International Journal of Mathematics and Mathematical Sciences 2017 vol.2017

Calculation of Precise Constants in a Probability Model of Zipf's Law Generation and Asymptotics of Sums of Multinomial Coefficients

Bochkarev V., Lerner E.

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2017 Vladimir Bochkarev and Eduard Lerner. Let ω 0 ,ω 1 ,⋯,ω n be a full set of outcomes (symbols) and let positive p i , i=0,⋯,n, be their probabilities (∑ i=0 n p i =1). Let us treat ω 0 as a stop symbol; it can occur in sequences of symbols (we call them words) only once, at the very end. The probability of a word is defined as the product of probabilities of its symbols. We consider the list of all possible words sorted in the nonincreasing order of their probabilities. Let pr be the probability of the rth word in this list. We prove that if at least one of the ratios log p i /log p j , i,j ∈ {1,⋯,n}, is irrational, then the limit lim r→∞ pr/r -1/γ exists and differs from zero; here γ is the root of the equation ∑ i=1 n p i γ =1. The limit constant can be expressed (rather easily) in terms of the entropy of the distribution (p 1 γ ,⋯,p n γ).

http://dx.doi.org/10.1155/2017/9143747

References

[1] R. Durrett, Random Graph Dynamics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2007.

[2] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," Internet Mathematics, vol. 1, no. 2, pp. 226-251, 2004.

[3] R. H. Baayen, Word Frequency Distributions, vol. 18 of Text, Speech and Language Technology, Kluwer Academic, Dordrecht, Netherlands, 2001.

[4] G. K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge, Mass, USA, 1949.

[5] B. Mandelbrot, "An informational theory of the statistical structure of languages," in Communication Theory, W. B. Jackson, Ed., pp. 486-502, 1953.

[6] B. Mandelbrot, "On recurrent noise limiting coding, in Proceedings of the symposium on information networks," in Proceedings of the Symposiumon Information Networks, pp. 205-221, Polytechnic Institute of Brooklyn, New York, NY, USA, April 1954.

[7] G. A. Miller, "Some effects of intermittent silence," The American Journal of Psychology, vol. 70, no. 2, pp. 311-314, 1957.

[8] W. Li, "Random texts exhibit Zipf 's-law-like word frequency distribution," IEEE Transactions on Information Theory, vol. 38, no. 6, pp. 1842-1845, 1992.

[9] R. Perline and R. Perline, "Two universality properties associated with the monkey model of Zipf 's law," Entropy, vol. 18, no. 3, article 89, 2016.

[10] B. Conrad and M. Mitzenmacher, "Power laws for monkeys typing randomly: the case of unequal probabilities," IEEE Transactions on Information Theory, vol. 50, no. 7, pp. 1403-1414, 2004.

[11] V. V. Bochkarev and E. Yu. Lerner, "The Zipf law for random texts with unequal letter probabilities and the Pascal pyramid," Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika, vol. 56, no. 12, pp. 30-33, 2012.

[12] R. Edwards, E. Foxall, and T. J. Perkins, "Scaling properties of paths on graphs," Electronic Journal of Linear Algebra, vol. 23, pp. 966-988, 2012.

[13] V. V. Bochkarev and E. Yu. Lerner, "Strong power and subexponential laws for an ordered list of trajectories of aMarkov chain," Electronic Journal of Linear Algebra, vol. 27, pp. 534-556, 2014.

[14] V. V. Bochkarev and E. Yu. Lerner, "Zipf and non-Zipf laws for homogeneous Markov chain," https://arxiv.org/abs/1207.1872.

[15] E. Artin, The Gamma Function, Translated by Michael Butler. Athena Series: Selected Topics in Mathematics, Holt, Rinehart and Winston, New York, NY, USA, 1964.

[16] Yu. Suhov and M. Kelbert, Probability and statistics by example. II, Markov Chains: A Primer in Random Processes and Their Applications, Cambridge University Press, Cambridge, UK, 2008.

[17] L. Kuipers and H. Niederreiter, Uniform Distribution of Sequences, Pure and Applied Mathematics, Wiley-Interscience [John Wiley & Sons], New York, NY, USA, 1974.

[18] R. Sedgewick and P. Flajolet, An Introduction to the Analysis of Algorithms, Addison-Wesley, Boston, Mass, USA, 1995.

[19] P. Flajolet, M. Roux, and B. Vallee, "Digital trees and memoryless sources: from arithmetics to analysis," in Proceedings of the 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA '10), Vienna, Austria, 2010.