LSPI CUSTOMIZED FOR OPTIMAL EXERCISE OF AMERICAN OPTIONS

ASHWIN RAO (STANFORD UNIVERSITY)

Let us first consider the general LSPI algorithm. In each iteration, we are given:

$$Q(s,a) = \sum_i w_i \cdot \phi_i(s,a)$$

and deterministic policy π (known as the target policy for that iteration) is given by:

$$\pi(s) = \arg\max_a Q(s,a)$$

The goal in the iteration is to solve for weights $\{w'_i\}$ such that we minimize

$$\sum_{s,a,r,s'} (r + \gamma \cdot Q'(s', \pi(s')) - Q'(s,a))^2$$

over a data-set comprising of a sequence of 4-tuples (s,a,r,s') with:

$$Q'(s,a) = \sum_i w'_i \cdot \phi_i(s,a)$$
$$Q'(s',\pi(s')) = \sum_i w'_i \cdot \phi_i(s',\pi(s'))$$

Therefore, we solve for $\{w'_i\}$ that minimizes:

$$\sum_{s,a,r,s'} (r + \gamma \cdot Q'(s', \pi(s')) - \sum_i w'_i \cdot \phi_i(s,a))^2$$

So we calculate the gradient of the above expression with respect to $\{w'_j\}$ and set it to 0 (semi-gradient). This gives:

$$\sum_{s,a,r,s'} (r + \gamma \cdot Q'(s', \pi(s')) - \sum_i w'_i \cdot \phi_i(s,a)) \cdot \phi_j(s,a) = 0 \text{ for all } j \tag{1}$$

Now we customize LSPI to the problem of Optimal Exercise of American Options. We have two actions: $a = c$ (continue the American Option) and $a = e$ (exercise the American Option). We consider a function approximation for $Q'(s,a)$ only for the case of $a = c$ since we know the exact expression for the case of $a = e$, given by the Payoff function (call it $g : \mathcal{S} \to \mathbb{R}$). Therefore,

$$Q'(s,e) = g(s)$$

We write the function approximation for $Q'(s,c)$ as:

$$Q'(s,c) = \sum_i w'_i \cdot \phi_i(s,c) = \sum_i w'_i \cdot x_i(s)$$

for feature functions $x_i : \mathcal{S} \to \mathbb{R}$ (i.e., functions of only state and not action).

Since we are learning the Q-Value function for only $a = c$, our experience policy μ is a constant function $\mu(s) = c$. Also, for American Options, the reward for $a = c$ is 0. Thus, when considering the 4-tuples (s, a, r, s') for training experience, we always have $a = c$ and $r = 0$. So the 4-tuples are $(s, c, 0, s')$ and so, we might as well simply consider 2-tuples (s, s') for training experience (since $a = c$ and $r = 0$ are locked).

Now consider 2 cases to customize Equation (1) to the problem of Optimal Exercise of American Options.

Case 1: If $\pi(s') = c$ (this happens when $\sum_i w_i \cdot x_i(s') \geq g(s')$), then Equation (1) reduces to:

$$\sum_{s,s'} (\gamma \cdot \sum_i w'_i \cdot x_i(s') - \sum_i w'_i \cdot x_i(s)) \cdot x_j(s) = 0 \text{ for all } j$$

$$\Rightarrow \sum_i w'_i \cdot \sum_{s,s'} x_j(s) \cdot (x_i(s) - \gamma \cdot x_i(s')) = \sum_{s,s'} 0 \text{ for all } j \tag{2}$$

Case 2: If $\pi(s') = e$ (this happens when $g(s') > \sum_i w_i \cdot x_i(s')$), then Equation (1) reduces to:

$$\sum_{s,s'} (\gamma \cdot g(s') - \sum_i w'_i \cdot x_i(s)) \cdot x_j(s) = 0 \text{ for all } j$$

$$\Rightarrow \sum_i w'_i \cdot \sum_{s,s'} x_j(s) \cdot (x_i(s)) = \sum_{s,s'} x_j(s) \cdot \gamma \cdot g(s') \text{ for all } j \tag{3}$$

Equations (2) and (3) can be united with a common equation $\mathbf{A} \cdot \mathbf{w}' = \mathbf{b}$. The term $x_j(s) \cdot (x_i(s) - \gamma \cdot x_i(s'))$ from Equation (2) (for Case 1) and the term $x_j(s) \cdot (x_i(s))$ from Equation (3) (for Case 2) contributes to $\mathbf{A}$ for each (s, s') in the training experience data-set. The term 0 from Equation (2) (for Case 1) and the term $x_j(s) \cdot \gamma \cdot g(s')$ from Equation (3) (for Case 2) contributes to $\mathbf{b}$ for each (s, s') in the training experience data-set.