STATS 710 – Sequential Decision Making with mHealth Applications Nov 10, 2016

Lecture 17: Reinforcement Learning with UCB: Redux

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Ian Fox*

1 Recap

The presentation is roughly (but not exactly) based on [AO05].

We define expected regret as follows:

$$R_M(L, D, \Pi) = E[\sum_{j=1}^{M} V^2_{\pi^*}(S_0) - \sum_{j=1}^{M} V^2_{\pi^{j-1}}(S_0)]$$

The RL version of the UCB algorithm in a simplified setting is as follows:

Algorithm 1 Simple RL UCB

```
for j=1 to M do
    Learner sees S_{j,0} = S_0
    for t=0,1 do
        Learner selects A_{j,k} using π^{j−1}
        Learner sees S_{j,t+1}
        Learner receives R_{j,t} = R(S_{j,t}, A_{j,t}, S_{j,t+1})
    end for
    Learner forms π^j
end for
```

We define the following statistics:

- $N^{m-1}_{s,a} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbf{1}[S_{j,t} = s, A_{j,t} = a]$
- $\rho^{m-1}_{s,a} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbf{1}[S_{j,t} = s, A_{j,t} = a] R_{j,t}$
- $p^{m-1}_{s,a,s'} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} \mathbf{1}[S_{j,t} = s, A_{j,t} = a, S_{j,t+1} = s']$
- UCB average reward: $\tilde{r}^{m-1}_a(s) = \frac{\rho^{m-1}_{s,a}}{N^{m-1}_{s,a}} + \sqrt{\frac{c \log(m)}{N^{m-1}_{s,a}}}$
- UCB transition probability: $\tilde{p}^{m-1}_a(s, s') = \frac{p^{m-1}_{s,a,s'}}{N^{m-1}_{s,a}} + \sqrt{\frac{c \log(m)}{N^{m-1}_{s,a}}}$

The UCB algorithm forms two policies for each j, one per time step. The policies are generated as follows:

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

where $\tilde{V}_{\pi_1}^{1,m-1}(s)$ is the value function when 1 time step remains, similarly $\tilde{V}_{\pi_0,\pi_1}^{2,m-1}(s)$ is the value when 2 time steps remain.

Optimal policies are defined:

$$\pi_1^*(s) = \operatorname*{argmax}_a E[R|S=s, A=a]$$
$$E[R|S=s, A=a] \triangleq V_{\pi^*}(s)$$
$$\pi_0^*(s) = \operatorname*{argmax}_a E[R_0|S_0=s, A_0=a] + E[V_{\pi^*(s_1)}^1(s_1)|S_0=s, A_0=a]$$
$$= \operatorname*{argmax}_a r_a(s) + \sum_{s'} p_a(s,s')V_{\pi^*(s)}^1(s')$$

Our goal is to calculate π_0^*, π_1^* for which $E_{\pi_0,\pi_1}[R_0 + R_1]$ is maximal. Note:

$$V_{\pi_0,\pi_1}^2(s) = r_{\pi_0(s)}(s) + \sum_{s'} p_{\pi_0(s)}(s,s')V_{\pi_1(s')}^1$$

2 UCB Regret

Theorem 1. $R_M(L, D, \Pi) = O(\frac{\log(M)}{\Delta})$ *where the gap parameter* $\Delta = \min_{\pi \neq \pi^*} V_{\pi^*}^2(s_0) - V_\pi^2(s_0)$

Proof: First, note that we can approach this RL problem like a simple MAB, treating each policy as an arm. Using this approach we achieve this regret bound, but with a constant proportional to $|A|^{|S|}$. The author of the paper claims they proved a bound polynomial in $|A|, |S|$, but did not explicitly state the proof. The proof given here doesn't achieve this polynomial bound.

The regret suffered is

$$\sum_{m=1}^M V_{\pi^*}^2(s_0) - V_{\pi^{m-1}}^2(s_0)$$
$$= \sum_{m=1}^M (V_{\pi^*}^2(s_0) - V_{\pi^{m-1}}^2(s_0))\mathbf{1}[A_{m,0} = \pi_0^{m-1}, A_{m,1} = \pi_1^{m-1}(s_{m-1})]\mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0]$$
$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^M \mathbf{1}[A_{m,0} = \pi_0^{m-1}, A_{m-1} = \pi_1^{m-1}(s_{m-1})]\mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0]$$
$$*\mathbf{1}[N_{s_0,\pi_0(s_0)}^{m-1} \geq l_{\pi_0}, N_{\pi_1}^{m-1} \geq l_{\pi_1}] + \sum_{\pi \neq \pi^*} \Delta_\pi l$$

where

$$N_{\pi_1}^{m-1} = \sum_{j=1}^{m-1} \mathbf{1}[\Delta_{j,1} = \pi_1(S_{j,1})]$$

$$= \sum_{j=1}^{m-1} \mathbf{1}[A_{j,1} = \pi_1(S_{j,1}), S_{j,1} = 1] + \mathbf{1}[A_{j,1} = \pi_1(S_{j,1}), S_{j,1} = 0]$$

