Zipf exponent of trajectory distribution in the hidden Markov model

V.V. Bochkarev and E.Yu. Lerner

Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: vbockarev@mail.ru, eduard.lerner@gmail.com

Abstract. This paper is the first step of generalization of the previously obtained full classification of the asymptotic behavior of the probability for Markov chain trajectories for the case of hidden Markov models. The main goal is to study the power (Zipf) and nonpower asymptotics of the frequency list of trajectories of hidden Markov frequencys and to obtain explicit formulae for the exponent of the power asymptotics. We consider several simple classes of hidden Markov models. We prove that the asymptotics for a hidden Markov model and for the corresponding Markov chain can be essentially different.

In mathematical linguistics, one of the most widely known frequency laws is the Zipf law [1]. According to this law, there is a power dependence of the probability of the rth word in an ordered frequency list on r. The exponent in the classical Zipf law equals -1. In recent years, opportunities for this research became much wider, owing to the appearance of new large data repositories, in particular, Google Books Ngram corpus [6]. The analysis of large data corpora shows that the power dependence with an exponent close to -1 can approximately describe only frequencies of most commonly used words [4]. At the same time, frequencies of the main and peripheral vocabularies can be approximately described by a power law whose exponent is modulo greater than 1 (its typical values lie within the interval 1.7–2). Moreover, in languages with hieroglyphic scripts the asymptotic decrease of frequencies is faster than power law [5]. The development of a model explaining these phenomena is an actual problem.

There were many attempts to explain the frequency distribution of word usage with the help of probabilistic frequencies of text generation. One has considered both the ”monkey model” with independent letters, and Markov models.In [2] one describes the dependence of the asymptotic behavior of the frequency distribution on the structure of the transition probability matrix and obtains an explicit formula for the exponent. Note that unlike the monkey model [3], where there is a great abundance (because all words consisting of the same letters have equal probabilities), in Markov models, owing to dependencies between states, there is no such abundance. We can observe even a less entropy (and, consequently, a less abundance) in a hidden Markov model (HMM), where some hidden symbols merge to form the same visible ones. In future, the use of HMM will allow one to construct more realistic text generation models, taking into account the syllabic and morphological word structure. Moreover, within HMM one can also take into account the presence of misprints and mistakes in word corpora (according to [6], up to 30 percent of unique word forms in Google Books Ngram appear as a result of word recognition errors). Another linguistic application of HMM is the statistics of various sentences. In the latter case, graphs of Markov chains (MC) represent Chomsky syntactic structures, while HMM do their realizations as sentences. Our goal is to study the power and nonpower asymptotics of the frequency list of trajectories of HMM and to obtain explicit formulae for the power asymptotics exponent. Note that the value of this exponent appears to be greater than one, which corresponds to true values, if we estimate the transition probability matrix in accordance with the word frequencies in Google Books Ngram. In particular, these models with lighter distribution tails (in comparison with the classical Zipf model with the exponent of -1) better predict frequencies of peripheral vocabularies and the small quantity of hapax legomena.

Let us give exact definitions. Consider a Markov chain (MC) whose state set is $E = \{E_0, \ldots, E_n\}$; here the state E_0 is absorbing, while the rest ones are nonrecurrent, i.e., having started with any state, we finally get at E_0. Consider a hidden Markov model (HMM), where each non-absorbing state is associated with a random variable that takes on values in some alphabet $X = \{x_1, \ldots, x_m\}$; thus we get words representing a sequence of letters of the alphabet X. A word ends, when we reach the state E_0. If the initial distribution $a = (a_1, \ldots, a_n)$ is known, then each word has a certain probability, and the sum of these probabilities equals one. We are interested in the distribution of probabilities over words, namely, we are interested in decrease rates of word probabilities in their complete list sorted in the non-increasing order of probabilities.

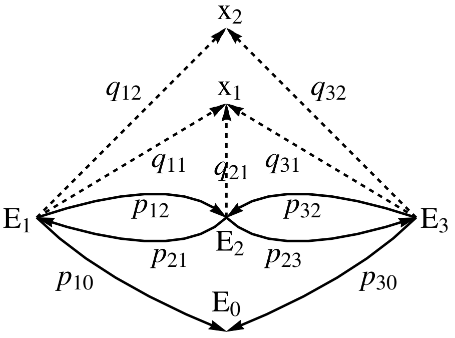

Figure 1. An example of the graph of an HMM with three states with a two-letter alphabet, $q_{21} = 1$.

