Lecture 3:

Sublinear time algorithms II

Course: Algorithms for Big Data

Instructor: Hossein Jowhari

Department of Computer Science and Statistics
Faculty of Mathematics
K. N. Toosi University of Technology

Spring 2021

Outline

- Deviation bounds: Markov, Chebyshev, Chernoff
- Estimating the average degree (continued)

Recap from previous lecture

We have a sequence of n (unknown) integers between 1 and $n-1$ (these are degrees of a graph on n nodes.)

$$d_1,\ d_2,\ d_3,\ d_4,\ \ldots, d_n$$

Want to estimate $d = \frac{d_1+d_2+\ldots+d_n}{n}$

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

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

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

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

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

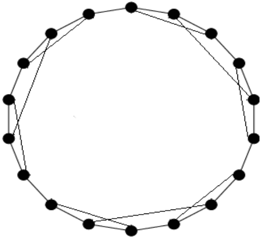
$Var[X_i] = E[X_i^2] - E^2[X_i] = (\frac{d_1^2}{n} + \ldots + \frac{d_n^2}{n}) - d^2$

$$
\begin{aligned}
Pr\Big(| X - E[X] | \geq \epsilon E[X]\Big) &\leq \frac{\frac{1}{s}\Big(\frac{d_1^2 + \ldots + d_n^2}{n} - d^2\Big)}{\epsilon^2 d^2} \\
&= \frac{1}{\epsilon^2 s}\Big(n \frac{d_1^2 + \ldots + d_n^2}{(d_1 + \ldots + d_n)^2} - 1\Big)
\end{aligned}
$$

How large the term $D = n\frac{d_1^2+\ldots+d_n^2}{(d_1+\ldots+d_n)^2}$ can be?

Lets consider two cases:

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

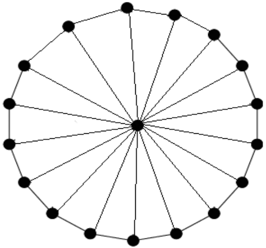


3-regular graph

$d = 3 \qquad D = 1$

$\Rightarrow \; s = 1$ is enough

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



wheel graph

$d \approx 4 \qquad D = O(n)$

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

- It can be shown that $D \leq \frac{d_{max}}{d}$ when $d_{max} = \max\{d_i\}$. It suggests $s = O(\frac{d_{max}}{\epsilon^2 d})$ is enough.
- The above cases tell us we need random $\Omega(n)$ degree queries to distinguish between $d = 3$ and $d \approx 4$.
- 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.
- Note that even if we set $\epsilon = \frac{2}{3}$ (for $\frac{1}{3}$ approximation), Chebyshev inequality needs $s = \Omega(n)$.

  $$Pr\Big(X \leq (1-\epsilon)E[X]\Big) \leq \frac{1}{\epsilon^2 s}\Big(D - 1\Big)$$

In our analysis using Chebyshev inequality, we did not use the fact that $d_1, \ldots, d_n$ is the degree sequence of an undirected graph. In other words, the numbers $d_1, \ldots, d_n$ have a certain relation. Consider the following sequence:

$$\underbrace{1, 1, 1, \ldots, 1, 1}_{n-t}, \underbrace{n, \ldots, n}_{t} \qquad \text{average } \approx t$$

Here we need $s = \Omega(n/t)$ samples for a $\alpha > 1/t$ approximation of the average. ☹

But wait! These numbers cannot be degree sequence of a graph.

