Policy Gradients

CS60077: Reinforcement Learning

Abir Das

IIT Kharagpur

Nov 09, 10, 2020

Agenda

§ Get started with the policy gradient methods.

§ Get familiar with naive REINFORCE algorithm and its advantages and disadvantages.

§ Getting familair with different variance reduction techniques.

§ Actor-Critic methods.

Resources

- § Deep Reinforcement Learning by Sergey Levine [Link]
- § OpenAI Spinning Up [Link]

Reinforcement Learning Setting

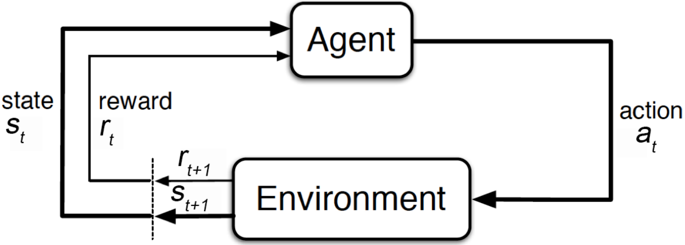


Figure credit: *[SB]*

Reinforcement Learning Setting

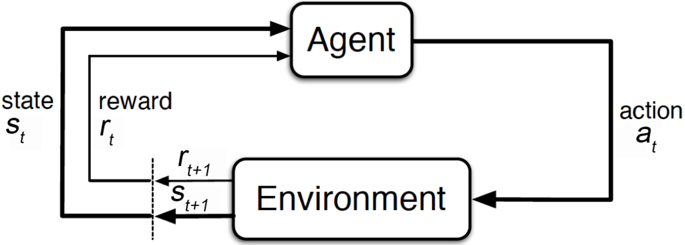


Figure credit: *[SB]*

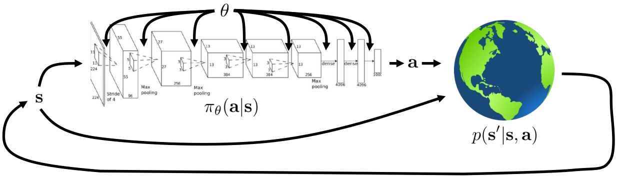


Figure credit: *[Sergey Levine, UC Berkeley]*

Reinforcement Learning Setting

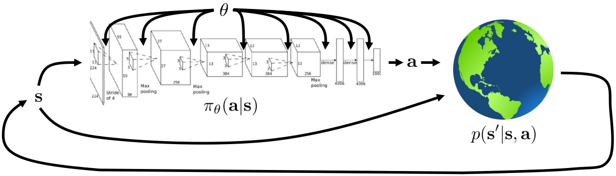


Figure credit: *[Sergey Levine, UC Berkeley]*

- In the middle is the 'policy network' which can directly learn a parameterized policy $\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s})$ (sometimes denoted as $\pi(\mathbf{a}|\mathbf{s};\boldsymbol{\theta})$) and provides the probability distribution over all actions given the state s and parameterized by $\boldsymbol{\theta}$.
- To distinguish it from the parameter vector $\mathbf{w}$ in value function approximator $\hat{v}(\mathbf{s};\mathbf{w})$, the notation $\boldsymbol{\theta}$ is used.

Reinforcement Learning Setting

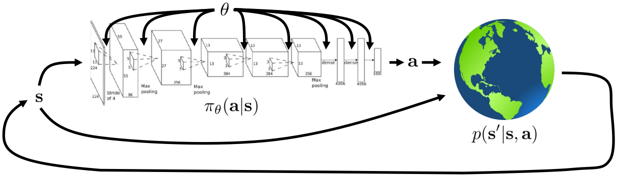


Figure credit: *[Sergey Levine, UC Berkeley]*

- Goal in RL Problem is to maximize the total reward "in expectation" over long run.
- A trajectory τ is defined as,

$$\tau = (\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \mathbf{s}_3, \mathbf{a}_3, \cdots)$$

- The probability of a trajectory is given by the joint probability of the state-action pairs.

$$p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \quad (1)$$

Reinforcement Learning Setting

- Proof of the above relation,

$$
\begin{aligned}
& p(s_{T+1}, s_T, a_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1) \\
&= p(s_{T+1}|s_T, a_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1)p(s_T, a_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1) \\
&= p(s_{T+1}|s_T, a_T)p(s_T, a_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1) \\
&= p(s_{T+1}|s_T, a_T)p(a_T|s_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1)p(s_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1) \\
&= p(s_{T+1}|s_T, a_T)\pi_{\boldsymbol{\theta}}(a_T|s_T)\boxed{p(s_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1)}
\end{aligned} \tag{2}
$$

- The boxed part of the equation is very similar to the left hand side. So, using similar argument repetitively, we get,

$$
\begin{aligned}
& p(s_{T+1}, s_T, a_T, s_{T-1}, a_{T-1}, \cdots, s_1, a_1) \\
&= p(s_{T+1}|s_T, a_T)\pi_{\boldsymbol{\theta}}(a_T|s_T)p(s_T|s_{T-1}, a_{T-1})\pi_{\boldsymbol{\theta}}(a_{T-1}|s_{T-1}) \\
&\qquad p(s_{T-1}, s_{T-2}, a_{T-2} \cdots, s_1, a_1) \\
&= p(s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t, a_t)\pi_{\boldsymbol{\theta}}(a_t|s_t)
\end{aligned} \tag{3}
$$

The Goal of Reinforcement Learning

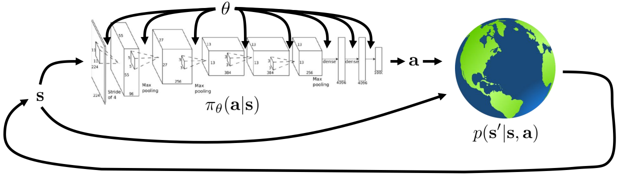


Figure credit: *[Sergey Levine, UC Berkeley]*

- We will sometimes denote the probability as $p_{\boldsymbol{\theta}}(\tau)$, *i.e.*,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

- The goal can be written as,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t)\right]}_{J(\boldsymbol{\theta})}$$

- Note that, for the time being, we are not considering discount. We will come back to that.

The Goal of Reinforcement Learning

§ Goal for a finite horizon setting:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t)} \left[r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

§ The same for the infinite horizon setting

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim p_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a})} \left[r(\mathbf{s}, \mathbf{a}) \right]$$

§ We will consider only finite horizon case in this topic.

