STATS 710 – Seq. Dec. Making with mHealth Applications Nov 03, 2016

Lecture 15: UCB Finite-Horizon Reinforcement Learning

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Nick Seewald*

1 Recap from Lecture 14

Consider the setting in which we observe M episodes, each with finite horizon T, and wish to learn an optimal policy π^*. We saw that π^* is vector-valued, $\pi^* = \left(\pi_0^*, \ldots, \pi_{T-1}^*\right)^\top$, and

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}}\, V_\pi^T(s),$$

where

$$V_\pi^T(s) = \mathbb{E}_{A \sim \pi}\left[\sum_{t=0}^{T} R_{j,t} \mid S_{j,0} = s\right]$$

for an episode $j = 1, \ldots, M$. Recall that the state-action T-tuples

$$\left\{S_{j,0}, \left\{S_{j,1}^{a_0}\right\}_{a_0 \in \mathcal{A}}, \left\{S_{j,2}^{a_1}\right\}_{a_1 \in \mathcal{A}}, \ldots, \left\{S_{j,T}^{a_{T-1}}\right\}_{a_{T-1} \in \mathcal{A}}\right\} \quad j = 1, \ldots, M$$

are i.i.d., so explicitly denoting dependence on the index j is not strictly required.

In Lecture 14, we saw that

$$\pi_{T-t}^*(s) := \underset{a}{\operatorname{argmax}}\, r_a(s) + \sum_{s'} p_a(s, s') V_{\pi_{T-t+1}^*}^{t-1}(s'),$$

where $V_{\pi_{T-t}^*}^t(s) := \max_a r_a(s) + \sum_{s'} p_a(s, s') V_{\pi_{T-t+1}^*}^{t-1}(s')$.

2 Auer's UCB Finite-Horizon Reinforcement Learning

We follow Auer and Ortner and introduce a UCB-type algorithm for the reinforcement learning problem [AO05]. We make several simplifications. First, we restrict the within-episode horizon to $T = 2$, and consider binary state and action spaces, $|\mathcal{A}| = |\mathcal{S}| = 2$. Let M be the number of episodes. Denote by $\pi^m = (\pi_0^m, \pi_1^m)^\top$ the policy learned over the past m episodes which is applied in the $m + 1^{\text{th}}$ episode. π^m is a function of the history through the m prior episodes, $\mathcal{H}_{m+1} = \{S_{j,0}, A_{j,0}, S_{j,1}, A_{j,1}, S_{j,2} : j \leq m\}$.

Auer's modified UCB algorithm is of the following form:

for $j = 1$ **to** M **do**
  Learner sees $S_{j,0} = s_0$
  for $t = 0$ **to** $T - 1$ **do**
    Learner selects $A_{j,t}$ using π^{j-1}
    Learner receives $S_{j,t+1}$
    Learner forms reward $R_{j,t} = R\left(S_{j,t}, A_{j,t}, S_{j,t+1}\right)$
  end for
  Learner forms π^j from j prior episodes.

9: **end for**

Notice that the starting state s_0 in line 2 does not depend on j: the same starting state is used for each episode. The UCB part of the algorithm comes in line 8, when the learner forms π^j.

Typically, we require $j = M_0$ episodes are run to ensure that we see every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. This is non-trivial, and makes an assumption that transition probabilities are bounded away from 0 and 1. We ignore this in the following.

2.1 Quantities Needed for the Learning Algorithm

We introduce three statistics:

$$N_{s,a}^{m-1} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbb{1}\left\{S_{j,t} = s, A_{j,t} = a\right\} \tag{1}$$

is the number of observations of state-action pair (s, a),

$$\rho_{s,a}^{m-1} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbb{1}\left\{S_{j,t} = s, A_{j,t} = a\right\} R_{j,t} \tag{2}$$

keeps a running sum of the rewards for (s, a), and

$$p_{s,a,s'}^{m-1} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbb{1}\left\{S_{j,t} = s, A_{j,t} = a\right\} \mathbb{1}\left\{S_{j,t+1} = s'\right\} \tag{3}$$

is used to form transition probabilities. We now define the UCB versions of two quantities:

$$\tilde{r}_a^{m-1}(s) = \frac{\rho_{s,a}^{m-1}}{N_{s,a}^{m-1}} + \sqrt{\frac{c \log m}{N_{s,a}^{m-1}}} \tag{4}$$

