Sieve of Eratosthenes

University of Waterloo

Peiran Tao

1 Introduction

Sieves are used to bound the size of a set after elements with certain "undesirable" properties have been removed. A basic example of a sieve is the method of inclusion-exclusion which gives an exact count for the number of elements in a set.

Suppose we are given $A = [1, x] \cap \mathbb{Z}$, the set of integers $\leq x$. We want to find all prime numbers in A. The following lemma gives us a neat way to do it

Lemma 1.1. Let $N \in \mathbb{N}$ be a positive integer and $n \in \mathbb{Z}$ with $2 \leq n \leq N$. If n is composite, then there is a prime divisor $p \mid n$ such that $p \leq \sqrt{N}$.

Proof: Suppose all prime divisors are $> \sqrt{N}$. Since n is composite, it means n has at least two prime factors (counting multiplicities), say p and q. Then $pq \mid n$ so $pq \leq n$, but

$$pq > \sqrt{N}\sqrt{N} = N$$

contradiction. □

This lemma tells us, if we can remove all the multiples of the primes in $[1, [\sqrt{N}]]$ in A, then the elements that remain are prime numbers between $[\sqrt{N}] + 1$ and N. But how do we do this? Here is an example:

Example. How many integers in $S = [1, 40]$ are not divisible by $2, 3$ or 5?

Let $A = 2\mathbb{Z} \cap S$ and $B = 3\mathbb{Z} \cap S$ and $C = 5\mathbb{Z} \cap S$ be integers that ARE divisible by $2, 3, 5$ in S. We wish to determine the size of the set

$$P = S \setminus (A \cup B \cup C)$$

It suffices to determine the size of $A \cup B \cup C$. We can do this by the inclusion-exclusion

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap B| - |B \cap C| + |A \cap B \cap C|$$

The size of each individual set is easy to determine

$$\begin{aligned} |A| &= [40/2] = 20 \\ |B| &= [40/3] = 13 \\ |C| &= [40/5] = 8 \\ |A \cap B| &= [40/6] = 6 \\ |A \cap C| &= [40/10] = 4 \\ |B \cap C| &= [40/15] = 2 \\ |A \cap B \cap C| &= [40/30] = 1 \end{aligned}$$

Then, the number of integers ≤ 40 that are not divisible by $2, 3$ or 5 is

$$40 - (20 + 13 + 8 - 6 - 4 - 2 + 1) = 10$$

1 is contained in P, but it not a prime. Also, 2,3,5 are not contained in P. So the total number of primes in $[1, 40]$ is $10 - 1 + 3 = 12$.

This method can be generalized in following ways:

1. Instead of doing sieve on the set $[1, x] \cap \mathbb{Z}$, we can do it on an arbitrary set.
2. Instead of removing the multiples of primes that are $< \sqrt{N}$, we can set pick a $z > 0$ and remove all the multiples of primes that are $< z$.

We make same definitions first.

Definition. Let A be a finite subset of $\mathbb{N}$, P a set of primes and $z > 0$ some real number. Define

$$P(z) = \prod_{\substack{p \in P \\ p < z}} p$$

For each $p \in P$, let there be an associated set $A_p \subseteq A$. Define $A_d = \bigcap_{p|d} A_p$ for $d \mid P(z)$. We define

$$S(A, P, z) = \left| \left(A \setminus \bigcup_{p|P(z)} A_p \right) \right|$$

to be the size of the set of all $a \in A$ that do not lie in A_p for all $p < z$.

