강화학습 PPO

Wan Ju Kang

KAIST LANADA material based on Huggingface Deep RL course

Policy Optimization

- Trust Region PO
- Maximum a Posteriori PO
- Proximal PO

- 모두 actor-critic 구조이지만 일부 자료에서는 policy gradient 계열로 소개하기도 함
  - 알고리즘의 독창성이 모두 actor 쪽의 구현이라서
- 기본적으로 학습을 진행하면서 정책이 나빠지지는 말게 하자는 공통된 목표를 위해 구현됨
  - Monotonic improvement guarantee
- 수렴 관련 성질의 우수함을 인정 받아 다양한 파생 연구가 진행됨

Proximal PO

- Proximal; approximate; proximity
- = 주변
- PPO = PG + Idea{1, 2, 3}

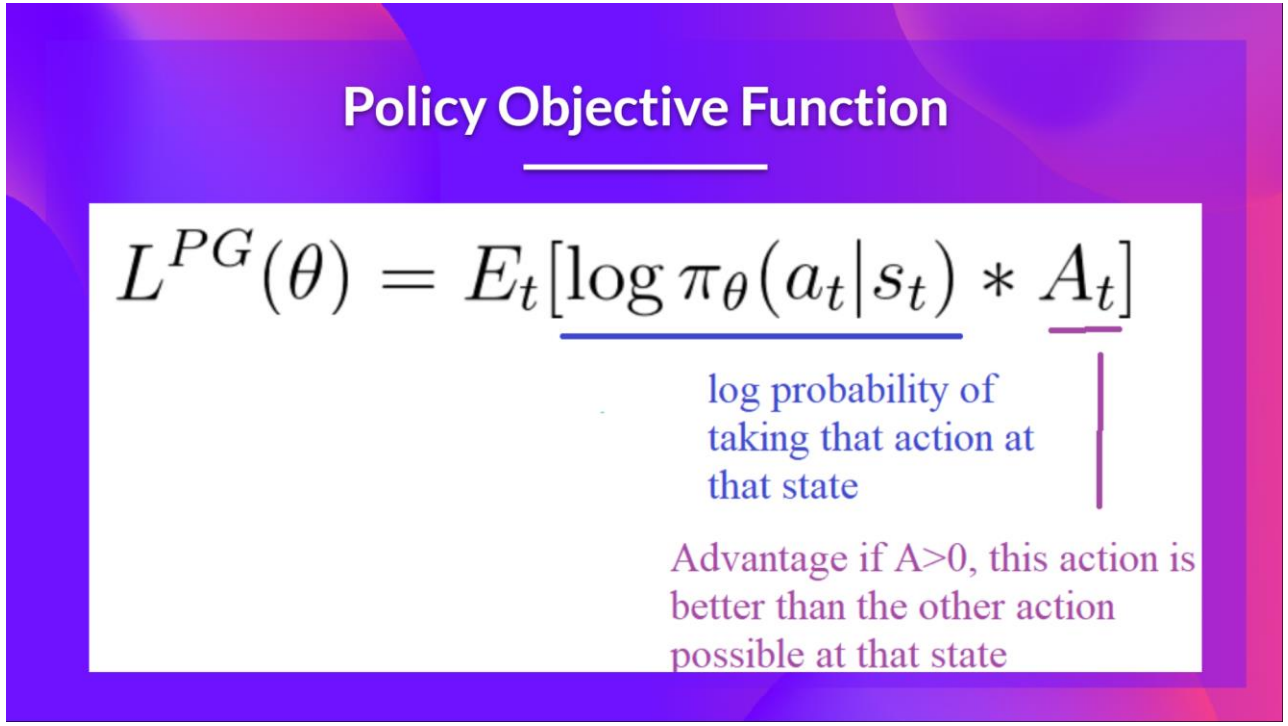
- Idea 1: Actor – 현재 정책의 주변을 탐사해서 좀 더 좋다면 옮겨가자

- Idea 2: Actor – 주변을 최대한 면밀히 탐사하자

- Idea 3: Critic – 그냥 하던 거 하자

PG Review

- RL slide 69
- Q값 자체 대신 Advantage라는 개념의 차용 → Advantage Actor-Critic (A2C)
  - A = Q – V



