벨만 방정식

Review

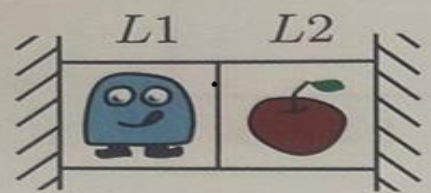
- 상태가치 함수 구하기

결정적 상황 vs 비결정적(확률적) 상황

2.4 MDP 예제

이번 절에서는 MDP에 속하는 구체적인 문제를 하나 살펴보겠습니다. 바로 [그림 2-13]과 같은 두 칸짜리 그리드 월드입니다.

그림 2-13 두 칸짜리 그리드 월드



칸이 두 개이고 좌우 끝은 벽으로 막혀 있습니다. 이 문제의 설정은 다음과 같습니다.

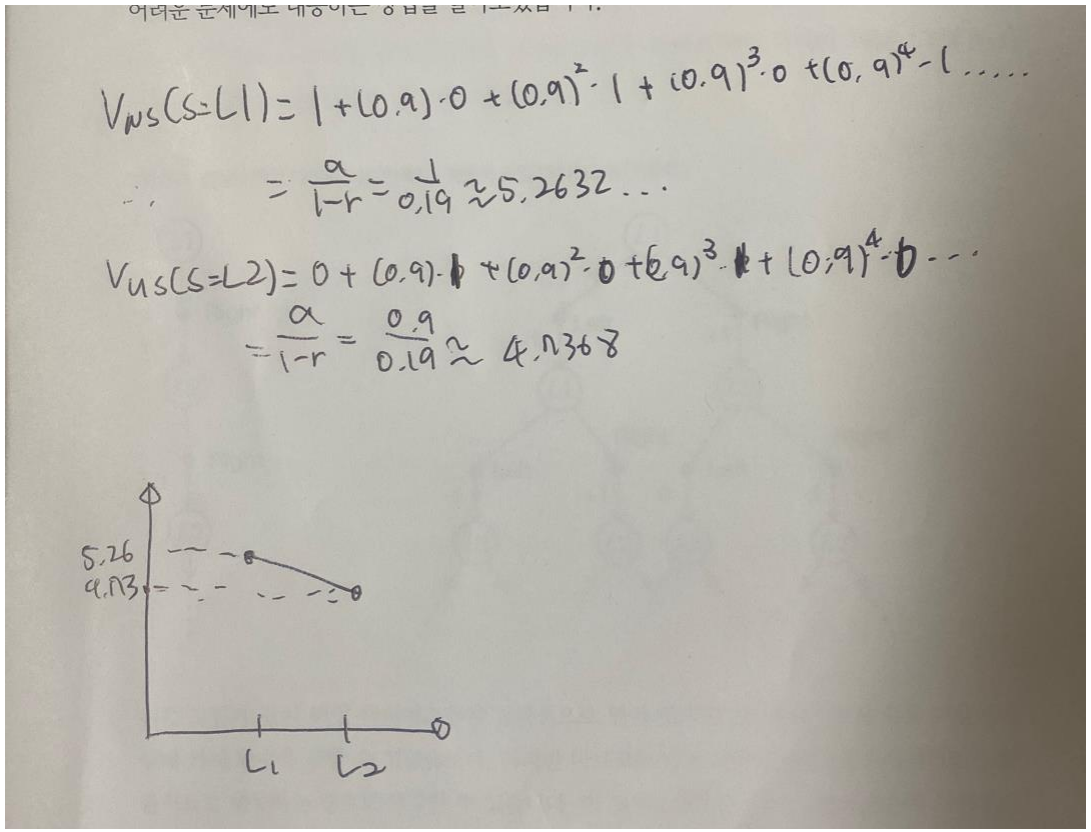
- 에이전트는 오른쪽이나 왼쪽으로 이동할 수 있다.
- 상태 전이는 결정적이다.
- 에이전트가 $L1$에서 $L2$로 이동하면 사과를 받아 +1의 보상을 얻는다.
- 에이전트가 $L2$에서 $L1$로 이동하면 사과가 다시 생성된다.
- 벽에 부딪히면 −1의 보상을 얻는다. 즉, 벌을 받는다. 예를 들어 에이전트가 $L1$에서 왼쪽으로 이동하면 −1의 보상을, $L2$에서 오른쪽으로 이동해도 마찬가지로 −1의 보상을 얻는다(이때 사과는 다시 생성되지 않는다).
- 지속적 과제, 즉 '끝이 없는' 문제다.

