A Theorist's Toolkit **(CMU 18-859T, Fall 2013)**

Lecture 1: Asymptotics

September 9, 2013

Lecturer: Ryan O'Donnell *Scribe: Anonymous*

In addition to the book references provided at the end of this document, two chapters of lecture notes on asymptotics by A.J. Hildebrand can be found at:

- http://www.math.uiuc.edu/~hildebr/595ama/ama-ch1.pdf
- http://www.math.uiuc.edu/~hildebr/595ama/ama-ch2.pdf

1 Asymptotic Notation

We all know that

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

If we want to describe the behavior of this function for large n, the quadratic term is the more important. We can write this as

$$\sum_{i=1}^{n} i = O(n^2).$$

Formally, the notation $f(x) = O(g(x))$ ("big-oh of $g(x)$") means that for x sufficiently large, there is a constant $C > 0$ such that $0 \leq f(x) \leq Cg(x)$.[1] This describes the asymptotic behavior of $f(x)$ as $x \to \infty$. Sometimes we will also talk about asymptotics of some function $f(x)$ as $x \to 0^+$. In that case, $f(x) = O(g(x))$ means that there is some $x_0 > 0$ such that for $0 \leq x \leq x_0$, $0 \leq f(x) \leq Cg(x)$. It will probably be clear from context which one is meant; generally, a variable named n goes to ∞, while a variable named ϵ goes to 0.

If we use $O(g(x))$ in an expression, such as $f(x) = 2^{O(g(x))}$, what we mean is that $O(g(x))$ can be replaced with some anonymous function of x which is $O(g(x))$. For example, we could write

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

One use of this that deserves special mention is $O(1)$: this is a function of x that is eventually bounded above by some constant.

In addition to O, there are several other symbols that can be used to say slightly different things:

[1] This is not entirely standard: some people allow $-Cg(x) \leq f(x) \leq Cg(x)$.

- $f(x) = \Omega(g(x))$ means that $g(x) = O(f(x))$: in other words, for x sufficiently large/small, there is a constant $C > 0$ such that $f(x) \geq Cg(x)$.

- $f(x) = \Theta(g(x))$ means that $f(x) = \Omega(g(x))$ and $f(x) = O(g(x))$ simultaneously: for x sufficiently large/small, there are constants $C_1, C_2 > 0$ such that $C_1 g(x) \leq f(x) \leq C_2 g(x)$.

- $f(x) \sim g(x)$ means that $\frac{f(x)}{g(x)} \to 1$ in the limit. For example, we could write
$$\sum_{i=1}^{n} i \sim \frac{1}{2}n^2.$$

- $f(x) = o(g(x))$ means that $\frac{f(x)}{g(x)} \to 0$, and $f(x) = \omega(g(x))$ means that $\frac{f(x)}{g(x)} \to \infty$. For example, we could write
$$\sum_{i=1}^{n} i = \frac{1}{2}n^2(1 + o(1)).$$

- $f(x) \leq \text{poly}(g(x)))$ means that $f(x) = g(x)^{O(1)}$: $f(x)$ is bounded by some polynomial function of $g(x)$.

- $f(x) = \widetilde{O}(g(x))$ means that $f(x) \leq g(x) \cdot \text{poly}(\log g(x))$: we forget about some polynomial in $\log g(x)$, which is insignificant compared to g itself. For example, we can write $n^5 \cdot 3^n = \widetilde{O}(3^n)$. Note that $n^5 \cdot 3^n \neq \widetilde{O}(2^n)$: the difference between 2^n and 3^n is too big. We can put a tilde on other things as well, such as $\widetilde{\Theta}(g(x))$.

