Chapter 4

Elementary results on the distribution of primes

4.1 Introduction

If $x > 0$ let $\pi(x)$ denote the number of primes not exceeding x. Then $\pi(x) \to \infty$ as $x \to \infty$ since there are infinitely many primes. The behaviour of $\pi(x)$ as the function of x has been the object of intense study by many celebrated mathematicians ever since the eighteenth century. Inspection of tables of primes led Gauss (1792) and Legendre (1798) to conjecture that $\pi(x)$ is asymptotic to $x/\log x$, that is

$$\lim_{x\to\infty} \frac{\pi(x)\log x}{x} = 1.$$

This conjecture was first proved in 1896 by Hadamard and de la Vallee Poussin and is known now as the *prime number theorem.*

Proofs of the prime number theorem are often classified as elementary or analytic. The proof of Hadamard and de la Vallee Poussin is analytic, using complex function theory and properties of the Riemann zeta function. An elementary proof was discovered by 1949 by A. Selberg and P. Erdös. Their proof makes no use of the zeta function nor of complex function theory but is quite intricate. In this class, we will concentrate on the first type of proof.

4.2 Chebycheff's Theorem and Merten's estimates

Definition. We define

$$\begin{aligned}\pi(x) &= \sum_{p\le x} 1,\\ \theta(x) &= \sum_{p\le x} \log p,\\ \psi(x) &= \sum_{n\le x} \Lambda(n) = \sum_{p^m\le x} \log p.\end{aligned}$$

Theorem 4.2.1 *There exist positive constants such that*

$$\frac{c_1 x}{\log x} \le \pi(x) \le \frac{c_2 x}{\log x} \quad (x \ge 2),$$

$$\begin{aligned} c_1 x \le \theta(x) \le c_2 x,\\ c_1 x \le \psi(x) \le c_2 x.\end{aligned}$$

Proof. Consider for $x \ge 4$, say,

$$S = \sum_{n\le x} \log n - 2\sum_{n\le \frac{x}{2}} \log n.$$

Using formula for $\sum_{n\le x} \log n$, we find that

$$S = x\log 2 + O(\log x).$$

Therefore,

$$\frac{x}{2} \le S \le x$$

provided $x \ge x_0 \ge 4$. Next, since $\log n = \sum_{d|n} \Lambda(d)$, we find that

$$\begin{aligned} S &= \sum_{n\le x}\sum_{d|n} \Lambda(d) - 2\sum_{n\le\frac{x}{2}}\sum_{d|n}\Lambda(d)\\ &= \sum_{d\le x} \Lambda(d)\left[\frac{x}{d}\right] - 2\sum_{d\le\frac{x}{2}} \Lambda(d)\left[\frac{x}{2d}\right]\\ &= \sum_{d\le\frac{x}{2}} \Lambda(d)\left\{\left[\frac{x}{d}\right] - 2\left[\frac{x}{2d}\right]\right\} + \sum_{\frac{x}{2}<d\le x} \Lambda(d)\left[\frac{x}{d}\right]\end{aligned}$$

Hence,

$$S = \sum_{d \le \frac{x}{2}} \Lambda(d)\theta_d + \sum_{\frac{x}{2} < d \le x} \Lambda(d).$$

Now, $\theta_d = 0$ or 1 since $[y] - 2[y/2] = 0$ or 1. Hence,

$$S \le \sum_{d \le x} \Lambda(d) = \psi(x).$$

and

$$S \ge \sum_{\frac{x}{2} < d \le x} \Lambda(d) = \psi(x) - \psi\left(\frac{x}{2}\right).$$

From these,

$$\psi(x) \ge S \ge \frac{x}{2} \quad (x \ge x_0).$$

Therefore,

$$\psi(x) \ge c_1 x.$$

Also,

$$\psi(x) - \psi\left(\frac{x}{2}\right) \le S \le x.$$

Therefore,

$$\begin{aligned} \psi(x) &\le x + \psi\left(\frac{x}{2}\right), \quad x \ge x_0 \\ &\le x + \frac{x}{2}\psi\left(\frac{x}{4}\right), \quad x \ge 2x_0 \\ &\vdots \\ &\le x + \frac{x}{2} + \cdots + \frac{x}{2^k} + \psi\left(\frac{x}{2^{k+1}}\right), \quad \frac{x}{2^{k+1}} < x_0 \le \frac{x}{2^k}. \end{aligned}$$

