Advantage Function

$$A_\pi(s,a) = q_\pi(s,a) - v_\pi(s)$$

Connection to Baselines

Reminder to Policy Gradients:

$$\nabla_\theta v(s_0) \propto \sum_s \mu(s) \sum_a q(s,a) \nabla_\theta \pi(a|s)$$

Generanlized Policy Gradient:

$$\nabla_\theta v_\pi(\cdot) \propto \sum_s \mu(s) \sum_a \Phi(s) \nabla_\theta \pi(a|s)$$

With Φ potentially (but not exclusive) as $\Phi(s) = \mathbb{E}[G] - b(s)$. With baseline b, this case is valid because:

$$\sum_s \mu(s) \sum_a \left[\mathbb{E}[G] - b(s)\right] \nabla_\theta \pi(a|s) =$$
$$\sum_s \mu(s) \sum_a \mathbb{E}[G] \nabla_\theta \pi(a|s) - \sum_s \mu(s) \sum_a b(s) \nabla_\theta \pi(a|s)$$

For the latter term being 0 since:

$$\sum_s \mu(s) \sum_a b(s) \nabla_\theta \pi(a|s) =$$
$$\sum_s \mu(s) b(s) \sum_a \nabla_\theta \pi(a|s) =$$
$$\sum_s \mu(s) b(s) \nabla_\theta \sum_a \pi(a|s)$$

And since $\sum_a \pi(a|s) = 1$ it follows that:

$$\sum_s \mu(s) b(s) \nabla_\theta 1 = \sum_s \mu(s) b(s) \cdot 0 = 0$$

And therefore:

$$\sum_s \mu(s) \sum_a \mathbb{E}[G] \nabla_\theta \pi(a|s) = \sum_s \mu(s) \sum_a \left[\mathbb{E}[G] - b(s)\right] \nabla_\theta \pi(a|s)$$

for any baseline b that is not dependent on any future action $a_t, a_{t+1}, \ldots, a_\infty$.

By chosing $b(s) = v_\pi(s)$ we reduce the bias without introducing variance ($\mathbb{V} \neq v_\pi$) for our optimization target; which is a nice property to have!

For $\Phi_{\text{baseline}} = q_\pi - v_\pi$ we get

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}_{s,a}[(\Phi_{\text{baseline}} - \mathbb{E}[\Phi_{\text{baseline}}])^2]$$

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}[(q_\pi(s,a) - v_\pi(s) - \mathbb{E}[q_\pi(s,a) - v_\pi(s))^2]$$

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}_{s,a}[(q_\pi(s,a) - v_\pi(s) - \mathbb{E}_{s,a}[q_\pi(s,a)] + \mathbb{E}_{s,a}[v_\pi(s)])^2]$$

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}_{s,a}[(q_\pi(s,a) - v_\pi(s) - v_\pi(s) + v_\pi(s))^2]$$

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}_{s,a}[(q_\pi(s,a) - v_\pi(s))^2]$$

For $\Phi_{\text{raw}} = q_\pi$ we get:

$$\mathbb{V}[\Phi_{\text{raw}}] = \mathbb{E}_{s,a}[(\Phi_{\text{raw}} - \mathbb{E}_{s,a}[\Phi_{\text{raw}}])^2]$$

$$\mathbb{V}[\Phi_{\text{raw}}] = \mathbb{E}_{s,a}[(q_\pi(s,a) - \mathbb{E}_{s,a}[q_\pi(s,a)])^2]$$

$$\mathbb{V}[\Phi_{\text{raw}}] = \mathbb{E}_{s,a}[(q_\pi(s,a) - v_\pi(s))^2]$$

Therefore we get:

$$\mathbb{V}[\Phi_{\text{baseline}}] = \mathbb{E}_{s,a}[A(s,a)^2] = \mathbb{V}[\Phi_{\text{raw}}]$$

We just outsmarted the bias/variance-tradeoff, by correcting the bias with the baseline without changing the variance. Turns out that we can actually have our cake and eat it too!

Using A_π as target for our policy optimization we implicitly chose the baseline $b(s) = v_\pi(s)$ with:

$$A_\pi = q_\pi - v_\pi = \mathbb{E}[G] - \mathbb{E}[G]$$

This results in an unbiased estimator for policy gradient acent on sampled trajectories!

Zero Expectation

Theorem:

$$\mathbb{E}_\pi[A_\pi(s,a)] = 0$$

Proof:

$$\mathbb{E}_\pi[A_\pi(s,a)] = \sum_{a \in \mathcal{A}(s)} \pi(a|s)[q_\pi(s,a) - v_\pi(s)]$$

$$\sum_{a\in\mathcal{A}(s)} \pi(a|s)\left[\sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s)] - \sum_{a'} \pi(a'|s)\sum_{s',r} p(s',r|s,a')[r+\gamma v_\pi(s)]\right]$$

$$\sum_{a\in\mathcal{A}(s)} \pi(a|s)\sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s)] - \\ \sum_{a\in\mathcal{A}(s)} \pi(a|s) \sum_{a'\in\mathcal{A}(s)} \pi(a'|s)\sum_{s',r} p(s',r|s,a')[r+\gamma v_\pi(s)]$$

Since $\sum_a \pi(a) = 1$ therefore $\sum_{a\in\mathcal{A}(s)} [\pi(a|s)X] = X$ as long as X is not a function of a:

$$\sum_{a\in\mathcal{A}(s)} \pi(a|s)\sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s)] - \\ \sum_{a'\in\mathcal{A}(s)} \pi(a'|s)\sum_{s',r} p(s',r|s,a')[r+\gamma v_\pi(s)] = 0 \quad \text{Q.E.D.}$$

