1 (100 PTS.) Balanced or not.

Let $\Sigma = \{a, b\}$. Consider a string $s \in \Sigma^*$ of length n. The ***depth*** of a string s is $d(s) = \#_a(s) - \#_b(s)$, where $\#_c(s)$ is the number of times the character c appears in s. The maximum depth of a string s is $d_{\max}(s) = \max_{p \text{ is any prefix of } s} d(p)$.

A string $t \in \{a, b\}^*$ is ***weakly balanced*** if $d(t) = 0$. The string t is ***balanced*** if it is weakly balanced, and for any prefix substring p of t, we have that $\#_a(p) \geq \#_b(p)$.

In the following, you can assume that $\forall x, y \in \Sigma^*$, we have $d(xy) = d(x) + d(y)$.

1.A. (20 PTS.) Let $s = s_1 s_2 \ldots s_n$ be the given string. For any i, let $s_{\leq i}$ be the prefix of s formed by the first i characters of s, where $0 \leq i \leq n$. For any i, let $f(i) = d(s_{\leq i})$. Prove that if there are indices i and j, such that $i < j$ and $f(i) = f(j)$, then $s_{i+1} s_{i+2} \ldots s_j$ is a weakly balanced string.

Solution:

We have that

$$d(s_{i+1} s_{i+2} \ldots s_j) = d(s_{\leq j}) - d(s_{\leq i}) = f(j) - f(i) = 0.$$

establishing the claim.

1.B. (40 PTS.) Prove (but not by induction please) that if a string $s \in \Sigma^*$ is balanced, then either:

(i) $s = \epsilon$,

(ii) $s = xy$ where x and y are non-empty balanced strings, or

(iii) $s = axb$, where x is a balanced string.

Solution:

Proof: If $|s| = 0$, then s is an empty string, $d(s) = 0$, and $s = \epsilon$, as claimed.

So, let $s = s_1 s_2 \ldots s_n$ be the given a string of length $n > 0$ that is balanced. It must be that $s_1 = a$, as otherwise $s_1 = b$, and then $d(s_1) < 0$, which contradicts that is balanced. As such, $d(s_1) = 1$, and $s_1 = a$.

It must be that $n > 1$, as otherwise $d(s) = 1$, which is impossible.

If there is any proper prefix x of s, such that $d(x) = 0$ (*proper* means $x \neq \epsilon$ and $x \neq s$), then $s = xy$, for some suffix $y = y_1, \ldots, y_k$, where $k = |y| > 0$. We have that for any i, that

$$d(y_1 \ldots y_i) = 0 + d(y_1 \ldots y_i) = d(x) + d(y_1 \ldots y_i) = d(x|y_1 \ldots y_i) \geq 0,$$

since $x|y_1 \ldots y_i$ is a prefix of s, s is balanced, and $d(x) = 0$. This also implies that $d(y) = d(x) + d(y) = d(s) = 0$. Namely y is a balanced string. Since x is a prefix of s, and $d(x) = 0$ it follows that x is also balanced. This implies that $s = xy$ with x and y, as claimed.

Otherwise, it must be that for any proper prefix $s_1 \ldots, s_i$ of s, we have that $d(s_1 \ldots s_i) > 0$. This implies that $s_n = b$, as otherwise s can not be balanced. This also implies that

$d(s_2 \ldots s_i) = d(s_1 \ldots s_i) - d(s_1) = d(s_1 \ldots s_i) - 1 \geq 0$. Namely, for any prefix x of $s' = s_2 \ldots s_{n-1}$, we have $d(x) \geq 0$. Furthermore, we have $d(s_2 \ldots s_{n-1}) = d(s) - d(s_1) - d(s_n) = 0 + 1 - 1 = 0$. Namely, s' is a balanced string, and $s = as'b$. ■

1.C. (40 PTS.) Prove that for any string $s \in \{a, b\}^*$ of length n, that is balanced, at least one of the following must happen:

(i) The maximum depth of s is $\geq \sqrt{n}$, or

(ii) s can be broken into m non-empty substrings $s = t_1|t_2|\cdots|t_m$, such that $t_2, t_3, \ldots t_{m-1}$ are weakly balanced strings, and $m \geq \sqrt{n} - 1$.

For example, the string *abaababaabaabbbaaaabbbbb* can be broken into substrings

a|ba|ab|ab|aabaabbb|aaaabbbb|b

Hint: Let $f(i) = d(s_{\leq i})$. Analyze the sequence $f(0), f(1), \ldots, f(n)$, and what happens if the same value repeats in this sequence many times.

