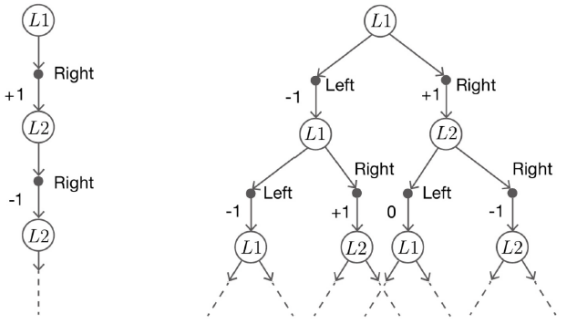
Chapter 3 벨만 방정식

2장에서는 '두 칸짜리 그리드 월드'라는 문제를 다루고, 여기에서 환경이 결정적이고 에이전트의 행동도 결정적이라고 가정했다. 따라서 백업 다이어그램은 왼쪽과 같이 하나의 직선으로 뻗어 있었다.



왼쪽 그림과 같이 백업 다이어그램이 일직선으로 뻗어 있다면 수식을 이용하여 상태 가치 함수를 구할 수 있다. 하지만 MDP에서는 에이전트가 확률적으로 행동하는 경우도 생각할 수 있다. 이 경우 백업 다이어그램이 오른쪽처럼 넓게 퍼져 나가고, 이때의 상태 가치 함수는 수식으로 구할 수 없다.

이번 장에서는 오른쪽 같은 상황에서도 상태 가치 함수를 구하는 것이고, 이때의 핵심이 **벨만 방정식**이다. 벨만 방정식은 마르코프 결정 과정에서 성립하는 가장 중요한 방정식이며 많은 강화 학습 알고리즘에 중요한 기초를 제공한다.

3.1. 벨만 방정식 도출

3.1.1. 확률과 기댓값(사전 준비)

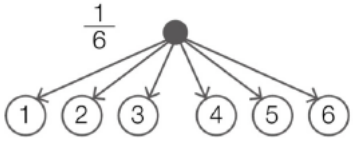
주사위를 예를 들어, 각각의 눈이 나올 확률이 정확하게 $\frac{1}{6}$씩인 이상적인 주사위라고 가정한다. 이때 눈 개수를 x라는 확률 변수로 표현하면 x는 1부터 6까지의 정수가 될 수 있다. 그리고 확률은 모두 $\frac{1}{6}$씩이니, 각 눈이 나올 확률은 다음 식으로 표현할 수 있다.

$$p(x)=\frac{1}{6}$$

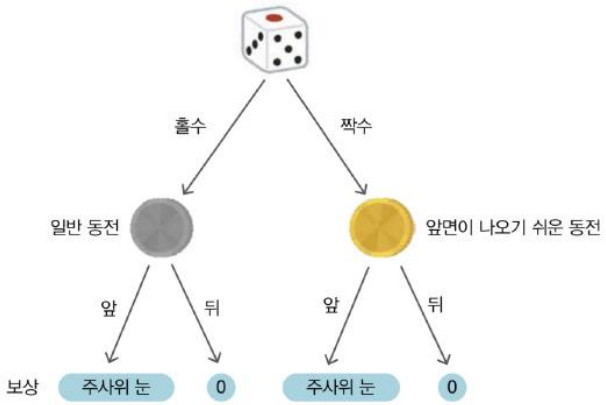
이제 주사위를 굴렸을 때 나올 눈의 기댓값은 다음처럼 계산하면 된다.

$$\begin{aligned}\mathbb{E}[x] &= 1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}\\ &= 3.5\end{aligned}$$

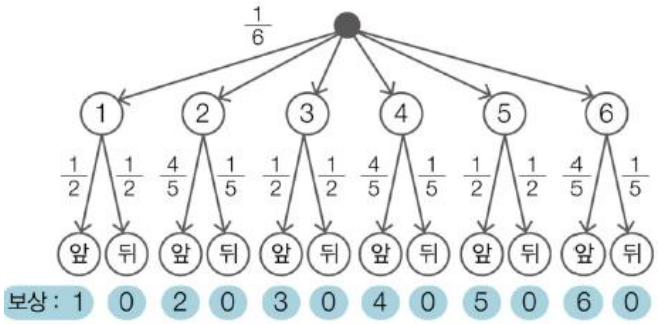
백업 다이어그램은 주사위 눈 개수가 두 번째 줄에 배치된 모습이 된다.



이번에는, 다음과 같은 문제를 생각해보자.



