MS&E338 Reinforcement Learning Lecture 14 - May 17, 2018

PSRL and UCRL

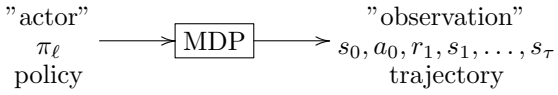
Lecturer: Ben Van Roy *Scribe: Isaac Faber, Ramon Iglesias, Travis Trammell*

The last few lectures have covered topics in general active learning. Today's topic will dive into the application of similar algorithms to MDPs and reinforcement learning. There are two techniques that will be explored in some detail.

1 PSRL

Posterior Sampling Reinforcement Learning (PSRL) is a method of learning that takes advantage of the posterior beliefs based on observed data.

$M = (S, A, R, P, p)$



During each episode, PSRL samples an MDP from the posterior and solves for the optimal policy. It then applies the policy throughout the episode and observe a trajectory, which is used to update the posterior. There is still some consideration about how you select prior distributions is this setting. One such approach is to use "uninformative" priors. We assume that $P_{s,a}$ are i.i.d. over (s, a) and $R_{s,a}$ are i.i.d. over (s, a).

1.1 Dirichlet

One popular type of prior/posterior pair is the Dirichlet/Categorical, which is a generalization of the Beta/Bernoulli distribution to allow for a range of discrete outcomes. The following are a list of characteristics:

-outcomes $1, 2, \ldots, M$
-Distribution over $(p_1, \ldots, p_M)$

$$\sum_{m=1}^{M} p_m = 1, \quad p_m \geq 0, \quad \forall m = 1, \ldots, M$$

-parameters $(\alpha_1, \ldots, \alpha_M)$, pdf $\propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots p_M^{\alpha_M - 1}$
-update $\alpha_m \leftarrow \alpha_m + 1$ if outcome m observed

$$\text{Prior } P_{s,a} \sim \text{Dirichlet } \big(\alpha_{s,a}(1), \alpha_{s,a}(2), \ldots, \alpha_{s,a}(|S|)\big)$$

This leads to easy application and updating where, with sufficient data observations, the distributions concentrate.

1.1.1 Prior Over Rewards

$$R_{s,a} \sim N\big(\overline{R}_{s,a}, \sigma_r^2\big)$$
$$\overline{R}_{s,a} \sim N\big(\mu_{s,a}, \sigma_{s,a}^2\big)$$

$R_{s,a}$ represents the rewards distribution. $\overline{R}_{s,a}$ is the mean reward and is not known. Priors are always generative models for the environment.

1.1.2 Update

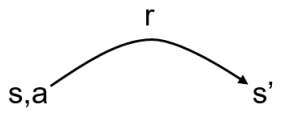


Figure 1: Updating Graphic

Here are the updating rules upon observing a transition (s, a, r, s').

$$\alpha_{s,a}(s') \longleftarrow \alpha_{s,a}(s') + 1$$

$$\mu_{s,a} \longleftarrow \frac{\frac{1}{\sigma^2_{s,a}}\mu_{s,a} + \frac{1}{\sigma^2_r} r}{\frac{1}{\sigma^2_{s,a}} + \frac{1}{\sigma^2_r}}$$

$$\sigma^2_{s,a} \longleftarrow \frac{1}{\frac{1}{\sigma^2_{s,a}} + \frac{1}{\sigma^2_r}}$$

Thus, we arrive at the following algorithm for PSRL.

Algorithm 1 PSRL

Start with Dirichlet and Gaussian priors
for $\ell = 1, 2, 3, \ldots$
Sample MDP $\hat{M}$ (sample $P_{s,a}$, $R_{s,a}$ $\forall(s, a)$)
Compute optimal policy $\hat{\pi}$
Apply $\hat{\pi}$ over one episode $\big(s_0, a_0, r_1, s_1, a_1, \ldots, s_\tau\big)$
for $t = 0, \ldots, \tau - 1$

$$\alpha_{s_t,a_t}(s_{t+1}) \longleftarrow \alpha_{s_t,a_t}(s_{t+1}) + 1$$

$$\mu_{s_t,a_t} \longleftarrow \frac{\frac{1}{\sigma^2_{s_t,a_t}}\mu_{s_t,a_t} + \frac{1}{\sigma^2_r} r_{t+1}}{\frac{1}{\sigma^2_{s_t,a_t}} + \frac{1}{\sigma^2_r}}$$

$$\sigma^2_{s_t,a_t} \longleftarrow \frac{1}{\frac{1}{\sigma^2_{s_t,a_t}} + \frac{1}{\sigma^2_r}}$$

*This is Thompson sampling for MDPs and works very well for accumulating rewards.

