Big-O notation

Big-O notation is used to describe the behavior of functions as the argument tends to infinity. In this class, we will often be interested in bounds and approximation and we'll use this notation frequently.

Definition 1. *For two functions $f(n), g(n)$, we say that $f(n) = O(g(n))$ if there exist positive constants c, N such that*

$$|f(n)| \leq c \cdot g(n), \quad \text{for all } n \geq N.$$

Definition 2. *For two functions $f(n), g(n)$, we say that $f(n) = o(g(n))$ if*

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

Definition 3. *For two functions $f(n), g(n)$, we say that $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$. This is equivalent to saying that $f(n) = O(g(n))$ but not $f(n) = o(g(n))$.*

Binomial and geometric distribution

The binomial and geometric distributions are two of the most commonly used discrete distributions. You will likely find them built-in to any software or coding language you use.

Definition 4. *A* binomial random variable *X, with parameters n and p, counts the number of successes in a series of n trials, where each trial succeeds with probability p independently of other trials. we have*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We have

$$\mathbb{E}(X) = np.$$

Definition 5. *A* geometric random variable *X, with parameter p, counts the number of trials until the first success, where each trial succeeds with probability p independently of other trials. we have*

$$\mathbb{P}(X = k) = p(1-p)^{k-1}.$$

We have

$$\mathbb{E}(X) = \frac{1}{p}.$$

You are encouraged to derive the above expected values yourself. You will need to look up the binomial theorem, in case you are not familiar with that yet.

Application: Dominating sets

A dominating set in a graph $G = (V, E)$ is a set S such that every vertex is in S or has a neighbor in S. Dominating sets have many applications, for example in communication routing in networks. Often, the goal is to find a smallest possible dominating set. This problem is NP-complete in general.
Let $d(v)$ be the degree of v in G, $\delta(G)$ be the minimum degree in G (and $\Delta(G)$ be the maximum degree in G).

Lemma 6. *Every graph G has a dominating set of at most size $\frac{\log(1+\delta)+1}{1+\delta}n$.*

Suppose $\delta \geq 1$. Create the set X by including each $v \in V$ with probability p i.i.d. Then create $Y = \{v \in V \backslash X : v \text{ has no neighbor in } X\}$. By definition, $X \cup Y$ is a dominating set in G. Since each v is included in X with probability p, the expected size of X is $\mathbb{E}(|X|) = np$. The expected size of Y is

$$\begin{aligned}\mathbb{E}(|Y|) &= \sum_{v \in V} \mathbb{E}(1_{v \in Y}) = \sum_{v \in V} \mathbb{P}(v \in Y) = \sum_{v \in V} \mathbb{P}(v \notin X, v\text{'s neighbors} \notin X) \\ &= \sum_{v \in V} (1-p)^{d(v)+1} \leq \sum_{v \in V} (1-p)^{\delta+1} = n(1-p)^{\delta+1}\end{aligned}$$

Since X and Y are disjoint we have

$$\mathbb{E}(|X \cup Y|) \leq np + n(1-p)^{\delta+1}.$$

Exercise 1. *Use the inequalities from the first set of notes to show that* 2

$$\mathbb{E}(|X \cup Y|) \leq np + ne^{-p(\delta+1)}.$$

Then, find the value of p that minimizes this function, and finish the proof of Lemma 6.

Application: Tournaments

A tournament is a directed complete graph: a copy of K_n such that each edge $\{u, v\}$ is directed $u \to v$ or $v \to u$. We can think of this the intuitive way: we have n players and play a tournament where every player plays against every other player exactly once, and there are no draws.

Exercise 2. *A (directed) Hamiltonian path in a tournament is a directed path that visits every vertex. Show that for every n there is a tournament on n vertices that has $n!2^{-(n-1)}$ Hamiltonian paths.* 1

We say that a tournament has the S_k property if for every subset $S \subseteq U$ with $|S| = k$ there is a vertex $v \in U \setminus S$ such that v beats every player in S (has edges towards all vertices in S). It is not obvious that such a tournament exists for any k, but we can use a probabilistic method to show that it does. We ask: What is the smallest value of n for which a tournament with the S_k property exists? ($k = 1$ is easy (rock-paper-scissors). $k = 2$ or $k = 3$ already much harder by hand.)

Theorem 7. *There exists a tournament on n vertices with the S_k property when*

$$\binom{n}{k}\left(1 - \left(\frac{1}{2}\right)^k\right)^{n-k} < 1.$$

Exercise 3. *Prove Theorem 7.* 2

Now, we will try to rewrite this as a bound on n. Let $f(k)$ be smallest possible n such that there exists a tournament with the S_k property on n vertices.

