A Note on the Distribution of the Number of Prime Factors of the Integers ⋆

Aravind Srinivasan [1]

Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742.

Abstract

The Chernoff-Hoeffding bounds are fundamental probabilistic tools. An elementary approach is presented to obtain a Chernoff-type upper-tail bound for the number of prime factors of a random integer in $\{1, 2, \ldots, n\}$. The method illustrates tail bounds in negatively-correlated settings.

Key words: Chernoff bounds, probabilistic number theory, primes, tail bounds, dependent random variables, randomized algorithms

1 Introduction

Large-deviation bounds such as the Chernoff-Hoeffding bounds are of much use in randomized algorithms and probabilistic analysis. Hence, it is valuable to understand how such bounds can be extended to situations where the conditions of these bounds as presented in their standard versions, do not hold: the most important such condition is *independence*. Here, we show one such

⋆ An earlier version of this work appeared in the *Proc. Hawaii International Conference on Statistics, Mathematics, and Related Fields*, 2005

Email address: srin@cs.umd.edu (Aravind Srinivasan).

URL: http://www.cs.umd.edu/∼srin (Aravind Srinivasan).

[1] This research was done in parts at: (i) Cornell University, Ithaca, NY (supported in part by an IBM Graduate Fellowship), (ii) Institute for Advanced Study, Princeton, NJ (supported in part by grant 93-6-6 of the Alfred P. Sloan Foundation), (iii) DIMACS Center, Rutgers University, Piscataway, NJ (supported in part by NSF-STC91-19999 and by support from the N.J. Commission on Science and Technology), and (iv) the University of Maryland (supported in part by NSF Award CCR-0208005).

extension, to the classical problem of the distribution of the number of prime factors of integers.

For any positive integer N, let $\nu(N)$ denote the number of prime factors of N, ignoring multiplicities. Let $\ln x$ denote the natural logarithm of x, as usual. It is known that the average value of $\nu(i)$, for $i \in [n] = \{1, 2, \ldots, n\}$, is $\mu_n \doteq \ln\ln n + O(1) \pm O(\ln^{-2} n)$, for sufficiently large n. See also the discussion in Alon & Spencer [1]. We are interested in seeing if there is a "significant" fraction of integers $i \in [n]$ for which $\nu(i)$ deviates "largely" from μ_n.

Formally, Hardy & Ramanujan [4] showed that for any function $\omega(n)$ with $\lim_{n\to\infty} w(n) = \infty$,

$$\frac{|\{i \in [n] : \nu(i) \geq \ln\ln n + \omega(n)\sqrt{\ln\ln n}\}|}{n} = o(1), \tag{1}$$

where the "$o(1)$" term goes to zero as n increases. Their proof was fairly complicated. Turán [9] gave a very elegant and short proof of this result; his proof is as follows. Let $E[Z]$ and $Var[Z]$ denote the expected value and variance of random variable Z, respectively. Define $\mathcal{P}_n$ to be the set of primes in $[n]$. For a randomly picked $x \in [n]$, define, for every prime $p \in \mathcal{P}_n$, X_p to be 1 if p divides x, and 0 otherwise. Clearly,

$$\nu(x) = \sum_{p\in\mathcal{P}_n} X_p.$$

Hence,

$$\mu_n = E[\nu(x)] = \sum_{p\in\mathcal{P}_n} E[X_p] = \sum_{p\in\mathcal{P}_n} \frac{\lfloor n/p \rfloor}{n}$$

and thus,

$$\mu_n \sim \mu'_n \doteq \sum_{p\in\mathcal{P}_n} \frac{1}{p} = \ln\ln n + 0.261\ldots \pm O(\ln^{-2} n),$$

where the last equality follows from Mertens' theorem (see, for instance, Rosser & Schoenfeld [7]). By Chebyshev's inequality,

$$Pr(|\nu(x) - \mu(n)| \geq \lambda) \leq \frac{Var[\nu(x)]}{\lambda^2}$$

and by obtaining good upper bounds on the variances $Var[X_p]$ and the co-

variances $\mathrm{Cov}[X_p, X_q]$, Turán obtains his result that

$$Pr(|\nu(x) - \mu_n| \geq \lambda) \leq O\left(\frac{\ln\ln n}{\lambda^2}\right), \tag{2}$$

