COMP 480/580 — Probabilistic Algorithms and Data Structure Aug 30, 2022

Lecture 3: Markov's, Chebyshev's, and Chernoff Bounds

Lecturer: Dr. Ben Coleman *Scribe By: Yufei Li, Linfeng Lou, Ziyang "Zion" Yang*

1 Motivation

In this lecture, we are focusing on the topic of how far away a value that the random variable can be taken from its mean.

The reason we care about this is the probability about the worst-case of an algorithm is important issue. It is possible that an algorithm, on average, takes 2n steps but on certain inputs, it takes $500n^2$ steps. Therefore, there are three bounds which will be introduced to solve this question.

2 Review of Basic Probabilities

Let's start from the definition of random variables. Consider an *experiment* and its associated *sample space*. We associate *probability* to subsets of the sample space. The intuition is that some *outcomes* in the sample space are more likely to happen than other outcomes. A ***random variable*** is a function from the sample space to the set of real numbers.

Example: Consider an experiment tossing coins for two times. The sample space is HH, HT, TH, TT where H stands for "head" and T stands for "tail". We can define a random variable X that map the outcome to the number of heads that ever appears in the experiment. That is, $X(HH) = 2$, $X(HT) = 1$, $X(TH) = 1$, $X(TT) = 0$.

Next we can define the ***expectation*** of a discrete random variable X as the sum of X weighted by the probability of its possible values:

$$\mathbb{E}(x) = \sum_{s \in S} P(s) \cdot X(s) \tag{1}$$

It follows from the definition that the operation of taking expectation is ***linear***. That is, given random variables (on the same sample space) $\{X_1, X_2, ..., X_n\}$, we have

$$\mathbb{E}(X_1 + X_2 + ... + X_n) = \mathbb{E}(X_1) + \mathbb{E}(x_2) + ... + \mathbb{E}(X_n) \tag{2}$$

$$\mathbb{E}(aX_i + b) = a\mathbb{E}(X_i) + b \tag{3}$$

On the other hand, we can define the ***variance*** $V(x)$ of the given random variable X. The intuition is that we need a notion to say the "sparsity" of the valuation of X.

$$V(x) = \mathbb{E}((X - E(X))^2) \tag{4}$$

and equivalently

$$V(x) = \mathbb{E}(X^2) - (\mathbb{E}(x))^2 \tag{5}$$

Next, let's prove a lemma for the Bienayme's formula.

Lemma: For a random variable X, $Var(aX+b) = a^2 Var(X)$.
Proof:

$$\begin{aligned} Var(aX+b) &= \mathbb{E}([aX+b-\mathbb{E}(aX+b)]^2) \\ &= \mathbb{E}([a(X-\mathbb{E}(x))]^2) \\ &= a^2\mathbb{E}((X-\mathbb{E}(x))^2) \end{aligned} \tag{6}$$

Finally we have the ***Bienayme's formula***. That is, given a set of pairwise independent random variables $\{X_1, X_2, ..., X_n\}$,

$$V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) \tag{7}$$

Proof:
Let $X' = X - \mathbb{E}(X)$, $Y' = Y - \mathbb{E}(Y)$. It follows from the definition that X' and Y' are independent, and $\mathbb{E}(X') = 0$, $\mathbb{E}(Y') = 0$. It follows from the lemma above that $Var(X) = Var(X')$, $Var(Y) = Var(Y')$.

$$\begin{aligned} Var(X+Y) &= Var(X'+Y') \\ &= \mathbb{E}((X'+Y')^2) + (\mathbb{E}(X'+Y'))^2 \\ &= \mathbb{E}(X'^2 + 2X'Y' + Y'^2) + 0 \\ &= \mathbb{E}(X'^2) + 2E(X'Y') + \mathbb{E}(Y'^2) \\ &= Var(X') + 2\mathbb{E}(X'Y') + Var(Y') \\ &= Var(X) + 2\mathbb{E}(X'Y') + Var(Y) \end{aligned} \tag{8}$$

Since X' and Y' are independent, $\mathbb{E}(X'Y') = \mathbb{E}(X')\mathbb{E}(Y') = 0$. Therefore $Var(X+Y) = Var(X) + Var(Y)$. QED.

3 Markov's Inequality

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

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{9}$$

3.1 Proof

By definition, we know $\mathbb{E}(\mathrm{X}) = \sum_x \mathrm{x} * \mathrm{P}(\mathrm{X} = \mathrm{x})$. Then, the equation can be divided into two part as follows.

