Finding exact constants in a Markov model of Zipf's law generation

V.V. Bochkarev, E.Yu. Lerner, A.A. Nikiforov, A.A. Pismenskiy
Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia
E-mail: vbochkarev@mail.ru, eduard.lerner@gmail.com, antnikiphorov@yandex.ru, aa.pismenskiy@gmail.com

Abstract. According to the classical Zipf law, the word frequency is a power function of the word rank with an exponent $-1$. The objective of this work is to find multiplicative constant in a Markov model of word generation. Previously, the case of independent letters was mathematically strictly investigated in [Bochkarev V V and Lerner E Yu 2017 International Journal of Mathematics and Mathematical Sciences Article ID 914374]. Unfortunately, the methods used in this paper cannot be generalized in case of Markov chains. The search of the correct formulation of the Markov generalization of this results was performed using experiments with different ergodic matrices of transition probability $P$. Combinatory technique allowed taking into account all the words with probability of more than $e^{-300}$ in case of 2 by 2 matrices.

It was experimentally proved that the required constant in the limit is equal to the value reciprocal to conditional entropy of matrix row with weights presenting the elements of the vector $\pi$ of the stationary distribution of the Markov chain.

1. Introduction. Formulation of the main result.

The Zipf law [1] determines dependence between frequency $f$ of usage of words and rank $r$ of these words (the number in the frequency list sorted by lack of growth). According to the classical Zipf law (empirically established by him for the most frequently used words)

$$f(r) \approx cr^{-1}.$$  

This law was justified in [6, 3] for random texts with independent letters (the so-called monkey model) with arbitrary letter probabilities. Let $p_1, p_2, \ldots, p_n, p_0$ be the probabilities of pressing letters of some alphabet $L = \{A_1, A_2, \ldots, A_n\}$ and a space (the symbol of the end of the word) on the keyboard, respectively, $\sum_{i=0}^{n} p_i = 1$. In this article, we will consider the case when the probability $p_0$ of pressing a space is fixed and sufficiently small. The classical Zipf law arises in the limiting case $p_0 \to 0$. First, let’s formulate the known results for this case.

The word $w$ is any finite sequence of letters:

$$w = (A_{i_1}, \ldots, A_{i_N}).$$  

(1)

The word weight (in the case of a model with independent letters) is determined by the formula:

$$P(w) = p_{i_1} \times p_{i_2} \times \ldots \times p_{i_N}.$$  

(2)

Let’s consider all kinds of words ordered by lack of growth of weights and matched with a rank $r$ (the position of a word in the sorted list). Let us denote the weight of a word of rank $r$ by $p(r)$. The statement was proved in [6, 3]:
Theorem 1 For any positive vector $\mathbf{p} = (p_1, \ldots, p_n)$, $n \geq 2$ satisfying the condition $\sum_{i=1}^{n} p_i = 1$, for some positive $c_1, c_2$, the following inequality occurs:

$$c_1 r^{-1} < p(r) < c_2 r^{-1}$$

(3)

It should be noted that when determining (2), the sum of the weights of all possible words is equal to infinity, and the sum of the weights of words of the same fixed length is equal to one. The infinite sum problem of the weights of all words was originally contained in the formulation of the Zipf law with $-1$ exponent of the power law, because of the divergence of the harmonic series. The papers [6, 3] that actually consider a model with symbols of letters and space does not have such problem. It leads to the same assertion with the exponent less than $-1$. To specify the probability of a word with a space, it is necessary to use not the formula (1), but a formula in which the weight $\mathbb{P}(w)$ is multiplied by the constant $p_0$. The basic inequality proved in the works [6, 3] has the following form:

$$c_1 r^{-\alpha} < p_0 p(r) < c_2 r^{-\alpha},$$

where $\alpha = 1/\gamma$, and $\gamma$ is the root of the equation $\sum_{i=1}^{n} p_i^\gamma = 1$ (in the case of $p_0 > 0$). Inequality (3) is an equivalent mathematical result. Also, it should be noted that $\alpha \to 1$ for $p_0 \to 0$.

Further, in sections 1 and 2 (as well as in the previous case with Theorem 1), we will work with models without a space ($p_0 = 0$), write about word weights (the sum of all weights is not equal to one) and correspondingly formulate the results of all the works cited below. To apply these results to the probabilities of the words normalized into the range $(0, 1]$, a modification of the case $p_0 > 0$ is required, which we consider in section 3.

