Counting Primes Using Entropy

Ioannis Kontoyiannis

Lecture given on Thursday, May 8 2008, at the 2008 IEEE Information Theory Workshop, Porto, Portugal

I. The Prime Number Theorem

Sometime before 300 BC someone showed that there are infinitely many prime numbers—we know this because a proof appears in Euclid's famous *Elements*. In modern notation, if we write $\pi(n)$ for the number of primes no greater than n, we can say that,

$$\pi(n) \to \infty, \quad \text{as } n \to \infty. \tag{1}$$

Here's a proof, based on the idea of an argument of Chaitin from 1979 [6]. Let N be a random integer distributed uniformly in $\{1, 2, \dots, n\}$, and write it in its unique prime factorization,

$$N = p_1^{X_1} \cdot p_2^{X_2} \cdot \cdots \cdot p_{\pi(n)}^{X_{\pi(n)}}, \tag{2}$$

where $p_1, p_2, \dots, p_{\pi(n)}$ are the primes up to n, and where each X_i is the largest power $k \geq 0$ such that p_i^k divides N. This defines a new collection of random variables $X_1, X_2, \dots, X_{\pi(n)}$, and, since $p_i^{X_i}$ divides N, we must have,

$$2^{X_i} \leq p_i^{X_i} \leq N \leq n,$$

or, writing log for $\log_2$,

$$X_i \leq \log n, \quad \text{for each } i. \tag{3}$$

Now here's a cool thing:

$$\begin{aligned}\log n &= H(N)\\ &= H(X_1, X_2, \dots, X_{\pi(n)})\\ &\leq H(X_1) + H(X_2) + \cdots + H(X_{\pi(n)})\\ &\leq \pi(n)\log(\log n + 1).\end{aligned} \tag{4}$$

The second equality comes from the uniqueness of prime factorization, that is, knowing N is the same as knowing the values of all the X_i; the last inequality comes from (3). Therefore,

$$\pi(n) \geq \frac{\log n}{\log(\log n + 1)}, \quad \text{for all } n \geq 2,$$

which not only proves that $\pi(n) \to \infty$, but also gives a lower bound on how *fast* it grows with n.

This is a tiny glimpse into a very, very long story: A large portion of number theory—and a very significant portion of modern mathematics at large—is devoted to quantifying (1). For a long time we've wanted to know:

How fast, exactly, does $\pi(n) \to \infty$, as n grows?

Enter Gauss. According to Apostol [1], in 1792, while inspecting tables of prime numbers, Gauss conjectured what has come to be known as the celebrated *prime number theorem*, namely that,

$$\pi(n) \sim \frac{n}{\log_e n}, \quad \text{as } n \to \infty, \tag{5}$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$. Apparently he was not able to prove it, and not because he was only 15 years old at the time—he kept trying, without success, for quite a while, and only disclosed his conjecture in a mathematical letter to Encke, over 50 years later.

In fact Gauss (still at 15) suggested that, for finite n, $\pi(n)$ is better approximated by the function,

$$\mathrm{Li}(n) = \int_2^n \frac{dt}{\log_e t},$$

sometimes called the *Eulerian logarithmic integral*. Since $\mathrm{Li}(n)$ asymptotically varies like $n/\log_e n$, the prime number theorem, henceforth PNT, can also be written,

$$\pi(n) \sim \mathrm{Li}(n), \quad \text{as } n \to \infty.$$

If you're not yet convinced that we should care all that much about how $\pi(n)$ behaves for large n, this should do it: Arguably the most important problem in mathematics today, the Riemann hypothesis, is equivalent to the following refined version of the PNT: For every $\epsilon > 0$,

$$\pi(n) = \mathrm{Li}(n) + O\left(n^{\frac{1}{2}+\epsilon}\right).$$

See [2] for more of the history and details.

II. Chebyshev's Attempt

The PNT was proved a little more than 100 years after Gauss conjectured it, but before talking about proofs (and attempted proofs), let's note that according to the PNT (5) our earlier estimate (3) was pretty loose. Can we do better?

