CS 388R: Randomized Algorithms Fall 2017

Lecture 7 — 09/21/, 2017

Prof. Eric Price *Scribe: Isidoros Tziotis, Nathan Guermond*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture we covered how to throw balls into bins with two choices.

In this lecture we begin by the problem of approximating the mean of an unknown distribution by sampling, and then we turn our attention towards Hash Tables and specifically Cuckoo Hashing and some of its properties. Cuckoo Hashing takes constant time for lookup and delete in the worst case and constant expected time for insertion. On the other hand it requires linear space.

2 Approximating the mean

Consider the following problem– we have an unknown distribution $\mathcal{D}$ over $\mathbb{R}$ with an unknown mean μ and variance σ^2. The goal is determine an approximation $\hat{\mu}$ for μ by sampling, so that

$$\mathbb{P}[|\hat{\mu} - \mu| \leq \epsilon\sigma] \geq 1 - \delta$$

for an appropriately chosen δ.

Say we have independent random samples $X_1, \ldots, X_n \sim \mathcal{D}$. A simple solution would be to take the average $\hat{\mu} = Z = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now, how can we bound $\hat{\mu}$ from the average μ? Chebyshev's inequality gives us the bound

$$\mathbb{P}[|Z - \mu| \leq t] \leq \frac{\sigma_Z^2}{t^2} = \frac{\sigma^2}{nt^2},$$

where σ_Z^2 is the variance for Z, and σ^2 is the variance of each variable X_i.

Set $t = \epsilon\sigma$, and the above bound gives us

$$\mathbb{P}[|Z - \mu| \leq \epsilon\sigma] \leq \frac{1}{n\epsilon^2} = \delta,$$

so we should choose $n = \frac{1}{\epsilon^2\delta}$. Can we do better than this (ie. is this bound tight)?

Let's first see what happens with the Gaussian distribution $Z \sim \mathcal{N}(\mu, \sigma^2/n)$, then one can show that

$$\mathbb{P}[|Z - \mu| \geq \frac{t\sigma}{\sqrt{n}}] \leq 2e^{-t^2/2},$$

so setting $\epsilon = \frac{t}{\sqrt{n}}$, we would need to choose $n \geq \frac{2}{\epsilon^2}\log\frac{\delta}{2}$.

To answer whether this is tight, let us first consider examples for which Markov's inequality is tight. Consider the distribution in which 0 is chosen with probability $1-p$ and k is chosen with probability p. Then for a random variable X, $\mu = kp$ and Markov tells us

$$p = \mathbb{P}[X \geq k] \leq \frac{\mu}{k} = p,$$

which is tight. We can do the same with Chebyshev's inequality. Consider the distribution in which $\alpha = \frac{1}{\sqrt{p}}\sigma$ and $-\alpha$ are each chosen with probability $p/2$, and 0 is chosen with probability $1-p$. Then for a random variable X, the variance is σ^2 and Chebyshev tells us

$$p = \mathbb{P}[|X| \geq \frac{1}{\sqrt{p}}\sigma] \leq \frac{\sigma^2}{\alpha^2} = p.$$

Now, for some chosen δ, n suppose we have the average $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$ where each X_i is distributed according to the preceding distribution with $p = \frac{2\delta}{n}$. Notice that

$$\mathbb{P}[|Z| \geq \frac{1}{n\sqrt{p}}\sigma] \geq \mathbb{P}[\exists! i \text{ s.t. } X_i \neq 0] = np(1-p)^{n-1} = 2\delta(1-\frac{2\delta}{n})^n \approx 2\delta e^{-2\delta} > \delta.$$

Now, in order for

$$\mathbb{P}[Z \geq \epsilon\sigma] > 2\delta$$

to be less than or equal to δ, we need we need $\epsilon \geq \frac{1}{n\sqrt{p}}\sigma = \frac{\sigma}{\sqrt{2\delta n}}$, and thus $n \geq \frac{2}{\epsilon^2\delta}$. This shows that our original bound is tight.

