are all nonincreasing. Differentiate (4) with respect to p. The terms containing derivatives of Es_{k-1} are nonpositive and so

$$\frac{d}{dp}Es_m \le \sum_k \binom{m}{k}\{k[p^{k-1}q^{m-k} - p^{m-k}q^{k-1}] - (m-k)[p^k q^{m-k-1} - p^{m-k-1}q 2^k]\}Es_{k-1}.$$

Use a binomial identify to obtain

$$\begin{aligned}\frac{d}{dp}Es_m &\le m\sum_k \binom{m-1}{k-1}(p^{k-1}q^{m-k} - p^{m-k}q^{k-1})Es_{k-1} \\ &\quad - m\sum \binom{m-1}{k}(p^k q^{m-k-1} - q^k p^{m-k-1})Es_{k-1} \\ &= m\sum_k \binom{m-1}{k}(p^k q^{m-k-1} - p^{m-k-1}q^k)E(s_k - s_{k-1}) \\ &= m \sum_{k<\frac{m-1}{2}} \binom{m-1}{k}(p^k q^{m-k-1} - p^{m-k-1}q^k) \\ &\quad \cdot[E|W_k| - E|W_{m-1-k}|].\end{aligned}$$

The square-bracketed term is nonpositive for $p \le q$ because $E|W_k| > E|W_j|$ for $k > j$. Then, Es_m is indeed nonincreasing.

References

[1] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[2] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, pp. 278–286, Mar. 1988.

[3] ——, "Compression, test of randomness, and estimating the statistical model of individual sequences," *SEQUENCES*, R. M. Capocelli, Ed. New York: Springer-Verlag, 1990, pp. 366–373.

[4] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75–81, Jan. 1976.

[5] J. Rissanen, "A universal data comparison system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–664, Sept. 1983.

Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution

Wentian Li

***Abstract*—It is shown that the distribution of word frequencies for randomly generated texts is very similar to Zipf's law observed in natural languages such as English. The facts that the frequency of occurrence of a word is almost an inverse power law function of its rank and the exponent of this inverse power law is very close to 1 are largely due to the transformation from the word's length to its rank, which stretches an exponential function to a power law function.**

***Index Terms*—Statistical linguistics, Zipf's law, power-law distribution, random texts.**

Manuscript received June 2, 1991; revised March 2, 1992. This work was supported by the MacArthur Foundation, the National Science Foundation under Grant PHY-87-14918, and the Department of Energy under Grant DE-FG05-88ER 25054. This work was performed at the Santa Fe Institute.

The author is with Rockefeller University, Box 167, 1230 York Avenue, New York, NY 10021.

IEEE Log Number 9201222.

Zipf observed long time ago [1]–[3] that the distribution of word frequencies in English, if the words are aligned according to their ranks, is an inverse power law with the exponent very close to 1. In other words, if the most frequently occurring word appears in the text with the frequency $P(1)$, the next most frequently occurring word has the frequency $P(2)$, and the rank-r word has the frequency $P(r)$, the frequency distribution is

$$P(r) = \frac{C}{r^\alpha}, \tag{1}$$

with $C \approx 0.1$ and $\alpha \approx 1$. This distribution, also called Zipf's law, has been checked for accuracy for the standard corpus of the present-day English with very good results [4].

The fall-off of the distribution as the rank is increased is obvious, because the more frequently occurring words are guaranteed to have larger frequencies than those less frequently occurring. Nevertheless, it seems to be a puzzle as why the decay is a power law instead of an exponential function or other faster decaying functions, and why the exponent is very close to 1 instead of 2 or even larger values. There are attempts to incorporate Zipf's law into the grander framework of "fractals" [5]–[7], but in doing so, little insight has been gained in understanding this particular "law."