is the UCB version of the conditional expectation of the immediate reward, and

$$\tilde{p}_a^{m-1}(s, s') = \frac{p_{s,a,s'}^{m-1}}{N_{s,a}^{m-1}} + \sqrt{\frac{c \log m}{N_{s,a}^{m-1}}} \tag{5}$$

is the naïve UCB version of transition probabilities. We say naïve since the $\tilde{p}_a^{m-1}(s, s')$ may not sum to 1. To account for this, Auer introduces the following transformation [AO05]:

$$\begin{aligned} \tilde{p}_a^{\pi_1, m-1}(s, 1) = {} & \mathbb{1}\left\{\tilde{r}_{\pi_1(1)}^{m-1}(1) \geq \tilde{r}_{\pi_1(0)}^{m-1}(0)\right\} \min\left\{1, \tilde{p}_a^{m-1}(s, 1)\right\} \\ & + \mathbb{1}\left\{\tilde{r}_{\pi_1(1)}^{m-1}(1) < \tilde{r}_{\pi_1(0)}^{m-1}(0)\right\} \left(1 - \min\left\{1, \tilde{p}_a^{m-1}(s, 0)\right\}\right). \end{aligned} \tag{6}$$

For a move from state s into state 0, we have $\tilde{p}_a^{\pi_1, m-1}(s, 0) = 1 - \tilde{p}_a^{\pi_1, m-1}(s, 1)$. The motivation behind these transformations is to privilege the state that looks best in the next stage. Notice that these probabilities are dependent on the policy π.

2.2 Regret Bound for the UCB Learning Algorithm

Recall our earlier notation $\pi_t^m(s)$, used to denote the policy at time t applied in the $m+1^{\text{th}}$ episode given state s. We define

$$\pi_1^{m-1}(s) = \operatorname*{argmax}_a \tilde{r}_a^{m-1}(s) = \operatorname*{argmax}_{\pi_1} \tilde{V}_\pi^{1,m-1}(s),$$

where $\tilde{V}_{\pi_1}^{1,m-1}(s) = \tilde{r}_{\pi_1(s)}^{m-1}(s)$, and

$$\pi_0^{m-1}(s) = \operatorname*{argmax}_a \tilde{r}_a^{m-1}(s) + \sum_{s'} \tilde{p}_a^{\pi_1^{m-1},m-1}(s,s')\tilde{r}_{\pi_1^{m-1}(s')}^{m-1}(s').$$

Define

$$\tilde{V}_\pi^{2,m-1}(s) = \tilde{r}_{\pi_0(s)}^{m-1}(s) + \sum_{s'} \tilde{p}_{\pi_0(s)}^{\pi_1,m-1}(s,s')\tilde{r}_{\pi_1(s')}^{m-1}(s').$$

The UCB algorithm uses $\pi_1^{m-1}(s)$ to select $A_{m,1}$ and $\pi_0^{m-1}(s)$ to select to $A_{m,0}$.

We define the expected regret at episode M as

$$\mathcal{R}_M\left(L,D,\Pi\right) = \mathbb{E}\left[\sum_{j=1}^M V_{\pi^*}^2(s_0) - \sum_{j=1}^M V_{\pi^{j-1}}^2(s_0)\right].$$

Susan proposes the following bound, which may be improved upon.

Proposition 1. *$\mathcal{R}_M(L,D,\Pi) = O\left(\log(M)/\min\left\{\Delta_1, \min_s \Delta_0(s)\right\}^2\right)$, where $\Delta_1 = \min_{\pi\neq\pi^*} V_{\pi^*}^2(s_0) - V_\pi^2(s_0)$ and $\Delta_0(s) = \min_{\pi_1\neq\pi_1^*} V_{\pi_1^*}^1(s) - V_{\pi_1}^1(s)$.*

Proof (first part plus sketch). We start by working with the regret. Specifically the term inside the expectation.

