Randomized Algorithms

南京大学

尹一通
Balls-into-bins model

throw \( m \) balls into \( n \) bins uniformly and independently

uniform random function

\[ f : [m] \rightarrow [n] \]

- The threshold for being 1-1 is \( m = \Theta(\sqrt{n}) \).
- The threshold for being on-to is \( m = n \ln n + O(n) \).
- The maximum load is

\[
\begin{cases}
O\left(\frac{\ln n}{\ln \ln n}\right) & \text{for } m = \Theta(n), \\
O\left(\frac{m}{n}\right) & \text{for } m = \Omega(n \ln n).
\end{cases}
\]

1-1 birthday problem
on-to coupon collector
pre-images occupancy problem
Stable Marriage

- Each man has a preference order of the $n$ women;
- Each woman has a preference order of the $n$ men;
- Solution: $n$ couples
- Marriages are stable!
Stable Marriage

- n men
- n women

**unstable (blocking pair):**
- A man and a woman, who prefer each other to their current partners

**stable:** no blocking pairs

- Local optimum
- Fixed point
- Equilibrium
- Deadlock
Proposal Algorithm
(Gale-Shapley 1962)

- Single man:
  - propose to the most preferable women who has not rejected him

- Woman:
  - upon received a proposal:
    - accept if she’s single or married to a less preferable man (divorce!)
Proposal Algorithm

- **woman**: once got married always married (will only switch to better men!)
- **man**: will only get worse ...
- once all women are married, the algorithm terminates, and the marriages are stable
- total number of proposals: $\leq n^2$

<table>
<thead>
<tr>
<th>Single man:</th>
</tr>
</thead>
<tbody>
<tr>
<td>propose to the most preferable women who has not rejected him</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Woman:</th>
</tr>
</thead>
<tbody>
<tr>
<td>upon received a proposal: if she’s single or married to a less preferable man (divorce!)</td>
</tr>
</tbody>
</table>

if “A” and “b” prefer each other than their current partners “a” and “B”, then “A” would have proposed to “b” before to “a”, and “b” should have accepted

this proves the existence of stable matching by construction
Average-case

- every man/woman has a uniform random permutation as preference list
- total number of proposals?

Looks very complicated!
Principle of Deferred Decisions

Principle of deferred decision

The decision of random choice in the random input is deferred to the running time of the algorithm.
Principle of Deferred Decisions

proposing in the order of a uniformly random permutation

at each time, proposing to a uniformly random woman who has not rejected him

decisions of the inputs are deferred to the time when Alg accesses them
Principle of Deferred Decisions

at each time, proposing to a uniformly random woman who has not rejected him

the man forgot who had rejected him (!)
Principle of Deferred Decisions

- uniformly and independently proposing to $n$ women
- Alg stops once all women got proposed.
- Coupon collector!
- Expected $O(n \ln n)$ proposals.
Tail Inequalities
Tail bound:
\[ \Pr[X > t] < \epsilon. \]

Thresholding:
- The running time of a Las Vegas Alg.
- Some cost (e.g. max load).
- The probability of extreme case.
Tail bound:
Pr\[X > t]\ < \epsilon.

n-ball-to-n-bin:
Pr [load of the first bin \(\geq t\)]
\[\leq \binom{n}{t} \left(\frac{1}{n}\right)^t\]
\[= \frac{n!}{t!(n-t)!n^t}\]
\[= \frac{1}{t!} \cdot \frac{n(n-1)(n-2)\cdots(n-t+1)}{n^t}\]
\[= \frac{1}{t!} \cdot \prod_{i=0}^{t-1} \left(1 - \frac{i}{n}\right)\]
\[\leq \frac{1}{t!}\]
\[\leq \left(\frac{e}{t}\right)^t\]
Tail bound:
\[ \Pr[X > t] < \epsilon. \]

Relate tail to some measurable characters of \( X \)

Reduce the tail bound to the analysis of the characters.

