CS5330: Randomized Algorithms

Problem Set 3—Solutions

Due: February 12, 6:30pm

Instructions. The *exercises* at the beginning of the problem set do not have to be submitted—though you may. They mainly cover topics related to Cuckoo hashing (which we covered in tutorial) and hash functions (which we will be talking about next week). There are four problems, related to different balls-and-bins process, the Coupon Collector problem, etc.

- Please submit the problem set on IVLE in the appropriate folder. (Typing the solution using latex is recommended.) If you want to do the problem set by hand, please submit it at the beginning of class.
- Start each problem on a separate page.
- If you submit the problem set on paper, make sure your name is on each sheet of paper (and legible).
- If you submit the problem set on paper, staple the pages together.

Collaboration Policy. The submitted solution must be your own unique work. You may discuss your high-level approach and strategy with others, but you must then: (i) destroy any notes; (ii) spend 30 minutes on facebook or some other non-technical activity; (iii) write up the solution on your own; (iv) list all your collaborators. Similarly, you may use the internet to learn basic material, but do not search for answers to the problem set questions. You may not use any solutions that you find elsewhere, e.g. on the internet. Any similarity to other students' submissions will be treated as cheating.

Exercises and Review

Exercise 1. Prove the following facts regarding Cuckoo Hashing. (Note: we have already covered these in tutorial; the goal here is to carefully write down a proof.) Assume throughout that $m = 4n$.

- Given a Cuckoo Graph containing n edges, show that any path of length p in the Cuckoo Graph generated by an insert operation that terminates has a simple subpath (with no repeated nodes) of length at least $p/3$.
- Given a Cuckoo Graph containing n edges, show that for all nodes i and j, for all ℓ, the probability that the graph contains a path of length exactly ℓ from i to j is at most $\frac{1}{4^\ell}\frac{1}{m}$.
- Prove that the expected cost of an insert operation is $O(1)$ (as long as the Cuckoo Hash Table contains at most n items).
- Given a Cuckoo Graph containing n edges, prove that the probability it contains a cycle is at most $1/2$.
- Whenever an insert does not terminate (i.e., ends in an infinite loop), you need to create a new Cuckoo Hash Table, choose new hash functions, and reinsert all n items. (If your insert operation continues for $2m$ steps, you can assume you have entered an infinite loop!). Assume that you continue this process, rehashing as necessary, until you successfully insert all n items. Prove that the expected total cost of this process is $O(n)$ (including the cost of the repeated rehashing).

Exercise 2. Consider a version of Cuckoo hashing that only uses one array A, but has two hash functions f and g. Each element x is inserted at either $A[f(x)]$ or $A[g(x)]$. As before during an insert, if a space is occupied, then the old item is kicked out and moved to its new location. Can you modify the existing analysis to work in this setting?

Exercise 3. Think about the variant where you have one array A and k hash functions $f_1, f_2, \ldots, f_k$. Each element x is inserted at some $A[f_j(x)]$, for some j. How would you design an insert algorithm for this variant? How do you decide where to move an item when it is evicted? What if you are allowed to store ℓ items in each slot in the array? What are the trade-offs involved?

Exercise 4. Define the following three random variables:

- A has value 100 with probability $1/2$ and value 200 with probability $1/2$.
- B has value 100 with probability $1/3$, value 120 with probability $1/6$, and value 200 with probability $1/2$.
- C has value 1000 with probability $1/4$ and value 0 otherwise.

Show that B stochastically dominates A, that B does *not* stochastically dominate C, and that C does not stochastically dominate B. (That is, B and C are incomparible.)

Problem 1. [Deploying a Sensor Network.]

Imagine you are deploying a sensor network in a flat, square field that is 1km by 1km. Each sensor has a range $r < 1km$, meaning that it can record all events that occur within distance r of the sensor. The sensors are deployed randomly in the field (e.g., imagine they are dropped from an airplane).

How many sensors should be deployed to ensure that with probability at least $(1 - \epsilon)$, every event in the field can be monitored, i.e., every point in the field is within range of at least one sensor. Assume ϵ is a fixed constant error parameter < 1. (*Hint:* a sensor in a square with side-length $r/\sqrt{2}$ can reach every point in the square.)

Solution: Divide the field into a grid of small squares with edge-length $r/\sqrt{2}$. Notice that if there is at least one sensor in each square, then it can cover the entire square (as the diagonal of the square is of length r). There are at most $2/r^2$ squares in the field, and each sensor is equally likely to be deployed to each square. Let $n = 2/r^2$ denote the number of squares.

