Statistical Bioinformatics, Makerere
Alignment, Random Walks, Significance

Timo Koski

TK

02.08.2018

KTH Matematik

Contents

This lecture is based on sections 7.1, 7.4, 7.5, 7.6, of Ewens and Grant. We start by simple properties of random walks and generalize them as needed in the BLAST calculations.

KTH Matematik

Introduction

In this lecture we study the theory of random walks in order to understand the significance calculations in BLAST (=Best Local Alignment Search Tool). BLAST finds alignments or does sequence similarity search by finding high scoring segments (or words) in sequences. If the sequences (a query sequence and a data base sequence) compared are long, then the probability of finding segments that are similar by chance increases. Hence it is important to be able to compute the significance of the maximal score.

KTH Matematik

Introduction

Pearson & Todd (2000) have written what follows: BLAST was developed to to identify distantly related - homologous- sequences based on sequence similarity. Two sequences are said to be homologous, if we believe that the two sequences diverged from a common ancestor in the past. Almost without exception, if two (protein) sequences share statistically significant similarity, they will share significant structural similarity.

KTH Matematik

Definitions for Random Walks

Let for $n = 1, \ldots,$

$$S_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } q = 1 - p. \end{cases}$$

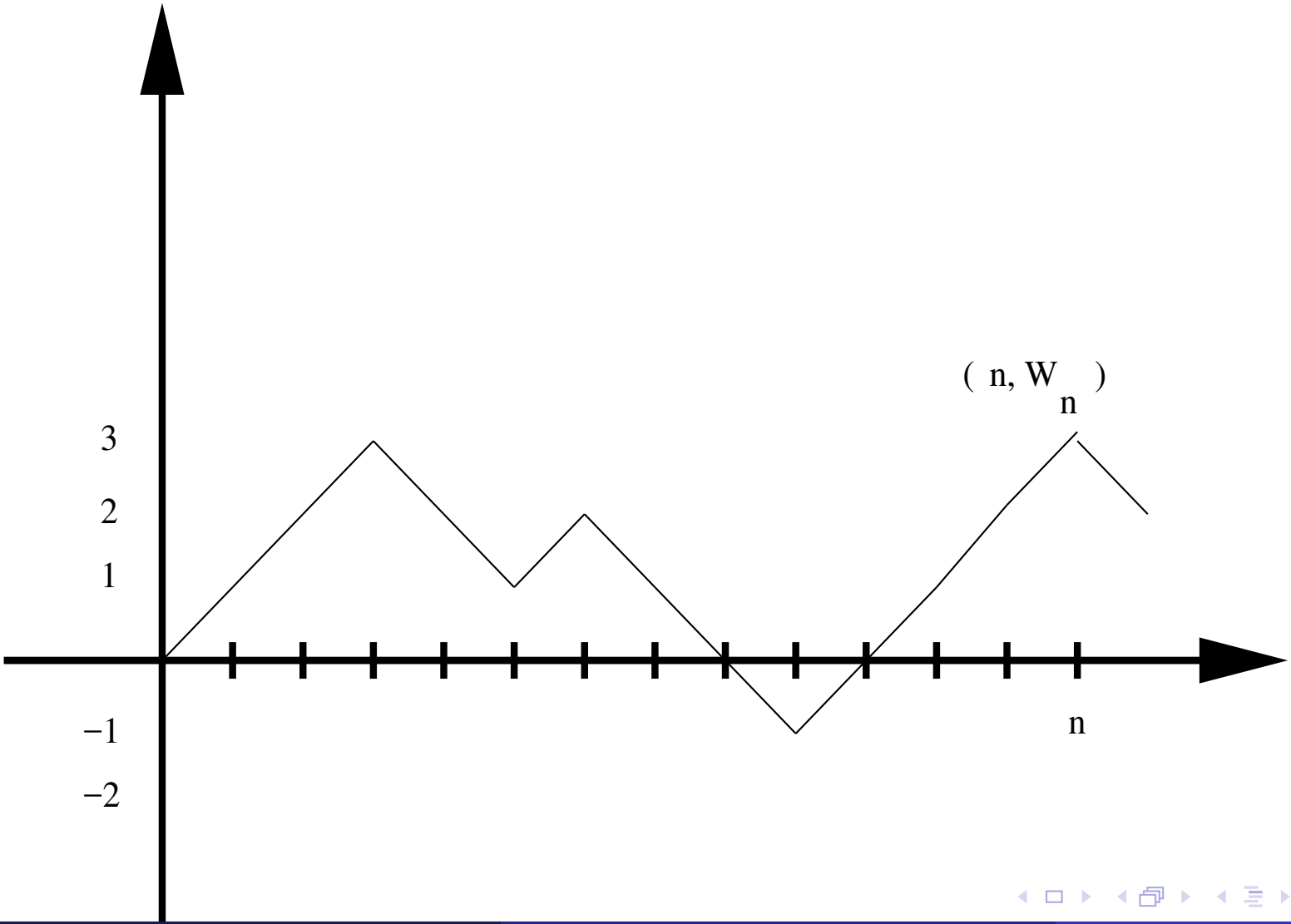
We take $W_0 = h$. Let

$$W_n = h + \sum_{i=1}^{n} S_i.$$

KTH Matematik

Definitions for Random Walks

We record the motion of the particle as the sequence $\{(n, W_n)|n =, 1, \ldots, \}$.



Random Walks

One interpretation of random walk in terms of bioinformatics is as an accumulated similarity score for two local segments of sequences:

