CMPT 210: Probability and Computation

Lecture 21

Sharan Vaswani

July 26, 2022

Recap

Tail inequalities bound the probability that the r.v. takes a value much different from its mean.

Markov's Theorem: If X is a non-negative random variable, then for all $x > 0$, $\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}$.

Chebyshev's Theorem: For a r.v. X and all $x > 0$, $\Pr[|X - \mathbb{E}[X]| \geq x] \leq \frac{\mathrm{Var}[X]}{x^2}$.

Randomized QuickSort

Given an array A of n distinct numbers, sort the elements in A in increasing order.

Algorithm Randomized QuickSort

```
function QuickSort($A$)
If Length(A) $= 1$, return A.
Select $p \in A$ uniformly at random.
Construct arrays Left $:= [x \in A | x < p]$ and Right $:= [x \in A | x > p]$.
Return the concatenation [QuickSort(Left), $p$, QuickSort(Right)].
```

Randomized QuickSort

If $A = [2, 7, 0, 1, 3]$ and according to the algorithm, $p \sim \text{Uniform}(A)$. Say $p = 3$. For this step, $\text{Left} = [2, 0, 1]$ and $\text{Right} = [7]$.

The algorithm will return the concatenation $[\text{QuickSort}([2, 0, 1]), 3, \text{QuickSort}([7])] = [\text{QuickSort}([2, 0, 1]), 3, 7]$.

Total number of comparisons $= 4$ (comparing every element to the pivot $= 3$.)

In the second step, for running the algorithm on $[2, 0, 1]$, say $p = 1$. For this step, $\text{Left} = [0]$ and $\text{Right} = [2]$ and the algorithm will return the concatenation $[\text{QuickSort}([0]), 1, \text{QuickSort}([2]), 3, 7] = [0, 1, 2, 3, 7]$.

Total number of comparisons $= 4$ (from step 1) $+ 2$ (comparing elements in Left to pivot $= 1$.)

Q: Run the algorithm if $p = 2$ in the first step?

Ans: $\text{Left} = [0, 1]$ and $\text{Right} = [7, 3]$. Running the algorithm on $[0, 1]$ will return $[0, 1]$ and on $[7, 3]$ will return $[3, 7]$. Hence the algorithm will return the concatenation $[0, 1, 2, 3, 7]$ thus sorting the array.

Randomized QuickSort

Claim: For a set A with n distinct elements, the expected (over the randomness in the pivot selection) number of comparisons for QuickSort is $O(n \ln(n))$.

Let us write the elements of A in sorted order, $a_1 < a_2 < \ldots < a_n$. Let X be the r.v. equal to the number of comparisons performed by the algorithm.

Observation: Every pair of elements is compared at most once since we do not include the pivot in the recursion.

For $i < j$, let $E_{i,j}$ be the event that elements i and j are compared, and define $X_{i,j}$ to be the indicator r.v. equal to 1 if event $E_{i,j}$ happens. Hence, $X = \sum_{1 \leq i < j \leq n} X_{i,j}$, and

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{1 \leq i < j \leq n} X_{i,j}\right] = \sum_{1 \leq i < j \leq n} \mathbb{E}[X_{i,j}] = \sum_{1 \leq i < j \leq n} \Pr[E_{i,j}] \qquad \text{(Linearity of expectation)}$$

Randomized QuickSort

Fix $i < j$ (meaning that $a_i < a_j$) and let $R = [a_i, \dots, a_j]$.

Claim: $E_{i,j}$ happens if and only if the first pivot selected from R is either a_i or a_j.

Elements a_i and a_j are compared if they are still in the same sub-problem at the time that one of them is chosen as the pivot. Elements a_i and a_j are split into different recursive sub-problems at precisely the time that the first pivot is selected from R. If this pivot is either a_i or a_j, then they will be compared; otherwise, they will not.

