Tarot Sham Proof

Per

May 25, 2017

Note This was produced as an internal working document of the ICPC World Finals 2017 judges. I've made it public since several people expressed an interest in it. This proof conveys very little intuition, and if someone has a proof that is more combinatorial, more intiuitive, or just plain simpler, I would be really interested in seeing it – in that case feel free to contact me at `austrin@kth.se`.

1 Basic setup

Notation:

- Δ denotes the first difference operator, i.e., for any function $f : \mathbb{Z} \to \mathbb{R}$ we define Δf to be the function $\Delta f(n) = f(n) - f(n-1)$.
- For integer s, $[s]$ denotes $\{1, \ldots, s\}$.

Let the random variable X be a uniformly random infinite string over R,P,S. Let $P = P_1 P_2 \ldots P_\ell$ be a pattern of length ℓ, and write $M_P(i)$ for the event that P occurs at position $i \geq 1$ in X.

What we are interested in is to understand the quantity

$$\Pr\left[\bigcup_{i=1}^{n} M_P(i)\right]$$

and how it depends on P (this quantity is the probability that P appears in a random string of length $n + \ell - 1$).

The *overlap set* of P is the set

$$S = \{r \in [\ell - 1] : P_{\geq r+1} = P_{\leq \ell - r}\} \subseteq [\ell - 1]$$

We can compute this from the KMP array quickly. Another way of thinking about it is that it is the set of non-trivial r:s so that when we append the last r letters of P to P, the resulting string ends with P.

Definition 1. *For $S \subseteq [\ell - 1]$ define*

$$f_S(n) = \frac{n - f_S(\leq n - \ell)}{3^\ell} - \sum_{i \in S} f_S(n - i) 3^{-i}$$

with initial values $f_S(n) = 0$ for $n \leq 0$.

Here $f_S(\leq x)$ is short-hand for the prefix sum $\sum_{i=0}^{x} f_S(i)$

Lemma 2. *If S is the overlap set of P then*

$$f_S(n) = \Pr\left[\bigcup_{i=1}^{n} M_P(i)\right]$$

Proof. Let $g(n)$ be the probability that X has an occurence of P in position 1 and no other occurences in positions 2 to n. Our goal is to prove that $f_S(n) = \sum_{i=1}^{n} g(i)$ (because we can view $g(i)$ is the probability that the last occurence of P in positions $1 \ldots n$ happens at position $n - i$) or equivalently that $g(n) = \Delta f_S(n)$.

We have

$$g(n) = \frac{1}{3^\ell} - \frac{g(\leq n - \ell)}{3^\ell} - \sum_{i \in S} \frac{g(n-i)}{3^i} \tag{1}$$

The first term is the probability of P occurring at position 1 in X.

For $j \in \{1, \ldots n-1\}$, the probability of the event "P occurs at position 1, and at position $j+1$, and at no positions after $j+1$" equals

$$\begin{cases} \frac{g(n-j)}{3^j} & \text{if } j < \ell,\ j \in S \\ \frac{g(n-j)}{3^\ell} & \text{if } j \geq \ell \\ 0 & \text{otherwise} \end{cases}$$

These events are disjoint for different values of j and are precisely the events we need to subtract from $1/3^\ell$ to obtain $g(n)$.

A quick inspection then shows that (1) equals $\Delta f_S(n)$. □

The function f_S is well-defined and seemingly well-behaved for any set $S \subseteq [\ell - 1]$, but ultimately we will only prove our end result for overlap sets. It is likely true for all S but there are a few places where overlap sets allow us to take short-cuts that would require more work for arbitrary S.

2 Some simple observations

Lemma 3. *For any S, $f_S(n)$ is a (weakly) monotone function, i.e., $\Delta f_S(n) \geq 0$ for all n.*

This is trivial for overlap sets, but it is also easy to prove for any S by induction using the recurrence.

Lemma 4. *For every overlap set S and n it holds that*

$$\Delta f_S(n) \geq \frac{2}{3} \Delta f_S(n-1)$$

This Lemma is probably true for all sets, not just overlap sets, but it's easier to prove for overlap sets using the combinatorial interpretation:

Proof. Note that $\Delta f_S(n)$ equals the probability that the pattern occurs at position 1 but not at any of the positions $2, \ldots, n$. This probability does not go down by more than a factor $2/3$ when increasing the length, since there is a probability $2/3$ that last symbol added does not match the last symbol of P. □

Lemma 5. *If an overlap set S contains i, then it contains all multiples of i up to $\ell - 1$. In particular, if $S \neq [\ell - 1]$ then $1 \notin S$.*

This is obviously not true for arbitrary sets.

