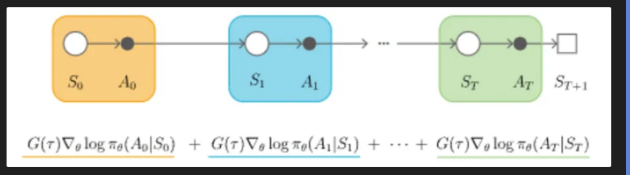
# 2025-1 단러닝클럽 활동보고서 (비교과형)

| 팀 이름 | RLab | 활동 회차 | _____6_____회 |
|---|---|---|---|
| 팀 유형 | ☒퓨리클럽 ( 튜터링: O / X ) | □두런클럽 | □글로벌클럽 ( 튜터링 : O / X ) |
| 활동 정보 | | | |
| 팀장 이름 | 정다훈 | 활동일자 | 2025.05.26 |
| 활동장소 | 퇴계기념도서관 도산라운지 | 활동시간 | 13:00 – 15:00 |
| 참석자 | 김민성, 구선주, 정다훈, 이호영, 정지욱, 최예림 | | |
| 결석자 | | 총 참여인원 | 6 |

| 활동내용 | |
|---|---|
| 주제 | 강화학습에 대해 알아본다. |
| 목표 | 정책 기반 강화학습의 핵심 기법인 정책 경사법의 수학적 원리와 학습 절차를 이해하고, 가치 기반 방법과의 차이점 및 장단점을 파악하여 보다 유연한 정책 최적화 기법을 익힌다. |
| 학습 내용 | 이번 회차에서는 정책 경사법(Policy Gradient Method)을 중심으로 정책 기반 강화학습의 핵심 개념들을 학습하였다. 정책 경사법은 가치 함수를 추정하는 대신 정책 자체를 직접 최적화하는 방식으로, 연속적인 행동 공간 또는 확률적 정책을 다룰 수 있는 유연성이 특징이다.<br>• 정책 함수 π(a\|s,θ)의 개념 및 파라미터 θ의 정의<br>• 정책 경사 정리(Policy Gradient Theorem) 유도 과정 및 수식 해석<br>• REINFORCE 알고리즘 구조 및 학습 과정 이해<br>• variance를 줄이기 위한 baseline 기법 (ex: 상태 가치 함수 V(s)) 소개<br>• 가치 기반 방식(Q-learning, DQN 등)과의 차이점 및 통합 기법<br>실습 시간에는 REINFORCE 알고리즘의 파이썬 구현을 분석하고, 간단한 환경에서의 정책 학습 과정을 시각화하였다. |
| 팀성찰 | 기존 가치 기반 학습과는 다른 방식으로 작동하는 정책 경사법에 대해 배우면서, 팀원들 모두 개념의 구조적 차이를 명확히 이해할 수 있었다. 정책을 직접 업데이트한다는 방식이 익숙하지 않았지만, 수식 유도와 코드 구현을 함께 살펴보며 자연스럽게 이해할 수 있었다.<br>또한 variance 문제와 이를 해결하기 위한 baseline 전략을 서로 설명하고 예시를 들며, 실제 정책 학습의 어려움과 해결책을 함께 고민하는 시간이었다. 전체적으로 강화학습의 두 축(정책 기반 vs 가치 기반)을 모두 비교하며 사고의 폭을 넓힌 회차였다. |
| 활동증빙 | 활동사진 | 활동자료 사진 |
| |  | |

<table>
<tr>
<td></td>
<td>  </td>
<td colspan="2">
<ul>
<li>몬테카를로법: 샘플링을 여러 번 하여 평균을 구하는 방법 → $\nabla_\theta J(\theta)$를 구할 때 사용
<ul>
<li>$\tau^{(i)}$: i번째 에피소드에서 얻은 궤적</li>
<li>$A_t^{(i)}$: i번째 에피소드의 시간 t에서의 행동</li>
<li>$S_t^{(i)}$: i번째 에피스도의 시간 t에서의 상태</li>
</ul>
</li>
<li>에이전트를 정책 $\pi_\theta$에 따라 실제로 행동하게 하여 n개의 궤적 $\tau$를 얻어냄</li>
<li>$\sum_{t=0}^{T} G(\tau^{(i)}) \nabla\theta \log \pi_\theta(A_t^{(i)}|S_t^{(i)})$를 구하고 평균을 내면 $\nabla_\theta J(\theta)$를 근사</li>
</ul>
샘플링: $\tau \sim \pi_\theta$
$$\nabla_\theta J(\theta) \cong \sum_{t=0}^{T} G(\tau) \nabla_\theta \log \pi_\theta(A_t|S_t)$$
<ul>
<li>n=1인 경우 단순화 가능 → 원리 이해를 쉽게 하기 위해 이 값을 대상으로 설명을 진행</li>
</ul>


