Expected Number of Distinct Subsequences in Randomly Generated Strings

Yonah Biers-Ariel Anant Godbole Elizabeth Kelley

ETSU

Rutgers University

University of Minnesota

Cumberland Conference, Vanderbilt University, May 20, 2017

Finding the Expected Value of Random Quantities

- Finding the expected value of random quantities is often non-trivial!

Finding the Expected Value of Random Quantities

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;

Finding the Expected Value of Random Quantities

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence.

Finding the Expected Value of Random Quantities

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence.
- This could be of biological relevance in the case of two DNA strings.

Finding the Expected Value of Random Quantities

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence.
- This could be of biological relevance in the case of two DNA strings.
- Subadditivity arguments are easy to apply to prove that $L = \lim_{n\to\infty} \frac{E(L_n)}{n}$ exists.

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.
- The variance is of order n and in 2014, Houdré proved a CLT

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.
- The variance is of order n and in 2014, Houdré proved a CLT
- More was known earlier about the length of the longest increasing subsequence of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson.

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.
- The variance is of order n and in 2014, Houdré proved a CLT
- More was known earlier about the length of the longest increasing subsequence of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson.
- But even here, calculation of the expected value was non-trivial.

More on the LCS and LIS problems

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \le L \le 0.82$.
- The variance is of order n and in 2014, Houdré proved a CLT
- More was known earlier about the length of the longest increasing subsequence of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson.
- But even here, calculation of the expected value was non-trivial.
- The combined results of Vershik and Kerov; Logan and Shepp from the 1970's gave

$$\lim \frac{EL_n}{\sqrt{n}} = 2.$$

Tracy Widom Distribution

This was followed by concentration results—due to Bollobas and Janson; Kim; and Frieze among others—that revealed that the standard deviation of the size of the longest monotone subsequence (LMS) is of order $\Theta(n^{1/6})$, and culminated with the work of Baik, Deift and Johansson that exhibited the limiting law of a normalized version of the LMS. This is often cited as one of the crowning achievements of Probability/Analysis of the 20th Century. An *AMS Notices* article of Aldous and Diaconis gives a great summary.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- We consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- We consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- We consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas
- The string 10110 contains the subsequences 0, 1, 01, 10, 11, 00, 100, 101, 110, 111, 011, 010, 1011, 1010, 1110, 0110, and 10110.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- We consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas
- The string 10110 contains the subsequences 0, 1, 01, 10, 11, 00, 100, 101, 110, 111, 011, 010, 1011, 1010, 1110, 0110, and 10110.
- What is the average case behavior?

In our submitted paper, we proved

Theorem

Let $s_1, s_2, \ldots$ be a sequence of independent and identically distributed random variables with $Pr(s_1 = j) = \alpha_j, j = 1, 2, \ldots, d, \sum_j \alpha_j = 1$. Set $\alpha = (\alpha_1, \ldots, \alpha_d)$. Let $\phi(S_n)$ be the number of distinct subsequences in $S_n = (s_1, \ldots, s_n)$. Let $\psi(n) = E(\phi(S_n))$. Then there exists $c = c_{d,\alpha} \geq 1$ such that

$$\psi(n)^{1/n} \to c; n \to \infty,$$

where $c = 1$ iff $d \geq 1$ and $\max_j \alpha_j = 1$.

Discussion

- The above theorem is hardly surprising, but raises other questions, namely as to whether the "true" numbers contain, additionally, polynomial factors as do several Stanley-Wilf limits in the theory of pattern avoidance (note that there are no polynomial factors in our next result with $d = 2$) Also, in general the existence of limits is not automatic, as seen by the following example:

Discussion

- The above theorem is hardly surprising, but raises other questions, namely as to whether the "true" numbers contain, additionally, polynomial factors as do several Stanley-Wilf limits in the theory of pattern avoidance (note that there are no polynomial factors in our next result with $d = 2$) Also, in general the existence of limits is not automatic, as seen by the following example:
- Assume that n balls are independently thrown into an infinite array of boxes so that box j is hit with probability $1/2^j$ for $j = 1, 2, \ldots$. Let π_n be the probability that the largest occupied box has a single ball in it. Then, as proved by several people in the 1990's, $\lim_{n\to\infty} \pi_n$ does not exist, and $\limsup_{n\to\infty} \pi_n$ and $\liminf_{n\to\infty} \pi_n$ differ in the fourth decimal place! Such behavior does not however occur in our context, as the theorem states.