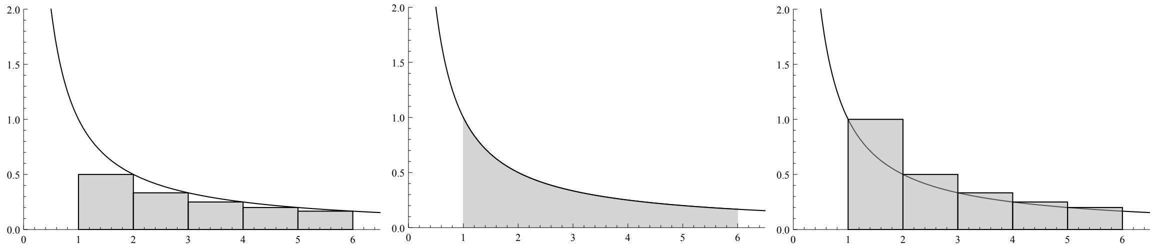
- If $x \to 0^+$, then $f(x) = \widetilde{O}(g(x))$ instead means $f(x) \leq g(x) \cdot \text{poly}(\log 1/g(x))$, since $g(x)$ is probably something small. For example, $\epsilon \cdot \log^2(1/\epsilon) = \widetilde{O}(\epsilon)$.

2 The Harmonic Number

Let $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$. This is called the n-th harmonic number. It comes up in many applications including, for example, the coupon collector problem: if there are n different coupons, and you pick coupons uniformly at random with replacement, you will need to look at $n \cdot H_n$ coupons (in expectation) before you have seen all of them at least once.

We can get a simple upper bound for H_n as follows: if $k = \lceil \log_2(n+1) \rceil$, then

$$\begin{aligned} H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &\leq 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2^k - 1} \end{aligned}$$



(a) $\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$ (b) $\int_1^n \frac{1}{x}\,dx$ (c) $1+\frac{1}{2}+\cdots+\frac{1}{n-1}$

Figure 1: Comparing the harmonic series to an integral

$$\begin{aligned}
&\le 1+\left(\frac{1}{2}+\frac{1}{3}\right)+\left(\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}\right)+\cdots+\left(\frac{1}{2^{k-1}}+\cdots+\frac{1}{2^k-1}\right)\\
&\le 1+\left(\frac{1}{2}+\frac{1}{2}\right)+\left(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}\right)+\cdots+\left(\frac{1}{2^{k-1}}+\cdots+\frac{1}{2^{k-1}}\right)\\
&\le \underbrace{1+1+\cdots+1}_{k}=k.
\end{aligned}$$

Therefore $H_n \le \lceil \log_2(n+1) \rceil \sim \log_2 n$.

A similar technique can give us a lower bound on H_n: if $k=\lfloor \log_2 n \rfloor$, then

$$\begin{aligned}
H_n &= 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}\\
&\ge 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{2^k}\\
&\le 1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)+\cdots+\left(\frac{1}{2^{k-1}+1}+\cdots+\frac{1}{2^k}\right)\\
&\le 1+\frac{1}{2}+\left(\frac{1}{4}+\frac{1}{4}\right)+\left(\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\right)+\cdots+\left(\frac{1}{2^k}+\cdots+\frac{1}{2^k}\right)\\
&\le 1+\underbrace{\frac{1}{2}+\frac{1}{2}+\cdots+\frac{1}{2}}_{k}=1+\frac{1}{2}k.
\end{aligned}$$

Therefore $H_n \ge 1+\frac{1}{2}\lfloor \log_2 n \rfloor \sim \frac{1}{2}\log_2 n$. This is already enough to conclude $H_n=\Theta(\log n)$. (Note that the exact base of the log is irrelevant in big-Θ notation.)

In this class, this may be the last time you see the base 2 in $\log_2 n$ written explicitly; we will assume $\log n$ is $\log_2 n$ (which is sometimes also written as $\lg n$); for the natural log, we will write $\ln n$ (pronounced "lon enn").

What should we do if we want to figure out how big H_n is more precisely?

We can approximate the sum $\sum_{i=1}^{n}\frac{1}{i}$ by an integral. A lower bound for the integral $\int_1^n \frac{1}{x}\,dx$ is $\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}=H_n-1$, and an upper bound is $1+\frac{1}{2}+\cdots+\frac{1}{n-1}=H_{n-1}$ (see Figure 1).