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x * P(X = x) \\ &= \sum_{x \geq a} x * P(X = x) + \sum_{x < a} x * P(X = x) \\ &\geq \sum_{x \geq a} a * P(X = x) + 0 \\ &= a \sum_{x \geq a} P(X = x) \\ &= aP(X \geq a)\end{aligned} \tag{10}$$

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

3.2 Example

Assume the expected time it takes Algorithm A to traverse a graph with n nodes is 2n. What is the probability that the algorithm takes more than 10 times that?

By using Markov's Inequality, the solution is

$$P(X \geq 10 * E(X)) \leq \frac{\mathbb{E}(X)}{10 * \mathbb{E}(X)} = \frac{1}{10}$$

The probability that the algorithm takes more than 10 times is not more than 10%.

3.3 Summary

Markov's inequality can give us the result of how far away a random value can be. But for distributions encountered in practice, Markov's inequality gives a very loose bound. It is because Markov's inequality only considers the expectation of the algorithm, but does not consider the variance of it.

4 Chebyshev's Inequality

Let X be a random variable. For every real number r > 0,

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{V(X)}{a^2} \tag{11}$$

4.1 Proof

Since we know that $\mathbb{E}((X - \mathbb{E}(X))^2) = V(X)$, we can proof Chebyshev's inequality by using Markov's inequality:

$$\begin{aligned}P((X - \mathbb{E}(X))^2 \geq a^2) &\leq \frac{E((X - \mathbb{E}(X))^2)}{a} \\ &= \frac{V(X)}{a^2}\end{aligned} \tag{12}$$

Since $(X - \mathbb{E}(X))^2 \geq a^2$ equals to $|X - \mathbb{E}(X)| \geq a$, we can get the result:

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{V(X)}{a^2} \tag{13}$$

Corollary: Let $\sigma = V(X)$, $a = k\sigma$. We have:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{14}$$

4.2 Comparison with Markov's Inequality

Markov's inequality: $P(X \geq k\mu) \leq 1/k$
Chebyshev's inequality: $P(|X - \mu| \geq k\sigma) \leq 1/k^2$

We can know Chebyshev's inequality provides a tighter bound as k increases since Chebyshev's inequality scales quadratically with k, while Markov's inequality scales linearly with k.

4.3 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?

By using Chebyshev's inequality, the solution is

$$P(|X(s) - 80| \geq 20) \leq \frac{100}{20^2} = \frac{1}{4}$$

The lower bound on the percentage of values is not more than 25%.

4.4 Illustration: Estimating π Using the Monte Carlo Method

In order to estimate the value of π, we have a square whose area is known (1 in this case), and inside which there's a circle whose radius is 1/2 with area equals $\pi/4$. So the only thing we don't know here is π.

If we throw darts at the square, the darts will land randomly inside the square, the probability that it lands inside the circle equals the ratio of the circle area to the square area ($\pi/4$). Therefore, to estimate π, we can calculate the darts landed in the circle, divide it by the number of darts we throw, and multiply it by 4, that should be the expectation of π.

However, the performance of this result depends on how many darts we throw, intuition, if infinite darts are thrown, the estimation of π will converge to its actual value.

Let X_i be the random variable that denotes whether the i-th dart landed inside the circle (X_i is an indicator variable, 1 if inside the circle, and 0 otherwise).

The expectation of X :

$$\mathbb{E}(X_i) = \frac{\pi}{4} \times 1 + (1 - \frac{\pi}{4}) \times 0 = \frac{\pi}{4}$$

$$\mathbb{V}(X_i) = \frac{\pi}{4} \times (1 - \frac{\pi}{4})$$

Then:

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

Algorithm 1: Monte Carlo π Estimation

```
Input: $n \in \mathbb{N}$
Output: Estimate $\hat{\pi}$ of $\pi$
for $i = 1...n$ do
    $a \leftarrow random(0,1)$; // random number in [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
        $X_i \leftarrow 1$ ; // the dart landed inside/on the circle
    end
end
$\hat{\pi} \leftarrow 4 \times (\sum_{i=1}^{n} X_i)/n$
return $\hat{\pi}$
```

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

This confirms that our estimation of π is unbiased.

If we have bunch of pairwise i.i.d random variables(X_i), the variances are going to decay with the increasing of n.

