Google Developers

Experts

정책 경사도 방법
Policy Gradient Methods

강화학습

송호연 @chris_loves_ai
sjhshy@gmail.com

정책 경사도 방법
Policy Gradient Methods

정책 기반 강화학습

Policy Based Reinforcement Learning

- 지금까지 우리는 **행동 가치 함수를 매개변수로 근사**했습니다.

$$V_\theta(s) \approx V^\pi(s)$$
$$Q_\theta(s, a) \approx Q^\pi(s, a)$$

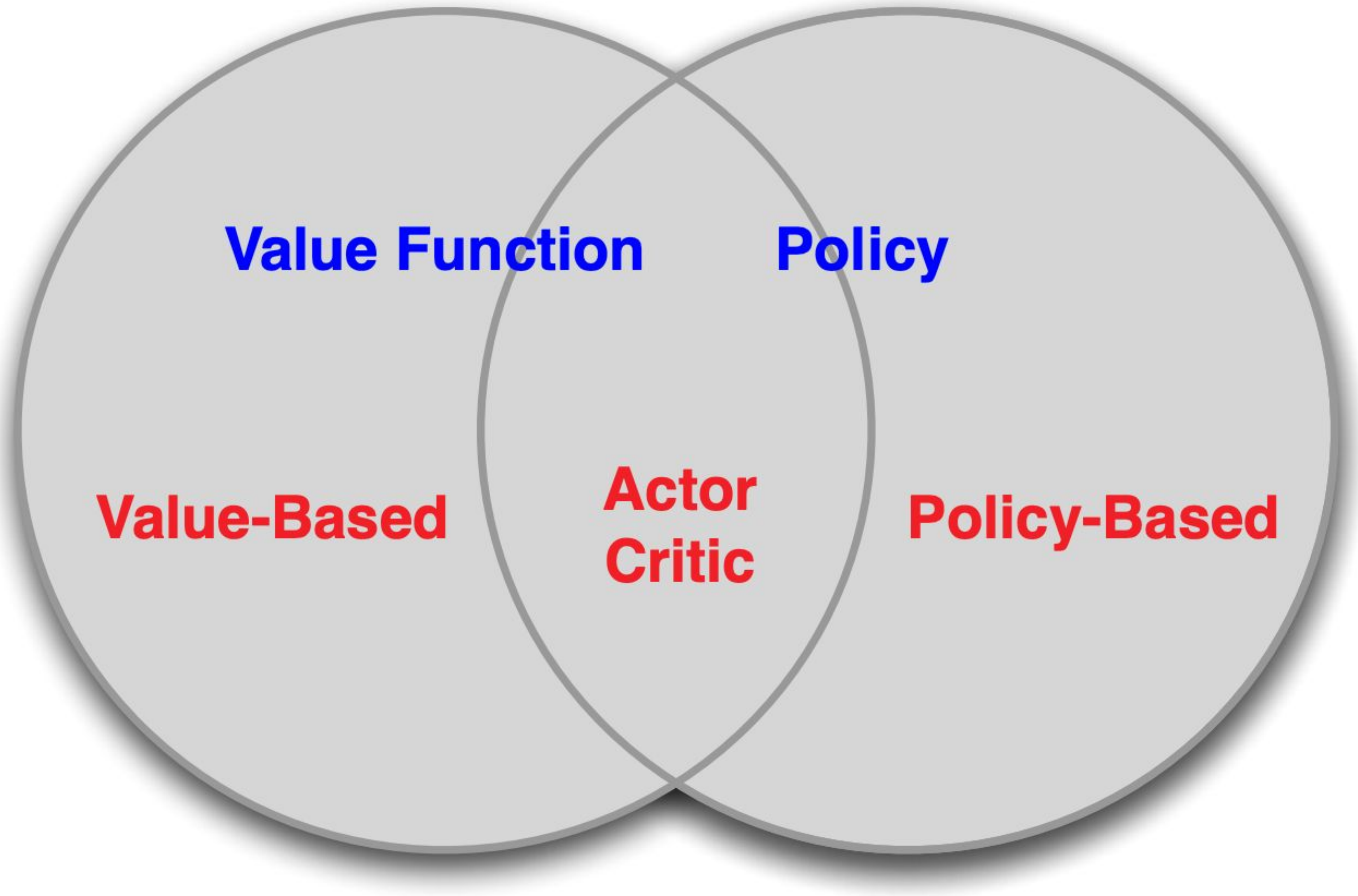
- 이번에 우리는 **정책 함수를 직접 매개변수화**시킬 것입니다.

$$\pi_\theta(s, a) = \mathbb{P}[a \mid s, \theta]$$

가치 기반, 정책 기반 강화학습

Value Based, Policy Based Reinforcement Learning

- **가치 기반 강화학습 (Value Based RL)**
  - 가치 함수 학습
  - 내포된 정책 함수
- **정책 기반 강화학습 (Policy Based RL)**
  - 가치 함수 없음
  - 정책 함수 학습
- **정책 비평가 강화학습 (Actor-Critic RL)**
  - 가치 함수 학습
  - 정책 함수 학습



