COMP 480/580 — Probabilistic Algorithms and Data Structure Jan 23, 2020

Lecture 4

Lecturer: Anshumali Shrivastava *Scribe By: Alan Ji*

1 Introduction

If we recall from two classes ago, we talked about chaining and linear probing, where the expected length of any chain was $1 + (m-1)/n$. What does this expectation mean? It gives us an 'average' of the random variable, but it doesn't give enough information to tell whether or not how 'bad' things can go beyond this expectation.

To recall, a **random variable** is a function from the sample space of an experiment/process to the set of real numbers. The **expected value** (also called the expectation or mean) of a (discrete) random variable X on the sample space S is

$$\mathbb{E}(X) = \sum_{s \in S} P(s) \cdot X(s) = \sum_{x} x \cdot P(X = x)$$

Linearity of Expectations states that if X_i, $i = 1, 2, ..., n$ are random variables on S, and if a and b are real numbers, then

$$\mathbb{E}(X_1 + X_2 + ... + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + ... + \mathbb{E}(X_n)$$
$$\mathbb{E}(aX_i + b) = a\mathbb{E}(X_i) + b$$

If X is a random variable on a sample space S, then the **variance** of X, denoted by $V(X)$, is

$$V(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Bienayme's Formula (shown below) is similar to the 'Linearity of Expectations', but for variance, provided that we have an extra assumption: that each random variable $X_i, i = 1, 2, ..., n$ is pairwise independent on S. Otherwise, covariance terms between the random variables will be non-zero.

$$V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i)$$

Furthermore, for constant a,

$$V(aX) = a^2 V(X)$$

2 Markov's Inequality

Let X be a random variable that takes only nonnegative values. Then, for every real number $a > 0$ we have

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

2.1 Example

Before we go into the proof, let's give an example from the chaining discussion 2 weeks ago. Suppose, for a nonnegative random variable X, $\mathbb{E}(X) \leq 2$ (if $n > m$). Then we can ask, what is the probability that $P(X \geq 10)$? Using Markov's Inequality, we can conclude that it is

$\leq \frac{2}{10} = \frac{1}{5}$. Great, then we know the system won't fail to this certain extent.

2.2 Proof

$$\begin{aligned}
\mathbb{E}(X) &= \sum_x x \cdot Pr(X = x) \\
&= (\sum_{x \geq a} x \cdot Pr(X = x)) + (\sum_{x < a} x \cdot Pr(X = x)) \\
&\geq (\sum_{x \geq a} a \cdot Pr(X = x)) + 0 \\
&= a \sum_{x \geq a} Pr(X = x) \\
&= a(Pr(X \geq a)) \\
&\therefore P(X \geq a) \leq \tfrac{\mathbb{E}(X)}{a}
\end{aligned}$$

2.3 Discussion

Unfortunately however, in general and for distributions encountered in practice, Markov's inequality gives us a very loose bound in which many times is not useful or interesting.

Typically, we want to understand how much our random variable deviates from our expected behavior. Call this deviation, δ, so that we get the following by Markov's:

$$Pr(X \geq \mathbb{E}(X) + \delta) \leq \tfrac{\mathbb{E}(X)}{\mathbb{E}(X)+\delta}$$

Now, if delta is small, this upper bound is generally equal to 1, which tells us nothing useful. If we wanted this probability to be 0.5, then we must ask whether we have deviated by twice its expected value. Therefore, we need more information to get these tighter bounds.

3 Chebyshev's Inequality

If we only know about a random variable's expected value, then Markov's upper bound is the only probability we can get. However, if we know the variance, then the tighter Chebyshev's can be achieved. For a random variable X, and every real number $a > 0$,

$$P(|X - \mathbb{E}(X)| \geq a) \leq \tfrac{V(X)}{a^2}$$

3.1 Proof

From Markov's we get

$$P(|X - \mathbb{E}(X)| \geq a) = P((X - \mathbb{E}(X))^2 \geq a^2)$$

We can redefine another random variable to be $(X - \mathbb{E}(X))^2$ (which is always nonnegative) and use Markov's again:

$$P((X - \mathbb{E}(X))^2 \geq a^2) \leq \tfrac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{a^2}$$

The numerator in this probability is the definition of variance, $V(X)$.

$$\therefore P(|X - \mathbb{E}(X)| \geq a) \leq \tfrac{V(X)}{a^2}$$

3.2 Discussion

To compare Markov's and Chebyshev's, we can see that Markov decays by $1/a$ while Chebyshev decays by $1/a^2$. In other terms, both inequalities can be described, respectively, below, where μ is the mean or expected value of the random variable, σ is its standard deviation, and k is a positive constant.

$$P(X \geq k\mu) \leq \frac{1}{k}$$
$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

3.3 Example

Assume we have a distribution whose mean is 80 and standard deviation is 10. What is a lower bound on the percentage of values that fall between 60 and 100 (exclusively) in this distribution?
We know the following: $\mathbb{E}(X) = 80$, $V(X) = 100$, and $a = 20$. Therefore, we can write our inequality as:

$$P(|X(s) - 80| \geq 20) \leq \frac{100}{20^2}$$

The lower bound in this case is 75%.

