Policy Gradient Methods

Khimya Khetarpal
Kushal Arora

1. Policy Gradient for a Mixture of Policies

1.1. Proof of Policy Gradient for Policy over options μ

$$v(s;\theta,w) = \sum_o \mu(o|s;\theta) \sum_a \pi(a|s,o;w)\bigg(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s';\theta,w)\bigg) \tag{1}$$

Deriving the gradient of the v with respect to parameter θ:

$$\begin{split}\frac{\partial v}{\partial \theta}(s;\theta,w) = \sum_o \mu(o|s;\theta) \sum_a \pi(a|s,o;w)\gamma \sum_{s'} P(s'|s,a)\frac{\partial v}{\partial \theta}(s';\theta,w) + \\ \sum_o \frac{\partial \mu}{\partial \theta}(o|s;\theta) \sum_a \pi(a|s,o;w)\bigg(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s';\theta,w)\bigg)\end{split} \tag{2}$$

Now considering

$$P_\gamma^{(1)}(s'|s) = \sum_o \mu(o|s;\theta) \sum_a \pi(a|s,o;w)\gamma P(s'|s,a) \tag{3}$$

Also, From the definition of action-value-function

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s';\theta,w) \tag{4}$$

Taking an expectation of (4) for all possible actions, we can write

$$v_\pi(s,o;\theta,w) = \sum_a \pi(a|s,o;w)\bigg(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s';\theta,w)\bigg) \tag{5}$$

Substituting (3) and (5) in (2):

$$\frac{\partial v}{\partial \theta}(s;\theta,w) = \sum_{s'} P_\gamma^{(1)}(s'|s)\frac{\partial v}{\partial \theta}(s';\theta,w) + \sum_o \frac{\partial \mu}{\partial \theta}(o|s;\theta)v_\pi(s,o;\theta,w) \tag{6}$$

Since here we have the recurrence term of v for current state s and next state s',
Now unrolling the gradient of v we can write (6) as follows:

$$\frac{\partial v}{\partial \theta}(s;\theta,w) = \sum_{s'} \sum_{k=0}^{\infty} P_\gamma^{(k)}(s'|s) \sum_o \frac{\partial \mu}{\partial \theta}(o|s;\theta)v_\pi(s,o;\theta,w) \tag{7}$$

Now, let us consider

$$\Omega(s'|s) = \sum_{k=0}^{\infty} P_{\gamma}^{(k)}(s'|s) \tag{8}$$

Substituting (8) in (7), the final equation becomes:

$$\boxed{\frac{\partial v}{\partial \theta}(s;\theta,w) = \sum_{s'} \Omega(s'|s) \sum_{o'} \frac{\partial \mu}{\partial \theta}(o'|s';\theta) v_{\pi}(s',o';\theta,w)} \tag{9}$$

1.2. Proof of Policy Gradient for Policy over actions π

Given the Bellman equation for the model:

$$v(s;\theta,w) = \sum_{o} \mu(o|s;\theta) \sum_{a} \pi(a|s,o;w)\Big(r(s,a) + \gamma \sum_{s'} P(s'|s,a) v(s';\theta,w)\Big) \tag{10}$$

Deriving the gradient of the v with respect to parameter w:

$$\begin{aligned}\frac{\partial v}{\partial w}(s;\theta,w) = \sum_{o} \mu(o|s;\theta) \sum_{a} \frac{\partial \pi}{\partial w}(a|s,o;w)\Big(r(s,a) + \gamma \sum_{s'} P(s'|s,a) v(s';\theta,w)\Big) + \\ \sum_{o} \mu(o|s;\theta) \sum_{a} \pi(a|s,o;w)\Big(\gamma \sum_{s'} P(s'|s,a) \frac{\partial v}{\partial w}(s';\theta,w)\Big)\end{aligned} \tag{11}$$

Here;

$$q_{\pi}(s,a,o;w,\theta) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) v(s';\theta,w) \tag{12}$$

Now again similar to before, substituting (3), and (12) in (11):

$$\frac{\partial v}{\partial w}(s;\theta,w) = \sum_{o} \mu(o|s;\theta) \sum_{a} \frac{\partial \pi}{\partial w}(a|s,o;w) q_{\pi}(s,a,o;w,\theta) + \sum_{s'} P_{\gamma}^{(1)} \frac{\partial v}{\partial w}(s';\theta,w) \tag{13}$$