The case of $d = 2$

Theorem

Suppose $Pr[s_i = 1] = \alpha \in [0,1]$ *for all* $1 \leq i \leq n$*, and* $Pr[s_i = 0] = 1 - \alpha$*,* $\alpha \neq 0, 1$*. Then we have*

$$\phi(S_n) = \frac{A + B}{2\sqrt{\alpha(1-\alpha)}},$$

where

$$A = \left(1 - 2\sqrt{\alpha(1-\alpha)}\right)\left(1 - \left(1 - \sqrt{\alpha(1-\alpha)}\right)^n\right)$$

and

$$B = \left(1 + 2\sqrt{\alpha(1-\alpha)}\right)\left(\left(1 + \sqrt{\alpha(1-\alpha)}\right)^n - 1\right)$$

Result was Previously Known for $\alpha = 0.5$

It was shown in a 2004 EJC paper of Flaxman et al.that when $\Pr[s_i = 1] = .5$ then $E[\phi(S_n)] \sim k(\frac{3}{2})^n$ for a constant k. Later, Collins improved this result by finding that $E[\phi(S_n)] = 2(\frac{3}{2})^n - 1$. We generalized this in the previous theorem to non-uniform letter generation. Moreover, our method for finding this formula is very different from that used by Collins. We defined a new property of a string - the number of new distinct subsequences - and then use these numbers as the entries in a binary tree. Our formula is then given as a weighted sum of the entries in this tree. This procedure is a modification of a 2008 method due to Elzinga, Rahmann, and Wang.

Arbitrary d and Two-state Markov Chains

We are done with strings on a binary alphabet generated by a random process in which the probability that any given element was 1 was fixed at α. In the paper, we generalized this in two ways. First, we considered strings on the alphabet $\{1, 2, ..., d\} = [d]$ where each letter is independently j with probability α_j for all $j \in [d]$. After that, we returned to binary strings, but those generated according to a two-state Markov chain; in particular, if a letter follows a 1, then it is 1 with probability α, but if it follows a 0, then it is 1 with probability β. In both these cases, we found recurrences for the expected new weight contributed by the n^{th} letter, which led to explicit matrix equations for that expected new weight. Unfortunately, we have not yet been able able to find a closed-form formula for the total expected number of subsequences like we did for $d = 2$ (independent case). But we know that in the first of these two generalizations, the limit exists!

Open Problem 1

One of the central questions in the Permutation Patterns community is that of packing patterns and words in larger ensembles; see, e.g., a paper by Burstein et al. In a similar vein, we have the question of superpatterns, i.e., strings that contain all the patterns or words of a smaller size; see, e.g., the same paper. A distinguished question in this area is the one posed by Alon, who conjectured that a random permutation on $[n] = \left\lceil \frac{k^2}{4}(1 + o(1)) \right\rceil$ contains all the permutations of length k with probability asymptotic to 1 as $n \to \infty$. In the present context, a similar question might be: What is the largest k so that each element of $\{0, 1\}^k$ appears as a subsequence of a binary random string with high probability?

Also, the basic question studied in this paper appears to not have been considered in the context of permutations; i.e., one might ask: What is the expected number of patterns present in a random permutation on $[n]$?

Open Question 2

In the baseline case of binary equiprobable letter generation, we have that $E(\phi(S_n)) \sim 2(1.5)^n$, which implies that the *average* number of occurrence of a subsequence is $\frac{1}{2}2^n/(1.5)^n = \frac{1}{2}(4/3)^n$. Now a subsequence such as 1 occurs "just" around $n/2$ times, and the sequence $11\ldots1$ with $n/2$ ones occurs an average of $\binom{n}{n/2} \cdot \frac{1}{2^{n/2}}$ times, which simplifies, via Stirling's formula, to around $\sqrt{2}^n$, ignoring constants and polynomial factors. The same is true of any sequence of length $n/2$; it is, on average, over-represented. We might ask, however, what length sequences occur more-or-less an average number $(1.33)^n$ of times. We can parametrize by setting $k = xn$ and equating the expected number of occurrences of a k-long sequence to $(1.33)^n$. We seek, in other words, the solution to the equation

$$\binom{n}{xn}\frac{1}{2^{xn}} = (1.33)^n.$$

Ignoring non-exponential terms and employing Stirling's approximation, the above reduces to

$$2^x x^x (1-x)^{1-x} = 0.75,$$

which, via Wolfram Alpha, yields the solutions $x = .123\ldots$ and $x = .570\ldots$!