Policy Gradient Proof

Alina Vereshchaka

CSE4/510 Reinforcement Learning
Spring 2020

avereshc@buffalo.edu

April 6, 2020

*Slides are adopted from Policy Gradient Algorithms by Lilian Weng & David Silver’s Course
Goal: given policy $\pi_\theta(s, a)$ with parameters $\theta$, find best $\theta$
Policy Objective Functions

**Goal:** given policy $\pi_{\theta}(s, a)$ with parameters $\theta$, find best $\theta$

But how do we measure the quality of a policy $\pi_{\theta}$?
Policy Objective Functions

Goal: given policy $\pi_\theta(s, a)$ with parameters $\theta$, find best $\theta$

But how do we measure the quality of a policy $\pi_\theta$?

- In episodic environments we can use the start value

$$ J_1(\theta) = V_{\pi_\theta}(s_1) = \mathbb{E}_{\pi_\theta}[v_1] $$
Policy Objective Functions

Goal: given policy \( \pi_\theta(s, a) \) with parameters \( \theta \), find best \( \theta \)

But how do we measure the quality of a policy \( \pi_\theta \)?

- In episodic environments we can use the start value

\[
J_1(\theta) = V_{\pi_\theta}(s_1) = \mathbb{E}_{\pi_\theta}[v_1]
\]

- In continuing environments we can use the average value

\[
J_{avV}(\theta) = \sum_s d_{\pi_\theta}(s)V_{\pi_\theta}(s)
\]
Policy Objective Functions

**Goal:** given policy $\pi_\theta(s, a)$ with parameters $\theta$, find best $\theta$

But how do we measure the quality of a policy $\pi_\theta$?

- In episodic environments we can use the start value

$$J_1(\theta) = V_{\pi_\theta}(s_1) = \mathbb{E}_{\pi_\theta}[v_1]$$

- In continuing environments we can use the average value

$$J_{avV}(\theta) = \sum_s d_{\pi_\theta}(s)V_{\pi_\theta}(s)$$

- Or the average reward per time-step

$$J_{avR}(\theta) = \sum_s d_{\pi_\theta}(s)\sum_a \pi_\theta(s, a)R_s^a$$
Policy Objective Functions

- Where $d^{\pi_\theta}(s)$ is a stationary distribution of Markov chain for $\pi_\theta$ (on-policy state distribution under $\pi$).

- Imagine that you can travel along the Markov chain’s states forever, and eventually, as the time progresses, the probability of you ending up with one state becomes unchanged — this is the stationary probability for $\pi_\theta$

  $$d^{\pi}(s) = \lim_{t \to \infty} P(s_t = s|s_0, \pi_\theta)$$

  is the probability that $s_t = s$ when starting from $s_0$ and following policy $\pi_\theta$ for $t$ steps.

- In the episodic case, $d^{\pi_\theta}(s)$ is defined to be:

  the expected number of time steps $t$ on which $S_t = s$ in a randomly generated episode starting in $s_0$ and following $\pi$ and the dynamics of the MDP
Recap: Derivation Tricks

- Log derivative trick

\[ \nabla_{\theta} \log p(x, \theta) = \frac{\nabla_{\theta} p(x, \theta)}{p(x, \theta)} \]

\[ \nabla_{\theta} p(x, \theta) = p(x, \theta) \nabla_{\theta} \log p(x, \theta) \]
\[ \nabla_\theta V^\pi(s) \quad \text{How to decompose it in terms of } Q^\pi(s, a)? \]
\[ \nabla_\theta V^\pi(s) \]

\[ = \nabla_\theta \left( \sum_{a \in A} \pi_\theta(a|s)Q^\pi(s, a) \right) \]

How to decompose it in terms of \( Q^\pi(s, a) \)?
\[ \nabla_\theta V^\pi(s) \]

\[ = \nabla_\theta \left( \sum_{a \in A} \pi_\theta(a|s) Q^\pi(s, a) \right) \]

\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_\theta Q^\pi(s, a) \right) \]

How to decompose it in terms of \( Q^\pi(s, a) \)?

; Derivative product rule.
\[ \nabla_{\theta} V^\pi(s) \]
\[ = \nabla_{\theta} \left( \sum_{a \in A} \pi_\theta(a|s) Q^\pi(s, a) \right) \]
\[ = \sum_{a \in A} \left( \nabla_{\theta} \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_{\theta} Q^\pi(s, a) \right) \]
\[ ; \text{ Derivative product rule.} \]
\[ ; \text{Extend } Q^\pi \text{ with future state value.} \]
\[ = \sum_{a \in A} \left( \nabla_{\theta} \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_{\theta} \sum_{s', r} P(s', r|s, a) (r + V^\pi(s')) \right) \]
\[ \nabla_\theta V^\pi(s) \]
\[ = \nabla_\theta \left( \sum_{a \in A} \pi_\theta(a|s)Q^\pi(s, a) \right) \]
\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s)Q^\pi(s, a) + \pi_\theta(a|s)\nabla_\theta Q^\pi(s, a) \right) \]

; Derivative product rule.

