Quicksort and Tail Bounds

1 Randomized Quicksort

Algorithm 1 Randomized Quicksort

1: Given a set S, choose a pivot p uniformly at random (or with median finding technique from last lecture)
2: In linear time, compute $S^- = \{x \in S | x \leq p\}$ and $S^+ = x \in S | x > p\}$ {
* Quicksort(S^-, S^+)
* Return [sorted S^-, p, sorted S^+]

Notice that, up to constant factors, the runtime of Quicksort is equal to the number of comparisons made (which is also true for deterministic versions). Let T be the random variable representing the number of element pairs compared. We also note that when elements i, j are compared, one of them must have been p, so they cannot be compared again. Define indicator random variables:

$$X_{ij} = \begin{cases} 1 & \text{if } i, j \text{ compared} \\ 0 & \text{o.w.} \end{cases}$$

Thus:

$$T = \sum_{i,j} X_{ij}$$

$$\mathbb{E}[T] = \sum_{i.j} \mathbb{E}[X_{ij}]$$

$$= \sum_{i,j} \Pr[i, j \text{ compared}]$$

Consider elements i, j in the final sorted order. Without loss of generality, set $i < j$. If i or j is selected as the pivot before any element between them, then $X_{ij} = 1$; otherwise they will be placed in different sets and $X_{ij} = 0$. Now, consider the first time a pivot in $\{i, i+1, \ldots, j-1, j\}$ is selected. At that point, $\Pr[i \text{ or } j \text{ is pivot}|\text{such a pivot is selected}] = \frac{2}{j+1-i}$.

$$\mathbb{E}[T] = \sum_{i,j} \frac{2}{j+1-i}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n-i} \frac{2}{k-1}$$

$$\leq 2\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{1}{k}$$

$$= 2nH(n) \text{ where } H(n) \text{ is the } n^{th} \text{ harmonic number}$$

$$\approx 2n\log n$$

So the expected runtime is $\Theta(n\log n)$.

2 Tail Bounds

2.1 Intro to Tail Bounds

Oftentimes, we are interested in high probability guarantees (i.e., that for some event $\mathcal{E}$, $\Pr[\mathcal{E}] \geq 1-p^{-1}$ with p polynomial in n). This is especially useful for bounding expectation of maxima and minima (something pretty useful as a computer scientist). **Tail bounds** are bounds which show that $\Pr[X < \alpha]$ is "small". For our purposes, we want bounds which guarantee that $\Pr[X$ is far from $\mathbb{E}[X]]$ is "small".

The general approach for tail bounds is to show that each $X_i \leq \alpha$ with probability $\geq 1 - m^{-2}$ or some other probability close to 1, where m is the number of variables. Then, by the union bound:

$$\Pr[\text{one or more } X_i > \alpha] \leq \sum_i \Pr[X_i > \alpha]$$

$$= \sum_i \frac{1}{m^2}$$

$$= \frac{1}{m}$$

Thus:

$$\Pr[\text{all } X_i \leq \alpha] \geq 1 - \frac{1}{m}$$

2.2 Markov's Inequality

Definition 1. Markov's Inequality: If $X \geq 0$, then:

$$\Pr[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}$$

This bound is pretty weak, but can be useful. For example, say we play a casino game where x is measured in integer dollars and:

$$\mathbb{E}[X_{t+1}|X_t = x] = 0.9x$$

How soon, starting from n dollars, until we are broke ($X_t = 0$) with high probability? First, we know that:

$$\mathbb{E}[X_t] = 0.9^t n$$

We set a convenient value of t to get a nice expectation and plug it in to Markov:

$$t = 2\log_{\frac{1}{0.9}} n \implies \mathbb{E}[X_t] = \frac{1}{n}$$

$$\Pr[X_t \neq 0] = \Pr[X_t \geq 1]$$

$$\leq \frac{\frac{1}{n}}{1} \text{ by Markov's inequality}$$

$$= \frac{1}{n}$$

Then, we know that at that t, we are broke with high probability:

$$\Pr[X_t = 0] = 1 - \frac{1}{n} \text{ when } t = 2\log_{\frac{1}{0.9}} n$$

There is an extended version of Markov's Inequality called Chebyshev bounds, which give better bounds when we can calculate $Var(X)$, and in particular if $Var(X)$ is $\mathcal{O}(\mathbb{E}[X])$.

2.3 Chernoff Bounds

Assume $X_1, \ldots, X_n$ are independent indicator random variables, and $X = \sum_i X_i$ with $\delta > 0$. Then, Chernoff bounds state:

$$\Pr[X > (1+\delta)\mu] < \frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}^{\mu} \quad \forall \delta \geq \mathbb{E}[X]$$

$$\Pr[X < (1+\delta)\mu] < \exp(-\frac{1}{2}\mu\delta^2) \ \forall \delta \leq \mathbb{E}[X]$$

Assume $X_1, \ldots, X_n$ are independent indicator random variables such that $X_i \in [a_i, b_i]$ always, and $X = \sum_i X_i$ with $\mu = \mathbb{E}[X]$. Then, Hoeffding bounds state:

$$\Pr[|X - \mu| \geq \Delta] \leq 2\exp(\frac{-2\Delta^2}{\sum_i (b_i - a_i)^2})$$

2.4 Tail Bounds on Balls and Bins

Given n balls and m bins, distribute the balls across the bins to minimize the maximum number of balls in any bin. This is a trivial problem if we just evenly distribute the balls, but we are interested in the randomized approach as a simple example for usage of Chernoff bounds. Let N_i be the random variable representing the number of balls in bin i. Define indicator random variables:

$$B_{ij} = \begin{cases} 1 & \text{if ball } j \text{ lands in bin } i \\ 0 & \text{o.w.} \end{cases}$$

Thus:

$$N_i = \sum_j B_{ij}$$

$$\mathbb{E}[N_i] = \sum_j \mathbb{E}[B_{ij}]$$

$$= \frac{n}{m}$$

Now, we want an α such that $\Pr[N_i > \alpha] < m^{-2}$, so we can apply the union bound. Notice that for a fixed i, B_{ij} are independent, so we can apply Chernoff bounds:

$$\Pr[X > (1+\delta)\frac{n}{m}] < \frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}^{\frac{n}{m}}$$

We want the right side to be less than m^{-2}, so we can reverse engineer the smallest δ for which this works. We'll consider two special cases; first, the case when $n = m$:

$$\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}} < m^{-2} \iff \delta[(1+\delta)\ln(1+\delta)] > 2\ln m$$

So the solution is:

$$\delta = \Theta(\frac{\log m}{\log\log m})$$

Then, because $\mu = 1$:

$$\Pr[N_i > \Theta(\frac{\log m}{\log\log m})] < m^{-2}$$

And by the union bound, all N_i are bounded by $\Theta(\frac{\log m}{\log\log m})$ with probability greater than $1 - m^{-1}$. For the second special case, when $n = 16m\log m$:

$$\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}^{16\log m} < m^{-2} \implies \delta = 1$$

Then, because $\mu = 16\log m$:

$$\Pr[N_i > 16\log m] < m^{-2}$$

And by the union bound, all N_i are bounded by $16\log m$ with probability greater than $1 - m^{-1}$.