Probably few people pay attention to a comment by Miller in his preface to Zipf's book [8] that randomly generated texts, which are perhaps the least interesting sequences and unrelated to any other scaling behaviors, also exhibit Zipf's law. What he said was that Zipf's law is not exclusive for English or any other natural language. Miller did not give a proof of his statement, and it is the purpose of this short paper to provide a very simple proof that random texts do indeed exhibit Zipf's-law-like word frequency distribution.

By "random texts," I mean the symbolic sequences generated by the following procedure: each symbol out of total $(M + 1)$ symbols is selected randomly and deposited at position i, and another symbol is randomly selected and deposited at position $i + 1$, and so on. There is no correlation between the selection of symbol at position i and that at position $i + 1$. Among the $(M + 1)$ symbols, one of them is called the "blank space." Any "nonblank" symbol string between two blank spaces is called a 'word," whereas a string of blank spaces is not. Taking the English alphabets for example, $M = 26$, and the words in random texts can be *a*, *b*, *c*,$\cdots$, *aa*, *ab*, *ac*,$\cdots$, *ba*, *bb*,$\cdots$, *aaa*, *aab*,$\cdots$, etc. If the following sequence, for example, is generated,

$$a\_mdf\_pwell\_\ \_werlppa\_re\_\_\_kkel\_,$$

it then contains the words *a* (suppose that the beginning of the sequence also plays the role of a blank space, (*mdf*, *pwell*, *werlppa*, *re*, and *kkel*.

The probability that one would see the string $\_a\_$ in a random text is proportional to $(1/27)^3$, which is equal to the product of the probability for the first symbol to be a blank space $(= 1/27)$, for the second symbol to be $a(= 1/27)$, and for the third symbol to be a blank space $(= 1/27)$. Similarly, the probability for finding the string $\_bsl\_$ is proportional to $(1/27)^5$. Since the first probability is also the frequency of occurrence for any word with length 1 (except a normalization factor), and the second probability is the frequency of occurrence for any word with length 3 (again, except a normalization factor), we have the general formula for the frequency of occurrence for any word with

length L:

$$P_i(L) = c\frac{1}{(M+1)^{L+2}}, \qquad i = 1,2,\cdots,M^L. \tag{2}$$

With C as the normalization factor, note that there are M^L words having length L.

The constant c can be determined from the normalization condition for the frequencies of occurrence of all words:

$$\sum_{L=1}^{\infty} M^L \frac{c}{(M+1)^{L+2}} = c\frac{M}{(M+1)^2} = 1, \tag{3}$$

so

$$c = \frac{(M+1)^2}{M}. \tag{4}$$

Inserting the value of c back to the (2), the frequency of occurrence for any particular word with length L is

$$P_i(L) = \frac{1}{M(M+1)^L} \tag{5}$$

and the frequency of occurrence for all words with length L is

$$P(L) = M^L P_i(L) = \frac{M^{L-1}}{(M+1)^L}. \tag{6}$$

Both are exponential functions of L.

In a random text, all words with the length L rank higher than words with the length $L+1$, because they have larger value of frequency of occurrence by (5). If we represent the rank of any word with length L by $r(L)$, we have

$$\sum_{l=1}^{L-1} M^l < r(L) \le \sum_{l=1}^{L} M^l \tag{7}$$

or

$$\frac{M}{M-1}(M^{L-1}-1) < r(L) \le \frac{M}{M-1}(M^L - 1). \tag{8}$$

For example, $0 < r(1) \le M$, $M < r(2) \le M + M^2$, and so on. Equation (8) represents the exponential transformation from word's length to word's rank. One implication of the transformation to be exponential is that the longer the L, the more "stretching" of the rank variable, since there are more number of words with longer lengths.

