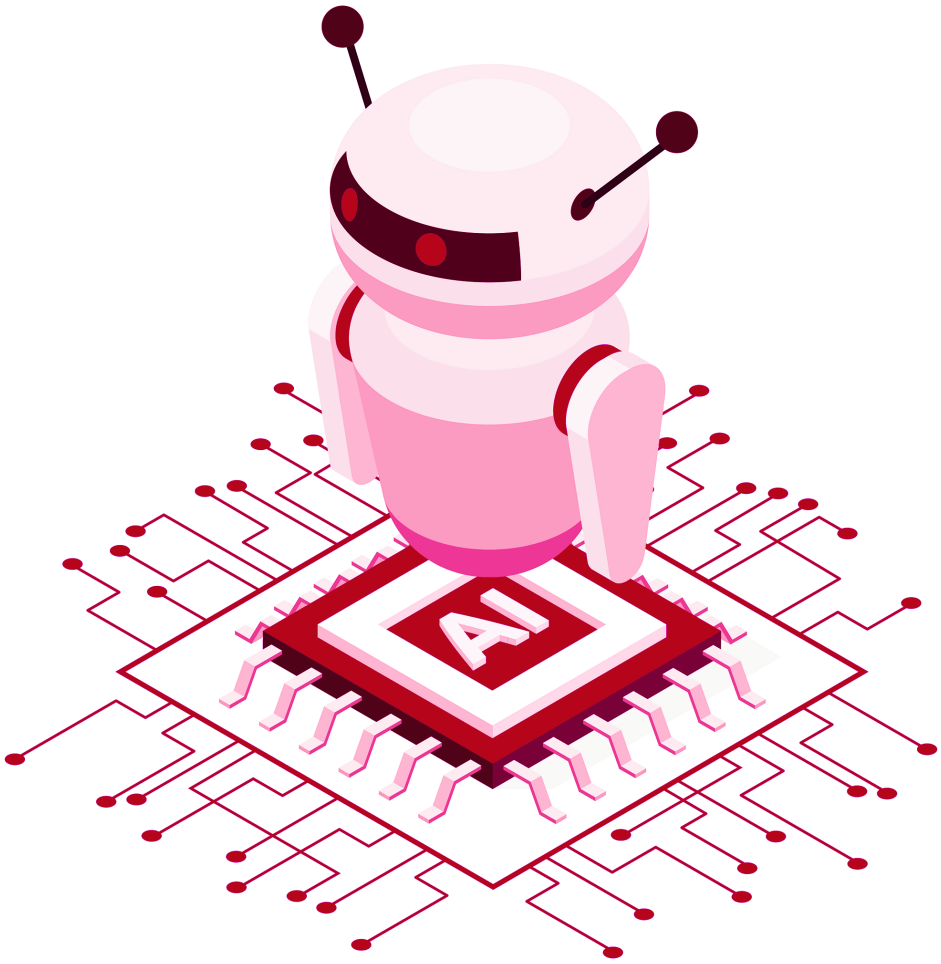

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 8:

Policy-Based Theory

By:

Amir Kooshan Fattah Hesari

401102191

RIML

Deep RL Course

Spring 2025

Contents

Grading

The grading will be based on the following criteria, with a total of 100 points:

Task	**Points**
Policy Gradient - Part (a)	20
Policy Gradient - Part (b)	10
Trust Region Policy Optimization - Part (a)	10
Trust Region Policy Optimization - Part (b)	5
Trust Region Policy Optimization - Part (c)	10
Trust Region Policy Optimization - Part (d)	20
Trust Region Policy Optimization - Part (e)	20
Trust Region Policy Optimization - Part (f)	5
Bonus: Writing your report in Latex	5

1 Policy Gradient Theorem

In this question, we will prove the policy gradient theorem and provide a set of sufficient conditions that allow us to use function approximations as a critic for the Q-value function so that the policy gradient using our function approximation remains exact.

