Expected hitting time of a pattern from a finite alphabet set

Setup:

Let Ω be a finite alphabet set

Consider a pattern ψ such that $\psi \in \Omega^m$, for some $m \in \mathbb{N}$

$\{X_n\}_{n \geq 0}$ for $X_n \overset{iid}{\sim} Uniform(\Omega)$

$T = \min\{k \geq m, k \in \mathbb{N} : (X_{k-m+1}, \ldots X_k) = \psi\}$

Objective: $E(T)$?

- Example 1 : Markov chain techniques, $\psi = ABA$
- Example 2 : Martingale theory, $\psi = ABRACADABRA$
- Result for arbitrary patterns

Markov chain technique :

Example 1 :

$\Omega = \{A, B, C, \dots, X, Y, Z\}$ and $\Psi = ABA$

X_n is iid Uniform(Ω) and $T = \min\{k \geq 3 : (X_{k-2}, X_{k-1}, X_k) = ABA\}$

$\dots * * * A\ B\ A$ $\qquad$ $*** \to **A \to *AB \to ABA$

We partition Ω^3 into $S = \{***,\ **A,\ *AB,\ ABA\}$

Let $Y_n = (X_n, X_{n+1}, X_{n+2})$, for $n \geq 0$. We will consider $\{Y_n\}_{n\geq 0}$ as a Markov chain on S with ABA being the absorbing state

	$***$	$**A$	$*AB$	ABA
$***$	$25/26$	$1/26$	0	0
$**A$	$24/26$	$1/26$	$1/26$	0
$*AB$	$25/26$	0	0	$1/26$
ABA	0	0	0	1

$|ABA| = 1$

$|*AB| = 26$

$|**A| = 26^2 - 1$

$|***| = 26^3 - 26^2 - 26$

Let $T' = \min\{k \geq 0 : Y_k = ABA\}$ and

$h(c) = E(T' \mid Y_0 = c)$, for $c \in \{***, **A, *AB, ABA\}$

Using one step probability calculation, $\qquad$ References: HW2 - Pb2

$h(ABA) = 0$

$h(***) = 1 + \frac{25}{26} h(***) + \frac{1}{26} h(**A) + 0 \times h(*AB) + 0 \times h(ABA)$

$h(**A) = 1 + \frac{24}{26} h(***) + \frac{1}{26} h(**A) + \frac{1}{26} h(*AB) + 0 \times h(ABA)$

$h(*AB) = 1 + \frac{25}{26} h(***) + 0 \times h(**A) + 0 \times h(*AB) + \frac{1}{26} h(ABA)$

$$\boxed{\begin{array}{ll} h(***) = 26^3 + 26 & h(**A) = 26^3 \\ h(*AB) = 26^3 - 26^2 + 26 & h(ABA) = 0 \end{array}}$$

It takes 3 steps to reach the initial state c

$$E(T) = E(T') + 3$$

$$E(T') = \sum_{c\in S} \mathbb{P}(Y_0 = c)\, E(T' \mid Y_0 = c) = \sum_{c\in S} \mathbb{P}(Y_0 = c)\, h(c)$$

But, $\mathbb{P}(Y_0 = c) = \frac{|c|}{|\Omega^3|} = \frac{|c|}{26^3}$ and

$$\mathbb{P}(Y_0 = ABA) = \frac{1}{26^3} \quad ; \quad \mathbb{P}(Y_0 = *AB) = \frac{26}{26^3}$$

$$\mathbb{P}(Y_0 = **A) = \frac{26^2-1}{26^3} \quad ; \quad \mathbb{P}(Y_0 = ***) = \frac{26^3 - 26^2 - 26}{26^3}$$

Then, $E(T') = 0 \times \frac{1}{26^3} + \frac{26(26^3 - 26^2 + 26)}{26^3} + \frac{(26^2-1)(26)}{26^3} + \frac{(26^3-26^2-26)(26^3+26)}{26^3}$

$$= \quad E(T') = 26^3 + 26 - 3$$

$$\therefore E(T) = 26^3 + 26$$

Process:

- Setup as a markov chain with ψ being the absorbing state.
- Calculate transition probabilities and category sizes.
- Do one step probability calculation to find $h(c)$.
- Calculate $E(T')$ and then $E(T)$.

Disadvantages:

- For pattern of length n, need to solve $(n+1)\times(n+1)$ system of linear equations.
- Can get very calculation intensive.

