강화학습 시험 해설 및 정답

구성: 정답 근거 → 오답 분석 → 참고식/수식 (가능 시) → 출처

문항 1 [MDP] · 난이도: hard
정답: D
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 상태 집합 S — 핵심 정의와 불일치하여 오답
B) 행동 집합 A — 핵심 정의와 불일치하여 오답
C) 전이 확률 P — 핵심 정의와 불일치하여 오답
D) 관찰 확률 O — 벨만 최적 연산/가치 정의에 부합하여 정답
출처: 개념 해설(출처 미기재)

문항 2 [Value/Action Value] · 난이도: easy
정답: A
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) V(s)=max_a Q(s,a) — 벨만 최적 연산/가치 정의에 부합하여 정답
B) Q(s,a)=min_s V(s) — 핵심 정의와 불일치하여 오답
C) V(s)=Σ_a Q(s,a) — 핵심 정의와 불일치하여 오답
D) Q(s,a)=γ·V(s) — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 3 [Bellman Equations] · 난이도: medium
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 탐욕적 정책의 무작위성 — 핵심 정의와 불일치하여 오답
B) 가치의 분해와 재귀적 정의 — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 모든 상태에서 동일 보상 — 핵심 정의와 불일치하여 오답
D) 할인율을 0으로 고정 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 4 [DP/MC/TD] · 난이도: easy
정답: C
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) MC는 부트스트래핑, TD는 에피소드 전체 반환 사용 — 핵심 정의와 불일치하여 오답
B) MC는 모델 필요, TD는 모델 불필요 — 핵심 정의와 불일치하여 오답
C) MC는 에피소드 종료 후 갱신, TD는 단계별 부트스트래핑 — 벨만 최적 연산/가치 정의에 부합하여 정답
D) 둘 다 항상 편향 없음 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 5 [Q-learning] · 난이도: medium
정답: C
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 현재 정책에 따른 다음 행동의 가치 — 핵심 정의와 불일치하여 오답
B) 무작위 행동의 평균 가치 — 핵심 정의와 불일치하여 오답
C) max_a' Q(s',a') — 벨만 최적 연산/가치 정의에 부합하여 정답
D) V(s') — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 6 [SARSA] · 난이도: medium
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) r + γ max_a' Q(s',a') — 핵심 정의와 불일치하여 오답
B) r + γ Q(s',a') (a'는 실제 선택) — 벨만 최적 연산/가치 정의에 부합하여 정답
C) r + γ V(s') — 핵심 정의와 불일치하여 오답
D) r + Q(s,a) — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 7 [Exploration (ε-greedy)] · 난이도: medium
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 탐색을 늘려 수렴을 방해 — 핵심 정의와 불일치하여 오답
B) 초기 탐험 후 수렴을 위해 탐색 축소 — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 항상 최적 행동만 선택하기 위해 — 핵심 정의와 불일치하여 오답
D) 보상을 0으로 만들기 위해 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 8 [Policy Gradient/Baselines] · 난이도: hard
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 편향을 증가 — 핵심 정의와 불일치하여 오답
B) 분산을 감소 — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 수렴 속도 저하 — 핵심 정의와 불일치하여 오답
D) 학습률 제거 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 9 [MDP] · 난이도: medium
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 정책 확률 π(a|s) — 핵심 정의와 불일치하여 오답
B) 전이 P(s'|s,a) — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 즉시 보상 r — 핵심 정의와 불일치하여 오답
D) 가치함수 V(s) — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 10 [Bellman Equations] · 난이도: medium
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 폴리시 이벨류에이션의 잔차 계산 — 핵심 정의와 불일치하여 오답
B) 벨만 백업(최적가치) — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 모델 프리 샘플 평균 — 핵심 정의와 불일치하여 오답
D) 정책 경사 추정 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 11 [DP/MC/TD] · 난이도: hard
정답: B
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 모델이 필요 없다 — 핵심 정의와 불일치하여 오답
B) 전이 확률과 보상 모델을 알고 있어야 한다 — 벨만 최적 연산/가치 정의에 부합하여 정답
C) 에피소드가 유한할 필요가 없다 — 핵심 정의와 불일치하여 오답
D) 오프폴리시만 가능하다 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 12 [Exploration (ε-greedy)] · 난이도: medium
정답: C
정답 근거: 정의 및 수식에 근거하여 해당 선택지가 옳습니다.
A) 탐색 부족 — 핵심 정의와 불일치하여 오답
B) 조기 수렴 — 핵심 정의와 불일치하여 오답
C) 과도한 무작위성으로 성능 저하 — 벨만 최적 연산/가치 정의에 부합하여 정답
D) 학습률 폭증 — 핵심 정의와 불일치하여 오답
출처: 개념 해설(출처 미기재)

