« Alireza Gargoori Motlagh _ 98102176 »

« Homework 2 _ Reinforcement Learning »

1 Policy Gradient

a) As we will prove in part (c) (also proved in class!), we have:

$$J(\theta) = \mathop{\mathbb{E}}_{\tau \sim P_\theta(\tau)} \left[\sum_t \gamma^t r(s_t, a_t) \right] = \mathop{\mathbb{E}}_{\tau \sim P_\theta(\tau)} \left[\gamma^t r(\tau) \right]$$

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim P_\theta(\tau)} \left[\nabla_\theta \log P_\theta(\tau)\, r(\tau) \right]$$

where

$$P_\theta(\tau_i) = P(s_{i,0}) \prod_{t=1}^{T} \pi(a_{i,t} | s_{i,t})\, P(s_{i,t+1} | a_{i,t}, s_{i,t})$$

$$r(\tau_i) = \sum_{t'=t}^{T} \gamma^t r(s_{i,t'}, a_{i,t'})$$

$$\Longrightarrow \nabla_\theta J_b(\theta) = \mathop{\mathbb{E}}_{\tau \sim P_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(\tau) \left(r(\tau) - b(s) \right) \right] = \dots$$

$$= \mathop{\mathbb{E}}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, r(\tau) \right] - \mathop{\mathbb{E}}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, b(s) \right]$$

$$* \quad \mathop{\mathbb{E}}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, b(s) \right] = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)\, b(s)\, d\tau =$$

$$\int p_\theta(\tau) \frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} b(s)\, d\tau = \int \nabla_\theta p_\theta(\tau)\, b(s)\, d\tau = b(s) \int \nabla_\theta p_\theta(\tau)\, d\tau =$$

$$b(s) \nabla \int p_\theta(\tau)\, d\tau = b \nabla_\theta 1 = 0 \implies$$

$$\nabla_\theta J_b(\theta) = \mathop{\mathbb{E}}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, r(\tau) \right] = \nabla_\theta J(\theta)$$

The above formula shows that $J(\theta)$ and the baselined-version $J_b(\theta)$ when using the true distribution $p_\theta(\tau)$ are the same; therefore their expectation by averaging would be the same and $\nabla_\theta J_b(\theta)$ is unbiased.

Therefore, subtracting a baseline (which only depends on states) is unbiased in expectation.

(b) $$\nabla_\theta J_b(\theta) = \mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\left(r(\tau) - b(s)\right)\right]$$

$$\mathrm{Var}\left(\nabla_\theta J_b(\theta)\right) = \mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\left(r(\tau) - b\right)\right)^2\right] - \left(\mathop{\mathbb{E}}_{\tau\sim p_\theta}\left(\nabla_\theta \log p_\theta(\tau)\left(r(\tau) - b(s)\right)\right)\right)^2$$

$$\overset{(a)}{=} \left(\mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right]\right)^2$$

(not a function of b)

$$\Rightarrow \frac{\partial \mathrm{Var}\left(\nabla_\theta J_b(\theta)\right)}{\partial b} = \frac{\partial}{\partial b} \mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\left(r(\tau) - b\right)\right)^2\right] =$$

$$\frac{\partial}{\partial b}\left(\mathop{\mathbb{E}}_{\tau}\left[\left(\nabla_\theta \log p_\theta\right)^2 r^2(\tau)\right] - 2\mathop{\mathbb{E}}_{\tau}\left[\left(\nabla_\theta \log p_\theta\right)^2 r(\tau) b\right] + \mathop{\mathbb{E}}_{\tau}\left[\left(\nabla_\theta \log p_\theta(\tau)\right)^2 b^2\right]\right)$$

$$= -2\mathop{\mathbb{E}}_{\tau}\left[\left(\nabla_\theta \log p_\theta(\tau)\right)^2 r(\tau)\right] + 2b\mathop{\mathbb{E}}_{\tau}\left[\left(\nabla_\theta \log p_\theta(\tau)\right)^2\right] = 0 \Rightarrow$$

$$\boxed{b^* = \frac{\mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\right)^2 r(\tau)\right]}{\mathop{\mathbb{E}}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\right)^2\right]}}$$

which is the weighted version of expected rewards by the magnitude of their log probability. (note that: $\nabla_\theta \log p_\theta(\tau) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)$)

