Longest common subsequences in binary sequences

John D. Dixon
Carleton University, Ottawa, Canada

May 7, 2014

Given two \{0, 1\}-sequences \(X\) and \(Y\) of lengths, say, \(m\) and \(n\), respectively, we write \(L(X,Y)\) to denote the length of the longest common subsequence (LCS). For example, if \(X = 0110110\) and \(Y = 101001011\), then a longest common subsequence is given by 110111 (010111 is another one), so \(L(X,Y) = 6\). We write \(X_k\) (respectively, \(Y_k\)) to denote the initial segments of length \(k\) for these sequences and for fixed \(X\) and \(Y\) we shall write \(l(i,j)\) as an abbreviation for \(L(X_i,Y_j)\) when \(i = 0,\ldots,m\) and \(j = 0,\ldots,n\). It is easily seen that

\[
l(i,j) = \begin{cases} 
  l(i-1,j-1) + 1 & \text{if } X_i \text{ and } Y_j \text{ end in the same symbol} \\
  \max(l(i,j-1),l(i-1,j)) & \text{if } X_i \text{ and } Y_j \text{ end in different symbols}
\end{cases}
\]

for \(1 \leq i \leq m\) and \(1 \leq j \leq n\). Together with the boundary conditions \(l(i,0) = l(0,j) = 0\) for all \(i, j\), this gives an efficient way to compute \(L(X,Y)\) for any given \(X, Y\) and is generally the way in which \(L(X,Y)\) is computed.

1 Distribution of \(L(X,Y)\) for random sequences of same length

Now suppose that \(X\) and \(Y\) are random binary sequences of lengths \(m\) and \(n\), respectively, and consider the random variable \(L(X,Y)\). Let \(L(m,n)\) denote its mean value. Then the following are know in the case \(m = n\) (see [8, Chap. 1]).

1. The ratio \(\gamma_n := L(n,n)/n\) converges to a limit \(\gamma\) as \(n \to \infty\) (see [3]). The constant \(\gamma\) is known as the Chvátal-Sankoff constant, and determination of its value is a longstanding open problem. The best bounds which have been proved so far are 0.788071 \(\leq \gamma \leq 0.826280\) (see [5]).

2. Numerical evidence suggests that the sequence \(\{\gamma_n\}\) is monotonic increasing, but this has not been proved. However, it is clear that

\[L(nk,nk) \geq kL(n,n)\] for \(k = 1, 2, \ldots\)
and so $\gamma \geq \gamma_n$ for all $n$. As we shall see below, computations indicate that $\gamma_{16384}$ is approximately 0.81110 so it seems very likely that $\gamma > 0.81$.

3. The Azuma-Hoeffding inequality shows that
\[
\Pr \{|L(X,Y) - L(n,n)| > \lambda \sqrt{n}\} \leq 2e^{-\lambda^2/8} \quad \text{for each } \lambda > 0
\]
so the values of the random variable $L(X,Y)$ are concentrated around the mean $L(n,n)$.

2 Embedding $X$ in a random binary sequence $Y$

Let $X$ be a fixed binary sequence of length $m$ and consider the probability $p(X,n)$ that $X$ can be embedded into a random binary sequence of length $n$ for some fixed $n \geq m$, that is, that $L(X,Y) = m$. Since it is equally likely that $X$ and $Y$ end in the same or different symbols, (1) shows that:
\[
p(X,n) = \frac{1}{2}p(X_{m-1},n-1) + \frac{1}{2}p(X,n-1).
\]
Induction now shows that $p(X,n)$ is independent of the particular sequence $X$ and depends only on its length, so we can put $p(m,n)$ in place of $p(X,n)$. In particular, we may assume that $X$ is the sequence of all 1’s and so $p(m,n)$ is the probability that a random binary sequence of length $n$ has at least $m$ 1’s. Thus, for all $n \geq m$ we have
\[
p(m,n) = 2^{-n} \sum_{k=m}^{n} \binom{n}{k}.
\]
In particular, the Azuma-Hoeffding inequality shows that
\[
p(m,n) > 1 - e^{-\lambda^2/8} \quad \text{if } m < \frac{1}{2}n - \lambda \sqrt{n}
\]
and
\[
p(m,n) < e^{-\lambda^2/8} \quad \text{if } m > \frac{1}{2}n + \lambda \sqrt{n}.
\]