Example. Fix $N \in \mathbb{N}$. If $A = [1, N] \cap \mathbb{Z}$ and P be the set of all prime numbers and let $z = [\sqrt{N} + 1]$. Also, for each $p < z$ let

$$A_p = \{a \in A : p \mid a\} = \{n \leq N : p \mid n\}$$

Note that $n \in A_p$ for some $p < z \iff (n, P(z)) > 1$. So $n \notin A_p$ for all $p < z \iff (n, P(z)) = 1$. Hence

$$S(A, P, z) = \sum_{\substack{n \leq N \\ (n, P(z))=1}} 1 = 1 + \sum_{\substack{1 < n \leq \sqrt{N} \\ (n, P(z))=1}} 1 + \sum_{\substack{\sqrt{N} < n \leq N \\ (n, P(z))=1}} 1$$

Note that the second sum is 0 as there is no $1 < n \leq \sqrt{N}$ that is coprime with $P(z)$. The last sum, by the above discussion, counts the primes between $\sqrt{N} + 1$ and N, thus

$$S(A, P, z) = 1 + \pi(N) - \pi(\sqrt{N})$$

Back to the definition above, recall the general principle of inclusion-exclusion:

Theorem 1.2 (Inclusion-Exclusion)**.** Let $A_1, \cdots, A_n$ be finite sets of a finite set A. For $I \subseteq \{1, \cdots, n\}$, define $A_I = \bigcap_{i \in I} A_i$, then we get

$$|A_1 \cup \cdots \cup A_n| = \sum_{\emptyset \neq J \subseteq \{1, \cdots, n\}} (-1)^{|J|+1} |A_J|$$

Equivalently, we have

$$|A \setminus (A_1 \cup \cdots \cup A_n)| = \sum_{J \subseteq \{1, \cdots, n\}} (-1)^{|J|} |A_J|$$

Proof: See [2]. □

With the set up in Definition 1.3, suppose all primes $< z$ are $\{p_1, \cdots, p_m\}$, then

$$S(A, P, z) = |A \setminus (A_{p_1} \cup \cdots \cup A_{p_n})| = \sum_{J \subseteq \{p_1, \cdots, p_m\}} (-1)^{|J|} |A_J|$$

If we identity each $J \subseteq \{p_1, \cdots, p_m\}$ with $d = \prod_{p \in J} p$, then $(-1)^{|J|} = \mu(d)$ and $|A_J| = |A_d|$, thus:

$$S(A, P, z) = \sum_{d | P(z)} \mu(d) |A_d| \tag{1}$$

This is known as the **Legendre's Identity**.

Now it all boils down to estimate $|A_d|$ for each $d \mid P(z)$. We will discuss this in Section 3. Let us see some examples first.

2 Twin Primes and Goldbach Numbers

In the setup in the Definition, we defined a set A and a set of primes P and some $z > 0$. So far the only example we have seen is the classic sieve of Eratosthenes. In this section we present some other examples.

Definition. We say a prime p is a **twin prime** if $p + 2$ is also a prime. Let $\pi_2(x)$ denote the number of twin primes $\leq x$.

Example. Fix $x > 0$. Define $A = \{n(n+2) : n \leq x\}$ and let P be the set of all primes. Let $z > 0$. Note that for $p < z$ and $n \geq z$, we have

$$p \mid n(n+2) \iff p \mid n \text{ or } p \mid (n+2)$$

Therefore

$$(p, n(n+2)) = 1 \iff p \nmid n(n+2) \iff p \nmid n \text{ and } p \nmid (n+2)$$

Therefore $S(A, P, z)$ counts all integers $z < n \leq x$ with this property, and all primes certainly satisfies this property, therefore

$$\pi_2(x) - \pi_2(z) = \sum_{\substack{z < p \leq x \\ p+2 \in P}} 1 \leq S(A, P, z)$$

It follows that

$$\pi_2(x) \leq \pi_2(z) + S(A, P, z) \leq z + S(A, P, z)$$

If we can give a good upper bound for $S(A, P, z)$, we can estimate the number of twin primes $\leq x$.

Definition. A positive even integer n is a **Goldbach number** if there are primes p, q such that $p + q = n$.

The Goldbach conjecture says that every even integer is a Goldbach number. That is, every even integer can be written as a sum of two prime numbers.

For $N > 2$ an even integer, let $R(N)$ denote the number of representations of N as the sum of two primes. The conjecture says $R(N) \geq 1$ for all N even.

