Chapter 13

Randomized Algorithms

By Sariel Har-Peled, December 2, 2021[1] Version: 0.26

13.1. Example: Estimating the median is sublinear time

You are given am array $X[1 \ldots n]$ of real numbers. Think about n as being huge (say, n is in the billions). You would like to estimate the median element of X. Cam one estimate the median number of X. The median of X is the number in X that half the elements of X are smaller than it, and half of them are bigger than it. Formally, an element $x \in X$ (interpret X as a set) has ***rank k*** if $|\{y \in X \mid y \leq x\}| = k$ (we assume here that all the elements of X are distinct).

A natural algorithm is to pick, say, k random numbers from X, say with replacement. Let Y be the resulting random sample. Compute the median of Y (say by sorting). Output the computed median as an estimate of the true median. How close is this to the true median?

13.2. Some Probability

Definition 13.2.1. (Informal.) A ***random variable*** is a measurable function from a probability space to (usually) real numbers. It associates a value with each possible atomic event in the probability space.

Definition 13.2.2. The ***conditional probability*** of X given Y is

$$\mathbb{P}[X = x \,|Y = y] = \frac{\mathbb{P}\Big[(X = x) \cap (Y = y)\Big]}{\mathbb{P}\Big[Y = y\Big]}.$$

An equivalent and useful restatement of this is that

$$\mathbb{P}\Big[(X = x) \cap (Y = y)\Big] = \mathbb{P}\Big[X = x \;\Big|\; Y = y\Big] * \mathbb{P}\Big[Y = y\Big].$$

Definition 13.2.3. Two events X and Y are ***independent***, if $\mathbb{P}[X = x \cap Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$. In particular, if X and Y are independent, then

$$\mathbb{P}\Big[X = x \;\Big|\; Y = y\Big] = \mathbb{P}[X = x].$$

Definition 13.2.4. The ***expectation*** of a random variable X is the average value of this random variable. Formally, if X has a finite (or countable) set of values, it is

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}[X = x],$$

where the summation goes over all the possible values of X.

One of the most powerful properties of expectations is that an expectation of a sum is the sum of expectations.

Lemma 13.2.5 (Linearity of expectation). *For any two random variables X and Y, we have $\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$.*

Proof: For the simplicity of exposition, assume that X and Y receive only integer values. We have that

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_x \sum_y (x+y)\,\mathbb{P}[(X=x)\cap(Y=y)] \\
&= \sum_x \sum_y x * \mathbb{P}[(X=x)\cap(Y=y)] + \sum_x \sum_y y * \mathbb{P}[(X=x)\cap(Y=y)] \\
&= \sum_x x * \sum_y \mathbb{P}[(X=x)\cap(Y=y)] + \sum_y y * \sum_x \mathbb{P}[(X=x)\cap(Y=y)] \\
&= \sum_x x * \mathbb{P}[X=x] + \sum_y y * \mathbb{P}[Y=y] \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

■

Another interesting function is the conditional expectation – that is, it is the expectation of a random variable given some additional information.

Definition 13.2.6. Given random variables X and Y, the ***conditional expectation*** of X given Y, is the quantity $\mathbb{E}\left[X \,\middle|\, Y\right]$. Specifically, you are given the value y of the random variable Y, and the condition expectation of X given Y is

$$\mathbb{E}\left[X \,\middle|\, Y\right] = \mathbb{E}\left[X \,\middle|\, Y=y\right] = \sum_x x * \mathbb{P}[X=x \mid Y=y].$$

Note, that for a random variable X, the expectation $\mathbb{E}[X]$ is a number. On the other hand, the conditional probability $f(y) = \mathbb{E}[X \mid Y=y]$ is a function. The key insight why conditional probability is the following.

