강화학습 요약 & 학습 로드맵 (Summary)

과목: Reinforcement Learning | 작성일: 2025-09-29

1. 전체 학습 로드맵

이 문서는 강화학습 핵심 주제를 정의→원리→수식→예시→응용→주의사항 순으로 정리합니다.
토픽은 MDP, 가치함수, 벨만 방정식, DP/MC/TD,
Q-learning/SARSA, 탐색전략(ε-greedy), 정책경사(Policy Gradient), Baseline/Advantage, Actor-Critic까지
포괄합니다.

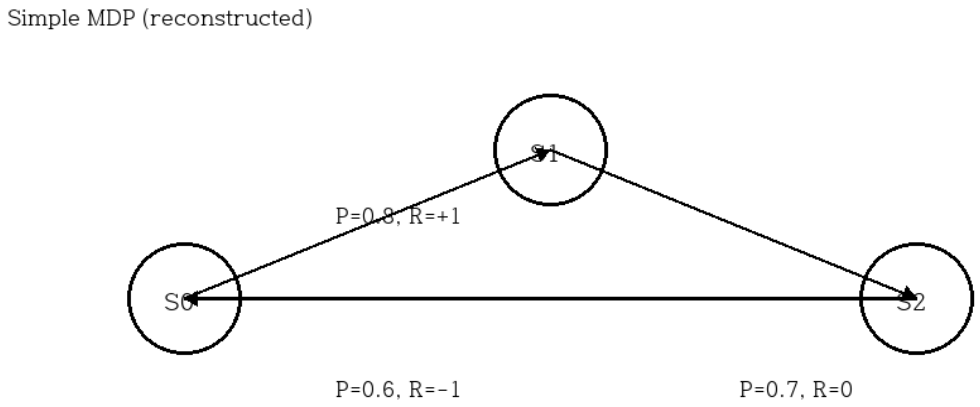

그림 1. 간단한 MDP 구조(재구성)

2. MDP (Markov Decision Process)

MDP는 (S, A, P, R, γ)로 정의됩니다. 마르코프성은 다음 상태가 현재 상태와 행동에만 의존함을 의미합니다. 목표는 누적 보상의 기댓값을 극대화하는 정책 π를
찾는 것입니다.

- 정의: 상태 S, 행동 A, 전이확률 P(s'|s,a), 보상 R(s,a,s'), 할인율 γ.
- 에피소드/무한지평 구분: 종료 상태 존재 여부.
- 예시: 그리드월드, 재고관리, 추천시스템 간단 모델링.

3. 가치함수 V, Q, Advantage

$V^{\pi}(s)$는 상태 s에서 정책 π가 따를 때의 기대 누적 보상입니다. $Q^{\pi}(s,a)$는 상태-행동 가치입니다.
Advantage
$A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s)$는 평균 대비 상대적 우수성을 나타냅니다.

- V는 상태 수준 평가, Q는 행동 비교에 직접적.
- A는 분산을 낮추고 정책경사에서 안정적 업데이트에 도움.

4. 벨만 방정식

가치함수는 재귀적으로 표현됩니다. 최적가치 V*와 Q*는 벨만 최적 방정식을 만족합니다.

- V*(s)=max_a Σ_s' P(s'|s,a)[r + γ V*(s')]
- Q*(s,a)=Σ_s' P(s'|s,a)[r + γ max_a' Q*(s',a')]
- 정책 평가/개선, 가치 반복의 이론적 기반.

