RANDOM WALKS

CHRISTIAN REMLING

1. Introduction

The material presented in these notes is taken from W. Feller, An Introduction to Probability Theory and its Applications, I.

We will continue to analyze the properties of sums of independent, identically distributed random variables

$$S_n = X_1 + X_2 + \ldots + X_n.$$

We will only deal with the very specific situation where X_j only takes the two values ± 1, with equal probability ($p = 1/2$). You can think of a *random walk,* performed as follows: at time j ($j = 1, 2, \ldots$) you toss a coin and then move one step to the left or to the right, according to the outcome of the coin toss. The random variable S_n then gives you your position at time n, if you start at the origin. Alternatively, you can think of betting a dollar on heads in a series of coin tosses. Then S_n records your winnings after n coin flips.

Recall that the central limit theorem (or its special case, the theorem of de Moivre-Laplace) says that

$$\lim_{n\to\infty} P(a\sqrt{n} \le S_n \le b\sqrt{n}) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

Exercise 1.1. Derive this more carefully from the CLT.

We also saw earlier that the (one-dimensional) random walk is *recurrent,* that is, given any $m \in \mathbb{Z}$, the probability that we will eventually visit m is one. In fact, with probability one, the random walk will visit every point infinitely many times.

Here, we want to study other asymptotic properties of S_n.

We need some notation. It is sometimes useful to think in terms of paths leading from the origin (say) to another point. We can represent these paths either as graphs in a coordinate system or by simply listing the values of $S_0, S_1, \ldots, S_n$, in this order. For example,

$$0 \to 1 \to 2 \to 1 \to 0 \to -1 \to 0$$

is a path of *length* 6 (meaning $n = 6$), starting and ending at the origin.

Using this language, all paths are equally likely, so we can find probabilities by counting paths. The total number of paths of length n, starting at a fixed point, equals 2^n because we have 2 choices at each step (left or right).

For integers k, l, n, let $N_n(k \to l)$ denote the number of paths of length n that lead from k to l. For instance, $N_2(1 \to 1) = 2$, because we have the two possibilities

$$1 \to 2 \to 1, \quad 1 \to 0 \to 1.$$

Sometimes we want to impose additional conditions on paths. We then use self-explanatory notation: for example, we write $N_n(k \to l, S_1 = k-1)$ for the number of paths of length n that start at k, then go to $k-1$ in the first step, and finally reach l in step n.

2. Some auxiliary results

Lemma 2.1 (The Reflection Lemma)**.** *Let k, l (and n) be positive integers. Then*

$$N_n(k \to l, S_j = 0 \textit{ for some } j = 1, 2, \ldots n-1) = N_n(-k \to l)$$

Proof. This is geometrically obvious, and the term *Reflection Lemma* suggests how to prove it: Given a path leading from k to l that visits the origin at some point, reflect the part up to the first visit to the origin about the x-axis. This establishes a one-to-one correspondence between paths from k to l visiting the origin and arbitrary paths from $-k$ to l. □

Exercise 2.1. Work this out in more detail (picture!).

The following is a useful consequence of the Reflection Lemma. Moreover, this statement is also of some independent interest.

Lemma 2.2 (The Ballot Theorem)**.** *Suppose that $l > 0$. Then*

$$N_n(0 \to l, S_j > 0, j = 1, \ldots, n) = N_{n-1}(0 \to l-1) - N_{n-1}(0 \to l+1)$$
$$= \frac{l}{n} N_n(0 \to l).$$

This is called the *Ballot Theorem* because of the following application: Suppose that in an election involving two candidates, candidate A scores l votes more than candidate B. Then, if there was a total of n votes, the probability that candidate A was ahead throughout the counting is $\frac{l}{n}$. This follows from Lemma 2.2 because the number of ways of doing the counting is the same as the number of paths from 0 to l (given that this final outcome is already known), and in order for candidate A to be ahead all the time, the path must not visit the origin again. By the identity from the Lemma, these paths are a fraction l/n of all paths leading to l.