; Extend \( Q^\pi \) with future state value.

\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s)Q^\pi(s, a) + \pi_\theta(a|s)\sum_{s', r} P(s', r|s, a)(r + V^\pi(s')) \right) \]

; Because \( P(s'|s, a) = \sum_r P(s', r|s, a) \)
\[ \nabla_\theta V^\pi(s) \]

How to decompose it in terms of \( Q^\pi(s, a) \)?

\[ = \nabla_\theta \left( \sum_{a \in A} \pi_\theta(a|s) Q^\pi(s, a) \right) \]

\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \nabla_\theta Q^\pi(s, a) \right) \]

; Derivative product rule.

; Extend \( Q^\pi \) with future state value.

\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \sum_{s', r} P(s', r|s, a)(r + V^\pi(s')) \right) \]

\[ = \sum_{a \in A} \left( \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) + \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi(s') \right) \]

; Because \( P(s'|s, a) = \sum_r P(s', r|s, a) \).
Policy Gradient Theorem

Now we have

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

This equation has a nice recursive form and the future state value function $V^\pi(s')$ can be repeated unrolled by following the same equation.
Let

$$
\phi(s) = \sum_{a \in A} \nabla_{\theta} \pi_{\theta}(a|s) Q^\pi(s, a)
$$

If we keep on extending $\nabla_{\theta} V^\pi(.)$ infinitely, it is easy to find out that we can transition from the starting state $s$ to any state after any number of steps in this unrolling process and by summing up all the visitation probabilities, we get $\nabla_{\theta} V^\pi(.)$.
\[ \nabla_\theta V^\pi(s) = \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi(s') \]
\[ \nabla_\theta V^\pi(s) \]
\[ = \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s, a) \nabla_\theta V^\pi(s') \]
\( \nabla_\theta V^\pi(s) \)

\[
\begin{align*}
&= \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s, a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \nabla_\theta V^\pi(s')
\end{align*}
\]
\[ \nabla_{\theta} V^{\pi}(s) \]
\[ = \phi(s) + \sum_{a} \pi_{\theta}(a | s) \sum_{s'} P(s' | s, a) \nabla_{\theta} V^{\pi}(s') \]
\[ = \phi(s) + \sum_{s'} \sum_{a} \pi_{\theta}(a | s) P(s' | s, a) \nabla_{\theta} V^{\pi}(s') \]
\[ = \phi(s) + \sum_{s'} \rho^{\pi}(s \rightarrow s', 1) \nabla_{\theta} V^{\pi}(s') \]
\[ = \phi(s) + \sum_{s'} \rho^{\pi}(s \rightarrow s', 1)[\phi(s') + \sum_{s''} \rho^{\pi}(s' \rightarrow s'', 1) \nabla_{\theta} V^{\pi}(s'')] \]
\[ \nabla_{\theta} V^\pi(s) \]
\[ = \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_{\theta} V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s, a) \nabla_{\theta} V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \nabla_{\theta} V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) [\phi(s') + \sum_{s''} \rho^\pi(s' \rightarrow s'', 1) \nabla_{\theta} V^\pi(s'')] \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \rightarrow s'', 2) \nabla_{\theta} V^\pi(s'') ; \text{ Consider } s' \text{ as the middle point for } s \rightarrow s'' \]
\[ \nabla_\theta V^\pi (s) \]
\[ = \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi (s') \]
\[ = \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s, a) \nabla_\theta V^\pi (s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \nabla_\theta V^\pi (s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \rightarrow s'', 2) \nabla_\theta V^\pi (s'') ; \text{ Consider } s' \text{ as the middle point for } s \rightarrow s'' \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \rightarrow s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \rightarrow s'', 2) \phi(s'') + \sum_{s'''} \rho^\pi(s \rightarrow s''', 3) \nabla_\theta V^\pi (s''') \]
\[ \nabla_\theta V^\pi(s) \]
\[ = \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) P(s'|s, a) \nabla_\theta V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \nabla_\theta V^\pi(s') \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1)[\phi(s') + \sum_{s''} \rho^\pi(s' \to s'', 1) \nabla_\theta V^\pi(s'')] \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \to s'', 2) \nabla_\theta V^\pi(s'') ; \text{Consider } s' \text{ as the middle point for } s \to s'' \]
\[ = \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \sum_{s''} \rho^\pi(s \to s'', 2) \phi(s'') + \sum_{s'''} \rho^\pi(s \to s''', 3) \nabla_\theta V^\pi(s''') \]
\[ = \sum_{x \in S} \sum_{k=0}^\infty \rho^\pi(s \to x, k) \phi(x) \]
The nice rewriting above allows us to exclude the derivative of Q-value function, $\nabla_\theta Q_\pi(s, a)$. By plugging it into the objective function $J(\theta)$, we are getting the following
\[ \nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) \]

; Starting from a random state \( s_0 \)
Policy Gradient

\[ \nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) \]