In our example, $A = [2, 7, 0, 1, 3]$ and suppose $a_i = 0$ and $a_j = 2$. After $p = 3$ is chosen, Left $= [2, 0, 1]$. Both 0 and 2 are compared to the pivot $p = 3$, and end up in the same sub-problem. Hence the elements in $R = [0, 1, 2]$ appear together.

For the next step, when recursing on Left, if $p = 1$, then Left $= [0]$ and Right $= [2]$ and elements 0 and 2 will never be compared. On the other hand, if $p = 2$, then since each element is compared to the pivot, 0 and 2 will be compared.

Hence, $E_{i,j}$ will happen if the first pivot selected from R is either a_i or a_j.

Randomized QuickSort

Claim: $\Pr[a_i$ or a_j is the first pivot selected from $R] = \frac{2}{|R|} = \frac{2}{j-i+1}$.

In our example, if $a_i = 0$ and $a_j = 2$ and say $p = 7$, then after the first step, Left $= [2, 0, 1, 3]$. Hence the elements in $R = [0, 1, 2]$ appear together in the same sub-problem.

For the second step, when recursing on $T = [2, 0, 1, 3]$, since p is chosen uniformly at random, conditioned on the event that $p \in R$, p is also uniformly random on R. Formally, for $x \in T$, $\Pr[p = x] = \frac{1}{|T|}$.

$$\Pr[p = x | p \in R] = \frac{\Pr[p = x \cap p \in R]}{\Pr[p \in R]} = \frac{\Pr[p = x]}{\Pr[p \in R]} \quad \text{(For all } x \notin R\text{, } \Pr[p = x \cap p \in R] = 0\text{)}$$

$$= \frac{1/|T|}{\sum_{x \in R} \Pr[p = x]} = \frac{1/|T|}{|R|/|T|} = \frac{1}{|R|}$$

Hence, the probability of selecting either 0 or 2 (a_i and a_j respectively) in a sub-array (T in the above example) that contains R ($[0, 1, 2]$ in the example) is $2/|R| = 2/(j - i + 1)$ (equal to $2/3$ in the example).

Randomized QuickSort

Putting everything together, $\Pr[E_{i,j}] = \frac{2}{j-i+1}$.

Hence, the expected number of comparisons is equal to

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{1 \le i < j \le n} \frac{2}{j-i+1} = \sum_{i=1}^{n-1} \left[\sum_{j=i+1}^{n} \frac{2}{j-i+1} \right] = 2 \sum_{i=1}^{n-1} \left[\frac{1}{2} + \frac{1}{3} \dots + \frac{1}{n-i+1} \right] \\
&< 2 \sum_{i=1}^{n-1} \left[\frac{1}{2} + \frac{1}{3} \dots + \frac{1}{n} \right] < 2n \left[\frac{1}{2} + \frac{1}{3} \dots + \frac{1}{n} \right] \\
&\le 2n \int_{1}^{n} \frac{dx}{x} = 2n \ln(n) \qquad \text{(Bounding the harmonic series similar to Lecture 14)}
\end{aligned}$$

Hence, the expected number of comparisons required for Randomized QuickSort is $O(n \ln(n))$.

Q: What is the number of comparisons for Randomized QuickSort in the worst-case? Similar to Randomized QuickSelect, for Randomized QuickSort, the worst-case happens when the pivot is selected to be the minimum (or maximum) element in the sub-array in each iteration. And hence the worst-case complexity is $O(n^2)$.

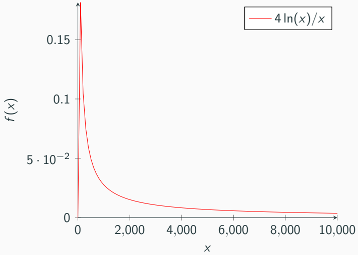
Markov's Theorem for Randomized QuickSort

Since X (the r.v. corresponding to the number of comparisons) is non-negative, we can use Markov's Theorem – For $x > 0$, $\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x} < \frac{2n\ln(n)}{x}$ If $x = 200n\ln(n)$, then, $\Pr[X \geq 200n\ln(n)] < \frac{2}{200} = 0.01$.