We will now see what happens if instead of taking the average, we take the median of the X_i. Note here that there is no ϵ dependence, ie. since all the X_i take values in ± 1, Z will also take values in ± 1. We now want to bound the probability that $|Z-\mu| \geq 2\sigma$. Note that since $|Z-\mu| = 1 = \sigma$, then for the median to not be in $\pm 2\sigma$ we need $n/2$ of the samples to be above or below $\pm 2\sigma$.

Now notice that

$$\mathbb{P}[\text{Any } |X_i| \leq 2\sigma] \geq 3/4,$$

so for Y_i the indicator function of whether $|X_i| \leq 2\sigma$ we have

$$\begin{aligned} \mathbb{P}[\text{At most } \frac{n}{2} \text{ of the } |X_i| \leq 2\sigma] &\leq \mathbb{P}[\sum_{i=1}^{n} Y_i \leq n/2] \\ &\leq \mathbb{P}[\sum_{i=1}^{n} Y_i \leq \mathbb{E}[2Y_i] - \frac{n}{4}] \\ &\leq 2^{-n/8} \leq \delta, \end{aligned}$$

so we would need to choose $n \geq 8\log\frac{1}{\delta}$.

Now, if we put it all together and combine the two methods and pick independent samples

$$\begin{array}{cccc} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \dots & X_{mn}, \end{array}$$

then we estimate μ by taking $\hat{\mu}_i = \text{mean}_{j\in[m]} X_{ij}$, and $\hat{\mu} = \text{median}_{i\in[m]}\hat{\mu}_i$. First, notice that using Chebyshev,

$$\mathbb{P}[\hat{\mu}_i - \mu| \leq \epsilon\sigma] \geq 1 - \frac{1}{n\epsilon^2} \geq 1 - \delta_1,$$

where we must choose $n \geq \frac{1}{\delta_1 \epsilon^2}$, and we will see later that it suffices to have $\delta_1 = \frac{1}{4}$.

We now consider the median of the $\hat{\mu}_i$'s. Let the random variable $Z_i = 1$ if $|\hat{\mu}_i - \mu| \leq \epsilon\sigma$ and 0 otherwise, then

$$\begin{aligned}
\mathbb{P}[|\text{median}_{i\in[m]}\hat{\mu}_i - \mu| \leq \epsilon\sigma] &\geq \mathbb{P}[\sum_{i=1}^{m} Z_i > m/2] \\
&= 1 - \mathbb{P}[\sum_{i=1}^{m} Z_i \leq m(1-\delta_1) - m(\frac{1}{2} - \delta_1)] \\
&\geq 1 - \mathbb{P}[\sum_{i=1}^{m} Z_i \leq \mathbb{E}[\sum_{i=1}^{m} Z_i] - m(\frac{1}{2} - \delta_1)] \\
&\geq 1 - \exp(-2(m(\frac{1}{2} - \delta_1))^2/m) \\
&\geq 1 - \delta_2
\end{aligned}$$

where we would need $m \geq \frac{1}{2(1/2-\delta_1)^2} \log \frac{1}{\delta_2}$, so if we choose $\delta_1 = \frac{1}{4}$, then we only need $m \geq 8 \log \frac{1}{\delta_2}$.

3 Cuckoo Hashing

- As we saw in previous lectures if we create a Hash Table and use random placement we get a worst case lookup time $O(\frac{\log n}{\log\log n})$.
- If instead we use the Power of Two Choices we get $O(\log\log n)$ which is much better.
- Aiming however for constant lookup time we turn our attention to Cuckoo Hashing.

In Cuckoo Hashing every cell of the hash table is considered a vertex and every element is mapped (from 2 different hash functions) to 2 vertices thus considered a (directed) edge.