Union bound

Let $E_1, \ldots, E_k$ be a collection of events. Then

$$Pr(E_1 \cup E_2 \cup \ldots \cup E_k) \leq \sum_{i=1}^{k} Pr(E_i)$$

Example: In the above sampling task, let E_i be the event that we sample an n in the i-th round.

$$E = E_1 \cup E_2 \cup \ldots \cup E_t \quad \Rightarrow \quad \text{at least one } nissampled$$

If we sample s times, by the union bound, we have

$$Pr(E) \leq \sum_{i=1}^{s} Pr(E_i) = \sum_{i=1}^{s} \frac{t}{n} = \frac{st}{n}$$

If $s < \frac{n}{2t} \quad \Rightarrow \quad Pr(E) < 1/2$

Lets try a different tool: Chernoff bound

Chernoff Bound: Let $0 \le \epsilon \le 1$. 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

$$Pr\Big(Y \le (1 - \epsilon) E[Y] \Big) \le e^{-\frac{\epsilon^2 E[Y]}{2}}$$

$$Pr\Big(Y \ge (1 + \epsilon) E[Y] \Big) \le e^{-\frac{\epsilon^2 E[Y]}{3}}$$

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

Proof: We present an incomplete proof shortly.

Lets apply Chernoff inequality to our problem.

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

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

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

$$E[Y] = \frac{s}{d_{max}} d$$

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

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

This is the same bound that we obtained using Chebyshev!

This is again what we expected because we have not yet used the fact that the numbers $d_1, \dots, d_n$ are the degree sequence of a graph.

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:

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

$$s > \frac{1}{\delta}\frac{d_{max}}{\epsilon^2 d}$$

Using Chernoff we should have:

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

$$s \geq 3\ln(\frac{1}{2\delta})\frac{d_{max}}{\epsilon^2 d}$$

An (incomplete) proof of Chernoff bound

Claim: Let $Y = Y_1 + \ldots + Y_t$ where Y_i's are independent random variables taking values in the interval $[0, 1]$. Let $\mu = E[Y]$. Then

$$Pr\Big(Y \geq (1+\epsilon)\mu\Big) \leq \Big(\frac{e^{\epsilon}}{(\epsilon+1)^{\epsilon+1}}\Big)^{\mu}$$

Proof: Fix $\theta > 0$.

$$Pr\Big(Y \geq (1+\epsilon)\mu\Big) = Pr\Big(e^{\theta Y} \geq e^{\theta(1+\epsilon)\mu}\Big)$$

because $f(x) = e^x$ is a monotone function.

$$Pr\Big(e^{\theta Y} \geq e^{\theta(1+\epsilon)\mu}\Big) \leq \frac{E\big[e^{\theta Y}\big]}{e^{\theta(1+\epsilon)\mu}} = \frac{E\big[e^{\theta(Y_1+\ldots+Y_t)}\big]}{e^{\theta(1+\epsilon)\mu}}$$

by Markov inequality.

Since Y_i's are independent, ($E[XY] = E[X]E[Y]$ when Y and X are independent.)

$$\frac{E[e^{\theta(Y_1+\ldots+Y_t)}]}{e^{\theta(1+\epsilon)\mu}} = \frac{E[e^{\theta Y_1}] \times \ldots \times E[e^{\theta Y_t}]}{e^{\theta(1+\epsilon)\mu}}$$

We show $E[e^{\theta Y_i}] \leq e^{(e^\theta - 1)E[Y_i]}$. Since $Y_i \in [0, 1]$,

$$E[e^{\theta Y_i}] \leq E[1 + (e^\theta - 1)Y_i] = 1 + (e^\theta - 1)E[Y_i] \leq e^{(e^\theta - 1)E[Y_i]}$$

Because for all $x \in [0, 1]$ and $\theta > 0$, we have

$$e^{\theta x} \leq 1 + (e^\theta - 1)x \leq e^{(e^\theta - 1)x}$$

$$\frac{\prod_{i=1}^{t} E[e^{\theta Y_i}]}{e^{\theta(1+\epsilon)\mu}} \leq \frac{\prod_{i=1}^{t} e^{(e^\theta - 1)E[Y_i]}}{e^{\theta(1+\epsilon)\mu}} = e^{(e^\theta - 1)\mu - \theta(1+\epsilon)\mu} = e^{((e^\theta - 1) - \theta(1+\epsilon))\mu}$$

The claim follows after setting $\theta = \ln(1 + \epsilon)$.

An 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

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