RECAP

CHERNOFF BOUNDS for sums of Bernoulli trials (independent indicator variables)

CHERNOFF'S BOUND 1: Let $X_1, X_2, \ldots, X_n$ be independent indicator variables (Bernoulli trials) with $\Pr(X_i = 1) = p_i = E[X_i]$, $0 < p_i < 1$, $1 \le i \le n$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = E[X] = \sum_{i=1}^{n} p_i$

Then $\boxed{\forall \delta > 0 : \quad \text{Prob}(X > (1+\delta)\mu) < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu}$

CHERNOFF'S BOUND 2: Let $X_1, X_2, \ldots, X_n$ be independent indicator variables (Bernoulli trials) with $\Pr(X_i = 1) = p_i = E[X_i]$, $0 < p_i < 1$, $1 \le i \le n$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = E[X] = \sum_{i=1}^{n} p_i$

Then $\boxed{\forall\, 0 < \delta < 1 : \quad \text{Prob}(X > (1+\delta)\mu) < e^{-\delta^2\mu/3}}$

CHERNOFF'S BOUND 3: Let $X_1, X_2, \ldots, X_n$ be independent indicator variables (Bernoulli trials) with $\Pr(X_i = 1) = p_i = E[X_i]$, $0 < p_i < 1$, $1 \le i \le n$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = E[X] = \sum_{i=1}^{n} p_i$

Then $\boxed{\forall\, 0 < \delta < 1 : \quad \text{Prob}(X < (1-\delta)\mu) < e^{-\delta^2\mu/2}}$

ANALYSIS OF RANDOMIZED QUICKSORT

Let S be a set of n distinct integers (w.l.o.g.). We will denote with

$$\boxed{SORT(S) = \langle x_1, x_2, \ldots, x_n \rangle}$$

increasing

the sorted sequence of elements in S.

ORDER STATISTICS (O.S.): $x_1 = \min(S)$, $x_n = \max(S)$, x_i = i-O.S.

We can write:

```
QUICKSORT(S)                         <x_1>    <∅>
  if |S| ≤ 1 then return <S>
  y ← CHOOSE-PIVOT(S)
  S_1 ← {x ∈ S : x < y}   }
  S_2 ← {x ∈ S : x > y}   } n-1
  X_1 ← QUICKSORT(S_1)
  X_2 ← QUICKSORT(S_2)
  return <X_1, y, X_2>
```

There are various ways of implementing CHOOSE-PIVOT:

1. $\boxed{\text{return } S[1]}$

(always return the first element of the unsorted sequence). If S is already sorted ($S[1..n] = \langle x_1, x_2, \ldots, x_n \rangle$) we obtain

$$\boxed{T_w(n) = T_w(\overset{|S_2|}{n-1}) + n - 1 \Rightarrow T_w(n) = \Theta(n^2)}$$

2. return MEDIAN(S)

(always return $x_{\lfloor \frac{n+1}{2} \rfloor}$). This is difficult to do, and requires a complicated algorithm running in time $c \cdot n$, with a very high constant c [CLRS, 220-222]

However, we obtain:

MEDIAN PARTITION

$$\boxed{\begin{aligned} T(n) &\le 2T\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + (c+1)\cdot n \\ \Rightarrow T(n) &\approx (c+1)\, n \log_2 n \end{aligned}}$$

High constant factor. Also, the resulting algorithm is not in place.