Similarly, if we want to investigate how likely is the worst-case behaviour, let us set $x = 2n^2$. In this case,

$$\Pr[X \geq 0.5n^2] < \frac{2n\ln(n)}{0.5n^2} = \frac{4\ln(n)}{n}$$

As n increases, the probability of worst-case behaviour decreases.



Questions?

Sums of Random Variables

If we know that the r.v X is (i) non-negative and (ii) $\mathbb{E}[X]$, we can use Markov's Theorem to bound the probability of deviation from the mean.

If we know both (i) $\mathbb{E}[X]$ and (ii) $\text{Var}[X]$, we can use Chebyshev's Theorem to bound the probability of deviation.

In many cases (the voter poll example), we know the distribution of the r.v. (for voter poll, $S_n \sim \text{Bin}(n, p)$) and can obtain tighter bounds on the deviation from the mean.

Chernoff Bound: Let $T_1, T_2, \ldots, T_n$ be mutually independent r.v's such that $0 \leq T_i \leq 1$ for all i. If $T := \sum_{i=1}^{n} T_i$, for all $c \geq 1$ and $\beta(c) := c \ln(c) - c + 1$,

$$\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\, \mathbb{E}[T])$$

If $T_i \sim \text{Ber}(p)$ and are mutually independent, then $T_i \in \{0, 1\}$ and we can use the Chernoff bound to bound the deviation from the mean for $T \sim \text{Bin}(n, p)$. In general, if $T_i \in [0, 1]$, the Chernoff Bound can be used even if the T_i's have different distributions!

Chernoff Bound – Binomial Distribution

Q: Bound the probability that the number of heads that come up in 1000 independent tosses of a fair coin exceeds the expectation by 20% or more.

Let T_i be the r.v. for the event that coin i comes up heads, and let T denote the total number of heads. Hence, $T = \sum_{i=1}^{1000} T_i$. For all i, $T_i \in \{0, 1\}$ and are mutually independent r.v's. Hence, we can use the Chernoff Bound.

We want to compute the probability that the number of heads is larger than the expectation by 20% meaning that $c = 1.2$ for the Chernoff Bound. Computing $\beta(c) = c\ln(c) - c + 1 \approx 0.0187$. Since the coin is fair, $\mathbb{E}[T] = 1000\,\frac{1}{2} = 500$. Plugging into the Chernoff Bound,

$$\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\mathbb{E}[T]) \implies \Pr[T \geq 1.2\mathbb{E}[T]] \leq \exp(-(0.0187)\,(500)) \approx 0.0000834.$$

Comparing this to using Chebyshev's inequality,

$$\begin{aligned}\Pr[T \geq c\mathbb{E}[T]] &= \Pr[T - \mathbb{E}[T] \geq (c-1)\mathbb{E}[T]] \leq \Pr[|T - \mathbb{E}[T]| \geq (c-1)\,\mathbb{E}[T]] \\ &\leq \frac{\mathrm{Var}[T]}{(c-1)^2\,(\mathbb{E}[T])^2} = \frac{1000\,\frac{1}{4}}{(1.2-1)^2(500^2)} = \frac{250}{0.2^2\,500^2} = \frac{250}{10000} = 0.025.\end{aligned}$$

Chernoff Bound – Lottery Game

Q: Pick-4 is a lottery game in which you pay \$1 to pick a 4-digit number between 0000 and 9999. If your number comes up in a random drawing, then you win \$5,000. Your chance of winning is 1 in 10000. If 10 million people play, then the expected number of winners is 1000. When there are 1000 winners, the lottery keeps \$5 million of the \$10 million paid for tickets. The lottery operator's nightmare is that the number of winners is much greater – especially at the point where more than 2000 win and the lottery must pay out more than it received. What is the probability that will happen?

