EE 126 — Probability and Random Processes
February 08, 2018 — Bounds and Limits

1 Reading

5.1. Markov and Chebyshev Inequalities

These inequalities use the mean and possibly the variance of a random variable in order to draw conclusions about the probabilities of certain events. They are useful for cases when the mean and/or variance are easily computable, but the distribution is not.

The **Markov inequality** asserts that if a nonnegative random variable has a small mean, then the probability that it takes a large value must also be small.

Markov Inequality
If a random variable X can only take nonnegative values, then

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad \text{for all } a > 0$$

This can be seen by defining a random variable Y_a over a fixed positive number a:

$$Y_a = \begin{cases} 0 & \text{if } X < a \\ a & \text{if } X \geq a \end{cases}$$

Then

$$\mathbb{E}[X] \geq \mathbb{E}[Y_a] = aP(Y_a = a) = aP(x \geq a)$$

However, the bounds provided by the Markov inequality can be rather loose. We continue with the **Chebyshev inequality**, which states that if a random variable has a small variance, then the probability that it takes a value far from its mean is also small.

Chebyshev Inequality
If X is a random variable with mean μ and variance σ^2, then

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} \quad \text{for all } c > 0$$

This can be seen by applying the Markov inequality to the nonnegative random variable $(X - \mu)^2$ with $a = c^2$:

$$P((X - \mu)^2 \geq c^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

and also noting that the event $(X - \mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$.

An alternative form of the Chebyshev inequality is obtained by letting $c = k\sigma$, where k is positive, which yields

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

Hence the probability that a random variable takes a value more than k standard deviations away from its mean is at most $1/k^2$.

W13.7. Bounds on Probabilities

Chernoff's inequality states that $P(X \geq a) \leq \mathbb{E}[e^{\theta(X-a)}]$ for all $\theta > 0$. **Jensen's inequality** states that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ for all $f(\cdot)$ that are convex.

2 Lecture

Recap: MGFs

$$M_X(s) = \mathbb{E}[e^{sX}] = \begin{cases} \sum_k e^{sk} P(X = k) & \text{for discrete } X \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) dx & \text{for continuous } X \end{cases}$$

We then have

$$\frac{d^n}{ds^n} M_X(s)\Big|_{s=0} = \mathbb{E}[X^n]$$

- If $X \sim \text{Exp}(\lambda) \implies f_X(x) = \lambda e^{-\lambda x}; x \geq 0$, then its MGF is $M_X(s) = \frac{\lambda}{\lambda - s}$ for $s < \lambda$.
- If $X \sim \text{Poisson}(\lambda) \implies P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}; k = 0, 1, ...$, then its MGF is $M_X(s) = e^{-\lambda + \lambda e^s}$.
- If $X \sim \mathcal{N}(0, 1)$, then its MGF is $M_X(s) = e^{s^2/2}$. If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then its MGF is $M_Y(s) = e^{s\mu + s^2\sigma^2/2}$.

If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(sa)$.

If $Y = X_1 + X_2 + ... + X_n$, where the X_i's are i.i.d., then

$$M_Y(s) = \prod_{i=1}^{n} M_{X_i}(s) = [M_X(s)]^n$$

If we want to find $f_Y(y)$ where $Y = X_1 + X_2$ (and X_1 and X_2 are independent), then

- $f_{X_1}(x) \xrightarrow{T} M_{X_1}(s)$
- $f_{X_2}(x) \xrightarrow{T} M_{X_2}(s)$
- $M_{X_1}(s) \cdot M_{X_2}(s) = M_Y(s)$
- $M_Y(s) \xrightarrow{T^{-1}} f_Y(y)$

This is the prescription for doing convolution, not only in probability, but also in other fields such as signal processing!

Example: convolving two Gaussians. Let $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(0, 1)$. Let X_1 and X_2 be independent. Then

$$\begin{aligned} Y &= X_1 + X_2 \\ M_{X_1}(s) &= e^{s^2/2} \\ M_{X_2}(s) &= e^{s^2/2} \\ M_Y(s) &= e^{s^2} \end{aligned}$$

from which we realize that $Y \sim \mathcal{N}(0, 2)$.

In general, if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, X_1 and X_2 are independent, and $Y = X_1 + X_2$, then

$$M_{X_1}(s) = e^{s^2\sigma_1^2/2+\mu_1 s}$$
$$M_{X_2}(s) = e^{s^2\sigma_2^2/2+\mu_2 s}$$
$$M_Y(s) = e^{s^2(\sigma_1^2+\sigma_2^2)/2+(\mu_1+\mu_2)s}$$

and

$$Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Bounds and Limits

We are interested in studying the limit behavior of a sequence of random variables, what the sequence converges to, and at what "rate?" We will start with the most elementary bound: the Markov bound.

Markov

The **Markov bound** asserts that if X is a nonnegative random variable, then

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

Proof.

$$\{1\}_{X \geq a} \leq \frac{X}{a}$$

Note that $\{1\}$ is an indicator function. It is 1 when $X \geq a$, and 0 otherwise. Taking the expectation $\mathbb{E}_X(\cdot)$ of both sides, we see

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

We can also prove the Markov bound pictorially, and by using the tail sum formula.

