

Lecture 1

Instructor: Sruthi Sekar

Scribe: Srijan Das

1 Asymptotics

1.1 Lecture Overview

The lecture begins with a discussion on the Coupon Collector Problem. While discussing the solution, we stopped at the expansion of the Harmonic number which motivated the discussion of asymptotics. This is followed by definition of several asymptotic notations. Examples have been provided wherever necessary. This led to the completed solution of the Coupon Collector Problem.

1.2 Motivation

Exercise 1. *Coupon Collector Problem : There are n different kind of coupons. In every draw a coupon is picked at random which is equally likely to be any of the n kinds, and independent of the previous coupons. What is the expected time T when we are done collecting all n kinds?*

Solution: Let us define a random variable T_i = first time we have i unique coupons $\forall i \in \{0, 1, \dots, n\}$. We have to find $\mathbb{E}[T_n]$.

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[T_1 + (T_2 - T_1) + (T_3 - T_2) + \dots + (T_n - T_{n-1})] \\ &= \mathbb{E}[T_1] + [T_2 - T_1] + \dots + \mathbb{E}[T_n - T_{n-1}] \quad (\text{from Linearity of Expectations})\end{aligned}$$

Now, $\mathbb{E}[T_i - T_{i-1}]$ refers to the expected number of steps required to get i^{th} unique coupon after getting the $(i-1)^{\text{th}}$ unique coupon. So, we have already seen $i-1$ unique coupons. Hence, there are $n - (i-1)$ unique coupons left. Thus the probability of getting a unique coupon is:

$$p = \Pr[\text{unique coupon}] = \frac{n - (i-1)}{n} = \frac{n - i + 1}{n} \quad (1)$$

Here the following conditions are met:

1. Trials are independent.
2. Each trial has one of two outcomes: success (getting the i^{th} unique coupon) or failure (repetition of the earlier $i-1$ coupons)
3. For each trial the probability of success p is constant across trials.

Thus, we can conclude that this is a geometric distribution with probability of success as given in equation

1. Hence, required expectation :

$$\mathbb{E}[T_i - T_{i-1}] = \frac{1}{p} = \frac{n}{n - i + 1}$$

which follows from the expectation result of a geometric distribution ($\mathbb{E}(X) = \frac{1}{p}$). Hence,

$$\begin{aligned}\mathbb{E}[T_n] &= \sum_{i=1}^n \frac{n}{n - i + 1} \\ &= n \sum_{i=1}^n \frac{1}{n - i + 1} \\ &= n \left(\frac{1}{n} + \frac{1}{n-1} + \dots + 1 \right) \\ &= n \times H_n\end{aligned} \quad (2)$$

The sum in bracket is referred to as the n^{th} harmonic number, H_n . Since we do not have a well formed formula for the harmonic number, we estimate this type of series using **Asymptotics**. Let us take a break from the solution and study about Asymptotics.

The word Asymptotics is rooted in Greek. Like the hyperbola can be approximated by its two asymptotes. Asymptotics is a method used to describe limiting behavior.

1.3 Asymptotic Notations

Definition 2.

$$f(n) \prec g(n) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$$

Example 3. $1 \prec \log \log n \prec \log n \prec n^\epsilon \prec n^c \prec n^{\log n} \prec c^n \prec n^n \prec c^{c^n}$ This can be graphically understood from figure 1.

Exercise 4. Show that $n^n \prec c^{c^n}$ if $c > 1$.

Solution: We know,

$$\begin{aligned} n \log_c n &< c^n \text{ for } c > 1 \\ \implies c^{n \log_c n} &< c^{c^n} \\ \implies n^n &< c^{c^n} (\because a^{\log_b c} = c^{\log_b a}) \\ \implies \lim_{n \rightarrow \infty} \frac{n^n}{c^{c^n}} &= 0 \\ \implies n^n &\prec c^{c^n} \end{aligned} \tag{3}$$

Definition 5.

$$f(n) \sim g(n) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1$$

i.e. the functions f and g have roughly the same growth rate.