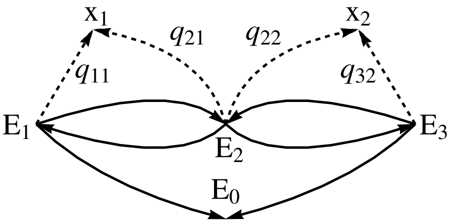

Figure 2. Another example of the graph of an HMM (modifying the example shown in Fig. 1), for which the hidden word is uniquely defined from the exposed one (see Theorem 3); here $q_{11} = q_{32} = 1$.

For example, let us consider a HMM, whose graph is given in Fig. 1. Here vertices correspond to states of the MC, solid arcs do to nonzero elements of the matrix of transition probabilities, and weights shown next to arcs are values of the corresponding elements of the matrix. Dashed arcs outgoing from non-absorbing vertices correspond to various variants of values of random variables associated with vertices, and weights shown next to them are probabilities of taking on the corresponding values. Evidently, for any non-absorbing vertex the sum of weights of all outgoing solid arcs, as well as that of all dashed ones, equals one. In particular, in Fig. 1, $q_{21} = 1$. In this example, assuming that at the initial time moment we certainly are situated at state E_2, possible words represent any sequences consisting of an even number of letters of the alphabet $\{x_1, x_2\}$, where odd positions are necessarily occupied by the symbol x_1. In particular, the probability of the word (x_1, x_1) equals $p_{21}q_{11}p_{10} + p_{23}q_{31}p_{30}$, that of the word (x_1, x_2) does $p_{21}q_{12}p_{10} + p_{23}q_{32}p_{30}$, and so on.

Introduce some denotations. We are interested in the decrease order of the function $p(\cdot)$, where $p(r)$ is the probability to obtain the rth word in the sorted list. Let G_0 be the graph of an MC (in Fig. 1 this graph consists of solid arcs), and let G be the subgraph of G_0 consisting of all its vertices and arcs, except the vertex E_0 and arcs entering it.

Evidently, if the graph G contains at least one cycle, then (and only in this case) the number of words is infinite. We are interested just in this case, namely, in the order of the asymptotics of the function $p(r)$.

Let P_G be a substochastic matrix of transition probabilities of the MC corresponding to the graph G (i.e., the transition matrix where the absorbing state is not taken into account) and let $P_G(\beta)$ be obtained from P_G by raising each its element to the power β. Earlier in [2] we considered ordinary MC (rather than HMM), i.e., the case when $X = \{E_1, \dots, E_n\}$ and each random variable is a deterministic identical map. We classified all possible variants of essentially different asymptotics of the function $p(r)$ and consider some techniques for finding parameters of these asymptotics for MC. In particular, we have proved the following assertion (we use the standard O-symbols, including the symbol Θ for denoting the asymptotic order and the symbol Ω for the lower estimate of the order):

Theorem 1 *Assume that the initial distribution a is such that the probability to reach (at least once) each state E_i, $i = 1, \dots, n$, is greater than zero. Then the following alternatives are possible:*

1. *If the graph G contains a vertex which go through to two different simple cycles, then $p(r) = \Omega(r^{-1/\beta})$, where β is a real value, with which the greatest modulo eigenvalue of the matrix $P_G(\beta)$ equals one. Note that such β exists, it is unique and belongs to the interval $(0, 1)$. Moreover, $p(r) = o(r^{-1/\gamma})$ for any $\gamma > \beta$. Note that the exact power order (i.e., the equality $p(r) = \Theta(r^{-1/\beta})$) takes place if and only if any simple path in the graph G belongs to no more than one strongly connected component H of this graph, for which the matrix $P_H(\beta)$ has unit eigenvalue.*
2. *If the graph G contains cycles, and each vertex of the graph G belongs to no more than one simple cycle, then $p(r) = \Omega(\alpha^r)$ and $p(r) = o(r^{-\lambda})$, where λ is any positive number and $\alpha \in (0, 1)$ is a constant depending on the matrix P (i.e., $p(r)$ is decreasing faster than any power function, but slower than some exponential one). Moreover, the exact exponential asymptotics ($p(r) = \Theta(\exp(-\nu r))$) for some $\nu > 0$ (see [2] for the calculation of ν) takes place if and only if any path in the graph G goes through vertices of no more than one cycle.*

Thus, for example, for a Markov chain with the graph G given in Fig. 1 (we pay no attention to dashed arcs) there is only one strongly connected component $p(r) = \Theta(r^{-1/\beta})$, where β is the root of the equation $p_{12}^{\beta} p_{21}^{\beta} + p_{23}^{\beta} p_{32}^{\beta} = 1$. The power asymptotics is typical; it defines the Zipf law [1] for frequencies of the occurrence of various words. Theorem 1 implies that if the matrix of transition probabilities is very sparse (more exactly, conditions of Theorem 1.2 are fulfilled), then $p(r)$ is decreasing faster than the power function. This property is valid for hieroglyphic scripts, in particular, for the Chinese language [5].