(c) $$V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \implies$$

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s)\, \nabla_\theta Q^{\pi_\theta}(a,s)$$

$$= \sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s)\, \nabla_\theta \sum_{s'} P(s'|s,a)\,(r + V^{\pi_\theta}(s'))$$

$$= \sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a)\, \nabla_\theta V^{\pi_\theta}(s')$$

iterate $$= \sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \times$$

$$\left(\sum_{a'} \nabla_\theta \pi(a'|s')\, Q^{\pi}(s',a') + \sum_{a'} \pi_\theta(a'|s') \sum_{s''} P(s''|s',a')\, \nabla_\theta V^{\pi}(s'') \right)$$

$$= \overset{\text{iteration}}{\cdots} =$$

$$= \sum_{s'} \sum_{t=0}^{\infty} Pr(s \to s', t, \pi_\theta) \sum_a \nabla_\theta \pi_\theta(a|s')\, Q^{\pi_\theta}(s',a)$$

$Pr(s \to s', t, \pi)$ is the probability of transitioning from s to s' with t steps under policy π_θ. (The same as Problem 3a and SB_RL Book). Also, using : $\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s)\, \nabla_\theta \log \pi_\theta(a|s)$

would be discounted

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_s \sum_{t=0}^{\infty} Pr(s_0 \to s, t, \pi_\theta) \sum_a \pi_\theta(a|s)\, \nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)$$

$$= \sum_s \sum_{t=0} \gamma^t P(s_t = s | s, \pi_\theta) \sum_a \pi_\theta(a|s)\, \nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)$$

Taking expectation on inital states distribution:

$$\Longrightarrow \nabla_\theta V^{\pi}(\mu) = \nabla_\theta \mathop{\mathbb{E}}_{s_0\sim\mu}\left[V^{\pi}(s_0)\right] = \mathop{\mathbb{E}}_{s_0\sim\mu}\left[\nabla_\theta V^{\pi_\theta}(s_0)\right] =$$

$$\sum_{s_0}\mu(s_0)\sum_{s}\sum_{t=0}^{\infty}\gamma^t \Pr(s_t=s\mid s_0,\pi_\theta)\sum_{a}\pi_\theta(a_t\mid s_t)\nabla_\theta \log\pi_\theta(a_t\mid s_t)\, Q^{\pi}(s_t,a_t)$$

$$= \mathop{\mathbb{E}}_{\mu,\pi}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta \log\pi_\theta(a_t\mid s_t)\, Q^{\pi_\theta}(a_t\mid s_t)\right] \xrightarrow{\tau\sim \Pr_\mu^{\pi}}$$

$$\boxed{\nabla_\theta V^{\pi_\theta}(\mu) = \mathop{\mathbb{E}}_{\tau\sim \Pr_\mu^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta \log\pi_\theta(a_t\mid s_t)\, Q^{\pi_\theta}(a_t\mid s_t)\right]}$$

(d) $$\mathop{\mathbb{E}}_{\tau\sim \Pr\pi}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] = \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s_0\sim d_{s_0}^{\pi_\theta}}\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot\mid s)}\left[f(s,a)\right]$$

Substituing $f(s_t,a_t)$ with : $\nabla_\theta \log\pi_\theta(a_t\mid s_t)\, Q^{\pi_\theta}(a_t\mid s_t)$ in the above equation yields :

$$\nabla_\theta V^{\pi}(\mu) = \mathop{\mathbb{E}}_{\tau\sim P_r^{\pi}_{\mu}} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi(a_t|s_t)\, Q^{\pi_\theta}(s_t,a_t) \right] =$$

$$\frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \right] \Longrightarrow$$

$$\boxed{\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \right]}$$

As we proved in part (a), $\nabla_\theta J(\theta) = \nabla_\theta V^{\pi}(\mu)$ is unbiased in expectation when subtracting a baseline which only depends on states, e.g. value of state $V^{\pi_\theta}(s)$! ; therefore:

$$\nabla_\theta V^{\pi}(\mu) = \mathop{\mathbb{E}}_{\tau\sim P_r^{\pi}_{\mu}} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi(a_t|s_t)\, Q^{\pi_\theta}(s_t,a_t) \right] =$$

$$\mathop{\mathbb{E}}_{\tau\sim P_r^{\pi}_{\mu}} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi(a_t|s_t) \left[Q^{\pi_\theta}(s_t,a_t) - V^{\pi_\theta}(s_t) \right] \right] =$$

$$\mathop{\mathbb{E}}_{\tau\sim P_r^{\pi}_{\mu}} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi(a_t|s_t)\, A^{\pi_\theta}(s_t,a_t) \right] \xrightarrow[(8)]{\text{using eq.}}$$

$$\boxed{\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s)\, A^{\pi_\theta}(s,a) \right]}$$

2 Value-based Methodes for Continous Actions

a) Since Q-learning is based on a greedy policy which involves finding the $\left[\arg\max_a Q(s,a) \;\&\; \max_a Q(s,a)\right]$ and it requires an iterative optimization process at every step. In other words, optimizing in the inner loop of an optimization problem would lead to massive increase in the computation time.