Evaluating the Objective

- We will see how we can optimize this objective - the expected value of the total reward under the trajectory distribution induced by the policy $\boldsymbol{\theta}$.
- But before that let us see how we can evaluate the objective in model free setting.

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right] \tag{4}$$

Evaluating the Objective

- We will see how we can optimize this objective - the expected value of the total reward under the trajectory distribution induced by the policy $\boldsymbol{\theta}$.
- But before that let us see how we can evaluate the objective in model free setting.

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right] \approx \frac{1}{N} \sum_i \sum_t r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) \quad (4)$$

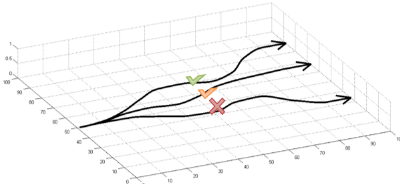


Figure credit: *[Sergey Levine, UC Berkeley]*

Maximizing the Objective

- Now that we have seen how to evaluate the objective, the next step is to maximize it.

Maximizing the Objective

- Now that we have seen how to evaluate the objective, the next step is to maximize it.
- Compute the gradient and take steps in the direction of the gradient.

Maximizing the Objective

- Now that we have seen how to evaluate the objective, the next step is to maximize it.
- Compute the gradient and take steps in the direction of the gradient.

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\overbrace{\sum_t r(\mathbf{s}_t, \mathbf{a}_t)}^{r(\tau)}\right]}_{J(\boldsymbol{\theta})}$$

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[r(\tau)\right] = \int p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$

Maximizing the Objective

- Now that we have seen how to evaluate the objective, the next step is to maximize it.
- Compute the gradient and take steps in the direction of the gradient.

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\overbrace{\sum_t r(\mathbf{s}_t, \mathbf{a}_t)}^{r(\tau)} \right]}_{J(\boldsymbol{\theta})}$$

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[r(\tau) \right] = \int p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \int \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$

- How to compute this complicated looking gradient!

Maximizing the Objective

- Now that we have seen how to evaluate the objective, the next step is to maximize it.
- Compute the gradient and take steps in the direction of the gradient.

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\overbrace{\sum_t r(\mathbf{s}_t, \mathbf{a}_t)}^{r(\tau)}\right]}_{J(\boldsymbol{\theta})}$$

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[r(\tau)\right] = \int p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \int \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$

- How to compute this complicated looking gradient! The log-derivative trick is our rescue.

Log Derivative Trick

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \frac{\partial \log p_{\boldsymbol{\theta}}(\tau)}{\partial p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) = \frac{1}{p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau)$$
$$\implies \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\tau) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) \tag{5}$$

Log Derivative Trick

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \frac{\partial \log p_{\boldsymbol{\theta}}(\tau)}{\partial p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) = \frac{1}{p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau)$$
$$\implies \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\tau) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) \tag{5}$$

§ So using eqn. (5) we get the gradient of the objective as,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \int \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau = \int p_{\boldsymbol{\theta}}(\tau) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau$$
$$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] \tag{6}$$

Log Derivative Trick

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) &= \frac{\partial \log p_{\boldsymbol{\theta}}(\tau)}{\partial p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) = \frac{1}{p_{\boldsymbol{\theta}}(\tau)} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) \\
\implies \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) &= p_{\boldsymbol{\theta}}(\tau) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) \qquad (5)
\end{aligned}
$$

§ So using eqn. (5) we get the gradient of the objective as,

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \int \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau = \int p_{\boldsymbol{\theta}}(\tau) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau \\
&= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau) \right] \qquad (6)
\end{aligned}
$$

§ Remember that

$$
J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[r(\tau) \right] = \int p_{\boldsymbol{\theta}}(\tau) r(\tau) d\tau
$$

Log Derivative Trick

- Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[r(\tau)\right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right]$$

Log Derivative Trick

- Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[r(\tau)\right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right]$$

- We have also seen,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

Log Derivative Trick

- Till now we have the following,

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\tau \sim p_\theta(\tau)} J(\theta); \quad J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[r(\tau) \right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau) r(\tau) \right]$$

- We have also seen,

$$p_\theta(\tau) = p_\theta(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$$

- Taking $\log$ both sides,

$$\log p_\theta(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T} \log p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) + \sum_{t=1}^{T} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$$

Log Derivative Trick

§ Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[r(\tau)\right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right]$$

§ We have also seen,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking $\log$ both sides,

$$\log p_{\boldsymbol{\theta}}(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T} \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) + \sum_{t=1}^{T} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking $\nabla_{\boldsymbol{\theta}}$ both sides,

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) =$$

Log Derivative Trick

§ Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[r(\tau) \right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau) \right]$$

§ We have also seen,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking $\log$ both sides,

$$\log p_{\boldsymbol{\theta}}(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T} \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) + \sum_{t=1}^{T} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking $\nabla_{\boldsymbol{\theta}}$ both sides,

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \cancelto{0}{\log p(\mathbf{s}_1)}$$

Log Derivative Trick

§ Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} [r(\tau)]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} [\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)]$$

§ We have also seen,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking log both sides,

$$\log p_{\boldsymbol{\theta}}(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T} \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) + \sum_{t=1}^{T} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$$

§ Taking $\nabla_{\boldsymbol{\theta}}$ both sides,

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \cancelto{0}{\log p(\mathbf{s}_1)} + \cancelto{0}{\sum_{t=1}^{T} \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}$$

Log Derivative Trick

§ Till now we have the following,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} J(\boldsymbol{\theta}); \quad J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[r(\tau) \right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau) \right]$$

§ We have also seen,

$$p_{\boldsymbol{\theta}}(\tau) = p_{\boldsymbol{\theta}}(\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \cdots, \mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T+1}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t)$$

§ Taking $\log$ both sides,

$$\log p_{\boldsymbol{\theta}}(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T} \log p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) + \sum_{t=1}^{T} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t)$$

§ Taking $\nabla_{\boldsymbol{\theta}}$ both sides,

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \cancelto{0}{\log p(\mathbf{s}_1)} + \cancelto{0}{\sum_{t=1}^{T} \log p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} + \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t)$$

Log Derivative Trick

§ Thus,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] \\ &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right]\end{aligned}$$

Log Derivative Trick

- Thus,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] \\ &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right]\end{aligned}$$

