Policy Gradient Theorem

We are interested in finding the gradient of the statevalue function $\nabla_\theta v_{\pi_\theta}(s)$ with respect to the policy parameters θ as a function of the policy gradient $\nabla_\theta \pi_\theta(a|s)$. By applying thre bellman equation we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \nabla_\theta \left[\sum_a \pi_\theta(a|s) q_{\pi_\theta}(s,a) \right]$$

By the identity $\nabla[a+b] = \nabla a + \nabla b$ we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_a \nabla_\theta [\pi_\theta(a|s) q_{\pi_\theta}(s,a)]$$

By the identity $\nabla[a \cdot b] = b \cdot \nabla a + a \cdot \nabla b$ we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta q_{\pi_\theta}(s,a)$$

By expanding the last $q_{\pi_\theta}(s,a)$ term according to the bellman equation we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta \left[\sum_{r,s'} p(s',r|s,a)[r + \gamma v_{\pi_\theta}(s')] \right]$$

By the identity $\nabla[a+b] = \nabla a + \nabla b$ we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \left[\sum_{r,s'} p(s',r|s,a)[\nabla_\theta r + \gamma \nabla_\theta v_{\pi_\theta}(s')] \right]$$

Since r is conditionally independent of θ i.e. $p(r|s,a,s',\theta) = p(r|s,a,s')$ we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \sum_{r,s'} p(s',r|s,a) \gamma \nabla_\theta v_{\pi_\theta}(s')$$

At this point we have a recursive relation ship of $\nabla_\theta v_{\pi_\theta}(s)$ to it self $\nabla_\theta v_{\pi_\theta}(s')$. Therefore we can substitute s with s' and self insert the recursive relation for one step.

$$\nabla_\theta v_{\pi_\theta}(s) = \left[\sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \left[\sum_{r,s'} p(s',r|s,a) \right.\right.$$

$$\gamma \left[\sum_{a'} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s', a') + \pi_\theta(a'|s') \left[\sum_{r', s''} p(s'', r'|s', a') \gamma \nabla_\theta v_{\pi_\theta}(s'') \right] \right] \Bigg] \Bigg]$$

Repeating this step leads to an infinite regression. For this regression we can seperate sums according to the distribution identity $\sum_x c(x) \cdot [f(x) + g(x)] = \sum_x c(x) \cdot f(x) + \sum_x c(x) \cdot g(x)$. By applying this step and pulling γ outside of the sums we get:

$$\begin{gathered} \nabla_\theta v_{\pi_\theta}(s) = \\ \gamma^0 \left[\sum_a \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s, a) \right] + \\ \gamma^1 \cdot \sum_a \pi_\theta(a|s) \sum_{r, s'} p(s', r|s, a) \left[\sum_{a'} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s', a') \right] + \\ \gamma^2 \cdot \sum_a \pi_\theta(a|s) \sum_{r, s'} p(s', r|s, a) \sum_{a'} \pi_\theta(a'|s') \sum_{r', s''} p(s'', r'|s', a') \left[\sum_{a''} \nabla_\theta \pi_\theta(a''|s'') q(s'', \right. \\ \vdots \end{gathered}$$

We can define the path probability of getting from state s to state $\hat{s}$ in exactly t timesteps under policy π_θ as $\mathrm{Pr}(s \to \hat{s}, t, \pi_\theta)$ with the following properties:

$$\mathrm{Pr}(s \to \hat{s}, 0, \pi_\theta) = \begin{cases} 1 & \text{if } s = \hat{s} \\ 0 & \text{else} \end{cases}$$

$$\mathrm{Pr}(s \to s'', t+1, \pi_\theta) = \sum_a \pi_\theta(a|s) \sum_{s'} p(s'|a, s) \mathrm{Pr}(s' \to s'', t, \pi_\theta)$$

By induction we can show that this path probability Pr can be used to substitue the factors of the infinite regression sum for $\nabla_\theta v_{\pi_\theta}(s)$.

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_{\hat{s}} \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s \to \hat{s}, t, \pi_\theta) \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

Since $\sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$ does not depend on the variable t we can write the previous equation as:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_{\hat{s}} \left[\sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s \to \hat{s}, t, \pi_\theta) \right] \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

The term $\sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s \to \hat{s}, t, \pi_\theta)$ is the discounted expected time spend in state $\hat{s}$ for an episode. It is a quantitative measure of how much state $\hat{s}$ contributes to state s in absolute terms. This measure will be abreviated as $\eta(\hat{s})$:

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_{\hat{s}} \eta(\hat{s}) \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

By inserting the independent factor $\frac{\sum_{\hat{s}} \eta(\hat{s})}{\sum_{\hat{s}} \eta(\hat{s})} = 1$ we get the following equation.

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_{\hat{s}} \frac{\sum_{\hat{s}} \eta(\hat{s})}{\sum_{\hat{s}} \eta(\hat{s})} \eta(\hat{s}) \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

$$\nabla_\theta v_{\pi_\theta}(s) = \left[\sum_{\hat{s}} \eta(\hat{s})\right] \sum_{\hat{s}} \frac{\eta(\hat{s})}{\sum_{\hat{s}} \eta(\hat{s})} \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

By substituting $\frac{\eta(\hat{s})}{\sum_{\hat{s}} \eta(\hat{s})}$ with $\mu(s)$ we get:

$$\nabla_\theta v_{\pi_\theta}(s) = \left[\sum_{\hat{s}} \eta(\hat{s})\right] \sum_{\hat{s}} \mu(\hat{s}) \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

By examining $\mu(\hat{s})$ we can see that μ gives the probability of being in state $\hat{s}$ over after infinite time steps where each step is weighted by the discountfactor γ to account for the relevance that the given state in time step t has for the overall value estimation. Therefore we can drop the proportionality constant $\sum_{\hat{s}} \eta(\hat{s})$ to get:

$$\nabla_\theta v_{\pi_\theta}(s) \propto \sum_{\hat{s}} \mu(\hat{s}) \sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)$$

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s}} \left[\sum_a \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)\right]$$

Further we can add the factor $\frac{\pi_\theta(a|\hat{s})}{\pi_\theta(a|\hat{s})} = 1$ to obtain:

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s}} \left[\sum_a \frac{\pi_\theta(a|\hat{s})}{\pi_\theta(a|\hat{s})} \nabla_\theta \pi_\theta(a|\hat{s}) q(\hat{s}, a)\right]$$

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s}} \left[\sum_a \pi_\theta(a|\hat{s}) \frac{\nabla_\theta \pi_\theta(a|\hat{s})}{\pi_\theta(a|\hat{s})} q(\hat{s}, a)\right]$$

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s}} \left[\mathbb{E}_a \frac{\nabla_\theta \pi_\theta(a|\hat{s})}{\pi_\theta(a|\hat{s})} q(\hat{s}, a)\right]$$

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s},a} \left[\frac{\nabla_\theta \pi_\theta(a|\hat{s})}{\pi_\theta(a|\hat{s})} q(\hat{s}, a)\right]$$

By the identity $\frac{\nabla x}{x} = \nabla \ln(x)$ we get the update for the **REINFORCE** algorithm:

$$\nabla_\theta v_{\pi_\theta}(s) \propto \mathbb{E}_{\hat{s},a}\left[q(\hat{s},a)\nabla_\theta \ln \pi_\theta(a|\hat{s})\right]$$