Example. We will find an upper bound for $R(N)$. Consider the set

$$A = \{n(N-n) : n \leq N\}$$

Let P be all primes and $z > 0$. Similar to the above example, note that for $z < n < N - z$ we have

$$p \mid n(N-n) \iff p \mid n \text{ or } p \mid N - n$$

$p \mid n$ or $p \mid (N-n)$ implies $n \notin P$ or $N - n \notin P$, since $n > z > p$ and $N - n > z > p$ (So they cannot be the prime p themselves). By contrapositive:

$$n \in P \text{ and } N - n \in P \implies (n(N-n), P(z)) = 1$$

with $z < n < N - z$. Therefore

$$\begin{aligned} R(N) &= \sum_{\substack{p\le z\\ N-p\in P}} 1 + \sum_{\substack{z<p<N-z\\ N-p\in P}} 1 + \sum_{\substack{p\ge N-z\\ N-p\in P}} 1 \\ &\le \sum_{\substack{p\le z\\ N-p\in P}} 1 + \sum_{\substack{z<n<N-z\\ (n(N-n),P)=1}} 1 + \sum_{\substack{p\ge N-z\\ N-p\in P}} 1 \\ &\le z + S(A,P,z) + z \\ &= 2z + S(A,P,z) \end{aligned}$$

Now it amounts to estimate $S(A, P, z)$ again.

3 Main Theorem

We need to approximate the size of $|A_d|$,

$$|A_d| = \frac{\omega(d)}{d}X + R_d$$

We should interpret X as the approximation to $|A|$, and interpret $\omega(d)/d$ as the estimation of the 'proportion' of A that are in A_d, and R_d as the error term to this estimation.

Another way to understand $\omega(d)$ is this: Recall that when $A_d = \{a \in A : d \mid a\}$, then $\omega(d) = 1$ for all d, as we only allow A_d to contain elements that lie in the residue class 0 modulo d. However, we can make it more general by 'distinguish' some residue classes, and think of A_d to be elements that lie in at least of the residue classes. Suppose there are $\omega(d)$ distinguished residue classes, then $\omega(d)/d$ can be understood as the proportion of A that are in A_d.

In practice, our estimations usually satisfy that $|R_d| = O(\omega(d))$, so assume it is true.

Lemma 3.1. With the setting above, we have

$$S(A,P,z) = XW(z) + O(F(z))$$

where $W(z)$ is defined by

$$W(z) := \prod_{p|P(z)} \left(1 - \frac{\omega(p)}{p}\right)$$

and $F(t, z)$ is defined by

$$F(z) := \sum_{d|P(z)} \omega(d)$$

Proof: By Legendre's Identity and the above setting, we have

$$\begin{aligned} S(A,P,z) &= \sum_{d|P(z)} \mu(d)|A_d| = \sum_{d|P(z)} \mu(d)\frac{\omega(d)}{d}X + \sum_{d|P(z)} \mu(d)R_d \\ &= X\sum_{d|P(z)} \mu(d)\frac{\omega(d)}{d} + O(F(z)) \end{aligned}$$

Now, recall an important equality that

$$\sum_{d|n} \mu(d)f(d) = \prod_{p|n}(1 - f(d))$$

provided that f is multiplicative. Here, $f(d) = \omega(d)/d$ is multiplicative, so

$$\sum_{d|P(z)} \mu(d)\frac{\omega(d)}{d} = \prod_{d|P(z)} \left(1 - \frac{\omega(p)}{p}\right) = W(z)$$

The result follows. □

After doing some careful analysis on the main term and error term, we can get the following remarkable theorem:

Theorem 3.2. Suppose for some $y > 0$ we have $|A_d| = 0$ for every $d > y$. If there is some $c \geq 0$ such that

$$\sum_{p|P(z)} \frac{\omega(p)\log p}{p} \leq c\log z + O(1)$$

Then we have

$$S(A, P, z) = XW(z) + O\left(\left(X + \frac{y}{\log z}\right)(\log z)^{c+1}\exp\left(-\frac{\log y}{\log z}\right)\right)$$

where

$$W(z) = \prod_{\substack{p\in P\\ p<z}} \left(1 - \frac{\omega(p)}{p}\right)$$

Proof: See pp. 70-72 of [1]. □

4 Applications

Recall from Example 2.2 we have:

$$\pi_2(x) \leq z + S(A, P, z)$$

with $A = \{n(n+2) : n \leq x\}$ and $P =$ all primes.

