QuickSort

**QuickSort** (Input $A[1], A[2], \ldots, A[n]$)
1. Pick an element from the array, the so-called pivot
2. If $|A| = 0$ or $|A| = 1$ then
   - return $A$
3. else
   - Create two subarrays $A_1$ and $A_2$ (without the pivot) such that:
     - $A_1$ contains the elements that are smaller than the pivot
     - $A_2$ contains the elements that are greater (or equal) than the pivot
   - QuickSort($A_1$)
   - QuickSort($A_2$)
4. return $A$

- **Example:** Let $A = (2, 8, 9, 1, 7, 5, 6, 3, 4)$ with $A[7] = 6$ as pivot.
  $A_1 = (2, 1, 5, 3, 4)$ and $A_2 = (8, 9, 7)$
- Worst-Case Complexity (number of comparisons) is $\Theta(n^2)$,
  while Average-Case Complexity is $O(n \log n)$.

We will now give a proof of this “well-known” result!
How to pick a good pivot? We don't. Just pick one at random.

This should be your standard answer in this course 😊

Let us analyse QuickSort with random pivots.
1. Assume \( A \) consists of \( n \) different numbers, w.l.o.g., \( \{1, 2, \ldots, n\} \)
2. Let \( H \) be the deepest level where element \( i \) appears in the tree.
   Then the number of comparison is \( H = \sum_{i=1}^{n} H_i \)
3. We will prove that there exists \( C > 0 \) such that
   \[
   P[H \leq C n \log n] \geq 1 - n^{-1}.
   \]
4. Actually, we will prove something slightly stronger:
   \[
   P[\bigcap_{i=1}^{n} \{H_i \leq C \log n\}] \geq 1 - n^{-1}.
   \]

Randomised QuickSort: Analysis (3/4)

- Consider now any element \( i \in \{1, 2, \ldots, n\} \) and construct the path
  \( P = P(i) \) one level by one
- For \( P \) to proceed from level \( k \) to \( k + 1 \), the condition \( s_k > 1 \) is necessary

How far could such a path \( P \) possibly run until we have \( s_k = 1 \)?

- We start with \( s_0 = n \)
- **First Case**, good node: \( s_{k+1} \leq \frac{2}{3} \cdot s_k \)
- **Second Case**, bad node: \( s_{k+1} \leq s_k \)

\[ \Rightarrow \] There are at most \( T = \frac{\log n}{\log(3/2)} < 3 \log n \) many good nodes on any path \( P \).
- Assume \( |P| \geq C \log n \) for \( C := 24 \)
  \[ \Rightarrow \] number of bad vertices in the first \( 24 \log n \) levels is more than \( 21 \log n \).

Let us now upper bound the probability that this "bad event" happens!

Randomised QuickSort: Analysis (4/4)

- Consider the first \( 24 \log n \) vertices of \( P \) to the deepest level of element \( i \).
- For any level \( j \in \{0, 1, \ldots, 24 \log n - 1\} \), define an indicator variable \( X_j \):
  - \( X_j = 1 \) if the node at level \( j \) is good,
  - \( X_j = 0 \) if the node at level \( j \) is bad.
- \( P[X_j = 1 | X_0 = x_0, \ldots, X_{j-1} = x_{j-1}] \leq \frac{2}{3} \)
- \( X := \sum_{j=0}^{24 \log n - 1} X_j \) satisfies relaxed independence assumption (Lecture 2)

**Question**: Edge Case: What if the path \( P \) does not reach level \( j \)?

Randomised QuickSort: Analysis (2/4)

- Let \( P \) be a path from the root to the deepest level of some element
  - A node in \( P \) is called good if the corresponding pivot partitions the array into two subarrays each of size at most \( 2/3 \) of the previous one
  - otherwise, the node is bad
- Further let \( s_i \) be the size of the array at level \( i \) in \( P \).

![Diagram](image_url)
Randomised QuickSort: Analysis (4/4)

- Consider the first $24 \log n$ vertices of $P$ to the deepest level of element $i$.
- For any level $j \in \{0, 1, \ldots, 24 \log n - 1\}$, define an indicator variable $X_j$:
  - $X_j = 1$ if the node at level $j$ is bad,
  - $X_j = 0$ if the node at level $j$ is good.
- $P \{ X_j = 1 \mid X_0 = x_0, \ldots, X_{j-1} = x_{j-1} \} \leq \frac{2}{3}$
- $X := \sum_{j=0}^{24 \log n - 1} X_j$ satisfies relaxed independence assumption (Lecture 2)

**Question:** Edge Case: What if the path $P$ does not reach level $j$?

**Answer:** We can then simply define $X_j$ as 0 (deterministically).

Randomised QuickSort: Final Remarks

- A classical result: expected number of comparison of randomised QUICKSORT is $2n \log n + O(n)$ (see, e.g., book by Mitzenmacher & Upfal)

**Exercise:** [Ex 2-3.6] Our upper bound of $O(n \log n)$ whp also immediately implies a $O(n \log n)$ bound on the expected number of comparisons!

- It is possible to deterministically find the best pivot element that divides the array into two subarrays of the same size.
- The latter requires to compute the median of the array in linear time, which is not easy...
- The presented randomised algorithm for QUICKSORT is much easier to implement!

Outline

Application 2: Randomised QuickSort

Extensions of Chernoff Bounds

Applications of Method of Bounded Differences

Appendix: More on Moment Generating Functions (non-examinable)
Method of Bounded Differences

- Besides sums of independent Bernoulli random variables, sums of independent and bounded random variables are very frequent in applications.
- Unfortunately the distribution of the $X_i$ may be unknown or hard to compute, thus it will be hard to compute the moment-generating function.
- Hoeffding’s Lemma helps us here: You can always consider $X' = X - E[X]$.

