Chapter 6. Concentration Inequalities

6.1: Markov and Chebyshev Inequalities

(From "Probability & Statistics with Applications to Computing" by Alex Tsun)

When reasoning about some random variable X, it's not always easy or possible to calculate/know its exact PMF/PDF. We might not know much about X (maybe just its mean and variance), but we can still provide **concentration inequalities** to get a bound of how likely it is for X to be far from its mean μ (of the form $\mathbb{P}(|X-\mu|>\alpha)$), or how likely for this random variable to be very large (of the form $\mathbb{P}(X \geq k)$).

You might ask when we would only know the mean/variance but not the PMF/PDF? Some of our distributions that we use (like Exponential for bus waiting time), are just modelling assumptions and are probably incorrect. If we measured how long it took for the bus to arrive over many days, we could *estimate* its mean and variance! That is, we have no idea the true distribution of daily bus waiting times but can get good estimates for the mean and variance. We can use these concentration inequalities to bound the probability that we wait too long for a bus knowing just those two quantities and nothing else!

6.1.1 Markov's Inequality

We'll start with our weakest inequality, Markov's inequality. This one only requires us to know the mean, and nothing else! Again, if we didn't know the PMF/PDF of what we cared about, we could use the sample mean as a good estimate for the true mean (by the Law of Large Numbers from 5.7), and our inequality/bound would be pretty accurate still!
This first example will help build intuition for why Markov's inequality is true.

Example(s)

The score distribution of an exam is modelled by a random variable X with range $\Omega_X = [0, 110]$ (with 10 points for extra credit). Give an upper bound on the proportion of students who score at least 100 when the average is 50? When the average is 25?

Solution What would you guess? If the average is $\mathbb{E}[X] = 50$, an upper bound on the proportion of students who score at least 100 should be 50% right? If more than 50% of students scored a 100 (or higher), the average would already be 50% since all scores must be nonnegative (≥ 0). Mathematically, we just argued that:

$$\mathbb{P}(X \geq 100) \leq \frac{\mathbb{E}[X]}{100} = \frac{50}{100} = \frac{1}{2}$$

This sounds reasonable - if say 70% of the class were to get 100 or higher, the average would already be at least 70%, even if everyone else got a zero. The best bound we can get is 50% - and that requires everyone else to get a zero.

If the average is $\mathbb{E}[X] = 25$, an upper bound on the proportion of students who score at least 100 is:

$$\mathbb{P}(X \geq 100) \leq \frac{\mathbb{E}[X]}{100} = \frac{25}{100} = \frac{1}{4}$$

Similarly, if we had more than 30% students get 100 or higher, the average would already be at least 30%, even if everyone else got a zero. □

That's literally the entirety of the idea for Markov's inequality.

Theorem 6.1.1: Markov's Inequality

Let $X \geq 0$ be a **non-negative** random variable (discrete or continuous), and let $k > 0$. Then:

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

Equivalently (plugging in $k\mathbb{E}[X]$ for k above):

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Proof of Markov's Inequality. Below is the proof when X is continuous. The proof for discrete RVs is similar (just change all the integrals into summations).

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f_X(x)dx && \text{[because } X \geq 0\text{]} \\
&= \int_0^k x f_X(x)dx + \int_k^\infty x f_X(x)dx && \text{[split integral at some } 0 \leq k \leq \infty\text{]} \\
&\geq \int_k^\infty x f_X(x)dx && \left[\int_0^k x f_X(x)dx \geq 0 \text{ because } k \geq 0, x \geq 0 \text{ and } f_X(x) \geq 0\right] \\
&\geq \int_k^\infty k f_X(x)dx && \text{[because } x \geq k \text{ in the integral]} \\
&= k\int_k^\infty f_X(x)dx \\
&= k\mathbb{P}(X \geq k)
\end{aligned}$$

□

So just knowing that the random variable is non-negative and what its expectation is, we can bound the probability that it is "very large". We know nothing else about the exam distribution! Note there is no bound we can derive if X could be negative. Always check that X is indeed nonnegative before applying this bound!

The following example demonstrates how to use Markov's inequality, and how loose it can be in some cases.

Example(s)

A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times. Use Markov's inequality to bound the probability it lands on heads at least 16 times.

Solution We actually do know this distribution; the number of heads is $X \sim \text{Bin}(n = 20, p = 0.2)$. Thus, $\mathbb{E}[X] = np = 20 \cdot 0.2 = 4$. By Markov's inequality:

$$\mathbb{P}(X \geq 16) \leq \frac{\mathbb{E}[X]}{16} = \frac{4}{16} = \frac{1}{4}$$

Let's compare this to the actual probability that this happens:

$$\mathbb{P}(X \geq 16) = \sum_{k=16}^{20} \binom{20}{k} 0.2^k \cdot 0.8^{20-k} \approx 1.38 \cdot 10^{-8}$$

This is not a good bound, since we only assume to know the expected value. Again, we knew the exact distribution, but chose not to use any of that information (the variance, the PMF, etc.). □

Example(s)

Suppose the expected runtime of QuickSort is $2n\log(n)$ operations/comparisons to sort an array of size n (we can show this using linearity of expectation with dependent indicator variables). Use Markov's inequality to bound the probability that QuickSort runs for longer than $20n\log(n)$ time.

