COMS 4771 Machine Learning (Spring 2020) Problem Set #4

Joseph High - jph2185@columbia.edu

April 23, 2020

Problem 1: Finding the value of a state under a policy

Proof.

First, using the law of total expectation and conditioning over $a \in \mathcal{A}$:

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_a \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a] \cdot \underbrace{P\left(a_t = a | s_t = s\right)}_{=\ \pi(a|s)}$$

$$= \sum_a \pi(a \mid s)\, \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a]$$

Again, using the law of total expectation, but now conditioning over s':

$$= \sum_a \pi(a \mid s) \sum_{s'} \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a, S_{t+1} = s'] \cdot \underbrace{P\left(S_{t+1} = s' \mid S_t = s, a_t = a\right)}_{=\ P(s'|s,a)}$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a, S_{t+1} = s']$$

Substituting in the definition of G_t:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Pulling out the first summand, R_{t+1}, from the infinite series:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[R_{t+1} + \sum_{k=2}^{\infty} \gamma^{k-1} R_{t+k}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Re-indexing the infinite series and factoring out a γ :

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[R_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Using the linearity property of expectation:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[R_{t+1} + \gamma\, \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s']$$

Again, using the linearity property of expectation:

$$\begin{aligned} = \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\Big[& \mathbb{E}_\pi[R_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] \\ & + \gamma\, \mathbb{E}_\pi[G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s']\Big] \end{aligned}$$

By the Markov property, $\mathbb{E}_\pi[G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] = \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']$. Applying this to the above:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\Big[\underbrace{\mathbb{E}_\pi[R_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s']}_{=\, R_a(s,s')} + \gamma\, \underbrace{\mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']}_{=\, v_\pi(s')}\Big]$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') + \gamma v_\pi(s')\big]$$

Hence,

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') + \gamma v_\pi(s')\big]$$

$\square$

Problem 2: Solving for a value function using linear algebra

In Problem 1 it was shown that

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \;+\; \gamma v_\pi(s')\big]$$

Applying the assumption that all transitions are deterministic, i.e., $P(s' \mid s, a) = \mathbb{1}\{s' = next(s, a)\}$, we get

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} \mathbb{1}\{s' = next(s, a)\} \cdot \big[R_a(s, s') \;+\; \gamma v_\pi(s')\big]$$

That is, for each action a, given the current state s, there is exactly one subsequent state. Thus, for each action a, the second summation reduces to a single term:

$$\begin{aligned} v_\pi(s) \;&=\; \sum_a \pi(a \mid s)\big[R_a(s, s') \;+\; \gamma v_\pi(s')\big] \\ &=\; \sum_a \pi(a \mid s) R_a(s, s') \;+\; \gamma \sum_a \pi(a \mid s) v_\pi(s') \\ &=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \;+\; \gamma\, \mathbb{E}_\pi[v_\pi(s') \mid S_t = s] \end{aligned}$$

Rearranging, we have

$$\begin{aligned} & v_\pi(s) \;-\; \gamma\, \mathbb{E}_\pi[v_\pi(s') \mid S_t = s] \;=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \\ \implies \quad & v_\pi(s) \;-\; \gamma \sum_{s'} P(s'|s) v_\pi(s') \;=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \\ \implies \quad & v_\pi(s) \;-\; \gamma \sum_{s'} P(s'|s) v_\pi(s') \;=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \end{aligned}$$

where $P(s'|s)$ is the probability of transitioning from state s to the subsequent state s', and so $\sum_{s'} P(s'|s)$ is the sum of transition probabilities over all possible subsequent states.

The above can be expressed as a system of linear equations, where each equation is the value function at a particular current state s_i. All possible subsequent states are denoted by s_j (see below).

$$v_\pi(s_i) \;-\; \gamma \sum_{j=1}^{n} P(s_j|s_i) v_\pi(s_j) \;=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s]$$

In matrix form, this is

$$\begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix} \;-\; \gamma \begin{bmatrix} p(s_1|s_1) & \cdots & p(s_n|s_1) \\ \vdots & \ddots & \vdots \\ p(s_1|s_n) & \cdots & p(s_n|s_n) \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix} \;=\; \begin{bmatrix} \mathbb{E}_\pi[R_a(s_1, s')] \\ \vdots \\ \mathbb{E}_\pi[R_a(s_n, s')] \end{bmatrix}$$

That is,

$$\begin{aligned} \mathbf{v}_\pi - \gamma \mathbf{P} \mathbf{v}_\pi &= \mathbf{R}_\pi \\ \implies \quad (\mathbf{I} - \gamma \mathbf{P}) \, \mathbf{v}_\pi &= \mathbf{R}_\pi \end{aligned}$$

where $\mathbf{R}_\pi$ denotes the vector of expected returns on the right-hand side of the expression above.

Problem 3: Finding the value states "in the real world"

Problem 4: Finding an optimal value function

(a) From problem 1, the value function can be written as

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s,s') \;+\; \gamma v_\pi(s')\big]$$

Because a policy governs the actions taken and the associated probabilities, maximizing $v_\pi(s)$ for a given state s, over all policies π, is tantamount to finding the action a that achieves the maximum value (i.e., the immediate, expected and future rewards). Let a^* denote such an action. Then, $a^* = \arg\max_a\{v_\pi(s)\}$. For a given state s, an optimal policy will always distribute all of the weight to a^*. Then, because $\sum_a \pi(a \mid s) = 1$, for a given s, we have that for an optimal policy π^*:

$$\pi^*(a \mid s) = \begin{cases} 1 & \text{if } a = \arg\max\limits_a\{v_\pi(s)\} \\ 0 & \text{otherwise} \end{cases}$$

Because the value function depends on the values of future states, the value at all future states s' will necessarily be optimal.

$$\begin{aligned}
v_*(s) \;&=\; \max_\pi\{v_\pi(s)\} \\
&=\; \max_\pi \;\mathbb{E}_\pi[G_t \mid S_t = s] \\
&=\; \max_\pi \left\{ \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s,s') \;+\; \gamma v_\pi(s')\big] \right\} \\
&=\; \max_a \left\{ \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s,s') \;+\; \gamma v_\pi(s')\big] \right\} \\
&=\; \max_a \left\{ \sum_{s'} P(s' \mid s, a)\big[R_a(s,s') \;+\; \gamma v_\pi(s')\big] \right\}
\end{aligned}$$

Joe: Need to include $v_*(s')$ in expression and argue it. Clean this proof up and include this.

(b)

(c)

Problem 5: Finding the optimal policy using iterative methods

Problem 6: Find the optimal value function for gridworld

Problem 7: A model-free approach

Extra Credit