COMPSCI 687 Pop Quiz 5 — Solutions

Grading instructions. Each question is either correct or incorrect; there is no partial credit. Answers must be perfect down to the symbol, except for the accepted variants listed below. Quantifiers may be included or omitted. Periods at the end of answers are optional.

1. Give the Bellman equation for v^π as a recursive expression for v^π using R, p, π, and γ.

   Accepted variants: Any one of the following two forms is sufficient:

$$v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p(s,a,s') \left(R(s,a) + \gamma v^\pi(s') \right),$$

$$v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a,s') v^\pi(s') \right).$$

   Here and elsewhere, the sets in operator subscripts may be omitted; indices are required. Dummy indices may be renamed consistently without conflicting with free variables. Terms may be re-ordered as long as the expressions are precisely equivalent (e.g., γ can be moved right). All factors, arguments, superscripts, and iteration indices must be correct.

2. Give the Bellman optimality equation for v^* as a recursive expression for v^* using R, p, and γ.

   Accepted variants: Any one of the following two forms is sufficient:

$$v^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a,s') v^*(s') \right),$$

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s,a,s') \left(R(s,a) + \gamma v^*(s') \right).$$

COMPSCI 687 Pop Quiz 5 — Solutions (continued)

3. Give the dynamic programming policy evaluation update for a fixed policy π, using R, p, π, γ, and the current estimate v_k.

   Accepted variants: Any one of the following two forms is sufficient:

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p(s,a,s') \left(R(s,a) + \gamma v_k(s') \right),$$

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a,s') v_k(s') \right).$$

   Note the use of v_k on the right side.

4. In policy iteration, the current policy is π_k. Give the expression used to obtain the next deterministic policy greedily from v^{π_k}, using R, p, and γ.

   Accepted variants: Any one of the following two forms is sufficient:

$$\pi_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s,a,s') \left(R(s,a) + \gamma v^{\pi_k}(s') \right),$$

$$\pi_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a,s') v^{\pi_k}(s') \right).$$

   The answer must select a maximizing action, using arg max, with v^{π_k} on the right side. Any maximizing action may be chosen when there is a tie. You may write π' for $\pi_{k=1}$ and π for π_k, but only if you make both of these changes (you should not mix π' and π_k, for example).