This implies that

$$\psi(x) \le 2x + \psi(x_0) \le c_2 x.$$

The inequalities for $\theta(x)$ and $\pi(x)$ follows from

Theorem 4.2.2

(i) $\theta(x) = \psi(x) + O(\sqrt{x})$.

(ii) $\pi(x) = \dfrac{\psi(x)}{\log x} + O\left(\dfrac{x}{\log^2 x}\right)$.

Corollary 4.2.3 *The prime number theorem* $\pi(x) \sim \dfrac{x}{\log x}$ *as* $x \to \infty$ *is equivalent to each of the following relations :*

(i) $\theta(x) \sim x \quad (x \to \infty)$

(ii) $\psi(x) \sim x \quad (x \to \infty)$.

Proof.

$$\begin{aligned}
\psi(x) - \theta(x) &= \sum_{\substack{p^m \le x \\ m \ge 2}} \log p \\
&= \sum_{\substack{p \le \sqrt{x} \\ m=2}} \log p + \sum_{p \le x^{1/3}} \log p \sum_{3 \le m \le \frac{\log x}{\log p}} 1 \\
&\le \psi(\sqrt{x}) + \sum_{p \le x^{1/3}} \log p \frac{\log x}{\log p} \\
&\ll \sqrt{x} + x^{1/3} \log x \ll \sqrt{x}.
\end{aligned}$$

(ii). It suffices to prove

$$\pi(x) = \frac{1}{\log x}\theta(x) + O\left(\frac{x}{\log^2 x}\right)$$

by (i).

$$\begin{aligned}
\pi(x) - \frac{\theta(x)}{\log x} &= \sum_{p \le x} \left(1 - \frac{\log p}{\log x}\right) \\
&= \sum_{p \le x} \log p \left(\frac{1}{\log p} - \frac{1}{\log x}\right).
\end{aligned}$$

By partial summation with

$$a(n) = \begin{cases} 0 & \text{if } n \ne p \\ \log p & \text{if } n = p. \end{cases}$$

$$A(t) = \sum_{n \le t} a(n) = \theta(t) \ll t.$$

The last sum is

$$\begin{aligned} \theta(x)\left(\frac{1}{\log x} - \frac{1}{\log x}\right) - \int_2^x \theta(t)\left(\frac{1}{\log t} - \frac{1}{\log x}\right)' dt \\ = \int_2^x \frac{\theta(t)}{t \log^2 t}\, dt \ll \int_2^x \frac{dt}{\log^2 t} \\ = \int_2^{\sqrt{x}} \frac{dt}{\log^2 t} + \int_{\sqrt{x}}^x \frac{dt}{\log^2 t} \\ \ll \sqrt{x} + \int_{\sqrt{x}}^x \frac{dt}{\log^2 x} \ll \frac{x}{\log^2 x}. \end{aligned}$$

Our next Theorem shows that there are infinitely many primes by showing that $\sum_{p \le x} \frac{1}{p}$ diverges.

Theorem 4.2.4 (Merten's estimates)

(i) $\sum_{n \le x} \frac{\Lambda(n)}{n} = \log x + O(1)$,

(ii) $\sum_{p \le x} \frac{\log p}{p} = \log x + O(1)$,

(iii) $\sum_{p \le x} \frac{1}{p} = \log\log x + A + O\left(\frac{1}{\log x}\right)$,

(iv) (Merten's Theorem) $\prod_{p \le x}\left(1 - \frac{1}{p}\right) = e^{-A}\left(1 + O\left(\frac{1}{\log x}\right)\right)$,

where A is a constant.