이번 문제에서는 주사위를 먼저 던지고 이어서 동전을 던지는 방식으로 진행된다. 이때 주사위를 던져 짝수가 나오면 앞면이 잘 나오는 동전(확률 = 0.8)이 주어지고, 홀수가 나오면 일반 동전(확률 = 0.5)이 주어진다. 그런 다음 주어진 동전을 던져 앞면이 나오면 주사위의 눈 개수만큼을 보상으로 얻는다. 반대로 뒷면이 나오면 보상은 0이다. 이 문제의 백업 다이어그램은 다음과 같다.



그림을 보면 주사위 1이 나올 확률은 $\frac{1}{6}$이고, 이어서 동전의 앞면이 나올 확률은 $\frac{1}{2}$이다. 이때의 보상은 1이다. '보상 기댓값'을 구하려면 모든 경우에 대해 똑같이 계산하여 다 더하면 된다. 실제로 해보면 다음과 같다.

$$\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 1\right)+\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 0\right)+\left(\frac{1}{6}\cdot\frac{4}{5}\cdot 2\right)+\left(\frac{1}{6}\cdot\frac{1}{5}\cdot 0\right)+\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 3\right)+\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 0\right)+$$
$$\left(\frac{1}{6}\cdot\frac{4}{5}\cdot 4\right)+\left(\frac{1}{6}\cdot\frac{1}{5}\cdot 0\right)+\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 5\right)+\left(\frac{1}{6}\cdot\frac{1}{2}\cdot 0\right)+\left(\frac{1}{6}\cdot\frac{4}{5}\cdot 6\right)+\left(\frac{1}{6}\cdot\frac{1}{5}\cdot 0\right)$$
$$= 2.35$$

'보상의 기댓값'을 구하는 방법은 말단 노드가 발생할 확률과 그때의 보상을 곱하는 계산을 모든 후보에 수행한 다음 다 더하는 것이다.

주사위의 눈을 x, 동전의 결과를 y로 표기하면, 주사위 눈 개수에 따라 동전 앞면이 나올 확률이 달라진다. 이 설정은 조건부 확률로 표현하며 다음과 같다.

$$p(y = \text{앞} \mid x = 4) = 0.8$$
$$p(y = \text{뒤} \mid x = 3) = 0.5$$

또한, x와 y가 동시에 일어날 확률, '동시 확률'은 다음과 같다.

$$p(x, y) = p(x)\,p(y \mid x)$$

이번 문제에서 보상은 x와 y의 값에 의해 결정된다. 따라서 보상을 함수 $r(x,y)$로 나타낼 수 있다.

$$r(x=4, y=\text{앞}) = 4$$
$$r(x=3, y=\text{뒤}) = 0$$

기댓값은 '값 * 그 값이 발생할 확률'의 합이다. 그러므로 보상의 기댓값은 다음 식으로 나타낼 수 있다.

$$\mathbb{E}[r(x,y)] = \sum_x \sum_y p(x,y) r(x,y)$$
$$= \sum_x \sum_y p(x)\, p(y|x)\, r(x,y)$$

이 수식의 형태는 밸만 방정식에도 동일하게 등장한다.

3.1.2. 벨만 방정식 도출

수익을 다음과 같이 정의할 수 있다.

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

이번 절에서는 보상을 무한히 계속 받을 수 있는 지속적 과제를 가정한다. 위의 식에서 t에 t+1을 해보면

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

이는 시간 t+1 이후에 얻을 수 있는 보상의 합이다. 이 식을 이용하여 G_t와 G_{t+1}의 관계를 알 수 있다.

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

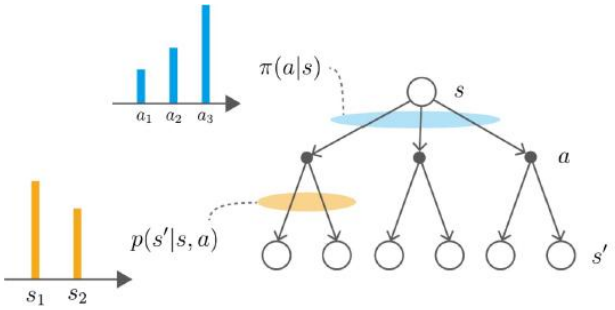
이 식을 상태 가치 함수의 수식에 대입할 수 있다. 상대 가치 함수는 수익에 대한 기댓값(기대 수익)이며, 다음 식으로 정의된다.

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

상태 s의 상태 가치 함수가 $v_\pi(s)$로 표현된다. 이 식의 G_t에 위에서 구한 보상의 식을 대입해보면 다음과 같다.

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

이 식의 항을 하나씩 구해보자. 첫 번째 항은 $\mathbb{E}[R_t | S_t = s]$이다.



