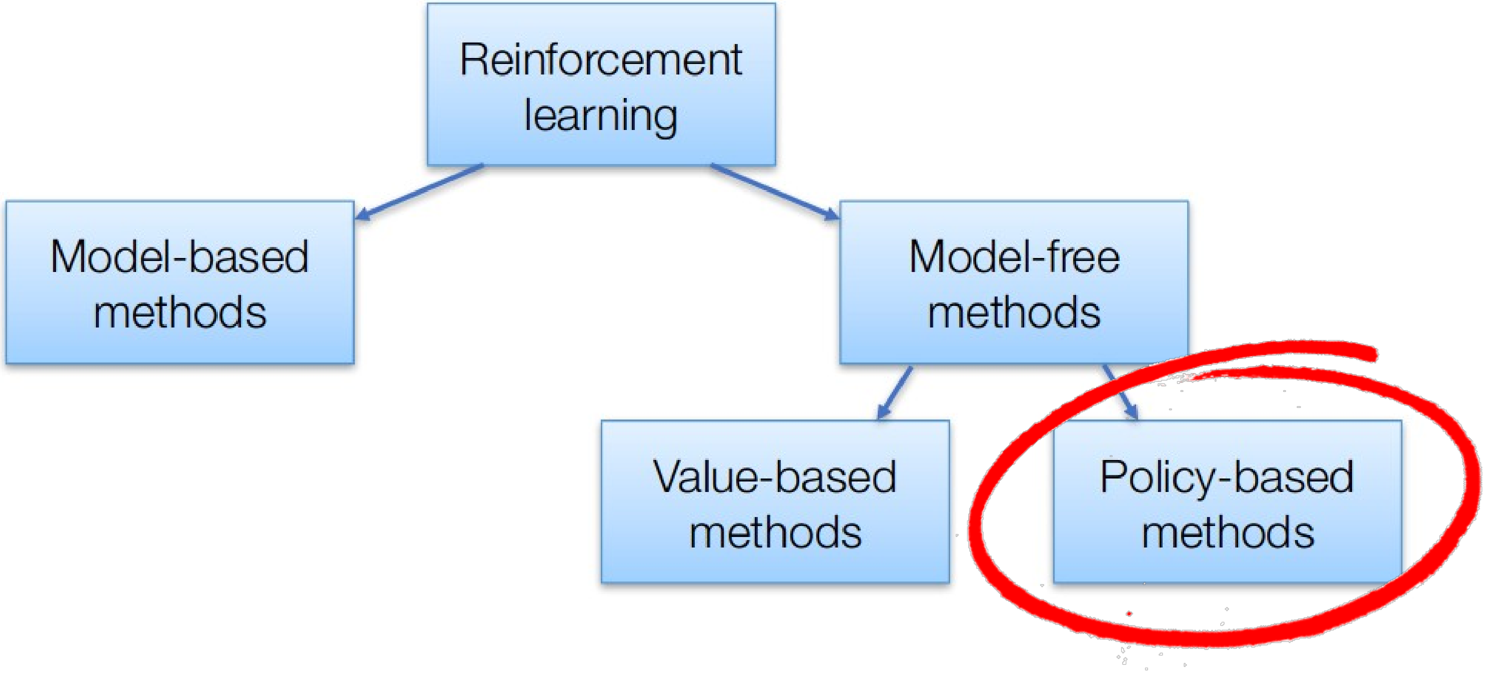
Policy Gradients and Actor Critic

Goal: Find an optimal policy that maximises expected return in a given environment

Policy Gradients

Goal: Find an optimal policy that maximises expected return in a given environment



Policy Gradients

Goal: Find an optimal policy that maximises expected return in a given environment

Value-based methods

Policy-based methods

Optimise the policy via a value function e.g. Q-learning

Optimise the policy directly

Policy Gradients

Goal: Find an optimal policy that maximises expected return in a given environment

Pros & cons of policy gradient methods (vs Q-learning)

- ➕ Learns policy directly (good when value function is very complex)
- ➕ Can learn stochastic policies
- ➕ Works with continuous action spaces
- ➕ Faster convergence

- ➖ Can converge to local optima
- ➖ High variance of rewards leads to low sample efficiency
- ➖ Less stable performance

Policy Gradients

Goal: Find an optimal policy that maximises expected return in a given environment

Definitions

$\tau = (s_0, a_0, s_1, a_1, ..., s_{H-1}, a_{H-1}, s_H)$ is a trajectory

$\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a stochastic policy with parameters θ

$R(s_t, a_t)$ is the expected reward for taking action a_t in state s_t

$\mathcal{R}(\tau) = \sum_{t=0}^{T-1} R(s_t, a_t)$ is the expected reward over a trajectory τ

Policy Gradients

Goal: Find an optimal policy that maximises expected return in a given environment