Let T_i be an indicator for the event that player i wins. Then $T := \sum_{i=1}^{n} T_i$ is the total number of winners. If we assume that the players' picks and the winning number are random, independent and uniform, then the indicators T_i are independent, as required by the Chernoff bound.

We wish to compute $\Pr[T \geq 2000] = \Pr[T \geq 2\mathbb{E}[T]]$. Hence $c = 2$ and $\beta(c) \approx 0.386$. By the Chernoff bound,

$$\Pr[T \geq 2\mathbb{E}[T]] \leq \exp(-\beta(c)\mathbb{E}[T]) = \exp(-(0.386)\,1000) < \exp(-386) \approx 10^{-168}$$

Chernoff Bound – Proof

We want to compute $\Pr[T \geq c\mathbb{E}[T]] = \Pr[f(T) \geq f(c\mathbb{E}[T])]$ where f is a one-one function. For $c \geq 1$, choosing $f(T) = c^T$ and using Markov's Theorem,

$$\begin{aligned}
\Pr[T \geq c\mathbb{E}[T]] &= \Pr[c^T \geq c^{c\mathbb{E}[T]}] \leq \frac{\mathbb{E}[c^T]}{c^{c\mathbb{E}[T]}} \\
&\leq \frac{\exp((c-1)\,\mathbb{E}[T])}{c^{c\mathbb{E}[T]}} \qquad\qquad (\text{To prove next: } \mathbb{E}[c^T] \leq \exp((c-1)\,\mathbb{E}[T])) \\
&= \frac{\exp((c-1)\,\mathbb{E}[T])}{\exp(\ln(c^{c\mathbb{E}[T]}))} = \frac{\exp((c-1)\,\mathbb{E}[T])}{\exp(c\mathbb{E}[T]\,\ln(c))} = \exp\left(-(c\ln(c) - c + 1)\,\mathbb{E}[T]\right) \\
\implies \Pr[T \geq c\mathbb{E}[T]] &\leq \exp\left(-\beta(c)\,\mathbb{E}[T]\right)
\end{aligned}$$

The proof would be done if we prove that $\mathbb{E}[c^T] \leq \exp((c-1)\,\mathbb{E}[T])$ and we do this next.

Chernoff Bound – Proof

Claim: $\mathbb{E}[c^T] \leq \exp((c-1)\,\mathbb{E}[T])$

$$\mathbb{E}[c^T] = \mathbb{E}[c^{\sum_{i=1}^n T_i}] = \mathbb{E}\left[\prod_{i=1}^n c^{T_i}\right] = \prod_{i=1}^n \mathbb{E}[c^{T_i}]$$

(Expectation of product of mutually independent r.v's is equal to the product of the expectation.)

For two variables, the proof of the above statement is in Lecture 15 and can be easily generalized.

$$\leq \prod_{i=1}^n \exp((c-1)\,\mathbb{E}[T_i]) \qquad \text{(To prove next: } \mathbb{E}[c^{T_i}] \leq \exp((c-1)\,\mathbb{E}[T_i])\text{)}$$

$$= \exp\left((c-1)\sum_{i=1}^n \mathbb{E}[T_i]\right) = \exp\left((c-1)\,\mathbb{E}\left[\sum_{i=1}^n T_i\right]\right)$$

(Linearity of Expectation)

$$\implies \mathbb{E}[c^T] \leq \exp((c-1)\mathbb{E}[T])$$

The proof would be done if we prove that $\mathbb{E}[c^{T_i}] \leq \exp((c-1)\,\mathbb{E}[T_i])$ and we do this next.