3 Distribution of $L(X,Y)$ for sequences of different lengths

To simplify notation we write $L(r,s)$ (for any two positive reals $r,s$) to mean $L(\lfloor r \rfloor, \lfloor s \rfloor)$ where $\lfloor \rfloor$ represents the floor function. Fix $\alpha > 0$ and consider the sequence $L(\alpha n,n)/n$ ($n = 1, 2, ...$). With an argument similar to the argument used to prove the existence of the limit for $L(n,n)/n$ (see Section 1) we can show that $L(\alpha n,n)/n$ converges as $n \to \infty$ for all $\alpha \geq 0$ and denote the limit by $\psi(\alpha)$. The function $\psi$ has the following properties.
1. Since 
\[ L(\alpha n, n)/n = \alpha L(n, \alpha n)/\alpha n = \alpha L(\alpha^{-1} r, r) / r \] 
where \( r = \alpha n \), we conclude that \( \psi(\alpha) = \alpha \psi(\alpha^{-1}) \) for \( \alpha > 0 \). Clearly \( \psi \) is increasing.

2. Since \( L(m_1, n_1) + L(m_2, n_2) \leq L(m_1 + m_2, n_1 + n_2) \) it follows that \( \lambda \psi(\alpha) + (1 - \lambda) \psi(\beta) \leq \psi(\lambda \alpha + (1 - \lambda) \beta) \) when \( \alpha, \beta, \lambda \) and \( 1 - \lambda \) are all nonnegative. Thus \( \psi \) is concave (see also [1]). Since \( \psi \) is bounded and concave in the open interval \((0, \infty)\) it has the following properties (see [4]): (i) \( \psi \) is continuous; (ii) \( \psi \) has a right-hand derivative and a left-hand derivative at each point with the right-hand derivative not less than the left-hand derivative; and (iii) these one-sided derivatives are monotonically decreasing.

3. It follows from the previous section that \( \psi(\alpha) = \alpha \) for \( 0 \leq \alpha < 1/2 \) and so by 1. we have \( \psi(\alpha) = 1 \) for \( \alpha > 2 \). It seems that \( \psi \) is at least twice differentiable except perhaps at \( \alpha = 1/2 \) and \( \alpha = 2 \) (see the graph shown below).

4. Let \( X \) and \( Y \) be two infinite random sequences. Then using the Azuma-Hoeffding inequality we have for each \( \varepsilon > 0 \):

\[
\Pr \left\{ \max_m \left| \frac{L(X_m, Y_n)}{n} - \psi\left(\frac{m}{n}\right) \right| > \varepsilon \right\} \to 0 \text{ as } n \to \infty.
\]

5. The function \( L^*(m, n) := \sqrt{(4mn - n^2 - m^2)/3} \) appears to be a close approximation to \( L(X_m, Y_n) \) for \( \frac{1}{2} n \leq m \leq 2n \) for “random” \( X \) and \( Y \). The graph \( m \mapsto L^*(m, n) \) is the arc of an ellipse tangential to the line \( y = \frac{1}{2} x \) at \( x = n \) and to \( y = n \) at \( x = 2n \). Its value at \( m = n \) is \( L^*(n, n) = n \sqrt{2/3} \) which is approximately 0.816496 and within all known bounds for \( \gamma n \) ([2] claims that \( \gamma = 0.812653 \) but I suspect that the latter estimate is unreliable). It seems possible that \( \psi(\alpha) \) is equal to \( \psi^*(\alpha) := \sqrt{(4x - x^2 - 1)/2} \) and \( \gamma = \sqrt{2/3} \). I conjecture that at any rate \( \psi^* \) is an upper bound to \( \psi \).