\[ \Pr[X > t] < f(t, I) \]
Markov’s Inequality

**Markov’s Inequality:**
For nonnegative $X$, for any $t > 0$, 

$$\Pr[X \geq t] \leq \frac{E[X]}{t}.$$ 

**Proof:**
Let 

$$Y = \begin{cases} 
1 & \text{if } X \geq t, \\
0 & \text{otherwise.} 
\end{cases}$$

$$\Rightarrow Y \leq \left\lfloor \frac{X}{t} \right\rfloor \leq \frac{X}{t},$$

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

tight if we only know the expectation of $X$
Las Vegas to Monte Carlo

- **Las Vegas**: running time is random, always correct.
- **A**: Las Vegas Alg with worst-case expected running time $T(n)$.
- **Monte Carlo**: running time is fixed, correctness is random.
- **B**: Monte Carlo Alg ...

<table>
<thead>
<tr>
<th>B(x):</th>
</tr>
</thead>
<tbody>
<tr>
<td>run A(x) for $2T(n)$ steps;</td>
</tr>
<tr>
<td>if A(x) returned</td>
</tr>
<tr>
<td>return A(x);</td>
</tr>
<tr>
<td>else return “yes”;</td>
</tr>
</tbody>
</table>

one-sided error!

\[
\Pr[\text{error}] \leq \Pr[T(A(x)) > 2T(n)]
\]

\[
\leq \frac{\mathbb{E}[T(A(x))]}{2T(n)} \leq \frac{1}{2}
\]
Markov’s Inequality:

For nonnegative $X$, for any $t > 0$,

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$
A Generalization of Markov’s Inequality

**Theorem:**
For any $X$, for $h : X \mapsto \mathbb{R}^+$, for any $t > 0$,

$$\Pr[h(X) \geq t] \leq \frac{\mathbb{E}[h(X)]}{t}.$$
Chebyshev’s Inequality:
For any $t > 0$,

$$\Pr [|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}[X]}{t^2}.$$
Variance

**Definition (variance):**

The variance of a random variable $X$ is

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$ 

The standard deviation of random variable $X$ is

$$\delta[X] = \sqrt{\text{Var}[X]}$$
Covariance

**Theorem:**

\[
\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y);
\]

\[
\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i] + \sum_{i \neq j} \text{Cov}(X_i, X_j).
\]

**Definition (covariance):**

The covariance of \(X\) and \(Y\) is

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].
\]
Covariance

**Theorem:**
For independent $X$ and $Y$, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

**Theorem:**
For independent $X$ and $Y$, $\text{Cov}(X, Y) = 0$.

**Proof:**
\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]
= \mathbb{E}[X - \mathbb{E}[X]] \mathbb{E}[Y - \mathbb{E}[Y]]
= 0.
\]
Variance of sum

**Theorem:**
For independent $X$ and $Y$, $\text{Cov}(X, Y) = 0$.

**Theorem:**
For pairwise independent $X_1, X_2, \ldots, X_n$,

$$\text{Var} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \text{Var}[X_i].$$
Variance of Binomial Distribution

- **Binomial distribution**: number of successes in $n$ i.i.d. Bernoulli trials.

- $X$ follows binomial distribution with parameter $n$ and $p$

\[
X = \sum_{i=1}^{n} X_i \\
X_i = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p 
\end{cases}
\]

\[
\text{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = p - p^2 = p(1 - p)
\]

\[
\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = p(1 - p)n \quad \text{(independence)}
\]
Chebyshev’s Inequality

Chebyshev’s Inequality:
For any $t > 0$,

$$
\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}[X]}{t^2}.
$$

Proof:
Apply Markov’s inequality to $(X - \mathbb{E}[X])^2$

$$
\Pr[(X - \mathbb{E}[X])^2 \geq t^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2}
$$
Selection Problem

Input: a set of $n$ elements
Output: the $k$th smallest element

straightforward alg:

- sorting, $\Omega(n \log n)$ time

sophisticated deterministic alg:

- median of medians, $\Theta(n)$ time

simple randomized alg:

- LazySelect, $\Theta(n)$ time, find the median w.h.p.
Selection by Sampling

distribution:

Naive sampling:

uniformly choose an random element

make a wish it is the median
Selection by Sampling

distribution:

\[ R: \]

sample a small set \( R \),
selection in \( R \) by sorting

roughly concentrated, but not good enough
Selection by Sampling

distributions:

Find such \( d \) and \( u \) that:

- Let \( C = \{ x \in S \mid d \leq x \leq u \} \).
- The median is in \( C \).
- \( C \) is not too large (sort \( C \) is linear time).
LazySelect
(Floyd & Rivest)

Size of $R$: $r$

Offset for $d$ and $u$ from the median of $R$: $k$

Bad events: median is not between $d$ and $u$; too many elements between $d$ and $u$. (inefficient to sort)
1. Uniformly and independently sample $r$ elements from $S$ to form $R$; and sort $R$.

2. Let $d$ be the $(\frac{r}{2} - k)$th element in $R$.

3. Let $u$ be the $(\frac{r}{2} + k)$th element in $R$.

4. If any of the following occurs then FAIL.
   \[
   |\{x \in S \mid x < d\}| > \frac{n}{2};
   \]
   \[
   |\{x \in S \mid x > u\}| > \frac{n}{2};
   \]
   \[
   |\{x \in S \mid d \leq x \leq u\}| > s;
   \]
   \[
   \text{Pr}[\text{FAIL}] < ?
   \]

5. Find the median of $S$ by sorting $\{x \in S \mid d \leq x \leq u\}$.
Bad events:

1. \(|\{x \in S \mid x < d\}| > \frac{n}{2} ;
2. \(|\{x \in S \mid x > u\}| > \frac{n}{2} ;
3. \(|\{x \in S \mid d \leq x \leq u\}| > s;

\text{or} \quad \(|\{x \in S \mid d \leq x \leq \frac{n}{2}\}| > \frac{s}{2} ;
\quad \(|\{x \in S \mid \frac{n}{2} \leq x \leq u\}| > \frac{s}{2} ;

\text{or} \quad \(|\{x \in S \mid d \leq x \leq \frac{n}{2}\}| > \frac{s}{2} ;
\quad \(|\{x \in S \mid \frac{n}{2} \leq x \leq u\}| > \frac{s}{2} ;
Bad events:

1. $|\{x \in S \mid x < d\}| > \frac{n}{2}$;
2. $|\{x \in S \mid x > u\}| > \frac{n}{2}$;
3. $|\{x \in S \mid d \leq x \leq u\}| > s$;

or

- $|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$;
- $|\{x \in S \mid x > u\}| < \frac{n}{2} - \frac{s}{2}$;

Symmetry!

Bad events for $d$:

- $d$ is too large: $|\{x \in S \mid x < d\}| > \frac{n}{2}$
- $d$ is too small: $|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$
Bad events for $d$:

- **$d$ is too large:**
  \[
  \{|x \in S \mid x < d\| > \frac{n}{2}
  \]

- **$d$ is too small:**
  \[
  \{|x \in S \mid x < d\| < \frac{n}{2} - \frac{s}{2}
  \]

### Bad events for $R$:

- The sample of rank $\frac{r}{2} - k$ in $R$ is ranked $> \frac{n}{2}$ in $S$.
- The sample of rank $\frac{r}{2} - k$ in $R$ is ranked $\leq \frac{n}{2} - \frac{s}{2}$ in $S$.

### Diagram:

- **$S$**
  - $d$
  - $u$
  - $k$ offset

- **$R$**
  - $r$ samples
  - $d$
  - $u$

### Notes:

- $R$: $r$ uniform and independent samples from $S$. 
Bad events for $d$:

$d$ is too large:
$$|\{x \in S \mid x < d\}| > \frac{n}{2}$$

$d$ is too small:
$$|\{x \in S \mid x < d\}| < \frac{n}{2} - \frac{s}{2}$$

Bad events for $R$:

\(< \frac{r}{2} - k \) samples are among the smallest half in $S$.