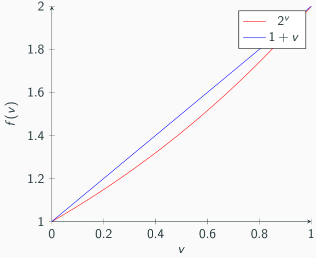
Chernoff Bound – Proof

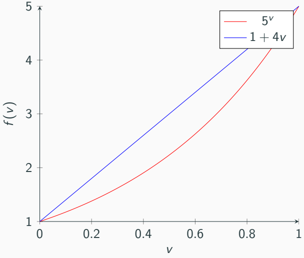
Claim: $\mathbb{E}[c^{T_i}] \leq \exp((c-1)\,\mathbb{E}[T_i])$

$$\mathbb{E}[c^{T_i}] = \sum_{v \in \mathrm{Range}(T_i)} c^v \Pr[T_i = v] \leq \sum_{v \in \mathrm{Range}(T_i)} (1 + (c-1)v) \Pr[T_i = v]$$

(Since $T_i \in [0,1]$ and $c^v \leq 1 + (c-1)v$ for all $v \in [0,1]$.)

For $c = 2$ and $c = 5$,



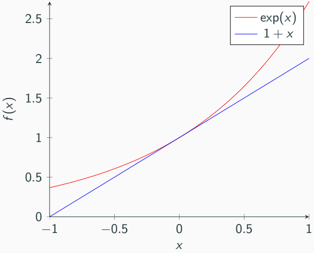


Chernoff Bound – Proof

$$\mathbb{E}[c^{T_i}] \leq \sum_{v \in \mathrm{Range}(T_i)} \Pr[T_i = v] + (c-1) \sum_{v \in \mathrm{Range}(T_i)} v \Pr[T_i = v]$$

$$= 1 + (c-1)\mathbb{E}[T_i] \leq \exp((c-1)\mathbb{E}[T_i]) \qquad \text{(Since } 1 + x \leq \exp(x) \text{ for all } x\text{)}$$

$$\implies \mathbb{E}[c^{T_i}] \leq \exp((c-1)\,\mathbb{E}[T_i])$$



Hence we have proved the Chernoff Bound!

Comparing the Bounds

For r.v's $T_1, T_2, \ldots T_n$, if $T_i \in \{0, 1\}$ and $\Pr[T_i = 1] = p_i$. Define $T := \sum_{i=1}^{n} T_i$. By linearity of expectation, $\mathbb{E}[T] = \sum_{i=1}^{n} p_i$. For $c \geq 1$,

Markov's Theorem: $\Pr[T \geq c\mathbb{E}[T]] \leq \frac{1}{c}$. Does not require T_i's to be independent.

Chebyshev's Theorem: $\Pr[T \geq c\mathbb{E}[T]] \leq \frac{\mathrm{Var}[T]}{(c-1)^2(\mathbb{E}[T])^2}$. If the T_i's are pairwise independent, by linearity of variance, $\mathrm{Var}[T_i] = \sum_{i=1}^{n} p_i\,(1 - p_i)$. Hence, $\Pr[T \geq c\mathbb{E}[T]] \leq \frac{\sum_{i=1}^{n} p_i\,(1-p_i)}{(c-1)^2\left(\sum_{i=1}^{n} p_i\right)^2}$. If for all i, $p_i = 1/2$, then, $\Pr[T \geq c\mathbb{E}[T]] \leq \frac{1}{(c-1)^2}$.

Chernoff Bound: If T_i' are mutually independent, then,
$\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\,\mathbb{E}[T]) = \exp\left(-(c\ln(c) - c + 1)\,\left(\sum_{i=1}^{n} p_i\right)\right)$. If for all i, $p_i = 1/2$,
$\Pr[T \geq c\mathbb{E}[T]] \leq \exp\left(-\frac{n\,(c\ln(c)-c+1)}{2}\right) \approx \exp(-c\,n/2)$.

Questions?