The theorem 1 was extended for the case of Markov dependence in the papers [2, 4] (a somewhat weaker result was independently proved in [7]). Let there be some initial probability distribution $\mathbf{q} = (q_1, \ldots, q_n)$, $q_i \geq 0$, $\sum_{i=1}^{n} q_i = 1$ and a stochastic matrix $P$ of transient probabilities (with the size $n \times n$) with the following property: some power $m$ of this matrix exists so that all the elements of $P^m$ are not equal to zero (the matrix $P$ has ergodic property). Let us redefine the weight of the word (1) by the formula:

$$\mathbb{P}(w) = q_{i_1} \times p_{i_1,i_2} \times \ldots \times p_{i_{N-1},i_N},$$

(4)

where $p_{ij}$ are the corresponding elements of the matrix $P$. Using the formula (4) instead of the formula (2), we correspondingly redefine $p(r)$ — the weight of the $r$-th word in the sorted word list. In [2, 4] the following theorem was proved.

Theorem 2 Suppose that the matrix $P$ has the ergodicity property and the weight of the word $w$ is given by the formula (4). Thus, the inequality (3) with some positive constants $c_1, c_2$ is satisfied for any initial distribution $\mathbf{q}$.

The natural question connected with the inequality (3) is the existence of a limit

$$\lim_{r \to \infty} \frac{p(r)}{r^{-1}}.$$ 

(5)

It is not difficult to verify that in general case the limit may not exist. For example, this is true for independent letters having equal probabilities $p_1 = p_2 = \ldots = p_n = 1/n$. Nevertheless, the following theorem was proved [5, Theorem 2]:

Theorem 3 Suppose that the conditions of Theorem 1 are satisfied and the relation $\frac{\log p_i}{\log p_j}$ for some pair of letters $A_i, A_j$ is irrational. Thus, the limit (5) exists and is equal to $H^{-1}(\mathbf{p})$, where $H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$ is entropy of the vector $\mathbf{p}$. 

2
Unfortunately, the proof methods of Theorem 3 are not applicable in the Markov situation. The aim of this paper was to find the correct formulation of the Markov analogue of this theorem. We have established experimentally the validity of the following assertion:

**Statement 1** Let the elements of the stochastic matrix of transition probabilities \( P \) from Theorem 2 be generated using any multidimensional continuous distribution which guarantees with a single probability that the conditions of this Theorem are satisfied and the vector \( q \) is chosen arbitrarily. Then the limit (5) exists and is equal to \( H^{-1}_{\text{cond}}(P) \), where \( H_{\text{cond}}(P) \) is the conditional entropy of rows of the matrix \( P \) with probabilities equal to the corresponding elements of the stationary distributed vector \( \pi = (\pi_1, \ldots, \pi_n) \), that is

\[
H_{\text{cond}}(P) = - \sum_{i=1}^{n} \pi_i \sum_{j=1}^{n} p_{ij} \log p_{ij},
\]

where \( \pi P = \pi \) and \( \sum_{i=1}^{n} \pi_i = 1 \).

Obviously, Theorem 3 is a special case of Statement 1.

2. The procedure of the experimental proof of the Statement 1.

The graph of the function \( p(r) \) is a set of steps of natural length and positive heights arranged in a descending order. The steps have a height \( \leq 1 \) defined by the formula (4), but they are usually relatively “long” (their length is equal to the number of words having the same weight). We compute the length of each step from the combinatorial considerations for all different values of \( P(w), P(w) > p' \). Then we sort these steps in a descending order \( P(w) \) and as a result find all the pivot points of the graph of the function \( p(r) \) for \( p(r) > p' \). The value of \( p' \) is chosen from taking into account that the number of steps allows us to effectively perform all of the above operations on the computer. Thus, the main idea was to consider classes of words, having the same probability instead of considering all the words.

Besides, the proof of the Statement 1 for any initial distribution \( q \) is reduced to the initial distribution, concentrated in one of the letters. In fact, any stochastic vector \( q \) can be represented as a convex combination of unitary vectors. Correspondingly, the weight of the word (4) can be represented as a convex combination of weights for the cases of initial distributions \( q \) concentrated in one of the letters. Assume that the Statement 1 will be proved for these particular cases. This means that for each of these initial distributions, the weights of all words, except for a finite number, lie in the \( \varepsilon \)-neighborhood of the number \( H^{-1}_{\text{cond}}(P) \). Then the number of exceptional words will be finite for the convex combinations under consideration.

Thus, we will consider only the case of an alphabet containing two symbols \( L = \{A, B\} \) and an initial distribution concentrated in state \( A \). In Section 3, we discuss the reasons why we are sure of the validity of the Statement 1 for an alphabet containing more than 2 symbols.

