CS 70 Discrete Mathematics and Probability Theory
Fall 2020 HW 13

1 Random Cuckoo Hashing

Cuckoo birds are parasitic beasts. They are known for hijacking the nests of other bird species and evicting the eggs already inside. Cuckoo hashing is inspired by this behavior. In cuckoo hashing, when we get a collision, the element that was already there gets evicted and rehashed.

We study a simple (but ineffective, as we'll see) version of cuckoo hashing, where all hashes are random. Let's say we want to hash n pieces of data $d_1, d_2, \ldots, d_n$ into n possible hash buckets labeled $1, \ldots, n$. We hash the $d_1, \ldots, d_n$ in that order. When hashing d_i, we assign it a random bucket chosen uniformly from $1, \ldots, n$. If there is no collision, then we place d_i into that bucket. If there is a collision with some other d_j, we evict d_j and assign it another random bucket uniformly from $1, \ldots, n$. (It is possible that d_j gets assigned back to the bucket it was just evicted from!) We again perform the eviction step if we get another collision. We keep doing this until there is no more collision, and we then introduce the next piece of data, d_{i+1} to the hash table.

(a) What is the probability that there are no collisions over the entire process of hashing $d_1, \ldots, d_n$ to buckets $1, \ldots, n$? What value does the probability tend towards as n grows very large?

(b) Assume we have already hashed $d_1, \ldots, d_{n-1}$, and they each occupy their own bucket. We now introduce d_n into our hash table. What is the expected number of collisions that we'll see while hashing d_n? (*Hint*: What happens when we hash d_n and get a collision, so we evict some other d_i and have to hash d_i? Are we at a situation that we've seen before?)

(c) Generalize the previous part: Assume we have already hashed $d_1, \ldots, d_{k-1}$ successully, where $1 \le k \le n$. Let C_k be the number of collisions that we'll see while hashing d_k. What is $\mathbb{E}[C_k]$?

(d) Let C be the total number of collisions over the entire process of hashing $d_1, \ldots, d_n$. What is $\mathbb{E}[C]$? You may leave your answer as a summation.

Solution:

(a) When hashing d_i, there are $(n-i+1)$ empty buckets, as $(i-1)$ of them are already occupied by $d_1, \ldots, d_{i-1}$. If we want no collisions over this entire hashing process, we must choose an empty bucket on the first go for each d_i. This gives:

$$\mathbb{P}[\text{no collisions}] = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{1}{n} = \frac{n!}{n^n}$$

To understand what happens as n grows very large, we can upper bound the probability as follows:

$$\mathbb{P}[\text{no collisions}] = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{1}{n} \le 1 \cdot \ldots \cdot 1 \cdot \frac{1}{n} = \frac{1}{n}$$

We are upper bounding each term in the product above by 1, except the very last term, which we leave as $\frac{1}{n}$. When n is large, this upper bound goes to 0, so $\mathbb{P}[\text{no collisions}]$ will also tend to 0.

Another way to obtain the $\frac{n!}{n^n}$ probability is to see that considering the first bucket to which each datum gets hashed is a uniform sample space with size n^n. The number of sample points in our event (no collisions) is the number of ways of assigning each datum a unique bucket to be placed in, i.e. the number of ways to permute the datum within the buckets, or $n!$.

(b) Let C_n be the number of collisions experienced when hashing a single datum into a table with $(n-1)$ buckets already populated. (Note that we don't specify that we hash d_n in particular when defining C.)

First, it is possible that we end with 0 collisions. This happens with probability $\frac{1}{n}$. Otherwise, we get a collision, and we have to evict some other datum d_i. Now, we are back in the original situation; the number of collisions experienced after re-hashing d_i is also C because we are again in the situation of introducing a single datum into a table with $(n-1)$ buckets already populated. However, we do need to count the fact that we already had one collision–the one that evicted d_i. This gives us:

$$\mathbb{E}[C_n] = 0 \cdot \frac{1}{n} + (\mathbb{E}[C_n]+1) \cdot \frac{n-1}{n}$$

Solving for $\mathbb{E}[C_n]$ above, we get an expected $(n-1)$ collisions.

Remark: It is also perfectly valid to use an infinite sum based solution.

(c) We take a similar approach to the previous part. Let C_k be the number of collisions experienced when hashing a single datum into a table with $(k-1)$.

When we hash d_k we have probability $\frac{n-(k-1)}{n}$ of not getting a collision and finishing the process with 0 collisions. Otherwise, we evict some other datum and are left with the same situation. This gives us:

$$\mathbb{E}[C_k] = 0 \cdot \frac{n-k+1}{n} + (\mathbb{E}[C_k]+1) \cdot \frac{k-1}{n}$$

Solving for $\mathbb{E}[C_k]$ above, we get an expected $\frac{k-1}{n-k+1}$ collisions.

