CS234: Reinforcement Learning – Problem Session #5

Winter 2022-2023

Problem 1

One method for demonstrating provably-efficient exploration in multi-armed bandits (or reinforcement learning) is through a UCB-style analysis, where one constructs confidence sets around the reward (or value) of each action (or state-action pair) and shows that these sets shrink, eventually converging to ground truth over time. In this problem, we'll explore an alternative, general proof technique for establishing regret bounds using information theory, the branch of statistics dedicated to compression and communication. We will focus our attention on multi-armed bandits but these ideas can also be extended to the reinforcement-learning setting.

Let's start with a quick primer on information theory. Perhaps the most fundamental information-theoretic quantity is *entropy*, quantifying how much uncertainty there is in the outcome of a particular random variable. Let X be a discrete random variable taking values on a set $\mathcal{X}$ and with probability mass function $p(x) \in \Delta(\mathcal{X})$. Then, the entropy of X is defined as

$$\mathbb{H}(X) = \mathbb{E}\left[-\log\left(p(x)\right)\right] = -\sum_{x \in \mathcal{X}} p(x) \log\left(p(x)\right),$$

where the logarithm is in base 2 resulting in entropy measured in bits of information (using the natural logarithm with base e would be measured in nats).

For discrete random variables, we are guaranteed that the entropy is always non-negative $\mathbb{H}(X) \geq 0$, where the inequality is tight (that is, holds with equality) whenever X follows a Dirac delta distribution and places all probability mass on a single element of $\mathcal{X}$ (you may want to take a moment and convince yourself that this makes sense for a measure of uncertainty).

(a) Using Jensen's inequality, show that $\mathbb{H}(X) \leq \log\left(|\mathcal{X}|\right)$.

Solution:

$$\mathbb{H}(X) = \mathbb{E}\left[-\log\left(p(x)\right)\right] = \mathbb{E}\left[\log\left(\frac{1}{p(x)}\right)\right] \leq \log\left(\mathbb{E}\left[\frac{1}{p(x)}\right]\right) = \log\left(\sum_{x \in \mathcal{X}} p(x) \frac{1}{p(x)}\right) = \log\left(|\mathcal{X}|\right).$$

(b) Give a distribution $p(x)$ for X so that the previous inequality is tight.

Solution: Take $p(x) = \frac{1}{|\mathcal{X}|}$, $\forall x \in \mathcal{X}$ and observe that

$$\mathbb{H}(X) = \mathbb{E}\left[-\log\left(p(x)\right)\right] = \mathbb{E}\left[\log\left(\frac{1}{p(x)}\right)\right] = \mathbb{E}\left[\log\left(\frac{1}{\frac{1}{|\mathcal{X}|}}\right)\right] = \mathbb{E}\left[\log\left(|\mathcal{X}|\right)\right] = \log\left(|\mathcal{X}|\right).$$

While the entropy conveys our uncertainty about a random variable, it is often useful to think about uncertainty when conditioning on a second random variable. For a discrete random variable Y taking values on a set $\mathcal{Y}$, the residual uncertainty in X after observing Y is quantified by the *conditional entropy*

$$\mathbb{H}(X \mid Y) = \mathbb{E}\left[-\log\left(p(x \mid y)\right)\right] = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\left(\frac{p(x,y)}{p(y)}\right) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(y)p(x \mid y)\log\left(p(x \mid y)\right).$$

Like regular entropy, the conditional entropy is always non-negative for discrete random variables $\mathbb{H}(X \mid Y) \geq 0$. If $\mathbb{H}(X)$ gives our uncertainty in X and $\mathbb{H}(X \mid Y)$ provides our uncertainty in X after observing Y, then we can naturally quantify the information gained about X from Y via the *mutual information* between them:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y).$$

Now that we have a way to quantify how much information one random variable conveys about another, let's see how these quantities can be applied to analyze the regret of a multi-armed bandit.