1.1 Notations

Consider a normal finite MDP with bounded rewards. $P(s'|s,a)$ represents the transition model, which corresponds to the probability of transitioning from state s to s' due to action a. Also, the reward model is represented by $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ where $r(s,a)$ is the immediate reward associated with taking action a in state s. Parameter $\gamma \in [0,1)$ corresponds to the discount factor, and s_0 indicates the starting state of our MDP.

A parametrized policy π_θ induces a distribution over trajectories $\tau = (s_t, a_t, r_t)|_{t=0}^{\infty}$ where s_0 is the starting state, and for all subsequent timesteps t, $a_t \sim \pi(.|s_t)$, $s_{t+1} \sim P(.|s_t, a_t)$. The state value function and the state-action value (Q-value) functions are defined as follows by the Bellman operator:

$$V^{\pi_\theta}(s) = \mathbb{E}_{a\sim\pi_\theta(.|s)}[Q^{\pi_\theta}(s,a)]$$
$$Q^{\pi_\theta}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(.|s,a)}[V^{\pi_\theta}(s')]$$

We also define the discounted state visitation distribution $d_{s_0}^{\pi}$ of a policy π as:

$$d_{s_0}^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0), \tag{1}$$

where $Pr^{\pi}(s_t = s|s_0)$ is the state visitation probability that $s_t = s$, after we execute π starting at state s_0.

1.2 Proving the Policy Gradient Theorem

The objective function of our RL problem is defined as $J(\theta) = V^{\pi_\theta}(s_0)$. The policy gradient method uses the gradient ascent algorithm to optimize θ. This can be done by the direct differentiation of the objective function.

a) Prove the following identity, which is known as the Policy Gradient Theorem:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \tag{2}$$

Answer :

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta V^{\pi_\theta}(s_0) = \nabla_\theta \mathbf{E}_{a\sim\pi_\theta(.|s)}\left[Q^{\pi_\theta}(s,a)\right] \\
&= \nabla_\theta \sum_{a_0} \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0,a_0) \\
&= \sum_{a_0} \nabla_\theta \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0,a_0) + \pi_\theta(a_0|s_0)\nabla_\theta Q^{\pi_\theta}(s_0,a_0) \\
&= \sum_{a_0} \pi_\theta(a_0|s_0)(\nabla_\theta \log \pi_\theta(a_0|s_0)) Q^{\pi_\theta}(s_0,a_0) + \gamma \sum_{a_0,s_1} \pi_\theta(a_0|s_0) P(s_1|s_0,a_0)\nabla_\theta V^{\pi_\theta}(s_1)
\end{aligned}$$

In the last line we have used the fact that

$$\nabla_\theta Q^{\pi_\theta}(s_0, a_0) = r(s_0, a_0) + \gamma \mathbf{E}_{s' \sim P(.|s_0,a_0)} \left[V^{\pi_\theta}(s') \right]$$

If we define $\text{Pr}_\mu^\pi(\tau) = \mu(s_0)\pi(a_0|s_0)P(s_1|s_0, a_0)\pi(a_1|s_1)...$, we are going to have :

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbf{E}_{a_0 \sim \pi_(.|s_0)} \left[\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \right] + \gamma \mathbf{E}_{a_0 \sim \pi(.|s_0)} \mathbf{E}_{s_1 \sim P(.|s_0,a_0)} \left[\nabla_\theta V^{\pi_\theta}(s_1) \right] \\
&= (*) \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \right] + \gamma \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta V^{\pi_\theta}(s_1) \right] \\
&= \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \right] + \gamma \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_1|s_1) Q^{\pi_\theta}(s_1, a_1) \right] \\
&+ \gamma^2 \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_2|s_2) Q^{\pi_\theta}(s_2, a_2) \right] + ... \\
\nabla_\theta J(\theta) &= \sum_{t=0}^{\infty} \gamma^t \mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t, a_t) \right]
\end{aligned}$$

