Randomized Algorithms, Lecture 3

Jacob Holm (jaho@di.ku.dk)

April 30th 2019

Good Afternoon.

Today's Lecture

Moments and Deviations

- Occupancy problems
- Markov's and Chebyshev's Inequalities
- Randomized Selection
- Two-point Sampling

We have previously talked about expected values.
Now we want to show that things happen with high probability.
Occupancy problems = Balls and bins.
Markov + Chebyshev = Tail inequalities = Probability that X deviates by a given amount from its expectation..
Today's algorithm gives an efficient way to find the kth smallest element of a set. We will use tail inequalities to prove that it is fast with high probability.
Finally, we show a technique for getting more out our random bits.

Occupancy Problems

Imagine we have m indistinguishable objects ("balls"), that we randomly assign to n distinct classes ("bins").

- What is expected maximum number of balls in any bin?
- What is the expected number of bins with k balls?

These are called *occupancy problems*.

Occupancy Problems

Imagine we have m indistinguishable objects ("balls"), that we randomly assign to n distinct classes ("bins").

- What is expected maximum number of balls in any bin?
- What is the expected number of bins with k balls?

These are called *occupancy problems*.

Occupancy Problems

Imagine we have m indistinguishable objects ("balls"), that we randomly assign to n distinct classes ("bins").

- What is expected maximum number of balls in any bin?
- What is the expected number of bins with k balls?

These are called *occupancy problems*.

Occupancy Problems

Imagine we have m indistinguishable objects ("balls"), that we randomly assign to n distinct classes ("bins").

- What is expected maximum number of balls in any bin?
- What is the expected number of bins with k balls?

These are called *occupancy problems*.

Occupancy Problems

Let $m = n \geq 3$, and for $i = 1, \ldots, n$ let X_i be the number of balls in the ith bin.

We want to find k such that, with very high probability, no bin contains more than k balls.

Let $\mathcal{E}_j(k)$ be the event that bin j contains at least k balls ($X_j \geq k$). First consider $\mathcal{E}_1(k)$.

Occupancy Problems

This is the binomial distribution.

$$
\begin{aligned}
\Pr[X_1 = i] &= \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i} \\
&\le \binom{n}{i}\left(\frac{1}{n}\right)^i \le \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \\
\Pr[\mathcal{E}_1(k)] &\le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right) \\
&\le \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)
\end{aligned}
$$

Occupancy Problems

$$\Pr[X_1 = i] = \binom{n}{i}\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{n-i}$$

$$\leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i \left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$$

$$\Pr[\mathcal{E}_1(k)] \leq \sum_{i=k}^{n} \left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \cdots\right)$$

$$\leq \left(\frac{e}{k}\right)^k \left(\frac{1}{1 - \frac{e}{k}}\right)$$

Occupancy Problems

$$\Pr[X_1 = i] = \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i}$$

$$\leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$$

$$\Pr[\mathcal{E}_1(k)] \leq \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right)$$

$$\leq \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)$$

Occupancy Problems

$$
\begin{aligned}
\Pr[X_1 = i] &= \binom{n}{i}\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{n-i} \\
&\leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i \left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i
\end{aligned}
$$

$$
\begin{aligned}
\Pr[\mathcal{E}_1(k)] &\leq \sum_{i=k}^{n} \left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \cdots\right) \\
&\leq \left(\frac{e}{k}\right)^k \left(\frac{1}{1 - \frac{e}{k}}\right)
\end{aligned}
$$

Occupancy Problems

$$\Pr[X_1 = i] = \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i}$$

$$\leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$$

$$\Pr[\mathcal{E}_1(k)] \leq \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right)$$

$$\leq \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)$$

Occupancy Problems

$$
\begin{aligned}
\Pr[X_1 = i] &= \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i} \\
&\le \binom{n}{i}\left(\frac{1}{n}\right)^i \le \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \\
\Pr[\mathcal{E}_1(k)] &\le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right) \\
&\le \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)
\end{aligned}
$$

Occupancy Problems

$$
\begin{aligned}
\Pr[X_1 = i] &= \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i} \\
&\le \binom{n}{i}\left(\frac{1}{n}\right)^i \le \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \\
\Pr[\mathcal{E}_1(k)] &\le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right) \\
&\le \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)
\end{aligned}
$$

Occupancy Problems

$$
\begin{aligned}
\Pr[X_1 = i] &= \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i} \\
&\le \binom{n}{i}\left(\frac{1}{n}\right)^i \le \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \\
\Pr[\mathcal{E}_1(k)] &\le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right) \\
&\le \left(\frac{e}{k}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right)
\end{aligned}
$$

Occupancy Problems

Let $k^\star = \min\{n+1, \lceil \frac{2e}{e-1}\frac{\ln n}{\ln\ln n}\rceil\} \le \lceil 3.164 \frac{\ln n}{\ln\ln n}\rceil$, then

$$\Pr[\mathcal{E}_1(k^\star)] \le \left(\frac{e}{k^\star}\right)^{k^\star}\left(\frac{1}{1-\frac{e}{k^\star}}\right) \le n^{-2}$$

The same holds for all i, so

$$\Pr[\mathcal{E}_i(k^\star)] \le \left(\frac{e}{k^\star}\right)^{k^\star}\left(\frac{1}{1-\frac{e}{k^\star}}\right) \le n^{-2}$$

$$\Pr[\cup_{i=1}^n \mathcal{E}_i(k^\star)] \le \sum_{i=1}^n \Pr[\mathcal{E}_i(k^\star)] \le \frac{1}{n}$$

Book claims $k^\star = \lceil \frac{e \ln n}{\ln\ln n}\rceil$, but this fails for e.g. $n = 61$ and $n \ge 1895$.