Lemma 13.2.7. *For any two random variables X and Y (not necessarily independent), we have that $\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right]$.*

Proof: We use the definitions carefully:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[X \mid Y]\right] &= \mathop{\mathbb{E}}_y\left[\mathbb{E}[X \mid Y=y]\right] = \mathop{\mathbb{E}}_y\left[\sum_x x * \mathbb{P}[X=x \mid Y=y]\right] \\
&= \sum_y \mathbb{P}[Y=y] * \left(\sum_x x * \mathbb{P}[X=x \mid Y=y]\right) \\
&= \sum_y \mathbb{P}[Y=y] * \left(\sum_x x * \frac{\mathbb{P}[(X=x)\cap(Y=y)]}{\mathbb{P}[Y=y]}\right) \\
&= \sum_y \sum_x x * \mathbb{P}[(X=x)\cap(Y=y)] = \sum_x \sum_y x * \mathbb{P}[(X=x)\cap(Y=y)] \\
&= \sum_x x * \left(\sum_y \mathbb{P}[(X=x)\cap(Y=y)]\right) = \sum_x x * \mathbb{P}[X=x] = \mathbb{E}[X].
\end{aligned}$$

■

13.3. Preliminaries

Definition 13.3.1 (Variance and Standard Deviation). For a random variable X, let

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}\left[X^2\right] - \mu_X^2$$

denote the ***variance*** of X, where $\mu_X = \mathbb{E}[X]$. Intuitively, this tells us how concentrated is the distribution of X. The ***standard deviation*** of X, denoted by σ_X is the quantity $\sqrt{\mathbb{V}[X]}$.

Observation 13.3.2. *(i) For any constant $c \geq 0$, we have $\mathbb{V}\left[cX\right] = c^2\,\mathbb{V}\left[X\right]$.*
(ii) For X and Y independent variables, we have $\mathbb{V}\left[X + Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$.

Definition 13.3.3 (Bernoulli distribution). Assume, that one flips a coin and get 1 (heads) with probability p, and 0 (i.e., tail) with probability $q = 1 - p$. Let X be this random variable. The variable X is has ***Bernoulli distribution*** with parameter p.

We have that $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$, and

$$\mathbb{V}\left[X\right] = \mathbb{E}\left[X^2\right] - \mu_X^2 = \mathbb{E}\left[X^2\right] - p^2 = p - p^2 = p(1 - p) = pq.$$

Definition 13.3.4 (Binomial distribution). Assume that we repeat a Bernoulli experiment n times (independently!). Let $X_1, \ldots, X_n$ be the resulting random variables, and let $X = X_1 + \cdots + X_n$. The variable X has the ***binomial distribution*** with parameters n and p. We denote this fact by $X \sim \mathrm{Bin}(n, p)$. We have

$$b(k; n, p) = \mathbb{P}\left[X = k\right] = \binom{n}{k} p^k q^{n-k}.$$

Also, $\mathbb{E}[X] = np$, and $\mathbb{V}[X] = \mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] = npq$.

Observation 13.3.5. *Let $C_1, \ldots, C_n$ be random events (not necessarily independent). Than*

$$\mathbb{P}\left[\cup_{i=1}^{n} C_i\right] \leq \sum_{i=1}^{n} \mathbb{P}[C_i].$$

(This is usually referred to as the **union bound***.) If $C_1, \ldots, C_n$ are* disjoint *events then*

$$\mathbb{P}\left[\cup_{i=1}^{n} C_i\right] = \sum_{i=1}^{n} \mathbb{P}[C_i].$$

13.3.1. Geometric distribution

Definition 13.3.6. Consider a sequence $X_1, X_2, \ldots$ of independent Bernoulli trials with probability p for success. Let X be the number of trials one has to perform till encountering the first success. The distribution of X is ***geometric distribution*** with parameter p. We denote this by $X \sim \mathrm{Geom}(p)$.