Since here we have the recurrence term of v for current state s and next state s',
Now unrolling the gradient of v we can write (13) as follows:

$$\frac{\partial v}{\partial w}(s;\theta,w) = \sum_{s'} \sum_{k=0}^{\infty} P_{\gamma}^{(k)}(s'|s) \sum_{o} \mu(o|s;\theta) \sum_{a'} \frac{\partial \pi}{\partial w}(a'|s',o';w) q_{\pi}(s',a',o';w,\theta) \tag{14}$$

Again, let us consider

$$\Omega(s'|s) = \sum_{k=0}^{\infty} P_{\gamma}^{(k)}(s'|s) \tag{15}$$

Substituting (15) in (14), the final equation becomes:

$$\boxed{\frac{\partial v}{\partial \theta}(s;\theta,w) = \sum_{s'} \Omega(s'|s) \sum_{o'} \mu(o'|s';\theta) \sum_{a'} \frac{\partial \pi}{\partial w}(a'|s',o';w) q_{\pi}(s',a',o';w,\theta)} \tag{16}$$

2. Policy Gradient Hessian

The state-value function can be written in terms of the action-value function as follows;

$$V_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) \tag{17}$$

Note here state-value function is parameterized with θ. For Hessian, let us consider the two components of θ namely; θ_i and θ_j. Gradient of the state-value function with respect to θ_i can thus be written as:

$$\frac{\partial V}{\partial \theta_i}(s;\theta_i) = \sum_a \frac{\partial \pi}{\partial \theta_i}(a|s) q_\pi(s,a) + \sum_a \pi(a|s) \frac{\partial q_\pi}{\partial \theta_i}(s,a) \tag{18}$$

Now deriving (18) w.r.t. to θ_j, we get:

$$\begin{aligned}\frac{\partial^2 V}{\partial \theta_i \theta_j}(s;\theta) = \sum_a \frac{\partial^2 \pi}{\partial \theta_i \theta_j}(a|s) q_\pi(s,a) + \sum_a \frac{\partial \pi}{\partial \theta_i}(a|s) \frac{\partial q_\pi}{\partial \theta_j}(s,a) + \\ \sum_a \frac{\partial \pi}{\partial \theta_j}(a|s) \frac{\partial q_\pi}{\partial \theta_i}(s,a) + \sum_a \pi(a|s) \frac{\partial^2 q_\pi}{\partial \theta_i}(s,a)\end{aligned} \tag{19}$$

In (19), $q_\pi(s,a)$ can be written as:

$$q_\pi(s,a) = \sum_{s',r} p(s'|s,a) \Big(r + \gamma V(s';\theta) \Big) \tag{20}$$

Taking the gradient w.r.t θ_i

$$\frac{\partial q_\pi}{\partial \theta_i} = \sum_{s',r} p(s'|s,a) \Big(\gamma \frac{\partial V}{\partial \theta_i}(s';\theta) \Big) \tag{21}$$

Further, Taking the second order gradient w.r.t θ_i

$$\frac{\partial^2 q_\pi}{\partial \theta_i} = \sum_{s',r} p(s'|s,a) \Big(\gamma \frac{\partial^2 V}{\partial \theta_i}(s';\theta) \Big) \tag{22}$$

Substituting (22) in (19), we get;

$$\begin{aligned}\frac{\partial^2 V}{\partial \theta_i \theta_j}(s;\theta) = \sum_a \frac{\partial^2 \pi}{\partial \theta_i \theta_j}(a|s) q_\pi(s,a) + \sum_a \frac{\partial \pi}{\partial \theta_i}(a|s) \frac{\partial q_\pi}{\partial \theta_j}(s,a) + \\ \sum_a \frac{\partial \pi}{\partial \theta_j}(a|s) \frac{\partial q_\pi}{\partial \theta_i}(s,a) + \sum_{s',r} p(s'|s,a) \gamma \frac{\partial^2 V}{\partial \theta_i}(s',a)\end{aligned} \tag{23}$$

Unrolling the second order gradient of V and substituting the state distribution $\rho^\pi(s)$ in (23), we get the final equation to be:

$$\frac{\partial^2 V}{\partial \theta_i \theta_j}(s;\theta) = \sum_{s,a} \rho^\pi(s) \Bigg(\frac{\partial^2 \pi}{\partial \theta_i \theta_j}(a|s) q_\pi(s,a) + \sum_a \frac{\partial \pi}{\partial \theta_i}(a|s) \frac{\partial q_\pi}{\partial \theta_j}(s,a) + \sum_a \frac{\partial \pi}{\partial \theta_j}(a|s) \frac{\partial q_\pi}{\partial \theta_i}(s,a) \Bigg) \tag{24}$$