Randomized Load Balancing

Fussbook is a new social networking site oriented toward unpleasant people. Like all major web services, Fussbook has a load balancing problem: it receives lots of forum posts that computer servers have to process. If any server is assigned more work than it can complete in a given interval, then it is overloaded and system performance suffers. That would be bad, because Fussbook users are not a tolerant bunch.

The programmers of Fussbook just randomly assigned posts to computers, and to their surprise the system has not crashed yet.

Fussbook receives 24000 forum posts in every 10-minute interval. Each post is assigned to one of several servers for processing, and each server works sequentially through its assigned tasks. It takes a server an average of $1/4$ second to process a post. No post takes more than 1 second.

This implies that a server is definitely overloaded when it is assigned more than 600 units of work in a 10-minute interval. On average, for $24000 \times \frac{1}{4} = 6000$ units of work in a 10-minute interval, Fussbook requires at least 10 servers to ensure that no server is overloaded (with perfect load-balancing).

Randomized Load Balancing

Q: There might be random fluctuations in the load or the load-balancing is not be perfect. How many servers does Fussbook need to ensure that their servers are not overloaded with high-probability?

Let m be the number of servers that Fussbook needs to use. Recall that a server is overloaded if the load it is assigned exceeds 600 units. Let us first look at server 1 and define T be the r.v. corresponding to the number of seconds of work assigned to the first server.

Let T_i be the number of seconds server 1 spends on processing post i. $T_i = 0$ if the task is assigned to a different (not the first server). The maximum amount of time spent on post i is 1-second. Hence, $T_i \in [0, 1]$.

Since there are $n := 24000$ posts in every 10-minute interval, the load (amount of units) assigned to the first server is equal to $T = \sum_{i=1}^{n} T_i$. Server 1 will be definitely overloaded if $T \geq 600$, and hence we want to upper-bound the probability $\Pr[T \geq 600]$.

Since the assignment of a post to a server is independent of the time required to process the post, the T_i r.v's are mutually independent. Hence, we can use the Chernoff bound.

Randomized Load Balancing

We first need to estimate $\mathbb{E}[T]$.

$$\mathbb{E}[T] = \mathbb{E}[\sum_{i=1}^{n} T_i] = \sum_{i=1}^{n} \mathbb{E}[T_i] \qquad \text{(Linearity of expectation)}$$

$$\begin{aligned} \mathbb{E}[T_i] &= \sum_{i=1}^{n} \mathbb{E}[T_i | \text{server 1 is assigned post } i] \Pr[\text{server 1 is assigned post } i] \\ &\quad + \mathbb{E}[T_i | \text{server 1 is not assigned post } i] \Pr[\text{server 1 is not assigned post } i] \\ &= \frac{1}{4}\frac{1}{m} + (0)(1 - 1/m) = \frac{1}{4m}. \end{aligned}$$

$$\implies \mathbb{E}[T] = \sum_{i=1}^{n} \frac{1}{4m} = \frac{n}{4m} = \frac{6000}{m}.$$

Randomized Load Balancing

Recall the Chernoff Bound: $\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\,\mathbb{E}[T])$. In our case, $c\mathbb{E}[T] = 600 \implies c = \frac{m}{10}$. Hence,

$$\Pr[T \geq 600] \leq \exp\left(-\beta\left(\frac{m}{10}\right)\frac{6000}{m}\right)$$

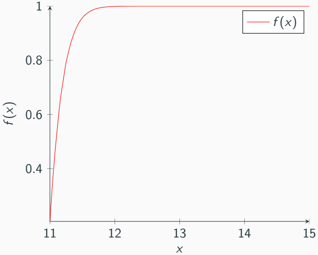
Hence, Pr[first server is overloaded] $= \Pr[T \geq 600] \leq \exp\left(-\beta\left(\frac{m}{10}\right)\frac{6000}{m}\right)$.

