The problem

We have a sequence of n (unknown) integers between

$$a_1,\ a_2,\ a_3,\ a_4,\ \ldots, a_n$$

Want to estimate $\overline{a} = \frac{a_1+a_2+\ldots+a_n}{n}$

We sample s integers (with replacement) and output the average.

Let X_i be a random variable associated with the i-th sample.

$$\text{Algorithm's output: } X = \frac{1}{s}(X_1 + \ldots + X_s)$$

$$\text{We know: } E[X] = \overline{a}$$

Deviation from Expectation

We want to know how often X deviates from $E[X]$ by a considerable degree.

In other words, we want to bound this probability ($\epsilon \geq 0$)

$$Pr\Big(\underbrace{|X - E[X]|}_{\text{the amount of deviation}} \geq \epsilon E[X] \Big)$$

We have some useful inequalities for this.

Deviation Bounds

Markov Inequality: For any non-negative random variable X,

$$Pr\big(X \geq t\big) \leq \frac{E[X]}{t} \quad \Rightarrow \quad Pr\big(X \geq tE[X]\big) \leq \frac{1}{t}$$

Chebyshev Inequality: For any random variable X and $t > 0$,

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

Specially (when $t = \epsilon E[X]$),

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

Proof: Apply Markov inequality to the random variable $Y = (X - E[X])^2$.

Applying Chebyshev

We need an upper bound on $Var[X]$.

Since X_i's are independent,

$Var[X] = Var[\frac{1}{s}(X_1 + \ldots + X_s)] = \frac{1}{s^2}(Var[X_1] + \ldots + Var[X_s])$

Since X_i's are identical, $Var[X] = \frac{1}{s^2} s Var[X_i] = \frac{1}{s} Var[X_i]$

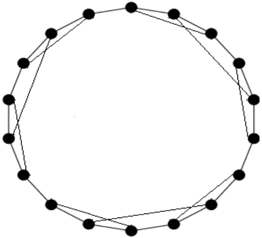
$Var[X_i] = E[X_i^2] - E^2[X_i] = (\frac{a_1^2}{n} + \ldots + \frac{a_n^2}{n}) - \overline{a}^2$

$$
\begin{aligned}
Pr\Big(|\, X - E[X] \,| \geq \epsilon E[X]\Big) \quad &\leq \quad \frac{\frac{1}{s}\Big(\frac{a_1^2 + \ldots + a_n^2}{n} - \overline{a}^2\Big)}{\epsilon^2 \overline{a}^2} \\
&= \quad \frac{1}{\epsilon^2 s}\Big(n \frac{a_1^2 + \ldots + a_n^2}{(a_1 + \ldots + a_n)^2} - 1\Big)
\end{aligned}
$$

How large the term $D = n\frac{a_1^2+\ldots+a_n^2}{(a_1+\ldots+a_n)^2}$ can be?

Lets consider two cases :
(when $a_1, \cdots, a_n$ are degrees of nodes of a graph)

$3, 3, 3, 3, \ldots, 3$

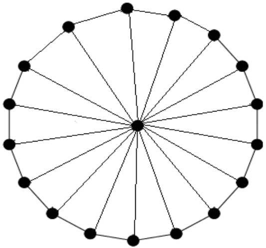


3-regular graph

$\overline{a} = 3 \qquad D = 1$

$\Rightarrow \; s = 1$ is enough

$3, 3, 3, \ldots, 3, n-1, 3, \ldots, 3$



wheel graph

$\overline{a} \approx 4 \qquad D = O(n)$

$\Rightarrow \; s = O(\frac{n}{\epsilon^2})$

- It can be shown that $D \le \frac{a_{max}}{\overline{a}}$ when $a_{max} = \max\{a_i\}$. It suggests $s = O(\frac{a_{max}}{\epsilon^2 \overline{a}})$ is enough.
- The above cases tell us we need random $\Omega(n)$ degree queries to distinguish between $\overline{a} = 3$ and $\overline{a} \approx 4$.
- This shows $\frac{3}{4} + \epsilon$ approximation is not possible using $o(n)$ degree queries.
- Uriel Feige showed that $O(\frac{\sqrt{n}}{\epsilon})$ random degree queries is enough to get a $\frac{1}{2} - \epsilon$ approximation of d.