Know that for $n > 1$, $\left(\frac{\ln n}{\ln\ln n}\right)^{\left(\frac{e}{e-1}\frac{\ln n}{\ln\ln n}\right)} \ge n$ (tight for $n = e^{e^e}$), and $\left(\frac{2}{e-1}\right)^k\left(\frac{1}{1-\frac{e}{k}}\right) \le 1$ for $k \ge 6$.

Let $k = \frac{2e}{e-1}\frac{\ln n}{\ln\ln n}$, so $k^\star = \max\{n+1, \lceil k\rceil\}$.

$$\begin{aligned}\left(\frac{e}{k}\right)^k = \left(\frac{k}{e}\right)^{-k} &= \left(\frac{2}{e-1}\right)^{-k}\left(\left(\frac{\ln n}{\ln\ln n}\right)^{\left(\frac{e}{e-1}\frac{\ln n}{\ln\ln n}\right)}\right)^{-2} \\ &\le \left(\frac{2}{e-1}\right)^{-k} n^{-2}\end{aligned}$$

For $n > e$, we have $k > 5$ so $k^\star \ge 6$ and

$$\left(\frac{e}{k^\star}\right)^{k^\star}\left(\frac{1}{1-\frac{e}{k^\star}}\right) \le \left(\frac{2}{e-1}\right)^{-k^\star} n^{-2}\left(\frac{1}{1-\frac{e}{k^\star}}\right) \le n^{-2}$$

Occupancy Problems

Let $k^\star = \min\{n+1, \lceil \frac{2e}{e-1} \frac{\ln n}{\ln\ln n} \rceil\} \leq \lceil 3.164 \frac{\ln n}{\ln \ln n} \rceil$, then

$$\Pr[\mathcal{E}_1(k^\star)] \leq \left(\frac{e}{k^\star}\right)^{k^\star} \left(\frac{1}{1-\frac{e}{k^\star}}\right) \leq n^{-2}$$

The same holds for all i, so

$$\Pr[\mathcal{E}_i(k^\star)] \leq \left(\frac{e}{k^\star}\right)^{k^\star} \left(\frac{1}{1-\frac{e}{k^\star}}\right) \leq n^{-2}$$

$$\Pr[\cup_{i=1}^{n} \mathcal{E}_i(k^\star)] \leq \sum_{i=1}^{n} \Pr[\mathcal{E}_i(k^\star)] \leq \frac{1}{n}$$

The last step uses an important principle. The Probability of a union is upper bounded by the sum of probabilities.

Occupancy Problems

We have shown:

Theorem

With probability at least $1 - \frac{1}{n}$*, every bin has less than* $k^\star = \min\{n + 1, \lceil \frac{2e}{e-1} \frac{\ln n}{\ln \ln n} \rceil\}$ *balls in it.*

Birthday Problem

Suppose m balls are randomly assigned to n bins. What is the probability that all balls land in distinct bins?

For $n = 365$ the question can be interpreted as "how large must a group of people be before it is likely two people have the same birthday"?

Birthday Problem

Why do we start with $i = 2$?
For $n = 365$ that is $m \geq 29$.

Let $\mathcal{E}_i$ be the event that the ith ball lands in an empty bin. From first lecture we know:

$$\begin{aligned}
\Pr[\cap_{i=2}^{m} \mathcal{E}_i] &= \Pr[\mathcal{E}_2] \Pr[\mathcal{E}_3 \mid \mathcal{E}_2] \cdots \Pr[\mathcal{E}_m \mid \cap_{i=2}^{m-1}] \\
&= \prod_{i=2}^{m} \left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=2}^{m} e^{-\frac{i-1}{n}} = e^{-\frac{m(m-1)}{2n}}
\end{aligned}$$

For $m \geq \lceil \sqrt{2n} + 1 \rceil$, the probability that all are distinct is at most $1/e$.

Markov's Inequality

Theorem

Let Y be a random variable taking only non-negative values. Then for all $t > 0$:

$$\Pr[Y \geq t] \leq \frac{\mathbb{E}[Y]}{t}$$

equivalently, for $k > 0$:

$$\Pr[Y \geq k\,\mathbb{E}[Y]] \leq \frac{1}{k}$$

Markov's Inequality, Proof

Let Z be indicator variable for the event $Y \geq t$. Then $Z \leq \frac{Y}{t}$, and thus

$$\Pr[Y \geq t] = \mathbb{E}[Z] \leq \mathbb{E}\left[\frac{Y}{t}\right] = \frac{\mathbb{E}[Y]}{t}$$

Setting $t = k\,\mathbb{E}[Y]$ we get

$$\Pr[Y \geq k\,\mathbb{E}[Y]] = \Pr[Y \geq t] = \frac{\mathbb{E}[Y]}{t} = \frac{1}{k}$$

Chebyshev's Inequality

Given a random variable X with expectation $\mathbb{E}[X] = \mu_X$, define its *variance* as $\sigma_X^2 := \mathbb{E}[(X - \mu_X)^2]$, and its *standard deviation* as $\sigma_X := \sqrt{\mathbb{E}[(X - \mu_X)^2]}$.