$$\begin{aligned}
&\Pr[\text{some server is overloaded}]\\
&= \Pr[\text{server 1 is overloaded} \cup \text{server 2 is overloaded} \cup \ldots \cup \text{server m is overloaded}]\\
&\leq \sum_{j=1}^{m} \Pr[\text{server j is overloaded}] && \text{(Union Bound)}\\
&= m\Pr[\text{server 1 is overloaded}] = m\exp\left(-\beta\left(\frac{m}{10}\right)\frac{6000}{m}\right)\\
& && \text{(Since all servers are equivalent)}
\end{aligned}$$

$$\implies \Pr[\text{no server is overloaded}] \geq 1 - m\exp\left(-\beta\left(\frac{m}{10}\right)\frac{6000}{m}\right).$$

Randomized Load Balancing

Plotting Pr[no server is overloaded] as a function of m.



Hence, as $m \geq 12$, the probability that no server gets overloaded tends to 1 and hence none of the Fussbook servers crash!

Questions?

Two-sided Chernoff Bound

From the Chernoff bound, we know that $\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\,\mathbb{E}[T])$.

By symmetry, one can also prove the **two-sided Chernoff Bound** to bound the probability that the r.v. T takes values (much) smaller or larger than the mean $\mathbb{E}[T]$:

$$\Pr\left[|T - \mathbb{E}[T]| \geq c\mathbb{E}[T]\right] \leq 2\exp\left(\frac{-c^2\mathbb{E}[T]}{3}\right)$$

A/B Testing

Fussbook is redesigning their website to try to make it more appealing to users. In order to see if the new redesigned website actually helps, Fussbook decides to do an experiment.

Given m users of the website, let $\mathcal{A}$ is the set of users who engage with the Fussbook posts (liking, sharing, etc) if *they are shown the old website*. Similarly, $\mathcal{B}$ is the set of users that engage if *they are shown the new website*.

Define f_A (and f_B) to be the fraction of users that engage with the website when shown the old (or new website respectively), i.e. $f_A := \frac{|\mathcal{A}|}{m}$ and $f_B := \frac{|\mathcal{B}|}{m}$. These fractions correspond to the proportion of users that prefer the old website vs the new.

The improvement on switching to the new website is defined in terms of the **lift** which is equal to $f_B - f_A$. If $f_B - f_A > 0$, then it makes sense to switch to the new website.

A/B Testing

In order to estimate f_A and f_B and see if the new website actually helps, Fussbook decides to run an **A/B** test on m customers.

Algorithm A/B Testing

```
function ABTest($m$)
$X_A = 0$, $X_B = 0$ {Initialize number of users that engaged when shown the old/new website}
for i = 1,2, . . . , m do
    $X_i \sim \text{Ber}(1/2)$
    if $X_i = 1$ then
        Show user $i$ old website.   If user engages, $X_A = X_A + 1$
    else
        Show user $i$ new website.   If user engages, $X_B = X_B + 1$
    end if
end for
return  $\hat{f}_A = \frac{2X_A}{m}$ and $\hat{f}_B = \frac{2X_B}{m}$.
```

A/B Testing

Claim: $\hat{f}_A$ and $\hat{f}_B$ are unbiased estimators of f_A and f_B respectively, i.e $\mathbb{E}[\hat{f}_A] = f_A$ and $\mathbb{E}[\hat{f}_B] = f_B$. Note that $X_A = \sum_{i \in \mathcal{A}} X_i$.

$$\mathbb{E}[\hat{f}_A] = \mathbb{E}\left[\frac{2X_A}{m}\right] = \frac{2}{m}\mathbb{E}\left[\sum_{i \in \mathcal{A}} X_i\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{A}} \mathbb{E}[X_i]\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{A}} \Pr[X_i = 1]\right] = \frac{2}{m}\frac{|\mathcal{A}|}{2} = f_A.$$

Similarly, $X_B = \sum_{i \in \mathcal{B}} (1 - X_i)$.