Computing \( L(X, Y) \) for random \( X, Y \) we obtained the graph in Figure 1.

4. From the discrete to the infinite

If \( X \) and \( Y \) are infinite random binary sequences, then [1] can be used to compute the values of \( l(m, n) (= L(X_m, Y_n)) \). Let \( l[m] \) denote the \( m \)th row of the infinite matrix \( [l(m, n)]_{k, n=0,1,...} \). For any infinite (real) vector \( v = (v(i))_{i=0,1,...} \) we define the \textbf{maximizer} \( \bigvee v \) such that \( \bigvee v \) is the vector whose \( j \)th entry is the maximum of the \( v(i) \) for \( i \leq j \).

Define \( T \) and \( \bar{T} \) as operators on vectors by

\[
(Tv)(i) := \begin{cases} 
    v(i - 1) + 1 & \text{if } y_i = 1 \\
    v(i) & \text{otherwise}
\end{cases}
\]

\[
(\bar{T}v)(i) := \begin{cases} 
    v(i - 1) - 1 & \text{if } y_i = 0 \\
    v(i) & \text{otherwise}
\end{cases}
\]
Figure 1: $L^*(m, 1000)/1000$ and $L(m, 1000)/1000$ with $0.5 < m/1000 < 2$
where $y_i$ is the $i$th entry of $Y$ and 

$$(	ilde{T}v)(i) := \begin{cases} v(i-1) + 1 & \text{if } y_i = 0 \\ v(i) & \text{otherwise} \end{cases}.$$ 

It follows from (1) that

$$l[m] = \sqrt{Tl[m-1]} \text{ if } x_m = 1$$

and

$$l[m] = \sqrt{\tilde{T}l[m-1]} \text{ if } x_m = 0.$$ 

This defines the successive rows of the LCS table using global operations. In particular, $l[m]$ is obtained from $l[0]$ by applying the nonlinear operators $\sqrt{T}$ and $\sqrt{\tilde{T}}$ in random order. In view of the relationship between the values of $l[m]$ and the values of $\psi$, this may give a hint as to the kinds of operators which leave $\psi$ fixed, and perhaps $\psi$ may be determined in this way.

Computations for this paper were done using the $J$-language developed by Iverson and Hui (see [3]). Although $J$ is an interpreted language it is fast because it is based on a large number of carefully integrated and optimized subroutines. The most efficient programs in $J$ turned out to be based on the global approach described above. $J$ is a very concise language and the full program to compute $L(X_m, Y) \ (m = 1, 2, \ldots)$ for two finite $\{0, 1\}$-lists $X$ and $Y$ is given as follows (lines beginning NB. are comments):

```j
LCS=: 3 : 0 "_ 0 _
NB. LCS Y; X returns a list of the lengths of the LCS
NB. for Y and all initial segments of X
' u v' =. y
val=. v*0
for_e. u do.
  val=. >./\(e=v)\} val,: }: 1, 1+val
end.
val
)
```

We estimated $L(n, n)$ by taking the mean value of 50 trials of LCS Y; X where $X$ and $Y$ were random $\{0, 1\}$-lists of length $n$ ($err$ is the standard deviation for the sample mean):

| $n$  | 64   | 128  | 256  | 512  | 1024 | 2048 | 4096 | 8192 | 16384 |
|------|------|------|------|------|------|------|------|------|-------|
| $L(n, n)$ | 0.77406 | 0.78266 | 0.79656 | 0.80121 | 0.80594 | 0.80711 | 0.80942 | 0.81031 | 0.81110 |
| $err$   | 0.00467 | 0.00286 | 0.00166 | 0.00108 | 0.00061 | 0.00052 | 0.00032 | 0.00021 | 0.00014 |
Table 1: State Transition Table. Rows are labelled by states and columns by pairs $(T_{ij}, Y_j)$.

|   | 00 | 01 | 10 | 11 |
|---|----|----|----|----|
| 0 | 0  | 0  | 1  | 1  |
| 1 | 1  | 0  | 1  | 1  |
| 2 | 0  | 2  | 1  | 3  |
| 3 | 1  | 2  | 1  | 3  |