Theorem

Let X be a random variable with expectation μ_X and standard deviation σ_X. Then for all $t > 0$:

$$\Pr[|X - \mu_X| \geq t\sigma_X] \leq \frac{1}{t^2}$$

Chebyshev's Inequality, Proof

Let $k = t^2$ and $Y = (X - \mu_X)^2$.
Then $\sigma_X^2 = \mathbb{E}[Y]$ (by definition) and

$$\begin{aligned}
\Pr[|X - \mu_X| \geq t\sigma_X] &= \Pr[(X - \mu_X)^2 \geq t^2\sigma_X^2] \\
&= \Pr[Y \geq k\,\mathbb{E}[Y]] \\
&\leq \frac{1}{k} \\
&= \frac{1}{t^2}
\end{aligned}$$

Summing 2-Independent Variances

Random variables $X_1, \ldots, X_m \in \mathcal{X}$ are *pairwise independent* iff for all $i \neq j$ and all $x, y \in \mathcal{X}$, $\Pr[X_i = x | X_j = y] = \Pr[X_i = x]$.

Lemma

Let $X_1, \ldots, X_m$ be pairwise independent random variables, and let $X = \sum_{i=1}^{m} X_i$. Then $\sigma_X^2 = \sum_{i=1}^{m} \sigma_{X_i}^2$.

While we are on the topic of variance, here is a small Lemma, which we will be needing later today. This is Lemma 3.4 and Exercise 3.8 in the book, and is a generalization of proposition C.9.

Summing 2-Independent Variances

Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_{i=1}^m \mu_i$. By definition,

$$\begin{aligned}
\sigma_X^2 = \mathbb{E}[(X-\mu)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^m (X_i - \mu_i)\right)^2\right] \\
&= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] \\
&= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[X_i-\mu_i]\,\mathbb{E}[X_j-\mu_j] \\
&= \sum_{i=1}^m \sigma_{X_i}^2 + 2\sum_{i<j} 0\cdot 0 \qquad \square
\end{aligned}$$

Summing 2-Independent Variances

Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_{i=1}^m \mu_i$. By definition,

$$\sigma_X^2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}\left[\left(\sum_{i=1}^m (X_i - \mu_i)\right)^2\right]$$

$$= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)]$$

$$= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[X_i-\mu_i]\,\mathbb{E}[X_j-\mu_j]$$

$$= \sum_{i=1}^m \sigma_{X_i}^2 + 2\sum_{i<j} 0\cdot 0 \qquad \square$$

Summing 2-Independent Variances

Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_{i=1}^m \mu_i$. By definition,

$$\begin{aligned}
\sigma_X^2 = \mathbb{E}[(X-\mu)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^m (X_i - \mu_i)\right)^2\right] \\
&= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] \\
&= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[X_i-\mu_i]\,\mathbb{E}[X_j-\mu_j] \\
&= \sum_{i=1}^m \sigma_{X_i}^2 + 2\sum_{i<j} 0\cdot 0 \qquad \square
\end{aligned}$$

Uses $\left(\sum_{i=1}^m a_i\right)^2 = \sum_{i=1}^m a_i^2 + 2\sum_{i<j} a_i a_j$ and linearity of expectation.

Summing 2-Independent Variances

Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_{i=1}^m \mu_i$. By definition,

$$\sigma_X^2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}\left[\left(\sum_{i=1}^m (X_i - \mu_i)\right)^2\right]$$

$$= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)]$$

$$= \sum_{i=1}^m \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[X_i-\mu_i]\,\mathbb{E}[X_j-\mu_j]$$

$$= \sum_{i=1}^m \sigma_{X_i}^2 + 2\sum_{i<j} 0\cdot 0 \qquad \square$$

Uses that for *independent* X, Y,

$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. (Proposition C.6 in the book)

Summing 2-Independent Variances

Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_{i=1}^{m} \mu_i$. By definition,

$$
\begin{aligned}
\sigma_X^2 = \mathbb{E}[(X-\mu)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{m}(X_i-\mu_i)\right)^2\right] \\
&= \sum_{i=1}^{m} \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] \\
&= \sum_{i=1}^{m} \mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j} \mathbb{E}[X_i-\mu_i]\,\mathbb{E}[X_j-\mu_j] \\
&= \sum_{i=1}^{m} \sigma_{X_i}^2 + 2\sum_{i<j} 0\cdot 0 \qquad \square
\end{aligned}
$$

Uses linearity of expectation on each term:
$\mathbb{E}[X_i-\mu_i] = \mathbb{E}[X_i] - \mathbb{E}[\mu_i] = \mu_i - \mu_i = 0.$

Selection Problem

We use same notation as when analyzing quicksort
$S_{(k)}$ is the kth element of S in sorted order.