Let us describe in more detail the considered steps and the method of calculating their lengths and heights. There are 3 kinds of words starting with the letter \( A \), i.e. corresponding to our initial distribution.

1. Words consisting only of the letters \( A \). Let \( A \) occur \( k \) times in a word \( w \). For each fixed \( k \geq 0 \) such word is unique (the length of the step is equal to one) and its weight is given by the formula: \( P_1(w) = p_{11}^k \).

2. Words starting with the letter \( A \) and ending with it, in which the letter \( B \) occurs a certain number of times. Let’s denote the number of transitions in such a word from the state \( A \) to \( B \) and back by \( x \) (\( x \geq 1 \)), the number of transitions from \( A \) to \( A \) (\( y \geq 0 \)) by \( Y \), and the number of transitions from \( B \) to \( B \) (\( z \geq 0 \)) by \( z \). Then, the weight of such a word \( w \) is given by the formula: \( P_2(w) = (p_{12}p_{21})^x p_{11}^y p_{22}^z \).
The number of such words is the number of ways to insert \( y \) transitions \( A \to A \) among \( x + y \) transitions from \( A \), multiplied by the number of ways to insert \( z \) transitions \( B \to B \) among \( x + z - 1 \) transitions from \( B \) (the last transition from \( B \) must be in \( A \)). Thus, the length of the corresponding step is equal to \( \binom{x+y}{y}\binom{x+z-1}{z} \).

3. Words beginning with the letter \( A \) and ending with \( B \). The weight of such a word \( w \) in the preceding paragraph is given by the formula: \( P_3(w) = (p_{12}p_{21})^zp_{11}^yp_{22}p_{12} \), \( x, y, z \geq 0 \). The number of such words (the length of the corresponding step) is \( (\binom{x+y}{y}\binom{x+z-1}{z}) \).

Note that it is convenient to plot not \( p(r) \), but \( -\log p(r) \) (“minus” since \( 0 < p(r) \leq 1 \)) on the ordinate axis. Let us introduce the following notations: \( a = -\log(p_{12}p_{21}) \), \( b = -\log p_{11} \), \( c = -\log p_{22} \), \( d = -\log p_{21} \). Thus, the logarithms of the weights of the words from Sections 1 to 3 are given by formulas:

\[
\begin{align*}
-\log P_1(w) &= (k - 1)b \\
-\log P_2(w) &= xa + yb + zc \\
-\log P_3(w) &= xa + yb + zc + d
\end{align*}
\]  
(7) (8) (9)

If \( p' = e^{-300} \), then to find the number of classes of words satisfying the condition \( P(w) > p' \) we should sort out all possible values \( k \geq 1 \) so that the right-hand side of the formula (7) is less than 300.

Besides, we need to sort out all possible values \( x \geq 1 \), \( y, z \geq 0 \) so that the right side of the formula (8) is less than 300. In the case the values \( p_{11}, p_{21} \) have uniform distribution within the interval, \( [0, 1] \) (in this case, the values of the remaining elements of the matrix \( P \) are uniquely determined: \( p_{12} = 1 - p_{11}, \ p_{22} = 1 - p_{12} \)) it turns out that the number of different classes of words (the number of different steps) has the order of \( 10^7 \).

To specify the graph of the function \( -\log p(r) \), we need to know the abscissas and ordinates of the pivot points of this graph. To calculate them, we sorted the steps in the order of increasing values given by formulas (7,8,9) (the ordinate of the pivot points). In this case, the lengths of the steps are given by combinatorial formulas from sections 1–3. Thus, as a result of sorting, we obtained a series of natural numbers — lengths of the steps. The abscissas of the pivot points are obviously partial sums of this series. Note that the Statement 1 is equivalent to the assertion

\[
\lim_{r \to \infty} (\log r + \log p(r)) = -\log H_{\text{cond}}(P).
\]  
(10)

Thus, to verify the Statement 1 experimentally, it is convenient to introduce a logarithmic scale along both the ordinate and the abscissa axes. Therefore, we took the logarithm of the abscissas of the pivot points found by the method described in the previous paragraph.

As a result, we obtained a step-by-step graph of the dependence of \(-\log p(r)\) on \(\log r\), which lies “along the bisector of the first coordinate angle”. The differences between the abscissa and ordinate values for the pivot points of this graph (that is, the values of \(\log r + \log p(r)\)) are shown in Fig. 1. It can be seen that for large values of \(\log r\), the oscillations of the graph are damped. Having calculated the average value \((\log r + \log p(r))\) by the last thousand of the pivot points, we experimentally obtained the required constant from the formula (10). Later we managed to formulate Statement 1 and the result coincided with the experimental one up to 3–4 digits after the decimal point (note that the values \(\log H_{\text{cond}}(P)\) for randomly generated matrices \(P\) are of the order of 1).

