Thesis page 2

📘 2. Policy Optimization (정책 최적화)

2.1 Policy Gradient Methods (정책 경사 방법)

- **정책 경사법**은 정책의 gradient(기울기) 추정치를 계산하고, 이를 통해 확률적 경사 상승(gradient ascent)을 수행하여 정책을 최적화함.
- 일반적으로 사용하는 gradient 추정 식은 다음과 같음:

$$\hat{g} = \mathbb{E}_t \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \hat{A}_t \right]$$

$$\pi_\theta$$

^ 확률적 정책 (state sts_t에서 action ata_t를 선택할 확률)

$$\hat{A}_t$$

^ Advantage function의 추정값

$$\mathbb{E}_t$$

^ 샘플에 대한 경험적 평균 (sampling → optimization 반복)

- Objective function:

$$L^{PG}(\theta) = \mathbb{E}_t \left[\log \pi_\theta(a_t \mid s_t) \hat{A}_t \right]$$

^ 이 함수를 미분해서 정책 gradient를 얻음.

- **주의사항**: 같은 trajectory(경로)로 여러 번 이 objective를 최적화하는 것은 잘못된 방식일 수 있으며, 너무 큰 policy 변화로 인해 성능이 떨어질 수 있음.

2.2 Trust Region Methods (신뢰 영역 기반 방법)

- **TRPO (Trust Region Policy Optimization)**: 정책 업데이트 시 정책이 지나치게 변하지 않도록 제한을 두는 방식.
- 최적화 목적:

$$\max_\theta \mathbb{E}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} A_t \right]$$

- 제약조건:

$$\mathbb{E}_t[\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \leq \delta$$

  → KL divergence가 작도록 (정책이 크게 변하지 않도록) 제한

- 이 문제는 선형 근사와 2차 근사를 통해 conjugate gradient 방법으로 효율적으로 해결할 수 있음.
- **실제로는 penalty 방식**도 사용함:

$$\max_\theta \mathbb{E}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \hat{A}_t - \beta \mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)] \right]$$

  - β : KL 패널티 계수
  - KL divergence에 penalty를 주는 방식은 더 유연하나, 좋은 β\beta 값을 찾는 것이 어렵기 때문에 TRPO는 일반적으로 hard constraint(제약) 방식을 채택.

🔑 요약

- **Policy Gradient**: 정책 기울기를 구해서 최적화하는 기본 방식. 간단하지만 큰 업데이트에 취약함.
- **TRPO**: 정책이 너무 급격히 변하지 않도록 제한(KL 제한)을 두어 안정성을 높임.
- 패널티 방식도 존재하지만 β\beta 조정이 어렵기 때문에 일반적으로 TRPO는 hard constraint를 선호함.