먼저 상황을 확인한다. 현재 상태가 s이고 에이전트는 정책 $\pi(a|s)$에 따라 행동한다. 예를 들어 다음의 세 가지 행동을 취할 수 있다.

$$\pi(a = a_1 \mid s) = 0.2$$
$$\pi(a = a_2 \mid s) = 0.3$$
$$\pi(a = a_3 \mid s) = 0.5$$

에이전트는 이 확률 분포에 따라 행동을 선택한다. 그러면 상태 전이 확률 $p(s'|s,a)$에 따라 새로운 상태 s'로 이동한다. 예를 들어 행동 a1을 수행했을 때 전이될 수 있는 상태 후보가 두 개라면 다음과 같은 값을 취한다.

$$p(s' = s_1 \mid s, a = a_1) = 0.6$$
$$p(s' = s_2 \mid s, a = a_1) = 0.4$$

그리고 마지막으로 보상은 $r(s,a,s')$ 함수로 결정된다.

구체적인 예로 에이전트가 0.2의 확률로 행동 a1을 선택하고 0.6의 확률로 상태 s1으로 전이된다고 가정해보자. 이 경우 얻게 되는 보상은 다음과 같다.

- $\pi(a = a_1 \mid s)p(s' = s_1 \mid s, a = a_1) = 0.2 \cdot 0.6 = 0.12$의 확률로
- $r(s, a = a_1, s' = s_1)$의 보상을 얻는다.

기댓값을 구하려면 모든 후보에 똑 같은 계산을 수행하여 다 더하면 된다.

$$\mathbb{E}_\pi[R_t \mid S_t = s] = \sum_a \sum_{s'} \pi(a \mid s)\, p(s' \mid s, a) r(s, a, s')$$

이와 같이 '에이전트가 선택하는 행동의 확률' $\pi(a|s)$와 '전이되는 상태의 확률' $p(s'|s,a)$ 그리고 '보상 함수' $r(s,a,s')$를 곱한다. 이 계산을 모든 후보에 수행한 다음 다 더했다. 이 것으로 첫번째 항의 전개는 끝났다.

$$v_\pi(s) = \mathbb{E}_\pi[R_t|S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1}|S_t = s]$$

첫 번째 항 전개

$$\sum_{a,s'} \pi(a|s) p(s'|s,a)\, r(s,a,s')$$

다음은 이쪽 항

이제 $\gamma \mathbb{E}[G_{t+1}|S_t = s]$를 살펴보자. 여기서 γ는 상수이므로 $\mathbb{E}[G_{t+1}|S_t = s]$에 대해서만 살펴본다. 이 식은 상

태 가치 함수의 정의식과 비슷하지만, G_{t+1} 부분이 다르다.

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

cf) 용어: 상태 가치 함수, 상태 전이 함수
상태 가치 함수: 특정 상태 s에서 시작했을 때, 앞으로 기대할 수 있는 보상의 총합
상태 전이 함수: 현재 상태 s에서 행동 a를 했을 때, 다음 상태 s'로 전이될 확률

상태 가치 함수의 식에서 t에 t+1을 대입한다.

$$v_\pi(s) = \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s]$$

이 식은 상태 S_{t+1} = s에서의 가치 함수이다. 우리의 관심은 $\mathbb{E}[G_{t+1}|S_t = s]$이다. 이 식은 현재 시간이 t일 때 한 단위 뒤 시간(t+1)의 기대 수익을 뜻한다. 문제 해결의 핵심은 조건인 S_t = s를 S_{t+1} = s 형태로 바꾸는 것이다. 즉, 시간을 한 단위만큼 흘려 보내는 것이다.

앞의 예처럼, 지금 에이전트의 상태는 S_t = s이다. 그리고 에이전트가 0.2의 확률로 a1 행동을 선택하고, 0.6의 확률로 s1 상태로 전이한다고 해보자. 그러면 다음과 같이 나타낼 수 잇다.

- $\pi(a = a_1 \mid s)\, p(s' = s_1 \mid s, a = a_1) = 0.2 \cdot 0.6 = 0.12$ 의 확률로
- $\mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s_1] = v_\pi(s_1)$ 로 전이된다.

이처럼 다음 단계의 시간을 '보는' 것으로 다음 상태의 가치 함수를 얻을 수 있다. 이제 기댓값 $\mathbb{E}[G_{t+1}|S_t = s]$를 구하려면 모든 후보에 이 계산을 수행하여 다 더한다.

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

앞서 전개한 식에 대입하면 다음 식이 도출된다.

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

이것이 벨만 방정식이다. 벨만 방정식은 '상태 s의 상태 가치 함수'와 '다음에 취할 수 잇는 상태 s'의 상태 가치 함수'의 관계를 나타낸 식으로 모든 상태 s와 모든 정책 π에 대해 성립한다.

cf) 이해하기: 벨만 방정식
벨만 방정식이란 현재 상태의 가치를 '즉시 보상 + 미래 가치'로 표현하는 공식이다. 어떤 상태에서 얻을 수 있는 최적의 총 보상을 계산하는 방법이다. 이 벨만 방정식을 이용해서 최적의 정책을 찾을 수 있다.