Table 2: Output Table

|   | 00 | 01 | 10 | 11 |
|---|----|----|----|----|
| 0 | 0  | 0  | 0  | 0  |
| 1 | 0  | 1  | 1  | 1  |
| 2 | 0  | 0  | 0  | 0  |
| 3 | 0  | 1  | 1  | 1  |

5 Generating LCS table with a finite state machine

For any two $(0,1)$-sequences $X$ and $Y$ we consider the (possibly infinite) table whose $(i,j)$th entry is $L(X_i, Y_j)$ $(i,j = 0,1,\ldots)$. The initial row and column of this table consists of 0’s and we can define a table $T$ with $(0,1)$-entries by

$$ T_{i,j} := L(X_i, Y_j) - L(X_{i-1}, Y_j) $$

for $i = 1,2,\ldots$ and $j = 0,1,\ldots$.

Evidently knowledge of the entries of $T$ determine the values of $L(X_i, Y_j)$. We can compute the rows of $T$ recursively with a finite state machine as follows.

To compute values of $T_{ij}$ with given $j > 0$ and $i = 1,2,\ldots$ we use the triple $(T_{i-1,j-1}, T_{i,j-1}, f)$ where $f$ is a flag equal to 0 or 1 which defines the state of the machine. As input we have the pair $T_{i-1,j}$ and $Y_j$. The machine computes $T_{ij}$, moves into a new state defined by $(T_{i-1,j}, T_{i,j}, \tilde{f})$ and outputs the value of $T_{ij}$. The flag represents the carry which is necessary when the maximizer is applied to the row in computation of $L(X_i, Y_j)$ described in Section 4. In this form the finite state machine requires $2^3$ states, but some of these turn out to be indistinguishable so we can reduce to four states. We do not give the details but provide the final tables for a fsm (see Tables 1 and 2).

6 Partially ordered sets and longest chains

Another way to describe the same problem is as follows. Given two binary sequences $X$ and $Y$ of lengths $m$ and $n$, respectively, we define the set $P :=$
\{(i,j) \mid x_i = y_j \text{ with } 1 \leq i \leq m \text{ and } 1 \leq j \leq n\}. \text{ We partially order } P \text{ with } < \text{ where } (i,j) < (i',j') \iff [i < i' \text{ and } j < j']. \text{ It can be verified that } (i_1,j_1) < (i_2,j_2) < \ldots < (i_k,j_k) \text{ is a chain in } (P,<) \text{ if and only if } (x_{i_1},x_{i_2},\ldots,x_{i_k}) = (y_{j_1},y_{j_2},\ldots,y_{j_k}) \text{ is a common subsequence of } X \text{ and } Y. \text{ In particular, the longest chain in } (P,<) \text{ has length } L(X,Y).

7 Bibliography

References

[1] S. Amsalu, H. Matzinger, M. Vachkovskaia, “Thermodynamical approach to the longest common subsequence problem”, J. Stat. Phys. 131 (2008), 1103–1120.

[2] R. Bundschuh, “High precision simulations of the longest common subsequence”, European Phys. J. Ser. B 22 (2001) 533–541.

[3] Vaclav Chvátal and David Sankoff, “Longest common subsequences of two random sequences”, J. Appl. Prob. 12 (1975) 306–315.

[4] G.H. Hardy, J.E. Littlewood and G. Pólya, “Inequalities” (2nd. ed.), Cambridge Univ. Press, Cambridge, 1959.

[5] Ken Iverson and Roger Hui, J6.02, Jsoftware, Inc. (www.jsoftware.com)

[6] George S. Lueker, “Improved bounds on the average length of longest common subsequences”, J. ACM. 56 (2009) No. 3, Art. 17.

[7] Peter Rabinovitch, “Expected length of the longest common subsequence” (ms. Dec. 13, 2007).

[8] J.M. Steele, “Probability Theory and Combinatorial Optimization”, SIAM, Philadelphia, 1997.