Given unsorted list S with $n = |S|$ distinct elements, and $k \in \{1, \dots, n\}$, find $S_{(k)}$.

For $y \in S$, let $r_S(y) := |\{y' \in S \mid y' \leq y\}|$ be the *rank* of y in S. The equivalent goal is to find $y \in S$ such that $r_S(y) = k$.

Observe that $r_S(S_{(k)}) = k$ and $S_{r_S(y)} = y$.

LazySelect

function LazySelect(S, k)

  repeat

    $R \leftarrow \lceil n^{3/4} \rceil$ elements from S, picked uniformly at random with replacement.

    Sort R in $\mathcal{O}(|R| \log |R|)$ steps.

    $x \leftarrow kn^{-1/4}$, $\ell \leftarrow \lfloor x - \sqrt{n} \rfloor + 1$, $a \leftarrow R_{(\ell)}$,
    $h \leftarrow \lceil x + \sqrt{n} \rceil - 1$, $b \leftarrow R_{(h)}$.
    By comparing a and b to every $s \in S$, find $r_S(a)$ and $r_S(b)$.

    $$P \leftarrow \begin{cases} \{y \in S \mid \quad y \leq b\} & \text{if } k < n^{3/4} \\ \{y \in S \mid a \leq y \quad\} & \text{if } k > n - n^{3/4} \\ \{y \in S \mid a \leq y \leq b\} & \text{if } k \in [n^{3/4}, n - n^{3/4}] \end{cases}$$

  until $S_{(k)} \in P$ and $|P| \leq 4n^{3/4} + 2$

  Sort P in $\mathcal{O}(|P| \log |P|)$ steps.

  return $P_{(k - r_S(a) + 1)}$ ▷ This is $S_{(k)}$.

We *sample* a subset R of the elements. For simplicity we allow the same element to be sampled multiple times.

We sort the samples in $\mathcal{O}(|R| \log |R|) \subseteq o(n)$ steps, using e.g. heapsort.

We compute $\ell, h, a = R_{(\ell)}, b = R_{(h)}$ so $|r_S(a) - r_S(b)|$ is expected to be small and $S_{(k)}$ is expected to be in $[a, b]$. If k is very small or very large, replace a or b with $\pm\infty$.

We compute $P = S \cap [a, b]$, and start over if we were unlucky. We can check this using the computed values of $r_S(a)$ and $r_S(b)$.

We sort P in $\mathcal{O}(|P| \log |P|) \subseteq o(n)$ steps, and then know where $S_{(k)}$ is.

LazySelect, Example

Start with this set S and $k = 17$.

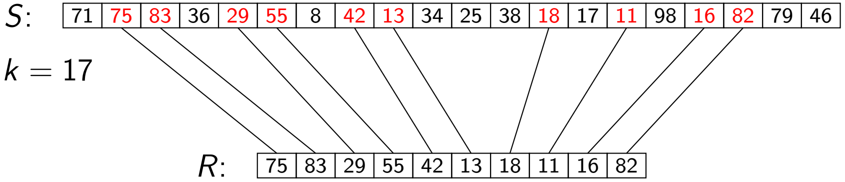
S:

71	75	83	36	29	55	8	42	13	34	25	38	18	17	11	98	16	82	79	46

$k = 17$

LazySelect, Example

Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.



LazySelect, Example

S:

71	75	83	36	29	55	8	42	13	34	25	38	18	17	11	98	16	82	79	46

$k = 17$

$R = R_{(\cdot)}$:

11	13	16	18	29	42	55	75	82	83

Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.

LazySelect, Example

S:

71	75	83	36	29	55	8	42	13	34	25	38	18	17	11	98	16	82	79	46

$k = 17$

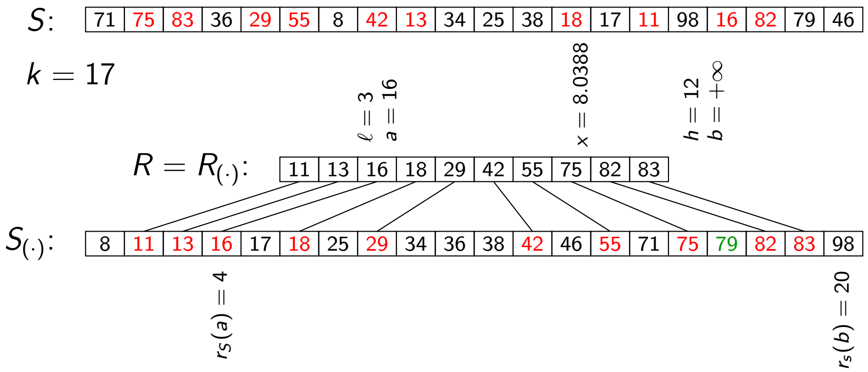
$\ell = 3$, $a = 16$ $\quad$ $x = 8.0388$ $\quad$ $h = 12$, $b = +\infty$