Example. X is the height of a random adult in Berkeley. We know that $\mathbb{E}[X] = 68$ inches. Then Markov says that $P(X > 144 \text{ in}) < \frac{68}{144} = 0.47$. This is a rather bad bound, but to be fair Markov doesn't use much information – only the mean!

Markov is mostly only useful as a building block.

Example: flipping a fair coin across i.i.d. trials. Let $X = X_1 + X_2 + ... + X_n$, and let

$$X_i = \begin{cases} 1 & \text{if toss } i \text{ is heads} \\ 0 & \text{otherwise} \end{cases}$$

Then $\mathbb{E}[X] = n\frac{1}{2} = n/2$, meaning that if $n = 1000$, $P(X > 900) \leq \frac{500}{900} = 5/9$. Once again there is a really big gap between Markov and what the upper bound could be.

Example. Let $X \sim \text{Exp}(1)$. Thus $P(X > x) = e^{-x}; x > 0$, and Markov tells us that

$$P(X > x) \leq \frac{1}{x}$$

As a summary of Markov:

- It is a weak inequality
- It uses only the mean of the distribution
- When we only know the mean, perhaps it's the best we can do. But if we know more, we should use something else

Chebyshev

Chebyshev builds on Markov by taking the square function. If X is a random variable with finite mean and variance σ^2, Chebyshev says that

$$P(|X - \mathbb{E}[X]| \geq c) \leq \frac{\sigma^2}{c^2} \qquad \forall\ c > 0$$

As a special case,

$$P(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2} \qquad \forall\ k > 0$$

Proof.

$$P(|X - \mathbb{E}[X]| \geq c) = P(|X - \mathbb{E}[X]|^2 \geq c^2) \leq \mathbb{E}[|X - \mathbb{E}[X]|^2]/c^2 = \sigma^2/c^2$$

Example: adult height in Berkeley. We have $\mathbb{E}[X] = 68$ in, and $\sigma_X^2 = 49\text{in}^2$. Chebyshev tells us that

$$P(X \geq 144in) \leq P(|X - 68| \geq 76) \leq \frac{49}{76^2} = 0.0084$$

which is better than Markov!

Example: random walk. A drunk guy starts from a bar. Wherever he is, he flips a coin and either goes left by one step or right by one step. We want to know the value of $P(\text{after } n = 10000 \text{ steps, drunk is more than 400 steps from the bar})$.

We can compute a bound on it as follows: $X = \sum_{i=1}^{n} X_i$, so $\mathbb{E}[X_i] = 0$ and $var(X_i) = \mathbb{E}[X_i^2] = 1$. Thus $E[X] = 0$ and $var(X) = n$. Chebyshev tells us that

$$P(|X| > k\sqrt{n}) \leq \frac{n}{k^2 n} = \frac{1}{k^2}$$

meaning $P(\text{greater than 400 steps away}) < \frac{1}{16}$. This is still not that great of a bound.

Example. Let $X \sim \mathcal{N}(0, 1)$. What is the probability we're three standard deviations away from the mean? Chebyshev gives us $P(|X| > 3) \leq 1/9 = 0.111$, but this is at least ten times worse than what it actually is (≈ 0.001).

The problem with our bounds so far is that they're only using one or two moments. We should use all the moments! This is where moment-generating functions come in.

Chernoff

The Chernoff bound also builds on Markov. $P(X \geq a) \leq E[X]/a$ for $a > 0$, but we can pick any (X, a) we want! Let's pick $X = e^{sY}$ and $a = e^{sb}$. This gives us

$$P(e^{sY} \geq e^{sb}) \leq \mathbb{E}[e^{sY}]/e^{sb} = M_Y(s) \cdot e^{-sb}$$

Chernoff picks the best s in the best way and gets the best bound.

- For $s \geq 0$, $P(Y \geq b)$ (the upper tail) $= P(e^{sY} \geq e^{sb}) \leq e^{-sb} M_Y(s)$.
- For $s < 0$, $P(Y \leq b)$ (the lower tail) $= P(e^{sY} \geq e^{sb}) \leq e^{-sb} M_Y(s)$.

Note that the left-hand side is not a function of s, but the right-hand side is! We can optimize over s to get the tightest bound.

The Chernoff bound is about as powerful as it gets. Qualitatively, it's better because $M_Y(s)$ basically contains all of the information about the distribution. We're using not just the mean, not just the variance, but all of the more subtle characteristics captured by the MGF.

$$P(Y \geq b) \leq e^{-sb} M_Y(s) = e^{-sb} e^{\ln M_Y(s)} = e^{-(sb - \ln M_Y(s))}$$

The best bound is found by maximizing the exponent!

Example. Let $Y \sim \mathcal{N}(0,1) \implies M_Y(s) = e^{s^2}/2$. Then $\max_{s\geq 0}(sb - s^2/2)$ is at $s = b$. Trying it out, we have $P(Y \geq b) \leq e^{-b^2/2}$.

Example. With Chebyshev, $P(|\mathcal{N}(0,1)| > 3) \leq 1/9 \approx 0.111$. With Chernoff, this becomes $P(|\mathcal{N}(0,1)| > 3) \leq e^{-4.5} \approx 0.0111$.

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific books. Athena Scientific, 2002.

[2] Jean Walrand. *Probability in Electrical Engineering and Computer Science: An Application-Driven Course.* Quorum Books, Westport, CT, USA, 2014.