Learning Goals

- Apply Chernoff bound in typical scenarios
- Understand analysis of Quicksort
- Develop quantitative understanding of the balls and bins process

Quicksort

- The proof for the expected height of a binary search tree for randomly arriving elements implies the expected running time of Quicksort is $O(n \log n)$.

Quicksort

- The proof for the expected height of a binary search tree for randomly arriving elements implies the expected running time of Quicksort is $O(n \log n)$.
- We now show that the running time of Quicksort is $O(n \log n)$ *with high probability* by applying the Chernoff bounds.

Quicksort

- The proof for the expected height of a binary search tree for randomly arriving elements implies the expected running time of Quicksort is $O(n \log n)$.
- We now show that the running time of Quicksort is $O(n \log n)$ *with high probability* by applying the Chernoff bounds.
- Recall: the procedure of Quicksort can be represented by a binary tree, where each node represents a pivoting, the two children being the two subsets resulting from comparisons with the pivoting element.

Quicksort

- The proof for the expected height of a binary search tree for randomly arriving elements implies the expected running time of Quicksort is $O(n \log n)$.
- We now show that the running time of Quicksort is $O(n \log n)$ *with high probability* by applying the Chernoff bounds.
- Recall: the procedure of Quicksort can be represented by a binary tree, where each node represents a pivoting, the two children being the two subsets resulting from comparisons with the pivoting element.
- The running time for each level in total is $O(n)$, so we will show that with high probability the height of the tree is $O(\log n)$.

Analysis of Quicksort

- Follow a path from the root, specified by “left” or “right” at each step.

Analysis of Quicksort

- Follow a path from the root, specified by "left" or "right" at each step.
- We say each step is "good" if the size of the array at the child is at most $\frac{3}{4}$ that of the parent; otherwise we say the step is "bad".

Analysis of Quicksort

- Follow a path from the root, specified by "left" or "right" at each step.
- We say each step is "good" if the size of the array at the child is at most $\frac{3}{4}$ that of the parent; otherwise we say the step is "bad".
- There can be at most $\log n / \log \frac{4}{3}$ good steps before we are at a leaf.

Analysis of Quicksort

- Follow a path from the root, specified by "left" or "right" at each step.
- We say each step is "good" if the size of the array at the child is at most $\frac{3}{4}$ that of the parent; otherwise we say the step is "bad".
- There can be at most $\log n / \log \frac{4}{3}$ good steps before we are at a leaf.
- Let's bound the probability that, in $12 \log n$ steps, there are fewer than $\log_{\frac{4}{3}} n$ good steps.

Analysis of Quicksort

- Follow a path from the root, specified by “left” or “right” at each step.
- We say each step is “good” if the size of the array at the child is at most $\frac{3}{4}$ that of the parent; otherwise we say the step is “bad”.
- There can be at most $\log n / \log \frac{4}{3}$ good steps before we are at a leaf.
- Let’s bound the probability that, in $12 \log n$ steps, there are fewer than $\log_{\frac{4}{3}} n$ good steps.
- Let X_i be the indicator variable for the i-th step being good, then $\mathbf{E}[X_i] \geq \frac{3}{4}$, and the X_i’s are i.i.d.

Analysis of Quicksort

- Follow a path from the root, specified by "left" or "right" at each step.
- We say each step is "good" if the size of the array at the child is at most $\frac{3}{4}$ that of the parent; otherwise we say the step is "bad".
- There can be at most $\log n / \log \frac{4}{3}$ good steps before we are at a leaf.
- Let's bound the probability that, in $12 \log n$ steps, there are fewer than $\log_{\frac{4}{3}} n$ good steps.
- Let X_i be the indicator variable for the i-th step being good, then $\mathbf{E}[X_i] \geq \frac{3}{4}$, and the X_i's are i.i.d.
- Let X be $\sum_{i=1}^{12 \log n} X_i$. By Chernoff bound, we have

$$\mathbf{Pr}\left[X < \frac{\log n}{\log \frac{4}{3}}\right] \leq \mathbf{Pr}\left[X < \mathbf{E}\left[X\right] - 5 \log n\right] \leq \exp\left(-2 \cdot \frac{25 \log^2 n}{12 \log n}\right) = n^{-25/6}.$$

Analysis of Quicksort (Cont.)

- There are n leaves.

Analysis of Quicksort (Cont.)

- There are n leaves.
- By the union bound, the probability that *any* leaf has depth more than $12 \log n$ is no more than $n \cdot n^{-25/6} = n^{-19/6}$.

Analysis of Quicksort (Cont.)

- There are n leaves.
- By the union bound, the probability that *any* leaf has depth more than $12 \log n$ is no more than $n \cdot n^{-25/6} = n^{-19/6}$.
- Therefore, with high probability, the height of the tree is bounded by $12 \log n$.

Analysis of Quicksort (Cont.)

- There are n leaves.
- By the union bound, the probability that *any* leaf has depth more than $12 \log n$ is no more than $n \cdot n^{-25/6} = n^{-19/6}$.
- Therefore, with high probability, the height of the tree is bounded by $12 \log n$.
- Obviously the constants in the analysis were not finetuned.