### Hoeffding’s Extension Lemma

Let $X$ be a random variable with mean 0 such that $a \leq X \leq b$. Then for all $\lambda \in \mathbb{R}$,

$$
E\left[e^{\lambda X}\right] \leq \exp\left(\frac{(b-a)^2 \lambda^2}{8}\right)
$$

We omit the proof of this lemma!

### Method of Bounded Differences

**Framework**

Suppose, we have independent random variables $X_1, \ldots, X_n$. We want to study the random variable:

$$f(X_1, \ldots, X_n)$$

Some examples:

1. $X = X_1 + \ldots + X_n$ (our setting earlier)
2. In balls into bins, $X_i$ indicates where ball $i$ is allocated, and $f(X_1, \ldots, X_m)$ is the number of empty bins
3. In a randomly generated graph, $X_i$ indicates if the $i$-th edge is present and $f(X_1, \ldots, X_m)$ represents the number of connected components of $G$

In all those cases (and more) we can easily prove concentration of $f(X_1, \ldots, X_n)$ around its mean by the so-called **Method of Bounded Differences**.

### Hoeffding Bounds

**Hoeffding’s Inequality**

Let $X_1, \ldots, X_n$ be independent random variable with mean $\mu_i$ such that $a_i \leq X_i \leq b_i$. Let $X = X_1 + \ldots + X_n$, and let $\mu = E[X] = \sum_{i=1}^n \mu_i$. Then for any $t > 0$

$$
P[X \geq \mu + t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n \lambda^2}\right),
$$

and

$$
P[X \leq \mu - t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n \lambda^2}\right).
$$

**Proof Outline (skipped):**

- Let $X_i' = X_i - \mu_i$ and $X' = X_1' + \ldots + X_n'$, then $P[X \geq \mu + t] = P[X' \geq t]$
- $P[X' \geq t] \leq e^{-\lambda t} \prod_{i=1}^n E\left[e^{\lambda X_i'}\right] \leq \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right)$
- Choose $\lambda = \frac{2t}{\sum_{i=1}^n (b_i - a_i)^2}$ to get the result.

This is not magic! You just need to optimise $\lambda$!

### McDiarmid’s Inequality

Let $X_1, \ldots, X_n$ be independent random variables. Let $f$ be Lipschitz with parameters $c = (c_1, \ldots, c_n)$ if for all $i = 1, 2, \ldots, n$,

$$|f(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i,$$

where $x_i$ and $\tilde{x}_i$ are in the domain of the $i$-th coordinate.

**McDiarmid’s inequality**

Let $X_1, \ldots, X_n$ be independent random variables. Let $f$ be Lipschitz with parameters $c = (c_1, \ldots, c_n)$. Let $X = f(X_1, \ldots, X_n)$. Then for any $t > 0$,

$$
P[X \geq \mu + t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right),
$$

and

$$
P[X \leq \mu - t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).
$$

- Notice the similarity with Hoeffding’s inequality! [Exercise 2/3.14]
- The proof is omitted here (it requires the concept of martingales).
### Application 3: Balls into Bins (again...)

Consider again $m$ balls assigned uniformly at random into $n$ bins.

- Enumerate the balls from 1 to $m$. Ball $i$ is assigned to a random bin $X_i$.
- Let $Z$ be the number of empty bins (after assigning the $m$ balls).
- $Z = Z(X_1, \ldots, X_m)$ and $Z$ is Lipschitz with $c = (1, \ldots, 1)$.
  (If we move one ball to another bin, number of empty bins changes by $\leq 1$.)
- By McDiarmid's inequality, for any $t \geq 0$,
  \[
P \left[ |Z - E[Z]| \geq t \right] \leq 2 \cdot e^{-2t^2/m}.
\]
  This is a decent bound, but for some values of $m$ it is far from tight and stronger bounds are possible through a refined analysis.

### Application 4: Bin Packing

- We are given $n$ items of sizes in the unit interval $[0, 1]$.
- We want to pack those items into the fewest number of unit-capacity bins.
- Suppose the item sizes $X_i$ are independent random variables in $[0, 1]$.
- Let $B = B(X_1, \ldots, X_n)$ be the optimal number of bins.
- The Lipschitz conditions holds with $c = (1, \ldots, 1)$. Why?
- Therefore
  \[
P \left[ |B - E[B]| \geq t \right] \leq 2 \cdot e^{-2t^2/n}.
\]
  This is a typical example where proving concentration is much easier than calculating (or estimating) the expectation!
Moment Generating Functions (non-examinable)

Moment-Generating Function

The moment-generating function of a random variable $X$ is

$$M_X(t) = E\left[e^{tX}\right], \quad \text{where } t \in \mathbb{R}.$$  

Using power series of $e$ and differentiating shows that $M_X(t)$ encapsulates all moments of $X$.

Lemma

1. If $X$ and $Y$ are two r.v.'s with $M_X(t) = M_Y(t)$ for all $t \in (-\delta, +\delta)$ for some $\delta > 0$, then the distributions $X$ and $Y$ are identical.

2. If $X$ and $Y$ are independent random variables, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Proof of 2:

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E\left[e^{tX} \cdot e^{tY}\right] (i) = E\left[e^{tX}\right] \cdot E\left[e^{tY}\right] = M_X(t)M_Y(t).$$