3.4 Corollary

The corollary of Chebyshev's Inequality states that for $X_1, X_2, ..., X_n$ independent random variables with $\mathbb{E}(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, then for any $a > 0$,

$$P(|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i| \geq a) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{a^2}$$

This is derived directly from Chebyshev's Inequality, utilizing both Linearity of Expectations and Bienayme's Formula.

3.5 Weak Law of Large Numbers

Following the corollary, we can show the property of the Weak Law of Large Numbers. Suppose $X_1, X_2, ..., X_n$ are i.i.d random variables, where the unknown expected value μ is the same for all variables, and their variance is finite. Therefore, for any $\epsilon > 0$, we have

$$P(|(\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu| \geq \epsilon) \xrightarrow{n \to \infty} 0$$

4 Estimating π Using the Monte Carlo Method

To estimate π, suppose we have a square of area (and length) 1 that encloses a circle of radius 0.5 and area $\pi/4$. If we were to throw darts at this square, we can essentially estimate the value of π by counting the proportion of times that the darts landed inside the circle, and then multiply by 4.

To verify this method more rigorously, let X_i be the random variable that denotes whether the i^{th} dart landed inside the circle (indicator variable of 1 if it did, and 0 otherwise).
Then, $\hat{\pi}(n) = 4 \cdot \frac{\sum_{i=1}^{n} X_i}{n}$

Algorithm 1: MonteCarlo π Estimation

```
Input: $n \in \mathbb{N}$
Output: Estimate $\hat{\pi}$ of $\pi$
for $i = 1...n$ do
    $a \leftarrow random(0,1)$;
    $b \leftarrow random(0,1)$;
    $X_i \leftarrow 0$;
    if $\sqrt{(a-0.5)^2 + (b-0.5)^2} \leq 0.5$ then
        $X_i \leftarrow 1$;
$\hat{\pi} \leftarrow 4 \cdot (\sum_{i=1}^{n} X_i)/n$;
return $\hat{\pi}$;
```

We know that the probability of landing inside the circle is $p = \pi/4$. This represents the probability associated with each of our random variables X_i. Therefore, using the definitions of expectation and variance,

$$\mathbb{E}(X_i) = \pi/4 \cdot 1 + (1-\pi/4) \cdot 0 = \pi/4$$
$$V(X_i) = \pi/4 \cdot (1-\pi/4)$$

To confirm that our estimator for π ($\hat{\pi}$) is unbiased,

$$\mathbb{E}(\hat{\pi}) = \mathbb{E}(\tfrac{4}{n}\sum_{i=1}^{n} X_i) = \frac{4}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \pi$$

by Linearity of Expectation.
Similarly,

$$V(\hat{\pi}) = V(\tfrac{4}{n}\sum_{i=1}^{n} X_i) = \frac{16}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\pi(4-\pi)}{n}$$

4.1 Example

Suppose that we wanted to find the value of n so that the estimation error of π is within δ with probability at least ϵ. This comes straight from Chebyshev's Inequality:

$$P(|\hat{\pi}(n) - \pi| < \delta) > \epsilon$$

or equivalently,

$$P(|\hat{\pi}(n) - \pi| \geq \delta) \leq 1 - \epsilon$$

For the purposes of this example, we want the error (δ) to be small (0.001) and success probability (ϵ) to be ≥ 0.95. How many times would we need to repeat this experiment (n)?

We can plug these into Chebyshev's:

$$P(|\hat{\pi}(n) - \pi| \geq 0.001) \leq 1 - 0.95$$

We know our calculations before that $V(\hat{\pi}) = \frac{\pi(4-\pi)}{n}$. And furthermore, from Chebyshev's, that $V(\hat{\pi})/a^2 \leq 0.05$.

We can plug in a and $V(\hat{\pi})$ to get

$$V(\hat{\pi})/a^2 = \tfrac{\pi(4-\pi)}{n(0.001)^2} \leq 0.05$$

and solve for n, which turns out to be $n \geq 80,000,000$

5 Chernoff Bounds

To obtain an even tighter bound than Markov's and Chebyshev's, we need an additional level of independence. For example, if $X_1 \cdot X_2 = X_3$ (all three are binary, pairwise independent variables), then if we know any two of them, we can strictly determine the third. Again, these additions of correlation helps us form the tighter bounds for Chernoff Bounds.