$R = R_{(\cdot)}$:

11	13	16	18	29	42	55	75	82	83

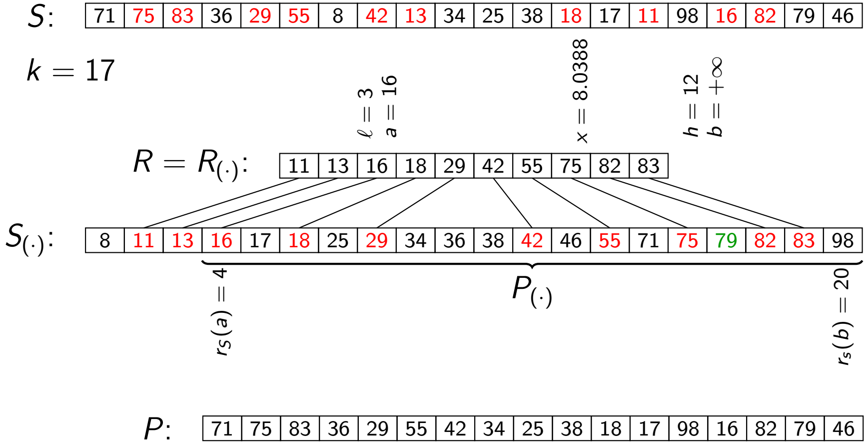
Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.

LazySelect, Example



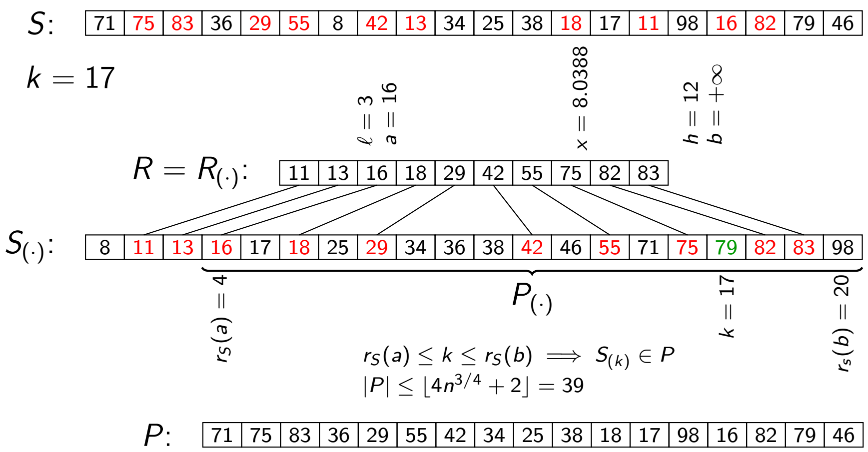
Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.
Compute $r_S(a) = 4, r_S(b) = 20$.

LazySelect, Example



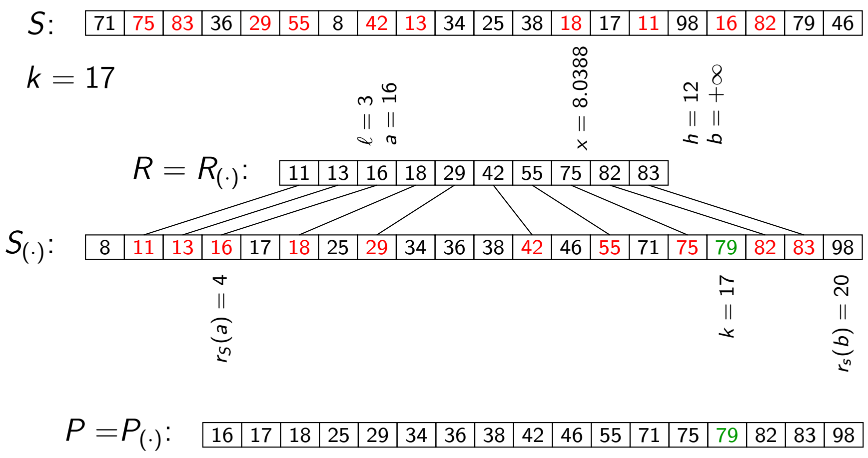
Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.
Compute $r_S(a) = 4, r_S(b) = 20$.
Compute $P = \{y \in S \mid a \leq y\}$ (not sorted).

