В. В. Бочкарев, **Э. Ю. Лернер** (Казань, К(П)ФУ)**. Закон Ципфа для случайных текстов с неравными вероятностями букв.**

Известно, что наиболее часто встречается в английском тексте слово «the», второе по частоте встречаемости — «of», таким образом, каждой словоформе w текста сопоставляется ее *ранг* $r(w)$ — номер в частотном списке. *Частота* слова $f(w)$ определяется как отношение количества слов w в тексте к длине этого текста. Закон Ципфа (открытый в первой половине 20-го века) утверждает, что произведение $r(w)$ на $f(w)$ есть константа, примерно равная $0,1$ для английского текста. В настоящее время благодаря Google Labs доступны данные по результатам распознавания 5% английских книг. Согласно нашим расчетам, прямая МНК, построенная по прологарифмированным (для удобства мы рассматривали десятичный логарифм) значениям r и f ста наиболее часто встречаемых слов в 2000 году, имеет вид $\lg f = -1,05182 - 1,00026 \lg r$. Таким образом, закон Ципфа выполняется с поразительной точностью.

Для объяснения закона Ципфа привлекались различные соображения, одно из популярных — модель обезьяны, каждый раз независимо либо с вероятностью p_0 нажимающей на пробел, либо равновероятно на одну из 26 английских букв [1, 2]. *Словом* называется последовательность букв между двумя пробелами. Очевидно, что все слова одинаковой длины получаются с одной и той же вероятностью, их ранги идут подряд. Обозначим $p(r)$ вероятность получения слова ранга r. Легко видеть, что $\exists\ c_1, c_2\colon\ c_1 \leqslant \ln p(r) - \alpha \ln r \leqslant c_2$, где в случае $p_0 = 1/27$ имеем $\alpha = \ln 26 / \ln 27$. На самом деле, частоты появления различных букв в тексте неодинаковы, для русского языка они приблизительно задаются законом, замеченным С. Гусейн-Заде [3] (вероятности пропорциональны средним значениями порядковых статистик показательного распределения). Обобщение модели обезьяны на случай неравновероятных букв было сделано относительно недавно в [4].

В докладе мы представим короткое доказательство степенного закона в модели обезьяны с неравновероятными буквами и опишем комбинаторный смысл закона Ципфа в этом случае.

Теорема. *Пусть вероятности нажатия букв равны* $p_1, p_2, \ldots, p_n$ $(\sum_i p_i = 1 - p_0)$, *а* γ *есть корень уравнения* $\sum_i p_i^\gamma = 1$. *Тогда* $\exists\ c_1, c_2\colon\ c_1 \leqslant \ln p(r) - \gamma \ln r \leqslant c_2$.

В доказательстве используется эквивалентность формулировки теоремы комбинаторному факту ограниченности разности $\ln Q(x) - \gamma x$, где $Q(x) = \sum_{k \geqslant 0\colon\ L_1 k_1 + \cdots + L_n k_n \leqslant x} (k_1 + k_2 + \cdots + k_n)!/(k_1!\, k_2! \cdots k_n!)$ (сумма по части пирамиды Паскаля), $L_i = -\ln p_i$. Важную роль играет функциональное уравнение для $Q(x)$.

Авторы благодарят В. Д. Соловьева за многократные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

1. *Mandelbrot B.* An informational theory of the statistical structure of languages. — In:,Communication Theory. / W. Jackson, eds. Betterworth, 1953, p. 486–502.
2. *Miller G. A.* Some effects of intermittent silence. — Amer. J. Psychology, 1957, v. 70, p. 311–314.
3. *Гусейн-Заде С. М.* О распределении букв русского языка по частоте встречаемости. — Пробл. передачи информ., 1989, т. 24, № 4, с. 102–107.
4. *Conrad B., Mitzenmacher M.* Power Laws for Monkeys Typing Randomly: The Case of Unequal Probabilities. — IEEE Transac., 2004, v. 50, p. 1403–1414.