Proof. If adding the last i letters of P to itself yields a string ending in P, then adding the last i letters again obviously also yields a string ending in P, and so on. □

3 Main result

Theorem 6. *Let S and T be two overlap sets, such that*

$$\min(T \setminus S) = r \qquad\qquad \min(S \setminus T) > r$$

i.e., when writing the elements in sorted order, T is lexicographically smaller than S, and the first item where they differ is r.

Then for all n we have $f_S(n) \geq f_T(n)$ with equality if and only if $n \leq r$.

Proof. Let $D(n) = f_S(n) - f_T(n)$. By the recurrence definition of f, we see that

$$\begin{aligned} D(n) &= -\frac{D(\leq n-\ell)}{3^\ell} + \sum_{i \in T} \frac{f_T(n-i)}{3^i} - \sum_{i \in S} \frac{f_S(n-i)}{3^i} \\ &= \sum_{i \in T \setminus S} \frac{f_T(n-i)}{3^i} - \sum_{i \in S \setminus T} \frac{f_S(n-i)}{3^i} - \frac{D(\leq n-\ell)}{3^\ell} - \sum_{i \in S \cap T} \frac{D(n-i)}{3^i} \\ &= \sum_{i \in T \setminus S} \frac{f_T(n-i)}{3^i} - \sum_{i \in S \setminus T} \frac{f_T(n-i)}{3^i} - \frac{D(\leq n-\ell)}{3^\ell} - \sum_{i \in S} \frac{D(n-i)}{3^i} \end{aligned}$$

From this it is easy to see that $D(n) = 0$ for $n \leq r$, and that $D(r+1) = 3^{-r} f_S(1) = 3^{-r-\ell}$ (the precise value doesn't really matter, just that it is positive).

We will prove by induction that $D(n) \geq \frac{2}{3} D(n-1)$. Since $D(r+1) > 0$ this shows that $D(n) > 0$ for all $n \geq r+1$.

Assume that this claim holds up to $n-1$ (clearly, it holds for the base cases, i.e., up to $r+1$).

We now compute

$$\Delta D(n) = \sum_{i \in T \setminus S} \frac{\Delta f_T(n-i)}{3^i} - \sum_{i \in S \setminus T} \frac{\Delta f_T(n-i)}{3^i} - \frac{D(n-\ell)}{3^\ell} - \sum_{i \in S} \frac{\Delta D(n-i)}{3^i} \tag{2}$$

Note that since $r \in T \setminus S$ we have (using Lemma 3)

$$\sum_{i \in T \setminus S} \frac{\Delta f_T(n-i)}{3^i} \geq \frac{\Delta f_T(n-r)}{3^r} \tag{3}$$

Furthermore, we have by Lemma 4 that $\Delta f_T(n-i) \leq (3/2)^{i-r} \Delta f_T(n-r)$ for all $i > r$, and since all elements of $S \setminus T$ are greater than r we get

$$\begin{aligned} \sum_{i \in S \setminus T} \frac{\Delta f_T(n-i)}{3^i} &\leq \Delta f_T(n-r) \sum_{i \in S \setminus T} \frac{(3/2)^{i-r}}{3^i} \\ &\leq \Delta f_T(n-r) \sum_{i=r+1}^{\infty} \frac{2^{r-i}}{3^r} = \frac{\Delta f_T(n-r)}{3^r} \end{aligned} \tag{4}$$

Plugging (3) and (4) into (2) they cancel each other out and we are left with

$$\Delta D(n) \geq -\frac{D(n-\ell)}{3^\ell} - \sum_{i \in S} \frac{\Delta D(n-i)}{3^i} \tag{5}$$

By the induction hypothesis we have that $D(n-i-1) \geq 0$ implying $\Delta D(n-i) \leq D(n-i)$ and so we get

$$\Delta D(n) \geq -\frac{D(n-\ell)}{3^\ell} - \sum_{i \in S} \frac{D(n-i)}{3^i} = -\sum_{i \in S \cup \{\ell\}} \frac{D(n-i)}{3^i} \geq -\sum_{i=\min(S)}^{\infty} \frac{D(n-i)}{3^i}$$

By Lemma 5[1] and the fact that $r \notin S$, it follows that $1 \notin S$ hence $\min(S) \geq 2$. Furthermore, the induction hypothesis also implies $D(n-i) \leq D(n-1) \cdot (3/2)^{i-1}$ and we get

$$\begin{aligned} \Delta D(n) &> -\sum_{i=2}^{\infty} \frac{D(n-1) \cdot (3/2)^{i-1}}{3^i} \\ &= -\frac{2}{3} D(n-1) \sum_{i=2}^{\infty} \frac{1}{2^i} = -\frac{1}{3} D(n-1). \end{aligned}$$

Finally we get

$$D(n) = D(n-1) + \Delta D(n) \geq D(n-1) \cdot (1 - 1/3) = \frac{2}{3} D(n-1)$$

□

[1]Here we use that S is an overlap set. We really shouldn't have to, but this was the first argument I came up with.