CS 70 Discrete Mathematics and Probability Theory
Fall 2018 Alistair Sinclair and Yun Song

HW 11

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?)

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:

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

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.

(b) Let C 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:

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

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

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

2 Markov's Inequality and Chebyshev's Inequality

A random variable X has variance $\text{var}(X) = 9$ and expectation $\mathbb{E}[X] = 2$. Furthermore, the value of X is never greater than 10. Given this information, provide either a proof or a counterexample for the following statements.

(a) $\mathbb{E}[X^2] = 13$.

(b) $\mathbb{P}[X \leq 1] \leq 8/9$.

(c) $\mathbb{P}[X \geq 6] \leq 9/16$.

(d) $\mathbb{P}[X \geq 6] \leq 9/32$.

Solution:

(a) TRUE. Since $9 = \text{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - 2^2$, we have $\mathbb{E}[X^2] = 9+4 = 13$.

(b) TRUE. Let $Y = 10 - X$. Since X is never exceeds 10, Y is a non-negative random variable. By Markov's inequality,

$$\mathbb{P}[10 - X \geq a] = \mathbb{P}[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a} = \frac{\mathbb{E}[10-X]}{a} = \frac{8}{a}.$$

Setting $a = 9$, we get $\mathbb{P}[X \leq 1] = \mathbb{P}[10 - X \geq 9] \leq 8/9$.

(c) TRUE. Chebyshev's inequality says $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \text{var}(X)/a^2$. If we set $a = 4$, we have

$$\mathbb{P}[|X-2| \geq 4] \leq \frac{9}{16}.$$

Now we observe that $\mathbb{P}[X \geq 6] \leq \mathbb{P}[|X-2| \geq 4]$, because the event $X \geq 6$ is a subset of the event $|X-2| \geq 4$.

(d) FALSE. Construct a random variable X that satisfies the conditions in the question but does not have an equal chance of being less than -2 or greater than 6. A simple example would be a random variable that takes on 2 values, where $\mathbb{P}[X = a] = p, \mathbb{P}[X = b] = 1 - p$. The expectation must be 2, so we have $pa + (1-p)b = 2$. The variance is 9, so $\mathbb{E}[X^2] = 13$ and $pa^2 + (1-p)b^2 = 13$. Solving for a and b. One example is $\mathbb{P}[X = 0] = 9/13, \mathbb{P}[X = 13/2] = 4/13$.

3 Easy A's

A friend tells you about a course called "Laziness in Modern Society" that requires almost no work. You hope to take this course next semester to give yourself a well-deserved break after working hard in CS 70. At the first lecture, the professor announces that grades will depend only on two homework assignments. Homework 1 will consist of three questions, each worth 10 points, and Homework 2 will consist of four questions, also each worth 10 points. He will give an A to any student who gets at least 60 of the possible 70 points.

However, speaking with the professor in office hours you hear some very disturbing news. He tells you that, in the spirit of the class, the GSIs are very lazy, and to save time the grading will be done as follows. For each student's Homework 1, the GSIs will choose an integer randomly from a distribution with mean $\mu = 5$ and variance $\sigma^2 = 1$. They'll mark each of the three questions with that score. To grade Homework 2, they'll again choose a random number from the same distribution, independently of the first number, and will mark all four questions with that score.

If you take the class, what will the mean and variance of your total class score be? Use Chebyshev's inequality to conclude that you have less than a 5% chance of getting an A when the grades are randomly chosen this way.

Solution:

Let X be the total number of points you receive in the class. Then $X = X_1 + X_2$ where X_1 is the number points received on Homework 1 and X_2 is the number of points received on Homework 2. Your Homework 1 score is generated as $X_1 = 3Y_1$, where the r.v. Y_1 represents the integer that the GSI chose when grading it. Similarly, $X_2 = 4Y_2$. The problem statement tells us that Y_1 and Y_2 are independent, both with mean 5 and variance 1, so $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 5$ and $\text{var}(Y_1) = \text{var}(Y_2) = 1$. Thus,

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 3\,\mathbb{E}[Y_1] + 4\,\mathbb{E}[Y_2] = 35,$$
$$\text{var}(X) = \text{var}(X_1) + \text{var}(X_2) = 9\,\text{var}(Y_1) + 16\,\text{var}(Y_2) = 25.$$

Using Chebyshev's Inequality, we get

$$\mathbb{P}[X \geq 60] \leq \mathbb{P}[|X - 35| \geq 25] \leq \frac{\text{var}(X)}{25^2} = \frac{1}{25}.$$

Unfortunately, any student will have at most a 4% chance of getting an A.

Note that although we calculated a bound for $\mathbb{P}[|X-35| \geq 25]$, which is the probability that you will get 60 or above or 10 or below, we cannot divide by 2 to refine our bound unless the distribution is symmetric about its mean. In this case, the distribution is not symmetric.

4 Confidence Interval Introduction

We observe a random variable X which has mean μ and standard deviation $\sigma \in (0,\infty)$. Assume that the mean μ is unknown, but σ is known.

We would like to give a 95% confidence interval for the unknown mean μ. In other words, we want to give a random interval (a,b) (it is random because it depends on the random observation X) such that the probability that μ lies in (a,b) is at least 95%.