\(\geq \frac{r}{2} - k \) samples are among the $\frac{n}{2} - \frac{s}{2}$ smallest in $S$.

$S$: 

$R$: $r$ uniform and independent samples from $S$
\( R: r \) uniform and independent samples from \( S \)

**Bad events for** \( R: \)

\( \mathcal{E}_1: \) \(< \frac{r}{2} - k \) samples are among the smallest half in \( S \).

\( \mathcal{E}_2: \) \( \geq \frac{r}{2} - k \) samples are among the \( \frac{n}{2} - \frac{s}{2} \) smallest in \( S \).

\[
\begin{align*}
X_i &= \begin{cases} 
1 & \text{ith sample ranks} \leq \frac{n}{2}, \\
0 & \text{otherwise.}
\end{cases} \\
X &= \sum_{i=1}^{r} X_i
\end{align*}
\]

\[
\begin{align*}
Y_i &= \begin{cases} 
1 & \text{ith sample ranks} \leq \frac{n}{2} - \frac{s}{2}, \\
0 & \text{otherwise.}
\end{cases} \\
Y &= \sum_{i=1}^{r} Y_i
\end{align*}
\]
\( R: r \) uniform and independent samples from \( S \)

**Bad events for** \( R: \)

\( \mathcal{E}_1: \) < \( \frac{r}{2} - k \) samples are among the smallest half in \( S \).

\( \mathcal{E}_2: \) \( \geq \frac{r}{2} - k \) samples are among the \( \frac{n}{2} - \frac{s}{2} \) smallest in \( S \).

\[
\begin{align*}
X_i &= \begin{cases} 
1 & \text{with prob } \frac{1}{2} \\
0 & \text{with prob } \frac{1}{2}
\end{cases} \\ 
X &= \sum_{i=1}^{r} X_i \\
Y_i &= \begin{cases} 
1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\
0 & \text{with prob } \frac{1}{2} + \frac{s}{2n}
\end{cases} \\
Y &= \sum_{i=1}^{r} Y_i
\end{align*}
\]

\( \frac{n}{2} - \frac{s}{2} \)

\( n \)

\( r \) samples
Bad events:

$\mathcal{E}_1 : X < \frac{r}{2} - k$

$\mathcal{E}_2 : Y \geq \frac{r}{2} - k$

$X = \sum_{i=1}^{r} X_i$

$X_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$

$Y = \sum_{i=1}^{r} Y_i$

$Y_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\ 0 & \text{with prob } \frac{1}{2} + \frac{s}{2n} \end{cases}$

$S$: $\frac{n}{2} - \frac{s}{2}$

$\frac{n}{2}$

$r$ samples
Bad events:

$\mathcal{E}_1: \quad X < \frac{r}{2} - k$

$\mathcal{E}_2: \quad Y \geq \frac{r}{2} - k$

$X$ and $Y$ are binomial!

$E[X] = \frac{r}{2}$

$E[Y] = \frac{r}{2} - \frac{sr}{2n}$

$Var[X] = \frac{r}{4}$

$Var[Y] = \frac{r}{4} - \frac{s^2 r}{4n^2}$

$X_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$

$X = \sum_{i=1}^{r} X_i$

$Y_i = \begin{cases} 1 & \text{with prob } \frac{1}{2} - \frac{s}{2n} \\ 0 & \text{with prob } \frac{1}{2} + \frac{s}{2n} \end{cases}$

