CSC 696H: Topics in Bandits and Reinforcement Learning Theory **Fall 2024**

Lecture 2: Basic Probability Tools - 8/24/23

Lecturer: Chicheng Zhang *Scribe: Chenyi Wang*

Outline:

1. Concentration inequalities
2. Generalization in supervised machine learning (ML)

1 Concentration Inequalities (Concentration of Measure)

Concentration inequalities are a set of important tools in establishing and measuring performance guarantees in RL. Let's consider a simple example of learning and estimation: given a biased coin represented by a Bernoulli random variable,

$$X = \begin{cases} 1 & \text{heads (w.p. } p\text{)}, \\ 0 & \text{tails.} \end{cases}$$

We have the coin at hand, so we are free to flip it (conduct experiments) and observe outcomes. We would like to estimate its bias p. Moreover, we also want to establish an error bound on the estimate, so we can quantitatively know how accurate our estimate is and have more confidence in face of uncertainties. To put into more context, in previous restaurant choice example, given a restaurant, X here can indicate whether we are satisfied with its service in a visit to it.

Suppose we observe the outcomes of n independent flips, represented by random variables (r.v.'s) $X_1, ..., X_n$. We can say that $X_1, \ldots, X_n$ are drawn independently and identically distributed (iid) from the distribution. An equivalent way of stating our goal is to provide an estimate of $p = \mathbb{P}(X = 1) = \mathbb{E}[X]$, with uncertainty quantification (error bound). To this end, we consider the standard sample mean estimator $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Question: How accurate is this estimate?

Generalizing this to general r.v.'s, we have the following setup:

- Given $X_1, ..., X_n$ iid random variables (RVs) from distribution D, define the mean $\mu = \mathbb{E}_{X\sim D}[X]$, and the variance $\sigma^2 = \mathrm{Var}_{X\sim D}[X]$. We would like to estimate how close $\bar{X}_n$ is to μ.
- Note that it is always possible that we are unlucky and draw unrepresentative examples, causing $\bar{X}_n$ to deviate from μ by a lot. Thus the best we can hope for is for $\bar{X}_n$ to be close to μ with high probability.

  Thus, we would like to establish a probabilistic statement like $\mathbb{P}\left(|\bar{X}_n - \mu| \geq f(n, \delta)\right) \leq \delta, \forall \delta > 0$, for appropriate choice of function f.
- Graphically, under the probability density function (PDF) of $\bar{X}_n$, we would like that the total probability of our estimator (the mean μ defined above) lying outside of $f(n, \delta)$ on either side of μ (a.k.a. tail probability) is less than or equal to δ.
- The question becomes one of finding valid $f(n, \delta)$ functions.

Question: What concentration inequalities may be appropriate?

Idea 1: Chebyshev's Inequality:

For any random variable Y and any $\epsilon > 0$, $\mathbb{P}(|Y - \mathbb{E}Y| \geq \epsilon) \leq \frac{Var(Y)}{\epsilon^2}$ by Chebyshev's inequality. Let $Y = \bar{X}_n$, Can Chebyshev's inequality give the probability statement we want? To this end, we need to first check if is it true that $\mathbb{E}[Y] = \mu$. We can verify that:

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} x_i] \\
&= \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} x_i] && \text{(Linearity property of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[x_i] \\
&= \mu.
\end{aligned}$$

Similarly, we want to obtain $\mathrm{Var}(Y)$:

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} x_i) \\
&= \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} x_i) && \text{(Note that } \mathrm{Var}(cZ) = c^2\,\mathrm{Var}(Z)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(x_i) && \text{(Covariance is zero due to iid)} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Plugging these into Chebyshev's inequality gives:

$$\forall \epsilon > 0, \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Let's set ϵ s.t. $\frac{\sigma^2}{n\epsilon^2} = \delta$. Through a little algebra, we have $\epsilon = \sqrt{\frac{\sigma^2}{n\delta}}$, i.e. $f(n, \delta) = \sqrt{\frac{\sigma^2}{n\delta}}$, which gives us our function for the probability interval in the setup formulation.