Lemma 13.3.7. *For a variable $X \sim \mathrm{Geom}(p)$, we have, for all i, that $\mathbb{P}[X = i] = (1 - p)^{i-1} p$. Furthermore, $\mathbb{E}[X] = 1/p$ and $\mathbb{V}[X] = (1 - p)/p^2$.*

Proof: The proof of the expectation and variance is included for the sake of completeness, and the reader is of course encouraged to skip (reading) this proof. So, let $f(x) = \sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$, and observe that $f'(x) = \sum_{i=1}^{\infty} ix^{i-1} = (1-x)^{-2}$. As such, we have

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i\,(1-p)^{i-1}p = pf'(1-p) = \frac{p}{(1-(1-p))^2} = \frac{1}{p},$$

$$\text{and } \mathbb{V}[X] = \mathbb{E}\big[X^2\big] - \frac{1}{p^2} = \sum_{i=1}^{\infty} i^2\,(1-p)^{i-1}p - \frac{1}{p^2}. = p + p(1-p)\sum_{i=2}^{\infty} i^2\,(1-p)^{i-2} - \frac{1}{p^2}.$$

Observe that

$$f''(x) = \sum_{i=2}^{\infty} i(i-1)x^{i-2} = \big((1-x)^{-1}\big)'' = \frac{2}{(1-x)^3}.$$

As such, we have that

$$\begin{aligned}\Delta(x) &= \sum_{i=2}^{\infty} i^2x^{i-2} = \sum_{i=2}^{\infty} i(i-1)x^{i-2} + \sum_{i=2}^{\infty} ix^{i-2} = f''(x) + \frac{1}{x}\sum_{i=2}^{\infty} ix^{i-1} = f''(x) + \frac{1}{x}(f'(x)-1)\\ &= \frac{2}{(1-x)^3} + \frac{1}{x}\left(\frac{1}{(1-x)^2} - 1\right) = \frac{2}{(1-x)^3} + \frac{1}{x}\left(\frac{1-(1-x)^2}{(1-x)^2}\right) = \frac{2}{(1-x)^3} + \frac{1}{x}\cdot\frac{x(2-x)}{(1-x)^2}\\ &= \frac{2}{(1-x)^3} + \frac{2-x}{(1-x)^2}.\end{aligned}$$

As such, we have that

$$\begin{aligned}\mathbb{V}[X] &= p + p(1-p)\Delta(1-p) - \frac{1}{p^2} = p + p(1-p)\left(\frac{2}{p^3} + \frac{1+p}{p^2}\right) - \frac{1}{p^2} = p + \frac{2(1-p)}{p^2} + \frac{1-p^2}{p} - \frac{1}{p^2}\\ &= \frac{p^3 + 2(1-p) + p - p^3 - 1}{p^2} = \frac{1-p}{p^2}.\end{aligned}$$

■

13.3.2. Some needed math

Lemma 13.3.8. *For any positive integer n, we have:*

(i) $1 + x \leq e^x$, *for all* x.

(ii) $(1+1/n)^n \leq e \leq \big(1+1/n\big)^{n+1}$.

(iii) $(1-1/n)^n \leq \frac{1}{e} \leq (1-1/n)^{n-1}$.

(iv) $n! \geq (n/e)^n$.

(v) *For any $k \leq n$, we have:* $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

Proof: (i) Let $h(x) = e^x - 1 - x$. Observe that $h'(x) = e^x - 1$, and $h''(x) = e^x > 0$, for all x. That is $h(x)$ is a convex function. It achieves its minimum at $h'(x) = 0 \implies e^x = 1$, which is true for $x = 0$. For $x = 0$, we have that $h(0) = e^0 - 1 - 0 = 0$. That is, $h(x) \geq 0$ for all x, which implies that $e^x \geq 1 + x$, see Figure 13.1.

(ii, iii) Indeed, $1 + 1/n \leq \exp(1/n)$ and $(1-1/n)^n \leq \exp(-1/n)$, by (i). As such

$$(1+1/n)^n \leq \exp(n(1/n)) = e \qquad \text{and} \qquad (1-1/n)^n \leq \exp(n(-1/n)) = \frac{1}{e},$$