There are multiple forms of Chernoff Bounds, which gets somewhat confusing in literature. Generally, let $X = X_1 + X_2 + ... + X_n$ where all X_i's are independent and $Bernoulli(p_i)$. Furthermore, let $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$.
Then, for $\delta > 0$,

$$P(|X - \mu| \geq \delta\mu) \leq 2e^{-\frac{\delta^2 \mu}{2+\delta}}$$

The bounds can also be written as

$$P(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2+\delta}} \text{ for } \delta > 0$$
$$P(X \geq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}} \text{ for } 1 > \delta > 0$$

5.1 Proof

We can utilize the following first two lemmas to prove the third, and then using the third for our overall proof of Chernoff Bounds.

Lemma 1: Given random variable $Bernoulli(p)$ random variable Y, we have for all $s \in \mathbb{R}$

$$\mathbb{E}(e^{sY}) \leq e^{p(e^s - 1)}$$

Lemma 2: Let $X_1, ..., X_n$ be independent random variables, and $X = \sum_{i=1}^{n} X_i$. Then, for $s \in \mathbb{R}$

$$\mathbb{E}(e^{sX}) = \prod_{i=1}^{n} \mathbb{E}(e^{sX_i})$$

Lemma 3: Let $X_1, ..., X_n$ be independent random variables (Bernoulli distributed), and $X = \sum_{i=1}^{n} X_i$ and $\mathbb{E}(X) = \sum_{i=1}^{n} p_i = \mu$. Then, for $s \in \mathbb{R}$

$$\mathbb{E}(e^{sX}) \leq e^{(e^s - 1)\mu}$$

To begin, we can use Markov's:

$$P(X \geq a) = P(e^{sX} \geq e^{sa}) \leq \frac{\mathbb{E}(e^{sX})}{e^{sa}} = \frac{e^{(e^s-1)\mu}}{e^{sa}}$$

Afterwards, we set $a = (1+\delta)\mu$ and $s = ln(1+\delta)$, which minimizes the upper bound of the inequality, to get

$$P(X \geq (1+\delta)\mu) \leq \frac{e^{(e^{ln(1+\delta)}-1)\mu}}{e^{ln(1+\delta)(1+\delta)\mu}} = (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{\mu}$$

Taking the natural log of this resulting term simplifies into $-\frac{\delta^2 \mu}{2+\delta}$, which is equivalent to the upper bound Chernoff inequality.

5.2 Discussion

We can easily see that these bounds are exponentially decaying (incredibly valuable), compared to the linear and quadratic versions from Markov and Chebyshev.

Another form of the Chernoff Bounds, given each independent X_i follows $a \leq X_i \leq b \; \forall i$ for some constants a and b.

$$P(X \geq (1+\delta)\mu) \leq e^{-\frac{2\delta^2 \mu^2}{n(b-a)^2}}$$
$$P(X \geq (1-\delta)\mu) \leq e^{-\frac{2\delta^2 \mu^2}{n(b-a)^2}}$$

Note that these random variables don't have to be identically distributed, just as long as they can squeeze between the a and b bounds.

6 Comparison and Discussion

Let's give an example where a fair coin is tossed 200 times. How likely is it to observe at least 150 heads? Let's assume there's an indicator variable $X_i, i = 1...200$ for each of these tosses: 1 if heads and 0 if tails. We can define the sum of these Bernoulli random variables as $X = \sum_{i=1}^{100} X_i$.

With just Markov and Linearity of Expectation, we calculate this probability as

$$P(X \geq 150) \leq \frac{\mathbb{E}(X)}{150} = 0.6666$$

Even with extremely correlated random variables, such as the second to 200th tosses all following the very first initial toss, this probability still holds true ($0.5 < 0.6666$) and Markov is still correct.

If we were to assume pairwise independence between these tosses, then we can calculate the variance easily, giving us a better bounds with Chebyshev's and Bienayme's

$$P(|X - 100| \geq 50) \leq \frac{V(X)}{50^2} = 0.02$$

With even more independence assumptions, we can use Chernoff's to get the tightest bounds

$$P(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2\mu}{2+\delta}}$$

where $\mu = 100$ and $\delta = 0.5$

$$P(X \geq 150) \leq e^{-\frac{0.5^2\cdot 100}{2+0.5}}$$

References

http://math.mit.edu/ goemans/18310S15/chernoff-notes.pdf