For a vertex v we understand a syllable $w(v)$ as a part of a word in an HMM which can be obtained by tracing some simple cycle in the graph G beginning at v. For example, for the HMM illustrated in Fig. 1 there are only two syllables in the form $w(E_2)$, namely, (x_1, x_1) and (x_1, x_2). We say that syllables $w_1(v)$ and $w_2(v)$ are essentially distinct, if no (multiple) concatenation of one syllable gives another one. For example, all letterwise distinct syllables of the same length are essentially distinct, while syllables composed of letters of the Latin alphabet (o, l, e) and $(o, l, e, o, l, e, o, l, e)$ are not so.

Theorem 2 *Assume that the initial distribution a is such that the probability to reach (at least once) each state E_i, $i = 1, \dots, n$, is greater than zero. For an HMM we have the bound $p(r) = \Omega(r^{-\beta'})$ for certain β' if and only if there exists a vertex v, for which one can find at least two essentially distinct syllables $w_1(v)$ and $w_2(v)$.*

Evidently, conditions of Theorem 2 are fulfilled for the HMM illustrated in Fig. 1.

Note that the fulfillment of conditions of Theorem 1.1 does not guarantee the fulfillment of conditions of Theorem 2. Similarly, the fulfillment of conditions of Theorem 2 does not guarantee the fulfillment of conditions of Theorem 1.1, i.e., the power order of the asymptotics of "exposed" and "hidden" words, generally speaking, is inherited neither from MC to HMM, nor vice versa.

Let us now consider the question on the exact order of the power asymptotics. Denote by the symbol Q a stochastic $n \times m$-matrix, whose elements q_{ij} ($i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$) are probabilities that the random variable corresponding to the state E_i takes on the value x_j. Denote by $\widehat{P}(\beta)$ the $n \times n$-matrix, whose elements $\widehat{p}_{ij}$ ($i, j \in \{1, \dots, n\}$) obey the formula $\widehat{p}_{ij} = p_{ij}^{\beta} \sum_{k=1}^{m} q_{jk}^{\beta}$.

Theorem 3 *Let conditions of Theorem 2 be fulfilled. For simplicity, we assume that the graph G consists of one strongly connected component. In addition, we assume that (with the considered initial distribution) for each exposed word w (i.e., a sequence of letters of the alphabet X or, in other words, values of random variables associated with states E_i), the realizing it hidden word (i.e., the set of states E_i visited by us) is defined uniquely. Then $p(r) = \Theta(r^{-1/\beta})$, where β is a real value, with which the greatest modulo eigenvalue of the matrix $\widehat{P}(\beta)$ equals one (it exists and belongs to the interval $(0, 1)$).*

Note that the HMM shown in Fig. 1 does not satisfy conditions of Theorem 3; for example, one can obtain the word (x_1, x_1) in two ways, namely, with the help of either the hidden word (E_2, E_1) or that (E_2, E_3). We can satisfy conditions of Theorem 3 by slightly changing directions of dashed arcs as is shown in Fig. 2. In this case any exposed word, whose $2l$th position is occupied by x_1, corresponds to a hidden word, whose $2l$th position is occupied by E_1; any exposed word, whose $2l$th position is occupied by x_2, corresponds to a hidden word, whose $2l$th position is occupied by E_3. In addition, odd positions in a hidden word are occupied by E_2. In this case by Theorem 3 we have $p(r) = \Theta(r^{-1/\beta})$, where β is defined from the equation

$$(p_{12}^{\beta} p_{21}^{\beta} + p_{23}^{\beta} p_{32}^{\beta})(q_{21}^{\beta} + q_{22}^{\beta}) = 1.$$

The condition for recovering a hidden word from an exposed one seems to be not very strong, however, it is just the case (due to the abundance of the language information) for the statistics of recognized (with errors) wordforms in the Google Books repository.

Corollary 1 *Assume that an HMM satisfies conditions of Theorem 2 (including the possibility to reach each state) and for hidden words its MC has the exponential asymptotics. Then $p(r) = \Theta(r^{-1/\beta})$, where β is a real value, with which the greatest modulo eigenvalue of the matrix $\widehat{P}(\beta)$ equals one.*

Let us now consider approaches to the calculation of β in the case, when the recovery of a hidden word from an exposed one may appear to be impossible. Let $w(v)$ be some syllable in an HMM; assume that one can obtain this syllable in several ways when tracing a cycle (cycles) in the graph G, beginning at a vertex v. We understand the weight of such a tracing c as the product of weights of all edges that enter in the cycle c multiplied by the product of probabilities of the corresponding values from the alphabet X in the corresponding state E_i. We understand the weight of a syllable $Pr(w(v))$ as the sum of weights calculated over all such tracings.