Why does this make intuitive sense?

Recall the Central Limit Theorem (CLT): $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{approx}{\sim} N(0,1)$. One nice property we know about the standard normal variable is that it has a very light tail. Thus, we can expect with high probability (w.h.p.) that $\left|\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right| \leq const$, which implies $|\bar{X}_n - \mu| \leq const \cdot \frac{\sigma}{\sqrt{n}}$.

Note that, the dependence on n, σ^2 is tight, while the dependence on δ is $\sqrt{1/\delta}$. This can be singificantly loose in many applications (e.g., when we establish generalization bounds in supervised learning, as in our next lecture). Fortunately, it can be significantly sharpened to $\sqrt{\ln \frac{1}{\delta}}$ with very mild assumptions, which is our next goal. To this end, we will aim for concentration inequalities like

$$\mathbb{P}\left(\bar{X}_n - \mu| \geq \epsilon\right) \leq \exp(-n\epsilon^2). \tag{1}$$

Idea 2: Chernoff Method (using Moment Generating Function)

Definition 1. *The moment generating function (MGF) of a random variable X is given by*

$$\begin{aligned}\phi_X(\lambda) &= \mathbb{E}[e^{\lambda X}] \\ &= \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(\lambda X)^i}{i!}\right] \\ &= \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\mathbb{E}[X^i]\end{aligned}$$

Definition 2. *A random variable X is sub-Gaussian with variance proxy b^2 (a.k.a. b^2-SG), if* $\forall\lambda$, $\phi_{X-\mathbb{E}X}(\lambda) = \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq e^{\frac{b^2\lambda^2}{2}}$.

Some useful facts:

1. Gaussian random vairables are also sub-Gaussian because $X \sim N(\mu, b^2) \Rightarrow \phi_{X-\mu}(\lambda) = e^{\frac{b^2\lambda^2}{2}}$, hence the name "SG" (MGF takes similar form, but is dominated by Gaussian)
2. X is bounded in [A,B] $\Rightarrow$ X is $\frac{(B-A)^2}{4}$-SG (highly non-trivial, see reading).
3. X is b^2-SG $\Rightarrow aX$ is a^2b^2-SG (Proof as an exercise).
4. X is b^2-SG, Y is c^2-SG, and X, Y are independent $\Rightarrow X + Y$ is $(b^2 + c^2)$-SG (Proof as exercise).
5. If X is b^2-SG, then $\sigma^2 = Var(X) \leq b^2$. For some r.v. X, it can be the case that $\sigma << b$.

Note that properties 3 and 4 show the similarities between variance proxy and variance. Property 5 says vairance is the smallest possible variance proxy and the gap between variance and variance proxy can be arbitrarily large (however, it is exactly the same when the random variable is Gaussian). **The intuition behind is that SG random variables are those whose concentration behavior are equal to or better than Gaussian.**

Next, let's introduce the foundational Hoeffding's inequality, which states the following:

Theorem 3. *(Hoeffding's Inequality) Suppose $X_1, ..., X_n$ are iid and $\mathbb{E}[X_i] = \mu$. Then the following holds:*

1. *If all X_i's are b^2-SG, then $\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2exp(-\frac{n\epsilon^2}{2b^2})$.*
2. *If all X_i's are in $[A, B]$,then $\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2exp(-\frac{2n\epsilon^2}{(B-A)^2})$ (This follows from the previous item, as well as Property 2 of SG random variables).*

Implication: When X_i's are in $[A, B]$, we can choose ϵ s.t. $2exp(-\frac{2n\epsilon^2}{(B-A)^2}) = \delta \Leftrightarrow \epsilon = (B-A)\sqrt{\frac{ln\frac{2}{b}}{2n}}$. This gives a valid choice of $f(n, \delta) := (B-A)\sqrt{\frac{ln\frac{2}{b}}{2n}}$, that has a much better dependence on δ as promised.