Introduction

Probabilistic Concepts
- Random Variables
- Indicator Function
- Notion of Probability
- Expected Value
- Variance
- Chebyshev's Inequality

Prime Divisors of n
- Estimation of $E(\#B)$
- Estimation of $V(\#B)$
- Conclusion
- References

Probabilistic Methods in the Analysis of the Distribution of Prime Numbers

José Fonseca

2nd year student at the Faculty of Engineering of the University of Porto

20th of July, 2013 - Gulbenkian Foundation Headquarters

Introduction

This small work aims to present an analysis of the distribution of prime number divisors of integers using rather elementary probabilistic tools, such as the *expected value* and *variance*.

I will present a result concerning the number of prime divisors of a natural number, using the above mentioned tools, which states that n has approximately $\log \log n$ prime divisors, with high probability.

This result was discovered, and proved, by Hardy and Ramanujan in the 1920s. Here we will give an overview of a simpler proof found by the hungarian mathematician Paul Turán, in 1934.

To do so, let us start by defining the probabilistic concepts needed for our goal.

Random Variables

Let Ω be some finite event space:

$$\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\}.$$

Then, a *random variable* is a real valued function $X(\omega_i)$ defined as

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega_i & \longrightarrow & X(\omega_i) \end{array}$$

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

Indicator Function

An important function that we will use in our discussion is the *indicator function*. It is defined as

$$I(\omega_i) = \begin{cases} 1 & \text{if } \omega_i \text{ occurs,} \\ 0 & \text{if } \omega_i \text{ does not occur.} \end{cases}$$

The cardinality of a certain subset B of a given set $A = \{a_1, a_2, ..., a_n\}$ can then be expressed by:

$$\#B = \sum_{i=1}^{n} I(a_i \in B)$$

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

Laplacian Probability

There are several definitions of probability. One of them is the traditional *Laplacian* definition, proposed by the french mathematician Pièrre Simon Laplace (1749 – 1827).

According to it, if E is the set of desired events, and if we denote by $P(E)$ the probability function of E, then

$$P(E) = \frac{\#E}{\#\Omega}$$

Axiomatic Theory of Probability

Other important definition of probability follows an axiomatic approach, which focuses on presenting the probability as a function that satisfies a certain set of *a priori* axioms, namely

I) The probability function is always a positive real number:

$$P(E) \geq 0, \quad \text{with } P(E) \in \mathbb{R}, \ \forall E \in \Omega.$$

II) The probability of the occurence of the set of all events, Ω, is equal to one:

$$P(\Omega) = 1.$$

III) Given a sequence of mutually exclusive events, we have that:

$$P(E_1 \cup E_2 \cup E_3 \cup ... \cup E_n) = \sum_{i=1}^{n} P(E_i), \quad \text{with } E_i \subseteq \Omega.$$

From the given axioms, one can derive:

- $0 \leq P(E) \leq 1, \ \forall E \in \Omega;$
- $P(\overline{E}) = 1 - P(E);$
- $P(\emptyset) = 0.$

Expected Value

Now that we have reviewed the basic concepts of probability, let us finish this section by presenting two cornerstone tools of probabilistic analysis. The first is the *expected value*.

The *expected value* of a given random variable X, on some finite event space Ω, is given by

$$E(X) = \sum_{i=1}^{n} X(\omega_i)P(\omega_i).$$

Furthermore, the expected value is a linear function: if Z is a random variable which is a linear combination of two other random variables, i.e. $Z = aX + bY$, then

$$E(aX + bY) = aE(X) + bE(X).$$

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

Variance

The expected value gives an idea of the 'centrality' of the values of X. For instance, if X follows a random uniform distribution, then $E(X)$ is the actual arithmetic average of the values of X.

If one, instead, is looking for an estimate of how far are the individual values of X of the central value estimate $E(X)$, one uses the *variance*, $V(X)$, which is defined as

$$V(X) = E(X - E(X))^2,$$

and is equivalent to

$$V(X) = E(X^2) - E(X)^2.$$

Variance

If one has a linear combination of pairwise independent random variables $X = X_1 + X_2 + \cdots + X_n$, then $V(X)$ is just the sum of the $V(X_i)$. Otherwise, one needs to add a sum relative to the so called *covariance*:

$$V(X) = \sum_{i=1}^{n} V(X_i) + \sum_{i,j \in [1,n]: i \neq j} Cov(X_i, X_j),$$

where the covariance is defined by

$$\begin{aligned} Cov(X_i, X_j) &= E\left((X_i - E(X_i))(X_j - E(X_j)\right) \\ &= E(X_i X_j) - E(X_i)E(X_j). \end{aligned}$$

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

Chebyshev's Inequality

From Markov's inequality we know that

$$P(X \geq a) \leq \frac{E(X)}{a}$$

holds. Now, if we make $Y = (X - E(X))^2$, with $a = k^2 V(X)$, we get

$$P[(X - E(X))^2 \geq k^2 V(X)] \leq \frac{E[X - E(X)]^2}{k^2 V(X)},$$

which simplifies to

$$P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$$

where $\sigma = \sqrt{V(X)}$ is the *standard deviation*.