$$W_n(= \bar{s}\left(\mathbf{x}[j:l], \mathbf{y}[j:l]\right)) = \sum_{i=j}^{l} s\left(x_i, y_i\right).$$

where

$$s(x,y) = \begin{cases} +1 & \text{if } x = y \\ -1 & x \neq y. \end{cases}$$

KTH Matematik

Properties of the simple random walk

Lemma

The simple random walk is spatially homogeneous, that is

$$P\left(W_n = j \mid W_0 = h\right) = P\left(W_n = j + b \mid W_0 = h + b\right).$$

Proof:

$$P\left(W_n = j \mid W_0 = h\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right)$$

$$P\left(W_n = j + b \mid W_0 = h + b\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right).$$

□

KTH VETENSKAP OCH KONST
KTH Matematik

Properties of the simple random walk

Lemma

The simple random walk is temporally homogeneous, that is

$$P\left(W_n = j \mid W_0 = h\right) = P\left(W_{n+m} = j \mid W_m = h\right).$$

Proof:

$$P\left(W_n = j \mid W_0 = h\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right) =$$

$$P\left(\sum_{i=m+1}^{m+n} S_i = j - h\right) = P\left(W_{n+m} = j \mid W_m = h\right).$$

KTH Matematik

Properties of the simple random walk

Lemma

The simple random walk has the Markov property, that is,

$$P\left(W_{n+1}=j \mid W_0, W_1, \ldots, W_n\right) = P\left(W_{n+1}=j \mid W_n\right).$$

Proof:

$$P\left(W_{n+1}=j \mid W_0, W_1, \ldots, W_n=a\right) =$$

$$= P\left(S_{n+1}=j-a\right) = P\left(S_{n+1}=j-a \mid W_n=a\right) = P\left(W_{n+1}=j \mid W_n=a\right)$$

□

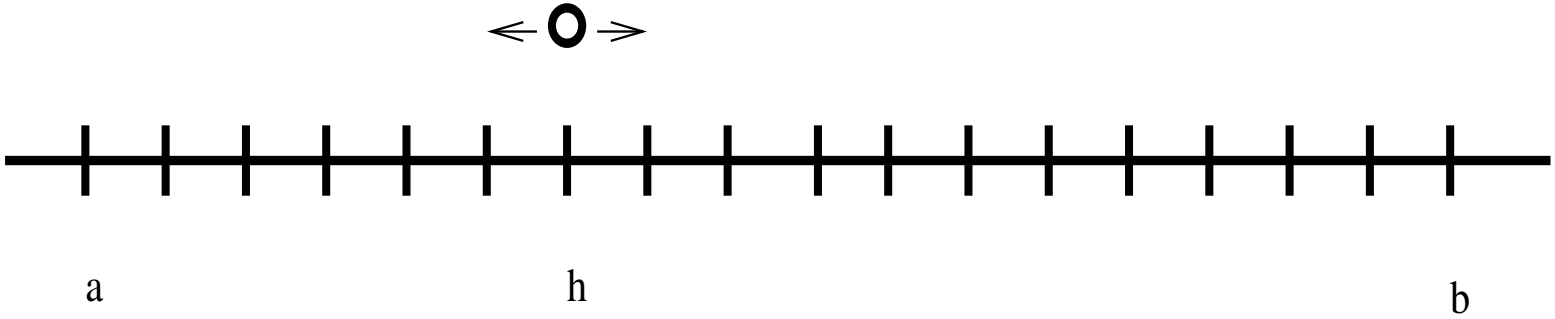
Properties of the simple random walk

Hence we are dealing with a Markov chain, whose state space is the set of integers and $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Absorption probabilities: the m.g.f. method

Let a and b be two integers, $a < b$. We take these points as barriers of absorption in the sense that when the simple random walk hits either of these points it will terminate there.