2 UCRL

Upper Confidence bound Reinforcement Learning (UCRL) is done by selecting optimistic policies consistent with some confidence set over the MDPs. In most cases the PSRL algorithm works much better than UCRL, however, it (UCRL) is quite popular in the literature. Often UCRL performance can be manipulated by selecting or engineering a tight confidence set. However, in the general case this algorithm has the same issues discussed in the previous lecture. Specifically, people often use box-shaped confidence sets instead of ellipsoidal ones for computational tractability while sacrificing statistical efficiency.

Algorithm 2 UCRL

Start with confidence set $\Theta^P_{s,a}$, $\Theta^R_{s,a}$ $\forall (s,a)$

for $\ell = 1, 2, 3, \ldots$

$$\hat{\pi} \in \operatorname*{argmax}_{\pi} \max_{\substack{P_{s,a} \in \Theta^P_{s,a} \\ R_{s,a} \in \Theta^R_{s,a}}} E\left[\sum_{t=0}^{\tau-1} r_{t+1} \,\middle|\, \pi, P, R\right]$$

Apply $\hat{\pi}$ over one episode $(s_0, a_0, r_1, s_1, a_1, \ldots, s_\tau)$

for $t = 0, \ldots, \tau - 1$

Update $\Theta^P_{s,a}$, $\Theta^R_{s,a}$ $\forall (s,a)$

To better understand the bounds on these algorithms it is useful to understand some related active learning concepts.

2.1 UCRL Confidence Sets

Here is a typical construction of confidence sets.

$$N(s,a) = \text{ visit count}$$
$$\hat{r}(s,a) = \text{ sample mean reward}$$
$$\hat{p}(s' \mid s,a) = \text{ sample estimate / empirical probability}$$

$$\Theta^{\overline{R}}_{s,a} = \left|\overline{R}_{s,a} - \hat{r}(s,a)\right| \le c_1 \sqrt{\frac{\log\left(|S||A|\ell/\delta\right)}{N(s,a)}}$$

$$\Theta^P_{s,a} = \left\|P_{s,a} - \hat{p}(\cdot \mid s,a)\right\|_1 \le c_2 \sqrt{\frac{|S| \log\left(2|A|\ell/\delta\right)}{N(s,a)}}$$

$$\delta = \text{ fudge factor}$$

There are several improvements to these confidence sets. One is KL-UCRL, where instead of 1 norm they use the KL divergence. This gives a more ellipsoidal set for each state action pair, but the confidence sets are still separate over state action pairs, which can be undesirable.

3 Regret Analysis

There are different ways for establishing regret bounds for Thompson sampling and UCB algorithms. Let's first briefly review the active learning framework.

$$x_t \longrightarrow \boxed{\theta} \longrightarrow y_t \longrightarrow r(y_t) = r_t$$

action	$x_t \in \mathcal{X}$
observations	$y_t \sim q_\theta(\cdot \mid x_t)$
reward	$r_t = r(y_t)$
prior	$\theta \sim p$

We define a shorthand

$$\overline{r}(x, \theta) = E\Big[r(y_t) \mid x_t = x, \theta\Big]$$

3.1 Upper Confidence Bound

Recall that UCB chooses action $\overline{x}_t$ based on upper confidence bounds U_t,

$$\overline{x}_t \in \operatorname*{argmax}_{x \in \mathcal{X}} U_t(x),$$

where U_t depends the prior p and past actions and observations x_1, y_1, x_2, y_2, $\ldots$, x_{t-1}, y_{t-1}.

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{T} \Big(\max_{x \in \mathcal{X}} \overline{r}(x, \theta) - \overline{r}(\overline{x}_t, \theta) \Big) \\
\overline{r}(x^*, \theta) - \overline{r}(\overline{x}_t, \theta) &= \overline{r}(x^*, \theta) - U_t(\overline{x}_t) + U_t(\overline{x}_t) - \overline{r}(\overline{x}_t, \theta) \\
&\leq \underbrace{\overline{r}(x^*, \theta) - U_t(x^*)\Big)}_{\substack{\text{pessimism} \\ \leq 0 \text{ with high probability}}} + \underbrace{\Big(U_t(\overline{x}_t) - \overline{r}(\overline{x}_t, \theta)\Big)}_{\substack{\text{width} \\ \to 0 \text{ at good rate}}}
\end{aligned}$$

The pessimism term is ≤ 0 with high probability if the confidence sets are chosen reasonably. The second term penalizes the width of the confidence sets. The major part in most analyses of UCB algorithms is showing that the width term goes to 0 at a desirable rate.

The next lecture will discuss some concrete results about Thompson sampling.