Chapter 2 State Value and Bellman Equation

State Value, Bellman Equation, Action Value

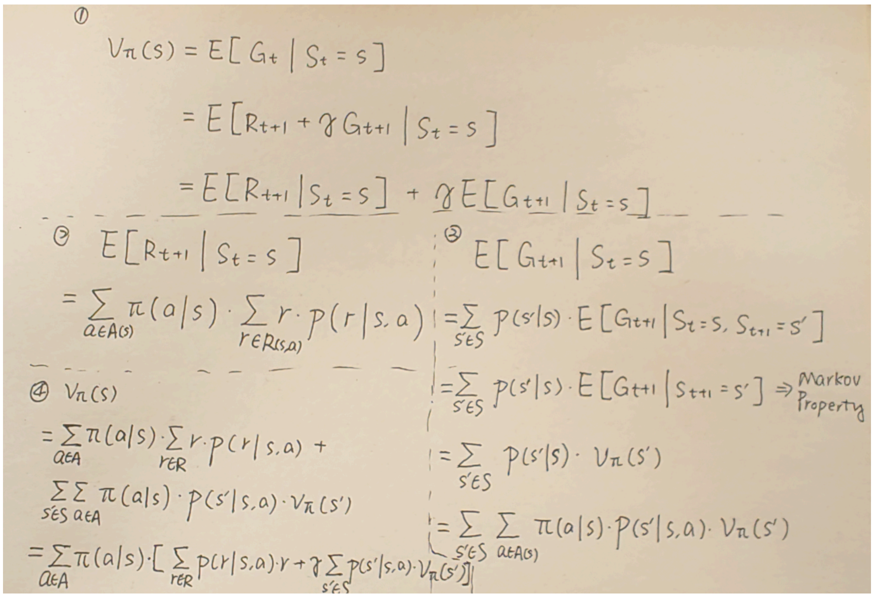
1. Why did we introduce State Value? Isn't return enough to valuate a policy?
2. Explain the understanding of Random Variables in RL.
3. Explain how state values depend on each other(Bootstrapping)?
4. What is G_t? What is the difference between G_t and v_t?
5. What does Bellman Equation tell us?
6. What is policy evaluation?
7. Derive the Bellman Equation. How to understand the ultimate format of Bellman Equation?
8. Did find any principle when dealing with multiple Σ ?
9. In the Bellman Equation, what are unknown, what represent the model, is π given?
10. Derive matrix-vector form of Bellman Equation.
11. Why need matrix-vector form beyond elementwise form?
12. How to solve Bellman Equation? Give 2.
13. Prove that the iterative solution is converged.
14. What is action value? Is action value based on state or has no relationship with state?
15. What is the relationship between state value and action value?
16. The elementwise form of action value.
17. Derive the matrix-vector form of action value.
18. In the matrix-vector form of state value and action value, we have vector $v = [v_1, v_2, v_3, \ldots, v_n]^{-1}$and $q = [q_1, q_2, q_3, \ldots, q_n]^{-1}$. Explain what do these 2 sequences represent for? Does sequence *v* represent for the state values of the whole state space? Does sequence *q* represent for the action values of the action space of a state?

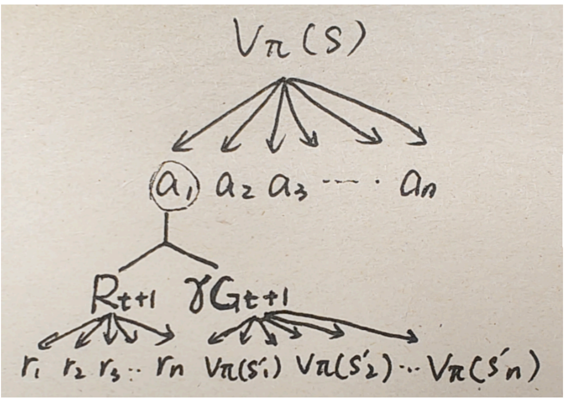
1. In stochastic system, the model (state transition and reward) is stochastic and can be described using conditional probabilities. Return can only describe a deterministic trajectory, but state value can described the expected return of a stochastic trajectory.
2. Random Variables do not have a deterministic value, but have a row of values with coresponded probabilities(Probability Distribution). Such as action, state transition, return, they are stochastic and are described with probability distribution.
3. $v_t = r_{t+1} + \gamma \cdot r_{t+2} + \gamma^2 \cdot r_{t+3} + \cdots = r_{t+1} + \gamma \cdot v_{t+1}$
4. G_t is a stochastic variable describes the discounted return of state s_t following a policy. v_t is the expectation of G_t and it is a deterministic value.