Absorption probabilities: the m.g.f. method

This is a Markov chain, whose state space is the set of integers $= \{a, \ldots, b\}$ and the transition probability matrix of which is of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & q & 0 & p \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Two problems

(i) finding the probability that the walk stops at b rather than at a.

(ii) finding the mean number of steps until the walk terminates.

We use the moment generating functions for both of these problems.

KTH VETENSKAP OCH KONST
KTH Matematik

Moment Generating Functions

The m.g.f. of any single step S is obviously

$$m(\theta) = qe^{-\theta} + pe^{\theta}.$$

We recall a theorem from the first lecture to show that there exists θ^* such that

$$m(\theta^*) = 1.$$

In fact this is a quadratic equation solved by

$$\theta^* = \log\left(\frac{q}{p}\right).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Markov times

Let

$$N = \min\{n \geq 0 \mid W_n = a \quad \text{or} \quad W_n = b\}.$$

Then N is the random number of steps until the simple random walk terminates and is an example of a **Markov time**. The m.g.f. of the total displacement

$$T_N = \sum_{i=1}^{N} S_i$$

after N steps is, since single steps are I.I.D., by a previous result

$$m(\theta)^N = \left(qe^{-\theta} + pe^{\theta}\right)^N$$

KTH VETENSKAP OCH KONST
KTH Matematik

Moment Generating Functions

Hence

$$m(\theta^*)^N = \left(qe^{-\theta^*} + pe^{\theta^*}\right)^N = 1.$$

We get also

$$w_h = P(W_N = b) = P(T_N = b - h),$$

and

$$u_h = P(W_N = a) = P(T_N = a - h) = 1 - w_h.$$

Moment Generating Functions

The moment generating function of T_N is thus

$$m_{T_N}(\theta) = w_h \cdot e^{(b-h)\theta} + u_h e^{(a-h)\theta}$$

$$= w_h \cdot e^{(b-h)\theta} + (1 - w_h) e^{(a-h)\theta}.$$

Then also

$$m_{T_N}(\theta^*) = m(\theta^*)^N = 1,$$

which gives

KTH Matematik

Moment Generating Functions

$$w_h \cdot e^{(b-h)\theta^*} + (1 - w_h) e^{(a-h)\theta^*} = 1$$

$$\Leftrightarrow$$

$$w_h \cdot \left(e^{b\theta^*} - e^{a\theta^*} \right) = e^{h\theta^*} - e^{a\theta^*}$$

$$\Leftrightarrow$$

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{\left(e^{b\theta^*} - e^{a\theta^*} \right)}.$$

KTH Matematik

The probability of absorption

Hence we have found the answer to (i) above.

Proposition

The probability of absorption at b is

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{\left(e^{b\theta^*} - e^{a\theta^*}\right)}. \tag{1}$$

□

KTH VETENSKAP OCH KONST
KTH Matematik

The expected time to absorption

In order to compute the mean number of steps until the walk terminates, i.e., $E[N]$, we need a result known as Wald's identity:

$$E\left[m(\theta)^{-N} e^{\theta T_N}\right] = 1 \tag{2}$$

for all θ that the m.g.f. exists.

The other problem

Next we differentiate Wald's identity with respect to θ.

$$\frac{d}{d\theta} E\left[m(\theta)^{-N} e^{\theta T_N}\right] = E\left[\frac{d}{d\theta} m(\theta)^{-N} e^{\theta T_N}\right] = 0,$$

where

$$\frac{d}{d\theta} m(\theta)^{-N} e^{\theta T_N} =$$

$$-Nm(\theta)^{-N-1} \frac{d}{d\theta} m(\theta) e^{\theta T_N} + m(\theta)^{-N} T_N e^{\theta T_N}$$

KTH VETENSKAP OCH KONST
KTH Matematik

The expected time to absorption

Hence

$$E\left[-Nm\left(\theta\right)^{-N-1}\frac{d}{d\theta}m\left(\theta\right)e^{\theta T_N}+m\left(\theta\right)^{-N}T_N e^{\theta T_N}\right]=0.$$

If we insert $\theta = 0$ in this equation we get

$$E\left[-NE(S)+T_N\right]=-E\left[N\right]\cdot E[S]+E\left[T_N\right]=0,$$

which gives

$$E\left[T_N\right]=E\left[N\right]\cdot E[S].$$

Hence

$$E\left[N\right]=\frac{E\left[T_N\right]}{E[S]}.$$

The expected time to absorption

The expected time to absorption is thus the ratio of the expected displacement at absorption to the expected step size. But we know that

$$E[T_N] = w_h(b-h) + u_h(a-h),$$

and

$$E(S) = p - q.$$

Hence we have found the solution to (ii) above.