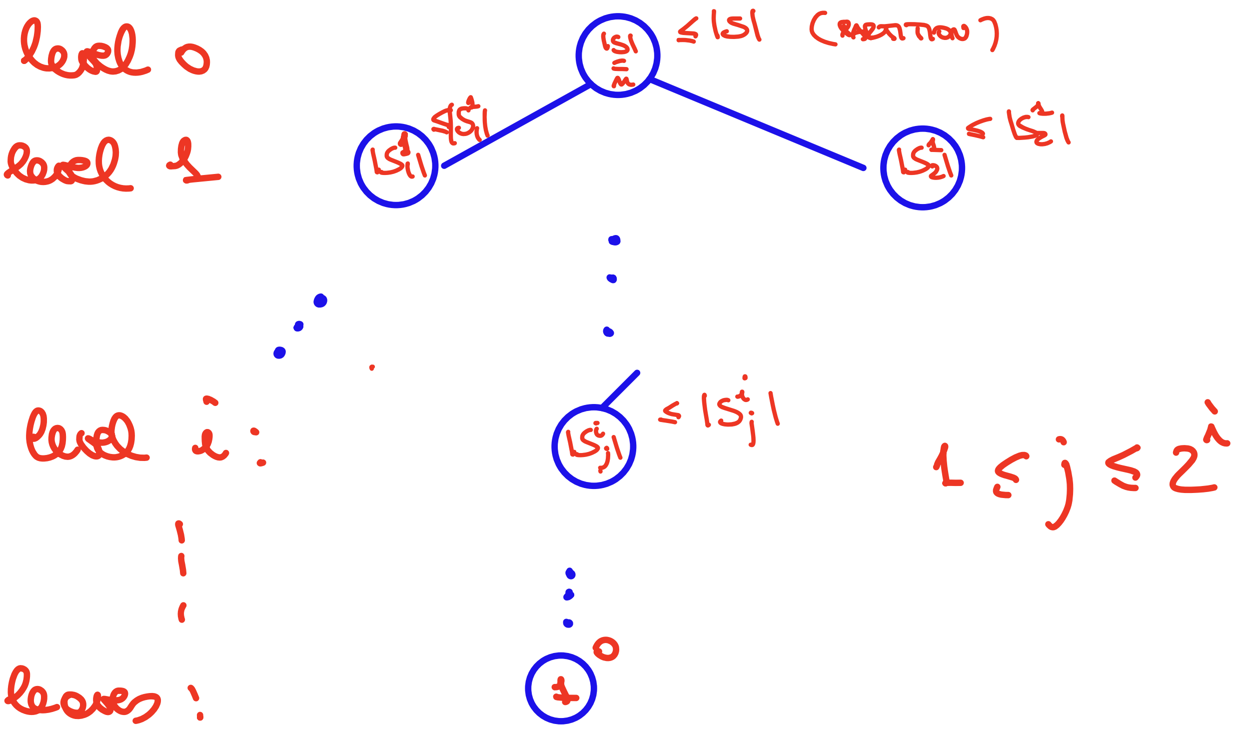
3. return RANDOM(S).

We will study this scenario in high probability.

BASIC IDEA The two scenarios 1. and 2. for CHOOSE-PIVOT are extreme (worst-1 vs best-2)

A random choice of the pivot does not guarantee a perfectly even split but will however split $S \setminus S^0$ into subsets S^1_1 and S^1_2 that are both "rather large", with "good" probability. As a consequence $Y \Leftarrow \text{RANDOM}(S)$ (3) behaves more like 2 than 1.

Consider the recursion tree of QUICKSORT:



- (We eliminate leaves corresponding to empty calls: not all internal nodes have two children)
  $\Rightarrow \leq n$ leaves
- For each internal node S^i_j, the sets associated to its children S', S'':

  $$\boxed{|S'| + |S''| = |S^i_j| - 1 \leq |S^i_j|}$$

  total work at level $i+1 \leq$ total work at level i

$\Rightarrow$ At each level i: total work at level $i \leq n$.

$$\Rightarrow \boxed{T_{QS} \leq n \cdot (\#\text{ levels})}$$

With an even split (given by $y \leftarrow \text{MEDIAN}(S)$):

$\Rightarrow \boxed{|S_j^i| \leq \lfloor n/2^i \rfloor \leq n/2^i}$

$$\boxed{\#\text{levels} \leq \log_2 n + 1 \Rightarrow T_{QS}(n) \leq (c+1) \cdot n(\log_2 n + 1)}$$

WEAKER BALANCING:

Assume that, rather than guaranteeing an even split (as given by $y = \text{MEDIAN}(S)$) we are happy with a slightly more unbalanced split: $\boxed{|S_1^1|, |S_2^1| \leq \frac{3}{4} n}$.