Proof. Clearly, if we want to go to $l > 0$ and also have to avoid the origin, the first step has to lead to 1. Thus the left-hand side of the identity we're trying to prove equals

$$N_{n-1}(1 \to l, S_j \neq 0, j = 1, \dots, n).$$

It now suffices to demand that $S_j \neq 0$ because if we start at 1 and end up at $l > 0$ and avoid the origin all the time, then automatically $S_j > 0$. By the Reflection Lemma, this number is equal to

(2.1)

$$N_{n-1}(1 \to l) - N_{n-1}(-1 \to l) = N_{n-1}(0 \to l-1) - N_{n-1}(0 \to l+1).$$

This proves the first part of Lemma 2.2.

Exercise 2.2. Prove that if there is a path of length n leading from 0 to k, then there are nonnegative integers a, b so that $n = a + b$, $k = a - b$.

Exercise 2.3. Use the fact that

$$N_n(0 \to k) = \binom{n}{a},$$

where a is as in Exercise 2.2 to prove that the last difference from (2.1) equals $(l/n)N_n(0 \to l)$.

The solution of this Exercise then completes the proof of Lemma 2.2. □

Lemma 2.3. *For a random walk starting at the origin, we have that*

$$P(S_j \neq 0, j = 1, 2, \dots, 2n) = P(S_{2n} = 0).$$

Proof. If the origin is never visited again, then either $S_j > 0$ for all $j > 0$ or $S_j < 0$ for all $j > 0$. By symmetry, these alternatives have the same probability, so

$$P(S_j \neq 0, j = 1, 2, \dots, 2n) = 2P(S_1 > 0, \dots, S_{2n} > 0)$$

Now by Lemma 2.2,

$$\begin{aligned} P(S_j > 0, \dots, S_{2n} > 0) &= \sum_{k=1}^{n} P(S_1 > 0, \dots, S_{2n-1} > 0, S_{2n} = 2k) \\ &= \sum_{k=1}^{n} 2^{-2n}(N_{2n-1}(0 \to 2k-1) - N_{2n-1}(0 \to 2k+1)) \\ &= 2^{-2n} N_{2n-1}(0 \to 1). \end{aligned}$$

The last line follows because the sum is telescoping. We can rewrite this as

$$2^{-2n}\binom{2n-1}{n} = 2^{-2n}\frac{(2n-1)!}{n!(n-1)!} = 2^{-2n-1}\binom{2n}{n} = \frac{1}{2}P(S_{2n}=0),$$

as claimed. □

We introduce the abbreviation

$$r_{2n} = P(S_{2n}=0) = \binom{2n}{n}2^{-2n}$$

for this probability (r as in return).

3. Last visits and the arc sine law

We can now find the probability distribution for the time of the last return to the starting point. Assume that we start the random walk at the origin. Let R_{2n} be the time of the last return to the origin for a random walk of length $2n$. In other words, $S_k = 0$ for $k = R_{2n}$, but $S_k \neq 0$ for $R_{2n} < k \leq 2n$. Note also that we can have $S_k = 0$ only for even k.

Theorem 3.1. $P(R_{2n} = 2k) = r_{2k}r_{2n-2k}$.

Proof. As noted above, $R_{2n} = 2k$ precisely if $S_{2k} = 0$, but $S_{2k+1}, \ldots, S_{2n} \neq 0$. The probability that this happens is

$$P(S_{2k}=0)P(S_{2k+1},\ldots,S_{2n}\neq 0|S_{2k}=0).$$

The first probability is just r_{2k}, and the conditional probability equals r_{2n-2k} by Lemma 2.3, because the individual steps of a random walk are independent, so this is the probability that a random walk starting at the origin, of length $2n-2k$, will never visit the origin again. □