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \ldots) \\ &= R_{t+1} + \gamma G_{t+1}, \end{aligned}$$

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t|S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s].
\end{aligned}
$$

5. Bellman Equation solve state values based on the model and a given policy. It explains the relationship between $v_\pi(s)$ and $v_\pi(s')$ based on the model. Also, Bellman Equation is fundamentally a formula calculating the AVERAGE DISCOUNTED RETURN on state s, it shows the general situation.
6. The process of solving Bellman Equantion to get state values is policy evaluation.
7.





$$
\sum_{a \in \mathcal{A}} \pi(a|s) \left[\sum_{r \in \mathcal{R}} p(r|s,a) r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v_\pi(s') \right]
$$

8. We can freely arrange the items and Σ. Just make sure that when the item and the Σ have the same variable notation, the items should follow behind the Σ but not before.
9. $v_\pi(s)$ and $v_\pi(s')$ is unknown, p(r|s,a) and p(s'|s,a) represent the model, π is a given policy and the process is policy evaluation because state values can evaluate the perfermance of the policy.
10.

$$v_\pi(s) = r(s) + \gamma v_\pi(s')$$
$$= r(s) + \gamma \cdot \sum_{s' \in S} p(s'|s) \cdot v_\pi(s')$$

$$\Rightarrow \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_n \end{bmatrix} + \gamma \cdot \begin{bmatrix} p(s_1|s_1) & p(s_2|s_1) & p(s_3|s_1) & \cdots & p(s_n|s_1) \\ p(s_1|s_2) & p(s_2|s_2) & p(s_3|s_2) & \cdots & p(s_n|s_2) \\ \vdots & & & & \\ p(s_1|s_n) & p(s_2|s_n) & p(s_3|s_n) & \cdots & p(s_n|s_n) \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix}$$

$$\vec{v} = \vec{r} + \gamma \vec{P} \cdot \vec{v}$$

11. Bellman Equation is a row of formula of every state. To solve it, we must put it together. Matrix-vector form can deal with this.
12. Closed-form solution and iterative solultion.
13.

define $\vec{\delta}_k = \vec{v}_k - \vec{v}_\pi$.

$$\vec{v}_{k+1} = \vec{r}_\pi + \gamma \vec{P}_\pi \vec{v}_k$$

$$\Rightarrow \vec{\delta}_{k+1} + \vec{v}_\pi = \vec{r}_\pi + \gamma \vec{P}_\pi (\vec{\delta}_k + \vec{v}_\pi)$$

$$\vec{\delta}_{k+1} = \gamma \vec{P}_\pi \cdot \vec{\delta}_k .$$

Thus, $\vec{\delta}_{k+1} = \gamma \vec{P}_\pi \cdot \gamma \vec{P}_\pi \vec{\delta}_{k-1}$

$$= \cdots$$
$$= \vec{\gamma}^{k+1} \vec{P}_\pi^{k+1} \cdot \vec{\delta}_0 \to 0$$

$0 < \gamma < 1$, then $\gamma^{k+1} \to 0$. $\vec{P}_\pi$ 特征值 ≤ 1, then $\vec{P}_\pi^{k+1}$ is finite.

14. Action value is the discounted return of taking a specific action at a specific state. Same action on different state may have different action value.
15. State value is the expectation value of every action value on this state.

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$$

16.

$$q_\pi(s, a) = \sum_{r \in \mathcal{R}} p(r|s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) v_\pi(s')$$

17. **Derive by myself. No record in the book. So maybe some mistakes.**

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

$$\begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} + \gamma \cdot \begin{bmatrix} p(s_1'|s,a_1) & p(s_1'|s,a_2) & \cdots & p(s_1'|s,a_n) \\ p(s_2'|s,a_1) & p(s_2'|s,a_2) & \cdots & p(s_2'|s,a_n) \\ \vdots & & & \\ p(s_m'|s,a_1) & p(s_m'|s,a_2) & \cdots & p(s_m'|s,a_n) \end{bmatrix}_{n \times m} \cdot$$

$$\begin{bmatrix} \pi(a_1|s_1') & \pi(a_1|s_2') & \cdots & \pi(a_1|s_m') \\ \pi(a_2|s_1') & \pi(a_2|s_2') & & \pi(a_2|s_m') \\ \vdots & \vdots & & \vdots \\ \pi(a_n|s_1') & \pi(a_n|s_2') & & \pi(a_n|s_m') \end{bmatrix}_{m \times n} \cdot$$

$$\begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}_{n \times 1}$$

$$\vec{q}_\pi = \vec{r} + \gamma \vec{P} \vec{\Pi} \vec{q}_\pi$$

18. They don't represent for the state values of the whole state space or the action values of the action space of a state. They come from bootstrapping. They are sequential records of the state values and action values obtained during a trajectory following the policy.