$$\begin{aligned}
&\sum_{m=1}^M V_{\pi^*}^2\left(s_0\right) - V_{\pi^{m-1}}^2\left(s_0\right) \\
=& \sum_{m=1}^M V_{\pi^*}^2(s_0) - V_{\pi_0^{m-1}\pi_1^*}^2(s_0) + V_{\pi_0^{m-1}\pi_1^*}^2(s_0) - V_{\pi^{m-1}}^2(s_0) \\
=& \sum_{m=1}^M \left(V_{\pi^*}^2(s_0) - V_{\pi_0^{m-1}\pi_1^*}^2(s_0)\right)\mathbb{1}\left\{A_{m,0} = \pi_0^{m-1}(s_0)\right\} \\
&+ \left(V_{\pi_0^{m-1}\pi_1^*}^2(s_0) - V_{\pi^{m-1}}^2(s_0)\right)\mathbb{1}\left\{A_{m,1} = \pi_1^{m-1}(S_{m,1})\right\} = *.
\end{aligned}$$

Notice that $V_{\pi^*}^2(s_0) \geq V_{\pi_0\pi_1^*}(s_0)$ for all π_0 and $V_{\pi_0\pi_1^*}^2(s_0) \geq V_{\pi_0\pi_1}^2(s_0)$ for all π_0, π_1. Now, define

$$\tilde{\Delta}_0 = \max_{\pi_0\neq\pi_0^*} V_{\pi^*}^2(s_0) - V_{\pi_0\pi_1^*}^2(s_0)$$

and

$$\tilde{\Delta}_1 = \max_{\pi_1\neq\pi_1^*,\pi_0} V_{\pi_0\pi_1^*}^2(s_0) - V_{\pi_0\pi_1}^2(s_0)$$

so we have

$$
\begin{aligned}
* &\leq \tilde{\Delta}_0 \sum_{m=1}^{M} \mathbb{1}\left\{A_{m,0} = \pi_0^{m-1}(s_0)\right\} + \tilde{\Delta}_1 \sum_{m=1}^{M} \mathbb{1}\left\{A_{m,1} = \pi_1^{m-1}(S_{m,1})\right\} \\
&\leq \tilde{\Delta}_0 \sum_{\pi_0 \neq \pi_0^*} \sum_{m=1}^{M} \mathbb{1}\left\{A_{m,0} = \pi_0(s_0)\right\} + \tilde{\Delta}_1 \sum_{\pi_1 \neq \pi_1^*} \sum_{m=1}^{M} \sum_{s} \mathbb{1}\left\{A_{m,1} = \pi_1(s), S_{m,1} = s\right\} \\
&\leq \tilde{\Delta}_0 \sum_{\pi_0 \neq \pi_0^*} \left(\ell_0 + \sum_{m=1}^{M} \mathbb{1}\left\{ A_{m,0} = \pi_0(s_0) \cap N_{s_0 \pi_0(s_0)}^{m-1} \geq \ell_0 \right\} \right) \\
&\quad + \tilde{\Delta}_1 \sum_{\pi_1 \neq \pi_1^*} \sum_{s} \left(\ell_1 + \sum_{m=1}^{M} \mathbb{1}\left\{ A_{m,1} = \pi_1(s) \cap N_{s,\pi_1(s)}^{m-1} \geq \ell_1 \right\} \right).
\end{aligned}
$$

The last inequality holds because, for example,

$$
\begin{aligned}
\sum_{m=1}^{M} \mathbb{1}\left\{A_{m,0} = \pi_0(s_0)\right\} = &\sum_{m=1}^{M} \mathbb{1}\left\{ A_{m,0} = \pi_0(s_0) \cap N_{s_0,\pi_0(s_0)}^{m-1} \geq \ell_0 \right\} \\
&+ \sum_{m=1}^{M} \mathbb{1}\left\{ A_{m,0} = \pi_0(s_0) \cap N_{s_0,\pi_0(s_0)}^{m-1} < \ell_0 \right\},
\end{aligned}
$$

and the second sum is at most ℓ_0.

The crux of this proof, which will be finished in a subsequent lecture, is ensuring we have enough data, i.e., enough observations of each state-action pair. We take $\ell_1 \simeq 8 \log M/\Delta_1$ and $\ell_0 \simeq 8 \log M/\Delta_0$. For the sums over m indicators, we take the expectation and show that they sum to a constant. □

Note: The theorem and proof will be slightly changed in the subsequent lecture (lecture 17) due to new insights.

References

[AO05] Peter Auer and Ronald Ortner. Online regret bounds for a new reinforcement learning algorithm. In Michael Zillich and Markus Vincze, editors, *1st Austrian Cognitive Vision Workshop*, pages 35–42. Austrian Computer Society, 2005.