CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 4 — August 28, 2023

Prof. Samson Zhou *Scribe: Shuo Xing*

1 Markov's Inequality

Theorem 1. *Let $X \geq 0$ be non-negative random variable. Then for any $t > 0$:*

$$Pr[X \geq t \cdot E[X]] \leq \frac{1}{t}.$$

Proof. By the definition of expectation of X, we have:

$$E[X] = \sum_{x \in \Omega} Pr[X = x] \cdot x \geq \sum_{x \geq t \cdot E[X]} Pr[X = x] \cdot x$$
$$\geq t \cdot E[X] \cdot \sum_{x \geq t \cdot E[X]} Pr[X = x] = t \cdot E[X] \cdot Pr[X \geq t \cdot E[X]]$$

The claimed inequality then follows from dividing by sides by $t \cdot E[X]$. ■

Comments about Markov's inequality:

- Markov's inequality bounds the probability that a random variable is "far away" from its expectation;
- Markov's inequality can not provide a tight bound, but it can help understand the guarantee of random variables and thus randomized algorithms;
- Markov's inequality also requires the assumption that the random variable X should be non-negative.

2 Variance

In order to obtain a more tight estimate of the difference between variable X and its expect value $E[x]$ based on Markov's inequality, we first introduce some notations and definitions about the *variance*.

Definition. For $p > 0$, the p-th moment of a random variable X over its sample space Ω is

$$E[X^p] = \sum_{x \in \Omega} Pr[X = x] \cdot x^p.$$

The variance of a random variable over its sample space Ω is

$$Var[X] = E[X^2] - (E[X])^2$$

Obviously, the 1-th moment of the variable X is its expected value. The variance is defined in terms of the *variable's* 2-*nd moment* - (*its* 1-*st moment*)^2. Additionally, the variance of X can be rewritten as

$$Var[X] = E[(X - E[X])^2],$$

since

$$E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - 2E[X]E[X] + (E[X])^2 = E[X^2] - (E[X])^2.$$

Definition. The standard deviation of a random variable X is $\sigma = \sqrt{Var[X]}$.

Comments on the standard deviation and the variance:

- The standard deviation (also the variance) measures how far apart the outcomes are (although the mean values of two different random variables might be identical, their variances (or standard deviations) could substantially differ, e.g., X and Y such that $Pr[X = 1] = Pr[X = -1] = 1/2$ and $Pr[Y = 100] = Pr[Y = -100] = 1/2$).
- The motivation to introduce standard deviation is that it is in the same unit as the data set.
- Linearity of variance for *independent* random variables: $Var[X + Y] = Var[X] + Var[Y]$.
- $Var[c \cdot X] = c^2 \cdot Var[X]$, where c is a constant (can be easily verified by the definition of the variance).

3 Chebyshev's Inequality

Since $|X| \geq t \iff X^2 \geq t^2$, we have

$$Pr[|X| \geq t] = Pr[X^2 \geq t^2]$$

By Markov's inequality, we know

$$Pr[|X| \geq t] = Pr[X^2 \geq t^2] \leq \frac{E[X^2]}{t^2}.$$

Plugging in $X - E[X]$ for X, we can obtain

$$Pr[|X - E[X]| \geq t] \leq \frac{E[(X - E[X])^2]}{t^2}.$$

By definition, the numerator in the right hand side of the above inequality is exactly the variance of variable X. Thus, the above inequality measures the absolute difference between random variable and its expectation in terms of its variance (deviation), which is known as *Chebyshev's Inequality* (as stated in the following Theorem 2).

Theorem 2. *Let X be a random variable with expected value $\mu := E[X]$ and variance $\sigma^2 := Var[X]$, we have*

$$Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}.$$

Unlike Markov's inequality, Chebyshev's inequality does not require any assumption about the random variable X.

4 Law of Large Numbers

Suppose we have a distribution and seek to estimate its true mean through empirical sampling (empirical mean), how accurate of an estimation can we achieve?

Let $X_1, \cdots, X_n$ be random variables that are independent identically distributed (i.i.d) with mean μ and variance σ^2, and $X = 1/n \sum_i X_i$ denote the sample average. By the linearity of expectation and variance (for independent variables), we have

$$E[X] = E\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n}E\left[\sum_i X_i\right] = \mu,$$

and

$$Var[X] = Var\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n^2}\sum_i Var\left[X_i\right] = \frac{\sigma^2}{n}.$$

Thus, by Chebyshev's inequality, we have

$$Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{nt},$$

which implies $Pr[|X - \mu| \geq t] \to 0$ as $n \to 0$ for all constant $t > 0$. Therefore we have the following Theorem 3 which is known as *Law of Large Numbers*.

Theorem 3. *[Informal] The sample average will always concentrate to the mean, given enough samples.*