	$s = L1$	$s = L2$
$\mu_1(s)$	Right	Right
$\mu_2(s)$	Right	Left
$\mu_3(s)$	Left	Right
$\mu_4(s)$	Left	Left

$$V_{\pi}(s) = \mathbb{E}[G_t | S_t = s]$$

$$E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots + | S_t = s]$$

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots$$



But 비결정적일땐??

상태전의함수까지 계산해줘야한다.

ex) 즉, 횟수 제한이 있다면 가능은 하겠지면 계산횟수가 지수적으로 늘어난다. ➔ 일반적인 상황에선 불가능이라 보면됨

+ 우리는 지금까지 정책을 고르고, 다 계산해서 최적을 찾았다.
이과정을 한번에 하는 식이 있다!!!

$$V(S_t) = \int_{a_t}^{a_\infty} G_t P(a_t, S_{t+1}, a_{t+1}, \dots \mid S_t) \frac{da_\infty}{da_t}$$

벨만 방정식

비결정적(확률적) 상황에서 유용하고, 결정적 상황에서도 식을 간결화 할수있다!!

벨만 방정식 유도방법

1. G(t)와 G(t+1) 관계식

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

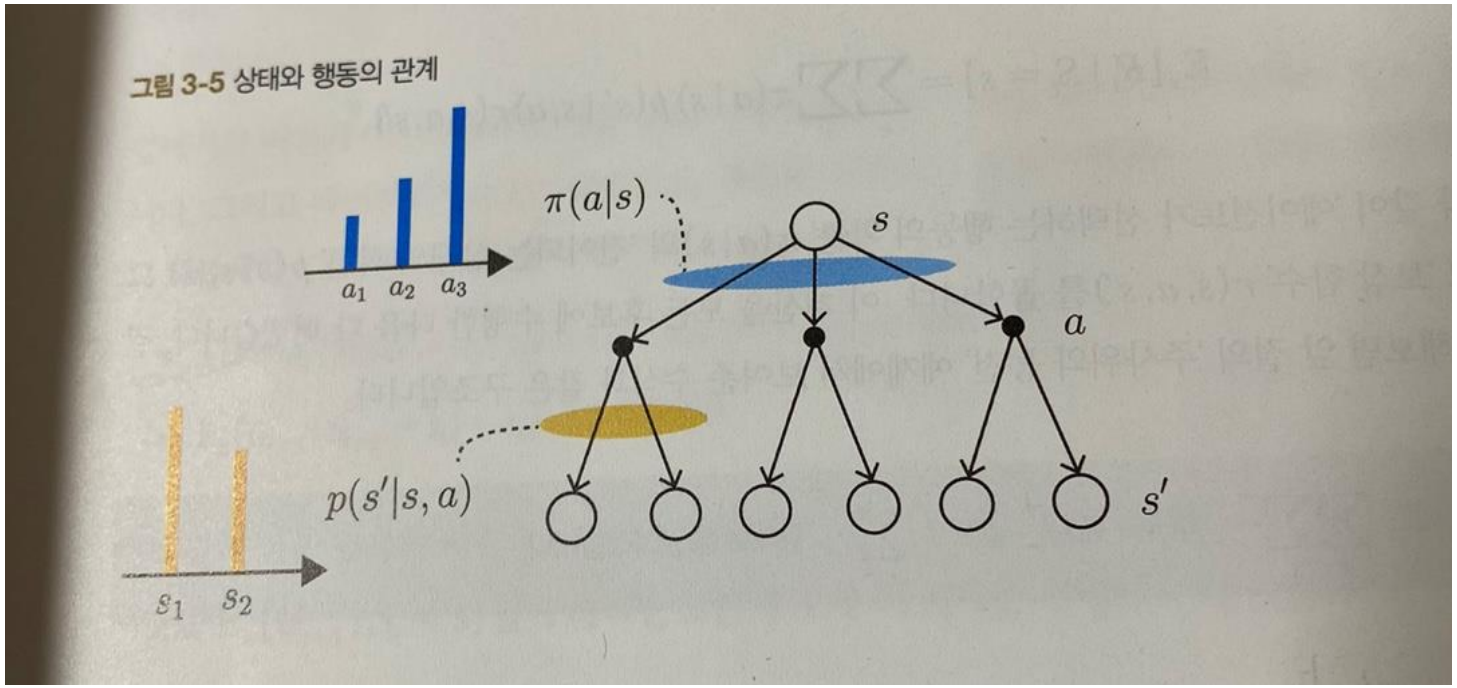
$$G_{t+1} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