강화학습에서는 현재 상태에서 미래에 받을 모든 보상(G)의 합을 고려한다. 보상은 여러 번 받을 수 있기 때문에 각 단계에서 받은 보상을 계속 더해주는 방식으로 정의한다. 하지만, 미래 보상은 현재보다 덜 중요하기 때문에 할인율 γ를 곱해준다.

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

강화학습에서는 미래 보상을 포함하여 현재 상태의 가치를 계산한다. 즉, 현재가치 V(s)는 현재 보상 R_t + 미래 가치 V(s')로 계산이 가능하다.

$$V(s) = \mathbb{E}[R_t + \gamma V(s')]$$

현실에서는 현재 상태에서 다음 상태로 무조건 한 가지 방향으로 이동하는게 아니라, 여러 가능성이 존재할 수 있다. 그래서 현재 상태에서 특정 행동을 했을 때, 다음 상태로 갈 확률(확률 전이 확률)을 고려 해야 한다.

$$V^{\pi}(s) = \sum \pi(a|s) \sum P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$$

즉, 모든 가능한 행동과 상태 전이를 고려하여 기대 보상을 고려해야 한다. 강화학습에서 목표는 최적의 정책을 찾는 것이기 때문에, 가장 좋은 행동을 선택했을 때의 가치를 계산해야 한다.

$$V^{*}(s) = \max_{a} \sum P(s'|s,a)[R(s,a,s') + \gamma V^{*}(s')]$$

현재 상태에서 가장 좋은 행동을 선택하면, 미래에 받을 보상이 최대가 된다. 강화학습은 이 방정식을 반복적으로 풀면서 최적의 행동을 찾아가는 과정이다.

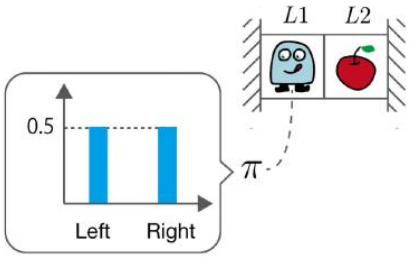
3.2. 벨만 방정식의 예

벨만 방정식을 이용하면 상태 가치 함수를 구할 수 있다.

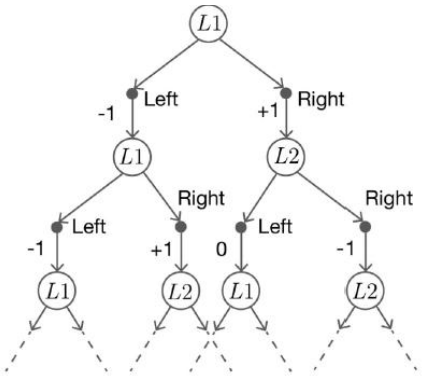
3.2.1. 두 칸짜리 그리드 월드

여기서는 '두 칸짜리 그리드 월드' 문제를 다룬다. 이번에는 에이전트가 무작위로 움직인다고 가정한다. 즉 50%의 확률로 오른쪽, 나머지 50%의 확률로는 왼쪽으로 이동한다.

그림 3-7 두 칸짜리 그리드 월드(벽에 부딪히면 –1, 사과를 얻으면 +1, 사과는 계속 생성)



$\nu_{\pi}(L1)$은 상태 L1에서 무작위 정책 π에 따라 행동했을 때 얻을 수 있는 기대 수익이다. 이 기대 수익은 앞으로 무한히 지속되는 보상의 총합이다. 백업 다이어그램은 다음과 같다. (상태는 결정적으로 전이됨)



그림과 같이 지금 문제는 무한히 분기되어 뻗어나가는 계산이다. 이처럼 무한히 분기하는 계산을 벨만 방정식을 이용하여 구할 수 있다. 벨만 방정식은 다음과 같다.

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

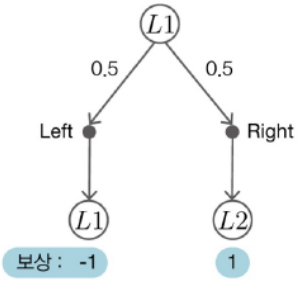
이번 문제에서는 상태는 결정적으로 전이되기 때문에, 상태 전이 확률은 $\int(s,a)$에 의해 결정된다.

