Naser Kazemi 99102059 HW3

Q1

A) There are two possible sources for the randomness in the cumulative reward:

1) The policy randomness. If the policy itself is stochastic, the we have to compute the expected value.
2) The environment dynamics. The transition between states after taking an action can be probablistic.

B) One way could be reparametrizing the actions. i,e, let a be an action sampled from the policy $\pi_\theta(a|s)$. we can reparametrize a as $a = f_\theta(s, \varepsilon)$ where $\varepsilon \sim p(\varepsilon)$. Let's assume p is a gaussian distribution, with mean $\mu_\theta(s)$ and variance $\sigma^2_\theta(s)$, then

$$a = f_\theta(s, \varepsilon) = \mu_\theta(s) + \sigma_\theta(s)\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$

Now we can see the dependency of rewards and θ, and we have

$$J_\theta = E\left[\sum_t \gamma^t r(s_t, f_\theta(s_t, \varepsilon))\right]$$

which is the expectation of a function of θ, and hence we can apply gradient ascent directly on the objective function.

Q2

A) I consider the following assumptions in my proof:

1) Trajectories are independent

2) Rewards and state-actions, i.e $r(s,a)$ are consistent

2) The policy $\pi_\theta(a|s)$ is differentiable.

We have

$$\nabla_\theta J(\theta) = \nabla_\theta E_{\tau\sim P(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$$

$$= \nabla_\theta \int P_\theta(\tau) R(\tau)\, d\tau \qquad R(\tau) = \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)$$

$$= \int \nabla_\theta P_\theta(\tau) R(\tau)\, d\tau \qquad \tau = (s_0, a_0, s_1, a_1, \dots)$$

$$= \int \frac{\nabla_\theta P_\theta(\tau)}{P_\theta(\tau)} P_\theta(\tau) R(\tau)\, d\tau$$

$$= E_{\tau\sim P_\theta(\tau)}\left[\nabla \log P_\theta(\tau) R(\tau)\right]$$

$$= E_{\tau\sim P_\theta(\tau)}\left[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(s_t,a_t)\right]$$

$$P_\theta(\tau) = \prod_{t=0}^{\infty}\pi_\theta(a_t|s_t)\, P(s_{t+1}|a_t,s_t) \Rightarrow \nabla_\theta \log P_\theta(\tau) = \sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)$$

Now on the other hand:

$$E_{\tau\sim P_\theta(\tau)}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_{i,\tau}|s_{i,t})\right)\left(\sum_{\tau=0}^{T}\gamma^t r(s_{i,t},a_{i,t})\right)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_{\tau\sim P_\theta(\tau)}\left[\left(\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_{i,\tau}|s_{i,t})\right)\left(\sum_{t=0}^{T}\gamma^t r(s_{i,t},a_{i,t})\right)\right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \bar{J}(\theta)$$

$$= \nabla_\theta J(\theta)$$

In this proof, we also assume that the maximum lenght of each trajectory is T.

B) First let's prove a lemma:

EGLP lemma: Suppose P_θ is a parmetrized distribution over a random variable, x. Then

$$E_{x \sim P_\theta}[\nabla_\theta \log P_\theta(x)] = 0$$

Proof: We know that

$$\int_x P_\theta(x)\, dx = 1$$

and

$$0 = \nabla_\theta 1 = \nabla_\theta \int_x P_\theta(x)\, dx$$

$$= \int_x \nabla_\theta P_\theta(x)\, dx$$

$$= \int_x \frac{\nabla P_\theta(x)}{P_\theta(x)} P_\theta(x)\, dx$$

$$= E_{x \sim P_\theta}[\nabla \log P_\theta(x)]$$

Now let's get back to the main problem.

$$E_\tau\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right)\right]$$

$$=\sum_{t=0}^{\infty} E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=0}^{\infty}\gamma^{t'} r(s_{t'},a_{t'})\right)\right]$$

We have:

$$E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=0}^{\infty}\gamma^{t'} r(s_{t'},a_{t'})\right)\right]$$

$$= E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=0}^{t-1}\gamma^{t'} r(s_{t'},a_{t'})\right]$$

$$+ E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=t}^{\infty}\gamma^{t'} r(s_{t'},a_{t'})\right]$$

the random variables $\sum_{t'=0}^{t-1}\gamma^{t'} r(s_{t'},a_{t'})$ and $\nabla \log \pi_\theta(a_t|s_t)$ are independent because of the Markove property, therefore we can write

$$E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=0}^{t-1}\gamma^{t'} r(s_{t'},a_{t'})\right]$$

$$= E_{s_0,a_0,\dots,s_{t-1},a_{t-1}}\left[\sum_{t'=0}^{t-1}\gamma^{t'} r(s_{t'},a_{t'})\right] E_{s_t,a_t,\dots}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$= 0$$

since by EGLP, $E_{s_t,a_t,\dots}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right] = 0$.