The Negative Binomial Distribution

- In the proof above, we wanted to bound the probability that, we take more than $12 \log n$ steps to see $\log_{4/3} n$ good ones; instead, we bounded the probability that, within a $12 \log n$ steps, there are fewer than $\log_{4/3} n$ good ones.

The Negative Binomial Distribution

- In the proof above, we wanted to bound the probability that, we take more than $12 \log n$ steps to see $\log_{4/3} n$ good ones; instead, we bounded the probability that, within a $12 \log n$ steps, there are fewer than $\log_{4/3} n$ good ones.
- Are these two probabilities equal?

The Negative Binomial Distribution

- In the proof above, we wanted to bound the probability that, we take more than $12 \log n$ steps to see $\log_{4/3} n$ good ones; instead, we bounded the probability that, within a $12 \log n$ steps, there are fewer than $\log_{4/3} n$ good ones.
- Are these two probabilities equal?
- Answer: Yes. A random variable counting the number of i.i.d. trials before seeing k successful ones is said to have the *negative binomial distribution*.

The Negative Binomial Distribution

- In the proof above, we wanted to bound the probability that, we take more than $12 \log n$ steps to see $\log_{4/3} n$ good ones; instead, we bounded the probability that, within a $12 \log n$ steps, there are fewer than $\log_{4/3} n$ good ones.
- Are these two probabilities equal?
- Answer: Yes. A random variable counting the number of i.i.d. trials before seeing k successful ones is said to have the *negative binomial distribution*.
- The probability that such a random variable is larger than n is equal to the probability that, within n i.i.d. trials we have not seen k successful ones.
  - The statement may seem obvious, but a formal argument needs either "coupling" or some careful calculations.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.
- In our first lecture, we considered n tasks sending requests uniformly at random to one of the servers.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.
- In our first lecture, we considered n tasks sending requests uniformly at random to one of the servers.
- Such scenarios arise very often in algorithmic analysis. It is helpful to develop an intuition for them.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.
- In our first lecture, we considered n tasks sending requests uniformly at random to one of the servers.
- Such scenarios arise very often in algorithmic analysis. It is helpful to develop an intuition for them.
- This is often abstracted as a *balls and bins* model, where we have n balls and m bins, and each ball is thrown uniformly at random to a bin.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.
- In our first lecture, we considered n tasks sending requests uniformly at random to one of the servers.
- Such scenarios arise very often in algorithmic analysis. It is helpful to develop an intuition for them.
- This is often abstracted as a *balls and bins* model, where we have n balls and m bins, and each ball is thrown uniformly at random to a bin.
- Any bin receives in expectation $\frac{n}{m}$ balls. If $m = n$, this is 1.

Bins and Balls

- When discussing hashing, we considered a naïve family of hash: mapping elements of U uniformly random to an address.
- In our first lecture, we considered n tasks sending requests uniformly at random to one of the servers.
- Such scenarios arise very often in algorithmic analysis. It is helpful to develop an intuition for them.
- This is often abstracted as a *balls and bins* model, where we have n balls and m bins, and each ball is thrown uniformly at random to a bin.
- Any bin receives in expectation $\frac{n}{m}$ balls. If $m = n$, this is 1.
- How about the bin that received the most balls? How many balls should we expect to see there?

Balls and Bins when $m = n$

- Let's consider a particular bin. Let X_i be the indicator variable for the event that the i-th ball falls in this bin.

Balls and Bins when $m = n$

- Let's consider a particular bin. Let X_i be the indicator variable for the event that the i-th ball falls in this bin.
- Then $\mathbf{Pr}[X_i = 1] = \frac{1}{n}$.

Balls and Bins when $m = n$

- Let's consider a particular bin. Let X_i be the indicator variable for the event that the i-th ball falls in this bin.
- Then $\mathbf{Pr}[X_i = 1] = \frac{1}{n}$.
- Let X be $\sum_i X_i$. Note that $\mathbf{E}[X] = 1$.

Balls and Bins when $m = n$

- Let's consider a particular bin. Let X_i be the indicator variable for the event that the i-th ball falls in this bin.
- Then $\mathbf{Pr}[X_i = 1] = \frac{1}{n}$.
- Let X be $\sum_i X_i$. Note that $\mathbf{E}[X] = 1$.
- For $t > 0$, we use Chernoff bound

$$\mathbf{Pr}\left[X > (1+t)\,\mathbf{E}\left[X\right]\right] \le \left(\frac{e^t}{(1+t)^{1+t}}\right)^{\mathbf{E}[X]} \le \left(\frac{e}{1+t}\right)^{1+t}.$$

Balls and Bins when $m = n$

- Let's consider a particular bin. Let X_i be the indicator variable for the event that the i-th ball falls in this bin.
- Then $\mathbf{Pr}[X_i = 1] = \frac{1}{n}$.
- Let X be $\sum_i X_i$. Note that $\mathbf{E}[X] = 1$.
- For $t > 0$, we use Chernoff bound

$$\mathbf{Pr}\left[X > (1+t)\,\mathbf{E}\left[X\right]\right] \le \left(\frac{e^t}{(1+t)^{1+t}}\right)^{\mathbf{E}[X]} \le \left(\frac{e}{1+t}\right)^{1+t}.$$

- We would like to find t to so that this probability is smaller than n^{-2}. Essentially we are asking what solves $x^x = n$.