$$G(\tau)\nabla_\theta \log \pi_\theta(A_0|S_0) + G(\tau)\nabla_\theta \log \pi_\theta(A_1|S_1) + \cdots + G(\tau)\nabla_\theta \log \pi_\theta(A_T|S_T)$$
$$\nabla_\theta \log \pi_\theta(A_t|S_t) = \frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_\theta(A_t|S_t)}$$
<ul>
<li>$\nabla_\theta \log \pi_\theta(A_t|S_t)$: $\nabla_\theta \pi_\theta(A_t|S_t)$라는 기울기 벡터에 $\frac{1}{\pi_\theta(A_t|S_t)}$를 곱한 값
<ul>
<li>모두 상태 $S_t$에서 행동 $A_t$를 취할 확률이 가장 높아지는 방향을 가리킴</li>
<li>그 방향에 대해 $G(\tau)\nabla_\theta \log \pi_\theta(A_t|S_t)$와 같이 $G(\tau)$라는 가중치가 곱해진다</li>
</ul>
</li>
</ul>
</td>
</tr>
<tr>
<td rowspan="5">개별성찰</td>
<td>구선주</td>
<td colspan="2">정책 경사법은 직관적으로는 이해하기 쉬우면서도 수학적으로는 복잡한 개념이라는 점에서 흥미로웠다. 수식 유도 과정을 따라가며 정책이 어떻게 직접 최적화되는지를 체계적으로 학습할 수 있었고, 특히 baseline을 활용해 variance를 줄이는 아이디어가 매우 인상 깊었다. 정책 기반 접근 방식이 실제로 어떤 문제 상황에서 더 적합한지를 고민해보며, 앞으로의 실습이나 프로젝트에도 적용해보고 싶다는 생각이 들었다.</td>
</tr>
<tr>
<td>김민성</td>
<td colspan="2">REINFORCE 알고리즘이 간단하면서도 정책 경사법의 핵심 개념을 잘 담고 있다는 점이 인상 깊었다. 기존 Q-learning 방식에서는 어려웠던 연속적인 행동 공간에서의 학습이 가능하다는 점에서, 이 방식이 현실 문제 해결에 더 적합한 경우도 많겠다는 생각이 들었다. 전체적으로 정책 기반 강화학습의 가능성과 확장성을 체감할 수 있는 유익한 시간이 되었다.</td>
</tr>
<tr>
<td>이호영</td>
<td colspan="2">처음에는 정책 경사 정리의 수식 유도가 다소 어렵게 느껴졌지만, 스터디를 통해 각 항의 의미와 전체 구조를 반복해서 설명 듣고 직접 정리하면서 점점 이해가 되었다. 코드 구현까지 이어지며 이론이 실제로 어떤 방식으로 적용되는지를 확인할 수 있었고, 앞으로 다양한 정책 기반 알고리즘(A2C, PPO 등)으로 확장해보고 싶다는 의욕이 생겼다.</td>
</tr>
<tr>
<td>정다훈</td>
<td colspan="2">이번 회차에서 가장 인상 깊었던 점은 variance 문제를 해결하기 위해 baseline을 사용하는 방식이었다. 단순히 정책을 업데이트하는 것뿐 아니라, 그 과정에서 학습 안정성을 높이는 다양한 아이디어들이 함께 작동한다는 점이 강화학습의 깊이를 느끼게 했다. 정책 경사법이 단순히 새로운 방법이 아니라, 실제 응용에서도 중요한 선택지가 될 수 있음을 알게 되었다.</td>
</tr>
<tr>
<td>정지욱</td>
<td colspan="2">정책 경사법은 직관적으로는 이해하기 쉬우면서도 수학적으로는 복잡한 개념이라는 점에서 흥미로웠다. 수식 유도 과정을 따라가며 정책이 어떻게 직접 최적화되는지를 체계적으로 학습할 수 있었고, 특히 baseline을 활용해 variance를 줄이는 아이디어가 매우 인상 깊었다. 정책 기반 접근 방식이 실제로 어떤 문제 상황에서 더 적합한지를 고민해보며, 앞으로의 실습이나 프로젝트에도 적용해보고 싶다는 생각이 들었다. 정책 경사법은 직관적으로는 이해하기 쉬우면서도 수학적으로는 복잡한 개념이라는 점에서 흥미로웠다. 수식 유도 과정을 따라가며 정책이 어떻게 직접 최적화되는지를 체계적으로 학습할 수 있었고, 특히 baseline을 활용해 variance를 줄이는 아이디어가 매우 인상 깊었다. 정책 기반 접근 방식이 실제로 어떤 문제 상황에서 더 적합한지를 고민해보며, 앞으로의 실습이나 프로젝트에도 적용해보고 싶다는 생각이 들었다.</td>
</tr>
</table>

| | | |
|---|---|---|
| | 최예림 | 기존의 Q-learning이나 DQN처럼 Q-value를 기반으로 하는 방식과 달리, 정책 자체를 직접 최적화한다는 개념이 흥미로웠다. 특히 확률적 정책과 연속 행동 공간에서의 유연한 활용이 가능하다는 점에서 정책 경사법의 실용성을 실감할 수 있었다. 실습을 통해 수식에서 직관으로 이어지는 연결고리를 찾을 수 있어 좋았다. |

**※ 작성 후 반드시 PDF파일로 저장하여 영웅스토리에 업로드 하세요.**