- So, to get the estimate of the gradient we take samples and average not only the sum of rewards but also average the sum of the gradients of the policy values.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})\right]$$

Log Derivative Trick

- Thus,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] \\ &= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right]\end{aligned}$$

- So, to get the estimate of the gradient we take samples and average not only the sum of rewards but also average the sum of the gradients of the policy values.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})\right]$$

- And the last bit is to update $\boldsymbol{\theta}$ along the gradient direction.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{7}$$

Fitting in Generic RL Pipeline

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \boxed{\frac{1}{N}\sum_{i=1}^{N}} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \boxed{\sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})}\right]$$

$$\boxed{\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}$$

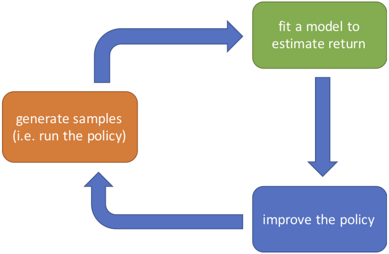


Figure credit: *[Sergey Levine, UC Berkeley]*

Fitting in Generic RL Pipeline

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \boxed{\frac{1}{N}\sum_{i=1}^{N}} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \boxed{\sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})} \right]$$

$$\boxed{\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}$$

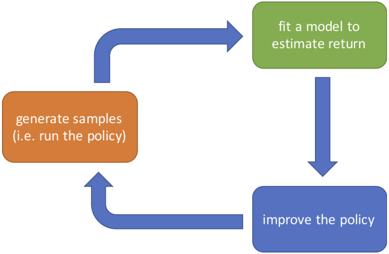


Figure credit: *[Sergey Levine, UC Berkeley]*

REINFORCE Algorithm

1. Sample $\{r^i\}$ from $\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})\sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})\right]$
3. $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
4. Repeat

Taking a Closer Look

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \cancel{\sum_{t=1}^{T} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})} \right]$$

- What is given by $\log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})$? - It is log of the probability of action $\mathbf{a}_{i,t}$ at state $\mathbf{s}_{i,t}$ under the distribution parameterized by $\boldsymbol{\theta}$.
- This gives the likelihood, *i.e.*, how likely, we are to see $\mathbf{a}_{i,t}$ as the action, if our policy is defined by the current $\boldsymbol{\theta}$ that we have.
- Computing the gradient and taking a step along the direction of the gradient, changes $\boldsymbol{\theta}$ in such a way that the likelihood of the action $\mathbf{a}_{i,t}$ increases.

Taking a Closer Look

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})\cancel{\sum_{t=1}^{T} r(\mathbf{s}_{i,t},\mathbf{a}_{i,t})}\right]$$

- What is given by $\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})$? - It is log of the probability of action $\mathbf{a}_{i,t}$ at state $\mathbf{s}_{i,t}$ under the distribution parameterized by $\boldsymbol{\theta}$.
- This gives the likelihood, *i.e.*, how likely, we are to see $\mathbf{a}_{i,t}$ as the action, if our policy is defined by the current $\boldsymbol{\theta}$ that we have.
- Computing the gradient and taking a step along the direction of the gradient, changes $\boldsymbol{\theta}$ in such a way that the likelihood of the action $\mathbf{a}_{i,t}$ increases.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})\sum_{t=1}^{T} r(\mathbf{s}_{i,t},\mathbf{a}_{i,t})\right]$$

- Now consider the case, when it is getting multiplied by $\sum_{t=1}^{T} r(\mathbf{s}_{i,t},\mathbf{a}_{i,t})$.
- Those actions with high rewards are getting more likely.

Taking a Closer Look

- Good stuff is made more likely.
- Bad stuff is made less likely.
- Formalizes the 'trial and error' learning.

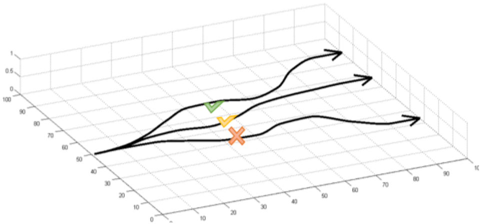


Figure credit: *[Sergey Levine, UC Berkeley]*

Taking a Closer Look

- Good stuff is made more likely.
- Bad stuff is made less likely.
- Formalizes the 'trial and error' learning.

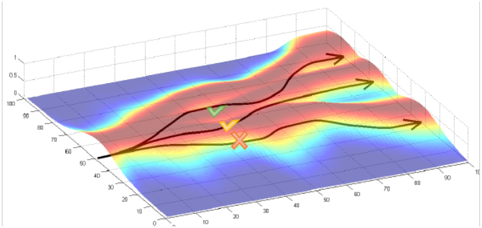


Figure credit: *[Sergey Levine, UC Berkeley]*

Taking a Closer Look

- Good stuff is made more likely.
- Bad stuff is made less likely.
- Formalizes the ‘trial and error’ learning.

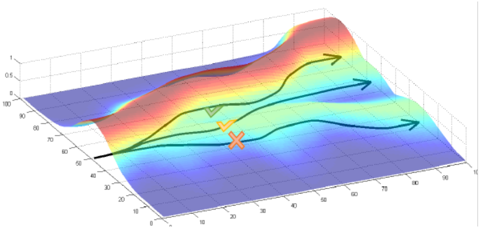


Figure credit: *[Sergey Levine, UC Berkeley]*

Taking a Closer Look

- Good stuff is made more likely.
- Bad stuff is made less likely.
- Formalizes the 'trial and error' learning.