Short Proof:

$$\mathbb{E}_\pi[A_\pi(s,a)] = \mathbb{E}_\pi[q_\pi(s,a) - v_\pi(s)]$$

$$\mathbb{E}_\pi[A_\pi(s,a)] = \mathbb{E}_\pi\left[q_\pi(s,a) - \mathbb{E}_\pi[q_\pi(s,a)]\right]$$

$$\mathbb{E}_\pi[A_\pi(s,a)] = \mathbb{E}_\pi[q_\pi(s,a)] - \mathbb{E}_\pi[q_\pi(s,a)] = 0$$

TD Expectation

TD-Error:

$$\delta = r + \gamma v_\pi(s') - v(s)$$

Theorem:

$$\mathbb{E}[\delta|s,a] = A_\pi(s,a)$$

Proof:

$$\mathbb{E}[\delta|s,a] = \mathbb{E}[r + \gamma v_\pi(s') - v_\pi(s)|s,a]$$

$$\mathbb{E}_\pi[\delta|s,a] = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s') - v_\pi(s)]$$

$$\mathbb{E}_\pi[\delta|s,a] = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s')] \\ - \sum_{r,s'} p(s',r|s,a) v_\pi(s)$$

Because $v_\pi(s)$ is not a function of r and s' we can pull $v_\pi(s)$ out.

$$\mathbb{E}_\pi[\delta|s,a] = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s')] \\ -v_\pi(s)\sum_{s',r} p(s',r|s,a)$$

By the identity $\sum_{s',r} p(s',r|s,a) = 1$ we can eliminate the sum.

$$\mathbb{E}_\pi[\delta|s,a] = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s')] - v_\pi(s)$$

Substituting for $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r+\gamma v_\pi(s')]$ we get:

$$\mathbb{E}_\pi[\delta|s,a] = q_\pi(s,a) - v_\pi(s') = A_\pi(s,a) \quad \text{Q.E.D.}$$

Short Proof:

$$\mathbb{E}[\delta|s,a] = \mathbb{E}[r+\gamma v_\pi(s') - v_\pi(s)|s,a]$$

$$\mathbb{E}[\delta|s,a] = \mathbb{E}[r+\gamma v_\pi(s')|s,a] - \mathbb{E}[v_\pi(s)|s,a]$$

$$\mathbb{E}[\delta|s,a] = q_\pi(s,a) - v_\pi(s) = A_\pi(s,a)$$

Generalized Advantage as a sum over TD Errors

Reminder:

$$\hat{A}_t^{(1)} = q(s_t,a_t) - v(s_t) = r_t + \gamma v(s_{t+1}) - v(s_t)$$

$$\hat{A}_t^{(2)} = q(s_t,a_t,a_{t+1}) - v(s_t) = r_t + \gamma r_{t+1} + \gamma^2 v(s_{t+2}) - v(s_t)$$

$$\hat{A}_t^{(3)} = q(s_t,a_t,a_{t+1},a_{t+2}) - v(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 v(s_{t+3}) - v(s_t)$$

$$\hat{A}_t^{(k)} = \sum_{l=0}^{k-1}[\gamma^l r_{t+l}] + \gamma^k v(s_{t+k}) - v(s_t)$$

Theorem:

$$\sum_{k=0}^{\infty} \gamma^k \delta_{t+k} = \lim_{k\to\infty} \hat{A}_t^{(k)}$$

Proof:

$$\sum_{k=0}^{\infty} \gamma^k \delta_k = \gamma^0 \delta_0 + \gamma^1 \delta_1 + \gamma^2 \delta_2 + \dots$$

$$\begin{gathered}\sum_{k=0}^{\infty} \gamma^k \delta_k = \\ \gamma^0 [r_0 + \gamma v(s_1) - v(s_0)] \\ + \gamma^1 [r_1 + \gamma v(s_2) - v(s_1)] \\ + \gamma^2 [r_2 + \gamma v(s_3) - v(s_2)] + \dots\end{gathered}$$

Pulling out the V terms from each δ yields:

$$\begin{gathered}\sum_{k=0}^{\infty} \gamma^k \delta_k = \\ \gamma^0 r_0 + \gamma^1 v(s_1) - \gamma^0 v(s_0) \\ + \gamma^1 r_1 + \gamma^2 v(s_2) - \gamma^1 v(s_1) \\ + \gamma^2 r_2 + \gamma^3 v(s_3) - \gamma^2 v(s_2) + \dots\end{gathered}$$

From this we can eliminate the redundant terms leaving only the first V term.

$$\begin{gathered}\sum_{k=0}^{\infty} \gamma^k \delta_k = \\ \gamma^0 r_0 + \cancel{\gamma^1 v(s_1)} - \gamma^0 v(s_0) \\ + \gamma^1 r_1 + \cancel{\gamma^2 v(s_2)} - \cancel{\gamma^1 v(s_1)} \\ + \gamma^2 r_2 + \cancel{\gamma^3 v(s_3)} - \cancel{\gamma^2 v(s_2)} + \dots\end{gathered}$$

$$\sum_{k=0}^{\infty} \gamma^k \delta_k = r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots - v(s_0) = \lim_{k \to \infty} \hat{A}_t^{(k)} \quad \text{Q.E.D.}$$