Now we estimate $|A_d|$ for each $d \mid P(z)$. It suffices to estimate $|A_p|$ for each prime $p < z$. Note that $n(n+2) \equiv 0 \pmod{p}$ has two solutions if $p > 2$ and 1 solution if $p = 2$.

Therefore $\omega(p) = 2$ if $p > 2$ and $\omega(2) = 1$. Then

$$\sum_{p|P(z)} \frac{2\log p}{p} = 2\sum_{p|P(z)} \frac{\log p}{p} \leq 2\log z + O(1)$$

so we can pick $c = 2$. Moreover, let $y = x + 2$ so that $|A_p| = 0$ for $p > x + 2$, because the biggest prime factor of $n(n+2)$ for $n \leq x$ is at most $x + 2$. Using x to approximate the size of A, we have:

$$S(A, P, z) = xW(z) + O\left(x(\log z)^3\exp\left(-\frac{\log x}{\log z}\right)\right) \tag{1}$$

Here we need to estimate $W(z)$,

$$W(z) = \prod_{p<z}\left(1 - \frac{2}{p}\right) \leq \exp\left(-\sum_{p<z}\frac{2}{p}\right) \ll (\log z)^{-2}$$

The second last inequality is by using the inequality $1 + x \leq e^x$ on each term in the product. The last inequality is by Merten's Theorem. Now we can pick our z to make this expression and (1) cleaner. Pick $z > 0$ such that

$$\log z = \frac{\log x}{A\log\log x}$$

for some $A > 0$ large. Then (1) becomes

$$S(A, P, z) \ll \frac{x(\log\log x)^2}{\log^2 x}$$

Combine it with the inequality we get from Example 2.2, we have

Theorem 4.1. Let $\pi_2(x)$ denote the number of twin primes $\leq x$, then

$$\pi_2(x) \ll \frac{x(\log\log x)^2}{\log^2 x}$$

as $x \to \infty$.

Recall the Dirichlet Theorem says that for $(a, k) = 1$, there are infinitely many primes p such that $p \equiv a \pmod{k}$. The trick of the proof of this theorem is to prove the series

$$\sum_{p \equiv a \pmod{k}} \frac{1}{p} = \infty$$

One may wonder if this trick works for twin primes. The answer is no.

Corollary 4.2 (Brun)**.** The sum of reciprocals of twin primes converges.

Proof: For fixed $x > 0$, consider the sum

$$S(x) = \sum_{\substack{p \leq x \\ p+2 \text{ is prime}}} \frac{1}{p} = \sum_{n \leq x} a_n f(n)$$

where $a_n = 1$ if n is prime and $n + 2$ is prime, and 0 otherwise and $f(t) = 1/t$. Partial summation yields

$$S(x) = \frac{A(x)}{x} + \int_2^x \frac{A(t)}{t^2}\, dt$$

where $A(x) = \sum_{n \leq x} a_n = \pi_2(x)$. By Theorem 4.1 we have

$$S(x) \ll \frac{(\log\log x)^2}{\log^2 x} + \int_2^x \frac{(\log\log t)^2}{t \log^2 t}\, dt$$

The first term goes to 0, and the integral converges. Therefore $S(x)$ is bounded. □

Now we look at sum of two squares. Recall a theorem from elementary number theory,

Theorem 4.3. A positive integer $n = p_1^{a_1} \cdots p_r^{a_r}$ can be expressed as a sum of two squares $\iff$ for all $1 \leq i \leq r$, we have $p_i \equiv 3 \pmod{4} \implies 2 \mid a_i$.

