Policy Gradient.

Measure the quality of a policy π_θ in continuing environments:

$$J_{avR}(\theta) = \sum_{s \in S} d^{\pi_\theta}(s) \cdot \sum_{a \in A} \pi_\theta(s,a) \cdot R_s^a,$$

where d^{π_θ} is the stationary distribution of the Markov chain induced by π_θ.

Definition 1. We call $\nabla_\theta \cdot \log \pi_\theta(s,a)$ the score function.

Proposition 2. The policy gradient of one-step MDP is

$$\nabla_\theta J(\theta) = \mathbb{E}_{\substack{s \sim d^{\pi_\theta}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} [\nabla_\theta \cdot \log \pi_\theta(s,a) \cdot R_s^a].$$

Proof.
$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \cdot \left(\sum_{s \in S} d^{\pi_\theta}(s) \cdot \sum_{a \in A} \pi_\theta(s,a) \cdot R_s^a \right) \\
&= \sum_{s \in S} d^{\pi_\theta}(s) \sum_{a \in A} \nabla_\theta \cdot \pi_\theta(s,a) \cdot R_s^a \\
&= \sum_{s \in S} \cdot d^{\pi_\theta}(s) \cdot \sum_{a \in A} \pi_\theta(s,a) \cdot \nabla_\theta \cdot \log \pi_\theta(s,a) \cdot R_s^a \\
&= \mathbb{E}_{\substack{s \sim d^{\pi_\theta}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} [\nabla_\theta \cdot \log \pi_\theta(s,a) \cdot R_s^a].
\end{aligned}$$

□

Theorem 3. For any differentiable policy $\pi_\theta(s,a)$, the policy gradient is

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

where we slightly abuse the notation by using $\mathbb{E}_{\pi_\theta}[\cdot]$ to denote $\mathbb{E}_{\substack{s \sim d^{\pi_\theta}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}}[\cdot]$.

- Monte-Carlo Policy Evaluation (REINFORCE).
- Actor-Critic (Variance Reduction *)
  - Use a critic to estimate the Q-value:

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

  - Actor-Critic algorithms follow an approximate policy gradient.

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

Theorem 4 (Compatible Function Approximation Theorem).

If ① the value function approximator (critic) is compatible to the policy:

$$\nabla_w Q_w(s,a) = \nabla_\theta \log \pi_\theta(s,a).$$

② the value function parameters w minimize the mean-squared error

$$\varepsilon = \mathbb{E}_{\pi_\theta}[(Q^{\pi_\theta}(s,a) - Q_w(s,a))^2],$$

then w is "compatible" to θ and the approximated policy gradient by actor-critic is exact:

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

Proof. $\nabla_w \varepsilon = \mathbb{E}_{\pi_\theta}[(Q_w(s,a) - Q^{\pi_\theta}(s,a)) \nabla_w Q_w(s,a)]$

$$= \mathbb{E}_{\pi_\theta}[(Q_w(s,a) - Q^{\pi_\theta}(s,a)) \nabla_\theta \log \pi_\theta(s,a)].$$

$$= 0$$

$$\Rightarrow \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) \cdot Q_w(s,a)] = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) \cdot Q^{\pi_\theta}(s,a)].$$

∎

- Advantage Function (Variance Reduction **).
  - Substract a baseline function $B(s)$ which may depend on state s but does not depend on action a to reduce the variance, but without changing the expectation:

$$\nabla_\theta \left(\sum_{s\in S} d^{\pi_\theta}(s) \sum_{a\in A} \pi_\theta(s,a) B(s)\right) = \sum_{s\in S} d^{\pi_\theta}(s) \nabla_\theta \sum_{a\in A} \pi_\theta(s,a) = \sum_{s\in S} d^{\pi_\theta}(s) \nabla_\theta 1 = 0.$$

  - A good choice for baseline function is the V-value function $V^{\pi_\theta}(s)$ which leads to the advantage function:

$$A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s).$$

  - Estimate the advantage function:
    - Use 2 function approximators with 2 parameter vectors.

$$A(s,a) = Q_w(s,a) - V_v(s).$$

    - Use 1 function approximator with 1 parameter vector since the true TD error of V-value is the unbiased estimate of the advantage function:

$$\mathbb{E}_{\pi_\theta}[\delta^{\pi_\theta} \mid s,a] = \mathbb{E}_{\pi_\theta}[R(s,a,s') + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s) \mid s,a]$$
$$= Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s).$$

    In practice, we could use the approximate TD error to estimate the true TD error:

$$\delta_v = R(s,a,s') + \gamma V_v(s) - V_v(s').$$

- Natural Policy Gradient
  - $\nabla_\theta^{nat} J(\theta) = F_\theta^{-1} \nabla_\theta J(\theta)$.
  - Specifically, if the critic parameterized by w is a linear

function approximator with $\nabla_\theta \log \pi_\theta(s,a)$ as the feature vector of state-action pair (s,a), then this critic has two nice properties:

(1) $\phi^T(s,a)\cdot w$ is an approximation of the advantage function $A^{\pi_\theta}(s,a)$

Proof. We only need to prove $\mathbb{E}_{\pi_\theta}[\phi(s,a)^T \cdot w - A^{\pi_\theta}(s,a)] = 0$, or equivalently, to prove $\mathbb{E}_{\pi_\theta}[\phi(s,a)^T \cdot w] = 0$ since $\mathbb{E}_{\pi_\theta}[A^{\pi_\theta}(s,a)] = 0$. This is clear since

$$\begin{aligned}
\mathbb{E}_{\pi_\theta}[\phi(s,a)^T \cdot w] &= \sum_{s\in S} d^{\pi_\theta}(s) \cdot \sum_{a\in A} \cdot \pi_\theta(s,a)\phi(s,a)\cdot w \\
&= \sum_{s\in S} d^{\pi_\theta}(s) \cdot \sum_{a\in A} \cdot \pi_\theta(s,a)\cdot \nabla_\theta \cdot \log \pi_\theta(s,a)^T \cdot w \\
&= \sum_{s\in S} d^{\pi_\theta}(s) \cdot \sum_{a\in A} \cdot \nabla_\theta \cdot \pi_\theta(s,a)^T \cdot w \\
&= \sum_{s\in S} d^{\pi_\theta}(s) w \nabla_\theta \cdot 1 \\
&= 0
\end{aligned}$$

(2) The natural gradient could be simplified as

$$\begin{aligned}
\nabla_\theta^{nat} \cdot J(\theta) &= F_\theta^{-1} \cdot \nabla_\theta \cdot J(\theta) \\
&= F_\theta^{-1} \cdot \mathbb{E}_{\pi_\theta}[\nabla_\theta \cdot \log \pi_\theta(s,a) \cdot \phi(s,a)^T \cdot w] \\
&= F_\theta^{-1} \cdot \mathbb{E}_{\pi_\theta}[\nabla_\theta \cdot \log \pi_\theta(s,a) \cdot \nabla_\theta \cdot \log \pi_\theta(s,a)^T \cdot w] \\
&= w.
\end{aligned}$$