Lets try a different tool: Chernoff bound

Chernoff Bound: Let $0 \leq \epsilon \leq 1$. Suppose $Y_1, \ldots, Y_t$ are independent random variables taking values in the interval $[0,1]$. Let $Y = \sum_{i=1}^{t} Y_i$. Then

$$Pr\Big(\, |Y - E[Y]| \, \geq \epsilon E[Y] \Big) \leq 2e^{-\frac{\epsilon^2 E[Y]}{3}}$$

Recall that $X = \frac{1}{s}(X_1 + \ldots + X_s)$ where $X_i \in \{1, \ldots, a_{max}\}$.

We define $Y_i = \frac{X_i}{a_{max}} \Rightarrow Y_i \in [0, 1]$.

$$Y = Y_1 + \ldots + Y_s \quad \Rightarrow \quad Y = \frac{s}{a_{max}} X$$

$$E[Y] = \frac{s}{a_{max}} \overline{a}$$

$$\begin{aligned} Pr\Big(|X - E[X]| \geq \epsilon E[X]\Big) \quad &= \quad Pr\Big(|\frac{sX}{a_{max}} - E[\frac{sX}{a_{max}}]| \geq \epsilon E[\frac{sX}{a_{max}}]\Big) \\ &= \quad Pr\Big(|Y - E[Y]| \geq \epsilon E[Y]\Big) \\ &\leq \quad 2e^{-\frac{\epsilon^2 E[Y]}{3}} = 2e^{-\frac{\epsilon^2 s}{3} \frac{\overline{a}}{a_{max}}} \end{aligned}$$

A direct application of Chernoff bound suggest $s = O(\frac{a_{max}}{\bar{a}\epsilon^2})$.

This is the same bound that we obtained using Chebyshev!

In comparison with Chebyshev inequality:

- Chernoff does not need a knowledge of the variance. It only needs the expectation.
- Chernoff gives a much higher probability of concentration.

Comparing Chebyshev and Chernoff

Suppose we want to have error probability $\delta < 0$.

Using Chebyshev we should have:

$$Pr\Big(|X - E[X]| \geq \epsilon E[X]\Big) \leq \frac{1}{\epsilon^2 s}\Big(D - 1\Big) < \frac{1}{\epsilon^2 s}\Big(\frac{a_{max}}{\overline{a}}\Big) \leq \delta$$

$$s > \frac{1}{\delta}\frac{a_{max}}{\epsilon^2 \overline{a}}$$

Using Chernoff we should have:

$$Pr\Big(|X - E[X]| \geq \epsilon E[X]\Big) \leq 2e^{-\frac{\epsilon^2 s}{3}\frac{\overline{a}}{a_{max}}} \leq \delta$$

$$s \geq 3\ln\Big(\frac{1}{2\delta}\Big)\frac{a_{max}}{\epsilon^2 \overline{a}}$$

Another application of Chernoff bound

Amplifying the success probability

Suppose we have a randomized algorithm A that processes the input data D and approximate some $f(D)$ where

$$|A(D) - f(D)| \leq \epsilon f(D) \text{ with probability at least } 3/4.$$

How to amplify the success probability of A?

We want to have a randomized algorithm A' with error probability $\delta << 1/4$.

Idea: Run A on input data D, $O(\ln(\frac{1}{\delta}))$ times and output the median of the outcomes.

Each (independent) repetition of A succeeds with probability $3/4$. Suppose a_i is the outcome of i-th repetition. We have

$$Pr(|a - f(D)| \geq \epsilon f(A)) \leq 1/4.$$

We define $X_i = 1$ if i-th repetition is good (its error is less than $\epsilon f(A)$), otherwise we let $X_i = 0$.

$X = X_1 + \ldots + X_t$ is the number of good outcomes in t repetitions.

The median of $\{a_1, \ldots, a_t\}$ is bad $\Rightarrow$ Less than $t/2$ repetitions are good. In other words, $X < t/2$.

By Chernoff bound, we have

$$Pr(\text{median is bad}) \leq Pr(X < t/2) \leq e^{O(-t)} \leq \delta \;\Rightarrow\; t = (\ln(\frac{1}{\delta}))$$