Introduction

Probabilistic Concepts
Random Variables
Indicator Function
Notion of Probability
Expected Value
Variance
Chebyshev's Inequality

Prime Divisors of n
Estimation of $E(\#B)$
Estimation of $V(\#B)$
Conclusion
References

Chebyshev's Inequality

It follows that

$$P(|X - E(X)| < k\sigma) > \frac{1}{k^2}$$

and, hence, one sees that, for a big k, the event

$$|X - E(X)| < k\sigma$$

is very likely, and using the usual assimptotic notation one has

$$X = E(X) + O(\sigma).$$

Introduction

Probabilistic Concepts
Random Variables
Indicator Function
Notion of Probability
Expected Value
Variance
Chebyshev's Inequality

Prime Divisors of n
Estimation of $E(\#B)$
Estimation of $V(\#B)$
Conclusion
References

Prime Divisors of n

Now that we possess the probabilistic tools needed, let us analyse the problem we want to solve.

Let

$$\nu(n) = \sum_{p \leq n} I(p|n)$$

be the function that gives the number of prime divisors of a natural number n. How are we going to compute it? Isn't it more reasonable to present an approximate result? If so, how are we going to do it?

Probabilistic Methods in the Analysis of the Distribution of Prime Numbers

José Fonseca

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

Prime Divisors of n

We are now ready to give a precise statement of the result of Hardy and Ramanujan mentioned above:

Let $\omega(n)$ be a function that tends to infinity in an arbitrarily slow manner. Then,

$$\#\left\{x \in [1, n] : |\nu(x) - \log\log x| > \omega(n)\sqrt{\log\log n}\right\} = o(n).$$

This theorem implies that $\nu(x) = \log\log(x) + O(\sqrt{\log\log(x)})$, with high probability, as claimed at the beginning.

Prime Divisors of n

What one has to show is that, if x is chosen uniformly at random in $\{1, 2, 3, \ldots, n\}$, then

$$P\left[|\nu(x) - \log\log(x)| > \omega(n)\sqrt{\log\log(n)}\right] = o(1).$$

Instead of $\nu(x)$, it is convenient to use $\#B$, where B given by

$$B = \{p \in [1, x] : p \text{ prime }, p \leq n^{\frac{1}{10}}, p \mid x\},$$

As $\#B \leq \nu(x) \leq \#B + 10$ and $\log\log(x) = \log\log(n) + O(1)$ with probability $1 - o(1)$, it is then enough to show that

$$P\left[|\#B - \log\log(n)| > \omega(n)\sqrt{\log\log(n)}\right] = o(1).$$

Now, this follows from Chebyshev's inequality after one shows that

$$E(\#B), V(\#B) = \log\log(n) + O(1).$$

Estimation of $E(\#B)$

Since $\#B = \sum_{b \in B} I(b \in B)$ and the expectation is linear,

$$E(\#B) = \sum_{p \leq n^{\frac{1}{10}}} P(p \mid x).$$

Using the Laplacian definition of probability for $P(p \mid x)$, and since, for $x \in [1, n]$, the number of possible events is n and the number of favourable events is $\left\lfloor \frac{n}{p} \right\rfloor = \frac{n}{p} + \theta$, for some $\theta \in [-1, 0]$, we get

$$P(p \mid x) = \frac{1}{p} + O\left(\frac{1}{n}\right).$$

Estimation of $E(\#B)$

We thus obtain

$$E(\#B) = \sum_{p \leq n^{\frac{1}{10}}} \left[\frac{1}{p} + O\left(\frac{1}{n}\right) \right].$$

From Merten's theorem, we know that

$$\sum_{p \leq n^{\frac{1}{10}}} \frac{1}{p} = \log\log(n) + O(1).$$

Thus, we are left with $\sum_{p \leq n^{\frac{1}{10}}} O\left(\frac{1}{n}\right)$ to compute.