Exercise 4. *Show that* 1

$$\binom{n}{k}\left(1 - \left(\frac{1}{2}\right)^k\right)^{n-k} < \left(\frac{ne}{k}\right)^k e^{-(n-k)/2^k}.$$

Exercise 5. *Show that there exists a constant c such that $f(k) \leq ck^2 2^k$.* 1

Application: coupon collector's problem

Suppose there are n types of coupons and at every time step we receive one coupon at random, uniformly sampled from the n types, always independently from previous coupons received. We are interested in the time until all coupon types are collected. Let

$$X_i = \text{time until } (i+1)\text{-th coupon collected, have } i \text{ types.}$$

At each time step we collect the $(i+1)$-th coupon with probability $p_i = (n-i)/n$. The probability that it takes k time steps to get the $(i+1)$-th coupon is

$$\mathbb{P}(X = k) = (1-p_i)^{k-1} p_i$$

since we need to not have received it in the first $k-1$ time steps, and then obtain it on the k-th time step. Hence X_i follows a geometric distribution with parameter p_i. Note that for a geometric distribution

$$\mathbb{E}(X_i) = \frac{1}{p_i} \qquad .$$

We are interested in $X = \sum_{i=0}^{n-1} X_i$, the time until all coupon types are collected. By linearity of expectation

$$\mathbb{E}(X) = \sum_{i=0}^{n-1} \mathbb{E}(X_i) = \sum_{i=0}^{n-1} \frac{1}{p_i} = \sum_{i=0}^{n-1} \left(\frac{n-i}{n}\right)^{-1}.$$

Making a change of index we can rewrite the above as

$$\mathbb{E}(X) = \sum_{i=1}^{n} \frac{n}{i} = n \sum_{i=1}^{n} \frac{1}{i}.$$

It is well known that the partial sums for the harmonic series grow as $\log n + O(1)$. (You can look up this derivation if you'd like, but you do not need to know how to do that for this class.) Thus the expected time until all coupon types are collected is

$$\mathbb{E}(X) = n(\log n + O(1)) = n \log n + O(n).$$

Application: isolated vertices in $G(n,p)$

The two models $G(n,m)$ and $G(n,p)$ are the two simplest models for generating a random graph on n vertices. The model $G(n,m)$ was introduced by Erdős and Rényi in a series of papers that form the foundation of random graph theory. The model $G(n,p)$ was introduced by Gilbert, although it is commonly referred to as the Erdős-Rényi model. The model $G(n,m)$ generates a graph uniformly at random over all simple (no multi-edges or self-loops) graphs on n vertices and m edges. There are $\binom{\binom{n}{2}}{m}$ such graphs. Therefore, for any graph G on n vertices and m edges, we have

$$\mathbb{P}(G(n,m) = G) = \binom{\binom{n}{2}}{m}^{-1}.$$

In this class, we will not talk much about the model $G(n,m)$, as it behaves very similarly to $G(n,p)$ in most interesting ways, but it is much harder to analyze. The model $G(n,p)$ is a graph on n vertices, where each of the $\binom{n}{2}$ possible edges is present with probability p, independently of other edges. This implies that for any graph G on n vertices and m edges, we have

$$\mathbb{P}(G(n,p) = G) = p^m(1-p)^{\binom{n}{2}-m}.$$

If we condition on the $G(n,p)$ having m edges, we get that

$$\begin{aligned}\mathbb{P}(G(n,p) = G \mid |E| = m) &= \frac{\mathbb{P}(G(n,p) = G)}{\mathbb{P}(|E| = m)} \\ &= \frac{\mathbb{P}(G(n,p) = G)}{\binom{\binom{n}{2}}{m}\mathbb{P}(G(n,p) = G)} \\ &= \binom{\binom{n}{2}}{m}^{-1} \\ &= \mathbb{P}(G(n,m) = G),\end{aligned}$$

because all graphs on m edges are equally likely in $G(n,p)$.

Exercise 6. *We said that $G(n,p)$ behaves similarly to $G(n,m)$ in most interesting ways, which is a bit of a vague statement. In what ways do you expect these to models to behave similarly, and in what ways do they behave differently?* 1

Exercise 7. *An* isolated vertex *is a vertex that has no neighbors (i.e. is not an endpoint of any edge in E). Let A_n be the event that a graph generated by $G(n,p)$ has an isolated vertex. Show that* 1

$$\mathbb{P}(A_n) \leq n(1-p)^{n-1}.$$

Exercise 8. *Show that for any $\epsilon > 0$, if $p = (1+\epsilon)\frac{\log n}{n}$, then* 2

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 0.$$

We will return to this topic later in the course, when we discuss thresholds for properties of $G(n,p)$. The value $p^* = \frac{\log n}{n}$ is a *threshold* for the disappearance of isolated vertices in the evolution of $G(n,p)$. This also means that if $p = (1-\epsilon)\frac{\log n}{n}$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$.

Exercise 9. *Fix a value of n. Use a computer to run a large number of simulations and create a plot of the average number of isolated vertices in $G(n,p)$ as p ranges from 0 to 1.* 3

Solutions to Selected Exercises

Exercise 1. *Use the inequalities from the first set of notes to show that*

$$\mathbb{E}(|X \cup Y|) \leq np + ne^{-p(\delta+1)}.$$

Then, find the value of p that minimizes this function, and finish the proof of Lemma 6.

