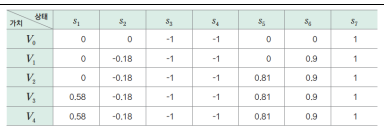
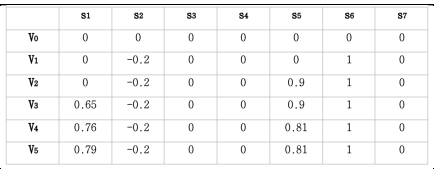
Do It! 강화학습 입문 정오표 (2021. 6. 26.)

위치	수정 전	수정 후
22 페이지	$P(s_1, a_{1\text{-}2}) = 0.8$ $P(s_1, a_{1\text{-}2}) = 0.2$	$P(s_5 \mid s_1, a_{1\text{-}2}) = 0.8$ $P(s_1 \mid s_1, a_{1\text{-}2}) = 0.2$
24 페이지	… 그에 따른 보상의 총합은 다음과 같이 나열할 수 있습니다.	… 그에 따른 보상의 총합은 다음과 같이 나열할 수 있습니다.[1] 1 각주: MDP 에서 보상은 상태와 행동의 함수 R(s, a)로 정의되고, 보상이 R(s) 즉 상태만의 함수로 정의되는 것은 마르코프 보상 과정 MRP, Markov reward process 라고 불립니다. 여기에서는 할인률의 개념을 간단하게 보여주기 위해 MDP 대신 MRP 의 수식을 사용하였습니다.
25 페이지	$G_t = \sum_{k=0}^{\infty} \gamma^k R(S_{t+k+1})$	$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
27 페이지	보상이 종단 상태에는 적혀 있지만 다른 상태에는 적혀 있지 않았죠? 종단 상태를 제외한 모든 상태의 가치를 0 으로 초기화하겠습니다. 종단 상태에는 더 이상 상태 전이가 없으니 보상은 곧 가치입니다. 즉, $V(s_3) = -1$, $V(s_4) = -1$, $V(s_7) = 1$ 입니다.	세 가지 종단 상태에는 보상이 주어져 있으며, 종단 상태를 제외한 다른 상태에는 별도의 보상이 주어지지 않습니다. 우선 모든 상태의 가치를 0으로 초기화하겠습니다.
27 페이지	s_3 의 가치: -1 가치의 기댓값: s_3 의 가치 -1 x 전이 확률 1 x 할인율 0.9 = -0.9	s_3 의 가치: 0 가치의 기댓값: s_3의 보상 -1 + s3의 가치 0 x 전이 확률 1 x 할인율 0.9 = -1
27 페이지	s_4 의 가치: -1 s_7 의 가치: 1 가치의 기댓값: ((s_4 의 가치 x 전이 확률 0.6) + (s_7 의 가치 x 전이 확률 0.4)) x 할인율 0.9 = -0.18	s_4 의 가치: 0 s_7 의 가치: 0 가치의 기댓값: [전이 확률 0.6 x (s_4의 보상 -1 + s_4의 가치 0 x 할인율 0.9) + 전이 확률 0.4 x (s_7의 보상 1 + s_7의 가치 0 x 할인율 0.9)] = -0.2
27 페이지	이 표를 참고하여 s_2에서 취할 수 있는 행동 $a_{2\text{-}1}$, $a_{2\text{-}2}$에서 가치의 기댓값을 비교해 볼까요? $a_{2\text{-}2}$를 취하면 -0.18이고 a2-1를 취하면 -0.9이므로	이 표를 참고하여s_2에서 취할 수 있는 행동$a_{2\text{-}1}$, $a_{2\text{-}2}$에서 가치의 기댓값을 비교해 볼까요? $a_{2\text{-}2}$를 취하면 -0.2이고 a2-1를 취하면 -1이므로 전자가 크다는 것을 알

	전자가 크다는 것을 알 수 있습니다. 이제 가치를 구할 수 있겠네요. s2 자체 보상은 0이므로 s2의 가치는 기댓값 -0.9에 현재 보상 0을 더해 -0.9임을 알 수 있습니다.	수 있습니다. 즉 s2의 가치는 취할 수 있는 행동이 가지는 기댓값 중 가장 높은 기댓값인 -0.2임을 알 수 있습니다.
27 페이지	$V(s)=R(s)+(V(s'))$	$V(s) = max_a \sum_{s'} P(s' \mid s,a)\,[R(s,a) + \gamma V(s')]$
28 페이지	그리고 종단 상태의 가치는 종단 상태의 보상과 같으므로 이미 알려진 값이지요.	(삭제) S6, a6-1: 0 S6, a6-2: 1
28 페이지	s에서 a를 선택해 도달하는 다음 상태 s' 의 가치 V(s')와 P(s' | s, a)를 곱함 도달 가능한 모든 상태 s '에 대해 P(s' |s, a) × V(s')를 합산한 기댓값을 구함	s에서 a를 선택해 도달하는 다음 상태 s' 의 가치 V(s')에 할인률 γ를 곱하고, 이 때의 보상인 R(s,a)를 더하고, 여기에 P(s'|s, a)를 곱함 도달 가능한 모든 상태 s '에 대해 P(s'|s, a)[R(s,a) ×γV(s')]의 기댓값을 구함
28 페이지	a_*의 기댓값에 할인율 c를 곱하고, 상태 s의 보상 R(s)를 더한 값 $V_n(s)$를 구함	a_*의 기댓값으로 값 $V_n(s)$를 업데이트
29 페이지		
56 페이지	알고리즘을 코드에 그대로 적용해 실습해 봅시다.	(삭제)
57~60 페이지		(전체 삭제)
61 페이지	02단계 코드 실행해 결과 살펴보기 다음 명령어를 실행하여 결과를 살펴봅시다. 이때 반복 횟수나 파라미터를 바꿔 보면서 결과가 어떻게 달라지는지도 확인해 보세요. (터미널 커맨드)	(삭제)
142 페이지	조금 더 자세히 설명하면, PPO는 DQN과 달리 완전한 온라인 학습 방법으로 에이전트가 환경으로부터 얻는 샘플을 바로 사용합니다. 샘플 효율성이 높다는 의미죠.	(삭제)