Recap :

Concentration Inequalities :

used for showing that random variable for a sample drawn from some dist.ⁿ is "close" to the expected value, with high prob.

1. Chebyshev's Ineq.

If $\text{Var}[X] = \sigma^2$, then

$$P_r\left[\,|X - \mu| \geq k\,\right] \leq \sigma^2/k^2$$

2. Chernoff Bound :

$$X = X_1 + \cdots + X_t$$

X_i s are all independent binary valued random variables

$$P_r\left[\,X \geq (1+\delta)\mu\,\right] \leq e^{-\mu\delta^2/(2+\delta)}$$

$$P_r\left[\,X \leq (1-\delta)\mu\,\right] \leq e^{-\mu\delta^2/2}$$

Plan :

- Other applications of Chernoff bounds.
- Wrap up probabilistic method (sum-free subset problem)

Application 1 : Exit Polls

n balls in a bag, each ball is red or blue. You are given that at least 60% balls are red, or at least 60% balls are blue. You want to estimate whether there are more blue balls, or more red balls, and you want the estimate to be correct w.p. at least $1 - 1/n$.

Algorithm :

1. Sample t balls, uniformly at random, with replacement. Note the color of i^{th} sampled ball.

2. If majority of sampled balls are red, estimate that there are more red balls. Else estimate that there are more blue balls.

Qn: How many samples are needed?

Ans: $O(\log n)$!

This should be a bit surprising. You only need $O(\log n)$ samples to estimate the majority correctly with high probability. This is the basic science behind exit polls conducted in election season.

Analysis of our algorithm's correctness:

Case 1: no. of red balls $\geq 0.6\,n$
no. of blue balls $\leq 0.4\,n$.

Suppose we take $t = c \ln n$ samples. for some constant c (to be fixed later). We want to show an upper bound on the prob. that our algorithm gives wrong output.

$$X_i = \begin{cases} 1 & \text{if } i\text{th sample is blue} \\ 0 & \text{otherwise.} \end{cases}$$

$$\Pr\left[\text{Algorithm gives wrong answer}\right]$$

$$= \Pr\left[X_1 + \dots + X_t > t/2\right]$$

Let $X = X_1 + \dots + X_t$

X_is are all independent random variables.

$$E[X_i] \leq 0.4$$

$$\mu = E[X] \leq 0.4t$$

$$Pr\left[X > t/2 \right]$$

$$\leq Pr[X > \mu + 0.1t]$$

$$= Pr[X > \mu(1 + t/10\mu)]$$

Similar to what we showed in last lecture, there exists a constant $c > 0$ s.t.

$$Pr[X > \mu(1 + t/10\mu)] \leq e^{-t/c}.$$

set $t = c \ln n$

Application 2: Coin Tossing

unbiased coin, tossed until we receive n heads.

X: number of coin tosses for n heads.

$$E[X] = ?$$

Suppose we toss an unbiased coin n times.
$E[\text{number of heads in } n \text{ tosses}] = n/2$

As a result, it may be tempting to immediately conclude that $E[X] = 2n$. The answer is correct, but the reasoning is not correct. It is important that you convince yourself that this reasoning is flawed.

Correct approach:

$$X_i = \begin{cases} \text{number of coin tosses after the } (i-1)^{th} \\ \text{heads in order to get } i^{th} \text{ heads.} \end{cases}$$

$$X_i \in \{1, 2, 3, \dots\}$$

$X = X_1 + X_2 + \cdots + X_n$. By linearity of exp.

$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n]$.

Recall, in Lecture 19, we showed the following:

Tossing a p-biased coin. $Pr[H] = p$.

X = Number of coin tosses needed to get the first H. $E[X] = 1/p$.

Therefore, $E[X] = 2n$ (since $E[X_i] = 2$ for all i).

$$Pr[X > 50n] = ?$$

Can't use Chernoff directly since X_is are not binary valued.

At this point, you have three options:

(i) is there a different event that has identical prob. as this one, or greater prob. than this one?

(ii) is there a different r.v. decomposition s.t. $X = \sum X_i'$, and X_i' are independent and binary valued?

(iii) is there a generalization of Chernoff bds. that allows X_is to be non binary?

For this problem, (i) suffices.

$$Pr\left[X > 50n\right] = Pr\left[\begin{array}{c}\text{number of heads in } 50n \\ \text{tosses is at most } n-1\end{array}\right]$$

Now consider a different experiment.
Toss an unbiased coin $50n$ times.

Y = no. of heads in the $50n$ coin tosses

$$Pr\underbrace{\left[X > 50n\right]}_{\mathcal{E}} = Pr\underbrace{\left[Y < n\right]}_{\mathcal{E}'}$$

Note that we have actually changed the sample space.

