A Nearest Integer Solution to the Longest Run of Randomly Generated Words

Kai F. Kristensen

Abstract. How rare is the event of observing more than a certain number of consecutive and identical letters of any kind somewhere in a randomly generated word? No one can deny that the use of generating functions is crucial for giving answers to questions like this. This paper, however, gives an answer, essentially based on elementary linear algebra. The derived formula is nevertheless simpler, has computational advantages and gives rise to a 'nearest integer' representation with an improved analytical range, as compared to earlier results.

1. INTRODUCTION. Let a sequence of n letters ($n \geq 1$), chosen with replacement from a k letter alphabet ($k \geq 2$), denote a word w_n of length n. If w_n contains a subsequence of r identical and consecutive letters ($r \geq 1$), we say that w_n has a run of length r. Now, let c_n denote the number of words w_n where the length of any run (of any letter) is at most r. Then we may divide the c_n words into groups according to the longest run at the end of the word, called the longest last run. If the longest last run consists of exactly m letters ($1 \leq m \leq r$), there must be $(k-1)c_{n-m}$ words in this category because the first $n-m$ letters may be organized in c_{n-m} legal ways, while there are $k-1$ ways to choose the letter present in the longest last run. Thus we have the recursion relation

$$c_n = (k-1)c_{n-1} + \cdots + (k-1)c_{n-r}, \quad \text{for } n > r. \tag{1}$$

If $n \leq r$, it is clear that $c_n = k^n$, constituting r initial conditions. The characteristic equation related to (1) is

$$\lambda^r = (k-1)\lambda^{r-1} + \cdots + (k-1)\lambda + (k-1). \tag{2}$$

Since the right hand side of (2) is a geometric series we also have

$$\lambda^r = (k-1) \cdot \frac{\lambda^r - 1}{\lambda - 1}, \quad \text{where } \lambda \neq 1. \tag{3}$$

We observe that $\lambda = 1$ is not a solution to (2). Multiplying each side of (3) by $\lambda - 1$, we get the equation

$$\lambda^r(\lambda - k) + k - 1 = 0. \tag{4}$$

Then we know except in the case $\lambda = 1$ that (4) will have the same solutions $\lambda_1, \ldots, \lambda_r$ as (2). If we define $F(\lambda) = \lambda^r(\lambda - k) + k - 1$, the fundamental theorem of algebra ensures that $F(\lambda)$ may be represented by the factorization

$$F(\lambda) = (\lambda - 1)(\lambda - \lambda_1) \cdots (\lambda - \lambda_r). \tag{5}$$

http://dx.doi.org/10.4169/amer.math.monthly.119.07.566
MSC: Primary 05A05, Secondary 15A06

Locating the roots of $F(\lambda)$ and their multiplicity is crucial to construct the solution of (1). To see that all the $r+1$ roots of $F(\lambda)$ are simple, we will argue by contradiction. Assume that $F(\lambda)$ has a root $\lambda = \lambda_i$ of multiplicity $m_i \geq 2$. Then the fundamental theorem of algebra allows us to write $F(\lambda) = (\lambda - \lambda_i)^{m_i} G(\lambda)$, where $G(\lambda)$ is a polynomial of degree $r + 1 - m_i$, resulting in $F'(\lambda_i) = 0$. On the other hand, we will also make use of the fact that $F'(\lambda) = \lambda^{r-1}((r+1)\lambda - kr)$ is a polynomial of degree r with r roots, counted with multiplicity. Since $\lambda = 0$ has multiplicity $r-1$, then $\lambda = kr/(r+1)$ must be simple. These constitute all the r roots of $F'(\lambda)$, and none of them are solutions to (2). Then, by contradiction, we may conclude that all the $r+1$ roots of $F(\lambda)$ must be simple.

In [**2**], K.A. Suman proves that exactly one of the r (for $r \geq 2$) simple solutions, λ_r, of (2) lies outside the unit circle $C = \{z : |z| = 1\}$. By inspection, we see that $k - 1 < \lambda_r < k$. The other $r-1$ solutions $\lambda_1, \ldots, \lambda_{r-1}$ must lie in the open unit disc $U = \{z : |z| < 1\}$. If there were solutions on C, then the requirement $|\lambda^r(\lambda - k)| = |\lambda - k| = k - 1$ would only be satisfied if $\lambda = 1$.

