

**Lemma 1.** *During any round, if a vertex  $w$  is marked then it is selected to be in  $S$  with probability at least  $1/2$ .*

**Proof.** The only reason a marked vertex  $w$  becomes unmarked is that one of its neighbors of degree at least  $d(w)$  is also marked. The probability of this happening is at most

$$\sum_{x \in \Gamma(w)} \frac{1}{2d(w)} \leq 1/2$$

□

**Lemma 2.** *The probability that a good vertex belongs to  $S \cup \Gamma(S)$  is at least  $(1 - e^{-1/6})/2$ .*

**Lemma 3.** *In a graph  $G = (V, E)$  the number of good edges is at least  $|E|/2$ .*

**Proof.** Direct the edges in  $E$  from the lower degree end-point to the higher degree end-point breaking ties arbitrarily. Let  $d_i(v)$  and  $d_o(v)$  be the in-degree and out-degree of  $v$ .

For each bad vertex  $v$ ,

$$d_o(v) - d_i(v) \geq d(v)/3 = \frac{d_o(v) + d_i(v)}{3}$$

Let  $E(S, T)$  be the set of edges directed from vertices in  $S$  to vertices in  $T$ ; and  $e(S, T) = |E(S, T)|$ .

The total degree of the bad vertices is given by

$$\begin{aligned} & 2e(V_B, V_B) + e(V_B, V_G) + e(V_G, V_B) \\ &= \sum_{v \in V_B} (d_o(v) + d_i(v)) \\ &\leq 3 \sum_{v \in V_B} (d_o(v) - d_i(v)) \\ &= 3 \sum_{v \in V_G} (d_i(v) - d_o(v)) \\ &= 3[(e(V_B, V_G) + e(V_G, V_G)) - (e(V_G, V_B) + e(V_G, V_G))] \\ &= 3[e(V_B, V_G) - e(V_G, V_B)] \\ &\leq 3[e(V_B, V_G) + e(V_G, V_B)] \end{aligned}$$

Thus,

$$e(V_B, V_B) \leq e(V_B, V_G) + e(V_G, V_B).$$

□

We now argue that the expected number of edges removed at a given iteration is at least a constant fraction of the number of edges present.

Let r.v.  $X$  denote the number of edges deleted in the current iteration and let  $E$  denote the current set of edges.

For each  $e \in E$ , let r.v.  $X_e$  indicate whether  $e$  is deleted or not.

$$X = \sum_{e \in E} X_e$$

$$E[X] = \sum_{e \in E} E[X_e] \geq \sum_{e \text{ is good}} E[X_e]$$

$$\geq \sum_{e \text{ is good}} (1 - e^{-1/6})/2 \geq (1 - e^{-1/6})|E|/4$$

# A Probabilistic Recurrence

Let  $g(x)$  be a monotone non-decreasing function from  $R^+$  to  $R^+$ . Consider a particle whose position changes at discrete time steps and is always at a positive integer. If the particle is currently at position  $m > 1$ , it proceeds at the next step to position  $m - X$ , where  $X$  is a random variable over integers  $1, \dots, m - 1$ . We are only given that  $E[X] \geq g(m)$  and that  $X$  is chosen independently of the past.

Assuming the particle starts at position  $n$ , what is the expected number of steps before it reaches position 1 ?

**Theorem 1.** *Let  $T$  be the random variable denoting the number of steps in which the particle reaches the position 1. Then  $E[T] \leq \int_1^n dx/g(x)$ .*

# Proof

By induction on  $n$ .

Suppose the theorem holds for values of  $m$  smaller than  $n$ . Let  $f(m) = \int_1^m dx/g(x)$  for  $m \geq 1$ .

Consider the first step, during which the particle proceeds from position  $n$  to position  $n - X$ , where  $X$  is chosen from a distribution for which  $E[X] \geq g(n)$ .

