2.5. Dynamic Programming

(1) Stochastic:

Using eqn 4.4, we can say:

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$
$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]$$

Using eqn 4.6, we can say

$$q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a]$$
$$= \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]$$

Using the above two eqns, we get:

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a).$$

(2) Deterministic :-

$$v_\pi(s) = q_\pi(s,a) \quad \forall\ a \in A : \pi(a|s) = 1$$

(2) The Bellman optimality eqⁿ for the action-value function:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]$$

The value-iteration for the state-value function was obtained by turning its Bellman eqⁿ to an update rule. [Barto Sutton]

Using the same for the action-value ~~function~~ iteration:-

$$q_{K+1}(s) = \sum_{s'} p(s', r | s, a)[r + \gamma \max_{a'} q_K(s', a')].$$

(3) We know:

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s, a) \quad \text{--- (1)}$$

$$q_\pi(s) = \sum_{s'} p(s', r | s, a)[r + \gamma v_\pi(s')] \quad \text{--- (2)}$$

Replacing $v_\pi(s)$ from (1) to (2)

$$q_\pi(s) = \sum_{s'} p(s', r | s, a)[r + \gamma \sum_{a \in A} \pi(a'|s) \cdot q_\pi(s, a')$$

~~(4) Using eqn 4.9, we can say~~

~~π~~

(4) Given:

$$\pi(s) \leftarrow \underset{a}{\text{argmax}} \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$$

We also know that:-

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a) \quad -(1)$$

Using (1) we can rewrite the policy improvement step:-

$$\pi(s) \leftarrow \underset{a}{\text{argmax}} \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \sum_{a' \in A} \pi(a'|s) \cdot q_\pi(s,a') \Big]$$

(5) The policy evaluation step on page 75 assumes that $\pi(s)$ has a probability distribution over $'a' \in A$. That's why a summation over $\forall\, a \in A$ is taken on page 75.

The evaluation step on page 80 assumes a deterministic or a pure greedy policy with respect to the actions. Hence, there is only one action that should be chosen for a state.