- $s' = f(s,a)$이면 $p(s' \mid s,a) = 1$
- $s' \neq f(s,a)$이면 $p(s' \mid s,a) = 0$

이를 벨만 방정식에 대입하면 $s' = \int(s,a)$를 만족하는 s'에 해당하는 항만 남는다. 따라서 다음과 같이 간소화 할 수 있다.

$$s' = f(s,a) \text{ 이면}$$
$$v_\pi(s) = \sum_{a} \pi(a \mid s)\{r(s,a,s') + \gamma v_\pi(s')\}$$

이 식을 이번 문제에 대입할 수 있다.



그림을 보면 백업 다이어그램이 두 갈래로 나뉘어 있다. 하나는 0.5의 확률로 행동 Left를 선택하고, 상태는 전이되지 않는다. 보상은 -1이다. 이때 할인율 γ를 0.9로 설정하면 위의 식에서 Left를 선택하는 경우는 다음과 같다.

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

그림에서 또 다른 가능성은 0.5의 확률로 행동 Right를 선택하여, 상태 L2로 전이하고 보상 1을 얻는 경우이다. 이로부터 다음 식을 얻을 수 있다.

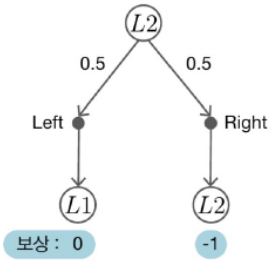
$$0.5\{1 + 0.9 v_\pi(L2)\}$$

지금까지의 내용을 벨만 방정식으로 나타내면 다음과 같다.

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

이 식이 상태 L1에서의 벨만 방정식이다.

이제 상태 L2에서의 벨만 방정식을 구해볼 수 있다.



$$v_\pi(L2) = 0.5\{0 + 0.9v_\pi(L1)\} + 0.5\{-1 + 0.9v_\pi(L2)\}$$
$$0.45\,v_\pi(L1) - 0.55v_\pi(L2) = 0.5$$

이제 알고 싶은 변수는 $v_\pi(L1)$과 $v_\pi(L2)$가 남았다. 그리고 다음의 두 방정식을 얻을 수 있다.

$$\begin{cases} -0.55v_\pi(L1) + 0.45v_\pi(L2) = 0 \\ 0.45v_\pi(L1) - 0.55v_\pi(L2) = 0.5 \end{cases}$$

연립 방정식이므로 각각을 구할 수 있다.

$$\begin{cases} v_\pi(L1) = -2.25 \\ v_\pi(L2) = -2.75 \end{cases}$$

이는 무작위 정책의 상태 가치 함수이다. 즉 상태 L1에서 무작위로 행동하면 앞으로 -2.25의 수익을 기대할 수 있다. 또한 $v_\pi(L1)$의 값이 $v_\pi(L2)$보다 큰 이유도 L1 옆에 사과가 있고 첫번째 행동에서 그 사과를 얻을 확률이 50%이기 때문이라고 이해할 수 있다.

3.2.2. 벨만 방정식의 의의

벨만 방정식을 통해 무한히 계속되는 계산을 유한한 연립방정식으로 변환할 수 있다. 이번처럼 행동이 무작위로 이루어지더라도 벨만 방정식을 이용하면 상태 가치 함수를 구할 수 있다.

또한 복잡한 문제라도 벨만 방정식을 이용해 연립방정식으로 표현할 수 있다. 그리고 연립방정식을 푸는 알고리즘을 이용하면 자동으로 상태 가치 함수를 구할 수 있다.

(벨만 방정식을 활용하면 상태가 n개 있는 문제에서 n개의 연립방정식으로 바꾸어 구할 수 있다.)

3.3. 행동 가치 함수 (Q 함수)와 벨만 방정식

행동 가치 함수는 상태 가치 함수와 마찬가지로 강화 학습 이론에 자주 등장하는 중요한 함수다. 지금까지는 상태 가치 함수를 사용하여 벨만 방정식을 도출했다. 이번 절에서는 행동 가치 함수의 정의식을 살펴본 후 행동 가치 함수를 이용한 벨만 방정식을 도출하겠다.

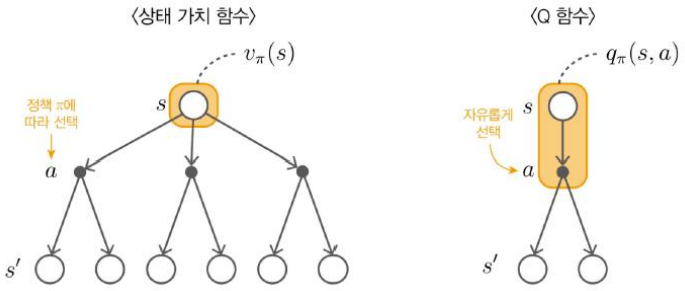
3.3.1. 행동 가치 함수

먼저 상태 가치 함수는 $\mathbb{E}[G_t|S_t = s]$로, 상태 가치 함수의 조건은 '상태가 s일 것'과 '정책이 π일 것' 두 가지이다. 그리고 조건에 '행동 a'를 추가할 수 있는데, 이것이 바로 행동 가치 함수(Q 함수)이다. 수식으로는 다음과 같다.