3. Discussion of the obtained results.
We have experimentally proved Statement 1 for \( n = 2 \). We considered approximately \( e^{300} \approx 10^{130} \) number of words (in fact, about \( 10^7 \) classes of words). We checked the validity of the main
Figure 1. The angular points of the graph of the function \( p(r) \) as \(- \log p(r) < 100\) for the matrix \( P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}\) the initial distribution \( q = \begin{pmatrix} 1 \\ 0 \end{pmatrix}\).

statement both for completely non-zero matrices \( P \) of dimension \( 2 \times 2 \) and for matrices which have \( p_{11} = 0 \) or \( p_{22} = 0 \).

Inequalities \(- \log P_2 < 300, \ - \log P_3 < 300\) (taking into account the relations (8, 9)) are analogous to the restrictions on the summation domain of combinations in the Pascal triangle, which were used in [5] for the proof of Theorem 3. This suggests that the Statement 1 is valid not only for randomly generated matrices \( P \), but in general for all matrices in which there is a pair of elements in one row with an irrational ratio of their logarithms.

M.D. Ward drew our attention that the Theorem 3 has an analog in the form of some statement for digital trees (also known as “tries”). In this case, the mean for which the statement is formulated satisfies a completely different type of recurrence relation, than \( p(r) \) (see [9, p.404]). After the experimental proof of Statement 1, we have found the article [8] with the same formula (6) for digital trees in the case of Markov dependence. We are going to apply the complex analysis technique used there, for mathematical proof of the Statement 1. We are also going to investigate the nature of fluctuations \( p(r) \) near its limit value (see Figure 1).

Finally, let us discuss the application of Statement 1 to the case when there is a certain small probability \( p_0 \) to get from each state into the space (the symbol of the end of the word). Then, the probability of each word is \( P(w) p_0 \), where the weight \( P(w) \) is given by the formula (4). Thus, the weight \( P(w) \) of the word can be restored by its probability, and the value \( p(r) \) of the weight of a word of rank \( r \) can be restored with the sorted list of probabilities. It was actually proved in [4] that in this case for some positive \( c_1, c_2 \) the following inequality is correct

\[ c_1 r^{-1/\gamma} < p(r) < c_2 r^{-1/\gamma}, \]

where \( \gamma \) is the power to which all elements of the matrix \( P \) must be raised, so that its spectral radius is equal to one. Note that for sufficiently small values \( p_0 \), the value \( \gamma \) is arbitrarily close to one. Assuming that the limit \( \lim_{r \to \infty} r^{-1/\gamma} p(r) \) exists and depends continuously on \( P \) and \( p_0 \), then, according to Statement 1 this limit will be arbitrarily close to \(- \sum_{i=1}^{n} \pi_i \sum_{j=1}^{n} p_{ij} \log p_{ij} \). In the case of sufficiently small values of \( p_0 \), where \( \pi \) is the eigenvector of the matrix \( P \) corresponding to its largest eigenvalue, \( \sum_{i=1}^{n} \pi_i = 1 \).
Acknowledgements
This work was supported by the Russian Foundation for Basic Research, Grant no. 15-06-07402. The research of the first author was supported by the Russian Government Program of Competitive Growth of Kazan Federal University.

References
[1] Baayen R H 2001 Word Frequency Distributions (Dordrecht: Kluwer Academic Publishers)
[2] Bochkarev V V and Lerner E Yu 2012 Zipf and non-Zipf Laws for Homogeneous Markov Chain Preprint arXiv:1207.1872
[3] Bochkarev V V and Lerner E Yu 2012 Russian Mathematics 50 25
[4] Bochkarev V V and Lerner E Yu 2014 Electron. J. Linear Algebra 27 534
[5] Bochkarev V V and Lerner E Yu 2017 International Journal of Mathematics and Mathematical Sciences 2017 Article ID 914374
[6] Conrad B and Mitzenmacher M 2004 IEEE Trans. Inform. Theory 50 no 7 1403
[7] Edwards R, Foxall E and Perkins T J 2012 Electron. J. Linear Algebra 23 966
[8] Jacquet Ph and Szpankowski W 1991 IEEE Trans. Information Theory 37 1470
[9] Sedgewick R and Flajolet P An Introduction to the Analysis of Algorithms 1995 Addison–Wesley Boston Mass