This means a prime number p is a sum of two squares $\iff p \equiv 1 \pmod{4}$.

Let $N(x)$ denote the number of positive integers $\leq x$ that can be expressed as a sum of two squares and let $N'(x)$ denote the number of positive squarefree integers $\leq x$ that are sum of two squares.

Note that $N'(x)$ is easier to study, because the exponents of each prime factor is 1, so it means if a squarefree integer is a sum of two squares, then all of its prime factor must be 2 or is $\equiv 1$ mod 4. Note that

$$N(x) = N'(x) + \sum_{k \leq x} N'\left(\frac{x}{k^2}\right)$$

using a fact that every $n \in \mathbb{N}$ can be decomposed into $n = k^2 q$ where q is squarefree.

Let P be the set of all primes that are $\equiv 3 \bmod 4$. Remember, when we are doing sieve, we are removing things that we do not want! This time we only need prime factors that are 2 or is $\equiv 1 \bmod 4$, so we are removing all prime factors that are $\equiv 3 \bmod 4$. Then

$$N'(x) \leq |\{n \leq x : n \text{ has no prime factor } p \equiv 3 \pmod 4\}| \leq S(A, P, z)$$

for all $z > 0$. Let us explain this inequality. Define:

$$B = \{n \leq x : n \text{ has no prime factor } p \equiv 3 \pmod 4\}$$

Then B is obtained from $\{n \leq x\}$ by removing ALL integers that have prime factor $\equiv 3 \pmod 4$. But $S(A, P, z)$ only counts the number of integers after removing SOME integers that have prime factor $\equiv 3 \pmod 4$. Therefore B removes more things from A, thus $|B| \leq S(A, P, z)$.

For each $p \in P$, we have:

$$A_p = \{n \leq x : p \mid n\} = \frac{x}{p} + O(1)$$

So we have $\omega(p) = 1$ for all p, thus $\omega(d) = 1$ for all d, and $R_d = O(1)$. Hence, by Lemma 3.1 we have:

$$S(A, P, z) \leq x \prod_{\substack{p \equiv 3 \pmod 4 \\ p < z}} \left(1 - \frac{1}{p}\right) + O(2^{\pi(z)})$$

The error term is:

$$O(F(z)) = O\left(\sum_{d \mid P(z)} \omega(d)\right) = O\left(\sum_{d \mid P(z)} 1\right) = O(2^{\pi(z)})$$

where the last equality is because $P(z)$ has $2^{\pi(z)}$ divisors. Pick $z = \log x$. Hence we have:

$$N'(x) \leq x \prod_{\substack{p \equiv 3 \pmod 4 \\ p < \log x}} \left(1 - \frac{1}{p}\right) + O(2^{\log x})$$

The product on the RHS actually converges to 0 as $x \to \infty$, therefore $N'(x)/x \to 0$ as $x \to \infty$. In other words, for any $\epsilon > 0$ we have:

$$N'(x) \leq \epsilon x \text{ for } x \text{ large enough}$$

Therefore, recall that $\sum_{k=1}^{\infty} \frac{1}{k^2} < 2$, so:

$$N(x) \leq N'(x) + \sum_{k \leq x} N'\left(\frac{x}{k^2}\right) \leq \epsilon x + \sum_{k \leq x} \frac{\epsilon x}{k^2} \leq x\left(\epsilon + \epsilon \sum_{k=1}^{\infty} \frac{1}{k^2}\right) < x(\epsilon + 2\epsilon) = (3\epsilon)x$$

Therefore $N(x) = o(x)$ as $x \to \infty$.

References

[1] Cojocaru, A.C. and Murty, M.R., An Introduction to Sieve Methods and their Applications. London Mathematical Society 66. Cambridge University Press, 2006.

[2] Eğecioğlu, Ö. and Garsia, A. M., Lessons in Enumerative Combinatorics. Graduate Texts in Mathematics. Springer, 2021.