Since the antiderivative of $\frac{1}{x}$ is $\ln x$, the integral is $\ln n - \ln 1 = \ln n$, so we have

$$H_n - 1 \le \ln n \le H_{n-1} \qquad \Leftrightarrow \qquad \ln(n+1) \le H_n \le 1 + \ln n.$$

We can write down a slightly less accurate but prettier estimate: $\ln n \le H_n \le 1 + \ln n$. In particular, $H_n \sim \ln n$.

What is the error when we replace $\ln(n+1)$ by $\ln n$?

$$\begin{aligned} \ln(n+1) &= \ln\left(n \cdot \left(1 + \frac{1}{n}\right)\right) \\ &= \ln n + \ln\left(1 + \frac{1}{n}\right) \\ &= \ln n + \Theta\left(\frac{1}{n}\right). \end{aligned}$$

The last step follows from the Taylor expansion of $\ln x$: for $-1 < x \le 1$.

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots.$$

Since x in our case is $\frac{1}{n}$, which goes to 0, the first term is most significant, and we can approximate $\ln\left(1+\frac{1}{n}\right)$ by $\frac{1}{n}$. The error bound in Taylor's theorem states that $\ln\left(1+\frac{1}{n}\right) = 1 + \frac{1}{n} - \frac{\xi^2}{2}$ for some $0 \le \xi \le \frac{1}{n}$, so actually we can say $\ln(n+1) = \ln n + \frac{1}{n} - O\left(\frac{1}{n^2}\right)$.

This is related to one of the most useful asymptotic approximations you will use: e^x is approoximately $1 + x$ for small x. To be more precise,

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ &= 1 + x + O(x^2). \end{aligned}$$

The Taylor expansion holds for all x, but only for small x is x^2 less significant than x. However, it's true for all x that $e^x \ge 1 + x$.

Although we will not prove this, H_n can actually be described even more precisely:

$$H_n = \ln n + \gamma + O\left(\frac{1}{n}\right),$$

where $\gamma \approx 0.577$ is known as the Euler-Mascheroni constant.

3 The Birthday Paradox

Suppose we have m bins which can hold balls, and we chuck n balls into the bins at random (to be precise, each ball chooses a bin uniformly at random, and the choices are independent). What is the probability that no two balls land in the same bin?

In the case $m = 366$, we get the birthday paradox: if you have n people, what is the probability that two will share a birthday? The reason it's a "paradox" is that the probability is surprisingly large: although only $n = 367$ guarantees that two people share a birthday, in a group of 30 people the probability is already over 70%.[2]

In general, the probability of no collisions between n balls thrown into m bins is

$$p_{n,m} = \left(1 - \frac{1}{m}\right)\left(1 - \frac{2}{m}\right)\cdots\left(1 - \frac{n-1}{m}\right).$$

A reasonable question to ask is: for what n, as a function of m, is $p_{n,m} \approx \frac{1}{2}$?

By using the inequality $e^x \geq 1 + x$ from the previous section, we can get a simple upper bound:

$$\begin{aligned} p_{n,m} &\leq e^{-1/m} \cdot e^{-2/m} \cdots e^{-(n-1)/m} \\ &= \exp\left(-\frac{1}{m} - \frac{2}{m} - \cdots - \frac{n-1}{m}\right) \\ &= \exp\left(-\frac{n(n-1)}{2m}\right). \end{aligned}$$

The $n - 1$ is somewhat annoying: we'd like to write $p_{n,m} \leq \exp\left(-\frac{n^2}{2m}\right)$, but this doesn't quite follow from the above. In any case, if we believe that this inequality is close to being the truth, we could answer the question "when is $p_{n,m} \approx \frac{1}{2}$" by solving this for n, obtaining $n \sim \sqrt{2 \ln 2} \cdot \sqrt{m}$.