Equation (8) can be converted to

$$L - 1 < \log_M\left(\frac{M-1}{M}r(L) + 1\right) \le L. \tag{9}$$

Raising $1/(M+1)$ to the power of all the terms gives

$$\frac{1}{(M+1)^{L-1}} > \left(\frac{1}{M+1}\right)^{\log_M\left(\frac{M-1}{M}r(L)+1\right)} \ge \frac{1}{(M+1)^L}; \tag{10}$$

multiplying all terms by $1/M$ gives

$$\frac{1}{M(M+1)^{L-1}} > \frac{1}{M}\left(\frac{1}{\frac{M-1}{M}r(L)+1}\right)^{\frac{\log(M+1)}{\log(M)}} \ge \frac{1}{M(M+1)^L}, \tag{11}$$

which can be written as

$$P_i(L) < \frac{C}{(r(L)+B)^\alpha} \le P_i(L-1), \tag{12}$$

with

$$\alpha = \frac{\log(M+1)}{\log(M)}, \qquad B = \frac{M}{M-1},$$
$$\text{and } C = \frac{1}{M}\frac{M^\alpha}{(M-1)^\alpha} = \frac{M^{\alpha-1}}{(M-1)^\alpha}. \tag{13}$$

The functional form

$$P(r) = \frac{C}{(r+B)^\alpha} \tag{14}$$

is also called the generalized Zipf's law by Mandelbrot [9]. Let us check how close the generalized Zipf's law for random texts can be to Zipf's law in English: since the number of alphabets is $M = 26$, we have $\alpha = 1.01158$ and $C = 0.04$. The exponent α is extremely close to what is observed in English, an amazing fact considering how little we have assumed. Even with the minimum number of symbols, $M = 2$ (if $M = 1$, the transformation from the word length to the word rank is linear, and no power-law distribution is expected), $\alpha = 1.58496$ is still not that far from 1.

The frequency of occurrence of words by their rank represented by (12) does not have the problem of divergence of the total probability typical for a power-law distribution, because the exponent $\alpha = 1.01158$ is strictly larger than 1—which takes care of the integration at the tail end; and there is a cutoff of the smallest word rank, i.e., $r = 1$—which takes care of the integration at the zero value of the rank.

Due to the assumption that each symbol appears in the sequence with exactly the same probability, all words with the same length have the same frequency of occurrence. In other words, $P(r)$ is a stepwise function having plateaus on $P_i(L)$'s. Fig. 1 shows a numerical result of the word frequency distributions for random texts with 2, 4, and 6 symbols, respectively. In the numerical simulation, a sequence of length N (which is 80 000, 200 000, and 600 000 for $M = 2$, 4, and 6) is generated with the M symbols and the blank space all having the equal probability. I also introduce a cutoff of the maximum possible word length $L_{\max}$ (6, 4, and 3 for $M = 2$, 4, and 6). The frequency of a word is derived by dividing the number of occurrence of that word with the total number of word countings (which is 16 306, 18 964, and 36 320 for $M = 2$, 4, and 6). Since we do not count words whose lengths are longer than the cutoff length, the normalization condition (3) now becomes

$$\sum_{L=1}^{L_{\max}} M^L \frac{c}{(M+1)^{L+2}} = \frac{c}{(M+1)^2} M\left(1 - \left(\frac{M}{M+1}\right)_{\max}\right) = 1, \tag{15}$$

which leads to a larger value of c

$$c = \frac{(M+1)^2}{M}\frac{1}{1-\left(\frac{M}{M+1}\right)^{L_{\max}}} \tag{16}$$

but the α estimated by (13) should be the same. To make a comparison with Zipf's law ($\sim 1/r$) as well as the power law with the exponent $2(\sim 1/r^2)$, these two functions are also plotted in Fig. 1. The numerical simulation confirms that the random texts