Thus, by the coupon collector's analysis, we need $O(n \log n) = O((2/r^2) \log(2/r^2)) = O(\frac{1-\log r}{r^2})$ sensors to monitor the field with high probability.

To get a probability in terms of ϵ, notice that the probability of a square being empty is at most $(1 - 1/n)^m \leq e^{-m/n}$, where m is the number of sensors being deployed. If we set $m = n \log(n/\epsilon)$, then $e^{-m/n} \leq e^{-\log(n/\epsilon)} \leq \epsilon/n$. By taking a union bound over the n squares, we see that the probability of any square being empty is at most ϵ.

Thus, since $\log(2/(\epsilon r^2)) = 1 - 2\log(\epsilon r)$, we need $n \log(n/\epsilon) = (1 - 2\log(\epsilon r))/r^2)$ sensors to ensure coverage with probability at least $1 - \epsilon$.

Problem 2. [Random Graphs.]
Assume you build a random graph $G(n, m)$ with n vertices and m edges where the edges are chosen uniformly at random from the set of all possible edges. (That is, the graph includes a random subset of the $\binom{n}{2}$ possible edges.) Via the principle of deferred decisions, we can imagine that you construct the graph as follows:

Repeat until the graph is connected:

- Choose a node u at random.
- Choose a node v at random.
- If (u, v) is not an edge in the graph, add edge (u, v) to the graph.

What is the expected number of edges you have to sample before the graph G becomes connected? Argue that from this you can derive a size m so that a random graph with m edges has at least a probability $1/2$ of being connected. (*Hint:* think about the graph in terms of its connected components, and apply the Coupon Collector's technique.)

Solution: This problem is quite similar to the Coupon Collector's analysis. Let $X_1, X_2, \ldots$ be random variables representing the number of connected components, where X_j is the number of connected components before the addition of edge j. Thus, X_1 is the number of connected components prior to adding any edges, i.e., $X_1 = n$.

We divide the sequence of random variables into segments that have equal value, i.e., let $S_k = \{X_j | X_j = k\}$. Let us focus our attention on a particular set S_k, i.e., the sequence of edges added to the graph while there were k connected components. Let $Y_1, Y_2, \ldots$ be indicator random variables specifying whether each edge in S_k connects two existing connected components. (Notice that if $Y_j = 1$, then edge j is the last edge in S_k, and $|S_k| = j$.)

We need to calculate $\Pr[Y_j]$. We can think of each edge as being chosen as follows: first we choose one endpoint v, and then we choose a second endpoint w. The probability $\Pr[Y_j = 1]$ is equal to the probability that the second endpoint w is in a different component as w. Since there are k connected components, we know that there are at least $k-1$ possible endpoints for w (out of the $n-1$ possible nodes) that would connect v to a new components. That is, $\Pr[Y_j = 1] \geq (k-1)/(n-1) \geq (k-1)/n$.

From this, we conclude the the expected number of edges that can be added before connecting two components is $\leq n/(k-1)$. That is, $\text{Exp}[|S_k|] \leq n/(k-1)$. By linearity of expectation, we conclude that $\text{Exp}\left[\sum_{k=n}^{2} |S_k|\right] \leq \sum_{k=2}^{n} \frac{n}{k-1} = O(n \log n)$.

Notice that one must be careful with the direction of the inequalities. If the connected components are large, then the probability of connecting two components may be larger than $(k-1)/n$; this is only providing an upper bound.

Problem 3. Group Assignment.

Professor Unfriendly wants to group students in his class in such a way that no pair of friends are in the same group. Luckily, Professor Unfriendly has access to the Facebook friend graph, and can tell who are friends. He has access to a graph $G = (V, E)$ where V is the set of n students in the class and each edge $e \in E$ indicates a pair of friends. The maximum degree of the graph is Δ, and Professor Unfriendly wants to create at most $T = 2\Delta$ groups.

The professor runs the following algorithm:

Repeat until every student is assigned a group:

- Iterate through all the students in order:
  1. Choose a group p uniformly at random from $[1, T]$.
  2. If the current student is not yet assigned to a group, and if none of his/her friends are in group p, then assign the student to group p.
  3. Otherwise, skip the current student and continue with the next one.

Your goal in this problem is to analyze the running time of this algorithm. To do that, we will count the number of times we execute the inner loop (i.e., choosing a group and trying to assign it).

Let X_j be an indicator random variable defined by the jth iteration of the inner loop: if the student in the jth iteration is assigned a group, then $X_j = 1$; otherwise $X_j = 0$. As soon as $\sum X_j = n$, we know that every student has been assigned a group.

