CMPT 210: Probability and Computing

Lecture 24

Sharan Vaswani

December 3, 2024

Sums of Random Variables

We have seen that:

• 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 random variable of interest is a sum of r.v's (e.g., for the voter poll), and we can use the Chernoff bound to 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 $\beta(c) := c\ln(c) - c + 1$,

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

Example: 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 indicator 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? (Assume that the players' picks and the winning number are random, independent and uniform)

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. Using the independence assumptions, we can conclude that 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,

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

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:

$$\Pr[T - \mathbb{E}[T] \geq x] \leq \Pr[|T - \mathbb{E}[T]| \geq x] \leq \frac{\mathrm{Var}[T]}{x^2}$$

$$\implies \Pr[T - \mathbb{E}[T] \geq (c-1)\mathbb{E}[T]] \leq \frac{\mathrm{Var}[T]}{(c-1)^2\,(\mathbb{E}[T])^2)} \qquad (x = (c-1)\,\mathbb{E}[T])$$

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

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, $p_i = 1/2$, $\Pr[T \geq c\mathbb{E}[T]] \leq \exp\left(-\frac{n\,(c\ln(c)-c+1)}{2}\right)$.

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 could be 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 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 may be overloaded if the load it is assigned exceeds 600 units. Let us first look at server 1 and define T to be the r.v. corresponding to the number of units 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 may be overloaded if $T > 600$, and hence we want to upper-bound the probability $\Pr[T > 600]$.

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 $\mathbb{E}[T]$.

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

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

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

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,

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

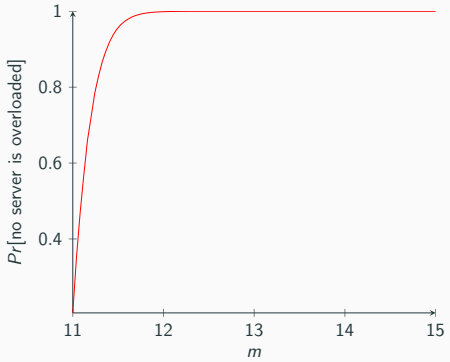
Hence, $\Pr[\text{first server is overloaded}] \leq \Pr[T \geq 600] \leq \exp\left(-\beta\left(\frac{m}{10}\right)\frac{6000}{m}\right)$.

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

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

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?

What's Next

- We have studied random variables that can take on discrete values.
- We have used these discrete distributions for designing randomized algorithms for verifying matrix multiplication, maximum cut in graphs, randomized Quick Select and voter poll.
- In many applications, it is often more natural to model quantities as continuous random variables, for example, the amount of time it takes to transmit a message over a noisy channel or study the distribution of income in a population.
- Continuous random variables are often used in distributed computing and for machine learning – fitting a model that can effectively explain the collected data.

What's Next

STAT 271: Probability and Statistics for Computing Science

- Continuous random variables and distributions
- Sampling and Parameter estimation
- Linear Regression
- Hypothesis testing
- Analysis of Variance