Thus

$$E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{\infty}\gamma^{t'} r(s_{t'},a_{t'})\right)\right]$$

$$= \gamma^t E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=t}^{\infty}\gamma^{t'-t} r(s_{t'},a_{t'})\right]$$

and

$$E_\tau\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t E_\tau\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=t}^{\infty}\gamma^{t'-t} r(s_{t'},a_{t'})\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t \int_\tau \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \prod_{t'=0}^{\infty} \pi_\theta(a_{t'}|s_{t'})\, p(s_{t'+1}|s_t,a_t) \left(\sum_{t'=t}^{\infty}\gamma^{t'-t} r(s_{t'},a_{t'})\right) d\tau$$

$$= \sum_{t=0}^{\infty}\gamma^t \int_\tau \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \prod_{t'=0}^{\infty} \pi_\theta(a_{t'}|s_{t'})\, p(s_{t'+1}|s_{t'},a_{t'})\, Q(s_t,a_t)\, d\tau$$

$$= \sum_{t=0}^{\infty}\gamma^t \int_{s,a} \nabla_\theta \log \pi_\theta(a_t=a|s_t=s)\, P(s_t=s)\, \pi_\theta(a_t=a|s_t=s)\, Q(s_t=s,a_t=a)\, ds\, da$$

$$= \sum_{t=0}^{\infty}\gamma^t \int_s P(s_t=s)\, ds \int_a \nabla_\theta \log \pi_\theta(a_t=a|s_t=s)\, \pi_\theta(a_t=a|s_t=s)\, Q(s_t=s,a_t=a)\, da$$

$$= \int_s \sum_{t=0}^{\infty}\gamma^t P(s_t=s)\, ds \int_a \nabla_\theta \log \pi_\theta(a_t=a|s_t=s)\, \pi_\theta(a_t=a|s_t=s)\, Q(s_t=s,a_t=a)\, da$$

$$= \int_s P_\pi(s)\, E_\pi\left[\nabla_\theta \log \pi_\theta(a|s)\, Q(s,a)\right]$$

$$= E_{s\sim\rho_\pi,\, a\sim\pi}\left[\nabla_\theta \log \pi_\theta(a|s)\, Q(s,a)\right]$$

c) Let $\nabla_\theta J(\theta) = E_{s\sim\rho_\pi,\, a\sim\pi}[g]$ where

$$g := \nabla_\theta \log \pi_\theta(a|s)\,(Q(s,a) - b(s))$$

$$
\begin{aligned}
Var(g) &= E_{\rho_\pi,\pi}[(g - E_{\rho_\pi,\pi}[g])^T (g - E_{\rho_\pi,\pi}[g])] \\
&= E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s)) b(s)^2 \\
&\quad - 2 E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s) \hat{Q}(s,a)] b(s) \\
&\quad + E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s) \hat{Q}(s,a)^2] \\
&\quad - E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)(\hat{Q}(s,a) - b(s))]^T \\
&\quad E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)(\hat{Q}(s,a) - b(s))]
\end{aligned}
$$

from one hand:

$$
\begin{aligned}
E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s) b(s)] &= E_{\rho_\pi}[E_{a\sim\pi}[\nabla_\theta \log\pi(a|s) b(s) | s]] \\
&= E_{\rho_\pi}\left[\int_a \nabla_\theta \log\pi(a|s) b(s) \, da \mid s\right] \\
&= E_{\rho_\pi}[b(s) \nabla_\theta \int_a \log\pi(a|s) \, da \mid s] \\
&= E_{\rho_\pi}[b(s) \nabla_\theta 1 \mid s] \\
&= E_{\rho_\pi}[0 | s] \\
&= 0
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
Var(g) &= E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s)) b(s)^2 \\
&\quad - 2 E_{\rho_\pi,\pi}[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s) \hat{Q}(s,a)] b(s)
\end{aligned}
$$

$$
\Rightarrow \frac{\partial}{\partial b(s)} Var[g] = 2 E_\pi[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s)) b^*(s)
$$
$$
- 2 E_\pi[\nabla_\theta \log\pi(a|s)^T \nabla_\theta \log\pi(a|s) \hat{Q}(s,a)] = 0
$$

$$\Longrightarrow b^*(s) = \frac{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s) \hat{Q}(s,a)]}{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s))}$$