Now define the Utility function $U(\theta)$ to be the expected return following policy π_θ

$$
\begin{aligned}
U(\theta) &= \mathbb{E}_{\pi_\theta}(\mathcal{R}(\tau)) \\
&= \sum_\tau P(\tau;\theta)\mathcal{R}(\tau)
\end{aligned}
$$

Goal: Find policy parameters that maximise $U(\theta)$:

$$
\arg\max_\theta (U(\theta)) = \arg\max_\theta \sum_\tau P(\tau;\theta)\mathcal{R}(\tau)
$$

Policy Gradients

Goal: Find policy parameters that maximise $U(\theta)$

Gradient ascent: $\theta \leftarrow \theta + \alpha \nabla_\theta \left[U(\theta) \right]$

Policy Gradients

The **REINFORCE** algorithm

$$\theta \leftarrow \theta + \alpha \nabla_\theta \left[U(\theta) \right]$$

with

$$\nabla_\theta \Big[U(\theta) \Big] = \mathbb{E}_{\pi_\theta} \left(\sum_{t=1}^{T-1} \Big(\nabla_\theta \left[\log \left(\pi_\theta(a_t | s_t) \right) \right] \Big) \mathcal{R}(\tau) \right)$$

$$\approx \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \mathcal{R}(\tau^{(i)}) \right)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Recall: $\theta \leftarrow \theta + \alpha \nabla_\theta \left[U(\theta) \right]$

Problems:

- High variance of $\nabla_\theta \left[U(\theta) \right]$
- (Potentially) slow to converge

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$\nabla_\theta\Big[U(\theta)\Big] \approx \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\mathcal{R}(\tau^{(i)})\right)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$\nabla_\theta\Big[U(\theta)\Big] \approx \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\mathcal{R}(\tau^{(i)})\right)$$
$$= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=1}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$
\begin{aligned}
\nabla_\theta \left[U(\theta) \right] &\approx \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \mathcal{R}(\tau^{(i)}) \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(\sum_{t'=1}^{T-1} R(s_{t'}^{(i)}, a_{t'}^{(i)}) \right) \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(\sum_{t'=1}^{t-1} R(s_{t'}^{(i)}, a_{t'}^{(i)}) + \sum_{t'=t}^{T-1} R(s_{t'}^{(i)}, a_{t'}^{(i)}) \right) \right)
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &\approx \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\mathcal{R}(\tau^{(i)})\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=1}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\cancel{\sum_{t'=1}^{t-1}R(s_{t'}^{(i)},a_{t'}^{(i)})}+\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$
\begin{aligned}
\nabla_\theta \left[U(\theta)\right] &\approx \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\mathcal{R}(\tau^{(i)})\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=1}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\cancel{\sum_{t'=1}^{t-1}R(s_{t'}^{(i)},a_{t'}^{(i)})}+\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(Q(s_t^{(i)},a_t^{(i)})\right)\right)
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &\approx \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\mathcal{R}(\tau^{(i)})\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=1}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\cancel{\sum_{t'=1}^{t-1}R(s_{t'}^{(i)},a_{t'}^{(i)})}+\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(\sum_{t'=t}^{T-1}R(s_{t'}^{(i)},a_{t'}^{(i)})\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{T-1}\left(\nabla_\theta\left[\log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\right]\right)\left(Q(s_t^{(i)},a_t^{(i)})\right)\right)
\end{aligned}
$$

Q Actor-Critic

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b: $$\nabla_\theta \Big[U(\theta)\Big] = \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta \left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau) - b\big)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$\nabla_\theta \Big[U(\theta) \Big] = \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \big(\mathcal{R}(\tau) - b \big)$$