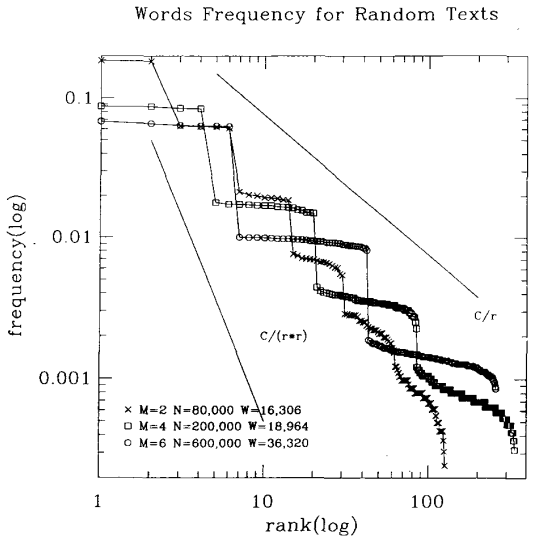

Fig. 1. Word frequency as the function of the word's rank for randomly generated sequences with the number of symbols $M = 2$, 4, and 6. There is a cutoff for the longest word length to be counted (the cutoffs L_{max} are 6, 4, and 3 respectively for $M = 2$, 4, and 6). All symbols including the blank space have the same probability to appear in the sequence. Frequency of occurrence of a word is the number of countings of that word divided by the number of countings of all words (they are 16 306, 18 964, and 36 320 respectively for $M = 2$, 4, and 6). Also shown are Zipf's scaling law (power law function with the exponent 1) and the power law function with the exponent 2.

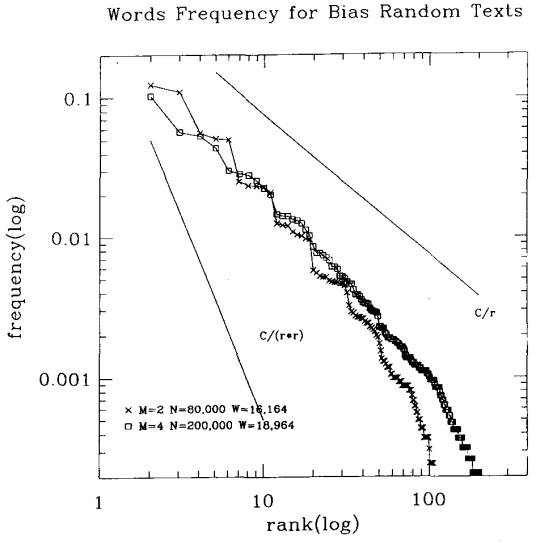

Fig. 2. Word frequency as the function of the word's rank for biased random sequences with the number of symbols $M = 2$ and 4. There is a cutoff for the longest word length to be counted ($L_{max} = 6$ and 4, respectively for $M = 2$ and 4). Different symbols as well as the blank space have different probability to appear in the sequence (see the text for their values). Frequency of occurrence of a word is the number of countings of that word divided by the number of countings of all words (they are 16 164 and 18 964, respectively for $M = 2$ and 4). Also shown are Zipf's scaling law (power law function with the exponent 1) and the power law function with the exponent 2.

exhibit a word frequency distribution very much the same with Zipf's law.

It is clear now that the existence of the Zip's-law-like word frequency distribution in random texts is purely due to the choice of the rank as the independent variable. By choosing the word rank rather than the word length, the exponential distribution which is typical for random texts becomes a power law function. This strongly suggests that the power law as expressed by Zipf's law in natural languages is also purely due to the choice of the rank as the independent variable. Actually, besides the cardinal number and the ordinal number, one can also use the third representation of the same frequency distribution: the distribution on a certain position of a digit, as related to the notorious "first digit problem" [10]. The transformation among the three representations is summarized by Gell-Mann [11].

Equations (13) also explains why the exponent in Zipf's law is close to 1, simply because $\log(M + 1) \approx \log(M)$ when M is large. As we have seen, even for the worse case of having two symbols ($M = 2$), the estimated $\alpha \approx 1.58$ is still smaller than 2. Only for $M = 1$ (the sequence is a binary sequence with one symbol and one blank space), no mechanism exists for stretching the frequency distribution from exponential to power law, and we fail to recover Zipf's law. If Zipf's law is observed for binary sequences, it indicates a "true" power law scaling, and one should expect other nontrivial scaling behaviors, such as $1/f$ noise and long-range correlations [12].