In LHS, $\Omega = \{$ sequences of H/T that end with H, and have exactly n H $\}$

In RHS, $\Omega' = \{H, T\}^{50n}$

$$\mathcal{E} \subset \Omega, \quad \mathcal{E}' \subset \Omega'.$$

There is a surjection from $\mathcal{E}$ to $\mathcal{E}'$.
Take any sequence in $\mathcal{E}$, note it has length greater than $50n$. Truncate it to length of $50n$.
Let us call this truncation operation $T : \mathcal{E} \to \mathcal{E}'$.

Obs. : Take any $\omega' \in \Omega'$. Let $\Gamma = T^{-1}(\omega')$.
Then $Pr[\omega'] = Pr[\Gamma]$.

$\Pr[Y < n]$ is easy to analyze using Chernoff bds.

$Y = Y_1 + \dots + Y_{50n}$ where $Y_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ coin toss is Heads} \\ 0 & \text{otherwise} \end{cases}$

$$E[Y] = 25n = \mu$$

$$\Pr[Y < n] = \Pr\left[Y < \left(1 - \frac{24}{25}\right)\mu\right] < e^{-\frac{25n}{2}\left(\frac{24}{25}\right)^2}$$

$$= e^{-n}.$$

Thm 21.1 : for all n,

$$\Pr\left[\begin{array}{c}\text{number of unbiased coin tosses to get } n \text{ heads} \\ > 50n\end{array}\right] < e^{-n}$$

Similarly, you can show that if the coin is biased but you know a lower bound i.e. $\Pr[\text{Heads}] \geq p$, then the number of coin tosses to get n heads can't be much larger than n/p.

Two exercises on using various concentration inequalities :

These exercises are fairly non-trivial, however the tools developed in last two lectures should be enough to analyze them. Solutions for these problems will not be provided. However, I am happy to discuss them with you

NDY

Exercise 1 : Push Protocol

a popular protocol in computer networks

There are n people, everyone knows each other. Person 1 makes up a rumor. The rumor spreads, in a completely distributed manner, as follows. If a person knows the rumor, then every day, he/she picks one of the n people uniformly at random, and tells the rumor.

How many days will it take for everyone to know the rumor?

X = number of days for everyone to know the rumor.

What is $E[X]$?

eg. 8 people.

Day 1: person 1 knows the rumor.
person 1 tells it to person 3.

Day 2: person 1 and person 3 know the rumor.
person 1 tells it to person 5.
person 3 tells it to person 7.

Day 3: persons 1, 3, 7 know the rumor.
$1 \longrightarrow 4$, $3 \longrightarrow 7$
$7 \longrightarrow 1$
note that 3 tells it to the same person twice.

Day 4: persons 1, 3, 4, 7 know the rumor.
$1 \longrightarrow 8$, $3 \longrightarrow 2$, $4 \longrightarrow 1$, $7 \longrightarrow 6$

Day 5: persons 1, 2, 3, 4, 6, 7, 8 know the rumor.

Day 6: persons 1, 2, 3, 4, 6, 7, 8 know the rumor.
$1 \longrightarrow 4$, $2 \longrightarrow 7$, $3 \longrightarrow 6$, $4 \longrightarrow 8$
$6 \longrightarrow 1$, $7 \longrightarrow 8$, $8 \longrightarrow 3$

Day 7: persons 1, 2, 3, 4, 6, 7, 8 know the rumor.
$1 \longrightarrow 3$, $2 \longrightarrow 5$, $3 \longrightarrow 4$, $4 \longrightarrow 7$
$6 \longrightarrow 2$, $7 \longrightarrow 1$, $8 \longrightarrow 5$

Everyone knows the rumor in 7 days.

Observation : $X \geq \log_2(n)$

and therefore $E[X] = \Omega(\log_2 n)$

Every day, the number of people knowing the rumor can at most double.

Prove : There exist constants c_1, c_2 s.t.

$$E[X] \leq c_1 \log n$$

$$Pr\left[X > c_2 \log n \right] < 1/n .$$

NDY

Exercise 2 : Distributed clients and servers.

There are n clients, n servers.

Every client has 1 job, that takes T unit of time.

At time steps $t = 0, T, 2T, 3T, \ldots,$ the clients and servers do the following :

Client: If its job is processed, then it does nothing.
Else it picks a unif. rand. server, and sends its job to server.

Server: Among the jobs received, picks one unif. at random and processes it. The remaining are sent back to the respective clients.

Assume all communication is instantaneous.

Example: 5 clients and servers

Rd. 1 C1 picks S4, C2 picks S1, C3 picks S4, C4 picks S5, C5 picks S1
S1 processes C5, S4 processes C1, S5 processes C4

Rd. 2 C3 picks S2, C2 picks S1. Both jobs get processed.

X: total time for all jobs to be completed.