Note: In some books the ratio is given as 1 while in some other it is written as a constant. For all our applications, it doesn't matter.

Definition 6. Big Oh Notation

$$f(n) = \mathcal{O}(g(n)) (\text{ as } n \rightarrow \infty) \iff \exists c, n_0 > 0 \text{ such that } |f(n)| \leq c|g(n)| \forall n \geq n_0$$

In other words, $f(n)$ is upper bounded by a constant multiple of $g(n)$.

Example 7. 1. $\sum_{i=1}^n i = \frac{n(n+1)}{2} = \frac{n^2}{2} + \frac{n}{2} = \mathcal{O}(n^2)$

2. $\sum_{i=1}^n i^2 = n(n + \frac{1}{2})(n + 1) = \mathcal{O}(n^3)$

Refer to figure 2 for a graphical example.

Remark 1 The use of equal to (=) sign here is an abuse of notation. This is because $\mathcal{O}(g(n))$ refers to a set of functions. The correct statement should be $f(n) \in \mathcal{O}(g(n))$. In most literature, equal to sign is used.

Remark 2 The big oh notation is also defined when $x \rightarrow \infty$.

$$f(x) = \mathcal{O}(g(x)) \text{ as } x \rightarrow \infty \iff \exists c, \epsilon > 0 \text{ such that } |f(x)| \leq c|g(x)| \forall x \leq \epsilon$$

Example 8. $\frac{1}{x} + \frac{1}{x^2} = \mathcal{O}(\frac{1}{x^2})$ as $x \rightarrow 0$.

Whenever this asymptotic notation is used, it should always be mentioned in context of $x \rightarrow \infty$ or $x \rightarrow 0$. This is also applicable for other asymptotic notations which are described further.