Interestingly, a small modification of our basic argument in (4) gives a slightly better bound. Suppose that, instead of the usual prime factorization, we express N as,

$$N = M^2 \cdot p_1^{Y_1} \cdot p_2^{Y_2} \cdot \cdots \cdot p_{\pi(n)}^{Y_{\pi(n)}}, \tag{6}$$

where $M \geq 1$ is the largest integer such that M^2 divides N, and the Y_i are now binary. Since M^2 divides N, we must have

$M^2 \leq N \leq n$, or $M \leq \sqrt{n}$, and noting that the representation (6) is also unique, arguing as before we get,

$$\begin{aligned}\log n &= H(N)\\ &= H(M, Y_1, Y_2, \ldots, Y_{\pi(n)})\\ &\leq H(M) + H(Y_1) + H(Y_2) + \cdots + H(Y_{\pi(n)})\\ &\leq \frac{1}{2}\log n + \pi(n)\,,\end{aligned}$$

which implies that $\pi(n) \geq \frac{1}{2}\log n$, for all $n \geq 2$. This is better than (3) but still pretty far from the optimal rate in (5).

I don't know how (or if it is possible) to twist this argument around further to get more accurate estimates, so let's get back to the classical proofs of the PNT. Another early player in this drama is Chebyshev (the one of the inequality), who also gave the PNT a go and, although he didn't succeed in producing a complete proof, he discovered a number of beautiful results along the way. One of them is the following unexpected asymptotic formula:

Theorem 1. Chebyshev (1852) [7], [8]

As $n \to \infty$,

$$C(n) \triangleq \sum_{p \leq n} \frac{\log p}{p} \sim \log n,$$

where the sum is over all primes p not exceeding n.

Actually Chebyshev came pretty close to proving the PNT. For example, using Theorem 1 in a slightly refined form, he was able to find explicit constants constants $A < 1 < B$ and n_0 such that:

$$A\frac{n}{\log_e n} \leq \pi(n) \leq B\frac{n}{\log_e n}, \quad \text{for all} \;\; n \geq n_0.$$

The PNT was finally proved in 1896 by Hadamard and, independently and almost simultaneously, by de la Vallée-Pousin. Both proofs were mathematically "heavy," relying on the use of Hadamard's theory of integral functions applied to the Riemann zeta function $\zeta(s)$; see [2] for details. In fact, for quite some time it was believed that no elementary proof would ever be found, and G.H. Hardy in a famous lecture to the Mathematical Society of Copenhagen in 1921 [5] went as far as to suggest that "*if anyone produces an elementary proof of the PNT … he will show that … it is time for the books to be cast aside and for the theory to be rewritten.*"

The announcement by Selberg and Erdös in 1948 that they had actually found such an elementary proof came as a big surprise to the mathematical world and caused quite a sensation; see [10] for a survey. What's particularly interesting for us, is that Chebyshev's result in Theorem 1 was used explicitly in their proof.

Thus motivated, we now discuss an elegant way to prove Theorem 1 using only elementary ideas from information theory and basic probability.

III. Entropy

Apparently the first person to connect prime-counting questions with information-theoretic ideas and methods is Patrick Billingsley. In 1973 he was invited to deliver the prestigious "Wald Memorial Lectures" at the IMS Annual Meeting in New York. Billingsley, a probabilist, has long been involved with entropy and information—and wrote a book [3] about it—and in the years before these lectures it appears he had developed a strong interest in "probabilistic number theory," that is, in the application of probabilistic techniques to derive results in number theory. In the transcript [4] of his 1973 lectures he describes a beautiful heuristic argument for proving Theorem 1 using simple computations in terms of the entropy. It goes like this.

Start as before with a random integer N uniformly distributed between 1 and some fixed $n \geq 2$, and write it in its unique prime factorization (2). What is the distribution of the induced random variables X_i? Let's first look at one of them. Since the number of multiples of p_i^k between 1 and n is exactly $\lfloor n/p_i^k \rfloor$, we have,

$$\Pr\{X_i \geq k\} = \Pr\left\{N \text{ is a multiple of } \; p_i^k\right\} = \frac{1}{n}\left\lfloor \frac{n}{p_i^k} \right\rfloor. \qquad (7)$$

Therefore, for large n,

$$\Pr\{X_i \geq k\} \approx \left(\frac{1}{p_i}\right)^k,$$

i.e., the distribution of each X_i is approximately geometric with parameter $1/p_i$. Similarly, since the number of multiples of $p_i^k p_j^\ell$ between 1 and n is $\lfloor n/p_i^k p_j^\ell \rfloor$, for the joint distribution of X_i, X_j we find,

$$\Pr\{X_i \geq k,\; X_j \geq \ell\} = \frac{1}{n}\left\lfloor \frac{n}{p_i^k p_j^\ell} \right\rfloor \approx \left(\frac{1}{p_i}\right)^k \left(\frac{1}{p_j}\right)^\ell,$$

so X_i and X_j are approximately independent. The same argument works for any finite sub-collection of the $\{X_i\}$. This intuition, that we can think of the $\{X_i\}$ as approximately independent geometrics, was well known for at least a few decades before Billingsley's lectures; see, e.g., Kac's classic gem [11].