The stepwise structure of the frequency of occurrence distribution in Fig. 1 can be removed by introducing bias among different symbols, i.e., different symbols have different probabilities to appear in the sequence. For example, symbol a can be more likely to appear in the sequence than symbol b; and consequently, word $\_a\_$ has a larger value of the frequency of occurrence than word $\_b\_$. The plateaus are then easily destroyed. In particular, a word with longer length can have a larger frequency than the words with shorter lengths; for example, word $\_aa\_$ ranks higher than word $\_b\_$ if the square of the probability for symbol a to appear in the sequence is larger than the probability for symbol b. Fig. 2 shows the numerical results for two biased random sequences with two and four symbols respectively. For the two-symbol sequence we choose the probability for having blank space to be 0.33, the probability for the first symbol is 0.47 and that for the second symbol is 0.2 (these numbers are arbitrarily chosen). For the four-symbol sequence, the probability for the blank space is 0.2, those for the remaining symbols are 0.5, 0.13, 0.1, and 0.07 (again, those are arbitrary numbers). A much smoother power law distributions show up in Fig. 2.

In conclusion, Zipf's law is not a deep law in natural language as one might first have thought. It is very much related the particular representation one chooses, i.e., rank as the independent variable. A symbolic sequence which exhibits Zipf's law does not have to exhibit other scaling phenomena such as the $1/f$ noise or long-range correlation. In fact, the long-range correlation and $1/f$ spectrum are absent in natural languages, as observed by the author that the mutual information function between two letters decays faster than power laws of small exponents [13]. Mandelbrot [5] seems to derive the same result that random texts exhibit the generalized Zipf's law by using lexicographic trees, and noticed that Zipf's law is "linguistically very shallow." But he still tries to link Zipf's law with other scaling phenomena. This correspondences provides a much intuitive derivation and emphasizes that Zipf's law does not share the common ground with other scaling behaviors.

ACKNOWLEDGMENT

The author would like to thank S. Burkie, M. Gell-Mann, and J. Bryngelson for helpful discussions.

REFERENCES

[1] G. Zipf, *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge, MA: MIT Press, 1932.
[2] ——, *Human Behavior and the Principle of Least-Effort*. Reading, MA: Addison-Wesley, 1965.
[3] ——, *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Cambridge MA: MIT Press, 1965).
[4] H. Kučera and W. Nelson Francis, *Computational Analysis of Present-Day American English*. Providence, RI: Brown Univ., 1967.
[5] B. Mandelbrot, *The Fractal Geometry of Nature*. New York: Freeman, 1982.
[6] ——, *Fractuals: Form, Chance and Dimension*. New York: Freeman, 1977.
[7] ——, *Les objects Fractal: Forme, Hasard et Dimension*. Flammarion, 1975.
[8] G. Miller, "Introduction," in *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press, 1965.
[9] B. Mandelbrot, "An informational theory of the statistical structure of language," in *Communication Theory*, W. Jackson, Ed. Betterworths, 1953.
[10] H. A. Raimi, "The peculiar distribution of first digits," *Scientific Amer.*, vol. 221, no. 6, pp. 109–115, Dec. 1969.
[11] M. Gell-Mann, unpublished notes.
[12] P. Manneville, "Intermittency, self-similarity and $1/f$ spectrum in dissipative dynamical systems," *Le Journal De Physique*, vol. 41, no. 11, pp. 1235–1243, 1980.
[13] W. Li, "Mutual information functions of natural language texts," Sante Fe Inst. preprint 89–009, 1989.

