Expected waiting times for strings of coin flips.

Coin flips are independent trials with $p(H) = p$, $p(T) = q = 1- p$, $0 < p < 1$.

1. Given a finite string s of H's and T's, find a simple algorithm for
E(s) = (expected waiting time for first occurrence of s in a sequence of coin flips).

2. Fix n. Find **Ex**(n) = (expected waiting time for a random string s of length n).

Notation

Let $\boldsymbol{R}(n) = \{s \mid s$ is a finite string of H's and T's of length $n\}$, so $|R(n)| = 2^n$.
The individual flips in s are denoted $s_1, s_2, ..., s_n$.
$\boldsymbol{< r, t >}$ denotes the concatenation of strings r and t.

The **frequency** of s is **F**(s) = $p^i q^{n-i}$ where s has length n and contains i many H's.

For fair coins with $p = q = 1/2$, $F(s) = \frac{1}{2^n}$ for all $s \in R(n)$.

Given a sequence C of (at least n) random coin flips, a random contiguous string of length n in C has probability $F(s)$ of being a copy of s.

Let $s \in R(n)$ and let $k < n$. For any $t \in R(k)$ and $s \in R(n)$, we say that s **overlaps** (itself) **at** t if the initial and final segments of length k in s are both copies of t. We also will say s **overlaps by** k in this case, and that t is an **overlapping segment** of s.

Examples
s = HTHTH overlaps at HTH and at H (or, by 1 and 3)
s = HTTT has no overlapping segments.
s = HTH overlaps at H

For $s \in R(n)$, let $\boldsymbol{V}$(s) = $\{ t \mid t \in R(k)$ for some $k < n$ and s overlaps at $t \}$.

Examples

V(HTHTH) = {H, HTH} V(HHH) = {H, HH} V(HTTT) = ∅

If $t = s_1 ... s_k$ is an initial segment of s, then $\boldsymbol{-t+s}$ denotes $s_{k+1} ... s_n$, so $< t, -t + s > = s$.
If $t = s_{n-k+1} ... s_n$ is a final segment of s, , then $\boldsymbol{s-t}$ denotes $s_1 ... s_{n-k}$, so $< s - t, t > = s$.

Observe then that if $t \in V(s)$, then $-t+s$ and $s-t$ are both defined and
$F(-t+s) = F(s)/F(t) = F(s-t)$

Example. s = HTHHT, t = HT. Then $-t+s$ = HHT, $s-t$ = HTH, and

F(HHT) = $p^2 q$ = F(HTHHT) / F(HT) = F(HTH).

Good and bad occurrences of *s*.

Let C be a sequence of coin flips, and let $s \in R(n)$. An occurrence of *s* in C is **good** (in C) if it does not overlap with a preceding good occurrence of *s*, else it is **bad** (in C). This is an inductive definition: the first occurrence G of *s* is good, all occurrences which overlap with G are bad, the first occurrence which is disjoint from G is good, etc.

Example. Let *s* = HHH, and let C = THTHH**H**HTTHTHH**H**TTHH**H**HH**H**HTT...

There are 8 occurrences of *s*, four of which are good (end with **bold**) and four of which are bad (end with underline). The **type** of a bad occurrence is the overlapping segment *t*. In the example, the first, second and fourth bad occurrences of *s* are of type HH, and the third is of type H.

For any bad occurrence B of *s*, there is a unique pair (G, *t*) with G a preceding good occurrence of *s,* and B a bad occurrence of type *t* overlapping with G. Conversely, given a good occurrence G of *s* and some *t* in *V(s)*, G will overlap with a bad occurrence of *s* of type *t* precisely when G is followed by the $n - k$ flips $-t+s$.

Example *s* = HTHTH, *t* = HTH. C = ... HTHTHxy . . .
A bad occurrence of *s* of type *t* will happen IFF xy = (-*t*+*s*) = TH.

Definition. Let $s \in R(n)$ and let $t \in V(s)$. $a_t(s)$ denotes the expected number of bad occurrences of *s* of type *t* overlapping a good occurrence G of *s*. $\boldsymbol{a = a(s)}$ denotes the expected total number of bad occurrences of *s* (of all types) which overlap with a good occurrence G of *s*.

Lemma 1. $a_t(s) = F(s)/F(t)$.
proof. Given a good occurrence G of *s*, it will overlap with either 0 or 1 bad occurrences of type *t*, so $a_t(s)$ is just the probability that the next $n - k$ flips are precisely $(-t + s)$. Thus $a_t(s) = F(-t + s) = F(s) / F(t)$.

