강화학습 Summary

강화학습 요약 리포트 (Summary Mode)
생성 시각: 2025-09-26 08:46
ZIP 내 PDF 수: 16
사용 폰트: HCR Batang

이 문서는 ZIP에 포함된 강화학습 강의/노트 PDF를 바탕으로
핵심 개념을 자립형 학습 자료로 요약·재구성한 것입니다.
정의→원리→수식→예시→응용→주의점의 흐름으로 정리하고,
필요한 경우 간단한 도식과 표를 덧붙였습니다.
(자동 추출된 텍스트가 부족한 경우, 개념 중심 요약으로 대체될 수 있습니다.)

목차

1. MDP 기본기

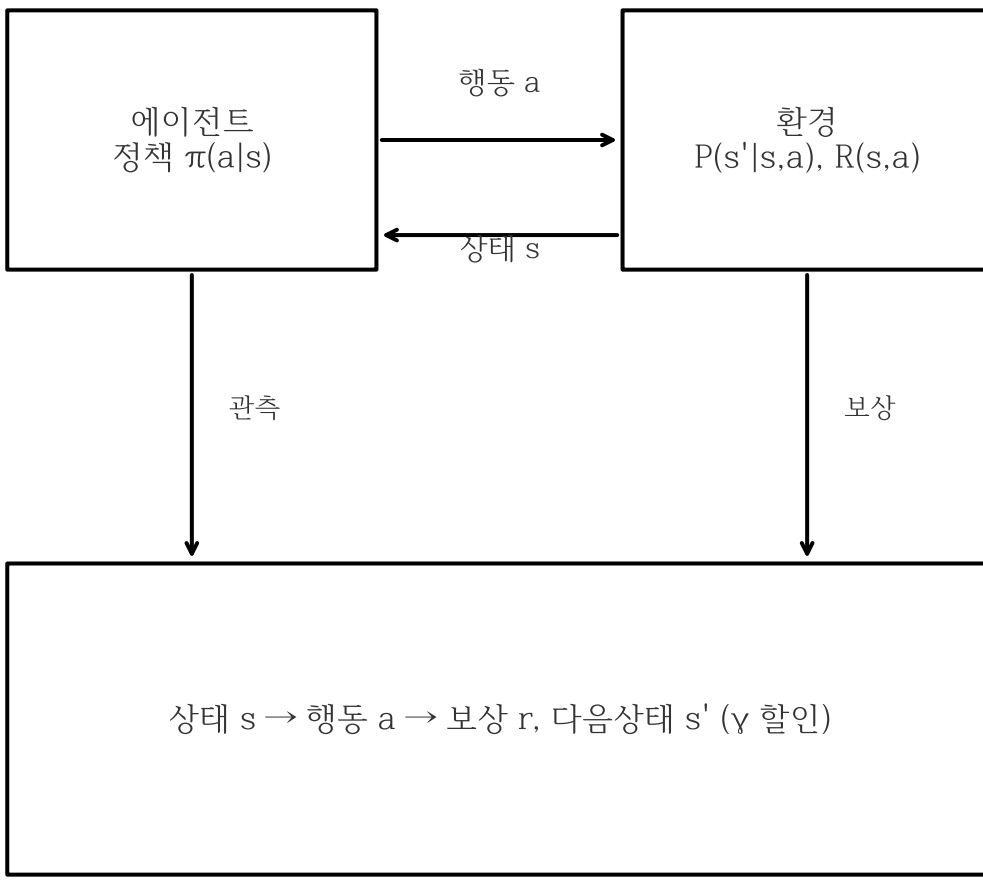
정의 및 배경:
마르코프 결정 과정(MDP)은 (S, A, P, R, γ)로 정의됩니다.
상태 S, 행동 A, 전이확률 P(s'|s,a), 보상 R(s,a), 감가율 γ로 구성되며
미래가 현재 상태에만 의존하는 마르코프 성질을 가정합니다.

수학적 기반:
목표는 누적 보상의 기대값을 최대화하는 정책 π*를 찾는 것입니다.
에이전트-환경 상호작용은 시계열로 전개되며, γ∈[0,1)에서 수렴성이 보장됩니다.

예시/응용:
격자 세계(Gridworld), 재고관리, 경로계획 등 다양한 의사결정 문제를 MDP로 모델링합니다.

주의점:
상태공간이 크면 정확한 모델링(P,R)을 구하기 어렵고, 근사/표현 학습이 필요합니다.

그림: MDP 상호작용 개요



2. 가치함수 (V, Q)

핵심 정의:
상태가치 Vπ(s)=Eπ[∑t γ^t r_t | s0=s],
행동가치 Qπ(s,a)=Eπ[∑t γ^t r_t | s0=s, a0=a]로
정의됩니다.

원리:
가치함수는 정책 품질을 정량화하는 지표로, 정책 개선(policy improvement)과
평가(policy evaluation)의 핵심입니다.

주의점:
표본 추정 분산, 편향-분산 트레이드오프, 함수근사 시 과적합과 안정성
이슈에 유의합니다.

3. 벨만 방정식

정의:
벨만 기대 방정식과 최적성 방정식은 가치함수의 재귀적 관계를 나타냅니다.
Vπ(s)=Eπ[r+ɣVπ(s')], Qπ(s,a)=Eπ[r+ɣ E_{a'~π}
Qπ(s',a')].
최적 V*와 Q*는 최대 연산을 포함합니다.

