2019 Mathematics of Machine Learning Summer School Brunskill and Zanette: Reinforcement Learning, #2

1 Effect of Modeling Errors in Policy Evaluation [20 pts]

Consider deploying a Reinforcement Learning (RL) agent on an episodic MDP M with a horizon of H timesteps. The true MDP M is never revealed to the agent except for the state and action space (S and A, respectively) and the time horizon H. In other words, the agent knows neither the expected reward $r(s,a)$ nor the transition dynamics $P(s' \mid s,a)$ for a generic state-action pair (s,a).

As the agent explores the environment in different episodes it builds an empirical model of the world (often via maximum likelihood) which we denote by $\hat{M}$. After collecting sufficient data we would expect the empirical MDP to be similar to the true MDP.

Since our MDP is episodic, the value of a state depends on the time to termination of the episode. Thus, for a given stochastic policy π, we require a different value function for each timestep, which we denote $V_1, \dots, V_H$ in M and $\hat{V}_1, \dots, \hat{V}_H$ in $\hat{M}$. In particular, for $i = 1, \dots, H$, let

$$V_i(s) = \mathbb{E}\left[\sum_{t=i}^{H} r(s_t, a_t) \;\middle|\; s_i = s\right] \text{ and } \hat{V}_i(s) = \mathbb{E}\left[\sum_{t=i}^{H} \hat{r}(s_t, a_t) \;\middle|\; s_i = s\right],$$

where the expectation is taken under following policy π. Suppose we use the empirical MDP $\hat{M}$ instead of M to evaluate policy π. Assuming $\hat{V}_{H+1} = V_{H+1} = \vec{0}$ show that for all $i = 1, \dots, H$:

$$\hat{V}_i(s) - V_i(s) = \sum_{t=i}^{H} \mathbb{E}\left[\hat{r}(s_t, a_t) - r(s_t, a_t) + \sum_{s'} (\hat{P}(s' \mid s_t, a_t) - P(s' \mid s_t, a_t))\hat{V}_{t+1}(s') \;\middle|\; s_i = s\right].$$

In the above equality the expectation is defined with respect to the states encountered in true MDP M upon starting from s_i and following stochastic policy π.

This result relates the value of policy π on $\hat{M}$ and M using the expected trajectories on M which we can compute easily. If it holds that $\hat{r}$ and $\hat{P}$ are close to r and P then this result can be used to conclude that the empirical value function $\hat{V}$ is also close to the true one V.

2 Expected Regret Bounds (35pts)

Assume a reinforcement learning algorithm A for discounted infinite-horizon MDPs has expected regret

$$\mathbb{E}_* \left[\sum_{t=1}^{T} r_t \right] - \mathbb{E}_A \left[\sum_{t=1}^{T} r_t \right] = f(T)$$

for all $T > 0$, where $\mathbb{E}_*$ is over the probability distribution with respect to the optimal policy π_* and $\mathbb{E}_A$ is the expectation with respect to the algorithm's behavior. We assume that $\gamma \in [0, 1)$ is the discount factor and that rewards are normalized, i.e., $r_t \in [0, 1]$.

(a) Let π be an arbitrary policy or algorithm. Show that for any $\varepsilon' > 0$ and $T' \geq \log_{\frac{1}{\gamma}} \frac{H}{\varepsilon'}$ where $H = 1/(1-\gamma)$, we have

$$\left| V_\pi(s) - \sum_{t=1}^{T'} \gamma^{t-1} \mathbb{E}_\pi[r_t | s_1 = s] \right| \leq \varepsilon' \,, \text{ for all state } s.$$

Note V_π is the value function associated with π and $\mathbb{E}_\pi$ is the expectation with respect to the randomization of π.

(b) From the regret guarantee of algorithm A and Part a), it follows that for any $\varepsilon' > 0$ and $T' \geq \log_{\frac{1}{\gamma}} \frac{H}{\varepsilon'}$, we have

$$\mathbb{E}_*[V_*(s_{T+1})] - \mathbb{E}_A[V_A(s_{T+1})] \leq f(T' + T) - f(T) + 2\varepsilon' \,, \text{ for } T > 0,$$

where V_A is the value function of the (possibly nonstationary) policy that algorithm A follows. Assume now $f(T) = \sqrt{T}$. Show that for any $\varepsilon > 0$ and $t \geq 1 + \frac{1}{\varepsilon^2} \left(\log_{\frac{1}{\gamma}} \frac{4H}{\varepsilon} \right)^2$, we have

$$\mathbb{E}_*[V_*(s_t)] - \mathbb{E}_A[V_A(s_t)] \leq \varepsilon \,.$$

Hint: It may be helpful to set ε' to be some function of ε and choose an appropriate value of T'.