Billingsley's insight was to bring the entropy into play. Combining the initial steps of our basic argument (4) with the observation that the X_i are approximately independent geometrics,

$$\begin{aligned}\log n &= H(N)\\ &= H(X_1, X_2, \ldots, X_{\pi(n)})\\ &\approx \sum_{i=1}^{\pi(n)} H(X_i) && (8)\\ &\approx \sum_{p \leq n}\left[\frac{\log p}{p-1} - \log\left(1 - \frac{1}{p}\right)\right], && (9)\end{aligned}$$

where in the last step we simply substituted the well-known [9] formula for the entropy of a geometric with parameter $1/p$. And since for large p the summands in (9) behave like

$$\frac{\log p}{p} + O\left(\frac{1}{p}\right),$$

from (9) we get the heuristic estimate,

$$C(n) = \sum_{p \le n} \frac{\log p}{p} \approx \log n, \quad \text{for large } \ n.$$

It would certainly be nice to have an actual information-theoretic proof of Theorem 1 along those lines—Billingsley suggests so too—but the obvious strategy doesn't work, or at least I wasn't able to make it work. The problem is that the approximation of the distribution of the $\{X_i\}$ by independent geometrics is not accurate enough to turn the two "≈" steps in (8) and (9) into rigorous bounds. That's the bad news. But there's also good news.

IV. An Information Theoretic Proof

As it turns out, it *is* possible to give an elementary information-theoretic proof of Theorem 1, albeit using somewhat different arguments from Billingsley's. Here's the more-beautiful-half of the proof; for the other half see [12].

Proof that $C(n)$ is asymptotically $\ge \log n$. The starting point is again our basic argument in (4):

$$\log n = H(N) = H(X_1, X_2, \ldots, X_{\pi(n)}) \le \sum_{i=1}^{\pi(n)} H(X_i).$$

Since the distribution of an integer-valued random variable X with mean $\mu > 0$ is maximized by the entropy

$$h(\mu) \triangleq (\mu + 1)\log(\mu + 1) - \mu \log \mu$$

of a geometric with the same mean, if we write $\mu_i = E(X_i)$ for the mean of X_i, then,

$$\log n \le \sum_{i=1}^{\pi(n)} h(\mu_i).$$

But from the distribution of X_i as expressed in (7) it is easy to get some useful information about μ_i:

$$\mu_i = \sum_{k \ge 1} \Pr\{X_i \ge k\} \le \sum_{k \ge 1} \left(\frac{1}{p_i}\right)^k = \frac{1/p_i}{1 - 1/p_i}.$$

Therefore, since $h(\mu)$ is an increasing function, we obtain,

$$\begin{aligned}\log n &\le \sum_{i=1}^{n} h\left(\frac{1/p_i}{1 - 1/p_i}\right)\\ &= \sum_{p \le n} \left[\frac{\log p}{p - 1} - \log\left(1 - \frac{1}{p}\right)\right],\end{aligned} \tag{10}$$

and that's basically it.

Since the summands above behave like $\frac{\log p}{p}$ for large p, an easy exercise in elementary calculus gives,

$$\liminf_{n \to \infty} \frac{C(n)}{\log n} \ge 1, \tag{11}$$

as claimed. □

V. Epilogue

It is very satisfying that elementary information-theoretic tools can produce optimal asymptotic estimates in number theory, like the lower bound (11) corresponding to Chebyshev's Theorem 1. In fact, from the actual result we derived in (10) it's also easy to deduce finite-n refinements of this lower bound, like, e.g.,

$$C(n) \ge \frac{86}{125} \log n - 2.35, \quad \text{for all } \ n \ge 16.$$

Unfortunately, it is not clear how to reverse the inequalities in the above proof to get a corresponding upper bound on $C(n)$. Nevertheless, a different information-theoretic argument does work, and shows that,

$$\sum_{p \le n} \frac{\log p}{p} \le \log n + 2 \log 2,$$

for all $n \ge 2$; see [12].