In the special case $r = 1$, we have $\lambda = k - 1$, supplying the general solution $c_n = B(k-1)^n$. The initial condition $c_1 = k$ implies $B = k/(k-1)$, so that $c_n = k(k-1)^{n-1}$, for $n \geq 1$. This result is intuitive since the first letter can be chosen in k ways, but every consecutive letter can differ from the preceding one in $k-1$ ways.

2. SOLVING THE RECURSION RELATION. Since we have r distinct and simple solutions to (2), the general solution to (1) is given by

$$c_n = a_1\lambda_1^n + a_2\lambda_2^n + \cdots + a_r\lambda_r^n.$$

Using the initial conditions $c_n = k^n$ when $1 \leq n \leq r$, we get the system

$$\begin{array}{ccccccc} a_1\lambda_1 & + & \cdots & + & a_r\lambda_r & = & k \\ a_1\lambda_1^2 & + & \cdots & + & a_r\lambda_r^2 & = & k^2 \\ \vdots & & & & \vdots & & \vdots \\ a_1\lambda_1^r & + & \cdots & + & a_r\lambda_r^r & = & k^r. \end{array} \tag{6}$$

The coefficient matrix A of (6), given by

$$A = \begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_r \\ \lambda_1^2 & \lambda_2^2 & \ldots & \lambda_r^2 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^r & \lambda_2^r & \ldots & \lambda_r^r \end{bmatrix},$$

is closely related to a Vandermonde matrix, which is a kind of matrix that rapidly appears in the study of polynomial interpolations. In [**3**], H.S. Wilf treats a similar but simpler case and is able to obtain a simple solution by inverting the corresponding Vandermonde matrix. In [**2**] Suman makes use of generating functions and partial fraction expansions to solve (6), but his solution doesn't exhibit the same simplicity. In the following, we will see what a solution based on straightforward Gauss elimination can do.

Let T be the augmented matrix of the system (6). If $\{R_i\}_{i=1}^r$ denote the rows of T, we perform operations, replacing R_i with $R_i - \lambda_1 R_{i-1}$ (for $i = 2, \ldots, r$), to get

$$T \sim \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_r & k \\ 0 & \lambda_2(\lambda_2-\lambda_1) & \dots & \lambda_r(\lambda_r-\lambda_1) & k(k-\lambda_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \lambda_2^{r-1}(\lambda_2-\lambda_1) & \dots & \lambda_r^{r-1}(\lambda_r-\lambda_1) & k^{r-1}(k-\lambda_1) \end{bmatrix}.$$

Proceeding in the same manner, replacing R_j with $\lambda_2 \cdot R_{j-1}$ $(j = 3, \dots, r)$ and so on, we get

$$T \sim \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_r & k \\ 0 & \lambda_2(\lambda_2-\lambda_1) & \dots & \lambda_r(\lambda_r-\lambda_1) & k(k-\lambda_1) \\ \vdots & \ddots & & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_r \prod_{j=1}^{r-1}(\lambda_r-\lambda_j) & k\prod_{j=1}^{r-1}(k-\lambda_j) \end{bmatrix},$$

so that

$$a_r = \frac{k\prod_{j=1}^{r-1}(k-\lambda_j)}{\lambda_r\prod_{j=1}^{r-1}(\lambda_r-\lambda_j)}.$$

Now it will not be necessary to find $a_1, \dots, a_{r-1}$ by back-substitution. We will instead use symmetry, arguing that the Gaussian elimination process is valid even if (a_r, λ_r) is interchanged with (a_i, λ_i), (for $i = 1, \dots, r-1$). Hence we have

$$a_i = \frac{k \cdot \prod_{1\le j\le r,\ j\ne i}(k-\lambda_j)}{\lambda_i \cdot \prod_{1\le j\le r,\ j\ne i}(\lambda_i-\lambda_j)} \quad (\text{for } i = 1, \dots, r), \tag{7}$$

resulting in

$$c_n = \sum_{i=1}^{r} \frac{k \cdot \prod_{1\le j\le r,\ j\ne i}(k-\lambda_j)}{\lambda_i \cdot \prod_{1\le j\le r,\ j\ne i}(\lambda_i-\lambda_j)} \cdot \lambda_i^n. \tag{8}$$

3. SIMPLIFICATIONS. The fact that the fractions of the solution (8) have much in common with the factorization of $F(\lambda)$ makes it possible to accomplish considerable simplifications.