Balls and Bins when $m = n$ (Cont.)

- To estimate the solution of $x^x = n$, we first take logarithm, $x \log x = \log n$, $\log x + \log \log x = \log \log n$.

Balls and Bins when $m = n$ (Cont.)

- To estimate the solution of $x^x = n$, we first take logarithm, $x \log x = \log n$, $\log x + \log \log x = \log \log n$.
- Note that $x < \log n$.

Balls and Bins when $m = n$ (Cont.)

- To estimate the solution of $x^x = n$, we first take logarithm, $x \log x = \log n$, $\log x + \log \log x = \log \log n$.
- Note that $x < \log n$.
- We have $2 \log x \geq \log x + \log \log x = \log \log n \geq \log x$, so

$$\frac{1}{2}x \leq \frac{\log n}{\log \log n} \leq x \Rightarrow x = \Theta\left(\frac{\log n}{\log \log n}\right).$$

Balls and Bins when $m = n$ (Cont.)

- To estimate the solution of $x^x = n$, we first take logarithm, $x \log x = \log n$, $\log x + \log \log x = \log \log n$.
- Note that $x < \log n$.
- We have $2 \log x \geq \log x + \log \log x = \log \log n \geq \log x$, so

$$\frac{1}{2}x \leq \frac{\log n}{\log \log n} \leq x \Rightarrow x = \Theta\left(\frac{\log n}{\log \log n}\right).$$

- Let the solution to $x^x = n$ be $\gamma(n)$, and let $1 + t = e\gamma(n)$, we have

$$\left(\frac{e}{1+t}\right)^{1+t} = \left(\frac{1}{\gamma(n)}\right)^{e\gamma(n)} = n^{-e} < n^{-2}.$$

Balls and Bins when $m = n$ (Cont.)

- To estimate the solution of $x^x = n$, we first take logarithm, $x \log x = \log n$, $\log x + \log \log x = \log \log n$.
- Note that $x < \log n$.
- We have $2 \log x \geq \log x + \log \log x = \log \log n \geq \log x$, so

$$\frac{1}{2}x \leq \frac{\log n}{\log \log n} \leq x \Rightarrow x = \Theta\left(\frac{\log n}{\log \log n}\right).$$

- Let the solution to $x^x = n$ be $\gamma(n)$, and let $1 + t = e\gamma(n)$, we have

$$\left(\frac{e}{1+t}\right)^{1+t} = \left(\frac{1}{\gamma(n)}\right)^{e\gamma(n)} = n^{-e} < n^{-2}.$$

- By union bound, with probability at least $1 - \frac{1}{n}$, no bin receives more than $e\gamma(n) = \Theta(\frac{\log n}{\log \log n})$ balls.

Balls and Bins When $n \gg m$

- As n grows, the number of balls concentrates more sharply around its means.

Balls and Bins When $n \gg m$

- As n grows, the number of balls concentrates more sharply around its means.
- E.g., take $n = 16n \log m$, with the previous notation, $\mathbf{E}[X] = 16 \log m$.

Balls and Bins When $n \gg m$

- As n grows, the number of balls concentrates more sharply around its means.
- E.g., take $n = 16n \log m$, with the previous notation, $\mathbf{E}[X] = 16 \log m$.

$$\mathbf{Pr}\left[X \geq 32 \log m\right] = \mathbf{Pr}\left[X \geq 2\,\mathbf{E}\left[X\right]\right] \leq e^{-\,\mathbf{E}[X]/3} = m^{-16/3} < \frac{1}{m^2};$$

$$\mathbf{Pr}\left[X \leq 8 \log m\right] = \mathbf{Pr}\left[X \leq \frac{1}{2}\,\mathbf{E}\left[X\right]\right] \leq e^{-\,\mathbf{E}[X]/8} = \frac{1}{m^2}.$$

Balls and Bins When $n \gg m$

- As n grows, the number of balls concentrates more sharply around its means.
- E.g., take $n = 16n \log m$, with the previous notation, $\mathbf{E}[X] = 16 \log m$.

$$\mathbf{Pr}\left[X \geq 32 \log m\right] = \mathbf{Pr}\left[X \geq 2\,\mathbf{E}\left[X\right]\right] \leq e^{-\mathbf{E}[X]/3} = m^{-16/3} < \frac{1}{m^2};$$

$$\mathbf{Pr}\left[X \leq 8 \log m\right] = \mathbf{Pr}\left[X \leq \frac{1}{2}\,\mathbf{E}\left[X\right]\right] \leq e^{-\mathbf{E}[X]/8} = \frac{1}{m^2}.$$

Theorem

For $n = \Omega(m \log m)$, with high probability, the number of balls every bin receives is between half and twice the average.