EE 178/278A Probabilistic Systems Analysis

Spring 2014 Tse/Hussami Lecture 9

BitTorrent Servers' Example

A video is broken down into m chunks. Each server has a random chunk, i.e. one out of m possible choices. We are interested in the number of servers we need to query to have the whole movie (meaning all m chunks). Let X be the number of servers we query before we get the m chunks. How do we analyze this problem? How do we compute $\mathbb{E}[X]$? Of course we will need to query at least m servers. But because some servers will store the same chunk, we may need more. The question is, how many more on the average?

There is a natural way of breaking this random variable into a sum of simpler random variables. We make progress whenever we get a new chunk. We write $X = X_1 + \cdots + X_m$ in order to use linearity of expectation. Let X_1 be the time to get the first new chunk. The first server we query will give a new unobserved chunk no matter what. Therefore $X_1 = 1$, and X_1 is a deterministic random variable. X_2 is the number of extra servers you query until you see the second new chunk, and similarly X_m is the number of servers you query until you get the last chunk. The chunks are not ordered here, it does not matter which one comes first as long as we have all m of them at the end.

Let us now determine the distribution of X_2 to compute its expectation. $X_2 \sim Geom(p)$ where p is the success probability. The probability of success is the probability of getting a new chunk. When you sample the next server, you will find a new chunk with a high probability at the beginning: $\frac{m-1}{m}$. The expected value of a geometric random variable is $\frac{1}{p}$. Therefore, $\mathbb{E}[X_2] = \frac{m}{m-1}$ which is slightly greater than 1. Similarly, $X_3 \sim Geom(\frac{m-2}{m})$. In general, $X_i \sim Geom(\frac{m-i+1}{m})$ and $\mathbb{E}[X_i] = \frac{m}{m-i+1}$. In particular, $X_m \sim Geom(\frac{1}{m})$.

Finally, we can compute the desired expectation

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_m] = \sum_{i=1}^{m} \frac{m}{m-i+1} = m[\frac{1}{m} + \frac{1}{m-1} + \frac{1}{m-2} + \cdots + \frac{1}{1}].$$

If m is very large, the number in brackets is roughly equal to $\ln m$ (which can be seen by looking at the area under the curve $\frac{1}{x}$). This means we have to query a factor of $\ln m$ more servers than the case where we know exactly where the chunks are (in which case we need only to query m servers.) This is the price of randomness in the protocol.

This is an important application of the geometric distribution. In general , once you figure out the relevant distribution of a random variable in a problem, you just need to determine the parameter of that distribution.

The Poisson Distribution

The Poisson distribution is a very widely accepted model for so-called "rare events", such as misconnected phone calls, radioactive emissions, crossovers in chromosomes, etc. Here, we motivate this distribution with the following example. Suppose Verizon wants to deploy a wireless network in a city. At any time, the number of people making calls is random and cannot be predicted ahead of time. There are many paying customers in the network, and all of them can potentially make a call during the same period of time. However, only a very small fraction of them actually will. Verizon needs to size the capacity of its network such that it can support a large number of users at any given point. Suppose an average call lasts for one minute. If we want to know how many people are using the same network at the same time, we need the number of people that initiate a call within the same minute. Focusing on a window of one minute, we are interested in the random variable which counts the number of calls initiated within that minute. What is a reasonable distribution to model this random variable?

Let X be the random variable that counts the number of calls initiated within a minute. To model the random variable X we need to make some reasonable assumptions. First, we need to collect some statistics. A useful quantity to measure is the rate at which calls are coming in. We count the number of calls generated in i minutes, and divide it by the number of minutes i to get a rate of call arrivals per minute. Let λ be the rate of call arrivals (with unit the number of calls per minute). Then $\lambda = \mathbb{E}[X]$, the expected number of arrivals in a minute. We would like to figure out the rest of the distribution.

Let us focus on that one minute interval and divide it into n very small subintervals. The first assumption we make on these subintervals is that the chance that we have more than one initiation within a very small subinterval is negligible. People are making the calls separately, and the chance of having a collision is small. More specifically, the first assumption we make is:

> 1- **The probability of having more than one arrival in a very small interval is negligible. Two events can happen in a small interval: either no initiations are made, or exactly one.**

Let X_i be the number of call initiations in the ith interval. Then $X = X_1 + \cdots + X_n$, where

$$X_i = \begin{cases} 0 & \text{if no calls are initiated in the } i\text{th sub-interval,} \\ 1 & \text{otherwise.} \end{cases}$$

What should the probability of $X_i = 1$ be? We need to express it in terms of λ and n. This probability must be $\frac{\lambda}{n}$ because $\mathbb{E}[X] = \lambda = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$. (Here we are assuming that each of the X_i's have the same distribution.) We are still missing the understanding of the relationship between the X_i variables. A reasonable model is that the event that a call is initiated in interval i is independent from the event that a call is initiated in interval j. Therefore, we make the following second assumption:

> 2- **The events of call initiation in different sub-intervals are mutually independent.**

The random variable X counts the number of call initiations in the 1-minute interval. Under the two assumptions, X becomes a binomial random variable: $X \sim Bin(n, \frac{\lambda}{n})$. The parameters of this binomial are related together by n. We now have that

$$\mathbf{P}(X = i) = \binom{n}{i} \left(\frac{\lambda}{n}\right)^i (1 - \frac{\lambda}{n})^{n-i}, \qquad i = 0, 1, \ldots, n.$$

Taking the limit as n goes to infinity, we get

$$\mathbf{P}(X=0) = (1-\frac{\lambda}{n})^n \to e^{-\lambda} \text{ as } n \to \infty$$

(using the fact that $\lim_{n\to\infty}(1+\frac{x}{n})^n = e^x$).