Beware, though, that it is not easy to compute $\Pr[X_j = 1]$, since it depends on the outcome of all the previous random choices. (For example, for the first student to be assigned a group, the probabiity is 1, i.e., $\Pr[X_1] = 1$. For the last student to be assigned a group, the probability is lower especially if they have a lot of friends!) So the X_j are not independent!

Problem 3.a. Define a new (independent) set of random variables $Y_1, Y_2, \ldots$ so that X_j stochastically dominates Y_j. (We will use the Y's to bound the running time, so we want them to be less likely to occur than the X's.)

Give the definition of Y_j and prove carefully (using the definition of stochastic domination) that X_j dominates Y_j.

Solution: Define $Y_j = 1$ with probability $1/2$ and $Y_j = 0$ with probability $1/2$. To show that Y_j stochastically dominates X_j, we need to show that for all values k, $\Pr[X_j \geq k] \geq \Pr[Y_j \geq k]$. We now consider the different cases:

- $k = 0$: $\Pr[X_j \geq 0] = 1$, and also $\Pr[Y_j \geq 0] = 1$. Thus the desired inequality holds.
- $k = 1$: $\Pr[Y_j \geq 1] = 1/2$, by definition. For X_j, notice that the student in question has at most $T/2$ friends, and the group is chosen from the set $[1, T]$. Therefore, we know that with probability at least $1/2$, the student is successfully assigned a group different from her friends. That is, $\Pr[X_j \geq 1] \geq 1/2 \geq \Pr[Y_j \geq k]$.
- $k \geq 2$: In this case, $\Pr[X_j = k] = \Pr[Y_j = k] = 0$, i.e., this is impossible.

Thus we have shown that the Y_j variables are stochastically dominated by the X_j variables.

Problem 3.b. Let $t = 2n \log n$, and define $Y = \sum_{j=1}^{t} Y_j$. Show that $Y \geq n$ with probability at least $1 - 1/n$. (If you cannot prove this, then you might want to revisit your definition of Y_j.)

Solution: There are several ways to show this, perhaps the most common being to just write out the binomial distribution and approximate. (Notice that, given how we have defined the Y_j variables, this is the same as showing that if you flip a coin $2n \log n$ times, then with probability you get at least n heads.) An alternate solution is to break the sequence up into collections of $2 \log n$ Y_j variables. The probability that all the Y_j in such a collection are 0 is at most $1/2^{2 \log n} = 1/n^2$. There are n such collections, so the probability that *any* collection is all 0 is at most $n/n^2 = 1/n$, by a union bound. Thus with probability at least $1 - 1/n$, each collection has at least one 1 and hence the sum is at least n.

Problem 3.c. Now conclude the proof by showing that, with high probability, the group assignment algorithm completes within time $O(n \log n)$.

You may assume the following fact:

- We define the term *unconditionally sequentially dominates* as follows: A sequence of random variables $(X_1, \ldots, X_n)$ *unconditionally sequentially dominates* another sequence of random variables $(Y_1, ..., Y_n)$ if for each j, $(X_j|$ arbitrary $X_1, ..., X_{j-1})$ stochastically dominates Y_j, i.e., if each X_j stochastically dominates Y_j, regardless of the outcome of all the previous $X_{j-1}, X_{j-2}, \ldots$ (i.e., unconditionally).
- If $X_1, ..., X_n$ are an arbitrary set of (discrete) random variables, and $Y_1, ..., Y_n$ are independent (discrete) random variables, if $(X_1, ..., X_n)$ unconditionally sequentially dominates $(Y_1, ..., Y_n)$ then $\sum(X_j)$ stochastically dominates $\sum(Y_j)$.

Solution: First, we note that the X_j unconditionally sequentially dominate the Y_j, since regardless of the other random variables, we see that each X_j has a probability of at least $1/2$ of being true.

As before, define $t = 2n \log n$, and let $X = \sum_{j=1}^{t} X_j$ and $Y = \sum_{j=1}^{t} Y_j$. By the fact stated above, we know that X stochastically dominates Y. That means that $\Pr[X \geq n] \geq \Pr[Y \geq n] \geq 1 - 1/n$. Thus we know that within the first t iterations of the loop, at least n students are assigned to groups (i.e., all of them), and so the algorithm terminates. Therefore, we conclude that the algorithm runs in $O(n \log n)$ time with high probability.

Problem 4. Contention Resolution.

One of the major problems in distributed and parallel systems is contention resolution: there are a collection of agents that want to coordinate access to a shared resource. For example, the resource may be an ethernet connection or a wireless channel (where only one device can broadcast a message at a time). Or the shared resource may be a lock that multiple concurrent threads are trying to access in order to update a data structure.

