Assignment 3
Policy Gradient Methods - Track 1
COMP767: Reinforcement Learning

Simon Guiroy

27th March, 2018

1. Policy Gradient for a Mixture of Policies

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

We want to derive the policy gradient according to μ the policy over options, and to the individual policies π. We can think about this model in a hierarchic way, seeing μ as a policy itself, but branching to other policies, rather than actions. This model could theoretically have an arbitrary hierarchical depth.

In the formal definition of a policy gradient, the objective function is the state value function of that policy, for a given initial state $v_{\pi_w}(s_0)$, which is equivalent to the average reward. The policy gradient is defined as the gradient of that objective function with respect to the vector parametrizing the policy, here w. If we apply this definition of the policy gradient to the hierarchically higher level which is our policy over options μ, we would have as a policy gradient $\nabla_\theta \; v_{\mu_\theta}(s_0)$.

a)

We first redefine the state value function of μ in the same form as for a single policy π, in order to "abstract" its hierarchic level, and treat the individual policies as if they were actions:

$$v(s;\theta,w) = \sum_o \mu(o|s;\theta) q_\mu(o,s|w,\theta) \tag{2}$$

Here the quantity $v(s:\theta,w)$, the state value funciton for the whole model, indicates the parametrization by w, which here is not just a parametric vector but rather a matrix, since there is a different vector for each policy. However, μ itself, is only parametrized by the vector θ in the sense we don't need to know the individual vectors w to output from μ the probability of each policy, if we know θ (and of course the state-policy vectors, which are analogous to the state-action vectors).

We define the "state-policy" value function $q_\mu(o,s|w,\theta)$ as:

$$q_\mu(o,s|w,\theta) = \sum_a \pi(a|s,o;w)\left(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s';\theta,w)\right) \tag{3}$$

We can now proceed with the derivation of the policy gradient:

$$\frac{\partial}{\partial\theta_i}v(s;\theta,w) = \sum_o \left(\mu(o|s;\theta)\frac{\partial}{\partial\theta_i}q_\mu(o,s|w,\theta) + q_\mu(o,s|w,\theta)\frac{\partial}{\partial\theta_i}\mu(o|s;\theta)\right) \tag{4}$$

Let's isolate terms where the derived function is only parametrized by θ by defining:

$$m_\theta^{(i)} = \sum_o q_\mu(o,s|w,\theta)\frac{\partial}{\partial\theta_i}\mu(o|s;\theta) \tag{5}$$

Let's write the derivative of the state-policy value function:

$$\frac{\partial}{\partial\theta_i}q_\mu(o,s|w,\theta) = \frac{\partial}{\partial\theta_i}\sum_a \pi(a|s,o;w)\left(r(s,a) + \gamma\sum_{s'} P(s'|s,a)v(s';\theta,w)\right)$$

$$= \gamma\sum_{s',a}\pi(a|s,o;w)P(s'|s,a)\frac{\partial}{\partial\theta_i}v(s';\theta,w)$$

We thus have

$$\frac{\partial}{\partial\theta_i}v(s;\theta,w) = m_\theta^{(i)} + \sum_o \mu(o|s;\theta)\frac{\partial}{\partial\theta_i}q_\mu(o,s|w,\theta)$$

$$= m_\theta^{(i)} + \sum_o \mu(o|s;\theta)\gamma\sum_{s',a}\pi(a|s,o;w)P(s'|s,a)\frac{\partial}{\partial\theta_i}v(s';\theta,w)$$

Using $P(s'|s,a)$, the state transitions probability matrix from the environment, we can define the state transition probability matrix that incorporates P along with the policy over option μ and its child policies π_o (probability of transitioning from state s to state s' under policy π obtained by following μ, in our given environment):

$$P_{env,model}(s',s) = \sum_{o,a}\mu(o|s;\theta)\pi(a|s,o;w)P(s'|s,a)$$

From here we will redefine the subscript "$env, model$" as "e,m" for clarity. We can now redefine the gradient as such:

$$\frac{\partial}{\partial\theta_i}v(s;\theta,w) = g_\theta^{(i)}(s) = m_\theta^{(i)} + \gamma\sum_{s'} P_{e,m}(s',s)g_\theta^{(i)}(s')$$

If we write this equation in matrix form we get:

$$g_\theta^{(i)} = m_\theta^{(i)} + P_{e,m}g_\theta^{(i)}$$

$$\iff$$

$$g_\theta^{(i)} = m_\theta^{(i)}(I - \gamma P_{e,m})^{-1}$$

As seen in class, we can define

$$d_{e,m}(s|s_0) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0)$$

And note that

$$\sum_s d_{e,m}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_s P(s_t = s|s_0)$$
$$= \frac{1}{1-\gamma}$$

knowing that the sum of $P(s_t = s|s_0)$ over all states s gives 1, since it is a probability distribution. Since

$$g_\theta^{(i)}(s_0) = \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0) m_\theta^{(i)}$$

we can write the policy gradient of μ as:

$$g_\theta^{(i)}(s_0) = \frac{\partial}{\partial \theta_i} v(s_0; \theta, w) = \frac{1}{1-\gamma} \sum_s d_{e,m}(s|s_0)\ m_\theta^{(i)}$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{d_{e,m}(s|s_0)} \left[m_\theta^{(i)} \right]$$

b)