Proof. (i)

$$\begin{aligned}
\sum_{n\le x}\frac{\Lambda(n)}{n} &= \sum_{n\le x}\left\{\Lambda(n)\frac{1}{x}\left(\left[\frac{x}{n}\right]+O(1)\right)\right\}\\
&= \frac{1}{x}\sum_{n\le x}\Lambda(n)\left[\frac{x}{n}\right]+O\left(\frac{\sum_{n\le x}\Lambda(n)}{x}\right)\\
&= \frac{1}{x}\sum_{n\le x}(\Lambda * u)(n)+O(1),\\
&= \frac{1}{x}\sum_{n\le x}\log n+O(1)\\
&= \frac{1}{x}(x(\log x-1)+O(\log x))+O(1)\\
&= \log x+O(1).
\end{aligned}$$

(ii)

$$\begin{aligned}
0 \le \sum_{n\le x}\frac{\Lambda(n)}{n}-\sum_{p\le x}\frac{\log p}{p} &= \sum_{p\le x}\log p\sum_{\substack{m\ge 2\\ p^m\le x}}\frac{1}{p^m}\\
\ll \sum_{p\le x}\frac{\log p}{p^2} \ll \sum_{p\le x}\frac{p^{1/2}}{p^2} \ll 1.
\end{aligned}$$

(iii) First, we give a crude bound.

$$\begin{aligned}
\sum_{p\le x}\frac{1}{p} &= \sum_{p\le x}\left(\frac{\log p}{p}\right)\left(\frac{1}{p}\right)\\
&= A(x)\frac{1}{\log x}-\int_2^x A(t)\left(\frac{1}{\log t}\right)'\,dt\\
&= \frac{A(x)}{\log x}+\int_2^x\frac{A(t)}{t\log^2 t}\,dt\\
&= \frac{\log x+O(1)}{\log x}+\int_2^x\frac{\log t+O(1)}{t\log^2 t}\,dt \quad \text{by (ii)}\\
&= 1+O\left(\frac{1}{\log x}\right)+\int_2^x\frac{dt}{t\log t}+O\left(\int_2^x\frac{dt}{t\log^2 t}\right)\\
&= 1+\log\log x-\log\log 2+O(1),
\end{aligned}$$

since

$$\int_2^x \frac{dt}{t\log^2 t} = \int_{\log 2}^{\log x} \frac{du}{u^2} = O(1)$$

and

$$\int_2^x \frac{dt}{t\log t} = \int_{\log 2}^{\log x} \frac{du}{u} = \log\log x - \log\log 2.$$

To get a sharper bound, namely, $A + O(1/\log x)$ instead of $O(1)$, we need to show that

$$\int_2^x \frac{A(t)}{t\log^2 t}\,dt = \log\log x + A' + O\left(\frac{1}{\log x}\right).$$

Write $A(t) = \log t + R(t)$, $R(t) \ll 1, \quad t \geq 2$.

$$\begin{aligned}\int_2^x \frac{\log t + R(t)}{t\log^2 t}dt &= \int_2^x \frac{dt}{t\log t} + \int_2^x \frac{R(t)}{t\log^2 t}\\ &= \log\log x - \log\log 2 + \int_2^\infty \frac{R(t)}{t\log^2 t}dt - \int_x^\infty \frac{R(t)}{t\log^2 t}\,dt\\ &= \log\log x - \log\log 2 + A'' + O\left(\frac{1}{\log x}\right).\end{aligned}$$

(iv)

$$\begin{aligned}\log\prod_{p\leq x}\left(1-\frac{1}{p}\right) &= \sum_{p\leq x}\log\left(1-\frac{1}{p}\right)\\ &= \sum_{p\leq x}\left(-\frac{1}{p}+r_p\right), \quad r_p = \log\left(1-\frac{1}{p}\right)+\frac{1}{p}\\ &= \sum_{p\leq x} r_p - \sum_{p\leq x}\frac{1}{p}\\ &= -\log\log x + A + O\left(\frac{1}{\log x}\right) + \sum_p r_p - \sum_{p>x} r_p\\ &= -\log\log x + A' + O\left(\frac{1}{\log x}\right) + O\left(\sum_{p>x}\frac{1}{p^2}\right)\\ &= -\log\log x + A' + O\left(\frac{1}{\log x}\right) + O\left(\frac{1}{x-1}\right)\\ &= -\log\log x + A' + O\left(\frac{1}{\log x}\right).\end{aligned}$$