which, in particular, implies (1). Erdős & Kac [3] show that as $n \to \infty$, the tail of $\nu(x)$ (and of any function from a fairly broad class of functions of x) approaches that of the corresponding normal distribution, *i.e.,* that if ω is real and if $K_n = |\{i \in [n] : \nu(i) \leq \ln\ln n + \omega\sqrt{2\ln\ln n}\}|$, then

$$\lim_{n\to\infty} \frac{K_n}{n} = \pi^{-1/2} \int_{-\infty}^{w} e^{-u^2} du. \tag{3}$$

We strengthen the "upper-tail" part of (2) by showing that for any n and any parameter $\delta > 0$,

$$Pr(\nu(x) \geq \mu_n(1+\delta)) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu'_n}.$$

In contrast with (3), we get a bound for every n; thus, for instance, we get a concrete bound for deviations that are of an order of magnitude more than the standard deviation. We point out that strong upper- and lower-tail bounds are known using non-probabilistic methods [6]. The goal of this note is to show that a simple probabilistic approach suffices to derive exponential upper-tail bounds here. We also hope that the method and result may be of pedagogic use in showing the strength of probabilistic methods, and in the study of tail bounds for (negatively) correlated random variables.

2 Large Deviation Bounds

We first quickly review some salient features of the work of Schmidt, Siegel & Srinivasan [8].

2.1 Chernoff-Hoeffding type bounds in non-independent scenarios

The basic idea used in the Chernoff–Hoeffding (henceforth CH) bounds is as follows [2,5]. Given n random variables (henceforth "r.v."s) $X_1, X_2, \ldots, X_n$,

we want to upper bound the upper tail probability $Pr(X \geq a)$, where $X \doteq \sum_{i=1}^{n} X_i$, $\mu \doteq E[X]$, $a = \mu(1+\delta)$ and $\delta > 0$. For any fixed $t > 0$,

$$Pr(X \geq a) = Pr(e^{tX} \geq e^{at}) \leq \frac{E[e^{tX}]}{e^{at}};$$

by computing an upper bound $u(t)$ on $E[e^{tX}]$ and minimizing $\frac{u(t)}{e^{at}}$ over $t > 0$, we can upper bound $Pr(X \geq a)$. Suppose $X_i \in \{0, 1\}$ for each i, a commonly occuring case. In this case, a commonly used such bound is

$$Pr(X \geq \mu(1+\delta)) \leq F(\mu, \delta) \doteq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu} \tag{4}$$

(see, for example, [1]).

One basic idea of [8] when $X_i \in \{0, 1\}$ is as follows. Suppose we define, for $z = (z_1, z_2, \ldots, z_n) \in \Re^n$, a family of functions $S_j(z), j = 0, 1, \ldots, n$, where $S_0(z) \equiv 1$, and for $1 \leq j \leq n$,

$$S_j(z) \doteq \sum_{1 \leq i_1 < i_2 \cdots < i_j \leq n} z_{i_1} z_{i_2} \cdots z_{i_j}.$$

Then, for any $t > 0$, there exist *non-negative* reals $a_0, a_1, \ldots, a_n$ such that $e^{tX} \equiv \sum_{i=0}^{n} a_i S_i(X_1, X_2, \ldots, X_n)$. So, we may consider functions of the form

$$\sum_{i=0}^{n} y_i S_i(X_1, X_2, \ldots, X_n)$$

where $y_0, y_1, \ldots, y_n \geq 0$, instead of restricting ourselves to those of the form e^{tX}, for some $t > 0$. For any $y = (y_0, y_1, \ldots, y_n) \in \Re_+^{n+1}$, define $f_y(X_1, X_2, \ldots, X_n) \doteq \sum_{i=0}^{n} y_i S_i(X_1, X_2, \ldots, X_n)$. Then, it is easy to see that

$$Pr(X \geq a) = Pr\left(f_y(X_1, \ldots, X_n) \geq \sum_{i=0}^{a} y_i \binom{a}{i} \right) \leq \frac{E[f_y(X_1, \ldots, X_n)]}{\sum_{i=0}^{a} y_i \binom{a}{i}}.$$