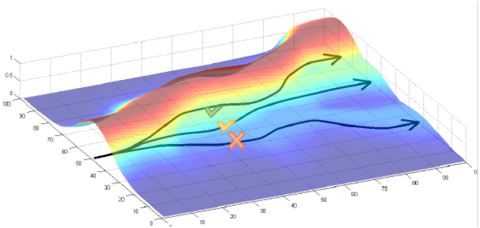


Figure credit: *[Sergey Levine, UC Berkeley]*

Bias and Variance in Estimation

- One way to work with values we do not know is to estimate them by experimenting repeatedly.
- Monte-Carlo methods provide the estimate of the true value and we have used Monte-Carlo methods to estimate the value functions and to estimate the gradient of the expected return.
- The estimator is a function of the data which itself are random variables. So the estimated value is subject to many possible outcomes if employed repeatedly, *i.e.*, if you conduct the experiment multiple times, in general, the estimator will provide different values.
- An estimator is good if,
  - On average the estimated values are close to the true value for different trials - (Bias)
  - The estimates do not vary much in each trial - (variance)

Unbiased Estimators

- An unbiased estimator is the one that yields the true value of the variable being estimated on average. With θ denoting the true value and $\hat{\theta}$ denoting the estimated value, and unbiased estimator is one with,

$$\mathbb{E}[\hat{\theta}] = \theta$$

- Naturally bias is defined as,

$$b = \mathbb{E}[\hat{\theta}] - \theta$$

Unbiased Estimators

- An unbiased estimator is the one that yields the true value of the variable being estimated on average. With θ denoting the true value and $\hat{\theta}$ denoting the estimated value, and unbiased estimator is one with,
$$\mathbb{E}[\hat{\theta}] = \theta$$
- Naturally bias is defined as,
$$b = \mathbb{E}[\hat{\theta}] - \theta$$
- Let us consider estimating a constant value (say temperature of this room) by some sensors which are not perfect. Consider the observations.

$x[n] = \theta + w[n] \quad n = 0, 1, \cdots, N-1. \quad w[n]$ is WGN with variance $= \sigma^2$.

Unbiased Estimators

- An unbiased estimator is the one that yields the true value of the variable being estimated on average. With θ denoting the true value and $\hat{\theta}$ denoting the estimated value, and unbiased estimator is one with,
$$\mathbb{E}[\hat{\theta}] = \theta$$
- Naturally bias is defined as,
$$b = \mathbb{E}[\hat{\theta}] - \theta$$
- Let us consider estimating a constant value (say temperature of this room) by some sensors which are not perfect. Consider the observations.
$x[n] = \theta + w[n] \quad n = 0, 1, \cdots, N-1.$ $w[n]$ is WGN with variance $= \sigma^2$.
- A reasonable estimator is the average value of $x[n]$ *i.e.*,
$\hat{\theta} = \frac{1}{N} \sum\limits_{n=0}^{N-1} x[n]$

Estimator Bias

- The sample mean estimator is unbiased.

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}[x[n]]$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\big([\theta + w[n]]\big) = \frac{1}{N}\sum_{n=0}^{N-1}\big(\mathbb{E}[\theta] + \mathbb{E}[w[n]]\big)$$

$$= \frac{1}{N}\sum_{n=0}^{N-1} = \big(\theta + 0\big) = \theta$$

Estimator Bias

- The sample mean estimator is unbiased.

$$\begin{aligned}\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}[x[n]] \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\big([\theta + w[n]]\big) = \frac{1}{N}\sum_{n=0}^{N-1}\big(\mathbb{E}[\theta] + \mathbb{E}[w[n]]\big) \\ &= \frac{1}{N}\sum_{n=0}^{N-1} = \big(\theta + 0\big) = \theta\end{aligned}$$

- Let us see what happens with a modified estimator, $x[n]$ *i.e.*, $\check{\theta} = \frac{1}{2N}\sum_{n=0}^{N-1} x[n]$

Estimator Bias

- The sample mean estimator is unbiased.

$$\begin{aligned}\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] &= \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}[x[n]]\\ &= \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\big([\theta + w[n]]\big) = \frac{1}{N}\sum_{n=0}^{N-1}\big(\mathbb{E}[\theta] + \mathbb{E}[w[n]]\big)\\ &= \frac{1}{N}\sum_{n=0}^{N-1} = \big(\theta + 0\big) = \theta\end{aligned}$$

- Let us see what happens with a modified estimator, $x[n]$ *i.e.*, $\check{\theta} = \frac{1}{2N}\sum_{n=0}^{N-1} x[n]$
- It is easy to see that $\mathbb{E}[\check{\theta}] = \frac{1}{2}\theta$.
- So the bias is $b = \mathbb{E}[\check{\theta}] - \theta = -\frac{1}{2}\theta$

Estimator Variance

- That an estimator is unbiased does not necessarily mean that it is a good estimator. It is reasonable to check by repeating the experiment how the results differ in successive trials.
- Thus the variance of the estimate is another measure of goodness of the estimator. And the aim will be to see how small we can make var$(\hat{\theta})$.
- Let us take the following 3 estimators for θ and see the variances of all these.

$$\hat{\theta}_a = 0$$
$$\mathbb{E}(\hat{\theta}_a) = 0$$
$$\text{var}(\hat{\theta}_a) = 0$$

$$\hat{\theta}_b = x[0]$$
$$\begin{aligned}\mathbb{E}(\hat{\theta}_b) &= \mathbb{E}(x[0]) \\ &= \mathbb{E}(\theta + w[0]) \\ &= \theta + 0 = \theta\end{aligned}$$
$$\text{var}(\hat{\theta}_b) = \text{var}(x[0]) = \sigma^2$$

$$\hat{\theta}_c = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$
$$\mathbb{E}(\hat{\theta}_c) = \theta \quad \text{(already seen)}$$
$$\text{var}(\hat{\theta}_c) = \mathbb{E}[(\hat{\theta}_c - \mathbb{E}[\hat{\theta}_c])^2]$$

(Continued on next slide.)

Estimator Variance

$$\text{var}(\hat{\theta}_c) = \mathbb{E}[(\hat{\theta}_c - \mathbb{E}[\hat{\theta}_c])^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} x[n] - \mathbb{E}[\hat{\theta}_c])^2] \tag{8}$$

$$= \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} \theta + w[n] - \theta)^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} w[n])^2] = \frac{1}{N^2}\mathbb{E}[(\sum_{n=0}^{N-1} w[n])^2]$$

Estimator Variance

$$\text{var}(\hat{\theta}_c) = \mathbb{E}[(\hat{\theta}_c - \mathbb{E}[\hat{\theta}_c])^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} x[n] - \mathbb{E}[\hat{\theta}_c])^2] \quad (8)$$

$$= \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} \theta + w[n] - \theta)^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} w[n])^2] = \frac{1}{N^2}\mathbb{E}[(\sum_{n=0}^{N-1} w[n])^2]$$

- Now,

$$\text{var}\Big(\sum_{n=0}^{N-1} w[n]\Big) = \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n] - \mathbb{E}\Big[\sum_{n=0}^{N-1} w[n]\Big]\Big)^2\right]$$

$$= \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n] - \sum_{n=0}^{N-1} \overbrace{\mathbb{E}\big[w[n]\big]}^{0}\Big)^2\right] = \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n]\Big)^2\right]$$

