R(s,a) 설계 보고서

1. 개요

본 보고서는 기업 간 협상 시뮬레이션을 위한 강화학습 에이전트의 보상함수 설계 내용을 상세히 기술합니다.
본 프로젝트의 에이전트는 구매자 역할을 수행하며, 목표는 판매자가 제시하는 가격을 최대한 낮추어 유리한 조건으로 협상하는 데 있습니다.

2. 보상함수 정의

에이전트가 상태 ss에서 행동 aa를 수행했을 때 받는 보상 R(s,a)R(s,a)는 다음과 같이 정의됩니다.

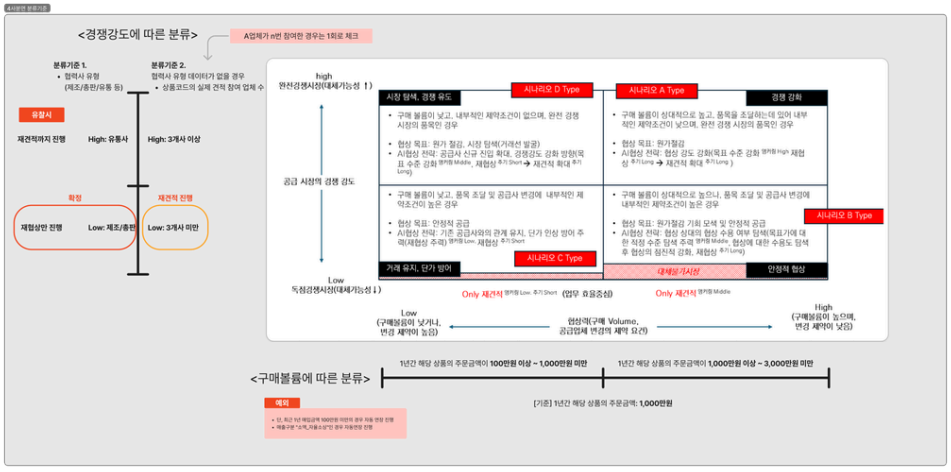
$$R(s,a) = W \times \frac{A}{P} + (1 - W) \times End$$

변수 설명

- A (앵커링 목표가): 구매자가 목표로 하는 기준 가격
- P (상대방 제안가): 판매자가 제시한 현재 가격
- End∈{0,1} (종료 여부): 협상이 종료되면 1, 진행 중이면 0
- W (가중치): 가격 보상과 종료 보상 간 비중 조절용 실수 값,

3. 가중치 W 계산 방식

$$W = \frac{S_n + PZ_n}{2}$$

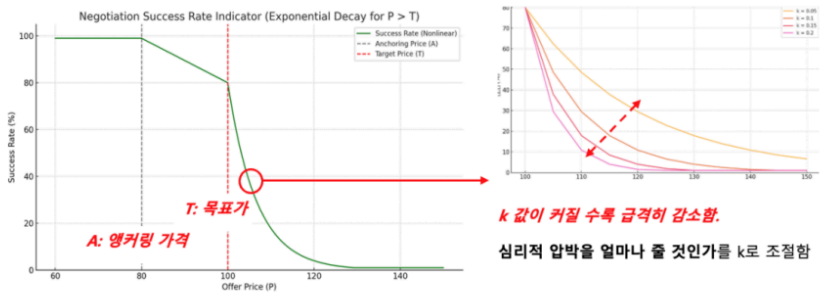
- S_n (시나리오 변수): 협상 상황별 가중치, 협상 종료와 가격 개선 간 중요도를 반영함
  - 

    - 값이 클수록 가격 개선이 더 중요함을 의미
      - S_1 = A,

    - S_2 = D
    - S_3 = C
    - S_4 = B
  - 설정값:

$$S_1 = 1,\ S_2 = 0.75,$$
$$S_3 = 0.5,\ S_4 = 0.25$$

- PZ_n (Price Zone 변수): 현재 가격 구간 가중치, 목표 가격과 상대 제안가 간 차이에 따른 영향도 반영
  - 값이 작을수록 제안 가격이 목표 가격에 가까움