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

Q 함수는 시간 t일 때 상태 s에서 행동 a를 취하고, 시간 t+1 부터는 정책 π에 따라 행동을 결정한다. 이 때 얻을 수 있는 기대 수익이 $q_\pi(s,a)$이다. 행동 a는 정책 π와 무관하게 결정할 수 있으며, 다음 행동부터 정책 π를 따른다.

Q 함수는 상태 가치 함수에 행동 a를 조건으로 추가한 것이다. 백업 다이어그램으로 비교해 볼 수 있다.



상태 가치 함수에서의 행동 a는 정책 π에 따라 선택된다. 반면 Q 함수에서는 행동 a는 자유롭게 선택할 수 있다. 따라서 만약 Q함수의 행동 a를 정책 π에 따라 선택하도록 설계하면 Q 함수와 상태 가치 함수는 완전히 같아진다. 예를 들어 상태 s에서 취할 수 있는 행동 후보가 {a1,a2,a3} 세 가지가 잇다고 가정하고, 이 때 정책 π에 따라 행동한다고 한다.

- $\pi(a_1|s)$의 확률로 행동 a_1을 선택하는 경우 Q 함수는 $q_\pi(s,a_1)$
- $\pi(a_2|s)$의 확률로 행동 a_2를 선택하는 경우 Q 함수는 $q_\pi(s,a_2)$
- $\pi(a_3|s)$의 확률로 행동 a_3을 선택하는 경우 Q 함수는 $q_\pi(s,a_3)$

이 경우 기대 수익은 Q 함수의 가중 합으로 구할 수 있다.

$$\pi(a_1|s)q_\pi(s,a_1)+\pi(a_2|s)q_\pi(s,a_2)+\pi(a_3|s)q_\pi(s,a_3)$$
$$=\sum_{a=\{a_1,a_2,a_3\}}\pi(a|s)q_\pi(s,a)$$

이 식은 상태 가치 함수와 똑같은 조건에서의 기대 수익이다. 따라서 다음식이 성립한다.

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

3.2.2. 행동 가치 함수를 이용한 벨만 방정식

이어서 행동 가치 함수(Q 함수)를 이용한 벨만 방정식을 도출하겠다. 먼저 다음과 같이 Q 함수를 전개한다.

$$q_\pi(s,a)=\mathbb{E}_\pi[G_t|S_t=s,A_t=a]$$
$$=\mathbb{E}_\pi[R_t+\gamma G_{t+1}|S_t=s,A_t=a]$$

여기서 상태 s와 행동 a는 정해져 있다. 그렇다면 다음 상태 s'로의 전이 확률은 $p(s'|s,a)$이고 보상은 $r(s,a,s')$ 함수에 의해 주어진다. 이 점을 고려하면 위의 식을 다음과 같이 전개할 수 있다.

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

위 식을 이용하면 상태 가치 함수 $v_\pi(s')$는 Q 함수를 사용하여 다음처럼 쓸 수 있다.

$$q_\pi(s,a) = \sum_{s'} p(s' \mid s,a)\left\{ r(s,a,s') + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a') \right\}$$

위 식에서 a'는 시간 t+1에서의 행동이다. 이는 행동 가치 함수(Q 함수)를 이용한 벨만 방정식이다.

cf) 이해하기: 행동 가치 함수
행동 가치 함수는 특정 상태에서 어떤 행동 a를 취했을 때 기대할 수 있는 보상을 계산하는 방식이다. 행동 가치 함수를 통해 각 행동이 얼마나 좋은 지 알 수 있다. 현재 상태에서 행동을 한 후, 미래에서 최적의 행동을 한다고 가정했을 때의 기대 보상을 나타내는 것이 행동 가치 함수의 벨만 방정식이다. 상태 가치 함수는 현재 상태에서 특정 정책에 따라 행동했을 때의 기대 보상을 말하고, 행동 가치 함수는 특정 행동을 했을 때의 기대 보상을 의미한다. 따라서 정책을 모르거나, 최적 행동을 찾고 싶다면 행동 가치 함수를 사용하는 것이 유리하다.

3.4. 벨만 최적 방정식

벨만 방정식은 어떤 정책 π에 대해 성립하는 방정식이다. 하지만, 우리가 궁극적으로 찾으려는 것은 최적 정책이다. 최적 정책이란 모든 상태에서 상태 가치 함수가 최대인 정책이다.