LazySelect, Example



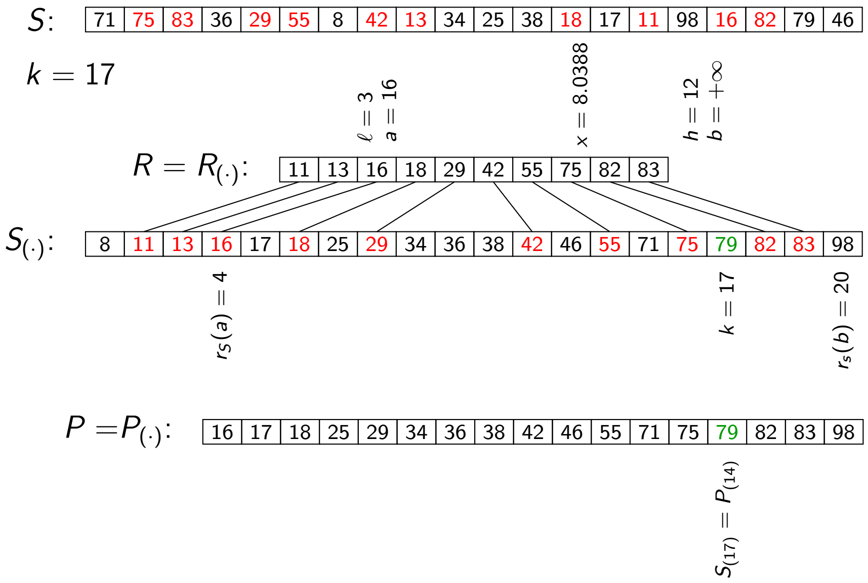
Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.
Compute $r_S(a) = 4, r_S(b) = 20$.
Compute $P = \{y \in S \mid a \leq y\}$ (not sorted).
Since $r_S(a) \leq 17$ we have $S_{(17)} \in P$ and since $|P| = 17 \leq \lfloor 4n^{3/4} + 2 \rfloor = 39$, exit the loop.

LazySelect, Example



Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.
Compute $r_S(a) = 4, r_S(b) = 20$.
Compute $P = \{y \in S \mid a \leq y\}$ (not sorted).
Since $r_S(a) \leq 17$ we have $S_{(17)} \in P$ and since $|P| = 17 \leq \lfloor 4n^{3/4} + 2 \rfloor = 39$, exit the loop.
Sort P.

LazySelect, Example



Start with this set S and $k = 17$.
Sample $\lceil n^{3/4} \rceil = 10$ elements into R.
Sort R.
Compute $x, \ell = 3, h = 12, a = R_{(\ell)}, b = R_{(h)}$. x is roughly the rank $S_{(k)}$ would get in R if sampled.
Compute $r_S(a) = 4, r_S(b) = 20$.
Compute $P = \{y \in S \mid a \leq y\}$ (not sorted).
Since $r_S(a) \leq 17$ we have $S_{(17)} \in P$ and since $|P| = 17 \leq \lfloor 4n^{3/4} + 2 \rfloor = 39$, exit the loop.
Sort P.
Return $P_{(k - r_S(a) + 1)} = P_{(14)} = S_{(17)}$.

LazySelect, Analysis

Theorem

With probability at least $1 - n^{-1/4}$, LazySelect *finds* $S_{(k)}$ *after only one run through the loop, and thus does only* $2n + o(n)$ *comparisons.*

Best known deterministic algorithm is complicated and uses $3n$ comparisons in the worst case.

LazySelect, Analysis

Theorem

With probability at least $1 - n^{-1/4}$*,* LazySelect *finds* $S_{(k)}$ *after only one run through the loop, and thus does only* $2n + o(n)$ *comparisons.*

Best known deterministic algorithm is complicated and uses $3n$ comparisons in the worst case.

LazySelect, Proof

The time bound is obvious from the algorithm. If it only does one run, the $2n$ comparisons come from computing $r_S(a)$ and $r_S(b)$. Each sort takes $\mathcal{O}(n^{3/4} \log n) \subseteq o(n)$ comparisons.

LazySelect, Proof

We assume for simplicity that $n^{3/4}$ is an integer.
We only prove the case where k is not near either end. The other two cases are similar, but simpler.

We need Pr[multiple runs] $\leq n^{-1/4}$.
Assume $k \in [n^{3/4}, n - n^{3/4}]$, then
$x \in [\sqrt{n}, n^{3/4} - \sqrt{n}]$.
Two ways to fail:

Type I: $S_{(k)} \notin P$

Type II: $S_{(k)} \in P \wedge |P| > \lfloor 4n^{3/4} + 2 \rfloor$

LazySelect, Proof

Let $k_\ell = \max\{1, k - 2n^{3/4}\}$ and $k_h = \min\{k + 2n^{3/4}, n\}$, then:

$$\Pr[S_{(k)} \notin P] = \Pr[S_{(k)} < a] + \Pr[S_{(k)} > b]$$

$$\Pr\left[S_{(k)} \in P \wedge |P| > \lfloor 4n^{2/3} + 2 \rfloor\right] \leq \Pr[S_{(k_\ell)} > a] + \Pr[S_{(k_h)} < b]$$

We will show that each of the probabilities on the right are bounded by $\frac{1}{4}n^{-1/4}$.

For the second inequality, note that by definition of k_ℓ and k_h, the condition $S_{(k)} \in P \wedge |P| > \lfloor 4n^{2/3} + 2 \rfloor$ implies that at least one of $S_{(k_\ell)} > a$ and $S_{(k_h)} < b$ is true. Thus we can upper bound the probability by the sum of the two probabilities.

LazySelect, Proof

Let $X_{(i)} = |\{y \in R \mid y \leq S_{(i)}\}|$.

Lemma

$$S_{(i)} < R_{(j)} \iff X_{(i)} < j$$

Proof.