  -
  - P < A → PZ_1
  - A < P < T → PZ_2
  - T < P → PZ_3
  - 설정값:

$$PZ_1 = 0.1,\ PZ_2 = 0.5,\ PZ_3 = 1$$

4. 보상함수 구성 요소별 해석

4.1 가격 비율 보상

- 목표 가격 대비 상대 가격 비율로, 가격이 낮을수록(좋을수록) 값이 커짐
- 가중치 W로 시나리오별 및 가격대별 중요도를 조절
- 구매자 관점에서 낮은 가격 유도에 따른 긍정적 피드백 역할 수행

4.2 종료 보상

- 협상 종료 시점에만 보상이 부여됨 (종료 = 1)
- 1-W는 가격 개선 보상의 보완적 역할 수행
- 협상 종료를 유도하는 인센티브 제공

5. 설계 의도 및 효과

- **가격 개선과 협상 종료 간 균형 조절**

- 시나리오별로 가격 개선과 종료 중 어느 쪽을 더 중시할지 가중치 조절 가능
- 가격이 목표에 가까워지면 종료 보상의 비중이 자연스럽게 커짐

- **협상 상황에 따른 동적 보상 조절**
  - S_n과 PZ_n 변수를 활용해 협상 환경에 맞는 유연한 보상 체계 제공

6. FQI + CQL 기반 Q-Learning 적용 계획

6.1 Offline RL 선택 배경

- 실시간 로그 수집이 제한됨 → Offline 데이터 기반 강화학습 적합
- 데이터셋에 없는 행동에 대해 과대 추정 방지 → Conservative Q-Learning(CQL) 도입

6.2 학습 데이터 구성

경험 튜플 (s, a, r, s', done) 구조

- s = (CardID, Scenario, PriceZone)
- a = 다음 카드 선택 (C1~C9)
- r = 보상 함수로 계산된 보상
- s' = 다음 상태
- done = 협상 종료 여부 (1 또는 0)

6.3 보상 함수 정의

`W = (S_n + PZ_n) / 2 R(s, a) = W * (A / P) + (1 - W) * End`

- A: 앵커링 목표가
- P: 상대방 제안가
- End: 협상 종료 여부 (0 또는 1)

6.4 손실 함수 (FQI + CQL)

총 손실 함수 = Bellman Error + Conservative Penalty

- Bellman Target:
  `y = r + γ * (1 - done) * max_a' Q_target(s', a')`
- 손실 함수:
  `Loss = (Q(s,a) - y)^2 + α * (E[Q(s,a) from π_max] - E[Q(s,a) from D])`
- γ: 할인율 (예: 0.97)
- α: CQL 보수성 하이퍼파라미터 (예: 1.0)

6.5 학습 하이퍼파라미터 예시

항목	값
Discount factor (γ)	0.97
Conservative factor (α)	1.0
Batch size	1024
Optimizer	Adam (lr=0.0003)
Target update rate (τ)	0.005

7. 시나리오 A~D 실사례 및 초기 Q값 구성

7.1 시나리오 가중치 (S_n)

시나리오	S_n 값	설명
A	1.0	가격 최우선
B	0.75	중간 우세
C	0.5	균형
D	0.25	종료 중시

7.2 가격 구간 가중치 (PZ_n)

Price Zone	PZ_n 값	정의
PZ1	0.1	P ≤ A
PZ2	0.5	A < P ≤ T
PZ3	1.0	P > T

7.3 보상 계산 사례

#	시나리오	카드	A	P	W	End	R = W∗A/P + (1-W)∗End
1	A	C5	100	115	1.0	0	0.87
2	B	C2	100	90	0.625	0	0.69
3	C	C7	100	83	0.3	1	1.06
4	D	C1	100	100	0.625	0	0.625