We can analyze these probabilities in the limiting case $n \to \infty$ in some detail. The argument is similar to (but actually easier than) the proof of the theorem of de Moivre-Laplace. In that proof, we saw that

$$r_{2m} \simeq \frac{1}{\sqrt{m\pi}}$$

for large m. Thus

$$r_{2k}r_{2n-2k} \simeq \frac{1}{\pi\sqrt{k(n-k)}},$$

at least if k is neither very small nor very close to n. We therefore expect that for $0 < a < b < 1$,

$$\begin{aligned} P\left(a \le \frac{R_{2n}}{2n} \le b\right) &\simeq \sum_{na \le k \le nb} \frac{1}{\pi\sqrt{k(n-k)}} \\ &\simeq \frac{1}{\pi} \int_{na}^{nb} \frac{dx}{\sqrt{x(n-x)}} \\ &= \frac{1}{\pi} \int_{a}^{b} \frac{dy}{\sqrt{y(1-y)}}. \end{aligned}$$

This analysis is correct (and can be made rigorous). Moreover, the integral can be solved explicitly. We obtain the following result:

Theorem 3.2 (Arc Sine Law for last visits)**.** *For $0 \le a < b \le 1$, we have that*

$$\begin{aligned} \lim_{n\to\infty} P\left(a \le \frac{R_{2n}}{2n} \le b\right) &= \frac{1}{\pi} \int_{a}^{b} \frac{dy}{\sqrt{y(1-y)}} \\ &= \frac{2}{\pi}\left(\arcsin\sqrt{b} - \arcsin\sqrt{a}\right). \end{aligned}$$

One remarkable feature about the limiting density is the fact that it becomes largest (in fact, infinite) as either $y \to 0$ or $y \to 1$. This means that the most probable scenarios are those where the last visit to the origin occurs either very early or very late.

An even more remarkable result states that the time spent on one side of the origin has the same limiting distribution.

Theorem 3.3 (Arc Sine Law for time spent on the positive side)**.** *Let N_n be the number of indices $k \in \{1, 2, \ldots, n\}$ for which $S_k > 0$. Then, for $0 \le a < b \le 1$,*

$$\lim_{n\to\infty} P\left(a \le \frac{N_n}{n} \le b\right) = \frac{1}{\pi} \int_{a}^{b} \frac{dy}{\sqrt{y(1-y)}}.$$

By the properties of this density observed above, that means random walks where approximately equal proportions of time are spent on either side of the origin are less likely than random walks where almost all the time is spent on one side.

We will not prove Theorem 3.3 here.

4. Gambler's ruin

You play the coin tossing game described at the beginning of these notes with a capital of a dollars. You decide that you will quit as soon as your accumulated winnings reach b dollars. What is the probability that you will go broke first? Let us denote this probability of the gambler's ruin by $r(a,b)$. In the language of random walks, we are asking: What is the probability that a random walk starting at the origin will reach $-a$ before it hits b?

This problem can be attacked by the method of difference equations: We temporarily switch notations and write r_0 instead of $r(a,b)$. More generally, we let r_k be the probability that a random walk starting at k will visit $-a$ before it visits b. In the original problem, the first step leads to either -1 or 1, with probability $1/2$ in each case, thus

$$r_0 = \frac{1}{2} r_{-1} + \frac{1}{2} r_1.$$

The same argument shows that, more generally,

$$r_k = \frac{1}{2}\left(r_{k-1} + r_{k+1}\right), \qquad k = -a+1, \ldots, b-1. \tag{4.1}$$

We also have the *boundary conditions* $r_{-a} = 1$, $r_b = 0$, which are obvious from the underlying problem, and these together with the *difference equation* (4.1) determine r_k uniquely. We can proceed as follows: An inspired guess shows that $r_k = 1$ and $r_k = k$ both solve (4.1), and the linear combination of these solutions gives us the general solution of (4.1):

$$r_k = C + Dk$$

(this can be derived more systematically by referring to general facts about such difference equations). As the final step, we still need to find the constants C, D from the boundary conditions. These now say that

$$C - aD = 1, \quad C + bD = 0,$$

and the unique solution of this system of linear equations is

$$C = \frac{b}{a+b}, \quad D = \frac{-1}{a+b}.$$

In particular, since $r(a,b) = r_0 = C = b/(a+b)$, we have indeed found the probability for the gambler's ruin:

Theorem 4.1 (Gambler's ruin)**.**

$$r(a,b) = \frac{b}{a+b}$$

We can slightly rephrase this as follows: The probability that the gambler will go broke against an adversary with fortune b is the quotient of b and all the money in play $(a+b)$, assuming that the game is continued until one of the players is ruined. This is exactly what we should have expected on intuitive grounds: If your opponent owns a fraction p of the total firepower, you will lose with probability p.

Similar methods can be used to find ruin probabilities for unfair games, where (let's say) $P(X_j = 1) = p$, $P(X_j = -1) = q$, with $p + q = 1$, but $p, q \neq 1/2$. The following exercises pursue this a little further:

Exercise 4.1. Show that the ruin probabilities r_k satisfy the difference equation

$$r_k = qr_{k-1} + pr_{k+1}$$

and the boundary conditions $r_{-a} = 1$, $r_b = 0$.

Exercise 4.2. Show that the difference equation is solved by $r_k = 1$ and $r_k = (q/p)^k$, so that the general solution is given by

$$r_k = C + D\left(\frac{q}{p}\right)^k .$$

Use the boundary conditions to find C and D and conclude that

$$r(a, b) = r_0 = \frac{s^b - 1}{s^b - s^{-a}}, \qquad s \equiv \frac{q}{p}.$$

Exercise 4.3. Suppose that the game is favorable for you ($p > q$), but your opponent has unlimited financial resources ($b \to \infty$). (This becomes relevant if you want to open a casino.) Show that then

$$r(a) = \lim_{b\to\infty} r(a, b) = \left(\frac{q}{p}\right)^a .$$

How much capital a would you need if you give yourself a hopefully inconspicuous edge of just $p = 0.51$ and you want to run a risk of at most $P = 0.01$ of going broke?

We can use ruin probabilities to derive a spectacular statement about expected visits to other sites before the first return to the starting point. Let us start out with the following simplified version of the problem: You start a random walk at the origin, and the game will stop after the first return to the origin. You get paid one dollar for each visit to the point 1. What would you have to bet to make this a fair game?

In other words, we want to know: What is the expected number of visits to 1 that occur *before* the first return to 0? Call this number X (so X is a random variable). Clearly, $P(X = 0) = 1/2$ because a visit to 1 is avoided precisely if the first step goes to -1. Similarly, $P(X = 1) = (1/2)^2$ because then the first step must go to 1, but the second step must lead back to 0, to avoid a second visit to 1. These are independent events and the individual probabilities both equal $1/2$.

More generally, this line of reasoning shows that $P(X = k) = 2^{-k-1}$ and thus

$$EX = \sum_{k=0}^{\infty} k2^{-k-1} = 1.$$

(To evaluate this series, the gadget of moment generating functions comes in handy:

$$M_X(t) = Ee^{tX} = \sum_{k=0}^{\infty} e^{tk}2^{-k-1} = \frac{1}{2-e^t} \quad (t < \ln 2),$$

so $EX = M_X'(0) = 1$.)

What happens if we replace 1 by another point: I will again start a random walk at the origin, and the game will stop as soon as I hit 0 again. You agree to pay me \$1 each time I visit $n = 1,000,000$ (say) while the game is still running. What do I have to wager to make this a fair game? Who has the edge if I bet 50c? Please take a guess before you read on.

This problem can actually be treated in the same way as the warm-up problem. Let X_n be the number of visits to n before the first return to 0. By symmetry, we can assume that $n > 0$. Also, we already discussed the case $n = 1$, so we will in fact assume that $n > 1$.

