Lecture Notes: Markov Inequality, Chebyshev Inequality, and Chernoff Bound

Yufei Tao
Chinese University of Hong Kong
taoyf@cse.cuhk.edu.hk

10 Feb, 2012

In this lecture, we will study three inequalities that are of paramount importance in analyzing randomized algorithms.

1 Markov inequality

Theorem 1 (Markov inequality)**.** *Let $X \geq 0$ be a random variable. For any $t > 0$, it holds that:*

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

Proof. Let $f(X)$ be the probability density function of X. It holds that:

$$\begin{aligned}
\mathbf{Pr}[X \geq t] &= \int_t^\infty f(X)dX \\
&= \frac{1}{t}\int_t^\infty t \cdot f(X)dX \\
(\text{as } t > 0) \quad &\leq \frac{1}{t}\int_t^\infty X \cdot f(X)dX \\
(\text{as } X \geq 0) \quad &\leq \frac{1}{t} \cdot \mathbf{E}[X]
\end{aligned}$$

as needed. □

Corollary 1. *Let $X \geq 0$ be a random variable. For any $t > 0$, it holds that:*

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

The above corollary is perhaps more intuitive (than Theorem 1): it says that the probability for X to be t times larger than its expected value is at most $1/t$.

Example: sampling. Consider a box of n balls, each of which is either black or white. We want to know the percentage p of the black balls. Each time we can look at a random ball. If we have seen k balls among which b balls are black, we estimate p to be b/k. The question is how large k needs to be before our estimate is close to p with a probability close to 1. To facilitate analysis, let us assume that after drawing a ball, we put it back into the box, so that it may be drawn again with the same chance of any other ball. This is called *sampling with replacement.*

At the end of the lecture, we will find a fairly good answer to the earlier question, but at this point, we will be content with the following (weaker) claim:

Lemma 1. *For any $k \geq 1$, the probability that $b/k \geq 2p$ is at most $1/2$.*

Proof. Define random variable x_i ($1 \leq i \leq k$) such that $x_i = 1$ if the i-th ball sampled is black, or 0 otherwise. Thus, $\mathbf{Pr}[x_i = 1] = p$, which implies that $\mathbf{E}[x_i] = p$. Clearly, $b = \sum_{i=1}^{k} x_i$. Hence:

$$\mathbf{E}[b] = \mathbf{E}\left[\sum_{i=1}^{k} x_i\right] = \sum_{i=1}^{k} \mathbf{E}[x_i] = kp.$$

By Markov inequality:

$$\begin{aligned} \mathbf{Pr}\big[b \geq 2 \cdot \mathbf{E}[b]\big] &\leq 1/2 \\ \Rightarrow \mathbf{Pr}[b \geq 2kp] &\leq 1/2 \\ \Rightarrow \mathbf{Pr}[b/k \geq 2p] &\leq 1/2. \end{aligned}$$

□

The lemma indicates that, we can over-estimate p by a factor of 2 with at most 50% probability, regardless of how many balls are sampled.

2 Chebyshev inequality

Theorem 2 (Chebyshev inequality)**.** *Let X be a random variable with expectation μ and variance σ^2 (namely, $\sigma > 0$ is the standard deviation). For any $t > 0$, it holds that:*

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

Proof.

$$\mathbf{Pr}[|X - \mu| \geq t\sigma] = \mathbf{Pr}[(X - \mu)^2 \geq t^2\sigma^2]$$

As $\sigma^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mathbf{E}[X^2] - \mu^2$, we have:

$$\mathbf{Pr}[(X - \mu)^2 \geq t^2\sigma^2] = \mathbf{Pr}[(X - \mu)^2 \geq t^2(\mathbf{E}[X^2] - \mu^2)]$$

Define $Y = (X - \mu)^2 = X^2 - 2X\mu + \mu^2$. Thus:

$$\begin{aligned} \mathbf{E}[Y] &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mu + \mu^2 \\ &= \mathbf{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbf{E}[X^2] - \mu^2. \end{aligned}$$

Hence:

$$\mathbf{Pr}[(X - \mu)^2 \geq t^2(\mathbf{E}[X^2] - \mu^2)] = \mathbf{Pr}\big[Y \geq t^2\mathbf{E}[Y]\big]$$

which is at most $1/t^2$ by Markov inequality. □

The theorem says that X can deviate from its expectation by at least t times the standard deviation with probability at most $1/t^2$.

