COMP 767 Winter 2018
Raihan Seraj
Student ID: 260752605
Gandharv Patil
Student ID: 260727335

1 Part 1

1. **Policy Gradient for a Mixture of Policies**

The Bellman equation of the model is given by

$$v(s;\theta,w) = \sum_{0} \mu(o|s;\theta) \sum_{a} \pi(a|s,o;w) \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) v(s';\theta,w) \right)$$

(a) The policy gradient for μ which is parameterized by θ is calculated by differentiating the expression for the value function with respect to θ.

$$\frac{\partial}{\partial\theta}v(s;\theta,w) = \frac{\partial}{\partial\theta}\sum_{o}\mu(o|s;\theta)\sum_{a}\pi(a|s,o;w)Q^{\pi\mu}(s,a;w,\theta) \tag{1}$$

$$\because r(s,a)+\gamma\sum_{s'}P(s'|s,a)v(s';\theta,w) = Q^{\pi\mu}(s,a;w,\theta) \tag{2}$$

$$= \sum_{a}\ \pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta)\frac{\partial}{\partial\theta}\sum_{o}\mu(o|s;\theta) \tag{3}$$

$$+\sum_{0}\mu(o|s;\theta)\sum_{a}\pi(a|s,o;w)\frac{\partial}{\partial\theta}(r(s,a)+\sum_{s'}\gamma P_{ss'}v(s';\theta,w))$$

$$= \sum_{o}\frac{\partial}{\partial\theta}\mu(o|s;\theta)\sum_{a}\ \pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta) \tag{4}$$

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

Unrolling the recursive value function as in Sutton et al. we get (5)

$$= \sum_{s}d^{\pi}(s)\sum_{o}\frac{\partial}{\partial\theta}\mu(o|s;\theta)\sum_{a}\ \pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta)$$

(b) The policy gradient of π is obtained by finding the derivative of the value function with respect to the parameters of the policy π. Therefore we can write the following

$$\frac{\partial}{\partial w}v(s;\theta,w) = \frac{\partial}{\partial w}\sum_{o}\mu(o|s;\theta)\sum_{a}\pi(a|s,o;w)Q^{\pi\mu}(s,a;w,\theta) \tag{6}$$

$$\because r(s,a)+\gamma\sum_{s'}P(s'|s,a)v(s';\theta,w) = Q^{\pi\mu}(s,a;w,\theta) \tag{7}$$

$$= \sum_{o}\mu(o|s;\theta)\sum_{a}\ \frac{\partial}{\partial w}\pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta) \tag{8}$$

$$+\sum_{o}\mu(o|s;\theta)\sum_{a}\pi(a|s,o;w)\frac{\partial}{\partial w}(r(s,a)+\sum_{s'}\gamma P_{ss'}v(s';\theta,w))$$

$$= \sum_{o}\mu(o|s;\theta)\sum_{a}\ \frac{\partial}{\partial w}\pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta) \tag{9}$$

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

$$= \sum_{o}\mu(o|s;\theta)\Bigg(\sum_{a}\ \frac{\partial}{\partial w}\pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta) \tag{10}$$

$$+\pi(a|s,o;w)\sum_{s'}\gamma P_{a}^{ss'}\frac{\partial}{\partial w}v(s';\theta,w)\Bigg)$$

The term inside the parathesis is the same as that in (11)

Policy Gradient Methods for Reinforcement Learning by Sutton et al.

$\therefore$ Unrolling the recursive value function as in Sutton et al. we get

$$= \sum_{o}\mu(o|s;\theta)\sum_{s}d^{\pi}(s)\sum_{a}\ \frac{\partial}{\partial w}\pi(a|s,o;w)Q^{\mu\pi}(s,a;w,\theta)$$

2 Part 2

Policy Hessian Theorem:
Here we start with the standard procedure of the first order derivation and then compute

the gradient of the result thus obtained.