문항 13 [Value/Action Value] · 난이도: easy
모범답안(요약): 핵심정의: V(s)는 상태 s에서의 기대 누적 보상, Q(s,a)는 상태-행동 쌍에서의 기대 누적 보상입니다. 예시/직관: 정책 개선은 행동 수준의 비교가 직접
가능해야 하며, Q는 행동 간 비교를 직접 제공합니다. 주의점: 정책에 따른 기대가치/행동가치의 의존성을 분명히 구분하고, 그리디 개선 시 Q의 최대값을
사용합니다.
출처: 개념 해설(출처 미기재)

문항 14 [DP/MC/TD] · 난이도: medium
모범답안(요약): 핵심정의: V(s)는 상태 s에서의 기대 누적 보상, Q(s,a)는 상태-행동 쌍에서의 기대 누적 보상입니다. 예시/직관: 정책 개선은 행동 수준의 비교가 직접
가능해야 하며, Q는 행동 간 비교를 직접 제공합니다. 주의점: 정책에 따른 기대가치/행동가치의 의존성을 분명히 구분하고, 그리디 개선 시 Q의 최대값을

사용합니다.
출처: 개념 해설(출처 미기재)

문항 15 [Q-learning] · 난이도: easy
모범답안(요약): 핵심정의: V(s)는 상태 s에서의 기대 누적 보상, Q(s,a)는 상태-행동 쌍에서의 기대 누적 보상입니다. 예시/직관: 정책 개선은 행동 수준의 비교가 직접
가능해야 하며, Q는 행동 간 비교를 직접 제공합니다. 주의점: 정책에 따른 기대가치/행동가치의 의존성을 분명히 구분하고, 그리디 개선 시 Q의 최대값을
사용합니다.
출처: 개념 해설(출처 미기재)

문항 16 [Policy Gradient/Baselines] · 난이도: hard
모범답안(요약): 핵심정의: V(s)는 상태 s에서의 기대 누적 보상, Q(s,a)는 상태-행동 쌍에서의 기대 누적 보상입니다. 예시/직관: 정책 개선은 행동 수준의 비교가 직접
가능해야 하며, Q는 행동 간 비교를 직접 제공합니다. 주의점: 정책에 따른 기대가치/행동가치의 의존성을 분명히 구분하고, 그리디 개선 시 Q의 최대값을
사용합니다.
출처: 개념 해설(출처 미기재)

문항 17 [Exploration (ε-greedy)] · 난이도: medium
모범답안(요약): 핵심정의: V(s)는 상태 s에서의 기대 누적 보상, Q(s,a)는 상태-행동 쌍에서의 기대 누적 보상입니다. 예시/직관: 정책 개선은 행동 수준의 비교가 직접
가능해야 하며, Q는 행동 간 비교를 직접 제공합니다. 주의점: 정책에 따른 기대가치/행동가치의 의존성을 분명히 구분하고, 그리디 개선 시 Q의 최대값을
사용합니다.
출처: 개념 해설(출처 미기재)

문항 18 [Bellman Equations] · 난이도: medium
정답: 6.5 value
문제 해석: 한 단계 최적 타깃 = r + γ·max_a' Q(s',a'). 공식/대입: 2 + 0.9×5.0 = 6.5000. 계산: 6.5000 → 반올림(소수1자리)
= 6.5. 단위 점검: value.
출처: 개념 해설(출처 미기재)

문항 19 [Q-learning] · 난이도: easy
정답: 3.8 value
업데이트 식: Q←Q + α[r + γ·max Q' − Q]. 대입: 3.0 + 0.5[1.0 + 0.9×4.0 − 3.0] = 3.0 + 0.5[1.0 + 3.6 −
3.0] = 3.0 + 0.5×1.6 = 3.8000. 반올림(소수1자리): 3.8 value.
출처: 개념 해설(출처 미기재)

문항 20 [Policy Gradient/Baselines] · 난이도: easy
개요(Outline): ① 기본 원리: 정책 π의 성능 지표 J(θ)에 대해 ∇J(θ)를 추정하여 업데이트. ② baseline/critic: REINFORCE는 Monte Carlo 반환으로,
Actor-Critic은 상태가치/어드밴티지로 분산을 낮춤. ③ 장단점: REINFORCE는 단순하나 분산 큼, Actor-Critic은 안정적이나 편향·튜닝 이슈. ④
한계/안정화: 학습률 스케줄, 엔트로피 정규화, advantage 정규화, trust-region/클립 손실 등.
모범답안(요약): 정의: 정책경사는 매개변수화된 정책 π_θ(a|s)의 성능 지표 J(θ)=E[∑γ^t r_t]에 대해 경사를 추정해 θ를 업데이트하는 방법이다. 비교: REINFORCE는
에피소드 반환을 이용해 ∇J를 추정하는 반면, Actor-Critic은 가치함수(또는 advantage) 추정치를 사용하여 분산을 낮춘다. 한계: 고분산/지연 보상 문제,
critic의 편향, 불안정한 공학적 튜닝. 실무 팁: baseline/advantage 정규화, KL 제약(TRPO/PPO 류), 타겟 네트워크/EMA, 학습률·배치
사이즈 스케줄링 등을 통해 안정성을 높인다.
출처: 개념 해설(출처 미기재)