$$\mathbb{E}[\hat{f}_B] = \mathbb{E}\left[\frac{2X_B}{m}\right] = \frac{2}{m}\mathbb{E}\left[\sum_{i \in \mathcal{B}} (1 - X_i)\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{B}} \mathbb{E}[1 - X_i]\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{B}} \Pr[X_i = 0]\right] = \frac{2}{m}\frac{|\mathcal{B}|}{2} = f_B.$$

A/B Testing

Claim: With probability 0.95, the algorithm ABTest(m) estimates the lift to an error equal to 0.1 if $m \approx 10517$.

$$\begin{aligned}
\Pr\left[|\hat{f}_A - f_A| \geq \epsilon\right] &= \Pr\left[\left|\frac{2X_A}{m} - f_A\right| \geq \epsilon\right] = \Pr\left[\left|X_A - \frac{m\,f_A}{2}\right| \geq \frac{m\epsilon}{2}\right] \\
&= \Pr\left[\left|X_A - \frac{m\,f_A}{2}\right| \geq \underbrace{\frac{\epsilon}{f_A}}_{c}\ \frac{mf_A}{2}\right] \\
&\leq 2\exp\left(-\frac{1}{3}\left(\frac{\epsilon}{f_A}\right)^2 \frac{mf_A}{2}\right) && \text{(Two-sided Chernoff Bound)} \\
&= 2\exp\left(-\frac{\epsilon^2}{6}\,\frac{m}{f_A}\right) \leq 2\exp\left(-\frac{\epsilon^2\,m}{6}\right)
\end{aligned}$$

(Since $f_A < 1$ and exp is a monotonically increasing function.)

Similarly, we can prove that, $\Pr\left[|\hat{f}_B - f_B| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{3}\ \frac{1}{m}\right)$.

A/B Testing

By the union-bound for two events A and B, $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$. Hence,

$$\Pr\left[|\hat{f}_A - f_A| \geq \epsilon \cup |\hat{f}_B - f_B| \geq \epsilon\right] \leq 4\exp\left(-\frac{\epsilon^2\, m}{6}\right)$$

$$\implies \Pr\left[|\hat{f}_A - f_A| < \epsilon \cap |\hat{f}_B - f_B| < \epsilon\right] \geq 1 - 4\exp\left(-\frac{\epsilon^2\, m}{6}\right)$$

$$\begin{aligned}|\text{estimated lift} - \text{true lift}| &= |[\hat{f}_B - \hat{f}_A] - [f_B - f_A]| = |[\hat{f}_B - f_B] + [\hat{f}_A - f_A]| \\ &\leq |[\hat{f}_B - f_B]| + |[\hat{f}_A - f_A]| \\ &\text{(Triangle Inequality: For any constants } a,\ b,\ |a+b| < |a| + |b|\text{)}\end{aligned}$$

$$\implies \Pr[|\text{estimated lift} - \text{true lift}| \leq 2\epsilon] \geq 1 - 4\exp\left(-\frac{\epsilon^2\, m}{6}\right).$$

Proof of Triangle inequality: For any a, b, $ab \leq |a||b|$.

$\implies a^2 + b^2 + 2ab \leq a^2 + b^2 + 2|a||b| = |a|^2 + |b|^2 + 2|a||b|$ (since $\forall x, x^2 = |x|^2$).

$\implies (a+b)^2 = (\,|(a+b)|\,)^2 \leq (|a| + |b|)^2 \implies |a+b| \leq |a| + |b|$.

A/B Testing

In the claim, we want the RHS to be equal to 0.95 and $2\epsilon = 0.01$. Hence,

$$4\exp\left(-\frac{\epsilon^2 m}{6}\right) = 0.05 \implies m = \frac{6\ln(80)}{\epsilon^2} = \frac{6\ln(80)}{(0.05)^2} \approx 10517.$$

Hence, Fussbook can estimate the lift upto an error of 0.1 accurately with probability 0.95 by running the A/B test on $m \approx 10517$ users.

Questions?