Two final remarks before closing. First, although Biilingsley in [4] does not produce any information-theoretic proofs *per se*, he does go in the "opposite" direction: He uses probabilistic techniques and results about the primes to compute the entropy of several relevant collections of random variables.

And lastly, we mention that in Li and Vitányi's text [13], an elegant argument is given for a more accurate lower bound on $\pi(n)$ than those we saw above. Using ideas and results from algorithmic information theory, they show that $\pi(n)$ asymptotically grows at least as fast as $\frac{n}{(\log n)^2}$. The proof, which they attribute to unpublished work by P. Berman (1987) and J. Tromp (1990), is somewhat involved, and uses tools very different to those developed here.

References

[1] T.M. Apostol, *Introduction to Analytic Number Theory*. Springer-Verlag, New York, 1976.

[2] P.T. Bateman and H.G. Diamond, *A hundred years of prime numbers*. Amer. Math. Monthly, vol. 103, no. 9, pp. 729–741, 1996.

[3] P. Billingsley, *Ergodic theory and information*. John Wiley & Sons Inc., New York, 1965.

[4] P. Billingsley, "The probability theory of additive arithmetic functions," *Ann. Probab.*, vol. 2, pp. 749–791, 1974.

[5] H. Bohr, "Address of Professor Harold Bohr," In *Proceedings of the International Congress of Mathematicians (Cambridge, 1950)*, vol. 1, pages 127–134, Amer. Math. Soc., Providence, RI, 1952.

[6] G.J. Chaitin, "Toward a mathematical definition of "life"," In *Maximum entropy formalism (Conf., Mass. Inst. Tech., Cambridge, Mass., 1978)*, pages 477–498. MIT Press, Cambridge, Mass., 1979.

[7] P.L. Chebychev, "Mémoire sur les nombres premiers, *J. de Math. Pures Appl.*, vol. 17, pp. 366–390, 1852.

[8] P.L. Chebychev, "Sur la totalité des nombres premiers inférieurs à une limite donnée," *J. de Math. Pures Appl.*, vol. 17, pp. 341–365, 1852.

[9] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. J. Wiley, New York, 1991.

[10] H.G. Diamond, "Elementary methods in the study of the distribution of prime numbers," *Bull. Amer. Math. Soc. (N.S.)*, vol. 7, no. 3, pp. 553–589, 1982.

[11] M. Kac, *Statistical Independence in Probability, Analysis and Number Theory.* Published by the Mathematical Association of America. Distributed by John Wiley and Sons, Inc., New York, 1959.

[12] I. Kontoyiannis, Some information-theoretic computations related to the distribution of prime numbers. *Preprint, available online at*: http://aps.arxiv.org/abs/0710.4076, November 2007.

[13] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, second edition, 1997.

Latest Activities of the IT Student Committee

A. Yener, L. Sankar, I. Maric, B. Shrader

The Student Committee has been hard at work since we last reported our activities in the December issue. In addition to having and planning our usual conference activities, a couple of new and exciting initiatives are underway.

First, a few words about our most recent event at the Conference on Information Sciences and Systems (CISS) in Princeton, NJ are in order. On Thursday, March 20, we organized a research discussion round table event for all participating students. About 90 students attended the event held at the Friend Center convocation room of Princeton University. There were six research topics discussed and led by the student volunteers. The following is a list of research topics and team leaders.

1. "MIMO Channels," leader: Jimmy Chui, Princeton University.

2. "Network Coding," leader: Anna Pantelidou, University of Maryland, College Park (UMD-CP).

3. "Sparse Representations and Compressed Sensing," leader: Eugene Brevdo, Princeton University.

4. "Ad-hoc Networks," leader: Sharon Betz, Princeton University

5. "Network and Information Security," leaders: Lifeng Lai and Ruoheng Liu, Princeton University, and Prasanth Ananthapadmanabhan, UMD-CP.

Jimmy Chui and Sharon Betz hold cards identifying the topics discussed at their tables at the CISS 2008 event.

6. "Network Optimization," leaders: Chee Wei Tan, Princeton University, and Joydeep Acharya, WINLAB, Rutgers.

We thank Lalitha Sankar for coordinating the event, and the student volunteers in the above list for leading these lively discussions and helping serve the lunch boxes. Aside from the non-vegetarian sandwiches going a bit faster than we anticipated (but no one went hungry!), the event appeared to be successful. The meeting concluded with the advertisement of the First Annual School of Information Theory by Gerhard Kramer who was our guest at the event. More details on the event and papers discussed can be found on the stu-