So, the goal now is to minimize this upper bound over $(y_0, y_1, \ldots y_n) \in \Re_+^{n+1}$. Assuming that the X_i's are independent, it is shown in [8] that the optimum for the upper tail occurs roughly when: $y_i = 1$ if $i = \lceil \mu\delta \rceil$, and $y_i = 0$ otherwise. We can summarize this discussion by

Theorem 2.1 *([8]) Let bits $X_1, X_2, \ldots X_n$ be random with $X = \sum_i X_i$, and*

let $\mu = E[X]$, $k = \lceil \mu\delta \rceil$. *Then for any* $\delta > 0$,

$$Pr(X \geq \mu(1+\delta)) \leq \frac{E[S_k(X_1, X_2, \ldots, X_n)]}{\binom{\mu(1+\delta)}{k}}.$$

If the X_i*'s are independent, then this is at most*

$$\left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu.$$

2.2 Tail Bounds for $\nu(x)$

Returning to our original scenario, let n be our given integer. For a randomly picked $x \in [n]$, let X_p be 1 if p divides x, and 0 otherwise. As stated earlier,

$$\nu(x) = \sum_{p \in \mathcal{P}_n} X_p.$$

Let $\{\hat{X}_p \mid p \in \mathcal{P}_n\}$ be a set of *independent* binary random variables with $Pr(X_p = 1) = 1/p$. For any r and any set of primes $p_{i_1}, p_{i_2}, \ldots, p_{i_r}$, note that

$$\begin{aligned} E[\prod_{j=1}^{r} X_{p_{i_j}}] &= Pr(\bigwedge_{j=1}^{r} (p_{i_j} | x)) \\ &= \frac{\lceil n / \prod_{j=1}^{r} p_{i_j} \rceil}{n} \\ &\leq \frac{1}{\prod_{j=1}^{r} p_{i_j}} \\ &= E[\prod_{j=1}^{r} \hat{X}_{p_{i_j}}]. \end{aligned} \tag{5}$$

Thus we get

Theorem 2.2 *For any* $n \geq 2$ *and for any* $\delta > 0$,

$$Pr(\nu(x) \geq \mu_n(1+\delta)) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^{\mu'_n},$$

just by invoking Theorem 2.1.

3 Variants

Why does our approach not work directly for the lower tail of $\nu(x)$ also? The reason is that a direct negative-correlation result analogous to (5) does not appear to hold. It would be interesting to see if good lower-tail bounds also can be obtained by a short proof; as in [3,9], it may be possible to make quantitative use of the fact that the $\{X_p\}$ are all "almost independent". It is known that counting the prime divisors *including* multiplicity changes the functions a little [6], and it would be worth considering short (probabilistic) proofs for the tail behavior of this function also. More generally, can we concretely exploit the "near-independence" properties of additive number-theoretic functions [3]?

Acknowledgments. I thank Noga Alon, Eric Bach, Carl Pomerance, Christian Scheideler and Joel Spencer for valuable discussions & suggestions.

References

[1] N. Alon and J. Spencer. *The Probabilistic Method, Second Edition*. John Wiley & Sons, Inc., 2000.

[2] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, **23**, 493–509 (1952).

[3] P. Erdős and M. Kac. The Gaussian law of errors in the theory of additive number theoretic functions. American Journal of Mathematics, **62**, 738–742 (1940).

[4] G. H. Hardy and S. Ramanujan. The normal number of prime factors of a number n. Quarterly J. Math., **48**, 76–92 (1917).

[5] W. Hoeffding. Probability inequalities for sums of bounded random variables. American Statistical Association Journal, **58**, 13–30 (1963).

[6] C. Pomerance. On the distribution of round numbers. Number Theory Proceedings, Ootacamund, India, 1984. K. Alladi, ed., Lecture Notes in Math. **1122**, 173–200 (1985).

[7] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. Illinois Journal of Mathematics, **6**, 64–94 (1962).

[8] J. P. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-Hoeffding bounds for applications with limited independence. SIAM Journal on Discrete Mathematics, **8**, 223–250 (1995).

[9] P. Turán. On a theorem of Hardy and Ramanujan. Journal of the London Mathematics Society, **9**, 274–276 (1934).