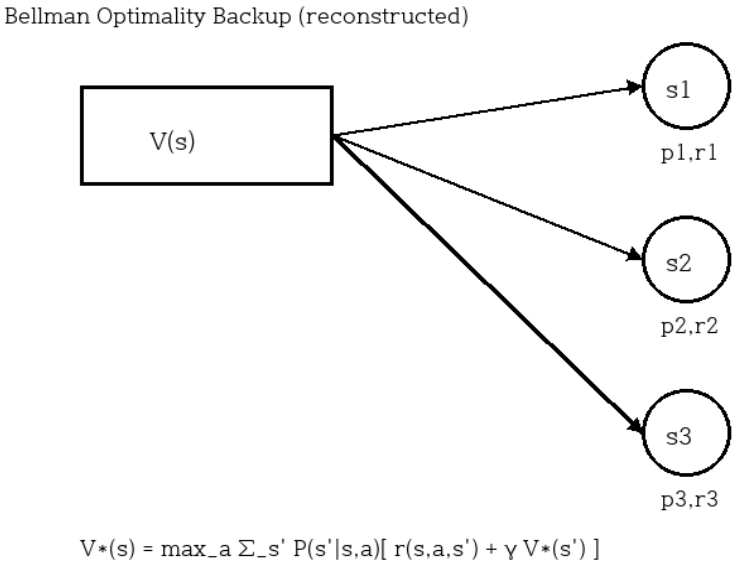

그림 2. 벨만 최적 백업(재구성)

5. Dynamic Programming (DP)

모델(P, R)을 알고 있을 때 가치 반복(Value Iteration)과 정책 반복(Policy Iteration)으로 최적 정책을 계산합니다.

- 정책 평가(벨만 기대 방정식 반복) ↔ 정책 개선(그리디).
- 수렴 보장(유한 MDP, $\gamma<1$).
- 상태공간이 크면 계산량 폭증 → 근사/표본기반 방법 필요.

6. Monte Carlo (MC)

모델 없이 에피소드 종료 후 반환을 평균하여 가치 추정. 부트스트래핑을 사용하지 않습니다.

- 장점: 편향 적고 구현 단순.
- 단점: 분산 큼, 에피소드 종료 필요.
- First-visit/Every-visit, 탐색 시작(Exploring Starts).

7. Temporal-Difference (TD)

부트스트래핑을 통해 한 단계 앞선 추정으로 즉시 업데이트합니다. TD(0), n-step, Eligibility Traces(λ-return) 등이 있습니다.

- 장점: 온라인/부트스트래핑, 데이터 효율 좋음.
- 단점: 편향 가능, 하이퍼파라미터 민감.
- SARSA/Q-learning의 기반 아이디어.

8. Q-learning (Off-policy)

타깃에 max_a' Q(s',a')를 사용하므로 오프폴리시입니다.

- 업데이트: Q←Q + α[r + γ·max_a' Q(s',a') − Q].
- 경험재현(Replay), 타깃네트워크로 안정성 향상(DQN 계열).
- 탐색은 별도의 ε-greedy 등으로 수행.

9. SARSA (On-policy)

실제 선택한 a'에 따른 Q(s',a')로 타깃을 구성합니다.

- 업데이트: Q←Q + α[r + γ·Q(s',a') − Q].
- 정책과 학습 타깃이 일치(온폴리시).
- ε-greedy와 같이 사용 시 안전한 경향.

10. 탐색 전략 (ε-greedy 등)

ε-greedy는 확률 ε로 무작위 행동을 선택해 탐색을 보장합니다. ε는 점차 감소(선형/지수)시키는 것이 일반적입니다.

• ε가 너무 크면 무작위성↑ 성능 저하, 너무 작으면 탐색 부족.
• 대안: 소프트맥스/볼츠만, UCB, 탐색 보너스, ε 스케줄 설계.

11. 정책경사 (Policy Gradient)

정책을 직접 미분해 업데이트합니다. REINFORCE는 반환 G_t를 사용하고, Actor-Critic은 baseline/critic으로 분산을 낮춥니다.

• 목표 J(θ)=E[∑ γ^t r_t], 경사 추정 ∝ E[∇θ log π_θ(a|s)·(G_t − b(s))].
• 엔트로피 정규화로 탐색 유지, KL 제약(TRPO/PPO)로 안정화.