The reason that we can change $a_0 \sim \pi(.|s_0)$ to $\tau \sim \text{Pr}_\mu^\pi$ is shown below :

$$\begin{aligned}
&\mathbf{E}_{\tau \sim Pr_\mu^\pi} \left[\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \right] \\
&= \sum_{a_0} \sum_{s_1} \sum_{a_1} ... \mu(s_0)\pi(a_0|s_0)P(s_1|a_0, s_0)\pi(a_1|s_1) ... \nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \\
&= \sum_{a_0} \pi(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \Big\{ \sum_{s_1} \sum_{a_1} ... \mu(s_0)P(s_1|a_0, s_0)\pi(a_1|s_1) ... \Big\} \\
&= \sum_{a_0} \pi(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \cdot 1 \\
&= \mathbf{E}_{a_0 \sim \pi(a_0|s_0)} \left[\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0) \right]
\end{aligned}$$

Now if we take the term inside the expectation as $f(t) = \nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0, a_0)$ and taking $\text{Pr}_\mu^\pi = \mathbf{P}(S_t = s, A_t = a \,|S_0 \sim \mu, \pi_\theta)$ we have

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{\infty} \gamma^t \sum_s \sum_a \mathbf{P}(S_t = s, A_t = a \,|S_0 \sim \mu, \pi_\theta) f(t) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_s \sum_a \mathbf{P}(S_t = s \,|S_0 \sim \mu, \pi_\theta) \pi_\theta(a|s) f(t) = \sum_{t=0}^{\infty} \gamma^t \sum_s \mathbf{P}(S_t = s \,|S_0 \sim \mu, \pi_\theta) \sum_a \pi_\theta(a|s) f(t) \\
&= \sum_s \underbrace{\Big\{ \sum_{t=0}^{\infty} \gamma^t \mathbf{P}(S_t = s \,|S_0 \sim \mu, \pi_\theta) \Big\}}_{\frac{d_{s_0}^\pi(s)}{1-\gamma}} \sum_a \pi_\theta(a|s) f(t) = \frac{1}{1-\gamma} \sum_s d_{s_0}^\pi(s) \sum_a \pi_\theta(a|s) f(t) \\
&= \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_{s_0}^\pi(s)} \mathbf{E}_{a \sim \pi_\theta(.|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \right]
\end{aligned}$$

1.3 Compatible Function Approximation Theorem

Now, consider the case in which Q^{π_θ} is approximated by a learned function approximator. If the approximation is sufficiently good, we might hope to use it in place of Q^{π_θ} in equation 2. If we use the function approximator $Q_\phi(s, a)$, the convergence of our method is not necessarily maintained due to the fact that our gradient will not be exact anymore. The following theorem provides sufficient conditions for our function approximator so that our gradient using the approximator remains exact.

Theorem 1.1. *(Compatible Function Approximation). If the following two conditions are satisfied for any function approximator with parameter ϕ:*

1. *Critic gradient is compatible with the Actor score function, i.e.,*
$$\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$$
2. *Critic parameters ϕ minimize the following mean-squared error[1]:*
$$\epsilon = \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2]$$

Then, the policy gradient using critic $Q_\phi(s,a)$ is exact, i.e.,
$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)]$$

b) Prove theorem 1.1.

Answer : We expand the ϵ term and have that :

$$\begin{aligned}
\epsilon &= \mathbf{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(.|s)} \left[(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2\right] \\
&= \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s)(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2
\end{aligned}$$

$$\begin{aligned}
\nabla_\phi \epsilon &= \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s) \nabla_\phi (Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2 \\
&= -2 \cdot \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s)(Q^{\pi_\theta}(s,a) - Q_\phi(s,a)) \nabla_\phi Q_\phi(s,a) = 0
\end{aligned}$$

and from the first condition we know that the critic gradient is compatible with the Actor score function , so

$$\begin{aligned}
&-2 \cdot \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s)(Q^{\pi_\theta}(s,a) - Q_\phi(s,a)) \nabla_\phi Q_\phi(s,a) = \\
&-2 \cdot \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s)(Q^{\pi_\theta}(s,a) - Q_\phi(s,a)) \nabla_\theta \log \pi_\theta(a|s) = 0 \\
&\sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \quad (*) \\
&= \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s) Q_\phi(s,a) \nabla_\theta \log \pi_\theta(a|s) \quad (**)
\end{aligned}$$