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

Question: How big should n be for us to get a good estimate (with a given very small percentage of error)?

We want to find the value of n so that the estimation error of π is within δ with probability of at least ϵ. (Of course, we want δ to be very small and ϵ to be as close to 1 as possible, we are bounding the tails.)

We can use Chebyshev's Inequality to bound the deviation:

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

equivalently,

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

For $\delta = 0.001$ and $\epsilon = 0.95$, we seek n such that:

$$p(|\hat{\pi}(n) - \pi| \geq 0.001) \leq 0.05$$

Combined with Chebyshev's Inequality and our former calculations, $\mathbb{E}(\hat{\pi}) = \pi$, $\mathbb{V}(\hat{\pi}) = \frac{\pi(4-\pi)}{n}$, $a = 0.001$, $\mathbb{V}/a^2 = 0.05$.

So, we would like n such that:

$$\frac{\pi(4-\pi)}{n(0.001)^2} \leq 0.05$$

We know that $\pi(4-\pi) \leq 4$, therefore we can add a condition in the inequality:

$$\frac{\pi(4-\pi)}{n(0.001)^2} \leq \frac{4}{n(0.001)^2} \leq 0.05$$

Solving it, turns out that $n \geq 80,000,000$

4.5 A Corollary of Chebyshev's Inequality

Let $X_1, X_2, ..., X_n$ be **independent** random variables with $\mathbb{E}(X_i) = \mu_i$ and $\mathbb{V}(X_i) = \sigma_i^2$, Then, for any $a > 0$:

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

This can be derived from the Linearity of Expectations and Variances.

4.6 The Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be independently and identically distributed (i.i.d.) random variables, where the (unknown) expected value μ is the same for all variables (that is, $\mathbb{E}(X_i) = \mu$) and their variance is finite. Then, for any $\epsilon > 0$, we have:

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

Proof:

Consider $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$. It obvious that $\mathbb{E}(Y) = \mu$ and $Var(Y) = n\frac{1}{n^2}Var(X) = \frac{1}{n}Var(X)$. By Chebyshev's inequality, we have

$$P(|Y - \mathbb{E}(Y)| \geq \epsilon) \leq \frac{Var(Y)}{\epsilon^2} \tag{15}$$

and it follows that

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq \frac{\frac{1}{n}Var(X)}{\epsilon^2} = \frac{Var(X)}{n\epsilon^2} \tag{16}$$

It's clear from (16) that given any ϵ, as n approaches infinity, the right-hand-side approaches 0.

Note that The only assumptions taken here are finite expectations and finite variances, it doesn't show how fast the result converges to 0. Intuitively, the more sample we have, the better estimation of mean we can get. It only converges in probability.

5 Chernoff Bounds

Chernoff Bound answers the question about how tight the bound we can get when having more information about the distribution of the random variables. (It is probably the strongest bound can get (much stronger than Chebyshev), but might be less useful.) It also bound the deviation from the mean.

Let $X = X_1 + X_2 + ... + X_n$, where all the X_i's are independent and $X_i \sim$ *Bernoulli* (p_i), $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$

We got an exponential term. (These are the setup for this specific Chernoff Bound.) Then, for $\delta > 0$,

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

The bound can also be written as follows:

For $\delta > 0$, (this work for the upper tail):

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

For $0 < \delta < 1$, (this work for the lower tail):

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

5.1 Proof

Lemma 1: Given random variable $Y \sim \mathit{Bernoulli}(p)$, 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 $s \in \mathbb{R}$,

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

Proof:

To establish the result, use Markov's inequality:

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

Set $a = (1+\delta)\mu$ and $s = \ln(1+\delta)$,

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

Taking the natural logarithm of the right-hand side yields:

$$\mu(\delta(1+\delta)\ln(1+\delta))$$

Use the inequality $\ln(1+x) \geq \frac{2x}{2+x}$ for $x > 0$, we obtain:

$$\mu(\delta(1+\delta)\ln(1+\delta)) \leq -\frac{\delta^2}{2+\delta}\mu$$

Hence, we have the desired bound for upper tail:

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

5.2 Another Chernoff Bound

Let $X = X_1 + X_2 + ... + X_n$, where all the X_i's are independent and $a \leq X_i \leq b$ for all i.

Let $\mu = \mathbb{E}(x)$. Then, for $\delta > 0$

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

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

The more we know about distribution, the more powerful statement we can make.