We will use a confidence interval of the form $(X-\varepsilon, X+\varepsilon)$, where $\varepsilon > 0$ is the width of the confidence interval. When ε is smaller, it means that the confidence interval is narrower, i.e., we are giving a more *precise* estimate of μ.

(a) Using Chebyshev's Inequality, calculate an upper bound on $\mathbb{P}\{|X-\mu| \geq \varepsilon\}$.

(b) Explain why $\mathbb{P}\{|X-\mu| < \varepsilon\}$ is the same as $\mathbb{P}\{\mu \in (X-\varepsilon, X+\varepsilon)\}$.

(c) Using the previous two parts, choose the width of the confidence interval ε to be large enough so that $\mathbb{P}\{\mu \in (X-\varepsilon, X+\varepsilon)\}$ is guaranteed to exceed 95%.

[Note: Your confidence interval is allowed to depend on X, which is observed, and σ, which is known. Your confidence interval is not allowed to depend on μ, which is unknown.]

(d) The previous three parts dealt with the case when you observe one sample X. Now, let n be a positive integer and let $X_1, \ldots, X_n$ be i.i.d. samples, each with mean μ and standard deviation $\sigma \in (0,\infty)$. As before, assume that μ is unknown but σ is known.

Here, a good estimator for μ is the *sample mean* $\bar{X} := n^{-1}\sum_{i=1}^{n} X_i$. Calculate the mean and variance of $\bar{X}$.

(e) We will now use a confidence interval of the form $(\bar{X}-\varepsilon, \bar{X}+\varepsilon)$ where $\varepsilon > 0$ again represents the width of the confidence interval. Imitate the steps of (a) through (c) to choose the width ε to be large enough so that $\mathbb{P}\{\mu \in (\bar{X}-\varepsilon, \bar{X}+\varepsilon)\}$ is guaranteed to exceed 95%.

To check your answer, your confidence interval should be *smaller* when n is larger. Intuitively, if you collect more samples, then you should be able to give a more *precise* estimate of μ.

Solution:

(a) Since $\mathbb{E}[X] = \mu$ and $\operatorname{var} X = \sigma^2$, then by Chebyshev's Inequality,

$$\mathbb{P}\{|X-\mu| \geq \varepsilon\} \leq \frac{\operatorname{var} X}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

(b) Note that $|X-\mu|<\varepsilon$ if and only if $-\varepsilon<X-\mu<\varepsilon$, if and only if $\mu-\varepsilon<X<\mu+\varepsilon$. However, the first inequality says that $\mu<X+\varepsilon$ and the second inequality says that $\mu>X-\varepsilon$, that is, $X-\varepsilon<\mu<X+\varepsilon$, which is the same thing as saying $\mu\in(X-\varepsilon,X+\varepsilon)$. So, the events $\{|X-\mu|<\varepsilon\}$ and $\{\mu\in(X-\varepsilon,X+\varepsilon)\}$ are identical.

(c) We want $\mathbb{P}\{\mu\in(X-\varepsilon,X+\varepsilon)\}\geq 0.95$, which is equivalent to

$$\mathbb{P}\{|X-\mu|\geq\varepsilon\}=1-\mathbb{P}\{|X-\mu|<\varepsilon\}=1-\mathbb{P}\{\mu\in(X-\varepsilon,X+\varepsilon)\}\leq 0.05.$$

However, we have the bound $\mathbb{P}\{|X-\mu|\geq\varepsilon\}\leq\sigma^2/\varepsilon^2$, so we just need to choose ε big enough so that $\sigma^2/\varepsilon^2\leq 0.05$. To do this, we want $\varepsilon^2\geq 20\sigma^2$, or $\varepsilon\geq\sqrt{20}\sigma\approx 4.47\sigma$. Our confidence interval is therefore $(X-4.47\sigma,X+4.47\sigma)$.

(d) For the mean, use linearity of expectation. We have

$$\mathbb{E}[\bar{X}]=\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}X_i\Big]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]=\frac{1}{n}\sum_{i=1}^{n}\mu=\frac{1}{n}\cdot n\mu=\mu.$$

For the variance, recall two facts. One is that for a constant c, the scaling of the variance is $\operatorname{var}(cX)=c^2\operatorname{var}X$. The second fact is that $X_1,\ldots,X_n$ are independent, so they are pairwise uncorrelated, that is, for any distinct $i,j\in\{1,\ldots,n\}$, $\operatorname{cov}(X_i,X_j)=0$; this implies that $\operatorname{var}(\sum_{i=1}^n X_i)=\sum_{i=1}^n\operatorname{var}X_i$. Using these facts,

$$\operatorname{var}\bar{X}=\operatorname{var}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i\Big)=\frac{1}{n^2}\operatorname{var}\Big(\sum_{i=1}^{n}X_i\Big)=\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}X_i=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}.$$

(e) By Chebyshev's Inequality,

$$\mathbb{P}\{|\bar{X}-\mu|\geq\varepsilon\}\leq\frac{\operatorname{var}\bar{X}}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}.$$

We want $\sigma^2/(n\varepsilon^2)\leq 0.05$, and to do this, we choose $\varepsilon^2\geq 20\sigma^2/n$, or $\varepsilon\geq\sqrt{20}\sigma/\sqrt{n}$.