Solution Let X be the runtime of QuickSort, with $\mathbb{E}[X] = 2n\log(n)$. Then, since X is non-negative, we can use Markov's inequality:

$$\begin{aligned}
\mathbb{P}(X \geq 20n\log(n)) &\leq \frac{\mathbb{E}[X]}{20n\log(n)} && \text{[Markov's inequality]} \\
&= \frac{2n\log(n)}{20n\log(n)} \\
&= \frac{1}{10}
\end{aligned}$$

So we know there's at most 10% probability that QuickSort takes this long to run. Again, we can get this bound despite not knowing anything except its expectation! □

6.1.2 Chebyshev's Inequality

Chebyshev's inequality unlike Markov's inequality does not require that the random variable is non-negative. However, it also requires that we know the variance in addition to the mean. The goal of Chebyshev's inequality is to bound the probability that the RV is far from its mean (in either direction). This generally gives a stronger bound than Markov's inequality; if we know the variance of a random variable, we should be able to control how much if deviates from its mean better!

We'll actually prove the Weak Law of Large Numbers as well!

Theorem 6.1.2: Chebyshev's Inequality

Let X be any random variable with expected value $\mu = \mathbb{E}[X]$ and finite variance $\mathsf{Var}(X)$. Then, for any real number $\alpha > 0$:

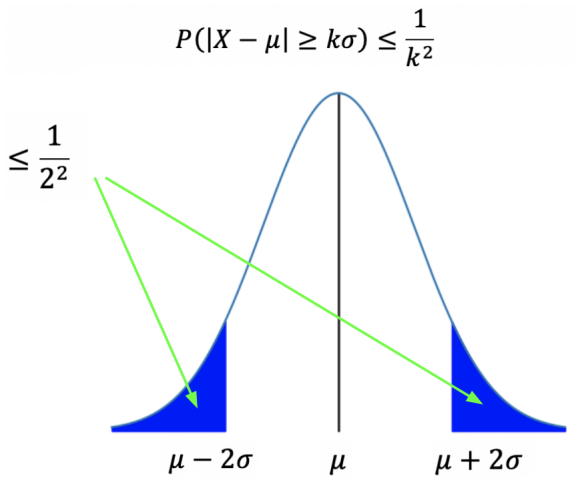
$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\mathsf{Var}(X)}{\alpha^2}$$

Equivalently (plugging in $k\sigma$ for α above, where $\sigma = \sqrt{\mathsf{Var}(X)}$):

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

This is used to bound the probability of being in the *tails*. Here is a picture of Chebyshev's inequality

bounding the probability that a Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ is more than $k = 2$ standard deviations from its mean:



Proof of Chebyshev's Inequality. X is a random variable, so $(X - \mathbb{E}[X])^2$ is a **non-negative** random variable. Hence, we can apply Markov's inequality.

$$\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}\left[X\right]| \geq \alpha\right) &= \mathbb{P}\left(\left(X - \mathbb{E}\left[X\right]\right)^2 \geq \alpha^2\right) && \text{[square both sides]} \\
&\leq \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]}{\alpha^2} && \text{[Markov's inequality]} \\
&= \frac{\mathsf{Var}\left(X\right)}{\alpha^2} && \text{[def of variance]}
\end{aligned}$$

□

While in principle Chebyshev's inequality asks about distance from the mean in either direction, it can still be used to give a bound on how often a random variable can take large values, and will usually give much better bounds than Markov's inequality. This is expected, since we also assume to know the variance - and if the variance is small, we know the RV can't deviate too far from its mean.

Example(s)

Let's revisit the example in Markov's inequality section earlier in which we toss a weighted coin independently with probability of landing heads $p = 0.2$. Upper bound the probability it lands on heads at least 16 times out of 20 flips using Chebyshev's inequality.

Solution Because $X \sim \text{Bin}(n = 20, p = 0.2)$:

$$\mathbb{E}\left[X\right] = np = 20 \cdot 0.2 = 4$$

and:

$$\mathsf{Var}\left(X\right) = np(1-p) = 20 \cdot 0.2 \cdot (1 - 0.2) = 3.2$$

Note that since Chebyshev's asks about the difference in either direction of the RV from its mean, we must weaken our statement first to include the probability $X \leq -8$. The reason we chose -8 is because

Chebyshev's inequality is symmetric about the mean (difference of 12; 4 ± 12 gives the interval $[-8, 16]$):

$$\begin{aligned}
\mathbb{P}(X \geq 16) &\leq \mathbb{P}(X \geq 16 \cup X \leq -8) && \text{[adding another event can only increase probability]} \\
&= \mathbb{P}(|X - 4| \geq 12) && \text{[def of abs value]} \\
&= \mathbb{P}(|X - \mathbb{E}[X]| \geq 12) && [\mathbb{E}[X] = 4] \\
&\leq \frac{\mathsf{Var}(X)}{12^2} && \text{[Chebyshev's inequality]} \\
&= \frac{3.2}{12^2} = \frac{1}{45}
\end{aligned}$$

This is a much better bound than given by Markov's inequality, but still far from the actual probability. This is because Chebyshev's inequality only takes the mean and variance into account. There is so much more information about a RV than just these two quantities! □

We can actually use Chebyshev's inequality to prove an important result from 5.7: The Weak Law of Large Numbers. The proof is so short!

6.1.3 Proof of the Weak Law of Large Numbers

Theorem 6.1.3: Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables with mean μ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\overline{X}_n$ converges in probability to μ. That is, for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) = 0$$

Proof. By the property of the expectation and variance of sample mean consisting of n iid variables: $\mathbb{E}\left[\overline{X}_n\right] = \mu$ and $\mathsf{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$ (from 5.7). By Chebyshev's inequality:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \leq \lim_{n\to\infty} \frac{\mathsf{Var}\left(\overline{X}_n\right)}{\epsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

□