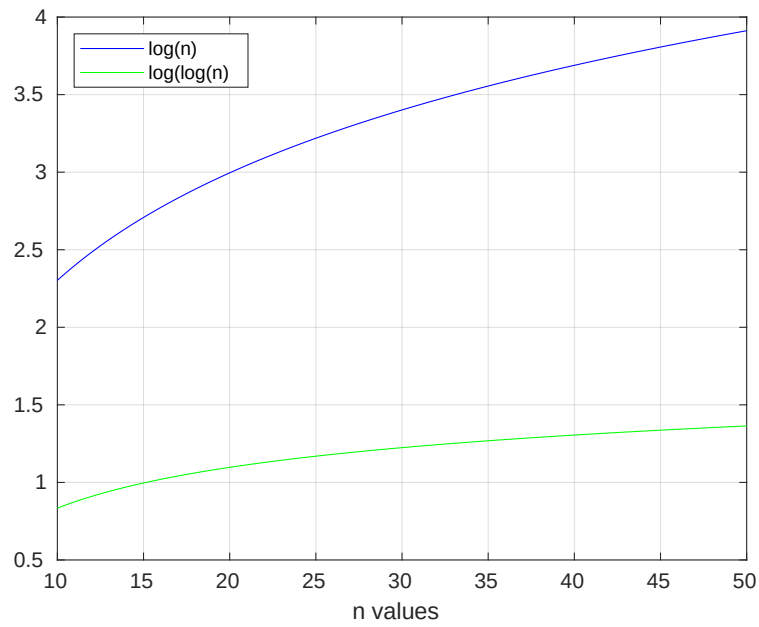


Figure 1: Comparison between $\log n$ and $\log \log n$

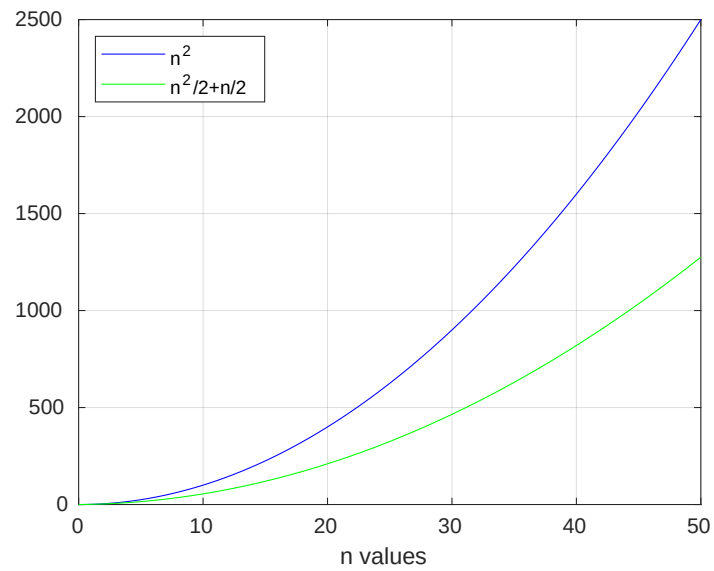


Figure 2: Example of $\mathcal{O}$ notation

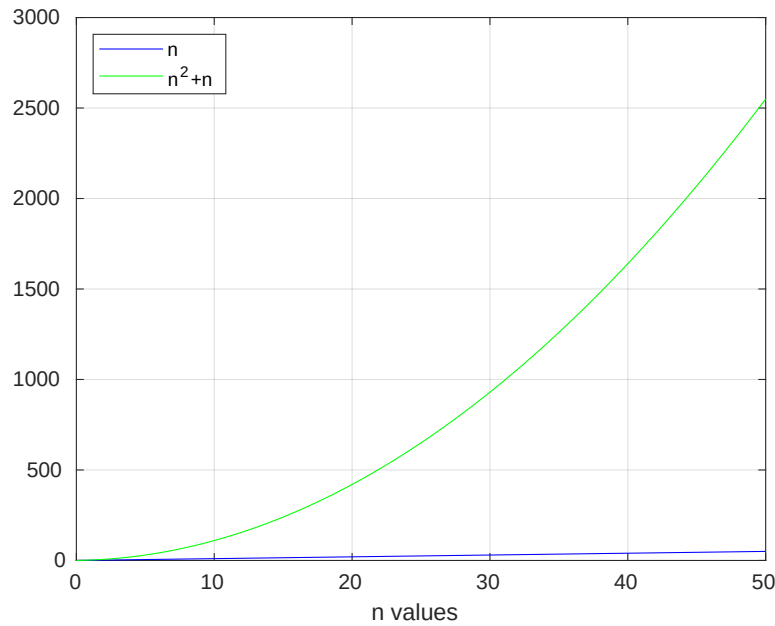


Figure 3: Example of Ω notation

Definition 9. Big Omega Notation

$$f(n) = \Omega(g(n)) \text{ (as } n \rightarrow \infty) \iff \exists c, n_0 > 0 \text{ such that } |f(n)| \geq c|g(n)| \forall n \geq n_0$$

Example 10.

$$n^2 + n = \Omega(n) \text{ (Refer to figure 3)}$$

Definition 11. Theta Notation

$$f(n) = \Theta(g(n)) \text{ (as } n \rightarrow \infty) \iff f(n) = \Omega(g(n)) \text{ and } f(n) = \mathcal{O}(g(n))$$

Example 12.

$$\sum_{i=1}^n i = \frac{n(n+1)}{2} = \frac{n^2}{2} + \frac{n}{2} = \Theta(n^2) \text{ (Refer to figure 4)}$$

Definition 13. Small Oh Notation

$$f(n) = o(g(n)) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0 \iff f(n) \prec g(n)$$