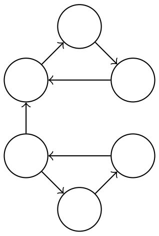
n vertices (bins)

m edges (balls)

Each element can occupy either end of the edge. If an element is mapped to 2 already occupied hash cells then we randomly evict one of them and repeat the same process until an open cell is found.

But things can go sour for our algorithm if a barbell appears in the graph.

Figure 1: Our algorithm fails if a barbell occurs in the graph.



In order to upper bound the probability our algorithm fails it suffices to compute the probability that any cycle appears in the graph.

Note: The analysis will borrow elements from Erdos Renyi $G(n, p)$ graphs and Galton-Wachon processes.

So what is the chance a cycle exists?

$$\mathbb{P}[\text{any cycle exists}] \leq \sum_{i=2}^{n} [\text{any length } i \text{ cycle exists}]$$

Without loss we can focus on undirected cycles and given that we have $\binom{n}{i}$ different cycles of length i we proceed:

$$\leq \sum_{i=2}^{n} n^i \, \mathbb{P}[\text{specific cycle of length } i \text{ exists}] \leq \sum_{i=2}^{n} n^i \, \mathbb{P}[\text{any particular set of } i \text{ edges exists}]$$

Focusing on the probability that any particular set of i edges exists we notice that there are $\binom{m}{i}$ possible edge assignments and each one of them is taking place with probability $\binom{n}{2}^{-i}$ thus

$$\sum_{i=2}^{n} n^i \, \mathbb{P}[\text{any particular set of } i \text{ edges exists}] \leq \sum_{i=2}^{n} n^i m^i \binom{n}{2}^{-i} \leq \sum_{i=2}^{n} O((\frac{m}{n})^i)$$

If we have that $m < cn$ for a sufficiently small c then

$$\sum_{i=2}^{n} O((\frac{m}{n})^i) = O((\frac{m}{n})^2) < \frac{1}{10000}$$

Thus for sufficiently small m,

$$\mathbb{P}[\text{Cuckoo Hashing fails}] < \frac{1}{10000}$$

If a cycle is encountered then we create new hash functions and rebuild the table. Obviously the expected times of rebuilding the table is $O(1)$.

$$\begin{aligned}\mathbb{E}[\text{time to build}] &= \sum_{i=1}^{m} \mathbb{E}[\text{time to insert the } i^{th} \text{ element}] \\ &\leq m \, \mathbb{E}[\text{the size of the component of any element}] \\ &\leq 2m \, \mathbb{E}[\text{size of component of any vertex}]\end{aligned}$$

Finally in order to prove that the expected size of the component of any vertex is constant we will use the analysis from Erdos Renyi $G(n,p)$ and Galton-Watson processes.

We pick arbitrarily any node u as the root. u has $n-1$ potential children-neigbors and each one of them has probability $m\binom{n}{2}^{-i}$. Thus the expected number of u's neighbors is at most $2m/n << 1$. Similarly any child v has itself $n-1-j \leq n-1$ potential children (where j are the nodes that are already in the same component as u) and each of them has probabilty at most $m\binom{n}{2}^{-i}$.

From Galton-Watson process analysis it follows that the expected size of the component of u is $O(1)$.

$$\begin{aligned} f(n,p) &= \mathbb{E}[\textit{size of component}] \\ &\leq 1 + p(n-1)f(n-1,p) \\ &\leq 1 + np + np^2 + ... \\ &\leq \frac{1}{1-np} \end{aligned}$$

which is obviously constant for $p = m/\binom{n}{2}$.

References

[1] P. Erdos and A. Renyi. On random graphs, i. *Publicationes Mathematicae (Debrecen)* 6:290-297, 1959.

[2] P. Erdos and A. Renyi. *On the evolution of random graphs.* Akad. Kiado, 1960.

[3] H. KESTEN, P. NEY, and F. SPITZER *Galton-Watson processes with mean one and finite variance* Theor. Probability Appl., Vol. 13, pp. 513-540, 1966.