$$(*) = (1-\gamma)\nabla_\theta J(\theta) = (**)$$

$$\begin{aligned}
(1-\gamma)\nabla_\theta J(\theta) &= \sum_s (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathsf{Pr}^\pi(s_t = s|s_0) \sum_a \pi_\theta(a|s) Q_\phi(s,a) \nabla_\theta \log \pi_\theta(a|s) \\
&= \mathbf{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(.|s)} \left[\nabla_\theta \pi_\theta(a|s) Q_\phi(s,a)\right] \\
\nabla_\theta J(\theta) &= \frac{1}{1-\gamma} \mathbf{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(.|s)} \left[\nabla_\theta \pi_\theta(a|s) Q_\phi(s,a)\right]
\end{aligned}$$

[1] Assume that the mean-squared error has only one critical point which corresponds to its minimum.

2 Trust Region Policy Optimization

In this question, we will dive deep into the mathematical theories behind the TRPO algorithm. As a roadmap, we first prove that minimizing a certain surrogate objective function guarantees policy improvement with non-trivial step sizes. Then, we make a series of approximations to the theoretically justified algorithm, yielding a practical algorithm, which has been called trust region policy optimization (TRPO).

2.1 Notations and Preliminaries

Let π denote a stochastic policy and let $\eta(\pi)$ denote its expected discounted reward:

$$\eta(\pi) = \mathbb{E}_{s_0,a_0,\ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t)]$$

where

$$s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t), s_{t+1} \sim P(s_{t+1}|s_t, a_t).$$

Also, we will use the following standard definitions of the state-action value function Q_π, the value function V_π, and the advantage function A_π:

$$\begin{aligned} Q_\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1},a_{t+1},\ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})] \\ V_\pi(s_t) &= \mathbb{E}_{a_t,s_{t+1},\ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})] \\ A_\pi(s, a) &= Q_\pi(s, a) - V_\pi(s) \end{aligned}$$

a) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{3}$$

Equation 3 basically shows that the difference between the expected total rewards of any two policies π' and π depends on the advantage function of policy π if the trajectory is sampled by running π'. We will use this equation to derive an optimization scheme further to maximize the expected total reward using the advantage function of policy π to obtain policy π'.

Let ρ_π be the unnormalized discounted visitation frequencies:

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots$$

Proof. We have that

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s) = \mathbf{E}_{s'\sim P(s'|s,a)}\left[r(s) + \gamma \cdot V_\pi(s') - V_\pi(s)\right]$$

$$\begin{aligned}
\mathbf{E}_{s_0,a_0,s_1,a_1,\ldots\sim\pi'}\left[\sum_t \gamma^t A_\pi(s,a)\right] &= \mathbf{E}_{\tau|\pi'}\left[\sum_t \gamma^t (r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t))\right] \\
&= \mathbf{E}_{\tau|\pi'}\left[-V_\pi(s_0) + \sum_{t=0} \gamma^t r(s_t)\right] \\
&= -\mathbf{E}_{s_0}\left[V_\pi(s_0)\right] + \mathbf{E}_{\tau|\pi'}\left[\sum_{t=0} \gamma^t r(s_t)\right] = -\eta(\pi) + \eta(\pi') \\
\Rightarrow \eta(\pi') &= \eta(\pi) + \mathbf{E}_{\tau|\pi'}\left[\sum_{t=0} \gamma^t A_\pi(s_t,a_t)\right] \quad \text{QED}
\end{aligned}$$

□

b) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{4}$$

Equation 4 can be used as an optimization objective in reinforcement learning. Note that this equation has been considered difficult to optimize directly due to the complex dependency of $\rho_{\pi'}(s)$ on π'. Instead, the following local approximation of η has been introduced for optimization:

$$L_\pi(\pi') = \eta(\pi) + \sum_s \rho_{\pi}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{5}$$

Note that L_π uses the visitation frequency ρ_π rather than $\rho_{\pi'}$, ignoring changes in state visitation density due to changes in the policy. In the next section, we will derive an algorithm to guarantee a monotonic improvement in our policy using equation 5 as our objective function, showing that equation 5 is good enough in our case.