We have

$$\begin{aligned} E[T] &\leq 1 + E[f(n - X)] \\ &= 1 + E\left[\int_1^n dy/g(y) - \int_{n-X}^n dy/g(y)\right] \\ &= 1 + f(n) - E\left[\int_{n-X}^n dy/g(y)\right] \\ &\leq 1 + f(n) - E\left[\int_{n-X}^n dy/g(n)\right] \\ &= 1 + f(n) - E[X]/g(n) \leq f(n) \end{aligned}$$

## Expectation is not everything....

Which gambling game would you prefer:

- We flip one coin, you win \$1 if head, loose one \$1 if tail.
- We flip 10 coins, you win  $\$2^{10}/2 = 0.5K$  if all heads, else you pay \$1.
- We flip 20 coins, you win  $\$2^{20}/2 = M/2$  if all heads, else you pay \$1.

# Bounding Deviation from Expectation

**Theorem 2. [Markov Inequality]** For any non-negative random variable and for any  $a > 0$

$$\Pr(X \geq a) \leq \frac{E[X]}{a}.$$

**Proof.** For any  $a > 0$ , let

$I$  be an indicator r.v. for the event  $X \geq a$ .

$$I \leq X/a. \quad E[I] \leq E[X]/a \quad \square$$

Example: What is the probability of getting more than  $\frac{3N}{4}$  heads in  $N$  coin flips?

$$\leq \frac{N/2}{3N/4} \leq \frac{2}{3}.$$

# Variance

**Definition 1.** *The variance of a random variable  $X$  is*

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

**Definition 2.** *The standard deviation of a random variable  $X$  is*

$$\sigma(X) = \sqrt{\text{Var}[X]}.$$

Example: Let  $X$  be a 0-1 random variable with  $Pr(X = 0) = Pr(X = 1) = 1/2$ .

$$E[X] = 1/2.$$

$$Var[X] = \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 = \frac{1}{4}.$$

# Chebyshev's Inequality

**Theorem 3.** *For any random variable*

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$

**Proof.**

$$Pr(|X - E[X]| \geq a) = Pr((X - E[X])^2 \geq a^2)$$

By Markov inequality

$$\begin{aligned} Pr((X - E[X])^2 \geq a^2) &\leq \frac{E[(X - E[X])^2]}{a^2} \\ &= \frac{Var[X]}{a^2} \end{aligned}$$

□

**Theorem 4.** *For any random variable*

$$\Pr(|X - E[X]| \geq a\sigma[X]) \leq \frac{1}{a^2}.$$

**Theorem 5.** *For any random variable*

$$\Pr(|X - E[X]| \geq \epsilon E[X]) \leq \frac{\text{Var}[X]}{\epsilon^2 (E[X])^2}.$$

**Theorem 6.** *If  $X$  and  $Y$  are independent random variable*

$$E[XY] = E[X] \cdot E[Y],$$

**Proof.**

$$E[XY] = \sum_i \sum_j i \cdot j \Pr((X = i) \cap (Y = j)) =$$

$$\sum_i \sum_j ij \Pr(X = i) \cdot \Pr(Y = j) =$$

$$\left( \sum_i i \Pr(X = i) \right) \left( \sum_j j \Pr(Y = j) \right).$$

□

**Theorem 7.** *If  $X$  and  $Y$  are independent random variable*

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

**Proof.**

$$\begin{aligned} \text{Var}[X + Y] &= E[(X + Y - E[X] - E[Y])^2] = \\ E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] &= \\ \text{Var}[X] + \text{Var}[Y] + 2E[X - E[X]]E[Y - E[Y]] \end{aligned}$$

Since the random variables  $X - E[X]$  and  $Y - E[Y]$  are independent.

$$\text{But } E[X - E[X]] = E[X] - E[X] = 0. \quad \square$$

## Back to Coin Flips

Assume again that we flip  $N$  coins. Let  $X$  be the number of heads.

$X_i = 1$  if the  $i$ -th flip was a head else  $X_i = 0$ .