$$\frac{\partial V_\pi(s)}{\partial\theta} := \frac{\partial}{\partial\theta}\sum_a \pi(s,a)Q^\pi(s,a) \tag{12}$$

$$\sum_a[\frac{\partial\pi(s,a)}{\partial\theta}Q^\pi(s,a) + \pi(s,a)\frac{\partial}{\partial\theta}Q^\pi(s,a)] \tag{13}$$

$$\text{We now take the gradient of (13) w.r.t } \theta \tag{14}$$

$$\therefore \frac{\partial^2 V^\pi(s)}{\partial\theta^2} = \sum_a \Bigg\{ [\underbrace{{}^\pi(s,a)\frac{\partial^2\pi(s,a)}{\partial\ \theta^2}}_{\text{A}} + \underbrace{\frac{\partial\pi(s,a)}{\partial\theta}\frac{\partial Q^\pi(s,a)}{\partial\theta}}_{\text{B}}] \tag{15}$$

$$+ \underbrace{\frac{\partial}{\partial\theta}Q^\pi(s,a)\frac{\partial\pi(s,a)}{\partial\theta}}_{\text{C}} + \underbrace{\pi(s,a)\frac{\partial^2}{\partial\theta^2}Q^\pi(s,a)}_{\text{D}} \Bigg\} \tag{16}$$

$$\text{We can see that the term B and C are equivalent} \tag{17}$$

$$\text{Hence the entire Hessian can be written as the following} \tag{18}$$

$$\frac{\partial^2 v(s)}{\partial\theta^2} = A + B + B + D \tag{19}$$

$$\frac{\partial^2 Q(s,a)}{\partial\theta^2} = \gamma\sum_{s'} P_a^{ss'}\frac{\partial^2 v(s')}{\partial\theta^2} \tag{20}$$

$$\text{We also define the following} \tag{21}$$

$$P_{a,b}(s'|s,a) = \sum_a(\pi(s,a)P(s'|s,a)) \tag{22}$$

$$\frac{\partial^v(s)}{\partial\theta^2} = k_\theta = A + B + B + \gamma\sum_{s'} P_{a,b}(s,s')k_\theta \tag{23}$$

$$\text{Hence the expression of } k_\theta \text{ is obtained as follows} \tag{24}$$

$$k_\theta = A + B + B + (I - \gamma P_{a,b})^{-1} \tag{25}$$

$$\tag{26}$$

Since $d^\pi = \sum_{t=0}^\infty \gamma^t P(s_t = s|s_0)$ is the weighted occupancy measure we therefore try to find a similar expression for policy gradient hessian in terms of d^π and obtain the followin

$$\frac{1}{1-\gamma}\mathbb{E}_{d^\pi}(A + B + B) \tag{27}$$

3 Part 3

Consider the following objective, under the usual discounted MDP formulation:

$$J_\alpha(\theta) = \mathop{\mathbb{E}}_{\alpha,\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)\right] - \eta \mathop{\mathbb{E}}_{\alpha,\theta}\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)\right] \tag{28}$$

where:
α = initial distribution over states.
θ = parameters of the stochastic policy π_θ
Our objective here is to find the θ to maximize the expected discounted return but also have to pay a cost C.
Solution The objective funtion given to us can be re-written as:

$$J_\alpha(\theta) = \mathop{\mathbb{E}}_{\alpha,\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(S_t, A_t) - \eta c(S_t, A_t)\right)\right] \tag{29}$$

- The above objective can be considered as a form of reward shaping, where an additional cost is incurred along with a obtained reward. The key difference between reward shaping and above modification is that, the optimal policy is invarient to reward shaping when the shaped reward is a potential function of the state as explained in Ng et al. in the paper *" Policy invariance under reward transformations: Theory and application to reward shaping "*

- The current modfication however does not qualify to be a potential function of the state and hence the policy learnt as a result of this function will be senstive to the cost(C) and η.

- We now define the state-action(Q)value-function in terms of the new reward:

$$Q^\pi(s,a) = r(s,a) + \eta c(s,a) + \gamma \sum_{s'} P_a^{ss'} V(s') \tag{30}$$

- Finally this new state-action(Q)value-function can be substituted in the result of the policy gradient theorem, which is given by:

$$\nabla_\theta \rho = \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(s,a) Q^\pi(s,a) \tag{31}$$