Proof.

$$\begin{aligned}
\eta(\pi') &= \eta(\pi) + \sum_{t=0}^{\infty} \sum_s P(s_t = s|\pi') \sum_a \pi'(a|s) \gamma^t A_\pi(s,a) \\
&= \eta(\pi) + \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi') \sum_a \pi'(a|s) A_\pi(s,a) \\
&= \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a) \quad \text{QED}
\end{aligned}$$

2.2 Monotonic Improvement Guarantee for General Stochastic Policies

In this section, we build the theoretical foundations to consider the policy optimization problem, assuming that the policy can be evaluated at all states. The ultimate goal of this section is to prove the following theorem:

Theorem 2.1. *Let π, π' be two stochastic policies. Then, the following bound holds:*

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi, \pi')$$
$$\textit{where } \epsilon = \max_{s,a} |A_\pi(s,a)|$$

During this section, we use the following definitions and inequality for the total variation and KL divergence:

$$D_{TV}(p||q) = \frac{1}{2}\sum_i |p_i - q_i|$$
$$D_{TV}^{\max}(\pi, \pi') = \max_s D_{TV}(\pi(.|s)||\pi'(.|s))$$
$$D_{KL}^{\max}(\pi, \pi') = \max_s D_{KL}(\pi(.|s)||\pi'(.|s))$$
$$D_{TV}(p||q)^2 \leq D_{KL}(p||q)$$

We will prove theorem 2.1 step by step, and you are required to complete the proof as indicated below. To begin the proof, we denote trajectories by τ and define $\bar{A}(s)$ as follows:

$$\bar{A}(s) = \mathbb{E}_{a\sim\pi'(.|s)}[A_\pi(s,a)]$$

Then we can rewrite equations 4 and 5 as follows:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t)] \tag{6}$$

$$L_\pi(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t)] \tag{7}$$

The only difference in these two equations is whether the states are sampled using π or π'. To bound the difference between $\eta(\pi')$ and $L_\pi(\pi')$, we first need to introduce a measure of how much π and π' agree. Specifically, we'll couple the policies so that they define a joint distribution over pairs of actions. We use the following definition of α-coupled policy pairs:

Definition 2.2. *(π, π') is an α-coupled policy pair if it defines a joint distribution $(a, a')|s$ such that $P(a \neq a'|s) \leq \alpha$ for all s. π and π' will denote the marginal distributions of a and a', respectively.*

c) Prove the following lemma:

Lemma 2.3. *Given that π, π' are α-coupled policies, for all s,*

$$|\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)|$$

Proof.

$$\bar{A}(s) = \mathbf{E}_{a' \sim \pi'} \left[A_\pi(s,a') \right] = \mathbf{E}_{(a,a') \sim (\pi,\pi')} \left[A_\pi(s,a') - A_\pi(s,a) \right]$$

The previous equality holds because we have :

$$\begin{aligned} \mathbf{E}_{a\sim\pi} \left[A_\pi(s,a) \right] = \sum_{a \in A} \pi(a|s)(Q_\pi(s,a) - V_\pi(s)) &= \mathbf{E}_{a\sim\pi} \left[Q_\pi(s,a) \right] - \mathbf{E}_{a\sim\pi} \left[\mathbf{E}_{a\sim\pi} \left[Q_\pi(s,a) \right] \right] \\ &= \mathbf{E}_{a\sim\pi} \left[Q_\pi(s,a) \right] - \mathbf{E}_{a\sim\pi} \left[Q_\pi(s,a) \right] = 0 \end{aligned}$$

And from the definition of an α-coupled policy pair we have

$$\bar{A}(s) = \mathbf{E}_{(a,a') \sim (\pi,\pi') | a \neq a'} \left[A_\pi(s,a') - A_\pi(s,a) \right] = \sum_{a,a'} P(a,a'|s) \left[A_\pi(s,a') - A_\pi(s,a) \right]$$

$$a = a' \Rightarrow \sum_{a,a'} P(a,a'|s)[A_\pi(s,a) - A_\pi(s,a)] = \sum_{a,a'} P(a,a'|s) \cdot 0 = 0$$

$$a \neq a' \Rightarrow \sum_{a \neq a'} P(a,a'|s) \left[A_\pi(s,a) - A_\pi(s,a') \right]$$