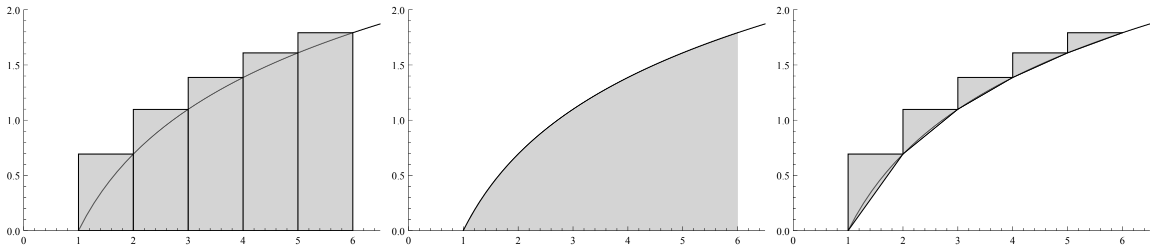
We also want a lower bound. The inequality $1 + x \leq e^x$ came from $\ln(1 + x) \leq x$, or $\ln x \leq x - 1$. By extending the Taylor series for $\ln x$, we can get a matching lower bound that looks like $1 + x \geq \exp(x - O(x^2))$. It follows that:

$$\begin{aligned} p_{n,m} &\geq \exp\left(-\frac{1}{m} - O\left(\frac{1}{m^2}\right)\right)\exp\left(-\frac{2}{m} - O\left(\frac{2^2}{m^2}\right)\right)\cdots\exp\left(-\frac{n-1}{m} - O\left(\frac{(n-1)^2}{m^2}\right)\right) \\ &= \exp\left(-\frac{1}{m} - \frac{2}{m} - \cdots - \frac{n-1}{m}\right)\exp\left(\frac{1}{m^2}O\left(1^2 + 2^2 + \cdots + (n-1)^2\right)\right) \\ &= \exp\left(-\frac{n(n-1)}{2m}\right)\exp\left(-O\left(\frac{n^3}{m^2}\right)\right). \end{aligned}$$

We get the same thing we got in the upper bound, multiplied by an error factor. But when $n = \Theta(\sqrt{m})$, this error factor looks like $\exp\left(-O\left(\frac{1}{\sqrt{m}}\right)\right)$, which is $1 - O\left(\frac{1}{\sqrt{m}}\right)$.

The implied constant in $n = \Theta(\sqrt{m})$ doesn't matter too much. But it often helps to know that when there are $O(\sqrt{m})$ balls and m bins, you probably get a collision.

[2] In our class, Alex Kazachkov shares a birthday with me (Anonymous).



(a) $\ln 2 + \cdots + \ln n$ (b) $\int_1^n \ln x \, dx$ (c) $\frac{1}{2} \ln n$

Figure 2: Comparing the sum $\ln 1 + \ln 2 + \cdots + \ln n$ to an integral

4 Stirling's Formula

Recall that $n! = n(n-1)(n-2) \cdot \cdots \cdot 2 \cdot 1$. How large is $n!$, asymptotically?

A very simple upper bound for $n!$: we can replace every factor by n, getting $n! \leq n^n$. For a lower bound, we could try several things:

- All factors except 1 are at least 2, so $n! \geq 2^{n-1}$.
- The first $\lceil \frac{1}{2} n \rceil$ factors are at least $\lceil \frac{1}{2} n \rceil$, so $n! \geq \lceil \frac{n}{2} \rceil^{\lceil n/2 \rceil} \geq \left(\frac{n}{2}\right)^{n/2}$.

Because all of these are very large products, we can get a better sense of how large they are by taking logs. We can then conclude that

$$2^{n-1} \leq \left(\frac{n}{2}\right)^{n/2} \leq n! \leq n^n$$

because

$$(n-1) \ln 2 \leq \frac{n}{2} \ln \left(\frac{n}{2}\right) \leq \ln(n!) \leq n \ln n.$$

This is already enough to say that $\ln(n!) = \Theta(n \ln n)$, and therefore $n! = 2^{\Theta(n \log n)}$.