$$G_t = R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots)$$

$$G_t = R_t + \gamma G_{t+1}$$

2. V(s)식에 대입

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



3. V(s)식 풀기

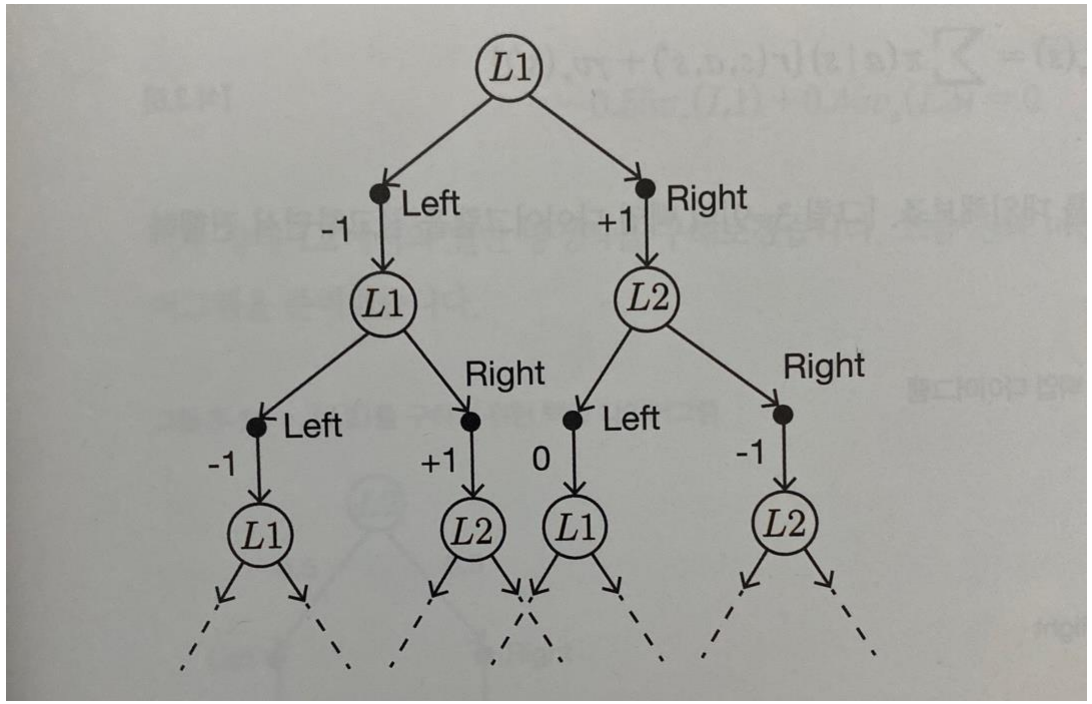
$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[R_t \mid S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s] && \text{[식 3.5]} \\
&= \sum_{a,s'} \pi(a \mid s)\, p(s' \mid s,a)\, r(s,a,s') + \gamma \sum_{a,s'} \pi(a \mid s)\, p(s' \mid s,a)\, v_\pi(s') \\
&= \sum_{a,s'} \pi(a \mid s)\, p(s' \mid s,a) \{ r(s,a,s') + \gamma v_\pi(s') \} && \text{[식 3.6]}
\end{aligned}
$$

현재 상태에서 정책에 따라 선택할 수 있는 모든 행동에 대한 **기대 보상**과 **미래 기대 가치**를 합산하여 상태 가치를 도출한다!