Example 14. $n^2 = o(n^3)$

From this definition we can conclude,

Theorem 15. $f(n) \sim g(n) \implies f(n) = g(n) + o(g(n))$

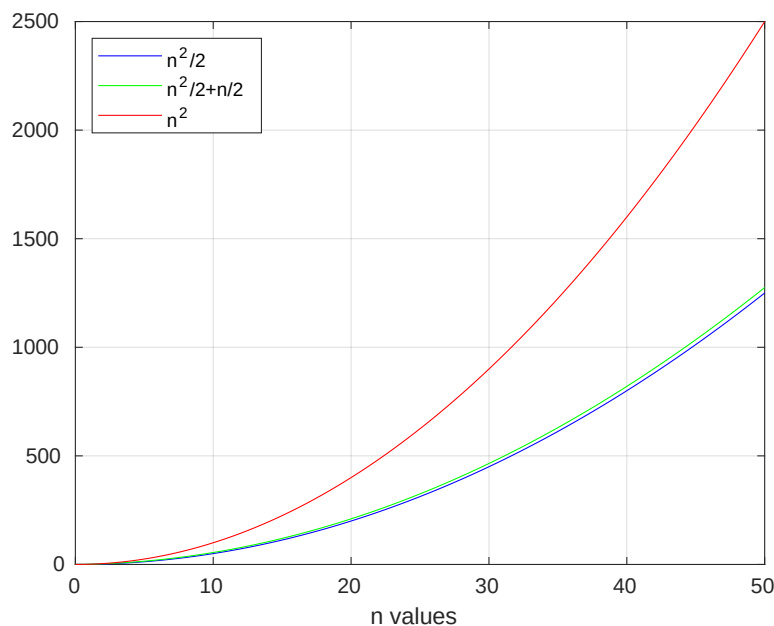


Figure 4: Example of Θ notation

Proof. Given, $f(n) \sim g(n)$. Hence,

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1 \text{ (from definition of } \sim \text{)}$$

Thus, $f(n)$ can be expressed in terms of $g(n)$ as:

$$\begin{aligned} f(n) &= g(n) + h(n) \text{ such that } \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1 \text{ for some } h(n) \\ \implies \lim_{n \rightarrow \infty} \frac{g(n) + h(n)}{g(n)} &= 1 \\ \implies 1 + \lim_{n \rightarrow \infty} \frac{h(n)}{g(n)} &= 1 \\ \implies \lim_{n \rightarrow \infty} \frac{h(n)}{g(n)} &= 0 \\ \implies h(n) &= o(g(n)) \text{ (from definition of small oh notation)} \end{aligned} \tag{4}$$

Therefore, $f(n) = g(n) + o(g(n))$. □

Definition 16. Small omega Notation

$$f(n) = \omega(g(n)) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \infty$$

Definition 17. Oh Tilde Notation This notation is used to hide log factors.

$$f(n) = \tilde{O}(g(n)) \iff f(n) = \mathcal{O}(g(n) \text{polylog}(g(n)))$$

Example 18. $n^2 \log^3 n = \tilde{O}(n^2)$

Exercise 19. Prove the following:

1. $n^m = \mathcal{O}(n^{m'})$ if $m < m'$
2. $\mathcal{O}(f(n)) + \mathcal{O}(g(n)) = \mathcal{O}(|f(n)| + |g(n)|)$

Solution

1. Given $m < m'$. Thus, for $n > 1$, $n^m < n^{m'}$.
So, $n^m = n^{m-m'} n^{m'} < n^{m'} (\because n^m < n^{m'})$
 $\therefore n^m < n^{m'} \forall n > 1$.
Hence, $n^m = \mathcal{O}(n^{m'})$.
2. Let $p(n) = \mathcal{O}(f(n))$ and $q(n) = \mathcal{O}(g(n))$.
Thus, $|p(n)| \leq c_1 |f(n)| \forall n > n_1$ and $|q(n)| \leq c_2 |g(n)| \forall n > n_2$ for some $c_1, c_2, n_1, n_2 > 0$.
Let $c' = \max(c_1, c_2)$ and $n' = \max(n_1, n_2)$.
Thus,

$$\begin{aligned}
 |p(n) + q(n)| &\leq |p(n)| + |q(n)| \text{ (property of modulus)} \\
 &\leq c_1 |f(n)| + c_2 |g(n)| \forall n > n' \text{ (from definition)} \\
 &\leq c' (|f(n)| + |g(n)|) \forall n > n' \text{ (} c_1, c_2 \leq c' \text{)} \\
 &\implies p(n) + q(n) = \mathcal{O}(|f(n)| + |g(n)|)
 \end{aligned} \tag{5}$$

Hence proved.