(d) Let C_k be the random variable denoting number of collisions which occur while hashing the k-th datum, d_k. Let C be the total number of collisions which occur over the entire process. That is, $C = C_1 + C_2 + \cdots + C_n$. Then we have:

$$\mathbb{E}[C] = \mathbb{E}\left[\sum_{k=1}^{n} C_k\right] = \sum_{k=1}^{n} \mathbb{E}[C_k] = \sum_{k=1}^{n} \frac{k-1}{n-k+1} = \sum_{k=0}^{n-1} \frac{k}{n-k}$$

The second step uses linearity of expectation, and the third step makes use of the result from the previous part.

2 Geometric and Poisson

Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Poisson}(\lambda)$ be independent. random variables. Compute $\mathbb{P}(X > Y)$. Your final answer should not have summations.

Solution: We condition on Y so we can use the nice property of geometric random variables that $\mathbb{P}(X > k) = (1-p)^k$, this gives

$$\begin{aligned}
P(X > Y) &= \sum_{y=0}^{\infty} P(X > Y | Y = y) \cdot P(Y = y) \\
&= \sum_{y=0}^{\infty} (1-p)^y \cdot \frac{e^{-\lambda}\lambda^y}{y!} \\
&= e^{-\lambda p} e^{\lambda p} \sum_{y=0}^{\infty} \frac{e^{-\lambda}(\lambda(1-p))^y}{y!} \\
&= e^{-\lambda p} \sum_{y=0}^{\infty} \frac{e^{-\lambda(1-p)}(\lambda(1-p))^y}{y!} \\
&= e^{-\lambda p}
\end{aligned}$$

To simplify the last summation we observed that the sum could be interpreted as the sum of the probabilities for a Poisson$(\lambda(1-p))$ random variable, which is equal to 1.

3 Exploring the Geometric Distribution

Suppose $X \sim \text{Geometric}(p)$ and $Y \sim \text{Geometric}(q)$ are independent. Find the distribution of $\min\{X,Y\}$ and justify your answer.

Solution:

X is the number of coins we flip until we see a heads from flipping a coin with bias p, and Y is the same as flipping a coin with bias q. Imagine we flip the bias p coin and the bias q coin at the same time. The min of the two random variables represents how many simultaneous flips occur before at least one head is seen.

The probability of not seeing a head at all on any given simultaenous flip is $(1-p)(1-q)$, so the probability that there will be a success on any particular trial is $p+q-pq$. Therefore, $\min\{X,Y\} \sim \text{Geometric}(p+q-pq)$.

We can also solve it algebraically. The probability that $\min\{X,Y\} = k$ for some positive integer k is the probability that the first $k-1$ coin flips for both X and Y were tails, then times the probability that we get heads on the k-th toss. Specifically,

$$((1-p)(1-q))^{k-1} \cdot (p+q-pq)$$

We recognize this as the formula for a geometric random variable with parameter $p+q-pq$.

4 Lunch Meeting

Alice and Bob agree to try to meet for lunch between 12 PM and 1 PM at their favorite sushi restaurant. Being extremely busy, they are unable to specify their arrival times exactly, and can say only that each of them will arrive (independently) at a time that is uniformly distributed within the hour. In order to avoid wasting precious time, if the other person is not there when they arrive they agree to wait exactly fifteen minutes before leaving. What is the probability that they will actually meet for lunch? (hint: Sketch the joint distribution of the arrival times of Alice and Bob. What parts of the distribution corresponds to them meeting for lunch?)

Solution:

Let the random variable A be the time that Alice arrives and the random variable B be the time when Bob arrives. Since A and B are both uniformily distributed, it is helpful to vizualize the distributiion graphically. Consider Figure 1, plotting the space of all outcomes (a,b): The arrival times are uniformily distributed over the box. The shaded region is the set of values (a,b) for which Alice and Bob will actually meet for lunch. Since all points in this square are equally likely, the probably they meet is the ratio of the shaded area to the area of the square. If the area of the square is 1, then the area of the shaded region is

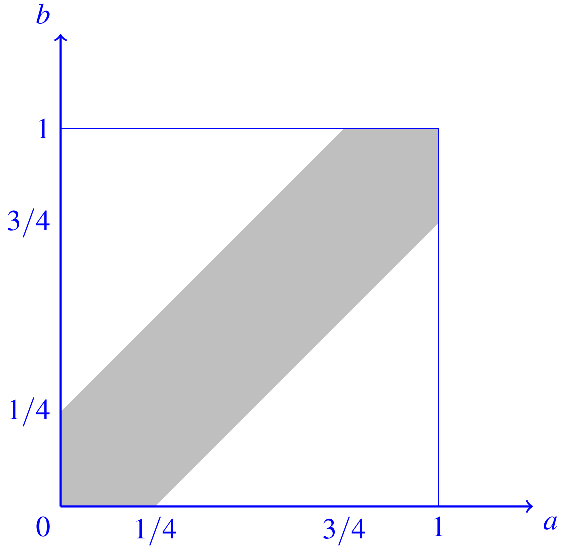

Figure 1: Visualization of joint probability density.

$$1 - 2 \times \left[\frac{1}{2} \times \left(\frac{3}{4}\right)^2\right] = \frac{7}{16},$$

since the area of the white triangle on the upper-left is $(1/2) \cdot (3/4)^2$, and the white triangle on the lower-right has the same area. Therefore, the probability that Alice and Bob actually meet is $7/16$.