Estimation of $E(\#B)$

From the very meaning of the assimptotic notation, we have that, for some constant k,

$$\sum_{p \leq n^{\frac{1}{10}}} \left| O\left(\frac{1}{n}\right) \right| \leq k \frac{1}{n} \sum_{p \leq n^{\frac{1}{10}}} 1.$$

But, from the Prime Number Theorem, we know that

$$\sum_{p \leq n^{\frac{1}{10}}} 1 \sim \frac{n^{\frac{1}{10}}}{\log(n^{\frac{1}{10}})},$$

from which we conclude that

$$\sum_{p \leq n^{\frac{1}{10}}} \left| O\left(\frac{1}{n}\right) \right| \leq \frac{k}{\log(n^{\frac{1}{10}})} n^{-\frac{9}{10}}.$$

Estimation of $E(\#B)$

Therefore:

$$\sum_{p \leq n^{\frac{1}{10}}} \left| O\left(\frac{1}{n}\right) \right| = O(n^{-\frac{9}{10}}).$$

So, we finally get

$$E(\#B) = \log\log(n) + O(1) + O(n^{-\frac{9}{10}}),$$

which for a large value of n is

$$E(\#B) \approx \log\log(n) + O(1),$$

like we intended to prove.

Estimation of $V(\#B)$

Let us now focus on the estimate for $V(\#B)$.

In the expression for the variance of a sum of random variables

$$V(X) = \sum_{i=1}^{n} V(X_i) + \sum_{i,j \in [1,n] : i \neq j} Cov(X_i, X_j),$$

the sum $\sum_{i=1}^{n} V(X_i)$ in the case $X = \#B$, becomes

$$V(I(p_1 \in B)) + V(I(p_2 \in B)) + \cdots + V(I(p_{\#B})).$$

Estimation of $V(\#B)$

Using $V(X) = E(X^2) - E(X)^2$, one gets

$$V(I(p_i \in B)) = E(I(p_i \in B)^2) - E(I(p_i \in B))^2 = P(p_i|x) - P(p_i|x)^2,$$

and therefore

$$V(\#B) = \sum_{p \leq n^{\frac{1}{10}}} P(p|x) - P(p|x)^2 + \sum_{p,q \leq n^{\frac{1}{10}} : p \neq q} Cov(I(p|x), I(q|x))$$

which is simply

$$V(\#B) = \sum_{i=1}^{n} \left(\frac{1}{p} - \frac{1}{p^2} \right) + \sum_{p,q \leq n^{\frac{1}{10}} : p \neq q} Cov(I(p|x), I(q|x))$$

For the first sum we already have an estimation by Merten's theorem. The second, for $\frac{1}{p^2}$, is convergent to $\frac{\pi^2}{6}$. Let us then focus on the covariance summation.

Estimation of $V(\#B)$

Applying the definition of covariance, we are led to

$$Cov(I(p|x), I(q|x)) = E(I(pq|x) - E(I(p|x))E(I(q|x)) = O\left(\frac{1}{n}\right).$$

The covariance summation is then

$$\begin{aligned}\sum_{p,q \le n^{\frac{1}{10}} : p \ne q} Cov(I(p|x), I(q|x)) &= \sum_{p \le n^{\frac{1}{10}}} \sum_{q \le n^{\frac{1}{10}}} Cov(I(p|x), I(q|x)) \\ &= \sum_{p \le n^{\frac{1}{10}}} \sum_{q \le n^{\frac{1}{10}}} O\left(\frac{1}{n}\right).\end{aligned}$$

Estimation of $V(\#B)$

Like we did with the expected value estimate, we have

$$\sum_{p \leq n^{\frac{1}{10}}} \sum_{q \leq n^{\frac{1}{10}}} O\left(\frac{1}{n}\right) \leq k\frac{1}{n} \sum_{p \leq n^{\frac{1}{10}}} \sum_{q \leq n^{\frac{1}{10}}} 1 = k\frac{1}{n}\left(\sum_{p \leq n^{\frac{1}{10}}} 1\right)^2$$

and applying the prime number theorem we are led to

$$\sum_{p \leq n^{\frac{1}{10}}} \sum_{q \leq n^{\frac{1}{10}}} Cov(I(p|x), I(q|x)) = O\left(n^{-\frac{8}{10}}\right).$$

Estimation of $V(\#B)$

Finally, applying all three results, the variance estimate, for large values of n yields

$$V(\#B) = \log\log(n) + O(1)$$

and the initial theorem is proven.

Conclusion

The final result is then that the number of prime divisors of a given number n is

$$\nu(n) = \log\log(n) + O\left(\sqrt{\log\log(n)}\right),$$

with higher and higher probability, as n goes to infinity.

Introduction

Probabilistic Concepts

Random Variables

Indicator Function

Notion of Probability

Expected Value

Variance

Chebyshev's Inequality

Prime Divisors of n

Estimation of $E(\#B)$

Estimation of $V(\#B)$

Conclusion

References

References

- GNEDENKO, Boris V., *Theory of Probability*, 6th edition, Moscow, Nauka, 1988.
- PAPOULIS, Athanasios, *Probability, Random Variables and Stochastic Processes*, 3rd edition, New York, McGraw-Hill, 1991.
- TAO, Terrence, VU, Van H., *Additive Combinatorics*, 1st edition, Cambridge, Cambridge Press, 2010