Asmt 1: Hash Functions and PAC Algorithms

Turn in (**a pdf**) through Canvas by 2:45pm:
Wednesday, January 15

Overview

In this assignment you will experiment with random variation over discrete events.

It will be very helpful to use the analytical results and the experimental results to help verify the other is correct. If they do not align, you are probably doing something wrong (this is a very powerful and important thing to do whenever working with real data).

As usual, it is recommended that you use LaTeX for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory: `http://www.cs.utah.edu/~jeffp/teaching/latex/`

1 Birthday Paradox (35 points)

Consider a domain of size $n = 5000$.

A: (5 points) Generate random numbers in the domain $[n]$ until two have the same value. How many random trials did this take? We will use k to represent this value.

B: (10 points) Repeat the experiment $m = 300$ times, and record for each time how many random trials this took. Plot this data as a *cumulative density plot* where the x-axis records the number of trials required k, and the y-axis records the fraction of experiments that succeeded (a collision) after k trials. The plot should show a curve that starts at a y value of 0, and increases as k increases, and eventually reaches a y value of 1.

C: (10 points) Empirically estimate the expected number of k random trials in order to have a collision. That is, add up all values k, and divide by m.

D: (10 points) Describe how you implemented this experiment and how long it took for $m = 300$ trials.

Show a plot of the run time as you gradually increase the parameters n and m. (For at least 3 fixed values of m between 300 and 10,000, plot the time as a function of n.) You should be able to reach values of $n =$ 1,000,000 and $m =$ 10,000.

2 Coupon Collectors (35 points)

Consider a domain $[n]$ of size $n = 300$.

A: (5 points) Generate random numbers in the domain $[n]$ until every value $i \in [n]$ has had one random number equal to i. How many random trials did this take? We will use k to represent this value.

B: (10 points) Repeat step **A** for $m = 400$ times, and for each repetition record the value k of how many random trials we required to collect all values $i \in [n]$. Make a cumulative density plot as in **1.B**.

C: (10 points) Use the above results to calculate the empirical expected value of k.

D: (10 points) Describe how you implemented this experiment and how long it took for $n = 300$ and $m = 400$ trials.

Show a plot of the run time as you gradually increase the parameters n and m. (For at least 3 fixed values of m between 400 and 5,000, plot the time as a function of n.) You should be able to reach $n = 20{,}000$ and $m = 5{,}000$.

3 Comparing Experiments to Analysis (30 points)

A: (15 points) Calculate analytically (using formulas from the notes in **L2** or M4D book) the number of random trials needed so there is a collision with probability at least 0.5 when the domain size is $n = 5000$. There are a few formulas stated with varying degree of accuracy, you may use any of these – the more accurate formula, the more sure you may be that your experimental part is verified, or is not (and thus you need to fix something).

[Show your work, including describing which formula you used.]

How does this compare to your results from **1.C**?

B: (15 points) Calculate analytically (using formulas from the notes in **L2** or M4D book) the expected number of random trials before all elements are witnessed in a domain of size $n = 300$? Again, there are a few formulas you may use – the more accurate, the more confidence you might have in your experimental part.

[Show your work, including describing which formula you used.]

How does this compare to your results from **2.C**?

4 BONUS : PAC Bounds (2 points)

Consider a domain size n and let k be the number of random trials run, where each trial obtains each value $i \in [n]$ with probability $1/n$. Let f_i denote the number of trials that have value i. Note that for each $i \in [n]$ we have $\mathbf{E}[f_i] = k/n$. Let $\mu = \max_{i \in [n]} f_i/k$.

Consider some parameter $\varepsilon \in (0, 1)$. As a function of parameter ε, how large does k need to be for $\mathbf{Pr}[|\mu - 1/n| \geq \varepsilon] \leq 0.05$? That is, how large does k need to be for *all* counts to be within $(\varepsilon \cdot 100)\%$ of the average with probability 0.05? *(Fine print: you don't need to calculate this exactly, but describe a bound as a function of ε for the value k which satisfies PAC property. Chapter 2.3 in the M4D book should help.)*

How does this change if we want $\mathbf{Pr}[|\mu - 1/n| \geq \varepsilon] \leq 0.005$ (that is, only 0.005 probability of exceeding ε error)?

[Make sure to show your work.]