D) Let's assume $\nabla_\theta \log \pi(a|s)$ and $Q(s,a)$ are independent, therefore:

$$b^*(s) = \frac{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s) Q(s,a)]}{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s))}$$

$$= \frac{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s)] E_{\pi}[Q(s,a)]}{E_\pi[\nabla_\theta \log \pi(a|s)^T \nabla_\theta \log \pi(a|s))}$$

$$= E_\pi[Q(s,a)]$$

$$= V^\pi(s)$$

Q3

A) Theorem: Let μ and ν be two probability distribution functions on $\mathcal{X}$. Then if $\forall x \in \mathcal{X}$, $|\mu(x) - \nu(x)| = \varepsilon$, there is a coupling $P(X, Y)$, such that $P(X \neq Y) = \varepsilon$.

Proof: Let

$$p = \sum_{x \in \mathcal{X}} \mu(x) \wedge \nu(x) = \sum_{\substack{x \in \mathcal{X} \\ \mu(x) \leq \nu(x)}} \mu(x) + \sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} \nu(x)$$

$$= \sum_{\substack{x \in \mathcal{X} \\ \mu(x) \leq \nu(x)}} \mu(x) + \sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} \nu(x) + \sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} \mu(x) - \sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} \mu(x)$$

$$= 1 - \left(\sum_{\substack{x \in \mathcal{X} \\ \mu(x) > \nu(x)}} \mu(x) - \nu(x) \right)$$

$$= 1 - |\mu(X) - \nu(x)|$$

because let $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A \subset \mathcal{X}$ be any event. then

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B)$$

Since for any $x \in A \cap B^c$, $\mu(x) - \nu(x)$, so the difference of probability without these terms is greater, and hence the first inequality. By the same argument, the second inequality can be prove. and in the same way we have

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$$

therefore $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$.

Now consider a coin with the probability of heads equal to ρ. Flip the coin

(1) If it comes up heads, then choose a random value z by the probability distribution:

$$\gamma_1(z) = \frac{\mu(z) \wedge \nu(z)}{\rho}$$

and set $X = Y = z$.

(2) else, choose x according the probability distribution

$$\gamma_2(x) = \begin{cases} \frac{\mu(x) - \nu(x)}{1-\rho} & , \text{ if } \mu(x) > \nu(x) \\ 0 & , \text{otherwise} \end{cases}$$

and set $X = x$. In a same way choose y according the probability distribution

$$\gamma_3(y) = \begin{cases} \frac{\nu(y) - \mu(y)}{1-\rho} & , \text{ if } \nu(x) > \mu(x) \\ 0 & , \text{otherwise} \end{cases}$$

and set $Y = y$. Obviously in this coupling, we have:

$$P\{X = Y\} = \rho = 1 - \varepsilon, \quad P\{X \neq Y\} = 1 - \rho = \varepsilon$$

and

$$\rho\gamma_1 + (1-\rho)\gamma_2 = \mu$$

$$\rho\gamma_1 + (1-\rho)\gamma_3 = \nu$$

thus the marginals of $P(X, Y)$, match μ and ν.

Now consider the above coupling for distributions $\pi_\theta(a_t|s_t)$ and $\pi_{\theta'}(a_t|s_t)$, where $|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)| \le \varepsilon$. By this coupling, the probability of taking different actions in each state s_t by π_θ and $\pi_{\theta'}$ would be at most ε. Thus for a state s_t,

$$P_{\theta'}(s_t) = P_{nm}^t P_\theta(s_t) + (1 - P_{nm}^t) P_m(s_t)$$

$P_{nm} = 1 - \varepsilon$ (probability of making no mistakes)

$P_m =$ probability of some distribution other than P_θ.

