Lecture 08: Concentration Bounds: Markov and Chebyshev Inequality

Overview

- In the previous lectures, we learned to compute estimates of the expected value of a random variable
- However, if the expected value a good representation of the random variable?
- If the random variable concentrates most of its probability mass around the expected value, then we consider the expected value to be a good representation of the random variable's behavior
- In the topics of concentration, we shall cover technique to argue the "typicality of a randomized experiment," i.e., say the mean of the median being a good representative of the random variable

Markov Inequality I

Theorem (Markov Inequality)

Let $\mathbb{X}$ be a r.v. over the sample space $\Omega \subseteq \mathbb{R}_{\geqslant 0}$ (i.e., the set of non-negative real numbers), and $\mu = \mathbb{E}[\mathbb{X}]$. Then, the following bound holds.

$$\mathbb{P}[\mathbb{X} \geqslant \lambda\mu] \leqslant \frac{1}{\lambda}$$

By substitution of variables, this bound is also equivalent to the following expression.

$$\mathbb{P}[\mathbb{X} \geqslant \lambda] \leqslant \frac{\mu}{\lambda}$$

Intuition: Suppose λ is large. Then, the probability that $\mathbb{X}$ deposits probability mass further than $\lambda\mu$ is unlikely. I present the proof only for discrete Ω. The case of general Ω is similar.

Markov Inequality II

Proof. If possible let, Markov inequality is false. That is, there exists $\lambda > 1$ such that $\mathbb{P}[\mathbb{X} \geqslant \lambda\mu] > 1/\lambda$. Then, let us lower-bound the expectation as follows.

$$\begin{aligned}
\mu = \mathbb{E}[\mathbb{X}] &= \sum_{i \in \Omega} i \cdot \mathbb{P}[\mathbb{X} = i] \\
&= \sum_{i \in \Omega\colon\, i < \lambda\mu} i \cdot \mathbb{P}[X = i] + \sum_{i \in \Omega\colon\, i \geqslant \lambda\mu} i \cdot \mathbb{P}[X = i] \\
&\geqslant \sum_{i \in \Omega\colon\, i < \lambda\mu} 0 \cdot \mathbb{P}[X = i] + \sum_{i \in \Omega\colon\, i \geqslant \lambda\mu} (\lambda\mu) \cdot \mathbb{P}[X = i] \\
&= 0 \cdot \mathbb{P}[\mathbb{X} < \lambda\mu] + (\lambda\mu) \cdot \mathbb{P}[\mathbb{X} \geqslant \lambda\mu] \\
&> (\lambda\mu) \cdot \frac{1}{\lambda} = \mu
\end{aligned}$$

So, we have obtained a contradiction that $\mu > \mu$.

Comments on Markov I

- We emphasize that for every μ and $\lambda \geqslant 1$, there is a distribution for which the Markov inequality is tight. Let $\mathbb{X}$ be a distribution such that $\mathbb{P}[\mathbb{X} = 0] = 1 - 1/\lambda$ and $\mathbb{P}[\mathbb{X} = \lambda\mu] = 1/\lambda$.
- If there exists B such that $\mathbb{P}[\mathbb{X} > B] = 0$, i.e., the sample space of $\mathbb{X}$ is bounded above, then we can also apply Markov inequality to the random variable $(B - \mathbb{X})$
- Think: The pigeon-hole principle states that if m balls are placed arbitrarily into n bins then there exists a bin with $\lceil m/n \rceil$ balls. How is Markov inequality equivalent to the pigeon-hole principle?

- Think: Consider the following problem. Suppose $(\mathbb{R}, \mathbb{C})$ is a joint distribution over $\Omega = \{1, \ldots, m\} \times \{1, \ldots, n\}$. Intuitively, think of a matrix with m-rows and n-columns. The r.v. associates probability to the cells. Suppose there is a Fun event and the following bound holds.

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun}\right] \geqslant \varepsilon$$

That is, if you sample a cell according to the joint distribution $(\mathbb{R}, \mathbb{C})$ then the probability of the Fun event occurring is at least ε. Consider the following expression.

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun} | \mathbb{R} = r\right]$$

This expression represents the probability of the Fun event happening if we restrict (condition) on the row $r \in \{1, \ldots, m\}$. Prove the following statement.

"The probability of sampling $r \sim \mathbb{R}$ such that it has

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun} \middle| \mathbb{R} = r\right] \geqslant \alpha$$

is at least ε/α."
Russel Impagliazzo referes to this result as the pigeon-hole principle. The proof of this result is similar to the proof of the Markov inequality. It is an excellent exercise to think of techniques to use this result for derandomization.

Theorem (Chebychev's Theorem)

For any random variable $\mathbb{X}$ *over real numbers, the following bound holds*

$$\mathbb{P}\left[|\mathbb{X}-\mu| \geqslant t\right] \leqslant \frac{\mathrm{Var}\left[\mathbb{X}\right]}{t^2},$$

where $\mu = \mathbb{E}\left[\mathbb{X}\right]$.

Proof Outline.

$$\begin{aligned}\mathbb{P}\left[|\mathbb{X}-\mu| \geqslant t\right] &= \mathbb{P}\left[(\mathbb{X}-\mu)^2 \geqslant t^2\right] \\ &\leqslant \frac{\mathbb{E}\left[(\mathbb{X}-\mu)^2\right]}{t^2} \\ &= \frac{\mathrm{Var}\left[\mathbb{X}\right]}{t^2}\end{aligned}$$

Chebyshev's Inequality II

- In the previous proof, we used the following fact.

$$\mathbb{P}\left[|\mathbb{X} - \mu| \geqslant t\right] = \mathbb{P}\left[(\mathbb{X} - \mu)^2 \geqslant t^2\right]$$

  In general, this is true for any monotonically increasing function function f. That is, for any monotonically increasing function $f \colon \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{P}\left[\mathbb{X} \geqslant t\right] = \mathbb{P}\left[f(\mathbb{X}) \geqslant f(t)\right]$$

  This trick is extremely crucial and shall be used in various other problems.

- Additionally, the random variable $(\mathbb{X} - \mu)^2$ is non-negative. Therefore, we could apply the Markov inequality to conclude the following

$$\mathbb{P}\left[(\mathbb{X} - \mu)^2 \geqslant t^2\right] \leqslant \frac{\mathbb{E}\left[(\mathbb{X} - \mu)^2\right]}{t^2}$$

Chebyshev's Inequality III

- So, we saw that "Markov studied the r.v. $\mathbb{X}$ and got a bound in $1/t$" and "Chebyshev studied the r.v. $\mathbb{X}^2$ and got a bound in $1/t^2$." Can we extrapolate this approach to use "higher powers of $\mathbb{X}$" (technically referred to as the moments) to obtain bounds that are "higher polynomials in $1/t$?"