Consider a multi-armed bandit problem over $T \in \mathbb{N}$ time periods with a finite number of arms $\mathcal{A}$ ($|\mathcal{A}| < \infty$) and a (potentially stochastic) reward function $r : \mathcal{A} \to \{0, 1\}$. If we knew the reward function exactly, then there would be no uncertainty in the optimal arm (and finding it would require one call to `numpy.argmax`). Consequently, it is our uncertainty in the underlying rewards of the environment that drives our uncertainty in the optimal action $A^\star = \arg\max_{a\in\mathcal{A}} r(a)$. Meanwhile, in each time period $t \in [T]$, the agent's policy samples an action A_t to observe a reward $R_t = r(A_t)$ and the cumulative regret over all T time periods is given by

$$\textsc{Regret}(T) = \sum_{t=1}^{T}\left(r(A^\star) - r(A_t)\right).$$

For Bayesian algorithms, we are interested in controlling the Bayesian regret

$$\textsc{BayesRegret} = \mathbb{E}\left[\textsc{Regret}(T)\right],$$

where the expectation accounts for our prior beliefs in the rewards of the environment. Ultimately, in this problem, we will prove that Thompson sampling obeys the following information-theoretic Bayesian regret bound:

$$\textsc{BayesRegret}(T) \leq \sqrt{\frac{1}{2}|\mathcal{A}|\mathbb{H}(A^\star)T}.$$

Solution: This result was originally shown by Russo and Van Roy [2016].

(c) What is the best-case regret for a Thompson sampling agent? Give a prior distribution over the optimal action $A^\star$ that achieves this best-case regret.

Solution: By assumption, Thompson sampling operates with a well-specified prior; that is, the prior must place non-zero probability mass on the true optimal action. So, if the agent's prior over the optimal action $A^\star$ is a Dirac delta distribution centered on the true optimal action, then $\mathbb{H}(A^\star) = 0$ and the Bayesian regret of Thompson sampling is also 0.

(d) What is the worst-case regret for a Thompson sampling agent? Give a prior distribution over the optimal action $A^\star$ that achieves this worst-case regret.

Solution: In the worst case, the entropy of the discrete random variable $A^\star$ can be no worse than $\mathbb{H}(A^\star) \leq \log\left(|\mathcal{A}|\right)$, which is achieved by having a uniform random prior over all actions: $\mathbb{P}(A^\star = a) = \frac{1}{|\mathcal{A}|}$, $\forall a \in \mathcal{A}$.

For any time period $t \in [T]$, let $E_t = (A_t, R_t)$ denote the observed experience at that timestep so that $H_t = (E_1, E_2, \ldots, E_{t-1})$ is the random history of agent interactions with the environment observed **at the start** of time period t, prior to observing E_t.

(e) Using the tower property of expectation, show that $\textsc{BayesRegret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[r(A^\star) - r(A_t) \mid H_t\right]\right]$.

Define the ratio

$$\Gamma_t = \frac{\mathbb{E}\left[r(A^\star) - r(A_t) \mid H_t\right]^2}{\mathbb{I}_t(A^\star; E_t)}, \qquad \forall t \in [T],$$

where the numerator is the squared expected regret in time period t and the denominator is the information gained about the optimal action $A^\star$ from the observed experience E_t. Note that, just like the numerator is conditioned on the current random history H_t, we use the t subscript in I_t to denote that information gain is also conditioned on H_t.

(f) Suppose we know that, for all time periods $t \in [T]$, $\Gamma_t \leq \overline{\Gamma}$ for some numerical constant $\overline{\Gamma} < \infty$. Then, show that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[r(A^\star) - r(A_t) \mid H_t\right]\right] \leq \sqrt{\overline{\Gamma}} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\mathbb{I}_t(A^\star; E_t)}\right]$.