문제 적용

정책이 0.5, 0.5이고

상태전의확률은 1일때



$$v_\pi(L1) = 0.5\{-1+0.9v_\pi(L1)\}+0.5\{1+0.9v_\pi(L2)\}$$

$$v_\pi(L2) = 0.5\{0+0.9v_\pi(L1)\}+0.5\{-1+0.9v_\pi(L2)\}$$

$s' = f(s,a)$ 이면

$$v_\pi(s) = \sum_a \pi(a \mid s)\{r(s,a,s') + \gamma v_\pi(s')\}$$

새로운 개념 행동가치함수

이때 까지는 정책에 대한 보상의 기댓값을 통해 어떤 정책이 좋은 정책인가에 대한 상태가치함수를 사용했었다.

이제는 어떤 state에서, 어떤 action이 최고일까에 대한 행동가치함수

$$V(s) = \mathbb{E}[G_t|S_t = s]$$

$$Q(s,a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \mid S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a] \\
&= \sum_{s'} p(s' \mid s,a) r(s,a,s') + \gamma \sum_{s'} p(s' \mid s,a) \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s'] \\
&= \sum_{s'} p(s' \mid s,a) \{ r(s,a,s') + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s'] \} \\
&= \sum_{s'} p(s' \mid s,a) \{ r(s,a,s') + \gamma v_\pi(s') \}
\end{aligned}
$$

[식

현재 보상과 **다음 상태에서의 기대 가치**로 나누어 표현한 것으로, 상태 전이 확률과 할인율을 고려하여 행동 가치 함수를 재귀적으로 계산할 수 있게 해줌

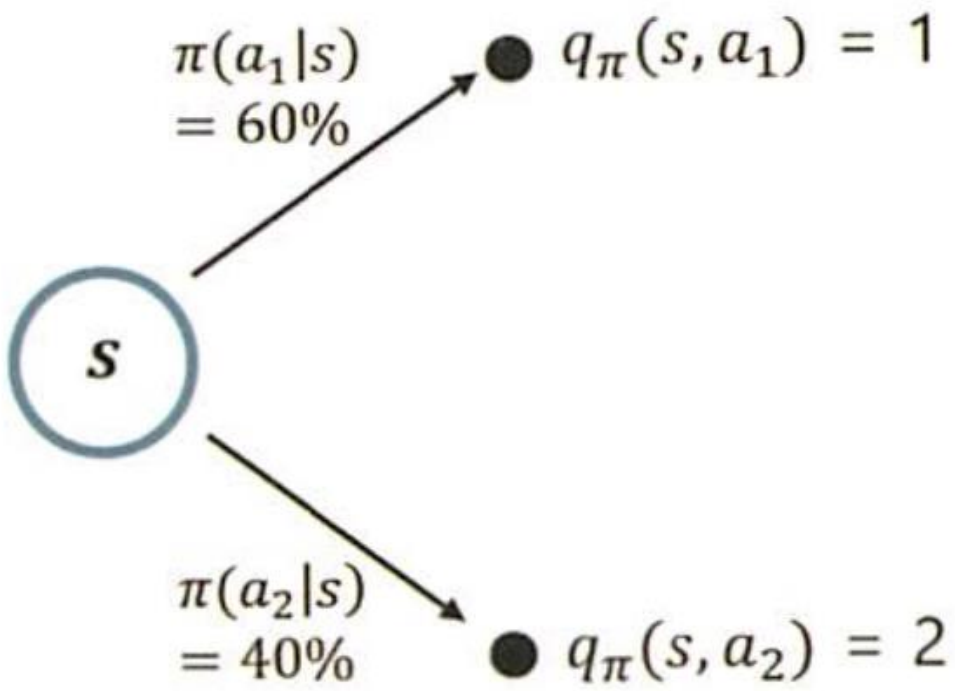
$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[R_t \mid S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s] \qquad \text{[식 3.5]} \\
&= \sum_{a,s'} \pi(a \mid s) p(s' \mid s,a) r(s,a,s') + \gamma \sum_{a,s'} \pi(a \mid s) p(s' \mid s,a) v_\pi(s') \\
&= \sum_{a,s'} \pi(a \mid s) p(s' \mid s,a) \{ r(s,a,s') + \gamma v_\pi(s') \} \qquad \text{[식 3.6]}
\end{aligned}
$$