First of all, what is $P(X_n = 0)$? This can happen in two ways: Either the first step goes to -1 (probability $1/2$), or it goes to $+1$ (again the probability that this happens is $1/2$), but then the random walk returns to 0 *before* it reaches n for the first time. This latter probability, however, is a ruin probability: it equals $r(1, n-1)$. Therefore

$$P(X_n = 0) = \frac{1}{2} + \frac{1}{2}r(1, n-1) = 1 - \frac{1}{2n}.$$

If we want to have $X_n = 1$, the random walk must go to n without visiting 0 again. As we just saw, this happens with probability $1/(2n)$ (we just discussed the complement of this event). Moreover, we must then go back from n to 0 without hitting n a second time. By symmetry and independence, this event has the same probability $1/(2n)$. Therefore

$$P(X_n = 1) = \left(\frac{1}{2n}\right)^2.$$

In general, in order to have $X_n = k$, we first must go straight to n (probability $1/(2n)$), then visit n again $k-1$ times without hitting 0 (probability $1 - 1/(2n)$ each time) and finally return to 0 without visiting n again (probability $1/(2n)$). We have thus proved the following:

Theorem 4.2.

$$P(X_n = 0) = 1 - \frac{1}{2n}, \quad P(X = k) = \left(\frac{1}{2n}\right)^2 \left(1 - \frac{1}{2n}\right)^{k-1} \quad (k \geq 1)$$

In particular, we have the following amazing consequence:

Corollary 4.1. *We have $EX_n = 1$, independently of n.*

Proof. Again, this can be obtained conveniently from the moment generating function of X_n. Note that the distribution of X_n has the form

$$P(X = 0) = p, \quad P(X = k) = (1-p)^2 p^{k-1} \quad (k \geq 1).$$

We will in fact show that $EX = 1$ for such a random variable, independently of p. We compute

$$M(t) = p + (1-p)^2 \sum_{k=1}^{\infty} e^{tk} p^{k-1} = p + \frac{e^t (1-p)^2}{1 - pe^t}$$
$$= p + \frac{(1-p)^2}{e^{-t} - p} \quad (t < -\ln p),$$

so $EX = M'(0) = 1$, as claimed. □

It is actually not hard to also understand this intuitively. While it is very unlikely that you will reach a big n ($n = 1,000,000$, say) before the game stops, if you manage to get there, you can expect huge winnings because now a return to the origin is just as unlikely. To substantiate this line of reasoning, let us take a look at $P(X \geq N | X \geq 1)$ (the probability that you will win at least N dollars, given that you win something). You are at n, and you now need to return at least $N - 1$ times, so

$$P(X \geq N | X \geq 1) = \left(1 - \frac{1}{2n}\right)^{N-1}.$$

(This can also be derived more formally from Theorem 4.2: you then use the fact that the conditional probability equals $P(X \geq N)/P(X \geq 1)$ and work out these probabilities.)

If $N = \alpha n$ with $\alpha > 0$ and n is large, then, since $(1 - 1/n)^n \to 1/e$, this says that

$$P(X \geq N | X \geq 1) \simeq e^{-\alpha/2}.$$

For example, if $n = 1,000,000$ and you win at all, then you will win at least \$1,000,000 with probability approximately $e^{-1/2} \simeq 0.61$. You will win \$5,000,000 or more with probability $\simeq e^{-5/2} \simeq 0.082$. On the other hand, if you win at all, you will win less than \$1,000 (or $\alpha = 10^{-3}$) only with probability $\simeq 1 - e^{-0.0005} \simeq 5 \cdot 10^{-4} = 0.05\%$.

Incidentally, this also shows that if someone foolishly offers you to play the game with $n = 1,000,000$ and a bet of just 50c, you should only accept if your adversary has a sufficiently large fortune to make the gigantic payoffs that inevitably will come in the long run. By the same token, you must have a large fortune yourself because you might have to wait for quite a while before your first winnings come in. Without that large fortune, you could be broke before you can reap the benefits of your superior command of probability theory.