$Y = \sum_{i=1}^{r} Y_i$
Bad events:

\[ \mathcal{E}_1 : \quad X < \frac{r}{2} - k \]

\[ \mathcal{E}_2 : \quad Y \geq \frac{r}{2} - k \]

\[ \mathbf{E}[X] = \frac{r}{2} \]

\[ \text{Var}[X] = \frac{r}{4} \]

\[ \mathbf{E}[Y] = \frac{r}{2} - \frac{sr}{2n} \]

\[ \text{Var}[Y] = \frac{r}{4} - \frac{s^2r}{4n^2} \]

\[ r = n^{3/4} \]

\[ k = n^{1/2} \]

\[ s = 4n^{3/4} \]
Bad events:

\[ \mathcal{E}_1 : \quad X < \frac{1}{2} n^{3/4} - \sqrt{n} \]

\[ \mathcal{E}_2 : \quad Y \geq \frac{1}{2} n^{3/4} - \sqrt{n} \]

\[ \mathbb{E}[X] = \frac{1}{2} n^{3/4} \]

\[ \text{Var}[X] = \frac{1}{4} n^{3/4} \]

\[ \mathbb{E}[Y] = \frac{1}{2} n^{3/4} - 2\sqrt{n} \]

\[ \text{Var}[Y] < \frac{1}{4} n^{3/4} \]

\[ r = n^{3/4} \]

\[ k = n^{1/2} \]

\[ s = 4n^{3/4} \]
Bad events:

\[ E_1: \quad X < \frac{1}{2} n^{3/4} - \sqrt{n} \]

\[ E_2: \quad Y \geq \frac{1}{2} n^{3/4} - \sqrt{n} \]

\[ \Pr[E_1] = \Pr \left[ X < \frac{1}{2} n^{3/4} - \sqrt{n} \right] \leq \Pr \left[ |X - \mathbb{E}[X]| > \sqrt{n} \right] \leq \frac{\text{Var}[X]}{n} \leq \frac{1}{4} n^{-1/4} \]

\[ \Pr[E_2] = \Pr \left[ Y \geq \frac{1}{2} n^{3/4} - \sqrt{n} \right] \leq \Pr \left[ |Y - \mathbb{E}[Y]| \geq \sqrt{n} \right] \leq \frac{\text{Var}[Y]}{n} \leq \frac{1}{4} n^{-1/4} \]

\[ \mathbb{E}[X] = \frac{1}{2} n^{3/4} \]

\[ \text{Var}[X] = \frac{1}{4} n^{3/4} \]

\[ \mathbb{E}[Y] = \frac{1}{2} n^{3/4} - 2\sqrt{n} \]

\[ \text{Var}[Y] < \frac{1}{4} n^{3/4} \]
Bad events:

\[ \mathcal{E}_1 : \quad X < \frac{1}{2} n^{3/4} - \sqrt{n} \]

\[ \mathcal{E}_2 : \quad Y \geq \frac{1}{2} n^{3/4} - \sqrt{n} \]

union bound:

\[ \Pr[d \text{ is bad}] \leq \Pr[\mathcal{E}_1 \vee \mathcal{E}_2] \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] \leq \frac{1}{2} n^{-1/4} \]

symmetry:

\[ \Pr[u \text{ is bad}] \leq \frac{1}{2} n^{-1/4} \]

union bound:

\[ \Pr[\text{FAIL}] \leq n^{-1/4} \]
1. Uniformly and independently sample $n^{3/4}$ elements from $S$ to form $R$; and sort $R$.

2. Let $d$ be the $(\frac{1}{2} n^{3/4} - \sqrt{n})$th element in $R$.

3. Let $u$ be the $(\frac{1}{2} n^{3/4} + \sqrt{n})$th element in $R$.

4. If any of the following occurs then FAIL.

\[
\begin{align*}
|\{x \in S \mid x < d\}| &> \frac{n}{2}; \\
|\{x \in S \mid x > u\}| &> \frac{n}{2}; \\
|\{x \in S \mid d \leq x \leq u\}| &> 4n^{3/4};
\end{align*}
\]

5. Find the median of $S$ by sorting $C$. 