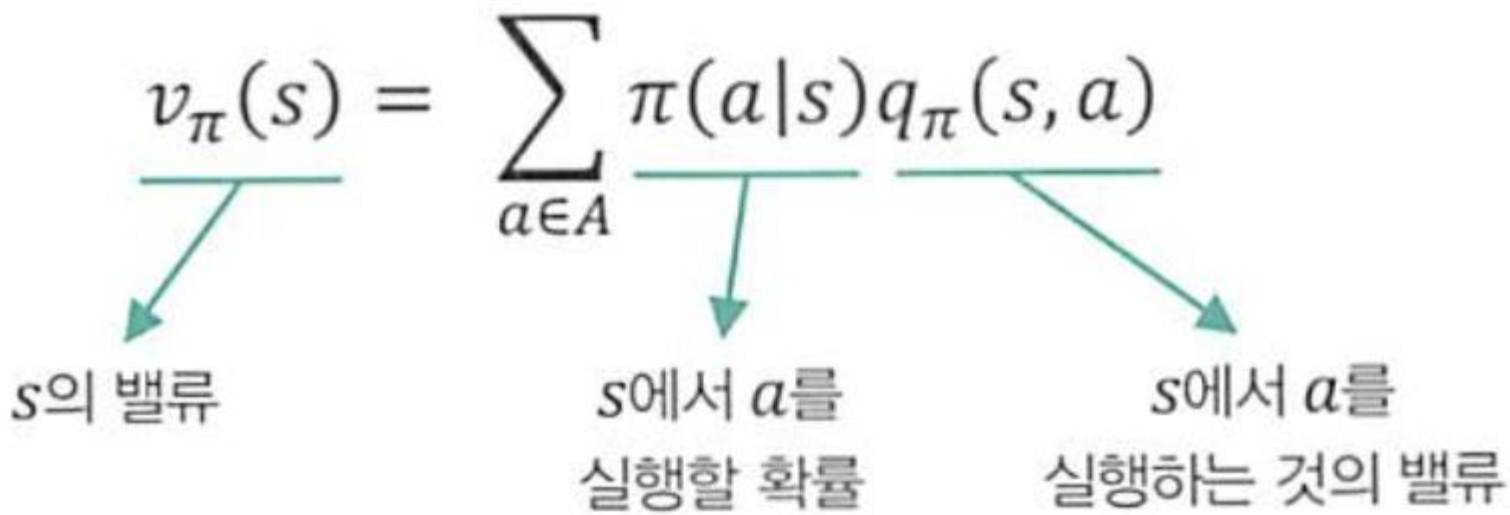
s의 상태 가치 함수는 a1을 취한 이후에 받는 반환값들과 a2를 취한 이후에 받는 반환값들을 모두 고려해서 기댓값을 구한 것