Also, discretizing the action space is limited as well and the number of actions increases exponentially with the degrees of freedom. (DDPG original paper)

(b)

1. $$Q_\phi(s,a) = -\frac{1}{2}\left(a-\mu_\theta(s)\right)^T P_\phi(s)\left(a-\mu_\theta(s)\right) + V_\phi(s)$$

$$\arg\max_a Q_\phi(s,a) : \frac{\partial Q_\phi}{\partial a} = 0 \Longrightarrow -\frac{1}{2}\times 2 \times P_\phi(s)\left(a-\mu_\theta(s)\right)=0$$

$$\Longrightarrow \quad P_{\phi(s)}\left(a - \mu_{\theta}(s)\right) = 0 \quad \overset{*}{\Longrightarrow} \quad a = \mu_{\theta}(s)$$

* It can have other answers if $P_{\phi(s)}$ is not positive definite.

$$\Longrightarrow \quad Q_{\phi}(s,a)\Big|_{a=\mu_{\theta}(s)} = 0 + V_{\phi}(s) = V_{\phi}(s)$$

$$\Longrightarrow \quad \boxed{\begin{aligned} &\arg\max_{a} Q_{\phi}(s,a) = \mu_{\theta}(s) \\ &\max_{a} Q_{\phi}(s,a) = V_{\phi}(s) \end{aligned}}$$

Using simple functions for state-action values, would result in a model which cannot capture the complex behaviour needed to execute in the environment; however, it results in a simpler model which needs less computation & etc.

(Somehow the same as underfitting!)

2.a DDPG is a model-free, off-policy, actor-critic method which alleviates the problem of a Q-learning in continous action spaces by interleaving learnig an approximator to $Q^*(s,a)$ with learnig an appoximator to $a^*(s) = \mu_\theta(s)$. Since the action space is continaw, the function $Q^*(s,a)$ is presumed to be differantiable arel we can approximate

$$\max_a Q(s,a) \approx Q(s, \mu_\theta(s)) .$$

The Q-learning pant minimizes (critic):

$$L(\phi, D) = \mathop{E}_{u \sim D}\left[\left(Q_\phi(s,a) - \left(r + \gamma \max_{a'} Q_{\phi_{targ}}(s', \mu_{\theta_{targ}}(s')) \right) \right)^2 \right]$$

where D is an experiance replay buffer (we are abk to do this sense Bellman eq. doesn't care which transition tuples are used arel etc, Since $Q^*(s,a)$ should satisfy Bellman optimality for all transitions.) Also, target networks are usecl for

Stability of learning by updating the target network by polyak averaging once per main network update: ($\tau \ll 1$)

$$\phi_{targ} \leftarrow \tau \phi + (1-\tau)\, \phi_{targ}$$

$$\theta_{targ} \leftarrow \tau \theta + (1-\tau)\, \theta_{targ}$$

The Objective function for actor is also:

$$\max_{\theta} \; \underset{s \sim D}{E} \left[Q_{\phi}(s, \mu_{\theta}(s)) \right]$$

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q)^2)$
    Update the actor policy using the sampled gradient:

$$\nabla_{\theta^\mu} \mu|_{s_i} \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$

$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

  end for
end for

2.6 In the DDPG algorithm, the actor and critic work with each other in the sense that the evaluation of critic is passed to the actor to maximize the performance through policy gradient and therefore, passing of gradients from critic to actor provides a better evaluation of value of the action.

However, in the classic REINFORCE algorithm, the trajectories are obtained from the current policy and the estimation of value of state has a high variance and values are not learnt through any critic network which results in a poor estimate and subsequent problems. Therefore actor-critic methods such as DDPG alleviate the problem of high variance estimate of value of states.