Corollary 2. *Let X be a random variable with expectation μ and variance σ^2 (namely, $\sigma > 0$ is the standard deviation). For any $t > 0$, it holds that:*

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

Sampling (cont.). We now utilize Chebyshev inequality to obtain a stronger claim about the sampling scenario described in the earlier section:

Lemma 2. *Let δ be any value satisfying $0 < \delta < 1$. With $k = \frac{1}{\epsilon^2 \delta}$, the probability that $|b/k - p| \leq \epsilon$ is at least $1 - \delta$.*

Proof. Define random variables $x_1, ..., x_k$ as in the proof of Lemma 1. For each $i \in [1, k]$, $\mathbf{E}[x_i] = p$ and $\mathbf{var}[x_i] = p(1-p)$. As $b = \sum_{i=1}^{k} x_i$ and $x_1, ..., x_k$ are mutually independent, we know:

$$\begin{aligned}
\mathbf{E}[b] &= kp \\
\mathbf{var}[b] &= \sum_{i=1}^{k} \mathbf{var}[x_i] = kp(1-p)
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathbf{Pr}[|b/k - p| \geq \epsilon] &= \mathbf{Pr}[|b - kp| \geq \epsilon k] \\
\text{(by Chebyshev inequality)} &\leq \frac{\mathbf{var}[b]}{\epsilon^2 k^2} = \frac{kp(1-p)}{\epsilon^2 k^2} = \frac{p(1-p)}{\epsilon^2 k} \\
&< 1/(\epsilon^2 k) \\
&= \delta
\end{aligned}$$

□

Note that the value of k in the above lemma is *independent* on n. In other words, a fixed number of samples is sufficient to achieve the probabilistic guarantee described by the same pair of ϵ and δ, regardless of the size of population (good news for surveying in a populous country).

3 Chernoff bound

Theorem 3 (Chernoff bound)**.** *Let $X_1, ..., X_k$ be k independent random variables such that, for each $i \in [1, k]$, X_i equals 1 with probability p, and 0 with probability $1 - p$. Let $X = \sum_{i=1}^{k} X_i$ and $\mu = kp$. For any ϵ satisfying $0 < \epsilon < 1$, it holds that:*

$$\mathbf{Pr}\Big[|X - \mu| \geq \epsilon\mu\Big] \leq 2e^{\frac{-\epsilon^2 \mu}{3}}.$$

We will skip the proof, which can be found in [1] and is a bit technically involved. Remember that this theorem demands $X_1, ..., X_k$ to be independent. When this condition is fulfilled, the Chernoff bound usually gives a tighter bound. Next, we demonstrate this in the sampling scenario of the previous sections.

Sampling (cont.). We will prove the following which significantly improves Lemma 2 when δ is small.

Theorem 4. *Let δ be any value satisfying $0 < \delta < 1$. With $k = \frac{3}{\epsilon^2} \ln \frac{1}{\delta}$, the probability that $|b/k - p| \leq \epsilon$ is at least $1 - \delta$.*

Proof. Define random variables $x_1, ..., x_k$ as in the proof of Lemma 1; recall that they are independent. For each $i \in [1, k]$, $\mathbf{E}[x_i] = p$. As $b = \sum_{i=1}^{k} x_i$, we know $\mathbf{E}[b] = kp$. Hence:

$$\begin{aligned}
\mathbf{Pr}[|b/k - p| \geq \epsilon] &= \mathbf{Pr}[|b - kp| \geq \epsilon k] \\
&= \mathbf{Pr}\left[|b - kp| \geq \frac{\epsilon}{p} kp\right] \\
&= \mathbf{Pr}\left[|b - \mathbf{E}[b]| \geq \frac{\epsilon}{p} \cdot \mathbf{E}[b]\right] \\
\text{(by Chernoff bound)} &= \leq 2e^{-\frac{\epsilon^2}{3p^2} kp} \\
&= 2e^{-\frac{\epsilon^2 k}{3p}}
\end{aligned}$$

To make the above at most δ, we need:

$$k \geq \frac{3p}{\epsilon^2} \ln \frac{2}{\delta}$$

As $p < 1$, taking $k = \frac{3}{\epsilon^2} \ln \frac{2}{\delta}$ guarantees the above. □

References

[1] T. Hagerup and C. Rub. A guided tour of chernoff bounds. *Information Processing Letters (IPL)*, 33(6):305–308, 1990.