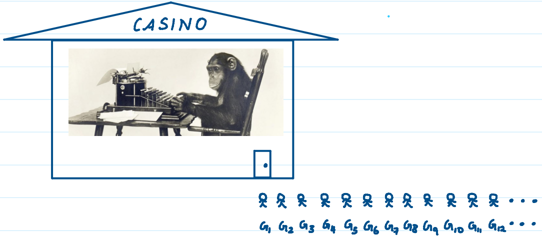
Martingale Method:

Example 2:

$\Omega = \{A, B, C, \dots, X, Y, Z\}$ and Ψ = ABRACADABRA

X_n is iid Uniform(Ω) and $T = \min\{k \geq 11 : (X_{k-10}, \dots, X_k) = \Psi\}$

Consider a casino with a monkey and a typewriter inside it. The typewriter has just the 26 english alphabets on it and the monkey types one letter randomly every second.



At every time k, $k \geq 1$, a new gambler walks into the casino and bets ₹1 that X_k, the next letter the monkey will type is going to be 'A'. If they are correct, they win ₹26 otherwise they leave the game.

Why ₹26? Expected profit $= \frac{1}{26}(26-1) + \frac{25}{26}(-1) = 0$

If the gambler was correct, then in the next time step (next round) they bet the entire ₹26 that the upcoming letter, X_{k+1} will be 'B' (for a prize of ₹26^2) and so on.

This goes on until the word ABRACADABRA appears and then the game stops.

Example: In the case when $X_1 = A$, $X_2 = B$, $X_3 = A$, $X_4 = D$, $X_5 = B$ . . .

Time	1	2	3	4	5	. . .
X_k	A	B	A	D	B	. . .
G_1	1	26	26^2	0	0	. . .
G_2	0	1	0	0	0	. . .
G_3	0	0	1	26	0	. . .
G_4	0	0	0	1	0	. . .
G_5	0	0	0	0	1	. . .
⋮	⋮	⋮	⋮	⋮	⋮	

Let Z_n^k be the amount of money the k^{th} gambler has at the end of time n

$$Z_n^k = \begin{cases} 1 & , \text{ if } n < k \\ 26\, Z_{n-1}^k \, 1_{\{X_n = \psi_{n-k+1}\}} & , \text{ if } k \le n \le k+10 \\ Z_{k+10}^k & , \text{ if } n \ge k+11 \end{cases}$$

Then, the net profit the casino makes from the k^{th} gambler at the end of the n^{th} round is: $Z_0^k - Z_n^k = 1 - Z_n^k$

Therefore, the total profit the casino makes from all the gamblers, at the end of the n^{th} round is:

$$M_n = \sum_{k=1}^{n} (1 - Z_n^k) = n - \sum_{k=1}^{n} Z_n^k$$

- $\{M_n\}_{n\ge1}$ is a martingale w.r.t the filtration $\mathcal{A}_n$

defined by the set of observable events till time n

Since $E(|M_n|) < \infty$, $\forall\, n \in \mathbb{N}$ and $E(M_n|\mathcal{A}_{n-1}) = M_{n-1}$

M_n is a function of $(X_1, X_2, \dots X_n)$

- T is a stopping time wrt $\{M_n\}_{n\ge1}$

We will find $E(T)$ using $E(M_T)$

How to find $E(M_T)$? We will use OST.

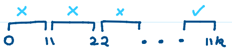
To be able to use OST, we have to show:

① $E(T)$ is finite

② $|M_n - M_{n-1}| \le c$, $\forall\, n \ge 1$, that is, M has bounded increments

To show $E(T)$ is finite:

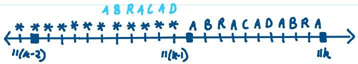
Consider blocks of size 11



Success if block $= \psi$ and failure otherwise

S is the number of blocks until the first success

So, $S = k \Rightarrow T \le 11k$



Hence, $T \le 11S \Rightarrow E(T) \le 11E(S)$

But $S \sim \text{Geometric}\left(\frac{1}{26^{11}}\right) \Rightarrow E(T) \le 11(26)^{11}$

$\therefore E(T) < \infty$ ①

To show M has bounded increments:

For $n>1$, consider $|M_n - M_{n-1}| = \left| n - \sum_{k=1}^{n} Z_n^k - n + 1 + \sum_{k=1}^{n-1} Z_{n-1}^k \right|$

$$\leq \left| \sum_{k=1}^{n-1} (Z_{n-1}^k - Z_n^k) \right| + |Z_n^n - 1|$$

$$\leq \left| \sum_{k=n-10}^{n-1} (Z_{n-1}^k - Z_n^k) \right| + 25 \quad \left[\text{Since } Z_{n-1}^k = Z_n^k,\ \text{for } n \geq k+10 \text{ and } Z_n^n \leq 26\right]$$

$$= \left| \sum_{k=n-10}^{n-1} \left(Z_{n-1}^k - 26\, Z_{n-1}^k \, 1_{\{U_n = W_{n-k+1}\}} \right) \right| + 25$$

$$\leq 25(1 + 26 + 26^2 + \dots + 26^{10}) \quad \left[\text{Since } Z_n^{n-k+1} \leq 26^k\right]$$

