Introduction

Finding repeated subsequences is a classical problem in computer science. Most variants of this problem rely on computing the longest common subsequence (LCS). Our interest is in the longest k-repeated subsequence problem (LkRS), which is the problem of returning the longest subsequence of a string w that is repeated k times. For any fixed $k > 0$, this can be naively implemented in $O(n^{2k-1})$ by running LCS on all possible k-partitions (see implementation [Git].) When $k = 2$, this problem is called longest tandem subsequence and can be solved in $O(n^k)$ due to Kosowski in 2004 [1] (see implemenation [Git]) However, for $k > 2$, there are no known faster algorithms.

Our contribution is an $O(n^2)$ algorithm that, for any $k > 0$, finds a k-repeated subsequence that is at least 2/3 the optimal length. The algorithm relies our analysis of the inherent combinatorial properties of the setting. The following is a proof of one such property.

Preliminaries

Consider a string w with length $n > 0$ such that p is the length of the longest common subsequence of any two disjoint substrings of w. We define the following notation:

- $L_w(i) : [1, n] \to [0, p]$:= the length of the common subsequence of the substrings created by splitting w at index i (the index after the i^{th} character). We refer to this as the "height" at index i Note: for all $i \in [1, n]$, $L_w(i) \in \mathbb{N}$ and $|L_w(i) - L_w(i + 1)| \leqslant 1$, i.e. adjacent indices have consecutive heights
- L_w:= the "landscape vector" corresponding to string w, i.e. an n-array whose i^{th} component is $L_w(i)$. Note: $\max(L_w) = p$.
- $c_w(i) : [0, p] \to [0, n]$:= the number of occurrences of height i in L_w.
- $s_w(i) = \sum_{j=0}^{i} c_w(j)$:= the number of indices in L_w with height less than or equal to i. Note that for any string w of length n and maximum height p, $s_w(p) = n$.
- $h_w = \sum_{i=0}^{p} i \cdot c_w(i)/n$:= the "average height" of w.

Claim 1: For all indices j such that $0 \leqslant j \leqslant p$, we have that $s_w(j) \leqslant 4j + 2$.

Suppose not. Then $s_w(j) \geqslant 4j + 3$, so there are $4j + 3$ indices x such that $L_w(x) \leqslant j$. Since each index must be separated by at least one distinct character, $n \geqslant 4j + 3$, of which $2j + 2$ must be the same character. Therefore, there is some index y, that that bisects these characters into two substrings each containing $j + 1$ of the same characters. So, $L_w(y) \geqslant j + 1$.

Claim 2: For all $0 < j < p$, $c_w(j) \geqslant 2$.

We have that $L_w(1) = 1$ and $L_w(n - 1) = 1$. Since $c_w(p) \geqslant 1$, there is some index i_p such that $L_w(i_p) = p$. In a discrete analogy of the intermediate value theorem, since adjacent components of L_w are consecutive numbers, for all heights $j \in [1, p - 1]$, there are indices $i_j^1 \in [1, i_p - 1]$ and $i_j^2 \in [i_p + 1, n - 1]$ such that $L_w(i_j^1) = j = L_w i_j^2$. So, $c_w(j) \geqslant 2$.

Claim 3: $\sum_{i=0}^{p} i \cdot c_w(i) = \left(np - \sum_{i=0}^{p-1} s_w(i) \right)$

We can represent $\sum_{i=1}^{p} i \cdot c_w(i)$ as an $p \times p$ matrix where each of the first i elements of row i is $c_w(i)$. Then, we can represent this matrix as the difference of the following two matrices:

$$\begin{bmatrix} 0 & \cdots & \cdots & 0 \\ c_w(1) & 0 & \cdots & 0 \\ c_w(2) & c_w(2) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ c_w(p-1) & c_w(p-1) & \cdots & c_w(p-1) \\ c_w(p) & c_w(p) & \cdots & c_w(p) \end{bmatrix} = \begin{bmatrix} c_w(0) & \cdots & c_w(0) \\ c_w(1) & \cdots & c_w(1) \\ c_w(2) & \cdots & c_w(2) \\ \vdots & \ddots & \vdots \\ c_w(p-1) & \cdots & c_w(p-1) \\ c_w(p) & \cdots & c_w(p) \end{bmatrix} - \begin{bmatrix} c_w(0) & \cdots & \cdots & c_w(0) \\ 0 & c_w(1) & \cdots & c_w(1) \\ 0 & 0 & \ddots & c_w(2) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_w(p-1) \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Consider the middle matrix, whose terms represent $\sum_{j=0}^{p} p \cdot c_w(j) = p \sum_{j=0}^{p} c_w(j) = p \cdot s_w(p) = np$.

For the rightmost matrix: its columns are precisely $0, s_w(1), \ldots, s_w(p-1)$, so the entries sum to $\sum_{j=0}^{p-1} s_w(j)$.

Therefore, $\sum_{i=0}^{p} i \cdot c_w(i) = np - \sum_{i=0}^{p-1} s_w(i)$

Our Construction

We define our construction as follows. Given an $n, p \in \mathbb{N}$ where $\lceil \frac{n-2}{4} \rceil \leqslant p \leqslant \lfloor \frac{n}{2} \rfloor$ define

$$\omega = \begin{cases} 0^{2p+1}1^{2p+1} & \text{if } n = 4p+2 \quad \text{“case 1”} \\ 0^{2p}1^{n-2p} & \text{otherwise} \quad \text{“case 2”} \end{cases}$$

Notice ω is the concatenation of an all-0 and all-1 substring, i.e. a “0 peak” and a “1 peak.”