If $S_{(i)} < R_{(j)}$, at most $j-1$ elements in R are $\leq S_{(i)}$, so $X_{(i)} \leq j-1$ and thus $X_{(i)} < j$. Conversely, if $S_{(i)} \geq R_{(j)}$, so are $R_{(1)}, \ldots, R_{(j)}$, thus $X_{(i)} \geq j$. $\square$

LazySelect, Proof (Type I)

Let X_i indicate that the ith element picked for R is $\leq S_{(k)}$. Then $\Pr[X_i = 1] = \frac{k}{n}$ and $X_{(k)} = X = \sum_{i=1}^{n^{3/4}} X_i$. X_i are *Bernoulli trials* with success probability $p = \frac{k}{n}$. Thus,

$$\mu_X = pn^{3/4} = kn^{-1/4} = x$$

$$\sigma_X^2 = n^{3/4} p(1-p) \leq \frac{n^{3/4}}{4}$$

$$\sigma_X \leq \frac{n^{3/8}}{2}$$

LazySelect, Proof (Type I)

Now

$$\begin{aligned}
\Pr[S_{(k)} < a] &= \Pr[X_{(k)} < \ell] && \text{(uses Lemma)}\\
&\le \Pr[|X - \mu_X| \ge \sqrt{n}]\\
&\le \Pr[|X - \mu_X| \ge 2n^{1/8}\sigma_X]\\
&\le \tfrac{1}{4}n^{-1/4}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Pr[S_{(k)} > b] &\le \Pr[S_{(k)} \ge R_{(h+1)}]\\
&= \Pr[X_{(k)} \ge h+1] && \text{(uses Lemma)}\\
&\le \tfrac{1}{4}n^{-1/4}
\end{aligned}$$

On the blackboard, if needed

$$\begin{aligned}
\Pr[X_{(k)} < \ell] &= \Pr[X < \lfloor x - \sqrt{n} \rfloor + 1]\\
&= \Pr[X \le \lfloor x - \sqrt{n} \rfloor]\\
&\le \Pr[X \le x - \sqrt{n}]\\
&= \Pr[X - \mu_X \le -\sqrt{n}] && \text{(uses } \mu_X = x\text{)}\\
&= \Pr[-(X - \mu_X) \ge \sqrt{n}]\\
&\le \Pr[|X - \mu_X| \ge \sqrt{n}]\\
\Pr[X_{(k)} \ge h+1] &= \Pr[X \ge h+1]\\
&= \Pr[X \ge \lceil x + \sqrt{n} \rceil - 1 + 1]\\
&= \Pr[X \ge \lceil x + \sqrt{n} \rceil]\\
&\le \Pr[X \ge x + \sqrt{n}]\\
&= \Pr[X - \mu_X \ge \sqrt{n}] && \text{(uses } \mu_X = x\text{)}\\
&\le \Pr[|X - \mu_X| \ge \sqrt{n}]
\end{aligned}$$

LazySelect, Proof (Type II)

By completely analoguous arguments,

$$\Pr[S_{(k_\ell)} > a] \leq \frac{1}{4}n^{-1/4}$$
$$\Pr[S_{(k_h)} < b] \leq \frac{1}{4}n^{-1/4}$$

Thus,

$$\Pr[\text{multiple runs}] \leq \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} = n^{-1/4}$$

Assume $k_\ell = \max\{1, k - 2n^{3/4}\} > 1$, otherwise $\Pr[S_{(k_\ell)} > a] = 0$. Let $X = X_{(k_\ell)}$, then

$$\mu_X = \frac{k_\ell}{n}n^{3/4} = x - 2\sqrt{n}$$
$$\sigma_X^2 = n^{3/4}\frac{k_\ell}{n}\left(1 - \frac{k_\ell}{n}\right) \leq \frac{n^{3/4}}{4}$$
$$\sigma_X \leq \frac{n^{3/8}}{2}$$

$$\begin{aligned}
\Pr[S_{(k_\ell)} > a] &\leq \Pr[S_{(k_\ell)} \geq a] \\
&= \Pr[X_{(k_\ell)} \geq \ell] && \text{(uses Lemma)} \\
&= \Pr[X \geq \lfloor x - \sqrt{n} \rfloor + 1] \\
&\leq \Pr[X \geq x - \sqrt{n}] \\
&= \Pr[X - \mu_X \geq \sqrt{n}] && \text{(uses } \mu_X = x - 2\sqrt{n}\text{)} \\
&\leq \Pr[|X - \mu_X| \geq \sqrt{n}] \\
&\leq \Pr[|X - \mu_X| \geq 2n^{1/8}\sigma_X] \leq \frac{1}{4}n^{-1/4}
\end{aligned}$$

LazySelect, Proof (Type II)

By completely analoguous arguments,

$$\Pr[S_{(k_\ell)} > a] \leq \frac{1}{4}n^{-1/4}$$
$$\Pr[S_{(k_h)} < b] \leq \frac{1}{4}n^{-1/4}$$

Thus,

$$\begin{aligned}\Pr[\text{multiple runs}] &\leq \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} + \frac{1}{4}n^{-1/4} \\ &= n^{-1/4}\end{aligned}$$