A Branching Process Analysis of the Stack Algorithm for Variable Channel Conditions

Marie-José Montpetit, *Member, IEEE,* David Haccoun, *Senior Member, IEEE*, and Gilles Deslauriers

***Abstract*—A branching process analysis in random environment is presented for bounding the average number of computations of sequential decoding over a finite state channel. Closed-form expressions applicable to specific cases are derived and evaluated. These unique bounds substantially reduce the need for lengthy simulations.**

***Index Terms*—Stack algorithm, variable channel, sequential decoding, branching processes, random environments.**

I. INTRODUCTION

It is well known that the number of computations necessary to decode one bit in sequential decoding has, asymptotically, a Pareto distribution. In spite of this variability, bounds on the average decoding effort of the stack algorithm over a binary symmetrical channel (BSC) have been found by using a multitype branching process model of the dynamics of the decoder [1]. However, over a channel with memory, a straightforward application of this analysis fails to provide good results. We now propose to model sequential decoding over a finite state channel (FSC) with branching processes in random environments (BPRE). By doing this, we establish the basis for a broadened examination of sequential decoding under variable conditions and we show that the stack algorithm is much more robust than expected under hostile conditions, provided we adapt it to these varying conditions. Our analysis differs from the one over the BSC [1], as we distinguish between generations in the branching process, since it is not homogeneous, and we use approximation matrices based on the varying channel statistics, to obtain closed-form expressions of the bounds. Although the proposed theoretical bounds and the simulated average number of computations of the stack algorithm are shown to generally agree, it can be said that the bounds found over the FSC are not as tight as those previously found over the BSC [1] for which more exact models can be used.

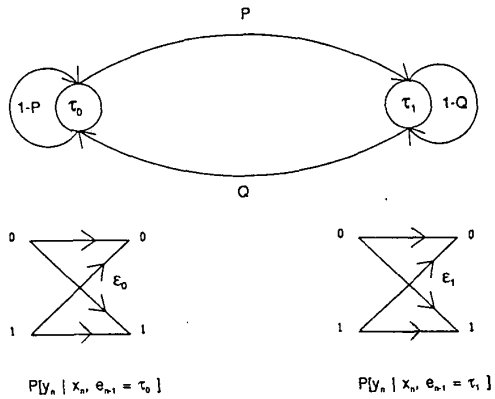

Fig. 1. Finite state channel.

Manuscript received March 6, 1990; revised December 9, 1991. This work was presented in part at the IEEE International Symposium on Information Theory, Ann Arbor, MI, October 1986.

The authors are with Ecole Polytechnique de Montréal, P.O. Box 6079, Station "A", Montréal, PQ, H3C 3A7, Canada.

IEEE Log Number 9201225.

II. BOUNDS ON THE AVERAGE NUMBER OF COMPUTATIONS

As previously mentioned, we are using the FSC, a channel which has two distinct states known as τ_0 and τ_1. As shown in Fig. 1, state τ_0 is a BSC with transition probability ϵ_0, and state τ_1 is a BSC with transition probability ϵ_1 where $\epsilon_1 \gg \epsilon_0$. P and Q are the transition probabilities between τ_0 and τ_1 which constitute the two states of a Markov chain with $P \ll Q \ll 1$. These probability assignments assure that both states will be persistent. Furthermore, most of the time the channel is in state τ_0. This simulates the behavior of a realistic bursty channel.

In sequential decoding, the Fano metric depends on the channel probability assignments and thus, in each state τ_i, the possible set of branch metrics are $+a_{0i}$, $-a_{1i}$, and $-a_{2i}$, $i = 0, 1$, for a coding rate of $1/2$. Since all channel states are persistent, we can consider the accumulated metric to be the sum of the individual branch metrics. On average, the total accumulated metric increases on the correct path. However, the metric falls rapidly in the presence of bursts of channel noise. Hence, we define the metric dip D_k as

$$D_k = \mu_k - \min_{i \geq k} \mu_i, \qquad \text{for } k = 0, 1, 2 \cdots . \tag{1}$$

D_k is the difference between the accumulated metric μ_k of a node on the correct path at depth k and its smallest succeeding value. When $D_k = 0$, node k is called a "breakout node" and will be decoded by a single computation. When $D_k > 0$, node k is nonbreakout and becomes the root of a tree of incorrect