Here we need to derive the policy gradient for a given policy π when μ has selected that policy with index o. We thus need to differentiate the objective function of our model with respect to a given vector w (there is a different vector w for each policy, but here for clarity we won't note w_o and simply w). Since μ is not parametrized by w, but only its state-policy value function q_μ is, we can start deriving the policy gradient with respect to w as such:

$$\frac{\partial}{\partial w_i} v(s; \theta, w) = \sum_o \mu(o|s; \theta) \frac{\partial}{\partial w_i} q_\mu(o, s|w, \theta)$$

We expand the derivative of q_μ:

$$\frac{\partial}{\partial w_i} q_\mu(o, s|w, \theta) = \sum_a \left(\pi(a|s, o; w) \frac{\partial}{\partial w_i} q_\pi(a, s|w, \theta) + q_\pi(a, s|w, \theta) \frac{\partial}{\partial w_i} \pi(a|s, o; w) \right)$$

Note here that q_π is also parametrized by θ, because it depends on the state value function of the model, which depends on θ, as we can see here:

$$q_\pi(a, s|w, \theta) = \gamma \sum_{s'} P(s'|s, a) v(s'; \theta, w)$$
$$\Longrightarrow$$
$$\frac{\partial}{\partial w_i} q_\pi(a, s|w, \theta) = \gamma \sum_{s'} P(s'|s, a) \frac{\partial}{\partial w_i} v(s'; \theta, w)$$

Here again, let's isolate the terms from the gradient where we have a derivative of a function only parametrized by the quantity of interest w, namely the policy π:

$$n_w^{(i)}(s) = \sum_{o,a} \mu(o|s;\theta) q_\pi(a,s|w,\theta) \frac{\partial}{\partial w_i} \pi(a|s,o;w)$$

We can now simplify the expression for the gradient:

$$\frac{\partial}{\partial w_i} v(s;\theta,w) = g_w^{(i)}(s) = n_w^{(i)}(s) + \sum_{o,a} \mu(o|s;\theta)\pi(a|s,o;w)\gamma \sum_{s'} P(s'|s,a) \frac{\partial}{\partial w_i} v(s';\theta,w)$$

We can define the state transition matrix $P_{e,m}(s',s)$ in the same fashion as in the last subquestion, and obtain:

$$g_w^{(i)}(s) = n_w^{(i)}(s) + \gamma \sum_{s'} P_{e,m}(s',s) \frac{\partial}{\partial w_i} g_w^{(i)}(s')$$

Again just like in the last subquestion, we can work with the matrix form of this expression and obtain:

$$g_\theta^{(i)} = n_w^{(i)} (I - \gamma P_{e,m})^{-1}$$

From the previous subquestion we know that

$$g_w^{(i)}(s_0) = \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0) n_w^{(i)}$$

and recalling that

$$d_{e,m}(s|s_0) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0)$$

We obtain the policy gradient for π using the same previous observations:

$$g_w^{(i)}(s_0) = \frac{\partial}{\partial w_i} v(s_0;\theta,w) = \frac{1}{1-\gamma} \sum_s d_{e,m}(s|s_0)\ n_w^{(i)}$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{d_{e,m}(s|s_0)} \left[n_\theta^{(i)} \right]$$

2. Policy Gradient Hessian

For the policy hessian, we first need to derive the first order derivatives of the state value function $v_{theta}(s)$ with respect to θ, which we obtain during the derivation of the proof of the Policy Gradient Theorem:

$$\frac{\partial}{\partial \theta_i} v_\theta(s) = \sum_a \left(q_\pi(a,s|\theta) \frac{\partial}{\partial \theta_i} \pi(a|s,\theta) + \pi(a|s,\theta) \frac{\partial}{\partial \theta_i} q_\pi(a,s|\theta) \right)$$

Let's write

$$\frac{\partial}{\partial \theta_j} \sum_a q_\pi(a,s|\theta) \frac{\partial}{\partial \theta_i} \pi(a|s,\theta)$$

$$= \sum_a \frac{\partial}{\partial \theta_j} q_\pi(a, s|\theta) \frac{\partial}{\partial \theta_i} \pi(a|s, \theta) + q_\pi(a, s|\theta) \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} \pi(a|s, \theta)$$

And define

$$A_{i,j} = \sum_a \frac{\partial}{\partial \theta_j} q_\pi(a, s|\theta) \frac{\partial}{\partial \theta_i} \pi(a|s, \theta)$$

$$B_{i,j} = \sum_a q_\pi(a, s|\theta) \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} \pi(a|s, \theta)$$