Since, $(1+a)^n \leq e^{an}$, for $a > -1,\ n \geq 0$, we have that

$$\mathbb{E}(|X \cup Y|) = np + n(1-p)^{\delta+1} \leq np + ne^{-p(\delta+1)}.$$

Then, we see that

$$\frac{d}{dp}(np + ne^{-p(\delta+1)}) = n - n(\delta+1)e^{-p(\delta+1)},$$

which strictly increases on the interval $0 \leq p \leq 1$, and is equal to 0 at

$$p^* = \frac{\log(1+\delta)}{1+\delta}.$$

Therefore, the smallest value of $np + ne^{-p(\delta+1)}$ is achieved at p^*, giving us the result that every graph G has a dominating set of at most size $\frac{\log(1+\delta)+1}{1+\delta}n$.

Exercise 3. *Prove Theorem 7.*

Generate a random tournament by directing edges uniformly and independently of each other. For a given subset S and vertex $v \in U \setminus S$, we have

$$\mathbb{P}(v \text{ beats } S) = \left(\frac{1}{2}\right)^k.$$

Therefore,

$$\mathbb{P}(\text{no } v \in U \setminus S \text{ beats } S) = \left(1 - \left(\frac{1}{2}\right)^k\right)^{n-k} \qquad \text{(independent events)}$$

Call a set S “"bad" if no $v \in U \setminus S$ beats it. Let X be the number of bad sets. Then

$$\mathbb{E}(\#\text{ bad sets}) = \binom{n}{k}\left(1 - \left(\frac{1}{2}\right)^k\right)^{n-k}.$$

If $\mathbb{E} < 1$ then a tournament with the S_k property exists, by the first moment method.

Exercise 4. *Show that* 1

$$\binom{n}{k}\left(1 - \left(\frac{1}{2}\right)^k\right)^{n-k} < \left(\frac{ne}{k}\right)^k e^{-(n-k)/2^k}.$$

This follows from the inequalities at the top of Notes 1.

Exercise 5. *Show that there exists a constant c such that $f(k) \leq ck^2 2^k$.* 1

We have

$$\mathbb{E}(X) \leq \left(\frac{en}{k}\right)^k e^{-(n-k)/2^k} \leq (1+o(1))n^k e^{-n/2^k},$$

and therefore we need

$$n > 2^k \log n^k = 2^k k \log n,$$

in order to satisfy $\mathbb{E}(X) < 1$. Using the hint, we see that

$$2^k k \log(2^k k^2) = O(2^k k^2),$$

and therefore there is some $c > 0$ such that $n = c2^k k^2$ satisfies $\mathbb{E}(X) < 1$. This gives the result that $f(k) = O(k^2 2^k)$.

Exercise 6. *We said that $G(n,p)$ behaves similarly to $G(n,m)$ in most interesting ways, which is a bit of a vague statement. In what ways do you expect these to models to behave similarly, and in what ways do they behave differently?*

We will see that these models behave very similarly for things such as connectivity and appearance of subgraphs: properties that depend on the approximate density of the graph. One way in which they are very different is for questions such as: what is the probability that G has exactly m edges? Or what is the probability that G has an even number of edges?

Exercise 7. *An* isolated vertex *is a vertex that has no neighbors (i.e. is not an endpoint of any edge in E). Let A_n be the event that a graph generated by $G(n,p)$ has an isolated vertex. Show that*

$$\mathbb{P}(A_n) \leq n(1-p)^{n-1}.$$

Let A_i be the event that vertex i in the random graph $G \sim G(n,p)$ has no neighbors, with $1 \leq i \leq n$. Then, $\mathbb{P}(A_i) = (1-p)^{n-1}$. We can now use the Union Bound to bound the probability of having **any** isolated vertices as

$$\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i) = n(1-p)^{n-1}.$$

Exercise 8. *Show that for any $\epsilon > 0$, if $p = (1+\epsilon)\frac{\log n}{n}$, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 0.$$

From Notes 1, we recall the fact that

$$\left(1 + \frac{a}{n}\right)^{bn} e^{ab}, \text{ as } n \to \infty.$$

This gives

$$\begin{aligned}
\mathbb{P}(\cup_i A_i) &\leq n(1-p)^{n-1} \\
&= n\left(1 - (1+\epsilon)\frac{\log n}{n}\right)^{n-1} \\
&= n\left(1 - \frac{1+\epsilon}{n/\log n}\right)^{n-1} \\
&= n\left(\left(1 - \frac{1+\epsilon}{n/\log n}\right)^{n/\log n}\right)^{((n-1)/n)\log n} \\
&\to ne^{-(1+\epsilon)\log n} = \frac{1}{n^\epsilon} \to 0.
\end{aligned}$$