KTH VETENSKAP OCH KONST
KTH Matematik

The expected time to absorption

Proposition

The mean time to absorption is

$$E[N] = \frac{w_h(b-h) + u_h(a-h)}{p-q}. \tag{3}$$

□

KTH Matematik

Asymptotic case

Let us suppose $h = 0$, $a = -1$ and let $b = y$. We are going to let $y \to +\infty$. By this we want to compute

(iii) the distribution of the maximum value of the walk before it hits $a = -1$.

(iv) the mean number of steps until the walk eventually terminates at $a = -1$.

We discuss (iii), i.e., the maximum value of the walk ever reaches before hitting -1.

Asymptotic case

We have

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{\left(e^{b\theta^*} - e^{a\theta^*}\right)},$$

which with $h = 0$, $a = -1$ and $b = y$ yields

$$w_0 = \frac{1 - e^{-\theta^*}}{\left(e^{y\theta^*} - e^{-\theta^*}\right)}.$$

Hence we can see that for large y

$$w_0 \propto \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic case

Hence, if Y is the maximum of the walk,

$$\text{Prob}\left(Y \geq y\right) \sim \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}$$

as $y \longrightarrow \infty$, this is a geometric-like probability, where we put

$$C \stackrel{\text{def}}{=} \left(1 - e^{-\theta^*}\right). \tag{4}$$

KTH Matematik

Asymptotic case

For (iv) we let

$$A \stackrel{def}{=} E[N]. \tag{5}$$

Then with $a = -1$, $h = 0$ and $b = y$

$$A = \frac{w_h(b-h) + u_h(a-h)}{p-q}$$

becomes

$$A = \frac{w_0 y - u_0}{p-q} = \frac{u_0 - w_0 y}{q-p}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic case

$$A = \frac{w_0 y - u_0}{p - q} = \frac{u_0 - w_0 y}{q - p}.$$

In the expression above $w_0 y \to 0$, as $y \to \infty$ and $u_0 = 1 - w_0 \to 1$, since $w_0 \to 0$. Hence

$$A = \frac{1}{q - p}$$

as $y \to \infty$.

KTH VETENSKAP OCH KONST
KTH Matematik

General Case

We generalize the results in the asymptotic case above to a general random walk with barrier at -1. In a general random walk the step size may be any of general finite number of integers.
The final goal is to show that the maximum excursion of the random walk will have a geometric like distribution

$$\text{Prob}\,(Y \geq y) \sim Ce^{-y\theta^*}$$

and to find C in this expression. We are going to use a renewal theorem.