v(s) =1.4

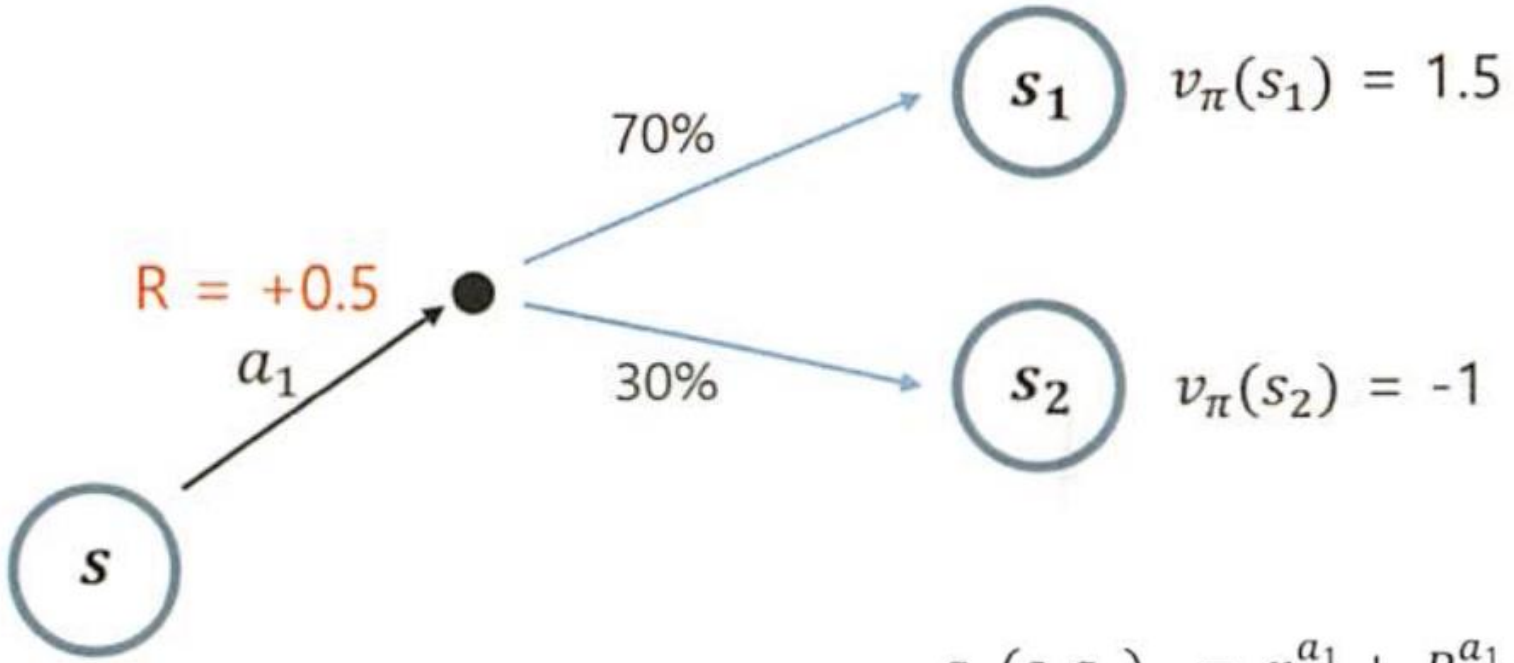
행동 가치 함수에 대해 생각해보면, 상태 s에서의 행동 가치함수는 2가지가 나옴

q(s,a1) =1, q(s,a2)=2




$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$
s의 밸류
s에서 a를
실행할 확률
s에서 a를
실행하는 것의 밸류

연습!



$$q_\pi(s, a_1) = r_s^{a_1} + P_{ss_1}^{a_1} * v_\pi(s_1) + P_{ss_2}^{a_1} * v_\pi(s_2)$$

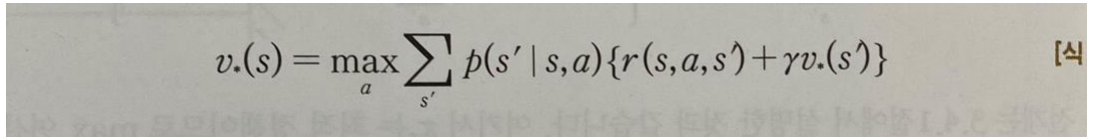
$$= 0.5 + 0.7 * 1.5 + 0.3 * (-1)$$

$$= 1.25$$

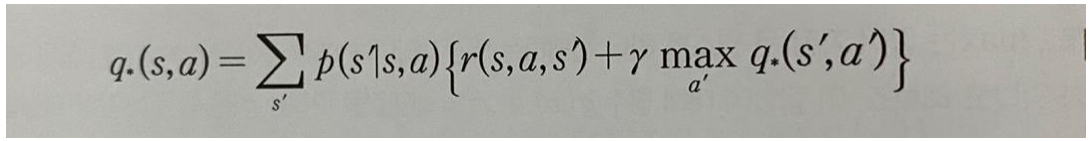
최적 밸만 방정식

$$v_*(s) = \max_{\pi} v_\pi(s)$$

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)$$



이 식은 현재 상태에서 최적의 행동을 선택했을 때, 기대할 수 있는 최대 가치를 계산



이 방정식은 **현재 보상**과 **다음 상태에서 최적의 행동을 선택했을 때의 가치**를 결합하여, 행동 가치 함수 계산