In general, $\mathbf{P}(X=i) \to \frac{\lambda^i e^{-\lambda}}{i!}$ as $n \to \infty$, giving the Poisson distribution of the random variable X. By choosing a large n, it is thus reasonable to model the number of call initiations during a one-minute period to be Poisson with parameter λ.

Definition 9.1 (Poisson distribution): A random variable X for which

$$\mathbf{P}(X=i) = \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda} \qquad \text{for } i=0,1,2,\ldots \tag{1}$$

is said to have the *Poisson distribution with parameter* λ. This is abbreviated as $X \sim \text{Poiss}(\lambda)$.

Chebyshev's inequality

There are two important numbers to calculate to describe a distribution: the mean and the variance. We have seen that, intuitively, the variance (or, more correctly the standard deviation) is a measure of "spread", or deviation from the mean. For instance if the variance is small, then the distribution of the variable is not very spread out meaning that the chance of getting an outcome that is far from the mean is small. Our next goal is to make this intuition quantitatively precise. What we can show is the following:

Theorem 9.1: **[Chebyshev's Inequality]** *For a random variable X with expectation $\mathbb{E}[X] = \mu$, and for any $a > 0$,*

$$\mathbf{P}(|X-\mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

We are interested in the quantity: $\mathbf{P}(|X-\mu| \geq a)$: what is the chance that the random variable is greater than $\mu + a$ or smaller than $\mu - a$? The claim is that if the variance of the random variable is small, then this probability is small. Chebyshev's inequality allows us to upper bound this probability by $\frac{\text{Var}(X)}{a^2}$. If we know the distribution, we can calculate this number exactly. However, we might not have the whole distribution, or it might be too complicated. Even without knowing the whole distribution, Chebyshev tells you that you can bound the probability of this event.

Before proving the inequality, let's pause to consider what it says. The upper bound decreases as a becomes larger. This means that the chance of deviating more and more from the mean gets smaller which seems reasonable. Moreover, the smaller the variance of the random variable, the smaller the probability of being far away from the mean which implies a smaller spread as we wanted to argue.

We now prove Chebyshev's inequality. We will show the equivalent statement that $\text{Var}(X) \geq a^2\mathbf{P}(|X-\mu| \geq a)$: if the probability at the two tails is large, then the variance has to be large as well. For a given tail probability, the smallest variance one can have for X is when we shift all the probability mass at values $x \geq \mu + a$ or $x \leq \mu - a$ to the boundaries at $\mu + a$ and $\mu - a$. Let then Y be a new random variable with the same distribution as X in the interval $(\mu - a, \mu + a)$, and all the mass at $x \geq \mu + a$ and $x \leq \mu + a$ concentrated at the points $\mu + a$ and $\mu - a$ respectively. The variance of X is greater than the variance of Y. Therefore,

$$\text{Var}(X) \geq \text{Var}(Y) \geq a^2\mathbf{P}(|X-\mu| \geq a)$$

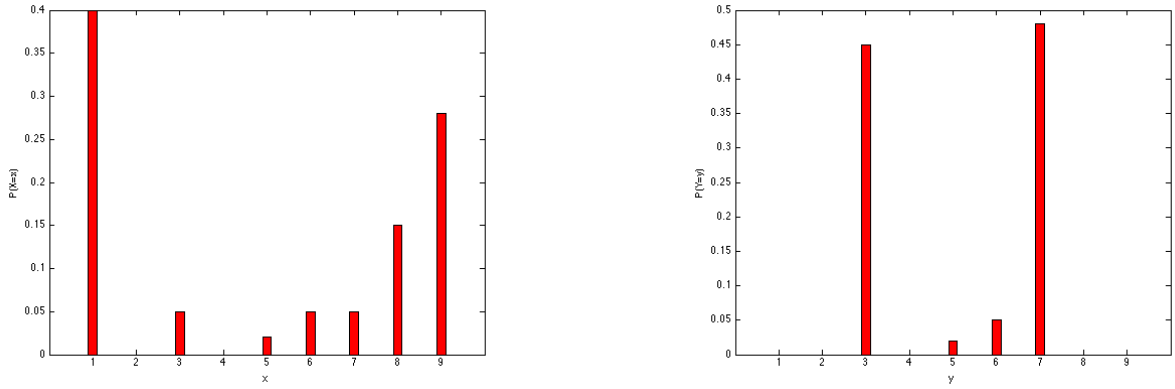

Figure 1: *Example of a distribution of variables X and Y with mean 5, where the mass in the tails of X is shifted to the values 3 and 7 to form the distribution of Y.*

where the second inequality follows from the definition of variance: the variance is a sum over all the masses, with each mass multiplied by the square of its distance to the mean. For Y, each of the two masses at $\mu - a$ and $\mu + a$ has a distance a from the mean, so the contribution of the two masses is $a^2\mathbf{P}(|X - \mu| \geq a)$, and the variance of Y is at least as much.

Application: Polling

Let us now show an application of Chebyshev's inequality. Suppose we want to figure out the preference of a population. For example, we want to know the fraction p of the number of democrats in California. How many people should we ask before we get a reliable answer? Let X_i model the answer of a person, where $X_i = 1$ if the person is a democrat and 0 otherwise. Then $X_i = 1$ with probability p.

We need to estimate the parameter p in the model from observing the outcome of the random experiment, which is a standard problem in statistics. An estimator takes the data collected from polling and outputs an estimate $\hat{p}$ of the parameter p. The data in this problem is the n responses $X_1, \ldots, X_n$ of the people. We estimate $\hat{p}$ from the n observations, which will be a function of the data. A reasonable estimate here is $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the fraction of the people polled who say they are democrats. How reliable is this estimate? We will explore this question in the next lecture and see how Chebychev's inequality will help us to answer this question.