Similarly, let's write

$$\frac{\partial}{\partial \theta_j} \sum_a \pi(a|s, \theta) \frac{\partial}{\partial \theta_i} q_\pi(a, s|\theta)$$

$$= \sum_a \frac{\partial}{\partial \theta_j} \pi(a|s, \theta) \frac{\partial}{\partial \theta_i} q_\pi(a, s|\theta) + \pi(a|s, \theta) \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} q_\pi(a, s|\theta)$$

and then define

$$C_{i,j} = \sum_a \frac{\partial}{\partial \theta_j} \pi(a|s, \theta) \frac{\partial}{\partial \theta_i} q_\pi(a, s|\theta)$$

$$D_{i,j} = \sum_a \pi(a|s, \theta) \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} q_\pi(a, s|\theta)$$

and use those definitions to write the policy hessian more compactly. Since $C_{i,j}$ is equivalent to $A_{j,i}$, we can write:

$$\frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} v_\theta(s) = A_{i,j} + A_{j,i} + B_{i,j} + D_{i,j}$$

Evaluating $\frac{\partial}{\partial \theta_j} \pi(a|s, \theta)$ is straightforward, as well as evaluating the second order partial derivative, since the policy is defined as

$$\pi(a|s, \theta) = \frac{e^{\theta^T x(s,a)}}{\sum_b e^{\theta^T x(s,b)}}$$

For example, for the first order derivative, we would get

$$\frac{\partial}{\partial \theta_i} \pi(a|s, \theta) = \frac{u'v}{v^2} - \frac{uv'}{v^2}$$

where

$$u = e^{\theta^T x(s,a)},\ u' = x^{(i)}(s, a) e^{\theta^T x(s,a)},\ v = \sum_b e^{\theta^T x(s,b)},\ v' = \sum_b x^{(i)}(s, b) e^{\theta^T x(s,b)}$$

Evaluating the first order derivative of q_π:

$$q_\theta(s, a|\theta) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) v_\theta(s')$$

$$\frac{\partial}{\partial\theta_i}q_\pi(a,s|\theta) = \sum_{s'} P(s'|s,a)\frac{\partial}{\partial\theta_i}v_\theta(s)$$

with

$$\frac{\partial}{\partial\theta_i}v_\theta(s) = \sum_s d_{m,e}(s)\sum_a \frac{\partial\pi_\theta(a|s;\theta}{\partial\theta_i}q_\pi(s,a|\theta)$$

obtained during the derivation of the policy gradient (using the trick in the the proof seen in class, which has already been applied for previous questions).

Each definition among $A_{i,j}$, $B_{i,j}$ and $D_{i,j}$ could be developed, but of all it is $D_{i,j}$ that allows us to define an equality for the second order derivative of $v_\theta(s)$, since it also contains a second order derivative with respect to $v_\theta(s)$. We thus go on computing this quantity:

$$q_\theta(s,a|\theta) = r(s,a) + \gamma\sum_{s'} P(s'|s,a)v_\theta(s')$$

$$\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}q_\theta(s,a|\theta) = \gamma\sum_{s'} P(s'|s,a)\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}v_\theta(s')$$

Again we can define the state transition probability matrix that incorporates P along with the policy π:

$$P_{e,m}(s',s) = \sum_a \pi(a|s;\theta)P(s'|s,a)$$

We can apply the same logic as in the last question

$$\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}v_\theta(s) = g_\theta^{(i,j)}(s) = A_{i,j} + A_{j,i} + B_{i,j} + \gamma\sum_{s'} P_{e,m}(s',s)g_\theta^{(i,j)}(s')$$

using the matrix form we obtain

$$g_\theta^{(i,j)} = (A_{i,j} + A_{j,i} + B_{i,j})\cdot(I - \gamma P_{e,m})^{-1}$$

recalling once again the definition

$$d_{e,m}(s|s_0) = \sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0)$$

And since

$$g_\theta^{(i,j)} = \sum_s\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0)(A_{i,j} + A_{j,i} + B_{i,j})$$

we end up, similarly as in previous questions, with the form

$$g_\theta^{(i,j)} = \frac{1}{1-\gamma}\sum_s d_{e,m}(s|s_0)\ (A_{i,j} + A_{j,i} + B_{i,j})$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{d_{e,m}(s|s_0)}\left[A_{i,j} + A_{j,i} + B_{i,j}\right]$$

3. Constrained Optimization/Intrinsic Rewards

We have an objective function defining Lagrangian for a relaxation of a constrained optimization problem, where we want to maximize the expected discounted return but also minimize some cost function:

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

Since we have our objective function defined here, to apply policy gradient methods we need to derive

$$\nabla_\theta J_\alpha(\theta)$$

One way could be to use the Lagrangian relaxation method to the constrained optimization problem:

$$max_\theta \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)\right]$$
$$s.t.$$
$$\eta\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)\right] \leq 0$$

Since we will inevitably have violations of the constraint inequality in this constrained optimization problem, we could use the Lagrangian relaxation method to solve it. Indeed as the cost function is understood, it can only produce a cost value greater or equal to zero, for a given state-action pair. The Lagrangian relaxation method will penalize violations of the inequality. These added penalties are used in the optimization problem, instead of the strict inequality constraint.

One other way that is pretty straightforward is simply to incorporate the intrinsic reward and cost functions into one single quantity:

$$z(S_t, A_t) = r(S_t, A_t) - \eta\ c(S_t, A_t)$$

and redefine the objective as

$$J_\alpha(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t z(S_t, A_t)\right]$$