Assume $k_h = \min\{k + 2n^{3/4}, n\} < n$, otherwise $\Pr[S_{(k_h)} < b] = 0$. Let $X = X_{(k_h)}$, then

$$\begin{aligned}
\mu_X &= \frac{k_h}{n}n^{3/4} = x + 2\sqrt{n} \\
\sigma_X^2 &= n^{3/4}\frac{k_h}{n}\left(1 - \frac{k_h}{n}\right) \leq \frac{n^{3/4}}{4} \\
\sigma_X &\leq \frac{n^{3/8}}{2}
\end{aligned}$$

$$\begin{aligned}
\Pr[S_{(k_h)} < b] &= \Pr[X_{(k_h)} < h] && \text{(uses Lemma)} \\
&= \Pr[X < \lceil x + \sqrt{n}\rceil - 1] \\
&\leq \Pr[X \leq x + \sqrt{n}] \\
&= \Pr[X - \mu_X \leq -\sqrt{n}] && \text{(uses } \mu_X = x + 2\sqrt{n}\text{)} \\
&= \Pr[-(X - \mu_X) \geq \sqrt{n}] \\
&\leq \Pr[|X - \mu_X| \geq \sqrt{n}] \\
&\leq \Pr[|X - \mu_X| \geq 2n^{1/8}\sigma_X] \leq \frac{1}{4}n^{-1/4}
\end{aligned}$$

LazySelect, Summary

We have shown that with *high probability* $= 1 - n^{-1/4}$, $\textsc{LazySelect}$ does only $2n + o(n)$ comparisons.

Two-point Sampling, Intro

A common technique we have seen for Monte Carlo algorithms is to run them several times to boost the probability of a correct result.

However, random bits can be expensive! Two-point sampling is to take just two random values in $\mathbb{Z}_n$ and turn them into many *pairwise independent* values.

Two-point Sampling, Idea

Proving this is Exercise 3.7.

Let n be prime, and let a, b be independent random variables uniformly chosen from $\mathbb{Z}_n = \{0, \ldots, n-1\}$.
Let $r_i = (a \cdot i + b) \bmod n$, then for any $i \neq j \pmod{n}$, r_i and r_j are independent and uniform in $\mathbb{Z}_n$.
Thus, $r_1, \ldots, r_n$ are pairwise independent.

Two-point Sampling, Application

Let $L \subseteq \Sigma^\star$ be some language, and let n be a prime.

A function $A : \Sigma^\star \times \mathbb{Z}_n \to \{0, 1\}$ is an **RP** algorithm for deciding L, if it runs in polynomial time for all inputs, and

If $x \in L$, then $A(x, r) = 1$ for at least half of all $r \in \mathbb{Z}_n$.

If $x \notin L$ then $A(x, r) = 0$ for all $r \in \mathbb{Z}_n$.

Two-point Sampling, Application

Running A with $t > 1$ independent random values from $\mathbb{Z}_n$ gives an error probability of at most 2^{-t}, but is expensive.

Lemma

Using two-point sampling, and running $A(x, r_1), \ldots, A(x, r_t)$ gives an error probability of at most $\frac{1}{t}$.

Note that it only holds for $t \leq n$.

Two-point Sampling, App Proof

Assume $x \in L$ (otherwise no error).

Let $Y_i = A(x, r_i)$ and $Y = \sum_{i=1}^{m} Y_i$. Then $\mu_{Y_i} = \mathbb{E}[Y_i] \geq \frac{1}{2}$, $\sigma_{Y_i}^2 = \mathbb{E}[(Y_i - \mu_{Y_i})^2] \leq \frac{1}{4}$, $\mu_Y = \sum_{i=1}^{t} \mu_i \geq \frac{t}{2}$ and $\sigma_Y^2 = \sum_{i=1}^{t} \sigma_{Y_i}^2 \leq \frac{t}{4}$, so $\sigma_Y \leq \frac{\sqrt{t}}{2}$.

An error means that $Y = 0 \leq \mu_Y - \frac{t}{2}$, so error probability is

$$\begin{aligned} \Pr[Y = 0] &\leq \Pr[|Y - \mu_Y| \geq \tfrac{t}{2}] \\ &\leq \Pr[|Y - \mu_Y| \geq \sqrt{t} \cdot \sigma_Y] \leq \tfrac{1}{t} \end{aligned}$$

Let $p_i = \Pr[Y_i = 1]$. $p_i \geq \frac{1}{2}$ because we assume at least half the values of $r \in \mathbb{Z}_n$ are witnesses.

$$\begin{aligned} \mu_{Y_i} &= p_i \\ \mathbb{E}[(Y_i - \mu_{Y_i})^2] &= (1 - p_i)\mu_{Y_i}^2 + p_i(1 - \mu_{Y_i})^2 \\ &= (1 - p_i)p_i^2 + p_i(1 - p_i)^2 \\ &= p_i(1 - p_i)(p_i + (1 - p_i)) \\ &= p_i(1 - p_i) \\ &\leq \tfrac{1}{4} \end{aligned}$$

In computing the variance σ_Y^2, we use the Lemma from earlier, and the fact that the r_i are pairwise independent.

Summary