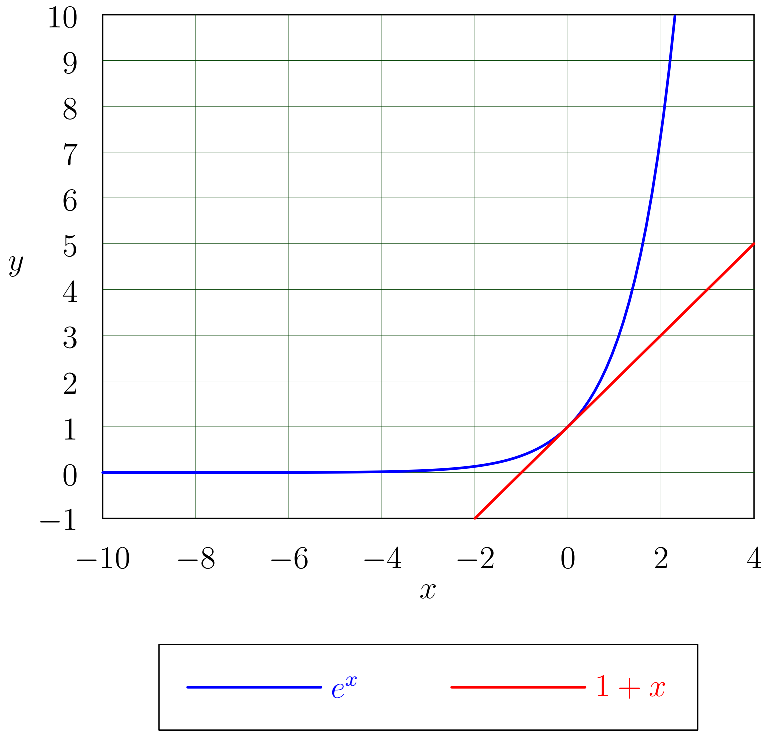

Figure 13.1

which implies the left sides of (ii) and (iii). These are equivalent to

$$\frac{1}{e} \leq \left(\frac{n}{n+1}\right)^n = \left(1 - \frac{1}{n+1}\right)^n \qquad \text{and} \qquad e \leq \left(1 + \frac{1}{n-1}\right)^n,$$

which are the right side of (iii) [by replacing $n+1$ by n], and the right side of (ii) [by replacing n by $n+1$].

(iv) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

(v) For any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$ since $kn - n = n(k-1) \leq k(n-1) = kn - k$. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k-1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-i}{k-i} \cdots \frac{n-k+1}{1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

As for the other direction, we have

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{(k/e)^k} = \left(\frac{ne}{k}\right)^k,$$

by (iii). ■

13.4. Sorting Nuts and Bolts

Problem 13.4.1 (Sorting Nuts and Bolts). You are given a set of n nuts and n bolts. Every nut have a matching bolt, and all the n pairs of nuts and bolts have different sizes. Unfortunately, you get the nuts and bolts separated from each other and you have to match the nuts to the bolts. Furthermore, given a nut and a bolt, all you can do is to try and match one bolt against a nut (i.e., you can not compare two nuts to each other, or two bolts to each other).

When comparing a nut to a bolt, either they match, or one is smaller than other (and you known the relationship after the comparison).

How to match the n nuts to the n bolts quickly? Namely, while performing a small number of comparisons.

The naive algorithm is of course to compare each nut to each bolt, and match them together. This would require a quadratic number of comparisons. Another option is to sort the nuts by size, and the bolts by size and then "merge" the two ordered sets, matching them by size. The only problem is that we can not sorts only the nuts, or only the bolts, since we can not compare them to each other. Indeed, we sort the two sets simultaneously, by simulating QuickSort. The resulting algorithm is depicted on the right.

MatchNutsAndBolts(N: nuts, B: bolts)
  Pick a random nut n_{pivot} from N
  Find its matching bolt b_{pivot} in B
  $B_L \leftarrow$ All bolts in B smaller than n_{pivot}
  $N_L \leftarrow$ All nuts in N smaller than b_{pivot}
  $B_R \leftarrow$ All bolts in B larger than n_{pivot}
  $N_R \leftarrow$ All nuts in N larger than b_{pivot}
  MatchNutsAndBolts(N_R,B_R)
  MatchNutsAndBolts(N_L,B_L)

13.4.1. Running time analysis

Definition 13.4.2. Let $\mathcal{RT}$ denote the random variable which is the running time of the algorithm. Note, that the running time is a random variable as it might be different between different executions on the *same input*.