Theorem 1. *Let c_n be the number of n letter words generated from an alphabet of $k \ge 2$ letters with at most $r \ge 1$ letters in any run. For each $n \ge 1$ we have*

$$c_n = \sum_{i=1}^{r} \frac{k(\lambda_i - 1)}{(k-1)((r+1)\lambda_i - kr)} \cdot \lambda_i^n,$$

where $\{\lambda_i\}_{i=1}^r$ are the solutions to (2).

Proof. The numerator in the solution formula of a_i, given in (7), can be written

$$k \cdot \prod_{1\le j\le r,\ j\ne i} (k-\lambda_j) = \frac{kF(k)}{(k-\lambda_i)(k-1)} = \frac{k}{k-\lambda_i}, \tag{9}$$

because $F(k) = k-1$. The denominator of a_i can be written

$$\lambda_i \cdot \prod_{1\le j\le r,\ j\ne i} (\lambda_i-\lambda_j) = \lim_{\lambda\to\lambda_i} \lambda_i \cdot \frac{F(\lambda)}{(\lambda-1)(\lambda-\lambda_i)}.$$

Applying L'Hôpital's rule, we obtain

$$\lambda_i \cdot \prod_{1\le j\le r,\ j\ne i} (\lambda_i-\lambda_j) = \frac{\lambda_i^r((r+1)\lambda_i - kr)}{\lambda_i - 1}. \tag{10}$$

Sutbstituting (9) and (10) into (7), generates

$$a_i = \frac{k}{k-\lambda_i} \cdot \frac{\lambda_i - 1}{\lambda_i^r((r+1)\lambda_i - kr)} = \frac{k(\lambda_i-1)}{\lambda_i^r(k-\lambda_i)((r+1)\lambda_i - kr)}.$$

Since $\lambda_i^r(k-\lambda_i) = k-1$ because of (4), we get

$$a_i = \frac{k(\lambda_i-1)}{(k-1)((r+1)\lambda_i - kr)} \quad (\text{for } i = 1,\dots,r)$$

and finally

$$c_n = \sum_{i=1}^{r} \frac{k(\lambda_i-1)}{(k-1)((r+1)\lambda_i - kr)} \cdot \lambda_i^n. \tag{11}$$

Substituting $r = 1$ and $\lambda_r = k-1$ into (11), the correct solution $c_n = k(k-1)^{n-1}$ is revealed. ∎

In [**2**] Suman derived the solution,

$$\tilde{c}_n = \sum_{i=1}^{r} \tilde{a}_i \lambda_i^n, \tag{12}$$

where

$$\tilde{a}_i = \frac{k(\lambda_i^{r+2} - 2\lambda_i^{r+1} + \lambda_i^r)}{(k-1)^2(\lambda_i^{r+1} - (r+1)\lambda_i + r)}.$$

Proving that $\tilde{c}_n = c_n$ may seem a bit complicated at first look, but (2) and (4) are adequate tools for reducing complexity. If we use the version $k-1 = \lambda_i^r(k-\lambda_i)$ of (4) in the denominator of $\tilde{a}_i$, we get

$$\tilde{a}_i = \frac{k\lambda_i^r(\lambda_i-1)^2}{(k-1)^2(\lambda_i^{r+1} - (r+1)\lambda_i + r)} = \frac{k(\lambda_i-1)^2}{(k-1)(k-\lambda_i)(\lambda_i^{r+1} - (r+1)\lambda_i + r)}.$$

Further on we find

$$\lambda_i^{r+1} - (r+1)\lambda_i + r = (\lambda_i - 1)(\lambda_i^r + \cdots + \lambda_i - r),$$

leading to

$$\tilde{a}_i = \frac{k(\lambda_i - 1)}{(k-1)(k-\lambda_i)(\lambda_i^r + \cdots + \lambda_i - r)}.$$

Since $\lambda_i^r + \lambda_i^{r-1} + \cdots + \lambda_i = \lambda_i^r + \lambda_i^r/(k-1) - 1$, because of (2) we obtain

$$\tilde{a}_i = \frac{k(\lambda_i - 1)}{(k-1)(k-\lambda_i)(\lambda_i^r + \lambda_i^r/(k-1) - 1 - r)}.$$

Again, helped by (4), we find

$$\tilde{a}_i = \frac{k(\lambda_i - 1)}{(k-1)(k-(1+r)(k-\lambda_i))} = a_i,$$

proving that $\tilde{c}_n = c_n$.

