EE 520: Random Processes **Fall 2020**

Lecture 9
Concentration of Random Variables

Instructor Name: John Lipor

Recommended Reading: Pishro-Nik: 6.2.0 - 6.2.3, 7.1; Gubner: 2.4, 3.3, 4.5, 5.6

In many cases, obtaining an exact probability of an event is too difficult, and we're satisfied with bounding a probability. We used the trivial bound $P(A) \leq 1$ on Homework 1, problem 4. Today, we'll look at more sophisticated tools, many of which are widely used within machine learning.

1 Bounding Random Variables

1.1 Markov's Inequality

The first inequality we'll consider is Markov's inequality. This relies on the least amount of information (requires the expectation only), but it also has the most strict assumptions and tends to be loose, i.e., may predict an upper bound that is much larger than the true probability.

Theorem 1. Let $X \geq 0$ be a RV. For all $t > 0$,

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

Proof. Any value of x can be written as

$$x = x\mathbb{1}_{\{x \geq t\}} + x\mathbb{1}_{\{x < t\}}.$$

Substitute the RV X for x above and take the expectation to see that

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[X\mathbb{1}_{\{X \geq t\}}\right] + \mathbb{E}\left[X\mathbb{1}_{\{X < t\}}\right] \\
&\geq \mathbb{E}\left[X\mathbb{1}_{\{X \geq t\}}\right] + 0 \\
&\geq \mathbb{E}[t\mathbb{1}_{\{X \geq t\}}] \\
&= tP(X \geq t),
\end{aligned}$$

where the bound of zero in the second line follows since $X \geq 0$ and the third line follows since $X \geq t$ by the indicator. The above indicates that

$$\mathbb{E}[X] \geq tP(X \geq t)$$

as desired. □

Example 1. Let $X \sim \text{Geom}(p)$. By direct calculation, we have

$$P(X \geq 3) = (1-p)^3.$$

Applying Markov's inequality, we get

$$P(X \geq 3) \leq \frac{1}{3p},$$

which can be very loose for small p. Note also that Markov's inequality has a *linear* dependence on t, i.e., it decays slowly as t increases. The next step up is quadratic dependence on t.

1.2 Chebyshev's Inequality

Chebyshev's inequality gives a quadratic dependence on t and allows us to talk about the deviation of a RV from its mean, i.e., the *concentration* of a RV. Two forms of this inequality are given below.

Theorem 2. Let X be a RV. For all $t > 0$,

$$P\left(|X| \geq t\right) \leq \frac{\mathbb{E}\left[X^2\right]}{t^2}.$$

In particular,

$$P\left(|X - \mathbb{E}[X]| \geq t\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{t^2} = \frac{\mathrm{var}(X)}{t^2}.$$

Note that by using second-order information (the second moment), we obtain faster decay in t.

Example 2. Toss a fair coin N times and let $S_N = \sum_{i=1}^{N} \mathbb{1}_{\{X_i\}}$ be the number of heads, where X_i is an indicator RV taking the value 1 if the ith toss is a head. What is $P\left(S_N \geq \frac{3}{4}N\right)$?

To apply Markov and Chebyshev, we need the first two moments. Since S_N has binomial distribution, we have

$$\mathbb{E}[S_N] = \frac{N}{2}, \quad \mathrm{var}(S_N) = \frac{N}{4}.$$

In this case, Markov's inequality gives us

$$P\left(S_N \geq \frac{3}{4}N\right) \leq \frac{N/2}{3N/4} = \frac{4}{6},$$

which does not depend on N. Intuitively, this is a bad thing, since we expect the number of heads to tend toward 1/2 as N grows. Chebyshev's inequality gives

$$\begin{aligned} P\left(S_N \geq \frac{3}{4}N\right) &= P\left(S_N - \frac{N}{2} \geq \frac{N}{4}\right) \\ &\leq P\left(\left|S_N - \frac{N}{2}\right| \geq \frac{N}{4}\right) \leq \frac{4}{N}. \end{aligned}$$

We see that Chebyshev gives us an answer that depends on N, which is a step in the right direction. But is this inequality tight? In other words, does this inequality reflect how S_N behaves, or does it tend to be closer to its mean?

