Inequality.

Pattern matching. Let $X = (X_1, \ldots, X_n)$ be a sequence of characters chosen independently and uniformly at random from an alphabet of size s. Let $B = (b_1, \ldots, b_k)$ be a fixed string of length k. Let F be the number of occurrences of B in X. Using linearity of expectation, it is easy to see that $E[F] = (n - k + 1)(1/s)^k$.

Let $Z_0 = E[F]$ and for $1 \leq i \leq n$, let $Z_i = E[F|X_1, \ldots, X_i]$. The sequence $Z_0, \ldots, Z_n$ is a Doob martingale and $Z_n = F$. Since each character in the string can participate in no more than k possible matches, it follows that $|Z_{i+1} - Z_i| \leq k$. In other words, the value of X_{i+1} can affect the value of F by at most k in either direction. So,

$$|E[F|X_1, \ldots, X_{i+1}] - E[F|X_1, \ldots, X_i]| = |Z_{i+1} - Z_i| \leq k.$$

By the Azuma-Hoeffding bound,

$$\mathsf{Pr}[|F - E[F]| \geq \varepsilon] \leq 2e^{-\varepsilon^2/2nk^2}.$$

From the corollary,

$$\mathsf{Pr}[|F - E[F]| \geq \lambda k\sqrt{n}] \leq 2e^{-\lambda^2/2}.$$

Slightly better bounds can be obtained by using the more general framework. Let $F = f(X_1, \ldots, X_n)$. Then changing the value of any input can change the value of F by at most k and so the function satisfies the Lipschitz condition. The stronger version of the Azuma-Hoeffding bound can now be applied to obtain

$$\mathsf{Pr}[|F - E[F]| \geq \varepsilon] \leq 2e^{-2\varepsilon^2/nk^2}.$$

This improves the value in the exponent by a factor of 4.

Balls and bins. Suppose m balls are thrown independently and uniformly at random into n bins. Let X_i be the random variables representing the bin into which the i-th ball falls. Let F be the number of empty bins after m balls are thrown. Then the sequence $Z_i = E[F|X_1, \ldots, X_i]$ is a Doob martingale.

The claim is that the function $F = f(X_1, X_2, \ldots, X_n)$ satisfies the Lipschitz condition with bound 1. Consider how placing the i-th ball can change the value of F. If the i-th ball falls into an otherwise empty bin, then changing the value of X_i to a non-empty bin increases the value of F by one; similarly, if the i-th ball falls into an non-empty bin, then changing the value of X_i such that the i-th ball falls into an empty bin decreases the value of F by one. In all other cases, changing X_i leaves F_i unchanged. So, using the Azuma-Hoeffding inequality, we obtain

$$\mathsf{Pr}[|F - E[F]|] \leq 2^{-\varepsilon^2/m}.$$

Note that $E[F] = n(1 - 1/n)^m$, but, it was possible to obtain a concentration bound for F without using $E[F]$. In fact, in many cases, it is possible to obtain a concentration bound for a random variable without knowing its expectation.