We observe that (11) is a considerably simpler solution than (12) because of the lower degree polynomials involved. Simplicity, having its own value in mathematics, also quite often will grant both numerical and analytical advantages.

In this case, we would consider it numerically unfavourable to apply a solution formula where fractions may contain unnecessary large numbers. If for example $k = 10$ and $r = 100$, the denominator of $\tilde{a}_r$ is approximately $8.1 \cdot 10^{102}$, a huge number compared to the corresponding denominator value of a_r which is about 90. If we are less modest when we choose the values of r and k, we can imagine how numbers will increase according to the expansion factor $\lambda_r^r(\lambda_r - 1)$.

The analytical advantages of (11) also become clear because, unlike the solution (12), (11) makes it easy to deduce a nearest integer formula of c_n which is valid for all $n \geq 1$ when $k \geq 5$.

4. A 'NEAREST INTEGER' FORMULA. We have defined λ_r to be the dominating solution of (2), satisfying $k - 1 < \lambda_r < k$ when $r \geq 2$. Then we can rewrite (11) to get

$$c_n = \sum_{i=1}^{r-1} \frac{k(\lambda_i - 1)}{(k-1)((r+1)\lambda_i - kr)} \cdot \lambda_i^n + \frac{k(\lambda_r - 1)}{(k-1)((r+1)\lambda_r - kr)} \cdot \lambda_r^n.$$

Since $\lim_{n\to\infty} \lambda_i^n = 0$ $(i = 1, \ldots, r-1)$, there must exist an integer N, such that

$$\left| \sum_{i=1}^{r-1} a_i \cdot \lambda_i^n \right| < \frac{1}{2}, \tag{13}$$

when $n \geq N$. Numerical computations done in [**2**] indicate that (13) should also hold when $n \geq 1$, but the author did not succeed in giving an analytical proof on the basis of (12). With the simpler solution (11) at hand, we will prove (13) to be true when $k \geq 5$, $r \geq 2$ and $n \geq 1$.

If we let the notation $[x]$ mean the 'nearest integer' to x, with the convention that $[m - 1/2] = m$ whenever m is a positive integer, we will be able to give a nice 'nearest integer' formula of c_n.

Theorem 2. *Let c_n be the number of n letter words generated from an alphabet of $k \geq 5$ letters with at most $r \geq 1$ letters in any run. For each $n \geq 1$ we have*

$$c_n = \left[\frac{k(\lambda_r - 1)}{(k-1)((r+1)\lambda_r - kr)} \cdot \lambda_r^n \right],$$

where λ_r is the unique solution to (2) *satisfying $k - 1 \leq \lambda_r < k$.*

Proof. First consider

$$M(\lambda) = \frac{\lambda - 1}{\lambda - K},$$

where $\lambda \in U$ and $K > 1$ is real. Then M is a linear fractional (Möbius) transformation (see for example [**1**, pp. 279–281]), transforming U onto an open circular disc. The fact that $M(1) = 0$, $M(0) = 1/K$ and $M(-1) = 2/(K+1)$ together with $M(\bar{\lambda}) = \overline{M(\lambda)}$ allow us to conclude that the image of U under M is the the open disc centered at $1/(K+1)$ with a radius of $1/(K+1)$. We have thus obtained

$$\left| \frac{\lambda - 1}{\lambda - K} \right| < \frac{2}{K+1},$$

when $\lambda \in U$. Now, with $K = kr/(r+1)$, $(r \geq 2)$, we get

$$\left| \sum_{i=1}^{r-1} \frac{k(\lambda_i - 1)}{(k-1)((r+1)\lambda_i - kr)} \cdot \lambda_i^n \right| \leq \sum_{i=1}^{r-1} \frac{k}{(k-1)(r+1)} \cdot \frac{2}{1 + k \cdot \frac{r}{r+1}}.$$