From the definition of conditional expectation we have that :

$$\mathbf{E}[\mathbf{Y}|\mathbf{Z}] = \sum_y y P(\mathbf{Y} = y|\mathbf{Z})$$

$$\mathbf{Y} = A_\pi(s,a') - A_\pi(s,a)\,,\ \mathbf{Z} = (a \neq a')\,,\ P(Y,Z) = P(Y|Z)P(Z)$$

$$\Rightarrow \sum_{a \neq a'} P(a \neq a'|s) \left[A_\pi(s,a) - A_\pi(s,a') \right]$$

$$= P(a \neq a'|s) \sum_{a \neq a'} \frac{P(a,a'|s)}{P(a \neq a'|s)} \left[A_\pi(s,a) - A_\pi(s,a') \right]$$

$$= P(a \neq a'|s) \mathbf{E}_{(a,a') \sim (\pi,\pi') | a \neq a'} \left[A_\pi(s,a') - A_\pi(s,a) \right]$$

And finally from the **triangle inequality** we have that :

$$|x - y| \leq |x| + |y| \Rightarrow$$

$$|\mathbf{E}_{(a,a')\sim(\pi,\pi')} \left[A_\pi(s,a') \right] - \mathbf{E}_{(a,a')\sim(\pi,\pi')} \left[A_\pi(s,a) \right]| \leq \underbrace{|\mathbf{E}_{(a,a')\sim(\pi,\pi')} \left[A_\pi(s,a') \right]|}_{\leq \max_{s,a} A_\pi(s,a)} + \underbrace{|\mathbf{E}_{(a,a')\sim(\pi,\pi')} \left[A_\pi(s,a) \right]|}_{\leq \max_{s,a} A_\pi(s,a)}$$

$$P(a \neq a'|s) \leq \alpha \Rightarrow |\bar{A}(s)| \leq \alpha \cdot 2 \cdot max_{s,a} A_\pi(s,a) \quad \text{QED}$$

d) Prove the following lemma:

Lemma 2.4. *Let (π, π') be an α-coupled policy pair. Then:*

$$|\mathbb{E}_{s_t \sim \pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t \sim \pi}[\bar{A}(s_t)]| \leq 4\alpha(1 - (1-\alpha)^t) \max_{s,a} |A_\pi(s,a)|$$

Given the coupled policy pair $(\pi, \tilde{\pi})$, we can also obtain a coupling over the trajectory distributions produced by π and $\tilde{\pi}$. We have pairs of trajectories $\tau, \tilde{\tau}$, where τ is obtained by taking actions from π, and $\tilde{\tau}$ is obtained by taking actions from $\tilde{\pi}$, where the same random seed is used to generate both trajectories.

$$E_{s_t\sim\tilde{\pi}}[\tilde{A}(s_t)] = P(n_t = 0)E_{s_t\sim\tilde{\pi}|n_t=0}[\tilde{A}(s_t)] + P(n_t > 0)E_{s_t\sim\tilde{\pi}|n_t>0}[\tilde{A}(s_t)]$$

The expectation decomposes similarly for actions are sampled using π:

$$E_{s_t\sim\pi}[A(s_t)] = P(n_t = 0)E_{s_t\sim\pi|n_t=0}[A(s_t)] + P(n_t > 0)E_{s_t\sim\pi|n_t>0}[A(s_t)]$$

Note that the $n_t = 0$ terms are equal:

$$E_{s_t\sim\tilde{\pi}|n_t=0}[\tilde{A}(s_t)] = E_{s_t\sim\pi|n_t=0}[A(s_t)]$$

because $n_t = 0$ indicates that π and $\tilde{\pi}$ agreed on all timesteps less than t.

$$E_{s_t\sim\tilde{\pi}}[\tilde{A}(s_t)] - E_{s_t\sim\pi}[A(s_t)] = P(n_t > 0)(E_{s_t\sim\tilde{\pi}|n_t>0}[\tilde{A}(s_t)] - E_{s_t\sim\pi|n_t>0}[A(s_t)])$$