2.C $$Q^{\mu}(s,\mu_\theta(s)) = r(s,\mu_\theta(s)) + \int_S \gamma\, p(s'|s,\mu_\theta(s))\, V^{\mu}(s')\,ds'$$

$$\Rightarrow \nabla_\theta V^{\mu}(s) = \nabla_\theta Q^{\mu}(s,\mu_\theta(s)) = \nabla_\theta \mu_\theta(s)\, \nabla_a r(s,\mu_\theta(s))\Big|_{a=\mu_\theta(s)} +$$ (Chain rule)

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

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

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

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

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

$$\int_S \gamma\, Pr(s \to s', 1, \mu_\theta) \int_S \gamma\, Pr(s' \to s'', 1, \mu_\theta(s'))\, \nabla_\theta V^{\mu}(s'')\,ds''\,ds' =$$

(Consider s' as the middle point to reach s'')

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

$$+ \int_S \gamma^2 \Pr(s \to s'', 2, \mu_\theta) \underbrace{\nabla_\theta V^\mu(s'')}_{\text{Iterate again}} ds'' = \dots =$$

where $\Pr(s \to s', k, \mu_\theta)$ is the probability of transition from state s to s' in k steps under policy μ_θ.

Since $\| \nabla_\theta V^\mu(s) \|$ is bounded, we can rewrite this :

$$\nabla_\theta V^\mu(s) = \int_S \sum_{t=0}^{H} \gamma^t \Pr(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^\mu(s,a)\Big|_{a=\mu_\theta(s')} ds'$$

Expectation over starting states distribution $p_0(s)$:

$$\nabla_\theta J(\mu_\theta) = \nabla_\theta \mathop{\mathbb{E}}_{s_0 \sim p_0}[V^\mu(s_0)] = \nabla_\theta \int_S p_0(s) V^\mu(s) ds = \int_S p_0(s) \nabla_\theta V^\mu(s) ds$$

$= \dots$

$$= \int_S p_0(s) \int_S \sum_{t=0}^{H} \gamma^t \, Pr(s \to s', t, \mu_\theta) \, \nabla_\theta \mu_\theta(s') \, \nabla_a Q(s', a) \Big|_{a=\mu_\theta(s')} ds' ds$$

$$= \int_S \int_S \sum_{t=0}^{H} \gamma^t \, p_0(s) \, Pr(s \to s', t, \mu_\theta) \, \nabla_\theta \mu_\theta(s') \, \nabla_a Q(s', a) \Big|_{a=\mu_\theta(s')} ds' ds$$

$$= \int_S \nabla_\theta \mu_\theta(s') \, \nabla_a Q(s', a) \Big|_{a=\mu_\theta(s')} ds' \int_S \underbrace{\sum_{t=0}^{H} \gamma^t \, p_0(s) \, Pr(s \to s', t, \mu_\theta) \, ds}_{\rho^\mu(s')}$$

$$= \int_S \rho^\mu(s') \, \nabla_\theta \mu_\theta(s') \, \nabla_a Q(s', a) \Big|_{a=\mu_\theta(s')} ds'$$

where $\rho^\mu(s') = \sum_t \gamma^t \, p_0(s) \, Pr(s \to s', t, \mu_\theta) \, ds$ is the discounted state distribution.

$\Longrightarrow$

$$\nabla_\theta J(\mu_\theta) = \int_S \rho^\mu(s) \, \nabla_\theta \mu_\theta(s) \, \nabla_a Q(s, a) \Big|_{a=\mu_\theta(s)}$$

$$= \mathop{\mathbb{E}}_{s \sim \rho^\mu} \left[\nabla_\theta \mu_\theta(s) \, \nabla_a Q(s, a) \Big|_{a=\mu_\theta(s)} \right]$$

3 New Optimization method by Charging Trust Region

(a) $$V_{\tilde{\pi}}(s_0) = (1-\gamma)\, \mathop{\mathbb{E}}_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t)\right] = \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\rho,\tilde{\pi}}\left[(1-\gamma) R(s_t,a_t)\right]$$

$$= \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\rho,\tilde{\pi}}\left[(1-\gamma) R(s_t,a_t) + V_{\pi}(s_t) - V_{\pi}(s_t)\right] =$$

$$\sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\rho,\tilde{\pi}}\left[(1-\gamma) R(s_t,a_t) + \gamma V_{\pi}(s_{t+1}) - V_{\pi}(s_t)\right] + V_{\pi}(s_0)$$