Lemma 2. $a(s) = F(s) \sum_{t \in V(s)} \frac{1}{F(t)}$.

proof.. By the discussion above, the number of bad occurrences overlapping G is just the sum of the number of bad occurrences of type *t* over all *t* in V(s). By linearity of expectation, we have $a(s) = \sum_{t \in V(s)} a_t(s) = \sum_{t \in V(s)} \frac{F(s)}{F(t)} = F(s) \sum_{t \in V(s)} \frac{1}{F(t)}$.

Examples. s = HHH, $V(s) = \{H, HH\}$ $a = p^3(\frac{1}{p} + \frac{1}{p^2}) = p^2 + p$ (=3/4 for fair coin)

s = HTT, $V(s) = \varnothing$, $a = pq^2(0) = 0$ (all occurrences of s are good)

s = HTHTH, V(s) = {H, HTH} $a = p^3q^2(\frac{1}{p} + \frac{1}{p^2q}) = p^2q^2 + pq$ (= 5/16 fair coin)

Expected waiting time for s

Theorem 1. Let *s* be a string of H's and T's. Then $E(s) = \frac{1}{F(s)} + \sum_{t \in V(s)} \frac{1}{F(t)}$.

proof. First, let *g=g(s)* denote the frequency (i.e. density) of good occurrences of *s* in a sequence C of coin flips, and let *b=b(s)* denote the frequency of bad occurrences of *s*. There are three simple observations which yield the proof.

i) $b = ag$ by the definition of *a* as the expected number of bad occurrences of *s* for each good one.

ii) $g + b = F(s)$, since every occurrence of *s* is either good or bad but not both.

iii) $E(s) = 1/g$, since repeatedly playing the "waiting for s" game is indistinguishable from marking good occurrences in a sequence C of flips. Thus the average length of the waiting game for *s* is simply the reciprocal of the density of good occurrences of *s* in C.

By (i) and (ii) we have $F(s) = g + ag$ and hence $g = F(s)/(1+a)$. Then by (iii) and Lemma 2, we get that $E(s) =$

$$\frac{1}{g} = \frac{1+a}{F(s)} = \frac{1}{F(s)} + \frac{a}{F(s)} = \frac{1}{F(s)} + \frac{1}{F(s)}(F(s) \sum_{t \in V(s)} \frac{1}{F(t)}) = \frac{1}{F(s)} + \sum_{t \in V(s)} \frac{1}{F(t)}.$$

Examples. $E(\text{HHH}) = \frac{1}{p^3} + (\frac{1}{p} + \frac{1}{p^2})$ [= 8 + 2 + 4 = 14 for fair coin]

$E(\text{HTT}) = \frac{1}{pq^2} + (0) = \frac{1}{pq^2} = \frac{1}{F(s)}$ since V(s) is empty [= 8 for fair coin]

$E(\text{HTHTH}) = \frac{1}{p^3q^2} + (\frac{1}{p} + \frac{1}{p^2q})$ [32 + 2 + 8 = 42 for fair coin]

$E(\text{HHH.....H}) = \frac{1}{p^n} + (\frac{1}{p} + \frac{1}{p^2} + ... \frac{1}{p^{n-1}}) = \frac{1 + p + ... + p^{n-1}}{p^n} = \frac{1 - p^n}{p^n q}$ [$= 2^{n+1} - 2$ for f.c.]

Corollary. For a fair coin and *s* of length *n*, let $W(s) = \{ k \mid k < n$ and *s* overlaps by *k* $\}$.

Then $E(s) = 2^n + \sum_{k \in W(s)} 2^k$.

Randomizing *s*

Determine the expected length of the following game. Fix *n*, and flip the coin *n* times, establishing your "goal" *s*. Now play the waiting game with *s*, your score being the number of flips (after establishing your goal *s*) until the first appearance of *s*. Since the relative frequency of *s* as the goal is *F(s),* we want to compute $\mathbf{Ex}(n) = \sum_{s \in R(n)} F(s)E(s)$.

Example: $n = 3$	*F(s)*	*E(s)*	*F(s)E(s)*	Total
HHT, THH	*ppq*	$1/(ppq)$	1	2
HTT, TTH	*pqq*	$1/(pqq)$	1	2
HHH	*ppp*	$1(ppp) + 1/p + 1/(pp)$	$1 + p + p^2$	$1 + p + p^2$
HTH	*ppq*	$1/(ppq) + 1/p$	$1+pq$	$1+pq$
THT	*pqq*	$1/(pqq) + 1/q$	$1+pq$	$1+pq$
TTT	*qqq*	$1/(qqq) + 1/q + 1/(qq)$	$1 + q + q^2$	$1 + q + q^2$

Then $Ex(3) = 8 + (p + q) + (p^2 + pq + pq + q^2) = 8 + (p + q) + (p + q)^2 = 8 + 1 + 1 = 10$.