A typical approach is to use a randomized strategy: when a device wants the resource, it randomly chooses one of the next T timeslots at random and tries to claim the resource in that randomly chosen slot. If it fails, it waits until all T slots elapse, and repeats the procedure. (The T slots are often referred to as the "window" and in a backoff protocol, the size of the window may be adjusted dynamically.)

Here we model this strategy as a simple balls and bins problem. Imagine you have n balls (which represents requests to use the shared resource) and b bins (which represent timeslots in the window). We play the following game:

> Repeat until all the balls have been removed:
>
> - Place each remaining ball in a bin chosen uniformly at random.
> - Every ball that lands in a bin by itself (with no other balls) is removed. (Since this ball/request was the only one in the bin/timeslot, it can safely access the resource during that timeslot.)
> - All the balls that lands in a bin containing more than one ball are collected and advance to the next round.

Intuitively, as long as b is sufficiently bigger than n, then in each round of the game we remove many of the balls and make progress toward completing the game. Our goal in this problem is to calculate the expected number of rounds until all the balls have been removed.

Assume throughout this problem that the initial number of balls $n \leq b/8$. Let n_j be the number of balls that remain after round j.

Continued on the next page.

Problem 4.a. Show that:

$$\mathrm{E}\left[n_j \mid n_{j-1} \leq x\right] \leq \frac{2x^2}{b} \ .$$

(*Hint:* remember that if $x \leq 1$, then $e^{-2} \leq (1-1/x)^x \leq e^{-1}$; identically, for $x \geq 1$, $e^{-2x} \leq (1-x) \leq e^{-x}$.)

Solution: The probability of a ball being removed is at least

$$(1-1/b)^{x-1} \geq (1-1/b)^x \geq e^{-2x/b} \geq (1-2x/b) \ .$$

Hence the expected number of balls that remain is $x(1-(1-2x/b)) = 2x^2/b$.

Problem 4.b. In the previous part, you showed that in expectation the number of balls decreases rapidly in each round. We now want to calculate the probability that we make progress. (Remember, in some rounds we may do better than the expectation; in some rounds we may do worse!)

Let I_j be the event that the number of remaining balls is at most $n/2^{2^j}$ at the end of a round. What is the probability that event I_0 does not occur in a round (if it has not previously occurred)? Show that I_0 occurs with probability at least $1/2$ (if it hasn't occurred previously).

Solution: Initially $n_0 = n \leq b/8$. As we showed in the previous part:

$$\mathrm{E}\left[n_1 | n_0 = n\right] \leq 2(n_0)^2/b \leq 2n(b/8)/b \leq n/4 \ .$$

Hence, by Markov's Inequality, the probability that I_0 does not occur at the end of the first round is $\Pr\left[n_1 \geq n/2 | n_0 = n\right] \leq 1/2$. The same is true for all the following rounds where I_0 has not occurred.

Problem 4.c. Given that event I_{j-1} has already occurred (and I_j has not already occurred), what is the probability of I_j not occuring in a round? Show that I_j occurs with probability at least $1/4$ (if it hasn't occurred previously, and I_{j-1} has occurred previously).

Solution: Let n' be the number of nodes at the beginning of the round t in question. Since I_{j-1} has occurred, we know that $n' \leq n/2^{2^{j-1}}$. From the previous part, we know that:

$$\mathrm{E}\left[n_t | n_{t-1} = n'\right] \leq 2n'^2/b \leq 2(n/2^{2^{j-1}})^2/b \leq 2n(b/8)(1/2^{2^j})/b \leq (1/4)n/2^{2^j} \ .$$

Thus by Markov's Inequality, the probability that I_j does not hold after round t is at most $\Pr\left[n_t \geq n/2^{2^j} | n_{t-1} = n'\right] \leq 1/4$.

Problem 4.d. What is the expected number of rounds until the game is over? (Hint: define X_j to be the number of rounds after I_{j-1} until I_j occurs.)

Solution: Define X_j as described, and observe that the game is over when event $I_{\log\log(n)}$ occurs, which implies that all n balls have been removed. Thus the expected time until the game is done is equal to $\sum_{j=1}^{\log\log n} X_j$. (This is similar to the coupon collector's analysis.)

We have already proven that $\mathrm{E}[X_j] \leq 2$. (In fact, aside from X_1, it is $\leq 4/3$.) The total time is $\sum_{j=1}^{\log\log n} X_j$, and so the expected running time is at most $2\log\log n$.