Definition 13.4.3. For a randomized algorithm, we can speak about the expected running time. Namely, we are interested in bounding the quantity $\mathbb{E}[\mathcal{RT}]$ for the worst input.

Definition 13.4.4. The ***expected running-time*** of a randomized algorithm for input of size n is

$$T(n) = \max_{U \text{ is an input of size } n} \mathbb{E}\big[\mathcal{RT}(U)\big],$$

where $\mathcal{RT}(U)$ is the running time of the algorithm for the input U.

Definition 13.4.5. The ***rank*** of an element x in a set S, denoted by $\operatorname{rank}(x)$, is the number of elements in S of size smaller or equal to x. Namely, it is the location of x in the sorted list of the elements of S.

Theorem 13.4.6. *The expected running time of MatchNutsAndBolts (and thus also of QuickSort) is $T(n) = O(n \log n)$, where n is the number of nuts and bolts. The worst case running time of this algorithm is $O(n^2)$.*

Proof: Clearly, we have that $\mathbb{P}\left[\mathrm{rank}(n_{pivot}) = k\right] = \frac{1}{n}$. Furthermore, if the rank of the pivot is k then

$$\begin{aligned} T(n) &= \mathop{\mathbb{E}}_{k=\mathrm{rank}(n_{pivot})}[O(n) + T(k-1) + T(n-k)] = O(n) + \mathop{\mathbb{E}}_{k}[T(k-1) + T(n-k)] \\ &= T(n) = O(n) + \sum_{k=1}^{n} \mathbb{P}[Rank(Pivot) = k] * (T(k-1) + T(n-k)) \\ &= O(n) + \sum_{k=1}^{n} \frac{1}{n} \cdot (T(k-1) + T(n-k)), \end{aligned}$$

by the definition of expectation. It is not easy to verify that the solution to the recurrence $T(n) = O(n) + \sum_{k=1}^{n} \frac{1}{n} \cdot (T(k-1) + T(n-k))$ is $O(n \log n)$. ■

13.4.1.1. Alternative incorrect solution

The algorithm MatchNutsAndBolts is lucky if $\frac{n}{4} \leq \mathrm{rank}(n_{pivot}) \leq \frac{3}{4}n$. Thus, $\mathbb{P}[\text{“lucky”}] = 1/2$. Intuitively, for the algorithm to be fast, we want the split to be as balanced as possible. The less balanced the cut is, the worst the expected running time. As such, the “Worst” lucky position is when $\mathrm{rank}(n_{pivot}) = n/4$ and we have that

$$T(n) \leq O(n) + \mathbb{P}[\text{“lucky”}] * (T(n/4) + T(3n/4)) + \mathbb{P}[\text{“unlucky”}] * T(n).$$

Namely, $T(n) = O(n) + \frac{1}{2} * \left(T(\frac{n}{4}) + T(\frac{3}{4}n)\right) + \frac{1}{2}T(n)$. Rewriting, we get the recurrence $T(n) = O(n) + T(n/4) + T((3/4)n)$, and its solution is $O(n \log n)$.

While this is a very intuitive and elegant solution that bounds the running time of QuickSort, it is also incomplete. The interested reader should try and make this argument complete. After completion the argument is as involved as the previous argument. Nevertheless, this argumentation gives a good back of the envelope analysis for randomized algorithms which can be applied in a lot of cases.

13.4.2. What are randomized algorithms?

Randomized algorithms are algorithms that use random numbers (retrieved usually from some unbiased source of randomness [say a library function that returns the result of a random coin flip]) to make decisions during the executions of the algorithm. The running time becomes a random variable. Analyzing the algorithm would now boil down to analyzing the behavior of the random variable $\mathcal{RT}(n)$, where n denotes the size of the input.In particular, the expected running time $\mathbb{E}[\mathcal{RT}(n)]$ is a quantity that we would be interested in.