Expanding the term *E(s)* by Theorem 1, we obtain the following formula for **Ex**(n) =

$$\sum_{s \in R(n)} F(s)E(s) = \sum_{s \in R(n)} F(s)\left(\frac{1}{F(s)} + \sum_{t \in V(s)} \frac{1}{F(t)}\right) = \sum_{s \in R(n)} \left(1 + \sum_{t \in V(s)} \frac{F(s)}{F(t)}\right)$$

$$= 2^n + \sum_{s \in R(n)} \sum_{t \in V(s)} F(-t + s)$$

The double sum is taken over all pairs (*s,t*) with *s* of length *n* and *t* in *V(s)*.

Define $\boldsymbol{U(n, k)} = \{(s, t) \mid s \in R(n),\ t \in V(s) \cap R(k)\}$. There are two cases, A and B:

Example A: $U(3,1) = \{$ (H**H***H*, H), (H**T***H*, H), (T**T***T*, T), (T**H***T*, T) $\}$.
Ex. B: $U(5,3) = \{$ **HH**HHH, HHH), (**HT**HTH, HTH), (**TH**THT, THT), (**TT**TTT, TTT) $\}$

Case A: $k \le n/2$. Then *t* is arbitrary since the initial and *final* copies of *t* in *s* are disjoint, and there is an arbitrary string **r** of length $n - 2k$ in between them. Thus in this case

$U(n,k) = \{ (< t, \mathbf{r}, t >, t) \mid t \in R(k),\ r \in R(n - 2k) \}$, and $-t+s = <\mathbf{r}, t>$.

Case B: $n/2 < k < n$. Then the copies of *t* inside *s* overlap with each other, and a simple inspection shows that that the first *n-k* entries in *s* are arbitrary, and then this initial sequence **u** must be repeated until *s* is filled (see example B above). So in case B we have

$U(n,k) = \{ (< \mathbf{u},\ t >, t) \mid\ u \in R(n - k),\ t = < u, u, \ldots, u, u^* > \}$ where
$u^* = u_1 \ldots u_m$ with $m = k \bmod (n - k) = n \bmod (n - k)$, and $s - t = \mathbf{u}$.

Lemma A. Let $k \le n/2$. Then $\sum_{(s,t) \in U(n,k)} F(-t+s) = 1$.

proof. $\sum_{(s,t) \in U(n,k)} F(-t+s) = \sum_{t \in R(k), r \in R(n-2k)} F(< r,t >) = \sum_{u \in R(n-k)} F(u) = 1$

since we are simply adding up the frequencies of all strings $u = < r, t >$ of length n-k.

Lemma B. Let $n/2 < k < n$. Then $\sum_{(s,t) \in U(n,k)} F(-t+s) = 1$

proof. $\sum_{(s,t) \in U(n,k)} F(-t+s) = \sum_{(s,t) \in U(n,k)} F(s-t) = \sum_{u \in R(n-k)} F(u) = 1$

again since we are adding all the frequencies of length (n-k). The first equality follows since $F(- t + s) = F(s - t))$.

So for each k with $1 \le k \le n-1$, we have $\sum_{(s,t) \in U(n,k)} F(-t+s) = 1$.

Summing over all such k we get $\sum_{k=1}^{n-1} \sum_{(s,t) \in U(n,k)} F(-t+s) = n - 1$, and since this double sum counts every pair (s,t) for $s \in R(n)$ and $t \in V(s)$ exactly once, we can re-index to get

$$\sum_{s \in R(n)} \sum_{t \in V(s)} F(-t+s) = n - 1.$$

Finally by the calculation at the start of this section we have

$$Ex(n) = \sum_{s \in R(n)} F(s)E(s) = \sum_{s \in R(n)} F(s)\left(\frac{1}{F(s)} + \sum_{t \in V(s)} \frac{1}{F(t)}\right) =$$

$$2^n + \sum_{s \in R(n)} \sum_{t \in V(s)} F(-t+s) = 2^n + n - 1$$

Note that $Ex(n)$ is independent of the weights p and q.

So we have shown

Theorem 2. For any coin with $0 < p < 1$, $Ex(n) = 2^n + n - 1$.

L- sided coins (e.g. finite sample spaces with independent trials). If we play the same games with an L-sided coin with positive probabilities $p_1,...,p_L$, and $F(s) = {p_1}^{i_1}...{p_L}^{i_L}$ when s contains i_j occurrences of outcome H_j, and overlaps are defined in the obvious way, then the same constructions work the same way to give

Theorem 1L: $E(s) = \frac{1}{F(s)} + \sum_{t \in V(s)} \frac{1}{F(t)}$

Theorem 2L: $Ex(n) = L^n + n - 1$, independent of $p_1,...,p_L$.