Estimator Variance

$$\text{var}(\hat{\theta}_c) = \mathbb{E}[(\hat{\theta}_c - \mathbb{E}[\hat{\theta}_c])^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} x[n] - \mathbb{E}[\hat{\theta}_c])^2] \tag{8}$$

$$= \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} \theta + w[n] - \theta)^2] = \mathbb{E}[(\frac{1}{N}\sum_{n=0}^{N-1} w[n])^2] = \frac{1}{N^2}\mathbb{E}[(\sum_{n=0}^{N-1} w[n])^2]$$

§ Now,

$$\text{var}\Big(\sum_{n=0}^{N-1} w[n]\Big) = \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n] - \mathbb{E}\big[\sum_{n=0}^{N-1} w[n]\big]\Big)^2\right]$$

$$= \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n] - \sum_{n=0}^{N-1} \overbrace{\mathbb{E}\big[w[n]\big]}^{0}\Big)^2\right] = \mathbb{E}\left[\Big(\sum_{n=0}^{N-1} w[n]\Big)^2\right]$$

§ Using the above in eqn. (8)

$$\text{var}(\hat{\theta}_c) = \frac{1}{N^2}\text{var}\Big(\sum_{n=0}^{N-1} w[n]\Big) = \frac{1}{N^2}\Big(\sum_{n=0}^{N-1} \text{var}(w[n])\Big) \quad \text{(WGN)}$$

$$= \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

Estimator Mean Square Error

§ The mean of the square error of estimation is,

$$\text{mse}(\hat{\theta}_c) = \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \;+\; \mathbb{E}[\hat{\theta}]-\theta)^2\big]$$

$$= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big]$$

Estimator Mean Square Error

- The mean of the square error of estimation is,

$$
\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \ + \ \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big]
\end{aligned}
$$

(why?) – (Hint: What is random here?)

Estimator Mean Square Error

- The mean of the square error of estimation is,

$$
\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \;+\; \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big] \\
&\quad (\text{why?}) - (\text{Hint: What is random here?}) \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\cancelto{0}{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])}
\end{aligned}
$$

Estimator Mean Square Error

§ The mean of the square error of estimation is,

$$
\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \;+\; \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big] \\
&\quad \text{(why?)} - \text{(Hint: What is random here?)} \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\cancelto{0}{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])} \\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}
$$

Estimator Mean Square Error

- The mean of the square error of estimation is,

$$
\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \;+\; \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big] \\
&\quad (\text{why?}) - (\text{Hint: What is random here?}) \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\cancelto{0}{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])} \\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}
$$

- So the mean square error in estimation, is composed of errors due to the variance of the esstimator as well as the bias.

Estimator Mean Square Error

- The mean of the square error of estimation is,

$$\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \; + \; \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big] \\
&\quad (\text{why?}) - (\text{Hint: What is random here?}) \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\cancelto{0}{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])} \\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}$$

- So the mean square error in estimation, is composed of errors due to the variance of the esstimator as well as the bias.
- Recall MC evaluation

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T \quad \text{and } v_\pi(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$\hat{v}_\pi(s) = \frac{1}{N}\sum_{i=1}^{N} G_t^{(i)}(S_t = s)$$

Estimator Mean Square Error

- The mean of the square error of estimation is,

$$\begin{aligned}
\text{mse}(\hat{\theta}_c) &= \mathbb{E}\big[(\hat{\theta}-\theta)^2\big] = \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}] \ + \ \mathbb{E}[\hat{\theta}]-\theta)^2\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}]-\theta)^2\big] + 2\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\big] \\
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\big]
\end{aligned}$$

(why?) − (Hint: What is random here?)

$$\begin{aligned}
&= \mathbb{E}\big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\big] + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\mathbb{E}[\hat{\theta}]-\theta)\cancelto{0}{(\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}])} \\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}$$

- So the mean square error in estimation, is composed of errors due to the variance of the esstimator as well as the bias.
- Recall MC evaluation

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T \quad \text{and } v_\pi(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$\hat{v}_\pi(s) = \frac{1}{N}\sum_{i=1}^{N} G_t^{(i)}(S_t = s)$$

- So $\hat{v}_\pi(s)$ is an unbiased estimator but with variance (inversely proportional to number of samples N.)

Bias and Variance of MC and TD

- One key contribution of variance in MC evaluation comes from the randomness at each timestep.
- This is not the case in TD as the G_t is estimated by bootstrapping,

$$\hat{G}_t = R_{t+1} + \gamma \hat{V}(S_{t+1})$$

- This makes the estimator suffer less from variance as randomness comes from only one random step taken. The rest is deterministic.
- But this introduces bias. The estimate always have the deterministic additive component $\gamma \hat{V}(S_{t+1})$

Reducing Variance in Policy Gradient Estimate

- We have seen,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

- Inside each trajectory, a lot of randomness is there.
- We can derive versions of this formula that eliminate terms to reduce variance.
- Let us apply the log derivative trick $(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \sum \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t))$ to compute the gradient for a single reward term.

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\tau} \left[r(\mathbf{s}_t, \mathbf{a}_t) \right] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\left(\sum_{t'=1}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{t'}|\mathbf{s}_{t'}) \right) r(\mathbf{s}_t, \mathbf{a}_t) \right] \tag{9}$$

- Note that the sum goes up to t. Why?

Reducing Variance in Policy Gradient Estimate

- We have seen,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

- Inside each trajectory, a lot of randomness is there.
- We can derive versions of this formula that eliminate terms to reduce variance.
- Let us apply the log derivative trick $\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) = \sum \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\right)$ to compute the gradient for a single reward term.

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\tau} \left[r(\mathbf{s}_t, \mathbf{a}_t) \right] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\left(\sum_{t'=1}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{t'}|\mathbf{s}_{t'}) \right) r(\mathbf{s}_t, \mathbf{a}_t) \right] \quad (9)$$

- Note that the sum goes up to t. Why? - The reward at timestep t depends on actions till $t' \leq t$. - **Causality**

Reducing Variance in Policy Gradient Estimate

§ Summing over time we get (with some reordering of the sums, last)

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\tau}\left[r(\tau)\right] &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_{t=1}^{T} r(\mathbf{s}_t,\mathbf{a}_t)\sum_{t'=1}^{t}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_{t'}|\mathbf{s}_{t'})\right]\\ &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T} r(\mathbf{s}_{t'},\mathbf{a}_{t'})\right]\end{aligned} \tag{10}$$

§ With less randomness inside each trajectory the variance is less, but what about bias?