Solution:

Assume that (i) does not hold – namely, that $\max_i f(i) < \sqrt{n}$. The set $V = \{f(i) \mid i = 0, \ldots, n\} = \{0, 1, \ldots, \ell\}$, for some $\ell < \sqrt{n}$. There must be $i_1, i_2, \ldots, i_m$ indices, such that $f(i_1) = f(i_2) = \ldots = f(i_m)$, where $m \geq (n+1)/(\ell+1) \geq n/\sqrt{n} \geq \sqrt{n}$. Let last step used the pigeon hole principle – if there are $n+1$ elements in the sequence, and they can take on only $\ell + 1$ values, then there must be a value that repeats at least as much as the average.

Let $t_1 = s_1 \ldots s_{i_1}$. For $j = 2, \ldots, m$, we take $t_j = s_{i_{j-1}+1} \ldots s_{i_j}$. Finally, set $t_{m+1} = s_{i_m+1} \ldots, s_n$.

Clearly, $s = t_1|t_2|\cdots|t_{m+1}$. By part (A), we have that the strings $t_2, \ldots, t_m$ are weakly balanced, and the claim follows.

1.D. (Harder + not for submission.) Prove that for any string $s \in \Sigma^*$ of length n, that is balanced, with maximum depth $< \sqrt{n}/2$, it must be that s can be broken into $2m + 1$ substrings as follows $s = t_1 t_2 t_3 \ldots t_{2m+1}$, such that the m substrings $t_2, t_4, t_6, \ldots t_{2m}$ are non-empty and balanced. Here m has to be at least $\sqrt{n}/2 - 2$.

Solution: Not for submission for you, not for submission for us.

Covered by this exercise: Basic proof techniques, string manipulation, pigeon hole principle. Not using induction.

2 (100 PTS.) How the first mega tribe was created.

According to an old African myth, in the beginning there were only $n > 1$ persons in the world, and each person formed their own tribe. There were all living in the same forest. Every once in a while two tribes would meet. These meeting tribes would always fight each other to decide which tribe is better, and after a short war, invariably, the tribe with the fewer people (that always lost) would merge into the bigger tribe (if the two tribes were of equal size, one of the tribes would be the losing side). Every person in the tribe that just lost, had to sacrifice a lamb to the forest god, for reasons that remain mysterious, as the lambs did nothing wrong. In the end, only one tribe remained.

Prove, that during this process, at most $n \log_2 n$ lambs got sacrificed. (You can safely assume that no new people were born during this period.)

Solution:

Let $p_1, \ldots p_n$ be the n persons. Assume that a merge a single happens at time t, for $t = 1, \ldots, n-1$. Let $\mathsf{size}(p_i, t)$ be the size of the tribe that contains p_i after the merge at time t happened. Initially $\mathsf{size}(p_i, 0) = 1$, for all i.

The key observation is that if the tribe of p_i is merged into a *bigger* (or equal size) tribe at time t, then $\mathsf{size}(p_i, t) \geq 2\mathsf{size}(p_i, t-1)$ – and only then p_i needs to sacrifice a lamb. Since $\mathsf{size}(p_i, n-1) = n$, it follows that the doubling of the size of the tribe that contains a person p_i can happen at most $\lfloor \log_2 n \rfloor$. This implies that p_i personally sacrifices at most $\log_2 n$ lambs, and at most $n \log_2 n$ lambs get scarified overall.

Solution:

Alternative solution.

Lemma 1.1. *Let $S(n)$ be the maximum number of lambs that get sacrificed for n persons. We have that $S(n) \leq n \log_2 n$.*

Proof: **Base of induction.** If $n = 1$ then no lamb gets sacrificed, and $S(n) = 0 = n \log_2 n = 2$.

If $n = 2$ then a single lamb gets sacrificed, and $1 \leq S(n) \leq n \log_2 n = 2$.

Inductive hypothesis. For $k > 2$, and any $n \leq k$, we have that the number of lambs sacrificed is at most $n \log_2 n$.

Inductive step: Assume $n = k + 1$, and consider the last step where two tribes of sizes n_1, n_2 (i.e., $n_1 + n_2 = n$) got merged into the bigger whole tribe, and $1 \leq n_1 \leq n_2$ (here, we are considering the worse case scenario where the maximum overall number of lambs are being sacrificed). We have that

$$
\begin{aligned}
S(n) &= S(n_1) + S(n_2) + n_1 \leq n_1 + n_1 \log_2 n_1 + n_2 \log_2 n_2 \\
&\leq n_1(1 + \log_2 n_1) + n_2 \log_2 n_2 \leq n_1 \log_2 2n_1 + n_2 \log_2 n_2 \leq n_1 \log_2 n + n_2 \log_2 n \\
&= n \log_2 n,
\end{aligned}
$$

since $2n_1 \leq n_1 + n_2 = n$. ∎

Covered by this exercise: Proof by induction, and amortized analysis.