$$= N_{1,\pi_1(0)}^{m-1} + N_{0,\pi(0)}^{m-1}$$

To put some words to this, we're following the proof strategy for the MAB UCB regret bound. We're assuming that we mess up, using suboptimal policy π, at least l_π times. We've manipulated the regret term to account for this. We define

$$l_\pi \triangleq \frac{36 * \log(M)}{\Delta_\pi^2}$$

. Now we'll use these goofs to control the probability of messing up in the future, by controlling:

$$\sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0]\mathbf{1}[N_{s_0,\pi_0(s_0)}^{m-1} \geq l_{\pi_0}, N_{\pi_1}^{m-1} \geq l_{\pi_1}]$$

We'll focus on $\mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0]$

$$\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0)$$

$$= \tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - V_\pi^2(s_0) - (\tilde{V}_{\pi^*}^{2,m-1}(s_0) - V_{\pi^*}^2(s_0)) + V_\pi^2(s_0) - V_{\pi^*}^2(s_0)$$

so equivalently

$$\mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0)]$$

$$= \mathbf{1}[\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - V_\pi^2(s_0) - (\tilde{V}_{\pi^*}^{2,m-1}(s_0) - V_{\pi^*}^2(s_0)) + V_\pi^2(s_0) - V_{\pi^*}^2(s_0) \geq 0]$$

Recall that

$$\mathbf{1}[A + B + C \geq 0] \leq \mathbf{1}[A \geq 0] + \mathbf{1}[B \geq 0] + \mathbf{1}[C \geq 0]$$

Let's focus on $\tilde{V}_{\pi^{m-1}}^{2,m-1}(s_0) - V_\pi^2(s_0)$ and write things out according to our definitions (warning:

things are getting messy):

$$
\begin{aligned}
&\tilde{V}^{2,m-1}_{\pi^{m-1}}(s_0) - V^2_{\pi}(s_0) = \\
&\frac{\rho^{m-1}_{s_0,\pi_0(s_0)}}{N_{s_0,\pi_0(s_0)}} - r_{\pi_0(s_0)} + \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}} \\
&+\sum_s p_{\pi_0(s)}(s_0,s)(\frac{\rho_{s_1,\pi_1(s)}}{N_{s,\pi_1(s_1)}} - r_{\pi_1(s)}(s) + \sqrt{\frac{c\log(m)}{N_{s,\pi(s)}}}) \\
&+(\frac{\rho^{m-1}_{s_0,\pi_0(1)}}{N_{s_0,\pi_0}} - p_{\pi_0(s_0)}(s_0,1) + \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}})(\tilde{r}_{\pi_1(1)}(1) - \tilde{r}_{\pi_1(0)}(0))^+ \\
&+(\frac{\rho^{m-1}_{s_0,\pi_0(0)}}{N_{s_0,\pi_0}} - p_{\pi_0(s_0)}(s_0,0) + \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}})(\tilde{r}_{\pi_1(0)}(0) - \tilde{r}_{\pi_1(1)}(1))^+ \\
&= \\
&\frac{\rho^{m-1}_{s_0,\pi_0(s_0)}}{N_{s_0,\pi_0(s_0)}} - r_{\pi_0(s_0)} - \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}} \\
&+\sum_s p_{\pi_0(s)}(s_0,s)(\frac{\rho_{s_1,\pi_1(s)}}{N_{s,\pi_1(s_1)}} - r_{\pi_1(s)}(s) + \sqrt{\frac{c\log(m)}{N_{s,\pi(s)}}}) \\
&+(\frac{\rho^{m-1}_{s_0,\pi_0(1)}}{N_{s_0,\pi_0}} - p_{\pi_0(s_0)}(s_0,1) - \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}})(\tilde{r}_{\pi_1(1)}(1) - \tilde{r}_{\pi_1(0)}(0))^+ \\
&+(\frac{\rho^{m-1}_{s_0,\pi_0(0)}}{N_{s_0,\pi_0}} - p_{\pi_0(s_0)}(s_0,0) - \sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}})(\tilde{r}_{\pi_1(0)}(0) - \tilde{r}_{\pi_1(1)}(1))^+ \\
&+2\sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}} + 2\sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}}(\tilde{r}_{\pi_1(1)}(1) - \tilde{r}_{\pi_1(0)}(0))^+ + 2\sqrt{\frac{c\log(m)}{N_{s_0,\pi_0(s_0)}}}(\tilde{r}_{\pi_1(1)}(0) - \tilde{r}_{\pi_1(0)}(1))^+
\end{aligned}
$$

This is where we ran out of time. From here things get nice, except that we can't control the visitation rate (rare transitions), leading to the exponential blowup.

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.