정책 기반 강화학습의 장점

Advantages of Policy-based RL

- **장점:**
  - 더 좋은 수렴성
  - 고 차원 행동 공간(action space)과 연속 행동 공간에서 효과적임
  - 확률적 정책(Stochastic policy)을 학습할 수 있음
- **단점:**
  - 글로벌 최적점보다 지역 최적점에 수렴할 수 있음
  - 정책을 평가하는 것은 분산이 높고 비효율적인 경우가 많음

예시: 가위 바위 보

Example: Rock-Paper-Scissors

- 두 사람이 가위 바위 보 게임을 합니다.
- 결정론적 정책은 쉽게 착취됩니다.
- 유니폼 정책이 가장 최적입니다. (내쉬 이퀄리브리엄)

정책 경사도 이론

Policy Gradient Theorem

- 정책 경사도 이론은 우도 비율이 멀티 스텝 MDP에 접근하는 것을 일반화합니다.
- 즉각적인 보상 r 대신 장기 가치 Q(s,a)를 사용합니다.

Theorem

For any differentiable policy $\pi_\theta(s,a)$,
for any of the policy objective functions $J = J_1, J_{avR},$ *or* $\frac{1}{1-\gamma} J_{avV}$,
the policy gradient is

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\ Q^{\pi_\theta}(s,a)\right]$$

정책 경사도 목적 함수

Policy Gradient Objective Functions

- 정책 경사도 목적 함수

$$J(\theta) = \sum_{s \in S} d^{\pi}(s) V^{\pi}(s) = \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi_{\theta}(a|s) Q^{\pi}(s,a)$$

$$
\begin{aligned}
&\nabla_\theta V^\pi(s) \\
=&\nabla_\theta \Big(\sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q^\pi(s,a) \Big) \\
=&\sum_{a\in\mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) + \pi_\theta(a|s) \nabla_\theta Q^\pi(s,a) \Big) && \text{; Derivative product rule.} \\
=&\sum_{a\in\mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) + \pi_\theta(a|s) \nabla_\theta \sum_{s',r} P(s',r|s,a)(r + V^\pi(s')) \Big) && \text{; Extend } Q^\pi \text{ with future state value.} \\
=&\sum_{a\in\mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) + \pi_\theta(a|s) \sum_{s',r} P(s',r|s,a) \nabla_\theta V^\pi(s') \Big) && P(s',r|s,a) \text{ or } r \text{ is not a func of } \theta \\
=&\sum_{a\in\mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) + \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^\pi(s') \Big) && \text{; Because } P(s'|s,a) = \textstyle\sum_r P(s',r|s,a)
\end{aligned}
$$

Now we have:

$$
\nabla_\theta V^\pi(s) = \sum_{a\in\mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) + \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^\pi(s') \Big)
$$

$$
\begin{aligned}
&\nabla_\theta V^\pi(s) \\
&= \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s,a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) [\phi(s') + \sum_{s''} \rho^\pi(s' \to s'', 1) \nabla_\theta V^\pi(s'')] \\
&= \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \to s'', 2) \nabla_\theta V^\pi(s'') \text{ ; Consider } s' \text{ as the middle point for } s \to s'' \\
&= \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \to s'', 2) + \sum_{s'''} \rho^\pi(s \to s''', 3) \nabla_\theta V^\pi(s''') \\
&= \dots \text{; Repeatedly unrolling the part of } \nabla_\theta V^\pi(.) \\
&= \sum_{x \in S} \sum_{k=0}^{\infty} \rho^\pi(s \to x, k) \phi(x)
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta V^\pi(s_0) && \text{; Starting from a random state } s_0 \\
&= \sum_s \sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k)\phi(s) && \text{; Let } \eta(s) = \textstyle\sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k) \\
&= \sum_s \eta(s)\phi(s) && \\
&= \Big(\sum_s \eta(s)\Big) \sum_s \frac{\eta(s)}{\sum_s \eta(s)}\phi(s) && \text{; Normalize } \eta(s), s \in \mathcal{S} \text{ to be a probability distribution.} \\
&\propto \sum_s \frac{\eta(s)}{\sum_s \eta(s)}\phi(s) && \textstyle\sum_s \eta(s) \text{ is a constant} \\
&= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) && d^\pi(s) = \frac{\eta(s)}{\sum_s \eta(s)} \text{ is stationary distribution.}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta J(\theta) &\propto \sum_{s\in S} d^\pi(s) \sum_{a\in A} Q^\pi(s,a) \nabla_\theta \pi_\theta(a|s) \\
&= \sum_{s\in S} d^\pi(s) \sum_{a\in A} \pi_\theta(a|s) Q^\pi(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \\
&= \mathbb{E}_\pi[Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s)] & \text{; Because } (\ln x)' = \tfrac{1}{x}
\end{aligned}
$$

몬테카를로 정책 경사도

Monte-Carlo Policy Gradient (REINFORCE)