$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) * Q_w(s,a)]$$

Idea 1: 주변만 탐사하기

- Clipped objective function
- 너무 급진적으로 다른 정책 말고 주변 가까운 정책부터 시도해보자
- TRPO는 clipping 대신 두 정책의 KL divergence를 계산
  - → 비쌈

PPO's Clipped surrogate objective function

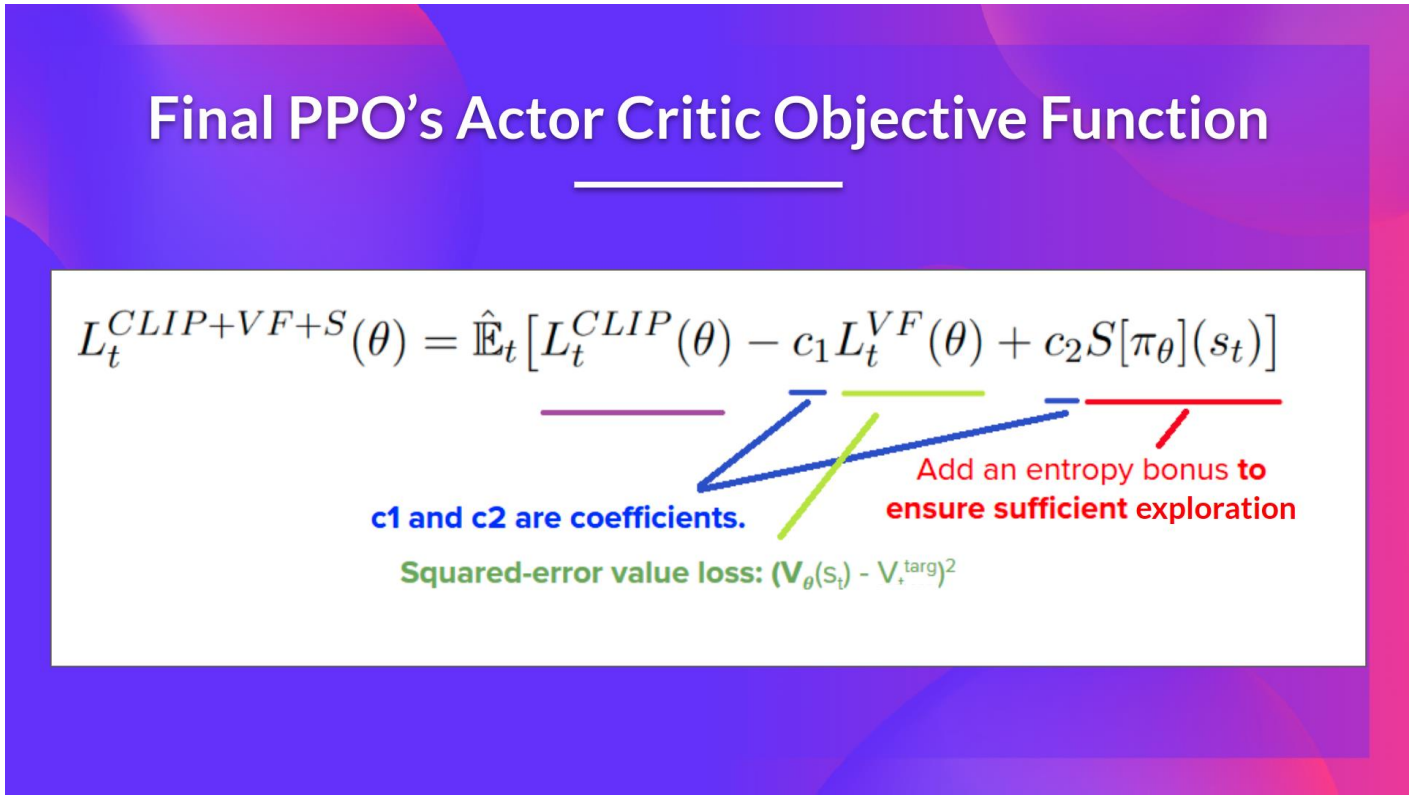
$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