After a short calculation it becomes clear that the inequality

$$\frac{k}{(k-1)(r+1)} \cdot \frac{2}{1 + k \cdot \frac{r}{r+1}} < \frac{1}{2(r-1)}$$

is satisfied when

$$k > 2 - \frac{5}{2r} + \sqrt{\left(2 - \frac{5}{2r}\right)^2 + 1 + \frac{1}{r}}.$$

As both $2 - 5/(2r)$ and $\sqrt{(2 - 5/(2r))^2 + 1 + 1/r}$ turn out to be increasing functions of r, we must require (when $r \to \infty$)

$$k > 2 + \sqrt{5}.$$

This proves that

$$\left| \sum_{i=1}^{r-1} \frac{k(\lambda_i - 1)}{(k-1)((r+1)\lambda_i - kr)} \cdot \lambda_i^n \right| < \frac{1}{2}$$

when $n \geq 1$, $r \geq 2$ and $k \geq 5$. By including the case $r = 1$, the proof of the assertion is complete. ∎

5. TOSSING DICE. Let X_n denote the length of the longest run in a randomly generated word of length n. Every letter in the alphabet of k letters has the same probability $1/k$ of occurring. The probability $P(X_n \geq r+1)$, that the longest run will have a length of at least $r+1$, is then given by

$$P(X_n \geq r+1) = 1 - c_n/k^n.$$

When $k \geq 5$, we have proved

$$P(X_n \geq r+1) = 1 - \frac{1}{k^n}\left[\frac{k(\lambda_r - 1)}{(k-1)((r+1)\lambda_r - kr)} \cdot \lambda_r^n\right]. \tag{14}$$

If we toss a die one hundred times, what is the probability of getting a longest run of at least length five? Applying formula (14), with $r = 4$, $n = 100$, $k = 6$ and $\lambda_r = 5.9961320107$, we get $P(X_n \geq r+1) \approx 0.06$. Table 1 supplies the probabilities of a number of other combinations.

Table 1. The table values are the probabilities $P(X_n \geq r+1)$.

$r \setminus n$	10	30	10^2	10^4	10^6	10^8
2	0.1813	0.4996	0.9107	1.0000	1.0000	1.0000
4	0.0040	0.0167	0.0601	0.9984	1.0000	1.0000
10	0	$2.8 \cdot 10^{-7}$	$1.2 \cdot 10^{-6}$	$1.4 \cdot 10^{-4}$	0.0137	0.7480
20	0	$2.3 \cdot 10^{-15}$	$1.8 \cdot 10^{-14}$	$2.3 \cdot 10^{-12}$	$2.3 \cdot 10^{-10}$	$2.3 \cdot 10^{-8}$

Looking for a nearly fair game, Table 1 gives the probability $P(X_{30} \geq 3) = 0.4996$, so in the long run it will not be a profitable to make a bet that the longest run, tossing a die 30 times, will be at least of length three. Maple simulated 30 tosses ten million times and gave 4997439 occurrences. The standard deviation of a binomial variable with success probability $p = 1/2$ and $n = 10^7$ trials is $\sqrt{np(1-p)} \approx 1580$. The expected number of occurrences with $p = P(X_{30} \geq 3)$ is 4996127. The difference between the expected and the observed values is therefore 1312, which is less than one standard deviation.

REFERENCES

1. W. Rudin, *Real and Complex Analysis*, third edition. McGraw-Hill Series in higher Mathematics, New York, 1986.
2. K. A. Suman, The longest run of any letter in a randomly generated word, in *Runs and Patterns in Probability: Selected Papers*, Kluwer Academic Publishers, The Netherlands, 1994. 119–130.
3. H. S. Wilf, The editor's corner: strings, substrings, and the 'nearest integer' function, *Amer. Math. Monthly* **94** (1987) 855–860; available at `http://dx.doi.org/10.2307/2322817`.

KAI F. KRISTENSEN received his cand.scient. (MSc) degree in mathematics from the University of Oslo in 1985. Most of his professional career he has been teaching mathematics and statistics to students of engineering and technology, but he also has experience with educating teachers and candidates of business administration. He likes to develop and show new aspects of ideas originating from teaching situations, everyday life and games. His means of doing so are often elementary stochastic processes and combinatorics.
Telemark University College, Faculty of Technology, N-3901 Porsgrunn, Norway
kai.f.kristensen@hit.no