KTH VETENSKAP OCH KONST
KTH Matematik

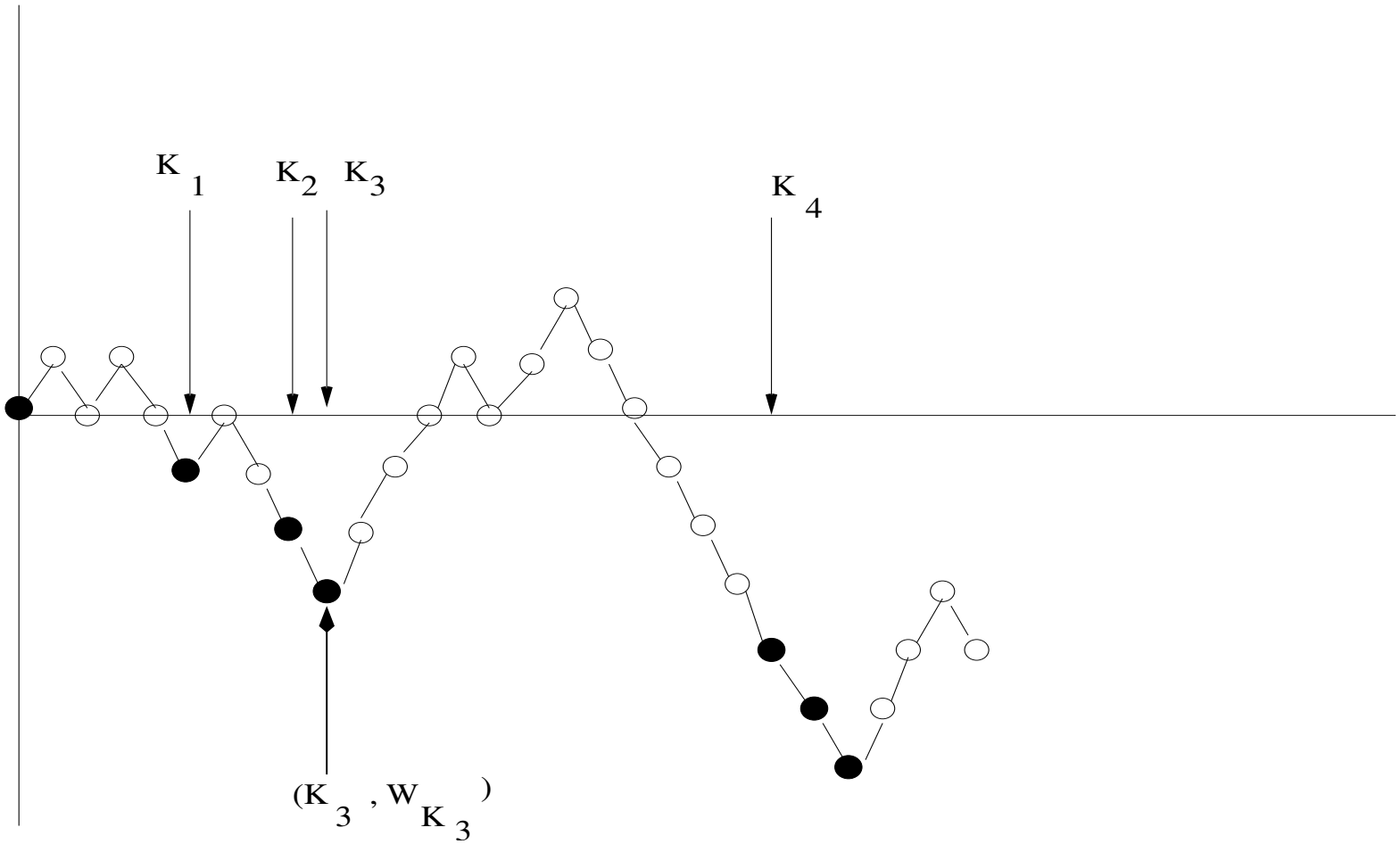
Ladder Points

Ladder points in a path of a random walk are defined as follows: **a ladder point is a point** (n, W_n)**, which is lower than than any previously reached point.**

In the figure (almost like the Figure 7.1 in Grant and Ewens p.220) the ladder points are the filled circles.