2 Limit Theorems

Recall that if $X_1, \ldots, X_N$ are independent, then

$$\mathrm{var}\left(X_1 + \cdots + X_N\right) = \sum_{i=1}^{N} \mathrm{var}(X_i),$$

and if they are i.i.d. with variance σ^2, we get

$$\mathrm{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{\sigma^2}{N}.$$

Hence, for our coin example, the variance goes to zero as N increases (thanks *a lot* Markov!). Another way of stating this is that the sample mean $M_N = S_N/N$ *converges* to the true mean. This fact is formalized in the *Law of Large Numbers*.

Theorem 3 (Weak Law of Large Numbers)**.** Let $X_1, \dots, X_N$ be uncorrelated RVs all having the same mean μ and variance σ^2. Then for every $\varepsilon > 0$,

$$\lim_{N\to\infty} P\left(\left|\frac{1}{N}S_N - \mu\right| \geq \varepsilon\right) = 0,$$

where $S_N = \sum_{i=1}^N X_i$.

An even stronger result shows that this convergence happens almost surely, i.e., with probability 1, when we assume the RVs are independent instead of only uncorrelated.

Theorem 4 (Strong Law of Large Numbers)**.** Let $X_1, \dots, X_N$ be i.i.d. RVs with mean μ and variance σ^2. Then as $N \to \infty$

$$\frac{1}{N}S_N \to \mu, \quad \text{with probability 1.}$$

Both the strong and weak laws of large numbers tell us where the sample mean goes as $N \to \infty$. As a next step, we can talk about what the distribution is as $N \to \infty$. This is done via the *central limit theorem* (CLT).

Theorem 5 (Lindberg-Lévy Central Limit Theorem)**.** Let $X_1, X_2, \dots$ be a sequence of i.i.d. RVs with mean μ and variance σ^2. Define

$$Z_N = \frac{S_N - \mathbb{E}[S_N]}{\sqrt{\mathrm{var}(S_N)}} = \frac{1}{\sigma\sqrt{N}}\sum_{i=1}^N (X_i - \mu).$$

Then as $N \to \infty$,

$$Z_N \to \mathcal{N}(0,1)$$

in distribution, i.e.,

$$\lim_{N\to\infty} F_{Z_N}(z) = \Phi(z),$$

where $\Phi(z)$ is the standard normal CDF.

Back to our coin tossing example, when N is large enough, we should have

$$P\left(S_N \geq \frac{3}{4}N\right) = P\left(Z_N \geq \sqrt{\frac{N}{4}}\right) \approx P\left(g \geq \sqrt{\frac{N}{4}}\right),$$

where $g \sim \mathcal{N}(0,1)$. The next question is how the tails of a Gaussian decay.

Proposition 1. Let $g \sim \mathcal{N}(0,1)$. Then for all $t \geq 1$,

$$P(g \geq t) \leq \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

Proof. Using the PDF of a Gaussian, we see that

$$\begin{aligned} P\left(g \geq t\right) &= \int_t^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-t^2/2}e^{-ty}e^{-y^2/2}dy \\ &\leq \frac{1}{\sqrt{2\pi}}e^{-t^2/2}\int_0^\infty e^{-ty}dy \\ &\leq \frac{1}{\sqrt{2\pi}}e^{-t^2/2}, \end{aligned}$$

where we have substituted $x = t + y$, used the fact that $e^{-y^2/2} \leq 1$, and noted that the integral in the last line is upper bounded by 1 since $t \geq 1$. □

Given the above bound, our new approximation says that

$$P\left(S_N \geq \frac{3}{4}N\right) \lesssim \frac{1}{\sqrt{2\pi}}e^{-N/8},$$

which decays very fast in N. Unfortunately, we cannot prove this rigorously using the CLT, because the approximation error in the CLT decays like $\frac{1}{\sqrt{N}}$. However, we can do this if we introduce a new tool.

3 Sub-Gaussian Random Variables

3.1 Chernoff's Bounding Method

Although Markov's inequality wasn't helpful in solving our problem, it is many practical uses. One of these is in deriving other concentration inequalities, as we now see.