From (24), the final result can also be written as following:

$$\boxed{Hessian(s;\theta) = \sum_{s,a} \rho^\pi(s) \Big(\nabla^2 \pi(a|s) Q^\pi(s,a) + \nabla \pi(a|s) \nabla Q^\pi(s,a)^T + \nabla Q^\pi(s,a) \nabla \pi(a|s)^T \Big)} \tag{25}$$

3. Constrained Optimization/ Intrinsic Rewards

Given the following objective function:

$$J_\alpha(\theta) = \mathbb{E}_{\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t)\right] - \eta\,\mathbb{E}_{\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t c(S_t, A_t)\right] \tag{26}$$

where; α: Initial distribution over states θ: Policy parameter

To maximize the expected discounted return but penalize cost c we can say

$$\begin{aligned}&\max_\theta J_\alpha(\theta),\\ &\text{where } J(\theta|d) = \mathbb{E}_{\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t) - \eta\gamma^t c(S_t, A_t)\right]\end{aligned} \tag{27}$$

Here, action-value function and the cost function can be seen as a single MDP whose reward is now the difference between the reward of the base MDP and the cost function as shown in (27).

Thus the above modified reward MDP can be interpreted as a set of Bellman equations which the new value function $V_\theta^c(s)$ equivalent to $Q^c(s,a)$ over the transformed reward. Thus, these should now satisfy the following rule:

$$V_\theta^c(s) = \sum_{a,s'}\pi_\theta(a|s)P(s'|s,a)\Big(r(s,a) - \eta c(s,a) + \gamma V_\theta^c(s')\Big) \tag{28}$$

Writing the modified transformed reward:

$$\hat{r}(s,a) = r(s,a) - \eta c(s,a) \tag{29}$$

Substituting (29) and taking a gradient of (28) w.r.t θ, we get:

$$\begin{aligned}\nabla V_\theta^c(s) = \sum_{a,s'}\frac{\partial\pi_\theta}{\partial\theta}(a|s)P(s'|s,a)\Big\{\hat{r}(s,a) + \gamma V_\theta^c(s)\Big\} +\\ \sum_{a,s'}\pi_\theta(a|s)P(s'|s,a)\Big\{\gamma\nabla V_\theta^c(s')\Big\}\end{aligned} \tag{30}$$

With the modified new reward, we can then write:

$$\begin{aligned}\nabla V_\theta^c(s) = \sum_{a}\frac{\partial\pi_\theta}{\partial\theta}(a|s)\Big\{\hat{r}(s,a) + \gamma\sum_{s'}P(s'|s,a)V_\theta^c(s)\Big\} +\\ \sum_{a,s'}\pi_\theta(a|s)P(s'|s,a)\Big\{\gamma\nabla V_\theta^c(s')\Big\}\end{aligned} \tag{31}$$

Now here $\hat{r}(s,a) + \gamma\sum_{s'}P(s'|s,a)V_\theta^c(s) = q_\pi(s,a)$

Substituting the above in (31), we get:

$$\nabla V_\theta^c(s) = \sum_{a}\frac{\partial\pi}{\partial\theta}(a|s)q_\pi(s,a) + \sum_{a,s'}\pi_\theta(a|s)P(s'|s,a)\gamma\nabla V_\theta^c(s') \tag{32}$$

Now unrolling $\nabla V^c_\theta(s')$, we get the following:

$$\begin{aligned}\nabla V^c_\theta(s) = \sum_a \frac{\partial \pi}{\partial \theta}(a|s) q_\pi(s,a) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \Bigg[\sum_{a'} \frac{\partial \pi}{\partial \theta}(a'|s') q_\pi(s',a') + \\ \sum_{a',s'} \pi_\theta(a'|s') P(s''|s',a') \gamma \nabla V^c_\theta(s'') \Bigg]\end{aligned} \tag{33}$$

With extension to k steps just as the usual policy gradient and above derivations, we get:

$$\nabla V^c_\theta(s) = \sum_{s'} \sum_{k=0}^{\infty} P^{(k)}_\gamma(s'|s) \sum_{a'} \frac{\partial \pi}{\partial \theta}(a'|s') q_\pi(s',a') \tag{34}$$

Substituting (15) in (34), we get the final equation as the following wherein the transformed reward function is given by $\hat{r}(s,a)$ as shown in (29)

$$\boxed{\frac{\partial V^c_\theta}{\partial \theta}(s) = \sum_{s'} \Omega(s'|s) \sum_{a'} \frac{\partial \pi}{\partial \theta}(a'|s') q_\pi(s',a')} \tag{35}$$