The ratio function

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

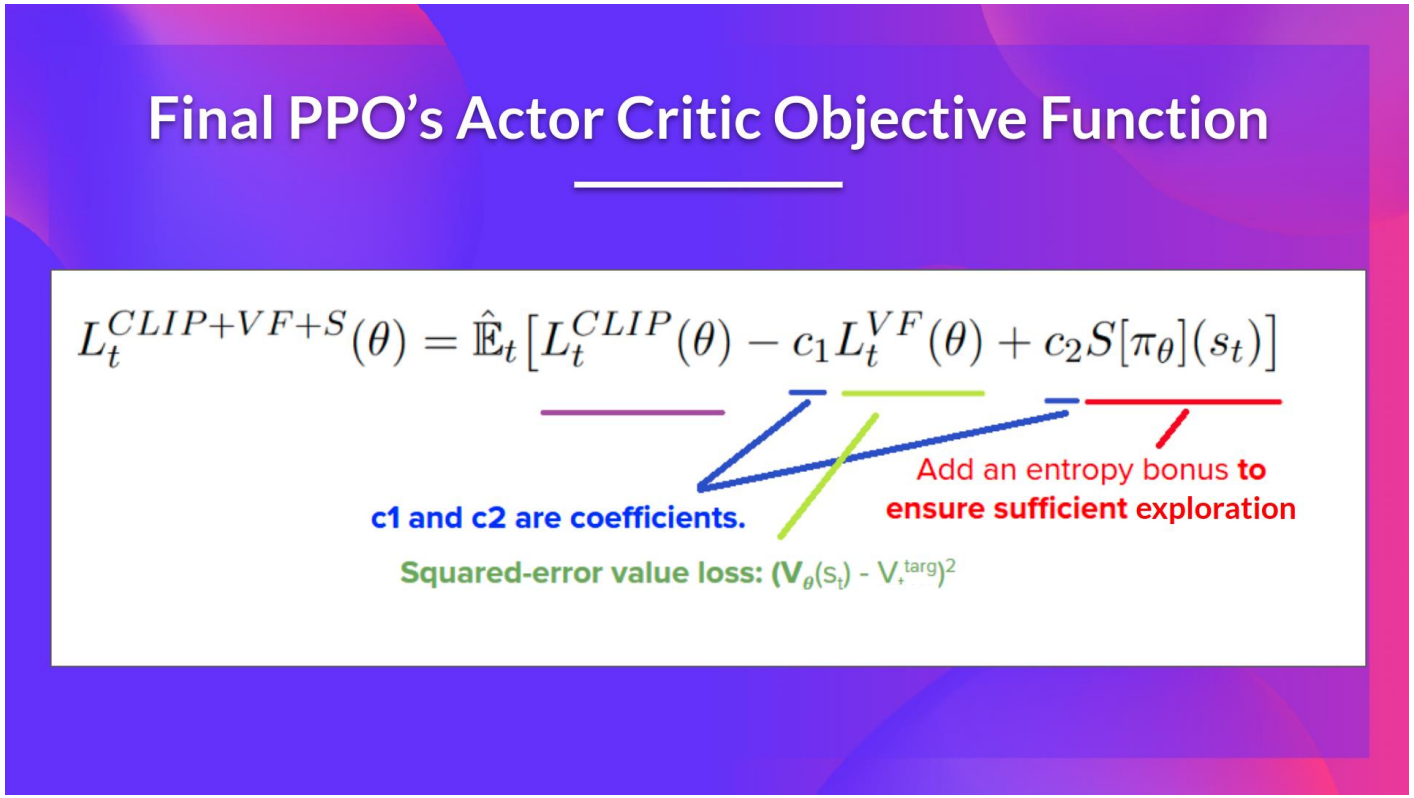
Idea 2: 주변은 다 탐사하기

- 일반적으로 엔트로피 항 S를 추가하면 보다 다양한 값을 바꿔 학습해보도록 진행됨



Idea 3: 비평가는 원래 하던 일

- 최종 손실 함수를 최대화하도록 PPO Agent 트레이닝
- L^{CLIP}: 새 정책의 advantage가 더 크되, epsilon 안에 있도록
- S: 엔트로피 크게 해서 epsilon 안은 최대한 모두 보도록
- L^{VF}: TD error



Thank you!

Any Questions?