So

$$|P_{\theta'}(s_t) - P_\theta(s_t)| = (1 - (1-\varepsilon)^t) |P_m(s_t) - P_\theta(s_t)|$$

By the Bernoulli identity, $(1-\varepsilon)^t \ge 1 - \varepsilon t$. Thus

$$|P_{\theta'}(s_t) - P_\theta(s_t)| \le \varepsilon t |P_m(s_t) - P_\theta(s_t)|$$

Also for any event $A \in \mathcal{X}$, we have

$$\begin{aligned} P_m(A) - P_\theta(A) &= P\{X \in A\} - P\{Y \in A\} \\ &\le P\{X \in A, Y \notin A\} \\ &\le P(X \ne Y) \\ &\le 1 \end{aligned}$$

$$\Rightarrow |P_m(s_t) - P_\theta(s_t)| = 2 |P_m(s_t) - P_\theta(s_t)|_{TV} \le 2$$

and

$$|P_{\theta'}(s_t) - P_{\theta}(s_t)| \leq 2\varepsilon t$$

B) Now assume $|\pi_{\theta'}(a_t|s_t) - \pi_{\theta}(a_t|s_t)| \leq \varepsilon$ for all s_t. therefore $|P_{\theta'}(s_t) - P_{\theta}(s_t)| \leq 2\varepsilon t$ for all s_t. Now consider an arbitrary function of s_t, such as $f(s_t)$. We have:

$$\begin{aligned}
E_{p_{\theta'}}[f(s_t)] &= \sum_{s_t} p_{\theta'}(s_t) f(s_t) \\
&= \sum_{s_t} (p_{\theta'}(s_t) + p_{\theta}(s_t) - p_{\theta}(s_t)) f(s_t) \\
&= \sum_{s_t} p_{\theta}(s_t) f(s_t) - \sum_{s_t} (p_{\theta}(s_t) - p_{\theta'}(s_t)) f(s_t) \\
&\geq \sum_{s_t} p_{\theta}(s_t) f(s_t) - |p_{\theta}(s_t) - p_{\theta'}(s_t)| \max_{s_t} f(s_t) \\
&\geq E_{p_{\theta}}[f(s_t)] - 2\varepsilon t \max_{s_t} f(s_t)
\end{aligned}$$

By setting $f(s_t) = E_{a_t \sim \pi_{\theta}(a_t|s_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t)\right]$ for limite horizon we obtain:

$$\begin{aligned}
&\sum_{t=0}^{T} E_{p_{\theta'}(s_t)}\left[E_{a_t \sim \pi_{\theta}(a_t|s_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t)\right]\right] \\
&\geq \sum_{t=0}^{T} E_{p_{\theta}(s_t)}\left[E_{a_t \sim \pi_{\theta}(a_t|s_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t)\right]\right] - \sum_{t=0}^{T} 2\varepsilon t C \\
&\geq \sum_{t=0}^{T} E_{p_{\theta}(s_t)}\left[E_{a_t \sim \pi_{\theta}(a_t|s_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t)\right]\right] - \sum_{t=0}^{T} 2\varepsilon t C
\end{aligned}$$

which $\sum_{t=0}^{T} 2\varepsilon t C = 2\varepsilon \frac{T(T+1)}{2} \cdot C$ where $C \le r_{max}$ so

$$\sum_{t=0}^{1} \underbrace{E_{p_\theta, (s_t)}}_{T} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\geq \sum_{t=0}^{T} E_{p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right] - O(T^2 r_{max})$$

And for infinite horizen:

$$\sum_{t}^{\infty} E_{p_\theta, (s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\geq \sum_{t} E_{p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right] - \sum_{t}^{\infty} 2\varepsilon t C(t)$$

$$C(t) = \max_{s_t, a_t} \left| \gamma^t A^{\pi_\theta}(s_t, a_t) \right| \le \gamma^t \max_{s,a} \left| A^{\pi_\theta}(s,a) \right| = \gamma^t C$$

$$\sum_{t}^{\infty} 2\varepsilon t C(t) \le \sum_{t} 2\varepsilon t \gamma^t C \approx \frac{d}{d\gamma} \sum_{t}^{\infty} 2\varepsilon \gamma^{t+1} C$$

$$= 2\varepsilon C \frac{d}{d\gamma}\left(\frac{1}{1-\gamma} - 1 \right)$$

$$= \frac{2\varepsilon C}{(1-\gamma)^2} = O\left(\frac{r_{max}}{(1-\gamma)^2} \right)$$

$$\sum_{t}^{\infty} E_{p_\theta, (s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\geq \sum_{t} E_{p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right] - O\left(\frac{r_{max}}{(1-\gamma)^2} \right)$$