$$= V_{\pi}(s_0) + \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\tilde{\pi},\rho}\left[A_{\pi}(s_t,a_t)\right] \Longrightarrow$$

$$V_{\tilde{\pi}}(s_0) = V_{\pi}(s_0) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\tilde{\pi},\rho}\left[A_{\pi}(s_t,a_t)\right] \xrightarrow[\rho_{\tilde{\pi}}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t=s|\tilde{\pi},\rho)]{\mathbb{E}_{s_0}}$$

$$\boxed{J(\tilde{\pi}) = J(\pi) + \int_S \int_A A^{\pi}_{(s,a)}\, d\tilde{\pi}(a|s)\, d\rho_{\tilde{\pi}}(s)}$$

(b) Original Optimization Problem:

$$\sup_{\tilde{\pi}\in\Pi} \int_S \int_A A^{\pi}(s,a)\, d\tilde{\pi}(a|s)\, d\rho_{\pi}(s)$$

$$\text{s.t. } \tilde{\pi}\in T_{\epsilon} := \left\{\tilde{\pi}\in\Pi : \int_S C(\pi(\cdot|s), \tilde{\pi}(\cdot|s))\, d\rho_{\pi}(s) \leq \epsilon\right\}$$

We rewrite the above eq. by duality and minimax:

$$\sup_{\tilde{\pi}\in\Pi} \inf_{\lambda\geq 0} \left\{ \int_S \int_A A^{\pi}(s,a)\, d\tilde{\pi}(a|s)\, d\rho_{\pi}(s) + \lambda\left(\epsilon - \int_S C(\pi(\cdot|s), \tilde{\pi}(\cdot|s))\, d\rho_{\pi}(s)\right)\right\}$$

$$\text{s.t. } \lambda \geq 0.$$

$$\leq$$

$$\inf_{\lambda\geq 0} \left\{ \lambda\epsilon + \sup_{\tilde{\pi}\in\Pi} \left\{ \int_S \int_A A^{\pi}(s,a)\, d\tilde{\pi}(a|s) - \lambda C(\pi(\cdot|s), \tilde{\pi}(\cdot|s)\, d\rho_{\pi}(s)\right\}\right\}$$

$$\leq$$

$$\star \quad \inf_{\lambda\geq 0} \left\{ \lambda\epsilon + \int_S \underbrace{\sup_{\tilde{\pi}\in\Pi} \left\{ \int_A A^{\pi}(s,a)\, d\tilde{\pi}(a|s) - \lambda C(\pi(\cdot|s), \tilde{\pi}(\cdot|s)\right\}}_{*} d\rho_{\pi}(s)\right\}$$

$$* : \lambda > 0 \quad : \quad \left(\text{if } \lambda = 0, \; * = \sup_{\tilde{\pi} \in \Pi} \int_A A^{\pi}(s,a) \, d\tilde{\pi}(a|s) \right)$$

Kantorovich Duality : $\left(\psi(\cdot) = \dfrac{A^{\pi}(s,\cdot)}{\lambda} \;\; \& \;\; \phi(\cdot) = \inf_{a' \in A} \{ c(\cdot, a') - \psi(a') \} \right)$

$$* = \sup_{\pi \in \Pi} \left\{ \int_A A^{\pi}(s,a) \, d\tilde{\pi}(a|s) - \lambda \sup_{\phi + \psi \leq c} \left\{ \int_A \phi(a) \, d\pi + \int_A \psi(a) \, d\tilde{\pi} \right\} \right\}$$

$$\phi(x) + \psi(x') = \inf_{a \in A} \{ c(x,a) - \psi(a) \} + \psi(x') \leq c(x,x')$$

$$\Longrightarrow * \leq \sup_{\pi \in \Pi} \left\{ \int_A - \inf_{a' \in A} \{ \lambda c(a,a') - A^{\pi}(s,a') \} \, d\pi(a|s) \right.$$

$$= \int_A \sup_{a' \in A} \{ A^{\pi}(s,a') - \lambda c(a,a') \} \, d\pi(a|s) \quad ** $$

** in $\Longrightarrow$ * Dual Problem

$$\inf \left\{ \lambda \epsilon + \int_S \int_A \sup_{a' \in A} \{ A^{\pi}(s,a') - \lambda c(a,a') \} \, d\pi(a|s) \, d\rho_{\pi}(s) \right.$$

$$\text{s.t. } \lambda \geq 0$$