$\left(\Rightarrow |S_1^1|, |S_2^1| \geq n - 1 - \frac{3}{4} n = \frac{n}{4} - 1\right)$

We have that at level i: $\boxed{|S_j^i| \leq \left(\frac{3}{4}\right)^i n}$

$\forall$ subinstance S_j^i.

The maximum level h is such that

$$h = \max\left\{ i : n \cdot \left(\frac{3}{4}\right)^i \geq 1 \right\}$$

$$n \left(\frac{3}{4}\right)^i \overset{?}{\geq} 1 \Rightarrow n \geq \left(\frac{4}{3}\right)^i \Rightarrow i \leq \log_{\frac{4}{3}} n$$

$$\Rightarrow h \leq \log_{4/3} n$$

$$\Rightarrow T(n) = O(n \cdot h) = O(n \cdot \log_{4/3} n) = O(n \log n)$$

The asymptotic running time stays the same even if we cannot guarantee perfect splits but only weaker balancing.

CRUCIAL IDEA: Weaker balancing is very easy to achieve, with good probability.

PROPERTY Let $\boxed{SORT(S) = \langle x_1, ..., x_m \rangle}$. If x_i is selected as a pivot, with

$$\boxed{\left\lfloor \frac{m}{4} \right\rfloor + 1 \leq i \leq \left\lceil \frac{3}{4} m \right\rceil}$$

then $\boxed{|S_1^1|, |S_2^1| \leq \frac{3}{4} m}$

PROOF: S_1^1 will surely contain $x_1, ..., x_{\lfloor \frac{m}{4} \rfloor}$

S_2^1 will contain $x_{\lceil \frac{3}{4} m \rceil + 1}, ..., x_m$. Thus:

$$|S_1^1| \geq \left\lfloor \frac{m}{4} \right\rfloor, \quad |S_2^1| \geq m - \left\lceil \frac{3}{4} m \right\rceil$$

But $|S_1^1| = m - 1 - |S_2^1| \leq m - 1 - m + \left\lceil \frac{3}{4} m \right\rceil =$

$= \left\lceil \frac{3}{4} m \right\rceil - 1 \leq \frac{3}{4} m$ $\qquad \left\lceil \frac{3}{4} m \right\rceil \leq \frac{3}{4} m + 1$

$|S_2^1| \leq m - 1 - |S_1^1| = m - 1 - \left\lfloor \frac{m}{4} \right\rfloor \leq$

$\left\lfloor \frac{m}{4} \right\rfloor \geq \left\lfloor \frac{m}{4} \right\rfloor - 1$

$\leq m - 1 - \frac{m}{4} + 1 \leq \frac{3}{4} m$

How many choices of the pivot guarantee weak balancing?

$$\boxed{\left\lceil \frac{3}{4} m \right\rceil - \left\lfloor \frac{m}{4} \right\rfloor \geq \frac{3}{4} m - \frac{m}{4} = \frac{m}{2}}$$

MORALE: If I pick the pivot at random from S, I achieve weak balancing with probability $\geq \frac{1}{2}$

INTUITION: On average, given a set S, I can reduce the size of all of the subinstances to $\leq \frac{3}{4}|S|$ in no more than two levels! Thus the maximum number of levels under 3) (y+RANDOM(S)) is $\boxed{\simeq c \cdot \log_{\frac{4}{3}} n}$

We transform this intuition into an high-probability analysis using Chernoff's bound 3.

Consider the recursion tree of a call of QUICKSORT(S) with random pivot. We know that:

1. the tree has $\leq n$ leaves $\Rightarrow$ there are $\leq n$ distinct root-to-leaf (r2l) paths.
2. The total work per level is $\leq n$

Let us bound the length of a fixed r2l path in the tree:

For $\boxed{a > 1}$, let $\boxed{t = a \log_{\frac{4}{3}} n}$. Let us study

$$\boxed{\Pr(\text{a fixed r2l path has length} > t)}$$

If "the r2l path has length > t" $\leftarrow e_1$:

$\Rightarrow$ "in the first $t = a \log_{4/3} n$ nodes on the path, the algorithm has made less than $\log_{4/3} n$ pivot choices that guarantee weak balancing (or the path would end by level t!)