Claim 4: $|\omega| = n$ and $\max(L_\omega) = p$

Notice $|\omega| = n$ is true by inspection.

In either case, notice that the longest common subsequence must be made up only of 0s or only of 1s. Therefore, in the first case $\max(L_\omega) = p = \lfloor (2p+1)/2 \rfloor = p$. In the second case, by the bounds of p, we know that $n \leqslant 4p+1$, so $n - 2p \leqslant 2p+1$. Thus, the height of the 0s peak is $\lfloor \frac{2p}{2} \rfloor = p = \lfloor \frac{2p}{2} \rfloor$. Further, $n - 2p < 2p$, so the height of the 0s peak is at most the height of the 1s peak. Therefore, $\max(L_\omega) = p = \lfloor \frac{2p+1}{2} \rfloor \geqslant$ height of the all 1 peak.

Claim 5a: $c_\omega(1), \ldots, c_\omega(\lfloor n/2 - p \rfloor - 1) = 4$.

In both cases, both the all-0 and all-1 peak reach height $\lfloor n/2 - p \rfloor - 1$. So, for all smaller heights i, L_w has height i once before and once after the point at height $\lfloor n/2 - p \rfloor$ in the all-0 substring. Likewise L_w has height i twice in the all-1 substring. These four indices are the four indices that are i indices away from points at height 0 (by convention, we say $L_w(0) = 0$, though we omit writing this in the landscape vector by convention).

Consider case 1, then $L_\omega(x) = i$ at $x \in \{i, 2p+1-i, 2p+1+i, n-i\}$.

Consider case 2, then $L_\omega(x) = i$ for $x \in \{i, 2p-i, 2p+i, n-i\}$.

Claim 5b: When $p < \frac{n}{2}$, $c_\omega(0) = 2$ and when $p = \frac{n}{2}$, $c_\omega(0) \leqslant 2$

In case 1, $L_\omega(2p+1) = L_\omega(n) = 0$. In case 2, $L_2(2p) = L_\omega(n) = 0$.

Consider the case where $p = n/2$. Then ω is a string of all 0s, therefore every two non-trivial substrings of ω have at least one character in common, so $c_\omega(0) = 1$. Otherwise, the 0-peak and 1-peak contain no common characters, therefore the length of longest common subsequence of their two substrings is 0.

Claim 6: $c_{\lfloor n/2-p \rfloor+1}, \ldots, c_{p-1} = 2$

In all cases the all 1 *substring reaches at most height* $\lfloor n/2 - p \rfloor$. *So, for all* $i \in \{\lfloor n/2 - p \rfloor + 1, \ldots, p-1\}$ *height* i *is only attained in the all* 0 *substring. Since for each* i, $i < p$, *we know that the all* 1 *substring has exactly two indices that result in substrings with longest common subsequences of length* i, *once at index* i *and once at index* $2p - i$. *So,* $c_\omega(i) = 2$.

Proof of Lowest Average Height

Let w be any string with length n and $\max(L_w) = p$ for analogously defined L_w. For each $i \in [1, n]$, also define $c_w(i), s_w(i), h(w)$ analogously.

Lemma 1. *For all* $i \in [0, p]$, $s_w(i) \leqslant s_\omega(i)$

First, consider $i \in \left[0, \lfloor \frac{n}{2} - p \rfloor - 1\right]$.

In case 1, $s_\omega(i) = 4i + 2$ by Claims 5a, 5b. Therefore, $s_w(i) \leqslant s_\omega(i)$ by Claim 1. This is also true in case 2, except when $p = n/2$, in which case, there is only one string of length n whose landscape reaches height p, so $w = \omega$, thus $s_w(i) = s_\omega(i)$.

Note that if $n = 4p + 2$, we're done since $\lfloor n/2 - p \rfloor - 1 = \lfloor p + 1 \rfloor - 1 = p$. So, henceforth we consider only case 2. Also note if $\lfloor n/2 - p \rfloor = p$, we're done since $s_\omega(p) = s_\omega(p) = n$, so we assume $\lfloor n/2 - p \rfloor < p$.

Now, consider $i \in [\lfloor n/2 - p \rfloor, p]$:

For ω: $s_\omega(i) = \ n - \sum_{j=i+1}^{p} c_\omega(i) = n - 2(p - i) - 1$ by Claim 6. For w: $\forall i < p$, $c_w(i) \geqslant 2$ and $c_w(p) \geqslant 1$ since L_w reaches height p. So, $s_w(i) = \ n - \sum_{j=i+1}^{p} c_w(i) \leqslant n - 2(p - i) - 1 = s_\omega(i)$. So, $s_w(i) \leqslant s_\omega(i)$.

Theorem 1. Given an n and p, the string with the lowest average landscape is of ω given by definition 1. In other words, $h_\omega \leqslant h_w$.

Proof.

$$h_\omega = \frac{1}{n} \sum_{i=1}^{p} i \cdot c_\omega(i) = \frac{1}{n} \left(np - \sum_{i=1}^{p-1} s_\omega(i) \right) \leqslant \frac{1}{n} \left(np - \sum_{i=0}^{p-1} s_w(i) \right) = \frac{1}{n} \sum_{i=0}^{p} i \cdot c_w(i) = h_w$$

Where the inequality holds by claim 3.

References

[1] Adrian Kosowski. "An efficient algorithm for the longest tandem scattered subsequence problem". In: *String Processing and Information Retrieval: 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004. Proceedings 11*. Springer. 2004, pp. 93–100.