Reducing Variance in Policy Gradient Estimate

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\Bigg[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\sum_{t=1}^{T}\Big(r(\mathbf{s}_t,\mathbf{a}_t)\Big)\Bigg] \\
&= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\Bigg[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\sum_{t'=1}^{T}\Big(r(\mathbf{s}_{t'},\mathbf{a}_{t'})\Big)\Bigg] \\
&= \mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\Bigg[\sum_{t=1}^{T}\sum_{t'=1}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)r(\mathbf{s}_{t'},\mathbf{a}_{t'})\Big)\Bigg] \\
&= \sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\overbrace{\Big[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)r(\mathbf{s}_{t'},\mathbf{a}_{t'})\Big]}^{f(t,t')} \qquad (11)
\end{aligned}
$$

Reducing Variance in Policy Gradient Estimate

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\Big[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\sum_{t=1}^{T}\Big(r(\mathbf{s}_t,\mathbf{a}_t)\Big)\Big] \\
&= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\Big[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\sum_{t'=1}^{T}\Big(r(\mathbf{s}_{t'},\mathbf{a}_{t'})\Big)\Big] \\
&= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\Big[\sum_{t=1}^{T}\sum_{t'=1}^{T}\Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) r(\mathbf{s}_{t'},\mathbf{a}_{t'})\Big)\Big] \\
&= \sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\overbrace{\big[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) r(\mathbf{s}_{t'},\mathbf{a}_{t'})\big]}^{f(t,t')}
\end{aligned} \tag{11}
$$

- Let us consider the term,

$$
\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\big[f(t,t')\big] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\big[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) r(\mathbf{s}_{t'},\mathbf{a}_{t'})\big] \tag{12}
$$

- We will show that for the case of $t' < t$ (reward coming before the action is performed) the above term is zero.

Reducing Variance in Policy Gradient Estimate

$$
\begin{aligned}
\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[f(t,t')\right] &= \int p(\tau)f(t,t')d(\tau) \\
&= \int p(s_1,a_1,\cdots,s_t,a_t,\cdots,s_{t'},a_{t'},\cdots)f(t,t') \\
&\qquad d(s_1,a_1,\cdots,s_t,a_t,\cdots,s_{t'},a_{t'},\cdots) \\
&= \int p(s_t,a_t,s_{t'},a_{t'})f(t,t')d(s_t,a_t,s_{t'},a_{t'})
\end{aligned}
\tag{13}
$$

§ The above comes from the property below.

$$
\begin{aligned}
\int\limits_X\int\limits_Y f(X)P(X,Y)dY\,dX &= \int\limits_X\int\limits_Y f(X)P(X)P(Y|X)dY\,dX \\
&= \int\limits_X f(X)P(X)dX \underbrace{\int\limits_Y P(Y|X)dY}_{1} \\
&= \int\limits_X f(X)P(X)dX
\end{aligned}
\tag{14}
$$

§ Taking $X=\{s_t,a_t,s_{t'},a_{t'}\}$ and Y the rest.

Reducing Variance in Policy Gradient Estimate

§ Till now we have,

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\left[f(t,t')\right] = \int p(s_t, a_t, s_{t'}, a_{t'}) f(t,t') d(s_t, a_t, s_{t'}, a_{t'}) \tag{15}$$

§ We will now use a variation of iterated expectation.

$$\begin{aligned}
\mathbb{E}_{A,B}[f(A,B)] &= \int P(A,B) f(A,B) dB dA \\
&= \int P(B|A) P(A) f(A,B) dB dA \\
&= \int P(A) \int P(B|A) f(A,B) dB \, dA \\
&= \int P(A) \mathbb{E}_B \left[f(A,B)|A\right] dA \\
&= \mathbb{E}_A \Big[\mathbb{E}_B \left[f(A,B)|A\right]\Big]
\end{aligned}$$

§ Taking $A = s_{t'}, a_{t'}$ and $B = s_t, a_t$, eqn. (15) can be written as,

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\left[f(t,t')\right] = \mathop{\mathbb{E}}_{s_{t'}, a_{t'}} \Big[\mathop{\mathbb{E}}_{s_t, a_t} \left[f(t,t') | \mathbf{s}_{t'}, \mathbf{a}_{t'}\right]\Big] \tag{16}$$

Reducing Variance in Policy Gradient Estimate

§ Putting the value of $f(t,t')$ back in eqn. (16), we get,

$$\begin{aligned}\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[f(t,t')\right] &= \mathop{\mathbb{E}}_{s_{t'},a_{t'}}\left[\mathop{\mathbb{E}}_{s_t,a_t}\left[f(t,t')|s_{t'},a_{t'}\right]\right] \\ &= \mathop{\mathbb{E}}_{s_{t'},a_{t'}}\left[\mathop{\mathbb{E}}_{s_t,a_t}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)r(\mathbf{s}_{t'},\mathbf{a}_{t'})|\mathbf{s}_{t'},\mathbf{a}_{t'}\right]\right] \\ &= \mathop{\mathbb{E}}_{s_{t'},a_{t'}}\left[r(\mathbf{s}_{t'},\mathbf{a}_{t'})\mathop{\mathbb{E}}_{s_t,a_t}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)|\mathbf{s}_{t'},\mathbf{a}_{t'}\right]\right]\end{aligned} \tag{17}$$

§ Let us take a closer look at the inner expectation,

$$\mathop{\mathbb{E}}_{s_t,a_t}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)|\mathbf{s}_{t'},\mathbf{a}_{t'}\right] = \int P(\mathbf{s}_t,\mathbf{a}_t|\mathbf{s}_{t'},\mathbf{a}_{t'})\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)d(\mathbf{a}_t,\mathbf{s}_t) \tag{18}$$