It is useful to compare the expected running time of a randomized algorithm, which is

$$T(n) = \max_{U \text{ is an input of size } n} \mathbb{E}[\mathcal{RT}(U)],$$

to the worst case running time of a deterministic (i.e., not randomized) algorithm, which is

$$T(n) = \max_{U \text{ is an input of size } n} \mathcal{RT}(U),$$

Caveat Emptor:[②]Note, that a randomized algorithm might have exponential running time in the worst case (or even unbounded) while having good expected running time. For example, consider the algorithm FlipCoins depicted on the right. The expected running time of FlipCoins is a geometric random variable with probability 1/2, as such we have that $\mathbb{E}[\mathcal{RT}(FlipCoins)] = O(2)$. However, FlipCoins can run forever if it always gets 1 from the RandBit function.

```
FlipCoins
    while RandBit= 1 do
        nothing;
```

This is of course a ludicrous argument. Indeed, the probability that FlipCoins runs for long decreases very quickly as the number of steps increases. It can happen that it runs for long, but it is extremely unlikely.

Definition 13.4.7. The running time of a randomized algorithm Alg is $O(f(n))$ with ***high probability*** if

$$\mathbb{P}[\mathcal{RT}(Alg(n)) \geq c \cdot f(n)] = o(1).$$

Namely, the probability of the algorithm to take more than $O(f(n))$ time decreases to 0 as n goes to infinity. In our discussion, we would use the following (considerably more restrictive definition), that requires that

$$\mathbb{P}[\mathcal{RT}(Alg(n)) \geq c \cdot f(n)] \leq \frac{1}{n^d},$$

where c and d are appropriate constants. For technical reasons, we also require that $\mathbb{E}[\mathcal{RT}(Alg(n))] = O(f(n))$.

13.5. Analyzing QuickSort

The previous analysis works also for QuickSort. However, there is an alternative analysis which is also very interesting and elegant. Let $a_1, ..., a_n$ be the n given numbers (in sorted order – as they appear in the output).

It is enough to bound the number of comparisons performed by QuickSort to bound its running time, as can be easily verified. Observe, that two specific elements are compared to each other by QuickSort at most once, because QuickSort performs only comparisons against the pivot, and after the comparison happen, the pivot does not being passed to the two recursive subproblems.

Let X_{ij} be an indicator variable if QuickSort compared a_i to a_j in the current execution, and zero otherwise. The number of comparisons performed by QuickSort is **exactly** $Z = \sum_{i<j} X_{ij}$.

Observation 13.5.1. *The element a_i is compared to a_j iff one of them is picked to be the pivot and they are still in the same subproblem.*

Also, we have that $\mu = \mathbb{E}\left[X_{ij}\right] = \mathbb{P}\left[X_{ij} = 1\right]$. To quantify this probability, observe that if the pivot is smaller than a_i or larger than a_j then the subproblem still contains the block of elements $a_i, \ldots, a_j$. Thus, we have that

$$\mu = \mathbb{P}\left[a_i \text{ or } a_j \text{ is first pivot} \in a_i, \ldots, a_j\right] = \frac{2}{j - i + 1}.$$

[②]Caveat Emptor - let the buyer beware (i.e., one buys at one's own risk)

```
QuickSelect(X, k)
   // Input:  X = {x_1, ..., x_n} numbers, k.
   // Assume x_1, ..., x_n are all distinct.
   // Task:  Return kth smallest number in X.
   y ← random element of X.
   r ← rank of y in X.
   if r = k then return y
   X_< = all elements in X < than y
   X_> = all elements in X > than y
   // By assumption |X_<| + |X_>| + 1 = |X|.
   if r < k then
      return QuickSelect( X_>, k − r )
   else
      return QuickSelect( X_≤, k )
```

Figure 13.2: QuickSelect pseudo-code.