KTH VETENSKAP OCH KONST
KTH Matematik

Ladder Points



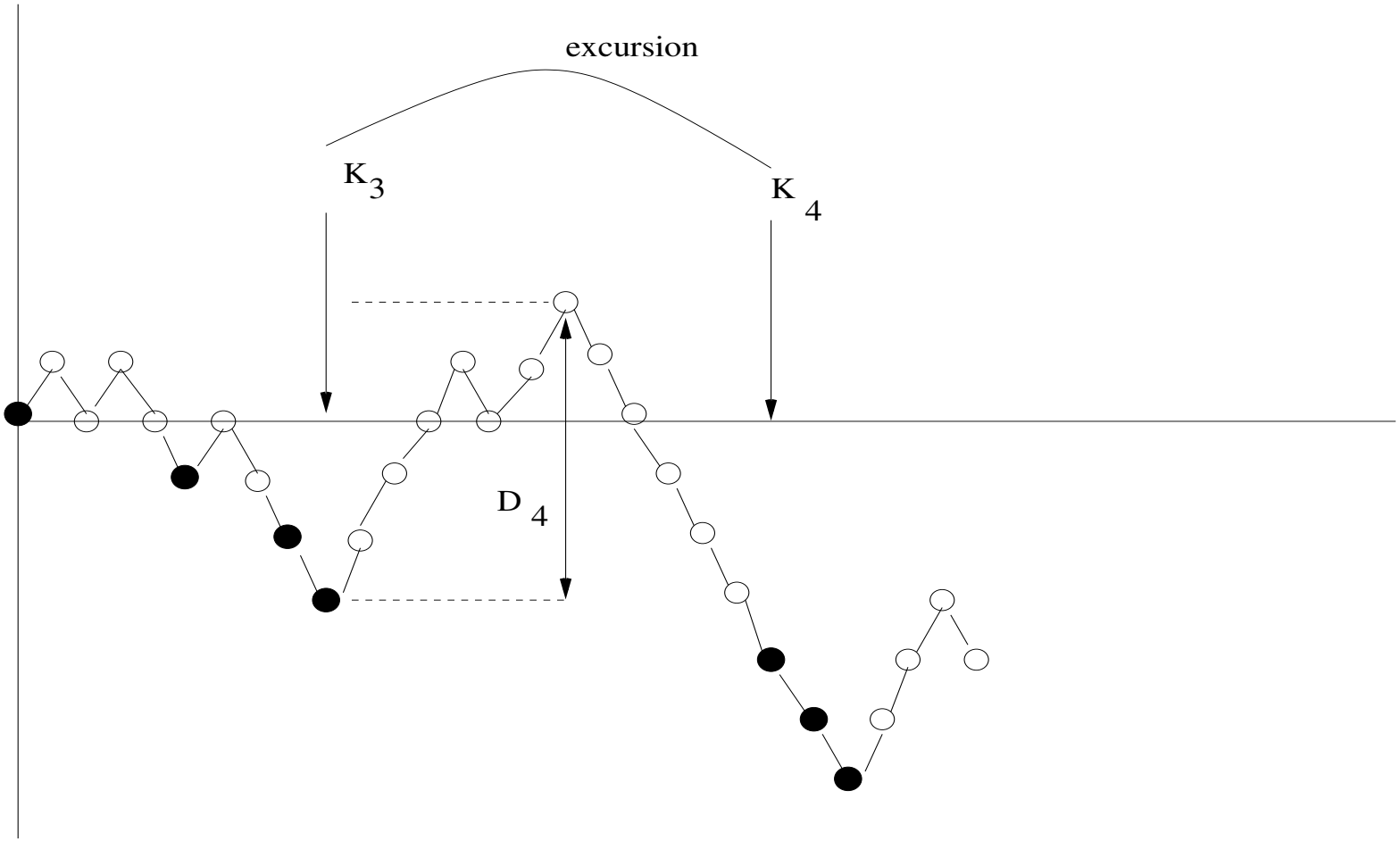
Excursions

An **excursion** is the part of the path of a random walk from a ladder point until the highest point attained before the next ladder point.

BLAST theory is concerned with the heights of these excursions: in the figure the height of the excursion indicated is 5. If one ladder point follows immediately after a ladder point, the height is taken as 0.

KTH VETENSKAP OCH KONST
KTH Matematik

Excursions



MaximaL displacement

$$D_v = \max_{K_{v-1} \leq l < K_v} (W_l - W_{K_{v-1}})$$

evaluates the **maximal displacement attained during the vth excursion**,

Path Maximum

$$M(n) = \max_{0 \le k \le l \le n} (W_l - W_k) .$$

Thus $M(n)$ **corresponds to a segment of the sequence** $S_1, S_2, \ldots, S_n$ **with maximal accumulated score**.

KTH VETENSKAP OCH KONST
KTH Matematik

Path Maximum

It follows from defnition of D_v that

$$M(n) = \max\left(D_1, D_2, \ldots, D_{R(n)}, D_n^*\right),$$

where $R(n)$ is the number of complete excursions over the time frame from 1 to n and D_n^* is the residual maximum during the last incomplete excursion up to time n.

KTH VETENSKAP OCH KONST
KTH Matematik

Path Maximum

We shall use the following notations

$$F(y) = P(D_1 \leq y)$$

and

$$A = E[K_1].$$

Path Maximum

We shall prove the following

$$\lim_{y \to \infty} \left[1 - F(y)\right] e^{\theta^* y} = C \cdot e^{-\theta^*}.$$

KTH Matematik

Unrestricted Walks

We start with the unrestricted general walks, that is, for random walks without barriers. Suppose that the possible step sizes in a random walk are

$$-c, -c+1, -c+2, \ldots, 0, \ldots, d-1, d$$

with the respective probabilities

$$p_{-c}, p_{-c+1}, p_{-c+2}, \ldots, 0, \ldots, p_{d-1}, p_d$$

KTH Matematik

Unrestricted Walks: assumptions

We assume that

(a) $p_{-c} > 0$ and $p_d > 0$.

(b) $E(S) = \sum_{j=-c}^{d} j p_j < 0$.

(c) the greatest common divisor of the step sizes that have non-zero probability is 1.

KTH VETENSKAP OCH KONST
KTH Matematik

Unrestricted Walks: m.g.f.