By definition of α, $P(\pi, \tilde{\pi}$ agree at timestep $i) \geq 1 - \alpha$, so $P(n_t = 0) \geq (1 - \alpha)^t$, and

$$P(n_t > 0) \leq 1 - (1 - \alpha)^t$$

Next, note that

$$\begin{aligned}\left|E_{s_t\sim\tilde{\pi}|n_t>0}[\tilde{A}(s_t)] - E_{s_t\sim\pi|n_t>0}[A(s_t)]\right| &\leq \left|E_{s_t\sim\tilde{\pi}|n_t>0}[\tilde{A}(s_t)]\right| + \left|E_{s_t\sim\pi|n_t>0}[A(s_t)]\right| \\ &\leq 4\alpha \max_{s,a} |A_\pi(s, a)|\end{aligned}$$

$$E_{s_t\sim\tilde{\pi}}[\tilde{A}(s_t)] - E_{s_t\sim\pi}[A(s_t)] \leq 4\alpha(1 - (1 - \alpha)^t) \max_{s,a} |A_\pi(s, a)|$$

e) Prove the following lemma:

Lemma 2.5. *Let (π, π') be an α-coupled policy pair. Then:*

$$|\eta(\pi') - L_\pi(\pi')| \leq \frac{4\alpha^2\gamma\epsilon}{(1 - \gamma)^2}$$

The preceding Lemma bounds the difference in expected advantage at each timestep t. We can sum over time to bound the difference between $\eta(\tilde{\pi})$ and $L_\pi(\tilde{\pi})$. Defining $\epsilon = \max_{s,a} |A_\pi(s, a)|$,

$$\begin{aligned}|\eta(\tilde{\pi}) - L_\pi(\tilde{\pi})| &= \left|\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau\sim\pi}[\tilde{A}(s_t)] - \mathbb{E}_{\tau\sim\tilde{\pi}}[\tilde{A}(s_t)]\right| \\ &\leq \sum_{t=0}^{\infty} \gamma^t \cdot 4\alpha\epsilon(1 - (1 - \alpha)^t) \\ &= 4\alpha\epsilon\left(\frac{1}{1 - \gamma} - \frac{1}{1 - \gamma(1 - \alpha)}\right) \\ &= \frac{4\alpha^2\gamma\epsilon}{(1 - \gamma)(1 - \gamma(1 - \alpha))} \\ &\leq \frac{4\alpha^2\gamma\epsilon}{(1 - \gamma)^2}\end{aligned}$$

f) Prove theorem 2.1. Hint: Use the fact that if we have two policies π and π' such that $D_{TV}^{\max}(\pi, \pi') \leq \alpha$, then we can define an α-couples policy pair (π, π') with appropriate marginals.
Answer : This has been proven in the previous parts step by step! [2]

Note that the inequality in theorem 2.1 becomes an equality in $\pi' = \pi$. Thus, the following optimization problem guarantees a non-decreasing expected return η:

$$\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi) - CD_{KL}^{\max}(\pi_i, \pi)$$
$$\text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$$
$$\text{and } L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s, a)$$

In practice, if we use the penalty coefficient C as recommended by the theory above, the step sizes would be very small. One way to take larger steps in a robust way is to use a constraint on the KL divergence between the two policies as a trust region:

$$\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi)$$
$$\text{subject to } D_{KL}^{\max}(\pi_i, \pi) \leq \delta$$

This problem imposes a constraint that the KL divergence is bounded at every point in the state space. While it is motivated by the theory, this problem is impractical to solve due to the large number of constraints. Instead, we can use a heuristic approximation by considering the average KL divergence. The following optimization problem has been proposed as the TRPO algorithm:

$$\pi_{i+1} = \arg\max_{\pi} L_{\pi_i}(\pi)$$
$$\text{subject to } \mathbb{E}_{s\sim\rho}[D_{KL}(\pi_i(.|s)||\pi(.|s))] \leq \delta$$

[2] There is no need to prove this hint!