$E[X_i] = 1/2$ .  $Var[X_i] = 1/4$ .

$$Pr(X \geq 3N/4) \leq Pr(|X - E[X]| \geq N/4) =$$

$$Pr(|X - E[X]| \geq E[X]/2) \leq \frac{Var[X]}{(E[X])^2(1/4)} =$$
$$\frac{N/4}{(N^2/4)(1/4)} = 4/N.$$

A significantly better bound than  $3/4$ .

# Bernoulli Trial

Let  $X$  be a 0-1 random variable such that

$$\Pr(X = 1) = p, \quad \Pr(X = 0) = 1 - p.$$

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p.$$

$$\begin{aligned} \text{Var}[X] &= p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)[1-p+p] = \\ & p(1-p). \end{aligned}$$

# A Binomial Random variable

Consider a sequence of  $n$  independent Bernoulli trials  $X_1, \dots, X_n$ . Let

$$X = \sum_{i=1}^n X_i.$$

$X$  has a **Binomial** distribution  $X \sim B(n, p)$ .

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

$$E[X] = np.$$

$$\text{Var}[X] = np(1 - p).$$

# Algorithm for Computing the Median

The **median** of a set  $X$  of  $n$  distinct elements is the  $\lceil \frac{n}{2} \rceil$  largest element in the set.

If  $n = 2k + 1$ , the median element is the  $k + 1$ -th element in the sorted order.

Easily computed through sorting in  $O(n \log n)$  time. There exists a complicated  $O(n)$  deterministic algorithm.

# Randomized Median Algorithm

**Input:** A set of  $n = 2k + 1$  elements from a totally ordered universe.

**Output:** The  $k + 1$ -th largest element in the set.

1. Pick a (multi)-set  $R$  of  $s = n^{3/4}$  elements in  $S$ , chosen independently and uniformly at random with replacement. Sort the set  $R$ .
2. Let  $d$  be the  $(\frac{1}{2}n^{3/4} - \sqrt{n})$ th smallest element in the sorted set  $R$ .
3. Let  $u$  be the  $(\frac{1}{2}n^{3/4} + \sqrt{n})$ th smallest element in the sorted set  $R$ .
4. By comparing every element in  $S$  to  $d$  and  $u$  compute the set  $C = \{x \in S : d \leq x \leq u\}$ , and the numbers  $\ell_d = |\{x \in S : x < d\}|$  and  $\ell_u = |\{x \in S : x > u\}|$ .
5. If  $\ell_d > n/2$  or  $\ell_u > n/2$  then FAIL.
6. If  $|C| \leq 4n^{3/4}$  then sort the set  $C$ , otherwise FAIL.
7. Output the  $(\lfloor \frac{n}{2} \rfloor - \ell_d + 1)$ st element in the sorted order of  $C$ .

# Intuition

- We can sort sets of size  $o(n)$  in linear time.
- The sample of  $R$  elements are spaced “more or less” evenly among the elements of  $X$ .
- W.h.p. more than  $\frac{1}{2}n^{3/4} - \sqrt{n}$  samples are smaller than the median.
- W.h.p. more than  $\frac{1}{2}n^{3/4} - \sqrt{n}$  samples are larger than the median.
- W.h.p. the median is in the set  $C$ , and  $|C| \leq 4n^{3/4}$ .

Let  $Y_1$  be the number of samples below or equal to the median.

Let  $Y_2$  be the number of samples above or equal to the median.

The algorithm computes the median in  $O(n)$  time if all the following three events hold:

1.  $E_1 : Y_1 \geq \frac{1}{2}n^{3/4} - \sqrt{n}$ .
2.  $E_2 : Y_2 \geq \frac{1}{2}n^{3/4} - \sqrt{n}$ .
3.  $E_3 : |C| \leq 4n^{3/4}$ .

What is the probability that the three random variables  $Y_1, Y_2$  and  $|C|$  are all within the required ranges?.

The sample space in execution of this algorithm is the set of all possible choices of  $n^{3/4}$  elements from  $n$ , with repetitions. (The sample space has  $n^{n^{3/4}}$  points.)

Each point in the sample space defines values for  $Y_1$ ,  $Y_2$  and  $|C|$ .