$\mu = E[X] = ?$ $\Pr[X > 100\mu] \leq \ldots$

Probabilistic Method : Sum free subsets.

$S \subseteq \mathbb{N}, \quad |S| = n$

Prove : There exists a subset $T \subseteq S$, $|T| \geq n/3$ s.t. $\forall\ a, b, c \in T, \quad a+b \neq c.$

This proof is a clever combination of probability and the number theory discussed earlier in the course. The solution is very elegant, and I will not expect you to solve such questions in quiz / exam. However, it is a demonstration of the power of prob. method.

Please go over this proof carefully.

$S \subseteq \mathbb{N}$, $|S| = n$. Every element of S is non-zero. Let p be a prime greater than all elements of S and $p \bmod 3 = 2$.
There are infinitely many primes of form $3k+2$, therefore we can take one such prime greater than all elements of S.

Magic #1 : where did this prime come from?

For any $\theta \in \mathbb{Z}_p$, consider the function

$$f_\theta : S \to \mathbb{Z}_p, \quad \text{where}$$

$$f_\theta(x) = x \times_p \theta \equiv \theta \cdot x \bmod p.$$

Observation 1: $\forall\ a, b, c \in S$, $\forall\ \theta \in \mathbb{Z}_p$,

$$a + b = c \implies f_\theta(a) +_p f_\theta(b) = f_\theta(c).$$

In other words, $\forall\ a, b, c \in S$, $\forall\ \theta \in \mathbb{Z}_p$,

$$\left[f_\theta(a) +_p f_\theta(b) \neq f_\theta(c) \right] \implies a + b \neq c$$

Observation 2: Let $\Gamma_2 = \{k+1, k+2, \ldots, 2k+1\}$.

$$|\Gamma_2| = k+1 > \frac{p}{3} \quad \text{since } p = 3k+2$$

$$\forall\ a, b, c \in \Gamma_2, \quad a +_p b \neq c.$$

Clearly, $a + b$ will not be in Γ_2. But note that $a + b \bmod p$ will also not be in Γ_2.

?

Note: For $\Gamma_3 = \{2k+2, \ldots, 3k+1\}$, we can say that $\forall\ a, b, c \in \Gamma_3$, $a + b \neq c$. However, $|\Gamma_3| < p/3$.

Observation 3: For any $x \in S$,

$$\Pr\left[f_\theta(x) \in \Gamma_2 \right] = \frac{k+1}{p} > \frac{1}{3}$$

where probability is over choice of θ, sampled uniformly from $\mathbb{Z}_p$.

Note: $x \neq 0$, $x \in \mathbb{Z}_p$ since $x < p$. For any $y \in \mathbb{Z}_p$, $\Pr[f_\theta(x) = y] = 1/p$ since there is a unique θ s.t. $f_\theta(x) = y$.

Observation 4: Let $f_\theta(S) = \{ f_\theta(x) : x \in S \}$

$$E\left[\left| f_\theta(S) \cap \Gamma_2 \right| \right] = \frac{k+1}{p} \cdot n > \frac{n}{3}.$$

$$X = \left| f_\theta(S) \cap \Gamma_2 \right|.$$

$$X = \sum_{w \in S} X_w \quad \text{where} \quad X_w = \begin{cases} 1 & \text{if } f_\theta(w) \in \Gamma_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr[X_w = 1] = \frac{k+1}{p} \Rightarrow E[X] > n/3.$$

Using prob. method, there exists $\theta \in \mathbb{Z}_p$ s.t.

$$\left| f_\theta(S) \cap \Gamma_2 \right| > \frac{n}{3}.$$

Fix this θ. Let $T = \{ w \in S \text{ s.t. } f_\theta(w) \in \Gamma_2 \}$

$$|T| > n/3.$$

Using Observation 1 and Observation 2, we conclude that for all $a, b, c \in T$, $a + b \neq c$. ∎

Exercise: Suppose you didnt know that there are infinitely many primes of form $3k+2$. Alter the above proof so that it works even when $p = 3k+1$.

We will take a break from probability and probabilistic method, and start with graph theory in Lecture 22. However, we will come back to prob. and prob. method at the end of the course.

Summary :

- Concentration inequalities
  - Markov ineq.
  - Chebyshev ineq.
  - Chernoff bounds.

- Chernoff requires $X = \sum X_i$ where all X_is are binary valued and independent. If not, then check if:

(i) there is a different event that has identical prob. as this one, or greater prob. than this one, and the new event can be analyzed using Chernoff bounds.

(ii) there is a different r.v. decomposition s.t. $X = \sum X_i'$, and X_i' are independent and binary valued

(iii) there is a generalization of Chernoff bds. that allows X_is to be non binary

- Probabilistic method : To show the existence of an object Θ in some space S,
  - sample from S using some distribution
  - show $Pr[\Theta] > 0$.