= \sum_s \sum_{k=0}^{\infty} \rho^\pi(s_0 \rightarrow s, k) \phi(s) \]

; Starting from a random state \( s_0 \)
\[ \nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) \]

\[ = \sum_{s} \sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k)\phi(s) \]

; Starting from a random state \( s_0 \)

; Let \( \eta(s) = \sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k) \)
\[ \nabla_{\theta} J(\theta) = \nabla_{\theta} V^\pi(s_0) \]

\[ = \sum_s \sum_{k=0}^{\infty} \rho^\pi(s_0 \rightarrow s, k) \phi(s) \]

\[ = \sum_s \eta(s) \phi(s) \]

; Starting from a random state \( s_0 \)

; Let \( \eta(s) = \sum_{k=0}^{\infty} \rho^\pi(s_0 \rightarrow s, k) \)
\[ \nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) \]

; Starting from a random state \( s_0 \)

\[ = \sum_s \sum_{k=0}^\infty \rho^\pi(s_0 \rightarrow s, k) \phi(s) \]

; Let \( \eta(s) = \sum_{k=0}^\infty \rho^\pi(s_0 \rightarrow s, k) \)

\[ = \sum_s \eta(s) \phi(s) \]

; Normalize \( \eta(s), s \in S \) to be a probability distribution.
Policy Gradient

\[ \nabla_{\theta} J(\theta) = \nabla_{\theta} V^{\pi}(s_0) \]

Starting from a random state \( s_0 \)

\[ = \sum_{s} \sum_{k=0}^{\infty} \rho^{\pi}(s_0 \rightarrow s, k) \phi(s) \]

\[ = \sum_{s} \eta(s) \phi(s) \]

Let \( \eta(s) = \sum_{k=0}^{\infty} \rho^{\pi}(s_0 \rightarrow s, k) \)

\[ = \left( \sum_{s} \eta(s) \right) \sum_{s} \frac{\eta(s)}{\sum_{s} \eta(s)} \phi(s) \]

\[ \propto \sum_{s} \frac{\eta(s)}{\sum_{s} \eta(s)} \phi(s) \]

Normalize \( \eta(s), s \in S \) to be a probability distribution.

\[ \sum_{s} \eta(s) \] is a constant
\[ \nabla_\theta J(\theta) = \nabla_\theta V^\pi(s_0) \]

\[ = \sum_s \sum_{k=0}^{\infty} \rho^\pi(s_0 \rightarrow s, k) \phi(s) \]

\[ = \sum_s \eta(s) \phi(s) \]

\[ = \left( \sum_s \eta(s) \right) \sum_s \frac{\eta(s)}{\sum_s \eta(s)} \phi(s) \]

\[ \propto \sum_s \frac{\eta(s)}{\sum_s \eta(s)} \phi(s) \]

\[ = \sum_s d^\pi(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) \]

; Starting from a random state \( s_0 \)

; Let \( \eta(s) = \sum_{k=0}^{\infty} \rho^\pi(s_0 \rightarrow s, k) \)

; Normalize \( \eta(s), s \in S \) to be a probability distribution.

\( \sum_s \eta(s) \) is a constant

\( d^\pi(s) = \frac{\eta(s)}{\sum_s \eta(s)} \) is stationary distribution.
Policy Gradient

Gradient can be written as

$$\nabla_{\theta} J(\theta) \propto \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} Q^{\pi}(s, a) \nabla_{\theta} \pi_{\theta}(a | s)$$
Gradient can be written as

\[ \nabla_{\theta} J(\theta) \propto \sum_{s \in S} d^\pi(s) \sum_{a \in A} Q^\pi(s, a) \nabla_{\theta} \pi_{\theta}(a|s) \]

\[ = \sum_{s \in S} d^\pi(s) \sum_{a \in A} \pi_{\theta}(a|s) Q^\pi(s, a) \frac{\nabla_{\theta} \pi_{\theta}(a|s)}{\pi_{\theta}(a|s)} \]
Gradient can be written as

\[
\nabla_\theta J(\theta) \propto \sum_{s \in S} d^\pi(s) \sum_{a \in A} Q^\pi(s, a) \nabla_\theta \pi_\theta(a|s)
\]

\[
= \sum_{s \in S} d^\pi(s) \sum_{a \in A} \pi_\theta(a|s) Q^\pi(s, a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}
\]

\[
= \mathbb{E}_\pi[Q^\pi(s, a) \nabla_\theta \ln \pi_\theta(a|s)]
\]

; Because \((\ln x)' = 1/x\)

Where \(\mathbb{E}_\pi\) refers to \(\mathbb{E}_{s \sim d_\pi, a \sim \pi_\theta}\) when both state and action distributions follow the policy \(\pi_\theta\) (on policy).
This vanilla policy gradient update has no bias but high variance. Many following algorithms were proposed to reduce the variance while keeping the bias unchanged.

\[ \nabla_\theta J(\theta) = \mathbb{E}_\pi [Q^\pi(s, a) \nabla_\theta \ln \pi_\theta(a|s)] \]