$E_1 \Rightarrow E_2 \Rightarrow \Pr(E_1) \leq \Pr(E_2)$

We can model this event as follows: consider t indicator variables (one per level): $X_1, X_2, \ldots, X_t$, where $\boxed{X_i = 1}$ if the pivot choice at the i-th node of the path yielded weak balancing. We have argued that

$$\boxed{\Pr(X_i = 1) = \frac{1}{2} \quad (\text{in fact}, \geq \frac{1}{2})}$$

since the X_i's are obtained by different calls to RANDOM the X_i's are independent (Bernoulli trials). We have to study:

$$\boxed{\leq \Pr\left(X = \sum_{i=1}^{t} X_i < \log_{4/3} n\right)} \qquad (1)$$

Observe that $\boxed{\mu = E[X] = \sum_{i=1}^{t} E[X_i] = t/2}$

Let's rewrite (1) in "Chernoff3" form!

Determine δ!

$$\boxed{\log_{\frac{4}{3}} n = (1-\delta)\mu = (1-\delta)\frac{t}{2} = (1-\delta)\frac{a}{2}\log_{\frac{4}{3}} n}$$

By fixing $\boxed{a = 8}$, we have $\boxed{\mu = \frac{t}{2} = 4\log_{\frac{4}{3}} n}$, therefore

$$\boxed{\underbrace{\log_{\frac{4}{3}} n}_{(1-\delta)\mu} = \frac{1}{4}\cdot\mu \Rightarrow (1-\delta) = \frac{1}{4} \Rightarrow \delta = \frac{3}{4}}$$

By applying Chernoff's bound 3:

$$\boxed{\begin{aligned}
&\Pr\left(X < \log_{\frac{4}{3}} n\right) = \Pr\left(X < (1-\delta)\mu\right) < e^{-\frac{\delta^2}{2}\mu} \qquad \Big|_{\delta = \frac{3}{4}} \\
&= e^{-\frac{\left(\frac{9}{16}\right)\left(4\log_{\frac{4}{3}} n\right)}{2}} = e^{-\frac{9}{8}\log_{\frac{4}{3}} n} < e^{-\log_{\frac{4}{3}} n} \\
&= e^{-\frac{\ln n}{\ln(4/3)}} = \left(e^{-\ln n}\right)^{\frac{1}{\ln(4/3)}} = \frac{1}{n^{(1/\ln(4/3))}} < \frac{1}{n^3},
\end{aligned}}$$

since $\boxed{1/\ln(4/3) = 3.47\ldots}$

We have just proved that the probability that a fixed root-to-leaf path is longer than $2\log_{\frac{4}{3}} n = 8\log_{\frac{4}{3}} n$ is $< 1/n^3$

However, we have $p \leq n$ such paths in the tree!

Let $\boxed{E_i}$ be the event that the i-th path has length $> 2\log_{\frac{4}{3}} n$. We have proved that $\Pr(E_i) < \frac{1}{n^3}$

We can upper bound the probability that there is a path of length $> 2\log_{\frac{4}{3}} n$ as

$$\boxed{\Pr\left(\bigcup_{i=1}^{n} E_i\right) \underset{\text{UNION BOUND}}{\leq} \sum_{i=1}^{n} \Pr(E_i) < \frac{n}{n^3} \leq \frac{1}{n^2}}$$

Finally we can say that

$$\Rightarrow \Pr\left(T_{RQS}(n) > 8n\log_{4/3} n\right) \le$$

$\Pr($the recursion tree has $> 8\log_{4/3} n$ levels$) \le$

$$\Pr\left(\bigcup_{i=1}^{P} E_i\right) \le \frac{1}{n^2}$$

or, equivalently

$$\Pr\left(T_{RQS}(n) \le 8n\log_{4/3} n\right) \ge 1 - \frac{1}{n^2}$$

HIGH PROBABILITY!