Theorem 6. Let X be a RV. For all $t > 0$,

$$\begin{aligned} P\left(X \geq t\right) &\leq \min_{s\geq 0} e^{-st}M_X(x) \\ &= \min_{s\geq 0} e^{-st}\mathbb{E}\left[e^{sX}\right]. \end{aligned}$$

Proof. Using Markov's inequality after exponentiating (since Markov only works with non-negative RVs), we have

$$\begin{aligned} P(X \geq t) &= P\left(sX \geq st\right) \\ &= P\left(e^{sX} \geq e^{st}\right) \\ &\leq e^{-st}\mathbb{E}\left[e^{sX}\right]. \end{aligned}$$

□

Now let's see what Chernoff tells us about the concentration of Gaussian RVs.

Example 3. Let $X \sim \mathcal{N}(0, \sigma^2)$. Then $M_X(s) = e^{s^2\sigma^2/2}$ and Chernoff's method gives us

$$P\left(X \geq t\right) \leq \min_{s\geq 0} e^{-st}e^{s^2\sigma^2/2}.$$

The above is easier to optimize if we take a logarithm of both sides, which doesn't change the minimizer because it is a monotone function. This gives

$$\log\left(P(X \geq t)\right) \leq \min_{s\geq 0} -st + \frac{s^2\sigma^2}{2}.$$

We can optimize the above by differentiating and setting the result to zero

$$\frac{d}{ds} - st + \frac{s^2\sigma^2}{2} = -t + s\sigma^2 = 0,$$

which gives

$$s = \frac{t}{\sigma^2}.$$

Plugging this value into the above, we get

$$P(X \geq t) \leq e^{-t^2/2\sigma^2},$$

i.e., that the tails of a Gaussian fall off exponentially fast. This fact is so important that we use it to define a whole class of RVs called *sub-Gaussian random variables.*

Definition 1. A RV X with mean $\mu = \mathbb{E}[X]$ is called **sub-Gaussian** if there exists a positive parameter σ such that

$$\mathbb{E}\left[e^{s(X-\mu)}\right] \leq e^{s^2\sigma^2/2}.$$

Any sub-Gaussian RV can be plugged into Chernoff's bounding method as above to achieve the same concentration result.

3.2 Hoeffding's Inequality

Using the above definition, we now state a strong concentration result for sub-Gaussian RVs.

Theorem 7. Let $X_1, \ldots, X_N$ be independent sub-Gaussian RVs with mean μ and parameter σ^2. Then for every $t \geq 0$,

$$P\left(\sum_{i=1}^{N}(X_i-\mu) \geq t\right) \leq \exp\left(-\frac{t^2}{2N\sigma^2}\right).$$

The two-sided bound also holds

$$P\left(\left|\sum_{i=1}^{N}(X_i-\mu)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2N\sigma^2}\right).$$

Aside from a Gaussian RV, an important example of a sub-Gaussian RV is that of a bounded RV.

Fact 1. Let X be a zero-mean RV supported on $[a,b]$. Then X is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$.

Back to our coin tossing experiment, we can now let $Y_i \in \{-1, 1\}$ be a symmetric Bernoulli (a.k.a. Rademacher) RV. We initially had

$$X_i = \begin{cases} 1 & \text{toss } i \text{ is a head} \\ 0 & \text{otherwise,} \end{cases}$$

which gave $\mathbb{E}[S_N] = \frac{N}{2}$. After the shift, we have $\mathbb{E}[S_N] = 0$ and our event of interest becomes

$$\sum_{i=1}^{N} Y_i \geq \frac{N}{2}.$$

Applying Hoeffding's inequality, we see that

$$\begin{aligned} P\left(\sum_{i=1}^{N} Y_i \geq \frac{N}{2}\right) &\leq \exp\left(-\frac{N^2/4}{2N}\right) \\ &= e^{-N/8}, \end{aligned}$$

which matches the result from our CLT approximation.

A few useful properties of sub-Gaussian RVs are given below. These reinforce the idea that the parameter σ^2 acts as a proxy for variance.

If X_1, X_2 are sub-Gaussian RVs with parameters σ_1^2 and σ_2^2, then

- cX_i is sub-Gaussian with parameter $c^2\sigma_i^2$ for all $c \in \mathbb{R}$
- $X_1 + X_2$ is sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$.

By the above, we see that the sample mean $M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ is σ^2/N-sub-Gaussian if the X_i's are independent sub-Gaussian RVs with parameter σ^2.