의미:
고정점 관점에서 수치적 반복(값 반복/정책 반복)의 수렴 근거를 제공합니다.

4. DP / MC / TD 비교

동적계획법(DP):
모델(P,R)을 안다는 가정 하에 정책평가/개선을 반복해 최적 정책을 찾습니다.

몬테카를로(MC):
에피소드 완결 후 반환값을 평균해 가치 추정. 분산이 크지만 비편향적입니다.

TD(Temporal-Difference):
한 스텝 앞선 예측으로 부트스트랩. MC와 DP 사이의 절충으로 온라인/증분 학습에 유리합니다.

비교:
MC는 완전반환, TD는 부트스트랩. 데이터/분산/편향 특성이 다릅니다.

5. Q-learning

Q-learning(오프-폴리시):
업데이트: $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a') - Q(s,a)]$.
탐험은 ε-greedy 등으로 수행하되, 목표는 최대행동 기준입니다.
수렴:
충분한 탐험과 학습률 조건 하에서 테이블형은 수렴이 알려져 있습니다.

Q-learning 절차(개요)

1. 상태 s 관측

↓

2. ε-greedy로 a 선택

↓

3. r, s' 관측

↓

4. $Q \leftarrow Q + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

↓

5. $s \leftarrow s'$, 반복

6. SARSA

SARSA(온-폴리시):
업데이트: $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma\ Q(s',a')-Q(s,a)]$,
여기서 a'는 실제로 취한 행동.
특징:
탐험 정책의 위험을 반영해 보수적인 학습이 가능하며, 특정 환경에서 더 안전한 경향이 있습니다.

SARSA 절차(개요)

1. 상태 s 관측

↓

2. ε-greedy로 a 선택

↓

3. r, s', a' 관측

↓

4. $Q \leftarrow Q + \alpha[r + \gamma\, Q(s',a') - Q(s,a)]$

↓

5. $s \leftarrow s'$, $a \leftarrow a'$, 반복

7. 탐색 전략 (ε-greedy 등)

탐색-활용 균형:
ε-greedy는 확률 ε로 랜덤탐색, 1−ε로 최고 Q를 선택합니다.
스케줄링(ε 감소), 소프트맥스(볼츠만 탐색), UCB 등 대안이 있습니다.

주의점:
탐험 부족은 지역해 정체, 과도한 탐험은 수렴 지연을 초래합니다.

8. 정책경사 & 베이스라인

정책경사:
목표 $J(\theta)=E[R]$에 대해 $\nabla\theta\, J(\theta)=E[\nabla\theta \log \pi\theta(a|s)\cdot G]$
형태.
REINFORCE는 에피소드 반환을 사용하며 분산이 큽니다.

베이스라인/어드밴티지:
분산 감소를 위해 상태별 베이스라인 $b(s)$를 빼서 $\nabla\theta\, J \approx E[\nabla\theta \log \pi\theta(a|s)\cdot(G-b)]$.
Actor-Critic은 함수근사로 정책과 가치(또는 어드밴티지)를 함께 학습합니다.

9. 학습 체크포인트

학습 체크포인트:
- MDP 다섯 요소와 마르코프 성질을 설명할 수 있는가?
- V, Q 정의와 벨만 방정식을 쓰고 해석할 수 있는가?
- DP/MC/TD의 차이와 장단점을 비교할 수 있는가?
- Q-learning과 SARSA 업데이트식을 정확히 외우고 설명 가능한가?
- ε-greedy/소프트맥스 탐색의 장단점을 이야기할 수 있는가?
- 정책경사 기본식과 베이스라인의 역할을 이해했는가?

10. 혼동 포인트 & 실수

혼동 포인트 & 자주 하는 실수:
- V와 Q의 정의 혼동, 벨만 기대 vs 최적성 구분 실수
- MC와 TD의 편향/분산 특성 혼동
- SARSA/Q-learning의 온/오프-폴리시 차이 누락
- ε 스케줄 미설계로 탐색 부족/과다
- 정책경사에서 리워드-노멀라이즈, 어드밴티지 추정 누락

11. 참고 문서 리스트 (ZIP)

- 강화학습_#11-1_(250512).pdf
- 강화학습_#11-2_(240513).pdf
- 강화학습_#12-1_(250519).pdf
- 강화학습_#13-1_(250527).pdf
- 강화학습_#13-1_필기.pdf
- 강화학습_#14-1_(250602).pdf
- 강화학습_#15-1_(250609).pdf
- 강화학습_#2-1_(250310).pdf
- 강화학습_#2-2_(250311).pdf
- 강화학습_#3-1_(250317).pdf
- 강화학습_#3-2_(250318).pdf
- 강화학습_#4-2_(250325).pdf
- 강화학습_#6-1_(250408).pdf
- 강화학습_#8-1_(250422).pdf
- 강화학습_#9-1_(250428).pdf
- 강화학습_#9-2_(250429).pdf

12. 심화 학습 자료 / 마무리

심화 학습 자료(개념 중심):
- Sutton & Barto, Reinforcement Learning: An Introduction
- Bertsekas & Tsitsiklis, Neuro-Dynamic Programming
- Silver의 UCL 강의 노트/동영상
(원문 PDF 내 정확한 페이지 인용은 ZIP 텍스트 추출 품질에 따라 제한될 수 있습니다.)

끝