정책 경사도 이론을 활용

v_t를 Q(s,a)의 무편향 샘플로 사용

function REINFORCE
    Initialise θ arbitrarily
    for each episode $\{s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
        for $t = 1$ to $T - 1$ **do**
            $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t$
        end for
    end for
    return θ
end function

비평자를 활용해 분산을 줄임

Reducing Variance Using a Critic

- 몬테 카를로 정책은 아직 높은 분산을 갖고 있음
- 우리는 행동 가치 함수를 추정하기 위해 비평가를 활용함

$$Q_w(s,a) \approx Q^{\pi_\theta}(s,a)$$

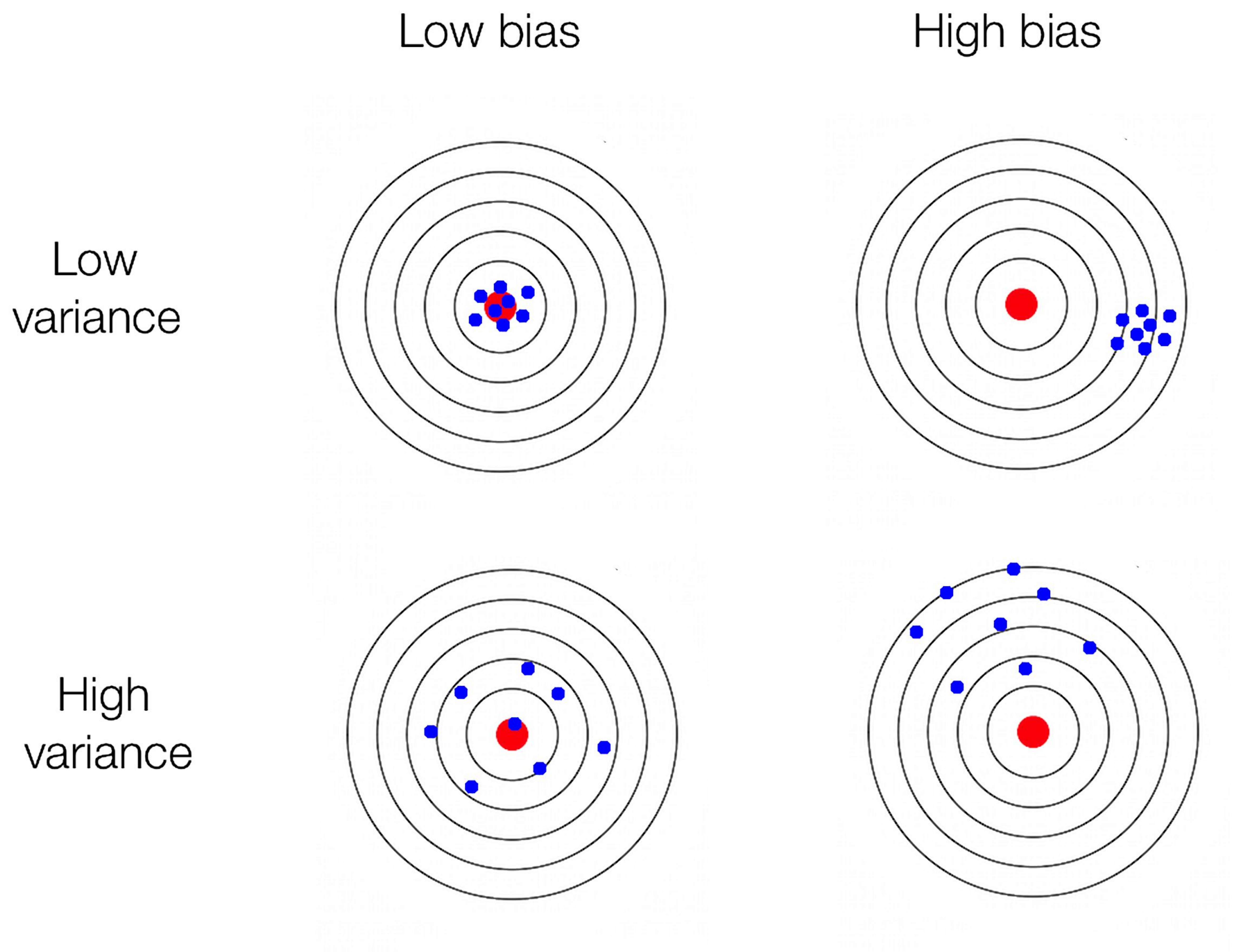
- 행동자-비평가 알고리즘은 두 세트의 파라미터를 보유 중임
  - 비평가: 파라미터로 행동 가치 함수를 업데이트
  - 행동가: 비평가가 제안하는 방향으로 정책 파라미터를 업데이트

- 행동자-비평가 알고리즘은 정책 경사도를 따름

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\ Q_w(s,a)\right]$$
$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s,a)\ Q_w(s,a)$$

비평자를 활용해 분산을 줄임

Reducing Variance Using a Critic



E.O.D.