3.4.1. 상태 가치 함수의 벨만 최적 방정식

벨만 방정식은 다음과 같다.

$$\begin{aligned} v_\pi(s) &= \sum_{a,s'} \pi(a \mid s)\, p(s' \mid s,a)\{r(s,a,s') + \gamma v_\pi(s')\} \\ &= \sum_{a} \pi(a \mid s) \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_\pi(s')\} \end{aligned}$$

벨만 방정식은 어떠한 정책에서도 성립한다. 따라서 최적 정책을 $\pi_*(a|s)$라고 하면 다음과 같은 벨만 방정식이 성립한다.

$$v_*(s) = \sum_{a} \pi_*(a \mid s) \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

이 식에서 최적 정책의 가치 함수는 $v_*(s)$이다. 이제 우리가 고민하고 싶은 문제는 최적 정책 $\pi_*(a|s)$에 의해 선택되는 행동 a이다. 최적 정책은 어떤 행동을 선택할까?

$$v_*(s) = \sum_{a} \pi_*(a|s) \sum_{s'} p(s'|s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

$$= \begin{cases} -2 & (a_1) \\ 0 & (a_2) \\ 4 & (a_3) \end{cases}$$

그림에는 세 개의 행동 후보 {a1, a2, a3}이 있고, 위의 값이 각각 -2, 0, 4라고 가정한다.

최적 정책이기 때문에 값이 최대인 행동 a3을 100% 확률로 선택해야 한다. 결정적 정책인 셈이다. 따라서 확률적 정책 $\pi_*(a|s)$는 결정적 정책 $\mu_*(s)$로 나타낼 수 있다. 그리고 항상 a3를 선택하기 때문에 $v_*(s)$의 값은 4가 된다.

이 예에서 최적 정책은 파란 부분의 값이 가장 큰 행동을 선택하고, 그 최댓값이 그대로 $v_*(s)$가 됨을 알 수

있다. 이를 수식으로 표현하면 다음과 같다.

$$v_*(s) = \max_a \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

이 식과 같이 최댓값은 max 연산자를 사용하여 표현할 수 있다. 그리고 이 식이 벨만 최적 방정식이다.

3.4.2. 행동 가치 함수의 벨만 최적 방정식

행동 가치 함수(Q 함수)에서도 벨만 최적 방정식을 구할 수 있다. Q 함수의 벨만 방정식은 다음과 같다.

$$q_\pi(s,a) = \sum_{s'} p(s' \mid s,a)\left\{r(s,a,s') + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\right\}$$

이 벨만 방정식은 모든 정책 π에 성립한다. 최적 정책 π_*에도 성립하므로 최적 정책 π_*를 대입할 수 있다.

$$q_*(s,a) = \sum_{s'} p(s' \mid s,a)\left\{r(s,a,s') + \gamma \sum_{a'} \pi_*(a' \mid s') q_*(s',a')\right\}$$

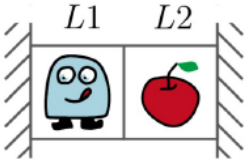
여기서는 π_*는 최적 정책이므로 max 연산자로 단순화 할 수 있다. 따라서 다음 식이 성립한다.

$$q_*(s,a) = \sum_{s'} p(s' \mid s,a)\left\{r(s,a,s') + \gamma \max_{a'} q_*(s',a')\right\}$$

이 식이 바로 Q 함수에 대한 벨만 최적 방정식이다.

3.5. 벨만 최적 방정식의 예

'두 칸짜리 그리드 월드'로 다시 문제를 다뤄볼 수 있다. 보상은 에이전트가 L1에서 L2로 이동할 때 +1, 벽에 부딪히면 -1이다. 사과는 몇 번이고 다시 생성된다.



3.5.1. 벨만 최적 방정식 적용

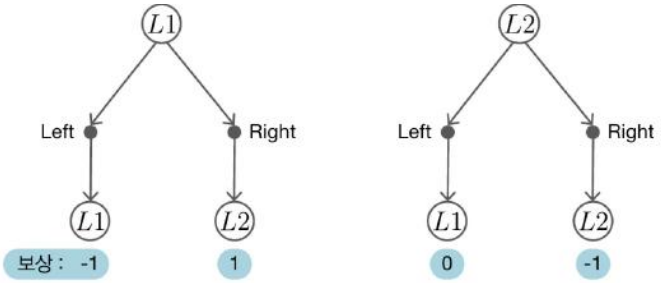
우리의 목표는 두 칸짜리 그리드 월드 문제에 벨만 최적 방정식을 적용하는 것이다. 벨만 최적 방정식은 다음과 같이 나타낼 수 있다.

$$v_*(s) = \max_a \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

또한 상태 전이가 결정적이라면 다음과 같이 단순화 할 수 있다.

$$s' = f(s,a)\text{일 때}$$
$$v_*(s) = \max_a \{r(s,a,s') + \gamma v_*(s')\}$$

이제 두 칸 짜리 그리드 월드에 위의 벨만 최적 방정식을 적용한다. 참고로 상태 L1과 L2를 시작 위치로 한 백업 다이어그램은 다음과 같다.