The m.g.f. is

$$m(\theta) = \sum_{j=-c}^{d} p_j e^{j\theta}.$$

By (a) and (b) we have that there is θ^* for which

$$\sum_{j=-c}^{d} p_j e^{j\theta^*} = 1.$$

Since $E(S) < 0$ the random walk will eventually drift down to $-\infty$. Before doing so, it might visit various positive values.

KTH VETENSKAP OCH KONST
KTH Matematik

Unrestricted Walks

The first goal is to find an equation for the probabilities Q_k, where

$$Q_k = \text{Prop}\,(\text{ the walk visits k before any other positive value }\,)$$

Since the largest positive step size is d we have $Q_k = 0$ for $k > d$. Obviously

$$\sum_{k=1}^{d} Q_k < 1.$$

We shall find the sought equation by introducing artificial barriers.

KTH VETENSKAP OCH KONST
KTH Matematik

Unrestricted Walks: Artificial barriers

A technical step: let us impose an artifical boundary at $+1$ and another at $-L$, where $L >> 0$. Then by the same argument as above and by applying Wald's identity we obtain

$$\sum_{k=-L-c+1}^{-L} Q_k(L)e^{k\theta^*} + \sum_{k=1}^{d} Q_k(L)e^{k\theta^*} = 1$$

where $Q_k(L)$ is the probability (depending on L) that the walk stops at the value k.

KTH Matematik

Unrestricted Walks: Artificial barriers

If the walk stops at $k > 0$, then k is the first positive value reached, as $+1$ is the boundary.

$$\lim_{L\to\infty} Q_k(L) = Q_k$$

Since $\theta > 0$ and $Q_k(L) < 1$, we that

$$\lim_{L\to\infty} \left[\sum_{k=-L-c+1}^{-L} Q_k(L) e^{k\theta^*} + \sum_{k=1}^{d} Q_k(L) e^{k\theta^*} \right]$$

$$= \sum_{k=1}^{d} Q_k e^{k\theta^*} = 1.$$

Probability for maximum excursion

Next we want to find an expression for

$$F_Y(y)$$

which is the probability that maximum excursion by the random walk is y or less.

Probability for maximum excursion

The event that the maximum upward excursion is y or less is the union of several non-overlapping events.
The first of these is the event that the maximum excursion never reaches positive values which has probability

$$\overline{Q} = 1 - Q_1 - Q_2 - \ldots - Q_d.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Probability for maximum excursion

The rest of the events are that the first positive value obtained by the excursion is k, $k = 1, 2, \ldots, y$ (with probability Q_k) and then, starting form the first positive value, the walk never achieves a further height exceeding $y - k$ (probability $F_Y(y - k)$).

KTH Matematik

Probability for maximum excursion

Then

$$F_Y(y) = \overline{Q} + \sum_{k=0}^{y} Q_k F_Y(y-k) \tag{6}$$

Let us introduce

$$V(y) \stackrel{\text{def}}{=} (1 - F_Y(y))\, e^{y\theta^*}$$

or

$$F_Y(y) = 1 - V(y) e^{-y\theta^*}. \tag{7}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Probability for maximum excursion

Then we get from the above

$$1 - V(y)e^{-y\theta^*} = \overline{Q} + \sum_{k=0}^{y} Q_k \left(1 - V(y-k)e^{-(y-k)\theta^*}\right)$$

which is reorganized as

$$V(y) = e^{y\theta^*}\left(Q_1 + Q_2 + \ldots + Q_d\right) + \sum_{k=0}^{y} e^{k\theta^*} Q_k V(y-k)$$

when $y < d$ and

$$V(y) = \sum_{k=0}^{d} e^{k\theta^*} Q_k V(y-k)$$

when $y \geq d$.

Probability for maximum excursion

These quantities are in the form applicable in a **renewal theorem**, where

$$f_k = e^{k\theta^*} Q_k$$

and

$$b_y = e^{y\theta^*} \left(Q_1 + Q_2 + \ldots + Q_d\right),$$

if $y < d$ and $b_y = 0$ if $y \geq d$.

KTH VETENSKAP OCH KONST

KTH Matematik

The Renewal Theorem

Suppose the three sequences $\{b_j\}_{j=0}^{\infty}$, $\{f_j\}_{j=0}^{\infty}$ and $\{u_j\}_{j=0}^{\infty}$ of nonnegative constants satisfy the equation

$$u_y = b_y + \left(\sum_{k=0}^{y} u_{y-k} f_y \right)$$

for all y. Suppose further that

$$B = \sum_{k=0}^{\infty} b_k < +\infty,$$

$$\sum_{k=0}^{\infty} f_k = 1,$$

and

$$\mu = \sum_{k=0}^{\infty} k f_k < +\infty,$$

and that the greatest common divisor of $\{f_j\}_{j=0}^{\infty}$ is one.