Solution: By definition of the information ratio Γ_t in each time period $t \in [T]$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[r(A^\star) - r(A_t) \mid H_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\Gamma_t \cdot \mathbb{I}_t(A^\star; E_t)}\right] \leq \sqrt{\overline{\Gamma}} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\mathbb{I}_t(A^\star; E_t)}\right].$$

(g) Recall that the Cauchy-Schwarz inequality says that for any two vectors $u, v \in \mathbb{R}^d$,

$$\left(\sum_{i=1}^{d} u_i v_i\right)^2 \leq \left(\sum_{i=1}^{d} u_i^2\right)\left(\sum_{i=1}^{d} v_i^2\right).$$

Using the Cauchy-Schwarz inequality and Jensen's inequality, show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\mathbb{I}_t(A^\star; E_t)}\right] \leq \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)\right]}.$$

Solution: Taking $u_t = 1$ and $v_t = \sqrt{\mathbb{I}_t(A^\star; E_t)}$ for the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\mathbb{I}_t(A^\star; E_t)}\right] \leq \mathbb{E}\left[\sqrt{T \cdot \sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)}\right] \leq \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)\right]}$$

As the agent progresses over time, not all experiences will carry new information about the optimal action $A^\star$, given everything that has already been seen up to that point. The *conditional mutual information* helps quantify the information gain between two random variables given the context of a third random variable Z:

$$\mathbb{I}(X; Y \mid Z) = \mathbb{H}(X \mid Z) - \mathbb{H}(X \mid Y, Z).$$

For our purposes, while $\mathbb{I}_t(A^\star; E_t)$ quantifies information given the random history experienced by the agent so far, we can take an expectation to average over all possible agent histories and get the conditional mutual information

$$\mathbb{E}\left[\mathbb{I}_t(A^\star; E_t)\right] = \mathbb{I}(A^\star; E_t \mid H_t).$$

(h) Show that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)\right] = \sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid H_t)$.

Solution: Applying the identity provided above after applying linearity of expectation immediately yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{I}_t(A^\star; E_t)\right] \sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid H_t).$$

One very useful property of mutual information is that it allows us to decompose the information gained from several random variables. For any sequence of $K \in \mathbb{N}$ random variables $Z_1, \ldots, Z_K$, the *chain rule of mutual information* says that $\mathbb{I}(X; Z_1, \ldots, Z_K) = \sum_{k=1}^{K} \mathbb{I}(X; Z_k \mid Z_1, \ldots, Z_{k-1})$.

(i) Using the chain rule of mutual information, show that $\sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid H_t) \leq \mathbb{H}(A^\star)$.

Solution: Directly applying the chain rule of mutual information with the history $H_t = (E_1, E_2, \ldots, E_{t-1})$ yields

$$\sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid H_t) = \sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid E_1, E_2, \ldots, E_{t-1}) = \mathbb{I}(A^\star; E_1, \ldots, E_T) = \mathbb{I}(A^\star; H_T) = \mathbb{H}(A^\star) - \underbrace{\mathbb{H}(A^\star \mid H_T)}_{\geq 0} \leq \mathbb{H}(A^\star).$$

(j) A fact (that you may use without proof) is that Thompson sampling applied to a multi-armed bandit problem like the one described above has $\Gamma_t \leq \frac{1}{2}|\mathcal{A}|$, for all time periods $t \in [T]$. Show that Thompson sampling obeys the following Bayesian regret bound:

$$\textsc{BayesRegret}(T) \leq \sqrt{\frac{1}{2}|\mathcal{A}|\mathbb{H}(A^\star)T}.$$

Solution: Putting all the previous parts together, we have

$$\begin{aligned}
\textsc{BayesRegret}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[r(A^\star) - r(A_t) \mid H_t\right]\right] \\
&\leq \sqrt{\overline{\Gamma}} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\mathbb{I}_t(A^\star; E_t)}\right] \\
&\leq \sqrt{\overline{\Gamma} T \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t(A^\star; E_t)\right]} \\
&= \sqrt{\overline{\Gamma} T \cdot \sum_{t=1}^{T} \mathbb{I}(A^\star; E_t \mid H_t)} \\
&\leq \sqrt{\overline{\Gamma} T \mathbb{H}(A^\star)} \\
&= \sqrt{\frac{1}{2}|\mathcal{A}|\mathbb{H}(A^\star)T}.
\end{aligned}$$

References

Daniel Russo and Benjamin Van Roy. An Information-Theoretic Analysis of Thompson Sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.