For example, for the HMM illustrated in Fig. 1 there are two syllables of type $w(E_2)$, namely, (x_1, x_1) and (x_1, x_2); we have $Pr(x_1, x_1) = p_{21}q_{11}p_{12} + p_{23}q_{31}p_{32}$, while $Pr(x_1, x_2) = p_{21}q_{12}p_{12} + p_{23}q_{32}p_{32}$. For the HMM illustrated in Fig. 2 there are four syllables, namely, all possible two-letter combinations.

Theorem 4 *Let conditions of Theorem 2 be fulfilled. Moreover, we assume that all cycles in the graph G contain some vertex v and all letterwise distinct syllables $w(v)$ are essentially distinct. Then $p(r) = \Omega(r^{-1/\beta})$, where β is determined as the root of the equation $\sum Pr(w(v))^{\beta} = 1$, and the sum is calculated over all possible syllables $w(v)$.*

Theorems 2, 3, and 4 cover particular cases of rather sparse matrices P or Q. Note that even in this case the asymptotics of the frequency list of hidden and visible states can be essentially different. Thus, in Theorem 3 it is proved that the exponent of the power asymptotics of visible states can be determined with the help of the matrix $P'(\beta)$, while the exponent of the power asymptotics of hidden states (see Theorem 1) can be determined with the help of the matrix $P(\beta)$. Since, evidently, $\widehat{p}_{ij} \geq p_{ij}^{\beta}$, for hidden states this exponent is (modulo) not less than that for visible ones (we assume that hidden states also have power asymptotics). An essentially different asymptotic behavior is also possible (e.g., Corollary 1). In a general case, frequencies of hidden words do not necessarily decrease faster than those of visible ones. Thus, for example, under conditions of Theorem 4, if several hidden syllables correspond to the same visible ones, then exponents of the power asymptotics satisfy the converse inequality.

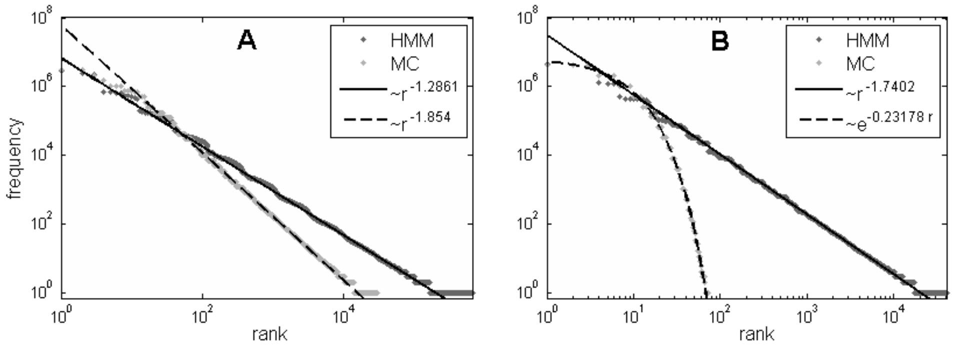

Figure 3. Examples of the frequency graph of an HMM and the corresponding MC. The total number of generated (nonunique) words equals 25 million.

In Fig. 3 we represent results of simulation experiments that demonstrate the essential distinction of asymptotics of frequencies of visible and hidden words; everywhere the number of hidden and visible states equals three. In case A, $P_G = \begin{pmatrix} 0.0007 & 0.007 & 0.6823 \\ 0.3493 & 0.0014 & 0.3493 \\ 0.007 & 0.6923 & 0.0007 \end{pmatrix}$, $Q = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$, in case B, $P_G = \begin{pmatrix} 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix}$, $Q = \begin{pmatrix} 0.75 & 0.15 & 0.1 \\ 0.15 & 0.75 & 0.1 \\ 0.1 & 0.15 & 0.75 \end{pmatrix}$. In case A, both the frequencies of hidden words and visible ones have power asymptotics, but their exponents are essentially different (they are calculated in accordance with Theorem 1 and results of the simulation experiment, correspondingly). In case B, frequencies of hidden words decrease exponentially, while those of visible ones have power asymptotics.

Thus, we have established explicit formulae for exponents of the power asymptotics for some variants of HMM. We show that asymptotics of frequency lists of an HMM and the corresponding MC can be essentially different.

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] Gerlach M and Altmann E G 2013 *Physical Review* **X**(3), 021006

[5] Lu L, Zhang Zi-Ke and Zhou T 2012 Scaling Laws in Human Language *Preprint* arXiv:1202.2903.

[6] Michel JB, Shen JB, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA and Aiden EL 2011 *Science* **331** 176