폴리시 그래디언트-> loss함수 최저화 대신 목표함수를 최대화하는 정책 기반 강화학습. J(theta)

J(theta)=V\_pi(S0) 이고 그래디언트 정리하면 결국

가 되고 여기서 정책기반은 출력층이 각 행동을 할 확률을 돌려주지 Q함수를 출력하는게 아니기에 이를 직접 식에 사용할 수 없게되는데 q대신 G로 대체해서 discounted reward를 계산해 더해주는게 REINFORCE고 액터크리틱은 큐함수를 근사하는 또 다른 가치신경망을 만들어서 critic이라 칭하는 것. (Actor-critic)

근데 여기서 오류함수는 정책신경망(actor)출력의 (log)와 큐함수(critic)출력의 곱인데 큐함수를 직접 쓰면 분산이 크기 때문에 베이스라인을 잡아주기 위해 V(S)를 뺴서 어드벤티지 함수를 정의. 근데 큐함수와 베이스라인인 가치함수 V를 따로 사용하면 비효율적이니 큐함수를 가치함수로 표현해서 어드밴티지 함수를 정의하는데 형태가 시간차에러와 같으므로 델타로 표현 (A2C)

A(St, At)= Qw(St,At)-Vv(St) =>

결국 액터크리틱의 업데이트 식은

그리고 가치신경망은 의 제곱(MSE)를 이용해 업데이트.

Actor : 출력을 softmax로 해서 각 행동을 할 확률을 출력해야함.

Critic : 가치함수 V를 출력해야하니 그냥 linear출력.