1.4 Harmonic Number

Let us derive the asymptotic expression of the n^{th} harmonic number. Initially, let us try to find a function $h(n)$ such that $H_n = \Theta(h(n))$. We know,

$$\begin{aligned}
 H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} \\
 \implies H_n &\leq 1 + \left(\frac{1}{2} + \frac{1}{2}\right) + \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right) + \dots \text{ (since } \frac{1}{3} \leq \frac{1}{2}, \frac{1}{5} \leq \frac{1}{4} \text{ and so on)} \\
 &\implies H_n \leq 1 + \lfloor \log_2 n \rfloor
 \end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned}
 H_n &\geq 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \dots \text{ (since } \frac{1}{3} \geq \frac{1}{4}, \frac{1}{5} \geq \frac{1}{8} \text{ and so on)} \\
 \implies H_n &\geq 1 + \frac{1}{2} \left[1 + \left(\frac{1}{2} + \frac{1}{2}\right) + \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right) + \dots \right] \\
 &\implies H_n \geq 1 + \frac{1}{2} \lceil \log_2 n \rceil
 \end{aligned} \tag{7}$$

From equations 6 and 7 we can conclude using big theta notation,

$$H_n = \Theta(\log_2 n)$$

Now, we claim:

$$H_n \sim \ln n$$

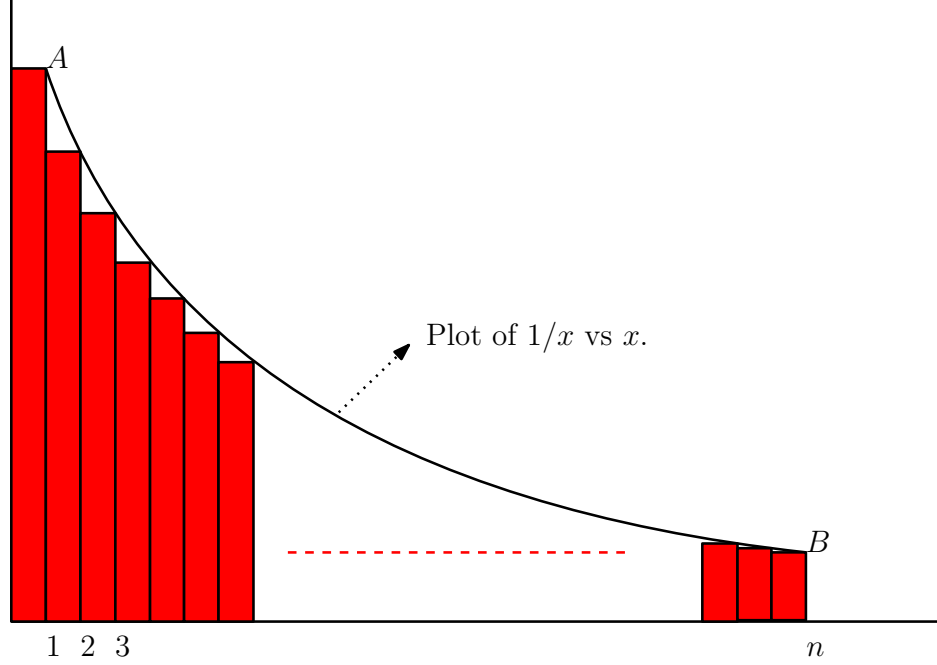


Figure 5: Upper bound of H_n

We will prove this claim using a graphical technique. Let us see the graph of the function $f(x) = \frac{1}{x}$ vs x for $x \in \{1, n\}$ (refer Fig. 5). Area of the red rectangles as shown in figure 5 is:

$$\begin{aligned} A_1 &= 1 + (2-1) \times f(2) + (3-2) \times f(3) + (4-3) \times f(4) + \dots + (n-(n-1)) \times f(n) \\ &= 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} = H_n \end{aligned} \quad (8)$$

Except the leftmost rectangle, all other rectangles lie strictly under the curve AB. So,

$$\begin{aligned} A_1 = H_n &\leq 1 + \text{Area under AB} = 1 + \int_1^n \frac{1}{x} dx = 1 + \ln n - \ln 1 = 1 + \ln n \\ \therefore H_n &\leq 1 + \ln n \end{aligned}$$