Hence,

$$\log \prod_{p\le x}\left(1-\frac{1}{p}\right) = -\log\log x + A' + O\left(\frac{1}{\log x}\right).$$

Exponentiating both sides :

$$\begin{aligned}\prod_{p\le x}\left(1-\frac{1}{p}\right) &= \exp\left(-\log\log x + A' + O\left(\frac{1}{\log x}\right)\right)\\ &= \frac{e^{A'}}{\log x}\exp\left(O\left(\frac{1}{\log x}\right)\right)\\ &= \frac{e^{A'}}{\log x}\left(1+O\left(\frac{1}{\log x}\right)\right),\end{aligned}$$

since $e^t = 1 + O(t)$.

4.3 Prime number Theorem and $M(\mu)$

Theorem 4.3.1 *The prime number theorem is equivalent to the relation*

$$\lim_{x\to\infty}\frac{1}{x}\sum_{n\le x}\mu(n) = 0, \quad i.e.\ M(\mu) = 0.$$

Proof.
(i) Prime number theorem implies $M(\mu) = 0$.
Define

$$M_1(\mu) = \lim_{x\to\infty}\frac{1}{x\log x}\sum_{n\le x}\mu(n)\log n.$$

Note that

$$M(\mu) = 0 \text{ iff } M_1(\mu) = 0,$$

since

$$|M_1(\mu) - M(\mu)| \le \frac{1}{x}\sum_{n\le x}\left|\frac{\log n}{\log x} - 1\right| \ll \frac{1}{\log x}.$$

Assume PNT (Prime Number Theorem) in the form $\theta(x) \sim x, \quad x \to \infty$.

$$\begin{aligned}
\sum_{n \le x} \mu(n) \log n &= \sum_{n \le x} \mu(n) \sum_{d \mid n} \Lambda(d) \\
&= \sum_{n \le x} \mu(n) \sum_{p \mid n} \log p \\
&= \sum_{p \le x} \log p \sum_{\substack{n \le x \\ p \mid n}} \mu(n) \\
&= \sum_{p \le x} \log p \sum_{\substack{n' \le x/p \\ p \nmid n'}} (-\mu(n')) \\
&= -\sum_{p \le x} \log p \sum_{n \le x/p} \mu(n) + O\left(\sum_{p \le x} \log p \sum_{\substack{n \le x/p \\ p \mid n}} 1 \right) \\
&= -\sum_{p \le x} \log p \sum_{n \le x/p} \mu(n) + O\left(x \sum_{p \le x} \frac{\log p}{p^2} \right) \\
&= -\sum_{n \le x} \mu(n) \sum_{p \le x/n} \log p + O(x) \\
&= -x \sum_{n \le x} \mu(n) \theta\left(\frac{x}{n}\right) + O(x), \quad \theta(y) = y + R(y) \\
&= -x \sum_{n \le x} \frac{\mu(n)}{n} - \sum_{n \le x} \mu(n) R\left(\frac{x}{n}\right) + O(x) \\
&= -\sum_{n \le x} \mu(n) R\left(\frac{x}{n}\right) + O(x).
\end{aligned}$$

Hence,

$$\frac{1}{x \log x} \left| \sum_{n \le x} \mu(n) \log n \right| \le \frac{1}{x \log x} \sum_{n \le x} \left| R\left(\frac{x}{n}\right) \right| + O\left(\frac{1}{\log x}\right).$$

Let $\epsilon > 0$ be given. By PNT, $R(y) \to 0$ as $y \to \infty$. Therefore, there exist a y_0 such that

$$|R(y)| \le \epsilon y \quad (y \ge y_0).$$

For $x \geq y_0$,

$$\begin{aligned}\sum_{n\leq x}\left|R\left(\frac{x}{n}\right)\right| &\leq \sum_{n\leq x/y_0}\epsilon\frac{x}{n}+\sum_{x/y_0<n\leq x}\max_{y\leq y_0}|R(y)| \\ &\leq \epsilon x\log x+O_\epsilon(x).\end{aligned}$$

Thus,

$$\overline{\lim}\,\frac{1}{x\log x}\left(\sum_{n\leq x}\left|R\left(\frac{x}{n}\right)\right|\right)\leq\epsilon.$$

Hence the result.