Another (and hopefully more intuitive) explanation for the above phenomena is the following: Imagine, that before running QuickSort we choose for every element a random priority, which is a real number in the range $[0, 1]$. Now, we reimplement QuickSort such that it always pick the element with the lowest random priority (in the given subproblem) to be the pivot. One can verify that this variant and the standard implementation have the same running time. Now, a_i gets compares to a_j if and only if all the elements $a_{i+1}, \ldots, a_{j-1}$ have random priority larger than both the random priority of a_i and the random priority of a_j. But the probability that one of two elements would have the lowest random-priority out of $j - i + 1$ elements is $2 * 1/(j - i + 1)$, as claimed.

Thus, the running time of QuickSort is

$$\mathbb{E}\big[\mathcal{RT}(n)\big] = \mathbb{E}\left[\sum_{i<j} X_{ij}\right] = \sum_{i<j} \mathbb{E}\big[X_{ij}\big] = \sum_{i<j} \frac{2}{j-i+1} = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{1}{j-i+1}$$
$$= 2\sum_{i=1}^{n-1}\sum_{\Delta=2}^{n-i+1} \frac{1}{\Delta} \leq 2\sum_{i=1}^{n-1}\sum_{\Delta=1}^{n} \frac{1}{\Delta} \leq 2\sum_{i=1}^{n-1} H_n = 2nH_n.$$

by linearity of expectations, where $H_n = \sum_{i=1}^{n} \frac{1}{i} \leq \ln n + 1$ is the nth harmonic number,

As we will see in the near future, the running time of QuickSort is $O(n \log n)$ with high-probability. We need some more tools before we can show that.

13.6. QuickSelect – median selection in linear time

Consider the problem of given a set X of n numbers, and a parameter k, to output the kth smallest number (which is the number with ***rank*** k in X). This can be easily be done by modifying QuickSort only to perform one recursive call. See Figure 13.2 for a pseud-code of the resulting algorithm.

Intuitively, at each iteration of QuickSelect the input size shrinks by a constant factor, leading to a linear time algorithm.

Theorem 13.6.1. *Given a set X of n numbers, and any integer k, the expected running time of QuickSelect(X, n) is $O(n)$.*

Proof: Let $X_1 = X$, and X_i be the set of numbers in the ith level of the recursion. Let y_i and r_i be the random element and its rank in X_i, respectively, in the ith iteration of the algorithm. Finally, let $n_i = |X_i|$. Observe that the probability that the pivot y_i is in the "middle" of its subproblem is

$$\alpha = \mathbb{P}\left[\frac{n_i}{4} \leq r_i \leq \frac{3}{4} n_i\right] \geq \frac{1}{2},$$

and if this happens then

$$n_{i+1} \leq \max(r_i - 1, n_i - r_i) \leq \frac{3}{4} n_i.$$

We conclude that

$$\begin{aligned}
\mathbb{E}[n_{i+1} \mid n_i] &\leq \mathbb{P}[y_i \text{ in the middle}] \frac{3}{4} n_i + \mathbb{P}[y_i \text{ not in the middle}] n_i \\
&\leq \alpha \frac{3}{4} n_i + (1-\alpha) n_i = n_i(1 - \alpha/4) \leq n_i(1 - (1/2)/4) = (7/8) n_i.
\end{aligned}$$

Now, we have that

$$\begin{aligned}
m_{i+1} &= \mathbb{E}[n_{i+1}] = \mathbb{E}[\mathbb{E}[n_{i+1} \mid n_i]] \leq \mathbb{E}[(7/8) n_i] = (7/8)\, \mathbb{E}[n_i] = (7/8) m_i \\
&= (7/8)^i m_0 = (7/8)^i n,
\end{aligned}$$

since for any two random variables we have that $\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}\Big[X \,\Big|\, Y\Big]\Big]$. In particular, the expected running time of QuickSelect is proportional to

$$\mathbb{E}\left[\sum_i n_i\right] = \sum_i \mathbb{E}[n_i] \leq \sum_i m_i = \sum_i (7/8)^i n = O(n),$$

as desired. ■

Bibliography