Now, extend the graph to $x = n+1$ to get curve AC (Fig. 6). Consider the set of green rectangles in the graph as shown in figure 6. Sum of areas of the green rectangles as shown in figure is:

$$\begin{aligned} A_2 &= (2-1) \times f(1) + (3-2) \times f(2) + (4-3) \times f(3) + \dots + (n+1-n) \times f(n) \\ &= 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} = H_n \end{aligned} \quad (9)$$

It is clear from the figure that A_2 is greater than the area under curve AC. Thus,

$$A_2 = H_n \geq \int_1^{n+1} \frac{1}{x} dx = \ln(n+1) - \ln 1 = \ln(n+1) \quad (10)$$

Therefore, we have:

$$\ln(n+1) \leq H_n \leq 1 + \ln n$$

Recall the Taylor series expansion of $\ln(1+x)$:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \quad (11)$$

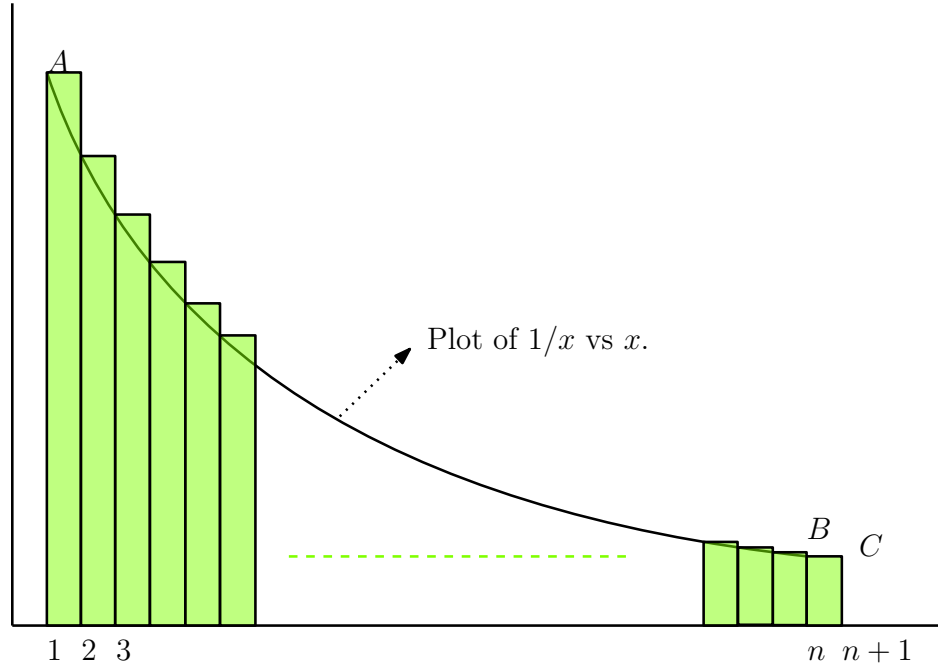


Figure 6: Lower Bound of H_n

Now,

$$\begin{aligned}
 \ln(n+1) &= \ln\left(n\left(1 + \frac{1}{n}\right)\right) = \ln n + \ln\left(1 + \frac{1}{n}\right) \\
 &= \ln n + \frac{1}{n} - \frac{1}{2n^2} + \frac{1}{3n^3} - \dots \text{(From equation 11)} \\
 &= \ln n + \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)
 \end{aligned} \tag{12}$$

Hence,

$$\ln n + \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^2}\right) \leq H_n \leq 1 + \ln n$$

Therefore,

$$\lim_{n \rightarrow \infty} \frac{H_n}{\ln n} = 1$$

Thus, $H_n \sim \ln n$. So, the solution to the coupon collector problem is :

$$\mathbb{E}[T_n] = n \times H_n \sim n \times \ln n$$

References

The content is based on the lecture covered in class. The images are drawn using MATLAB and ipe software.