The other direction depends on $\Lambda = \log *\mu$ and the decomposition

$$\log n = d(n) - 2C + r(n).$$

For $y \geq 1$,

$$\begin{aligned}\sum_{n\leq y} r(n) &= \sum_{n\leq y}\log n-\sum_{n\leq y}d(n)+2Cy+O(1)\\ &= y(\log y-1)+O(\log y)-(y\log y+(2C-1)y+O(\sqrt{y}))+2Cy+O(1).\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{n\leq x}\Lambda(n) &= \sum_{n\leq x}(\mu*\log)(n)\\ &= \sum_{n\leq x}(\mu*d)-2C(\sum_{n\leq x}\mu*u)+\sum_{n\leq x}(\mu*r)\\ &= [x]-2C+\sum_{n\leq x}(\mu*r)(n).\end{aligned}$$

since $\sum_{n\leq x}\mu*u*u=[x]$. Thus, PNT follows from $\psi(x)\sim x$ if

$$\lim_{x\to\infty}\frac{1}{x}\sum_{n\leq x}(\mu*r)(n)=0.$$

Now,

$$\begin{aligned}\sum_{n\leq x}\mu*r(n) &= \sum_{n\leq x}\sum_{d_1d_2=n}\mu(d_1)r(d_2)\\ &= \sum_{d_1\leq x}\sum_{\substack{d_2\leq x\\ d_1d_2\leq x}}\mu(d_1)r(d_2)\\ &= S_1+S_2-S_3,\end{aligned}$$

where

$$S_1 = \sum_{d_2 \le y} \sum_{d_1 \le x/d_2} \mu(d_1) r(d_2)$$
$$S_2 = \sum_{d_1 \le x/y} \sum_{d_2 \le x/d_1} \mu(d_1) r(d_2)$$
$$S_3 = \sum_{d_1 \le x/y} \sum_{d_2 \le y} \mu(d_1) r(d_2),$$

and y is a parameter in $[1, x]$ to be chosen.

For fixed y,

$$|S_1| \le \sum_{d_2 \le y} |r(d_2)| |\sum_{d_1 \le x/d_2} \mu(d_1)| \to 0 \quad \text{as } x \to \infty,$$

using the assumption that $\frac{1}{x} \sum_{n \le x} \mu(n) \to 0$ as $x \to \infty$. Next,

$$\begin{aligned} |S_2| &\le \sum_{d_1 \le x/y} \left| \sum_{d_2 \le x/d_1} r(d_2) \right| \le c \sum_{d_1 \le x/y} \sqrt{\frac{x}{d_1}} \\ &\le c\sqrt{x} \sum_{d_1 \le x/y} \frac{1}{\sqrt{d_1}} \le c\sqrt{x} \left(1 + \int_1^{x/y} \frac{dt}{\sqrt{t}} \right) \le c_1 \frac{x}{\sqrt{y}}. \end{aligned}$$

Also,

$$|S_3| \le \frac{x}{y} \left| \sum_{d_2 \le y} r(d_2) \right| \le \frac{x}{y} c_2 \sqrt{y} = c_2 \frac{x}{\sqrt{y}}.$$

Hence

$$\overline{\lim}_{x \to \infty} \frac{1}{x} \left| \sum_{n \le x} (\mu * r)(n) \right| \le 0 + c_1 \frac{1}{\sqrt{y}} + \frac{c_2}{\sqrt{y}}.$$

Since y is arbitrary,

$$\lim_{x \to \infty} \frac{1}{x} \sum_{n \le x} (\mu * r)(n) = 0.$$