REINFORCE 알고리즘

- 알고리즘 이름이 진짜 REINFORCE 알고리즘이다.

Policy Gradient 관계

- Policy Gradient 를 바탕으로 한다.
- 기댓값에서 큰수의 법칙으로 'N'이 충분하면 아래처럼표시 가능하다.
  - $\int_x x \cdot p(x)\,dx \approx \frac{1}{N}\sum_{i=1}^{N} x_i$
- 그럴따면 Policy Gradient도 적용시킬수있다.

$$\theta \leftarrow \theta + \alpha\left[\int_{\tau}\sum_{t=0}^{\infty}\left(\nabla_\theta \ln P_\theta(a_t|s_t)\times G_t\right)\cdot P_\theta(\tau)\cdot d\tau\right]$$

(x) sample ↓ $\int_\tau x\,p(\tau)\,d\tau$ ← $p(x)$

  - 만약충분한(N)이 있다면 큰수의법칙적용
    - → $\frac{1}{N}\sum_{i=1}^{\infty}\sum_{t=0}^{\infty}\nabla_\theta \ln P_\theta(a_t|s_t)\,G_t$
- 위의 처럼하면 가능은 하지만 1개의 에피소드 만해도 수만번③④ 때문에 계산이힘들다.

그래서(≈)는 힘들지만 (N=1)을 사용해서 미분진행한다.

- $\theta \leftarrow \theta + \alpha\sum_{t=0}^{\infty}\nabla_\theta \ln P_\theta(a_t|s_t)G_t$

- 알고리즘 순서

① 샘플 τ 생성 (즉 에피소드 생성) from $P_\theta(a_t|s_t)$

② $\theta \leftarrow \theta + \alpha\sum_{t=0}^{\infty}\nabla_\theta \ln P_\theta(a_t|s_t)G_t$

문제점

- (N=1)이다보니 결과적으로 biased, High variance 때문에 찾아가기힘들다. (high bias)
- Actor-Critic 으로 해결했다.