§ Now, let us consider the timestep t be greater than t', *i.e.*, the action occurs after the reward. In such a case, $P(\mathbf{s}_t,\mathbf{a}_t|\mathbf{s}_{t'},\mathbf{a}_{t'})$ can be broken down to $P(\mathbf{a}_t|\mathbf{s}_t)P(\mathbf{s}_t|\mathbf{s}_{t'},\mathbf{a}_{t'})$. Thus eqn. (18) becomes,

$$\begin{aligned}\mathop{\mathbb{E}}_{s_t,a_t}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)|\mathbf{s}_{t'},\mathbf{a}_{t'}\right] &= \int\int P(\mathbf{a}_t|\mathbf{s}_t)P(\mathbf{s}_t|\mathbf{s}_{t'},\mathbf{a}_{t'})\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)d\mathbf{a}_t d\mathbf{s}_t \\ &= \int P(\mathbf{s}_t|\mathbf{s}_{t'},\mathbf{a}_{t'})\int P(\mathbf{a}_t|\mathbf{s}_t)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)d\mathbf{a}_t d\mathbf{s}_t \\ &= \mathop{\mathbb{E}}_{s_t}\left[\mathop{\mathbb{E}}_{a_t}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)|\mathbf{s}_t\right]|\mathbf{s}_{t'},\mathbf{a}_{t'}\right]\end{aligned} \tag{19}$$

Reducing Variance in Policy Gradient Estimate

- Now we will use a neat trick known as '*Expected Grad Log Probability*' (EGLP) lemma which says $\mathbb{E}\big[\nabla_\theta \log p_\theta(x)\big] = 0$.

$$\mathop{\mathbb{E}}_{x\sim p_\theta(x)}\left[\nabla_\theta \log p_\theta(x)\right] = \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx = \int p_\theta(x)\frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}dx$$
$$= \int \nabla_\theta p_\theta(x)dx = \nabla_\theta \int p_\theta(x)dx = \nabla_\theta 1 = 0$$

- Thus the inner expectation in eqn. (19) is 0. This, in turn, means eqn. (17), (16) and (15) are all 0.
- That is, $\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\big[f(t,t')\big] = 0$ for $t > t'$.
- Now for $t \leq t'$, $P(\mathbf{s}_t, \mathbf{a}_t|\mathbf{s}_{t'}, \mathbf{a}_{t'})$ can **not** be broken down to $P(\mathbf{a}_t|\mathbf{s}_t)P(\mathbf{s}_t|\mathbf{s}_{t'}, \mathbf{a}_{t'})$, as past state ($\mathbf{s}_t$) will get conditioned on future state and actions ($\mathbf{s}_{t'}, \mathbf{a}_{t'}$) violating the Markov property.
- So, $\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\big[f(t,t')\big] \neq 0$ for $t \leq t'$.

Reducing Variance in Policy Gradient Estimate

§ So we began with,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{t'=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t,t') \right] \tag{20}$$

and have shown that

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t,t') \right] \begin{cases} = 0 & \text{if } \; t' < t \\ \neq 0 & \text{if } \; t' \geq t \end{cases}$$

Reducing Variance in Policy Gradient Estimate

§ So we began with,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{t'=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] \tag{20}$$

and have shown that

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] \begin{cases} = 0 & \text{if } \; t' < t \\ \neq 0 & \text{if } \; t' \geq t \end{cases}$$

§ So, the gradient of the total expected return can be written as,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{t'=t}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \sum_{t'=t}^{T} f(t, t') \right]$$

$$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t) \right) \sum_{t'=t}^{T} \left(r(\mathbf{s}_t, \mathbf{a}_t) \right) \right] \tag{21}$$

Reducing Variance in Policy Gradient Estimate

§ So we began with,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{t'=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] \tag{20}$$

and have shown that

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] \begin{cases} = 0 & \text{if } t' < t \\ \neq 0 & \text{if } t' \geq t \end{cases}$$

§ So, the gradient of the total expected return can be written as,

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{t'=t}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[f(t, t') \right] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \sum_{t'=t}^{T} f(t, t') \right]$$

$$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t) \right) \sum_{t'=t}^{T} \left(r(\mathbf{s}_t, \mathbf{a}_t) \right) \right] \tag{21}$$

§ This is the ‘reward to go’ formulation we have seen earlier and which has less variance. But this also is same as the total expected reward expression which is unbiased. So this is unbiased and less variance estimator of the total expected reward.

Baselines

- Good stuff is made more likely.
- Bad stuff is made less likely.

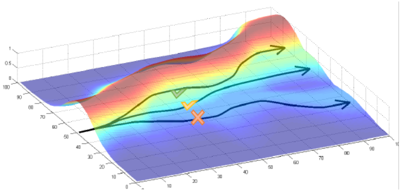


Figure credit: *[Sergey Levine, UC Berkeley]*

Baselines

- Good stuff is made more likely.
- Bad stuff is made less likely.
- What if all have high reward?

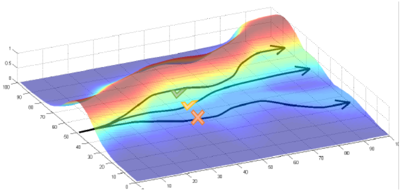


Figure credit: *[Sergey Levine, UC Berkeley]*

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right]$$

Baselines

- Good stuff is made more likely.
- Bad stuff is made less likely.
- What if all have high reward?

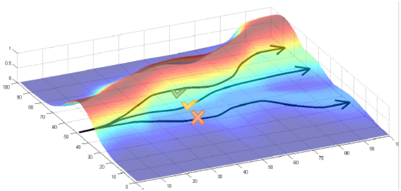


Figure credit: *[Sergey Levine, UC Berkeley]*

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) r(\tau)\right] = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right]$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)[r(\tau) - b]\right]$$

- Will it remain unbiased?
- Only if $\mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau) b\right] = b \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right] = 0$
- And $\mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right] = 0$ by EGLP Lemma.