Computing the probabilities directly is too complicated, instead we use bounds on deviation from the expectation.

$Y_1$  = the number of samples below or equal the median.

What is the probability that  $Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}$

Viewing  $Y_1$  as the sum of  $n^{3/4}$  independent 0-1 random variable, each with expectation  $1/2$  and variance  $1/4$  we prove:

$$E[Y_1] > \frac{1}{2}n^{3/4}.$$

$$\text{Var}[Y_1] < \frac{1}{4}n^{3/4}.$$

Applying Chebyshev Inequality we get:

$$\Pr(\bar{E}_1 : Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \Pr(|Y_1 - E[Y_1]| > \sqrt{n}) \leq$$

$$\frac{\text{Var}[Y_1]}{n} = \frac{n^{3/4}/4}{n} = \frac{1}{4}n^{-1/4}.$$

Similarly

$$\Pr(\bar{E}_2 : Y_2 < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \frac{1}{4}n^{-1/4}.$$

$$\Pr(\bar{E}_1 \cup \bar{E}_2) \leq \frac{2}{4}n^{-1/4}.$$

Recall:  $E_3 : |C| \leq 4n^{3/4}$ .

**Lemma 4.**

$$\Pr(\bar{E}_3) \leq \frac{1}{2}n^{-1/4}.$$

Define the following two events:

1.  $\mathcal{E}_{3,1}$ : at least  $2n^{3/4}$  elements of  $C$  are greater than the median;
2.  $\mathcal{E}_{3,2}$ : at least  $2n^{3/4}$  elements of  $C$  are smaller than the median.

If  $|C| > 4n^{3/4}$ , then at least one of the above two events occurs.

We bound  $\mathcal{E}_{3,1}$ : at least  $2n^{3/4}$  elements of  $C$  are greater than the median;

At least  $2n^{3/4}$  elements of  $C$  above the median  $\Rightarrow$

$u$  is at least the  $\frac{1}{2}n + 2n^{3/4}$  largest in  $S \Rightarrow$

$R$  had at least  $\frac{1}{2}n^{3/4} - \sqrt{n}$  samples among the  $\frac{1}{2}n - 2n^{3/4}$  largest elements in  $S$ .

Let  $X$  be the number of samples among the  $\frac{1}{2}n - 2n^{3/4}$  largest elements in  $S$ . Let  $X = \sum_{i=1}^{n^{3/4}} X_i$  where

$$X_i = \begin{cases} 1 & \text{the } i\text{-th sample in } \frac{1}{2}n - 2n^{3/4} \\ & \text{largest elements in } S \\ 0 & \text{otherwise.} \end{cases}$$

$$E[X_i] = E[(X_i)^2] = \frac{1}{2} - 2n^{-1/4}$$

$$\text{Var}[X_i] = E[(X_i)^2] - (E[X_i])^2 \leq \frac{1}{4}.$$

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

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

Applying Chebyshev's Inequality yields

$$\begin{aligned} \Pr(\mathcal{E}_{3,1}) &= \Pr(X \geq \frac{1}{2}n^{3/4} - \sqrt{n}) \\ &\leq \Pr(|X - E[X]| \geq \sqrt{n}) \\ &\leq \frac{\text{Var}[X]}{n} = \frac{\frac{n^{3/4}}{4}}{n} = \frac{1}{4}n^{-1/4}. \end{aligned}$$

Similarly,

$$\Pr(\mathcal{E}_{3,2}) \leq \frac{1}{4}n^{-\frac{1}{4}},$$

and

$$\Pr(\bar{E}_3) \leq \Pr(\mathcal{E}_{3,1}) + \Pr(\mathcal{E}_{3,2}) \leq \frac{1}{2}n^{-\frac{1}{4}}.$$

The probability that the algorithm succeeds is

$$\geq 1 - (\Pr(\bar{E}_1) + \Pr(\bar{E}_2) + \Pr(\bar{E}_3)) \geq 1 - \frac{1}{n^{1/4}}.$$