$$= \sum_\tau \Big(P(\tau; \theta) \big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \big) \big(\mathcal{R}(\tau) - b \big) \Big)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$\nabla_\theta \Big[U(\theta)\Big] = \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau) - b\big)$$
$$= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau) - b\big)\Big)$$
$$= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$
\begin{aligned}
\nabla_\theta \Big[U(\theta) \Big] &= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log\left(P(\tau;\theta) \right) \right] \Big) \big(\mathcal{R}(\tau) - b \big) \\
&= \sum_\tau \Big(P(\tau;\theta) \big(\nabla_\theta \left[\log\left(P(\tau;\theta) \right) \right] \big) \big(\mathcal{R}(\tau) - b \big) \Big) \\
&= \sum_\tau \Big(P(\tau;\theta) \big(\nabla_\theta \left[\log\left(P(\tau;\theta) \right) \right] \big) \mathcal{R}(\tau) \Big) - \sum_\tau \Big(P(\tau;\theta) \big(\nabla_\theta \left[\log\left(P(\tau;\theta) \right) \right] \big) b \Big) \\
&= \sum_\tau \Big(P(\tau;\theta) \big(\nabla_\theta \left[\log\left(P(\tau;\theta) \right) \right] \big) \mathcal{R}(\tau) \Big) - \sum_\tau P(\tau;\theta) \left(\frac{\nabla_\theta \left[\left(P(\tau;\theta) \right) \right]}{P(\tau;\theta)} \right) b
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau)-b\big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau)-b\big)\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau P(\tau;\theta)\left(\frac{\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]}{P(\tau;\theta)}\right)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]\Big)b
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau)-b\big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau)-b\big)\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau P(\tau;\theta)\left(\frac{\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]}{P(\tau;\theta)}\right)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]\Big)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot\nabla_\theta\Big[\sum_\tau\Big(\left(P(\tau;\theta)\right)\Big)\Big]
\end{aligned}$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau)-b\big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau)-b\big)\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau P(\tau;\theta)\left(\frac{\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]}{P(\tau;\theta)}\right)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]\Big)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot\nabla_\theta\Big[\sum_\tau\Big(\left(P(\tau;\theta)\right)\Big)\Big] \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot\nabla_\theta\Big[1\Big]
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau)-b\big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau)-b\big)\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big) \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau P(\tau;\theta)\left(\frac{\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]}{P(\tau;\theta)}\right)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - \sum_\tau \Big(\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]\Big)b \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot\nabla_\theta\Big[\sum_\tau\Big(\left(P(\tau;\theta)\right)\Big)\Big] \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot\nabla_\theta\Big[1\Big] \\
&= \sum_\tau \Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big) - b\cdot 0
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline b:

$$
\begin{aligned}
\nabla_\theta\Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau)-b\big)\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\big(\mathcal{R}(\tau)-b\big)\Big)\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-\sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)b\Big)\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-\sum_\tau P(\tau;\theta)\left(\frac{\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]}{P(\tau;\theta)}\right)b\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-\sum_\tau\Big(\nabla_\theta\left[\left(P(\tau;\theta)\right)\right]\Big)b\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-b\cdot\nabla_\theta\Big[\sum_\tau\Big(\left(P(\tau;\theta)\right)\Big)\Big]\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-b\cdot\nabla_\theta\Big[1\Big]\\
&= \sum_\tau\Big(P(\tau;\theta)\big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\big)\mathcal{R}(\tau)\Big)-b\cdot 0\\
&= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta\left[\log\left(P(\tau;\theta)\right)\right]\Big)\mathcal{R}(\tau)
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline $b(s)$:

$$\nabla_\theta \Big[U(\theta) \Big] = \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \big(\mathcal{R}(\tau) - b(s) \big)$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline $b(s)$: e.g $b(s) = V(s)$

$$\begin{aligned}\nabla_\theta \Big[U(\theta)\Big] &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta \left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau) - b(s)\big) \\ &= \mathbb{E}_{\pi_\theta}\Big(\nabla_\theta \left[\log\left(P(\tau;\theta)\right)\right]\Big)\big(\mathcal{R}(\tau) - V(s)\big)\end{aligned}$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

Introduce a baseline $b(s)$: e.g $b(s) = V(s)$

$$
\begin{aligned}
\nabla_\theta \Big[U(\theta) \Big] &= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - b(s) \right) \\
&= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - V(s) \right) \\
&= \dots \\
&\approx \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(Q(s_t^{(i)}, a_t^{(i)}) - V(s_t^{(i)}) \right) \right)
\end{aligned}
$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$$\boxed{A(s_t, a_t) = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)}$$

Introduce a baseline $b(s)$: e.g $b(s) = V(s)$

$$\begin{aligned}
\nabla_\theta \Big[U(\theta) \Big] &= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau;\theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - b(s) \right) \\
&= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau;\theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - V(s) \right) \\
&= \dots \\
&\approx \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(Q(s_t^{(i)}, a_t^{(i)}) - V(s_t^{(i)}) \right) \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(A(s_t^{(i)}, a_t^{(i)}) \right) \right)
\end{aligned}$$

Policy Gradients

Improvements to the **REINFORCE** algorithm

$A(s_t, a_t) = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$

Introduce a baseline $b(s)$: e.g $b(s) = V(s)$

$$
\begin{aligned}
\nabla_\theta \left[U(\theta) \right] &= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - b(s) \right) \\
&= \mathbb{E}_{\pi_\theta} \Big(\nabla_\theta \left[\log \left(P(\tau; \theta) \right) \right] \Big) \left(\mathcal{R}(\tau) - V(s) \right) \\
&= \dots \\
&\approx \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(Q(s_t^{(i)}, a_t^{(i)}) - V(s_t^{(i)}) \right) \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{T-1} \left(\nabla_\theta \left[\log \left(\pi_\theta(a_t^{(i)} | s_t^{(i)}) \right) \right] \right) \left(A(s_t^{(i)}, a_t^{(i)}) \right) \right)
\end{aligned}
$$

Advantage Actor-Critic