Baselines

- So subtracting a constant baseline keeps the estimate unbiased.
- A reasonable choice of baseline is average reward across the trajectories, $b = \frac{1}{N}\sum_{i=1}^{N} r(\tau)$
- What about variance?

$$\nabla_{\theta} J(\theta) = \mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \big[\nabla_{\theta} \log p_{\theta}(\tau)[r(\tau) - b]\big]$$

$$\text{var} = \mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \Big[\Big(\nabla_{\theta} \log p_{\theta}(\tau)[r(\tau) - b]\Big)^2\Big] - \Big(\mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \big[\nabla_{\theta} \log p_{\theta}(\tau)[r(\tau) - b]\big]\Big)^2$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \Big[\Big(\nabla_{\theta} \log p_{\theta}(\tau)[r(\tau) - b]\Big)^2\Big] - \Big(\mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \big[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau)\big]\Big)^2$$

$$\frac{\partial \text{var}}{\partial b} = \frac{\partial \mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \Big[\Big(\nabla_{\theta} \log p_{\theta}(\tau)[r(\tau) - b]\Big)^2\Big]}{\partial b} - 0$$

$$= \frac{\partial \mathop{\mathbb{E}}_{\tau \sim p_{\theta}(\tau)} \Big[\big(\nabla_{\theta} \log p_{\theta}(\tau)\big)^2 \big[r^2(\tau) - 2r(\tau)b + b^2\big]\Big]}{\partial b}$$

Baselines

$$\frac{\partial \text{var}}{\partial b} = \frac{\partial \underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2\left[r^2(\tau) - 2r(\tau)b + b^2\right]\right]}{\partial b}$$

$$= 0 - 2 \underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2 r(\tau)\right] + 2b \underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2\right]$$

§ For minimum variance,

$$\frac{\partial \text{var}}{\partial b} = 0$$

$$-\underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2 r(\tau)\right] + b \underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2\right] = 0$$

$$b = \frac{\underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2 r(\tau)\right]}{\underset{\tau\sim p_{\boldsymbol{\theta}}(\tau)}{\mathbb{E}}\left[\left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau)\right)^2\right]}$$

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \underbrace{\sum_{t'=t}^{T} \Big(r(\mathbf{s}_t, \mathbf{a}_t) \Big)}_{\widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)} \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]$$

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \right) \underbrace{\sum_{t'=t}^{T} \left(r(\mathbf{s}_t, \mathbf{a}_t) \right)}_{\widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t)} \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \right) \widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]$$

§ It would be good to have the true value of Q to be used in the equation.

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t) \right) \underbrace{\sum_{t'=t}^{T} \left(r(\mathbf{s}_t, \mathbf{a}_t) \right)}_{\widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)} \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t | \mathbf{s}_t) \right) \widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]$$

- It would be good to have the true value of Q to be used in the equation.
- But that is not available to us.

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \Bigg[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \underbrace{\sum_{t'=t}^{T} \Big(r(\mathbf{s}_t, \mathbf{a}_t) \Big)}_{\widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)} \Bigg]$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \Bigg[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \widehat{Q}^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \Bigg]$$

- It would be good to have the true value of Q to be used in the equation.
- But that is not available to us.
- Other alternatives are to estimate this value using methods that we have seen earlier - MC evaluation, Bootstrapped evaluation (TD), using function approximation for these.

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \Big[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - \mathop{\mathbb{E}}_{\mathbf{a}_t} [Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)] \big) \Big]$$

§ We can also use a baseline version of this.

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\Big[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t) - \mathop{\mathbb{E}}_{\mathbf{a}_t}\big[Q^{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t)\big]\big)\Big]$$

$$= \mathop{\mathbb{E}}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\Big[\sum_{t=1}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\Big)\big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t) - V^{\boldsymbol{\theta}}(\mathbf{s}_t)\big)\Big]$$

§ We can also use a baseline version of this.

Advantage Function

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \Big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - \mathop{\mathbb{E}}_{\mathbf{a}_t} [Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)] \Big) \right] \\
&= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \Big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - V^{\boldsymbol{\theta}}(\mathbf{s}_t) \Big) \right] \\
&= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) A^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]
\end{aligned}
$$

- We can also use a baseline version of this.
- This is called the ‘Advantage function’.

Advantage Function

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - \mathop{\mathbb{E}}_{\mathbf{a}_t} [Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t)] \big) \right] \\
&= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) \big(Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - V^{\boldsymbol{\theta}}(\mathbf{s}_t) \big) \right] \\
&= \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) A^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]
\end{aligned}
$$

- We can also use a baseline version of this.
- This is called the ‘Advantage function’.
- $A(\mathbf{s}_t, \mathbf{a}_t)$ can be approximated following the methods we used earlier (single sample backup or bootstrapping)

Advantage Function

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \Big[\sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \Big) A^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \Big]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \Big) A^{\boldsymbol{\theta}}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$$

- $Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \approx r(\mathbf{s}_t, \mathbf{a}_t) + V^{\boldsymbol{\theta}}(\mathbf{s}_t)$
- $A^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \approx r(\mathbf{s}_t, \mathbf{a}_t) + V^{\boldsymbol{\theta}}(\mathbf{s}_{t+1}) - V^{\boldsymbol{\theta}}(\mathbf{s}_t)$
- So we can use a neural network which learns to produce $V(\mathbf{s})$

Actor-Critic

An actor-critic algorithm

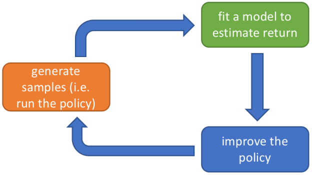
batch actor-critic algorithm:

1. sample $\{\mathbf{s}_i, \mathbf{a}_i\}$ from $\pi_\theta(\mathbf{a}|\mathbf{s})$ (run it on the robot)
2. fit $\hat{V}^\pi_{\mathsf{w}}(\mathbf{s})$ to sampled reward sums
3. evaluate $\hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i) = r(\mathbf{s}_i, \mathbf{a}_i) + \hat{V}^\pi_{\mathsf{w}}(\mathbf{s}'_i) - \hat{V}^\pi_{\mathsf{w}}(\mathbf{s}_i)$
4. $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta(\mathbf{a}_i|\mathbf{s}_i)\hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i)$
5. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

$$y_{i,t} \approx r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) + \hat{V}^\pi_{\mathsf{w}}(\mathbf{s}_{i,t+1})$$

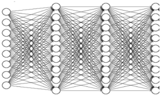
$$\mathcal{L}(\mathsf{w}) = \frac{1}{2}\sum_i \left\| \hat{V}^\pi_{\mathsf{w}}(\mathbf{s}_i) - y_i \right\|^2$$

fit $\hat{V}^\pi$



$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

$\mathbf{s}$



$\hat{V}^\pi(\mathbf{s})$

parameters w

$$V^\pi(\mathbf{s}_t) = \sum_{t'=t}^{T} E_{\pi_\theta}\left[r(\mathbf{s}_{t'}, \mathbf{a}_{t'})|\mathbf{s}_t\right]$$

Figure credit: *[Sergey Levine, UC Berkeley]*