그림을 참고하여 할인율이 0.9일 때의 벨만 최적 방정식은 다음과 같이 구할 수 있다.

$$v_*(L1) = \max\begin{Bmatrix} -1+0.9v_*(L1), \\ 1+0.9v_*(L2) \end{Bmatrix}$$
$$v_*(L2) = \max\begin{Bmatrix} 0.9v_*(L1), \\ -1+0.9v_*(L2) \end{Bmatrix}$$

이 연립방정식으로 $v_*(L1)$과 $v_*(L2)$를 구할 수 있다. $v_*(L1) = 5.26$ $v_*(L2) = 4.73$이다. 따라서 상태 L1에서 더 좋은 보상을 얻을 수 있다. 하지만 우리가 궁극적으로 알고 싶은 것은 최적 정책이다. 최적 정책에 대해 알아보자.

3.5.2. 최적 정책 구하기

최적 행동 가치 함수 $q_*(s,a)$를 알고 있다고 가정해보자. 그렇다면 상태 s에서의 최적 행동은 다음과 같이 구할 수 있다.

$$\mu_*(s) = \underset{a}{\operatorname{argmax}}\, q_*(s,a)$$

argmax는 최댓값이 아니라 최댓값을 만들어내는 인수(행동 a)를 반환한다. 이 식과 같이 최적 행동 가치 함수를 알고 있는 경우, 함수의 값이 최대가 되는 행동을 선택하면 된다. 그러면 그 행동을 선택하는 것이 최적 정책이다.

또한 이 식을 참고할 수 있다.

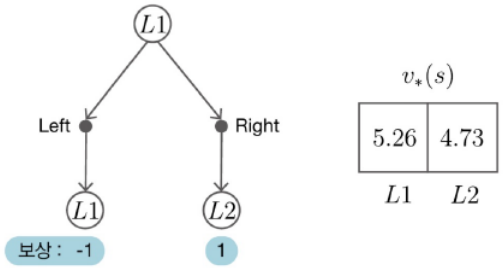
$$q_\pi(s,a) = \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_\pi(s')\}$$

이 식의 정책의 첨자 π를 최적 정책인 첨자 *로 대체할 수 있다. 그리고 위의 최적 행동 가치 함수의 식에 대입하면 다음과 같은 식이 만들어진다.

$$\mu_*(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

이렇게 최적 상태 가치 함수 $v_*(s)$를 사용하여 최적 정책 $\mu_*(s)$를 얻을 수 있다.

이 식을 이용하여 두 칸짜리 그리드 월드 문제의 최적 정책을 구할 수 있다. 앞서 구한 최적 상태 가치 함수인 $v_*(L1)$과 $v_*(L2)$를 참고하여 먼저 상태 L1에서의 최적 행동을 구해볼 수 있다.



그림에서 보듯 취할 수 있는 행동을 Left와 Right 두 가지이다. Left를 선택하면 상태 L1로 전이하여 보상 -1을 얻는다. 그러면 위의 식 $\sum_{s'} p(s' \mid s,a)\{r(s,a,s')+\gamma v_*(s')\}$ 부분의 값은 다음과 같다.

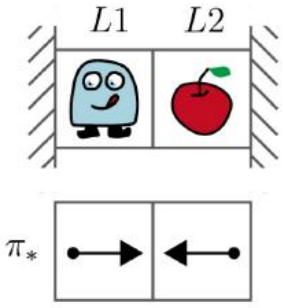
$$-1+0.9v_*(L1) = -1+0.9*5.26 = 3.734$$

한편, 행동 Right를 선택하여 상태 L2로 전이하여 보상 1을 얻는다. 이 경우의 값은 다음과 같다.

$$1+0.9v_*(L2) = 1+0.9*4.73 = 5.257$$

따라서 둘 중 값이 더 큰 행동은 Right이다. 상태 L1에서의 최적 행동은 Right라는 뜻이다. 같은 방식으로 상태 L2에서의 최적 행동을 찾으면 Left가 나온다.

최종적으로 최적 정책은 그림과 같이 L1에서는 오른쪽으로 L2에서는 왼쪽으로 이동하는 행동이 최적 정책이 된다.



이처럼 최적 상태 가치 함수를 알면 최적 정책을 구할 수 있다.