$\Rightarrow$ $|M_n - M_{n-1}| \leq 26^{11} - 1$, $\forall n \in \mathbb{N}$ and hence, M has bounded increments. — ②

So, ① and ② $\Rightarrow$ The conditions for OST, i.e., $E(M_T \mid T>n)\mathbb{P}(T>n) \to 0$ as $n \to \infty$ and $E(|M_T|) < \infty$

Hence, $E(M_T) = E(M_1)$

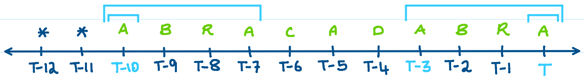
$$= 1 - E(Z_1^1) = 1 - \left(\frac{1}{26}(26) + \frac{25}{26}(0) \right)$$

$\Rightarrow$ $E(M_T) = 0$ ③

How are M_T and T related?

$$M_T = T - \sum_{k=1}^{T} Z_T^k = T - \sum_{k=T-10}^{T} Z_T^k$$

This is indeed the case, as for all gamblers who started betting before time $T-10$, $Z_T^k = 0$



Clearly, $Z_T^k = 0$ for $T-10 \leq k \leq T$ and $k \notin \{T, T-3, T-10\}$.

Then, $Z_T^T = 26$, $Z_T^{T-3} = 26^4$, $Z_T^{T-10} = 26^{11}$

$$\therefore M_T = T - 26^{11} - 26^4 - 26$$

Therefore, $E(M_T) = 0 = E(T - 26^{11} - 26^4 - 26)$

Hence, $E(T) = 26^{11} + 26^4 + 26$

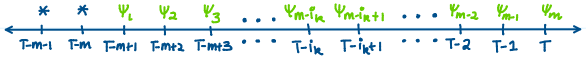
General Result:

$X_n \overset{iid}{\sim} \text{Uniform}(\Omega)$ and $\Psi \in \Omega^m$, for some $m \in \mathbb{N}$.

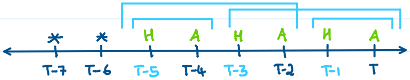
$T = \min\{k \geq m,\ k \in \mathbb{N} : (X_{k-m+1}, \ldots, X_{k-1}, X_k) = \Psi\}$

<u>Theorem 1</u>: Consider Ψ, a pattern of length $m \in \mathbb{N}$. Suppose there exist $0 < i_1 < i_2 < \ldots < i_j < m$, for $0 \leq j < m$, such that, the first i_k letters of the pattern Ψ is equal to the last i_k letters of the pattern Ψ, for $\forall\, 1 \leq k \leq j$. That is, $(\Psi_1, \Psi_2, \ldots, \Psi_{i_k}) = (\Psi_{m-i_k+1}, \Psi_{m-i_k+2}, \ldots, \Psi_m)$, $\forall\, 1 \leq k \leq j$. Then,

$$E(T) = |\Omega|^m + \sum_{k=1}^{j} |\Omega|^{i_k}$$



Examples: For $\Psi =$ HAHAHA, $E(T) = 26^6 + 26^4 + 26^2$



But for $\Psi =$ ABCDEF, $E(T) = 26^6$

Sketch of proof:

- Casino with monkey and typewriter setup as before.
- Casino closes when the word Ψ appears and incoming gamblers bet on the letters of the word Ψ.
- Z_n^k defined appropriately and $M_n = n - \sum_{k=1}^{n} Z_n^k$
- Show that $\{M_n\}_{n \geq 1}$ is a Martingale wrt $\mathcal{F}_n$
- Show that $E(T) < \infty$
- Show that M has bounded increments, i.e., $|M_n - M_{n-1}| \leq c$, $\forall\, n \geq 1$
- Using OST, we have $E(M_T) = E(M_1) = 0$
- Show that $M_T = T - |\Omega|^{|\Psi|} - \sum_{k=1}^{j} |\Omega|^{i_k}$
- And hence, $E(T) = |\Omega|^{|\Psi|} + \sum_{k=1}^{j} |\Omega|^{i_k}$