C) By the Pinsker's inequality:

$$|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)|_{tv} \le \sqrt{\frac{1}{2} D_{KL}(\pi_\theta(a_t|s_t) \| \pi_\theta(a_t|s_t))}$$

$$\Rightarrow 2|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)|_{tv}^2 \le D_{KL}(\pi_\theta(a_t|s_t) \| \pi_\theta(a_t|s_t))$$

D) let $\sum_t E_{p_\theta}[E_\pi[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t)]] = \bar{A}(\theta')$. So the taylor's approximation of $\bar{A}(\theta')$ is:

$$\bar{A}(\theta') \approx \bar{A}(\theta) + \nabla_\theta \bar{A}(\theta)^T(\theta' - \theta)$$

$$\nabla_\theta \bar{A}(\theta) = \nabla_\theta \sum_t E_{p_\theta}[E_\pi[\frac{\pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t)]]$$

$$= \sum_t E_{p_\theta}[E_\pi[\gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) A^{\pi_\theta}(s_t, a_t)]]$$

$$= \nabla_\theta J(\theta)$$

therefore:

$$\bar{A}(\theta') \approx \sum_t E_{p_\theta}[E_\pi[\gamma^t A^{\pi_\theta}(s_t, a_t)]] + \nabla_\theta J(\theta)^T(\theta' - \theta)$$

Q4

We know $Q^{\mu_\theta}(s, \mu_\theta(s)) = E_{a\sim\mu_\theta(s)}[Q^{\mu_\theta}(s,a)]$. Thus

$$\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta Q^{\mu_\theta}(s, \mu_\theta(s))$$

$$= \nabla_\theta \left(r(s, \mu_\theta(s)) + \gamma E_{s\sim p(s'|s,\mu_\theta(s))}[V^{\mu_\theta}(s')]\right)$$

$$= \nabla_\theta \left(r(s, \mu_\theta(s)) + \int_S \gamma p(s'|s, \mu_\theta(s)) V^{\mu_\theta}(s') ds'\right)$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$+ \int_S \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') ds' \quad \text{(chain rule)}$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a \left(r(s,a) + \int_S \gamma p(s'|s,a) V^{\mu_\theta}(s') ds'\right)|_{a=\mu_\theta(s)}$$

$$+ \int_S \gamma p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s\to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s\to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$

the S is compact and $\|\nabla_\theta V^{\mu_\theta}(s)\|$ and all the other ∇s are boun

↑

By recursively applying the above relation we will have:

$$\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s\to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$$

$$+ \int_S \gamma p(s\to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'$$

$$+ \int_S \gamma p(s\to s', 1, \mu_\theta) \int_S \gamma p(s'\to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'') ds'' ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^\theta(s,a)|_{a=\mu_\theta(s)}$$
$$+ \int_S \gamma \rho(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'$$
$$+ \int_S \gamma^2 \rho(s \to s', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$
$$\vdots$$
$$= \int_S \sum_{t=0}^{\infty} \gamma^t \rho(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q(s';a)|_{a=\mu(s)} ds'$$

Wow we have

$$\nabla_\theta J(\mu_\theta) = \nabla_\theta E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|\mu\right]$$
$$= \nabla_\theta E[V^{\mu_\theta}(s)]$$
$$= \nabla_\theta \int_S \rho_1(s) V^{\mu_\theta}(s) ds$$
$$= \int_S \rho_1(s) \nabla_\theta V^{\mu_\theta}(s) ds$$
$$= \int_S \int_S \sum_{t=0}^{\infty} \gamma^t \rho_1(s) \rho(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q(s';a)|_{a=\mu(s)} ds' ds$$
$$= \int_S \rho_\mu(s) \nabla_\theta \mu_\theta(s) \nabla_a Q(s,a)|_{a=\mu(s)} ds$$
$$= E_{s\sim\rho_\mu}[\nabla_\theta \mu_\theta(s) \nabla_a Q(s,a)|_{a=\mu(s)}]$$