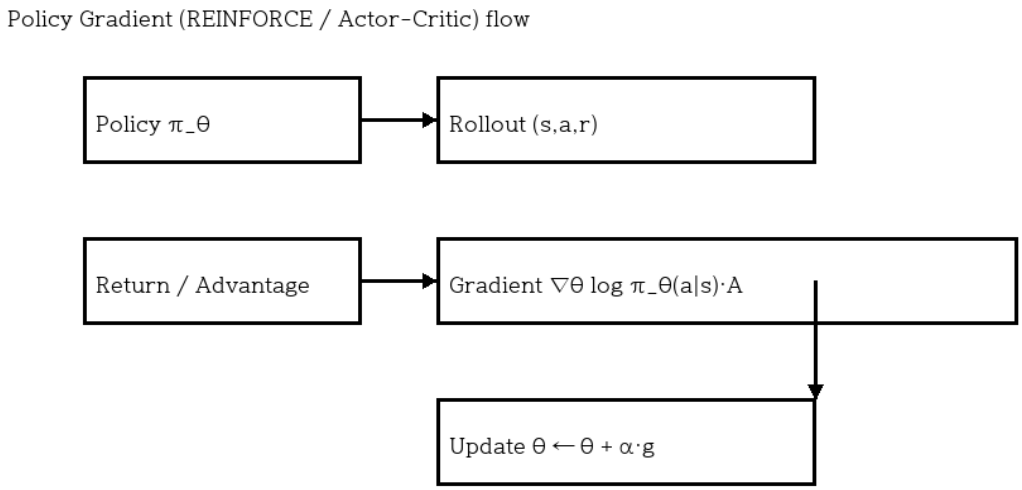

그림 3. 정책경사/Actor-Critic 흐름(재구성)

12. Baseline/Advantage & Actor-Critic

baseline(주로 V(s))은 분산을 낮추고 편향은 증가시키지 않습니다. Advantage는 상대적 품질을 반영합니다.

- A(s,a)=Q−V, GAE(λ)로 바이어스-분산 절충.
- Actor(정책) + Critic(가치) 공동 학습 구조.

13. 비교표: DP vs MC vs TD

항목	DP	MC	TD
모델 필요	필요(P,R)	불필요	불필요
부트스트래핑	예	아니오	예
업데이트 시점	반복적 계산	에피소드 종료 후	스텝별 온라인
편향/분산 경향	편향↓, 분산↓	편향↓, 분산↑	편향↑ 가능, 분산↓
대표 알고리즘	Policy/Value Iteration	MC 예측/제어	SARSA, Q-learning

14. 자주 하는 실수와 혼동 포인트

• Q-learning을 온폴리시로 착각: 타깃에 max_a'를 쓰므로 오프폴리시.
• MC와 TD의 차이를 '모델 유무'만으로 설명: MC는 부트스트래핑 미사용, TD는 사용.
• ε 스케줄을 고정: 초기에 크게, 점차 줄여 수렴을 유도.
• Advantage와 Q/V 혼동: A=Q−V이며 평균 대비 상대적 우수성.
• DP 적용 조건 간과: P, R 모델 가정 필요.

15. 학습 체크포인트

☐ MDP 5요소 정의를 정확히 말할 수 있는가?
☐ V/Q/Advantage의 차이와 역할을 설명할 수 있는가?
☐ 벨만 기대/최적 방정식을 유도 형태로 쓸 수 있는가?
☐ DP vs MC vs TD의 차이와 장단점을 사례와 함께 말할 수 있는가?
☐ Q-learning/SARSA 업데이트 식을 암기하지 않고도 유도할 수 있는가?
☐ ε 스케줄 설계(선형/지수)에서 장단점을 설명할 수 있는가?
☐ 정책경사의 경사 추정식과 baseline의 효과(분산 감소)를 이해하는가?

16. 추가 연습 방향

- Gridworld에서 DP/MC/TD 각각으로 V(s) 또는 Q(s,a) 추정 실습.
- ε 스케줄(선형, 지수, cosine) 비교 실험 그래프화.
- REINFORCE vs Actor-Critic(Advantage) 수렴 속도/분산 비교 실험.
- DQN의 타깃 네트워크/리플레이 버퍼 유무 비교.

17. 심화 학습 자료

- Sutton & Barto, Reinforcement Learning: An Introduction (2nd).
- David Silver의 강화학습 강의 시리즈.
- OpenAI Spinning Up: Policy Gradient, PPO, TRPO 개요.
- DQN, DDPG, A3C/A2C, PPO 등의 원 논문 및 튜토리얼.

※ 본 요약은 강의 PDF를 바탕으로 핵심 개념을 재구성한 학습용 문서입니다. 다이어그램은 관계만 보존하여 새롭게 그렸습니다.