3 (100 PTS.) A few recurrences.

3.A. Consider the recurrence

$$T(n) = 2n + T(\lfloor n/4 \rfloor) + T(\lfloor (3/4)n \rfloor),$$

where $T(n) = 1$ if $n < 10$. Prove by induction that $T(n) = O(n \log n)$.

Solution:

First, a reminder of what O notations means.

definition: For two positive functions $f, g : \mathbb{N} \to \mathbb{N}$, we denote that $f(n) = O(g(n))$ if there exists constants n_0, c such that for all $n \geq n_0$ we have that $f(n) \leq cg(n)$.

Importantly, O notations holds only about values that are sufficiently large (i.e., bigger than n_0). In our case, we prove the claim for $n > 1$.

Claim 1.2. *For any integer $n \geq 2$, we have that $T(n) \leq 6n \log_2 n$.*

Proof: Assume that $T(n) \leq cn \log_2 n$, for some constant c to be determined shortly. The claim clearly holds for $1 < n < 10$ if $c \geq 1$. Assume the claim holds for all $n \leq k$, and assume that $n = k + 1$. In induction, we have that

$$\begin{aligned}
T(n) &= 2n + T(\lfloor n/4 \rfloor) + T(\lfloor (3/4)n \rfloor) \\
&\leq 2n + c\lfloor n/4 \rfloor \log_2 \lfloor n/4 \rfloor + c\lfloor (3/4)n \rfloor) \log_2 \lfloor (3/4)n \rfloor), \\
&\leq 2n + c\frac{n}{4} \log_2 \frac{n}{4} + c\frac{3n}{4} \log_2 \frac{3n}{4}, \\
&\leq 2n + c\left(\frac{n}{4} + \frac{3n}{4}\right) \log_2 \frac{3n}{4}, \\
&= 2n + cn(\log_2 n + \log_2 (3/4)) \\
&\leq 2n + cn(\log_2 n - 0.4) \\
&= 2n - 0.4cn + cn \log_2 n \leq cn \log_2 n,
\end{aligned}$$

which holds if $0.4c = (2/5)c > 2$. Namely, this holds for $c > 5$, and for concreteness, we pick $c = 6$.

A subtle issue in the above is that $n = k + 1 \geq 11$. As such, we have that the two recursive evaluations $T(\lfloor n/4 \rfloor)$ and $T(\lfloor (3/4)n \rfloor)$ are done on values $\lfloor n/4 \rfloor \geq \lfloor 11/4 \rfloor = 2$ $\lfloor (3/4)n \rfloor \geq \lfloor 33/4 \rfloor \geq 8$. For both cases we can use the induction hypothesis. ∎

3.B. Consider the recurrence

$$T(n) = \begin{cases} T(\lfloor n/2 \rfloor) + T(\lfloor n/6 \rfloor) + T(\lfloor n/7 \rfloor) + n & n \geq 24 \\ 1 & n < 24. \end{cases}$$

Prove by induction that $T(n) = O(n)$.

(An easier proof follows from using the techniques described in section 3 of these notes on recurrences.)

Solution:

Assume that $T(n) \leq cn$, for some constant c to be determined shortly. The claim trivially holds for $n < 24$. Assume that the claim holds for all $n \leq k$, and let $n = k+1$. By induction, we have

$$\begin{aligned}
T(n) &= T(\lfloor n/2 \rfloor) + T(\lfloor n/6 \rfloor) + T(\lfloor n/7 \rfloor) + n \\
&\leq T(\lfloor n/2 \rfloor) + T(\lfloor n/6 \rfloor) + T(\lfloor n/7 \rfloor) + n \\
&\leq c\lfloor n/2 \rfloor + c\lfloor n/6 \rfloor + c\lfloor n/7 \rfloor + n \\
&\leq cn/2 + cn/6 + cn/7 + n \\
&= n\Big(c(1/2 + 1/6 + 1/7) + 1\Big) \\
&= n\left(c\frac{21+7+6}{42} + 1\right) \\
&= n\left(c\frac{17}{21} + 1\right) \leq cn,
\end{aligned}$$

if $c\frac{17}{21} + 1 \leq c \iff 1 \leq (1 - 17/21)c \iff 21/4 \leq c$.

As such, for concreteness, we pick $c = 6$.

Covered by this exercise: Proof by induction, and solving recurrences.