To get a better estimate, we can expand

$$\ln(n!) = \ln 2 + \ln 3 + \cdots + \ln n.$$

Just like we did for harmonic numbers, we can approximate this sum by an integral. The sum $\ln 2 + \cdots + \ln n$ (in Figure 2a) is an over-estimate of the integral $\int_1^n \ln x \, dx$ (in Figure 2b). The antiderivative of $\ln x$ is $x \ln x - x$, so we conclude

$$\ln(n!) \geq (n \ln n - n) - (1 \ln 1 - 1) = n \ln n - n + 1.$$

To get an upper bound on $\ln(n!)$, we can subtract off the triangles in Figure 2c, which will leave an under-estimate of the integral. Looked at sideways, the triangles have height 1 and their bases sum to $\ln n$, so their area is $\frac{1}{2} \ln n$, and therefore

$$\ln(n!) \leq n \ln n - n + \frac{1}{2} \ln n + 1.$$

If we take these bounds and exponentiate, we get:

$$e \cdot \left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n} \cdot \left(\frac{n}{e}\right)^n .$$

In particular, $n! = \widetilde{\Theta}\left(\left(\frac{n}{e}\right)^n\right)$.

We can still hope to do better. The error in our upper bound is the area of the slivers that form the overlap between Figure 2b and Figure 2c. It can be shown[3] that the total area of these slivers is $O(1)$. As a consequence,

$$n! = \Theta(\sqrt{n}) \cdot \left(\frac{n}{e}\right)^n .$$

This is known as Stirling's formula, of which a more precise variant states that

$$n! = \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n \left(1 \pm O\left(\frac{1}{n}\right)\right).$$

We will see where the $\sqrt{2\pi}$ comes from when we talk about the Central Limit Theorem.

We can use Stirling's formula to estimate the binomial coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ when $k = pn$ (to simplify notation, let $q = 1 - p$, so that $n - k = qn$). As $n \to \infty$ for constant p,

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} = \frac{n!}{(pn)!(qn)!} \\
&= \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 \pm O(\frac{1}{n}))}{\sqrt{2\pi pn}\left(\frac{pn}{e}\right)^{pn}\sqrt{2\pi qn}\left(\frac{qn}{e}\right)^{qn}(1 \pm O(\frac{1}{pn}))(1 \pm O(\frac{1}{qn}))} \\
&\sim \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{\sqrt{pq}}\left(\frac{1}{p^p q^q}\right)^n \\
&= \frac{1}{\sqrt{2\pi pq}} \cdot \frac{1}{\sqrt{n}} \cdot 2^{H(p)n}
\end{aligned}$$

where $H(p) = p \log \frac{1}{p} + q \log \frac{1}{q}$. This quantity is of independent interest and is known as the binary entropy of p. In particular, $H(1/2) = 1$, so we can conclude

$$\frac{\binom{n}{n/2}}{2^n} \sim \sqrt{\frac{2}{\pi n}} = \Theta\left(\frac{1}{\sqrt{n}}\right).$$

When n is even, this is the probability that if n fair coins are flipped, exactly half will come up heads.

[3] Cut the k-th triangle by the tangent line to $\ln x$ at $k+1$. This bounds the k-th sliver in a smaller triangle with height 1 and base $\ln(k+1) - \ln k - \frac{1}{k+1}$. The sum of the areas of the first n triangles is therefore $\frac{1}{2}(\ln(n+1) - H_{n+1} + 1)$, which converges to $\frac{1}{2}(1-\gamma)$ as $n \to \infty$.

When $k = o(n)$, things are slightly different: the error factor depending on pn is no longer comparable to the one depending on n. In this case, it's more accurate to approximate $\binom{n}{k}$ by $\frac{n^k}{k!}$. The error analysis is as follows:

$$\begin{aligned}\binom{n}{k} &= \frac{n(n-1)(\cdots)(n-k+1)}{k!} \\ &= \frac{n^k}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right).\end{aligned}$$

The error factor is exactly the probability $p_{k,n}$ we looked at in the Birthday Paradox section. We know that $p_{k,n}$ has a constant value for $k = \Theta(\sqrt{n})$, so for $k = o(\sqrt{n})$, $p_{k,n} = 1 - o(1)$, and we have

$$\binom{n}{k} \sim \frac{n^k}{k!}.$$

Since $n(n-1)(\cdots)(n-k+1) \leq n^k$, this is also an upper bound for all k.

References

[DB70] Nicolaas Govert De Bruijn. *Asymptotic Methods in Analysis*, volume 4. Dover Publications, 1970.

[GKP94] Ronald L Graham, Donald E Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., 1994.