KTH VETENSKAP OCH KONST
KTH Matematik

The Renewal Theorem

Then

$$u_y \to \frac{B}{\mu} \tag{8}$$

as $y \to +\infty$ □

Probability for maximum excursion

We must find

$$B = \sum_y b_y = \sum_{k=0}^{d} e^{k\theta^*} \left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right).$$

If we multiply this equation by $e^{\theta^*} - 1$, we have

$$\sum_{k=1}^{d} Q_k e^{k\theta^*} - (Q_1 + Q_2 + \ldots) = 1 - (Q_1 + Q_2 + \ldots) = \overline{Q}.$$

KTH Matematik

Probability for maximum excursion

Hence

$$B = \frac{\overline{Q}}{e^{\theta^*} - 1}.$$

The renewal theorem in Appendix A below, we should verify the conditions of the theorem, but we skip the details in this document) implies then that if

$$V = \lim_{y \to \infty} V(y),$$

then

$$V = \frac{\overline{Q}}{(e^{\theta^*} - 1) \sum_{k=0}^{d} k Q_k e^{k\theta^*}} \tag{9}$$

KTH VETENSKAP OCH KONST

KTH Matematik

Restricted Walks

Consider a restricted walk having a stopping boundary at -1. We assume that the walk starts at 0. Let

$$F^*_{Y_{\text{unr}}}(y) = 1 - F_{Y_{\text{unr}}}(y)$$

where

$$F_{Y_{\text{unr}}}(y)$$

is the probability of that maximum excursion by the unrestricted random walk is y or less, treated in the preceding subsection. Hence $F^*_{Y_{\text{unr}}}(y)$ is the probability that the maximum excursion by the unrestricted random walk exceeds y. Also

$$F^*_Y(y) = 1 - F_Y(y)$$

holds for the restricted random walk treated now (a stopping boundary at -1).

KTH VETENSKAP OCH KONST
KTH Matematik

Restricted Walks

Then from the above

$$F_{Y_{\text{unr}}}(y) = V(y)e^{-y\theta^*}$$

and

$$\lim_{y\to\infty} F_{Y_{\text{unr}}}(y)e^{y\theta^*} = V. \tag{10}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Restricted Walks

The size of an excursion of the unrestricted walk can exceed the value y either before or after the random walk has reached negatiev values (starting at 0). In the latter case the first native value reached by the random walk is one of $-1, -2, \ldots, -c$. If the probability that it is $-j$ is R_j, then

$$F^*_{Y_{\text{unr}}}(y) = F^*_Y(y) + \sum_{j=1}^{c} R_{-j} F^*_{Y_{\text{unr}}}(y+j)$$

(Yes, up till the time that the unrestricted random walk hits $-1, -2, \ldots, -c$, the unrestricted and restricted random walks are identical).

KTH Matematik

Restricted Walks

Then we multiply by $e^{y\theta^*}$ throughout in this equation and we get, since

$$F_{Y_{\text{unr}}}(y+j) = V(y+j)e^{-(y+j)\theta^*},$$

that

$$V = \lim_{y\to\infty} F_Y^*(y) + V\sum_{j=1}^{c} R_{-j}e^{-j\theta^*}$$

and from this

$$\lim_{y\to+\infty} F_Y^*(y) = V\left[1 - \sum_{j=1}^{c} R_{-j}e^{-j\theta^*}\right].$$

KTH VETENSKAP OCH KONST
KTH Matematik

Restricted Walks

$$\lim_{y\to\infty} F_Y^*(y) = \frac{\overline{Q}\left(1-\sum_{j=1}^{c} R_{-j}e^{-j\theta^*}\right)}{(e^{\theta^*}-1)\sum_{k=0}^{d} kQ_k e^{k\theta^*}}.$$

This we write as

$$\lim_{y\to\infty} F_Y^*(y)e^{y\theta^*} = C\cdot e^{-\theta^*}$$

where

$$C = \frac{\overline{Q}\left(1-\sum_{j=1}^{c} R_{-j}e^{-j\theta^*}\right)}{(e^{\theta^*}-1)\sum_{k=0}^{d} kQ_k e^{k\theta^*}}$$

KTH Matematik

Restricted Walks

We recall that

$$F_Y^*(y) = \text{Prob}\,(Y \geq y + 1)$$

and hence

$$\text{Prob}\,(Y \geq y) \sim Ce^{-y\theta^*}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

End of Lecture