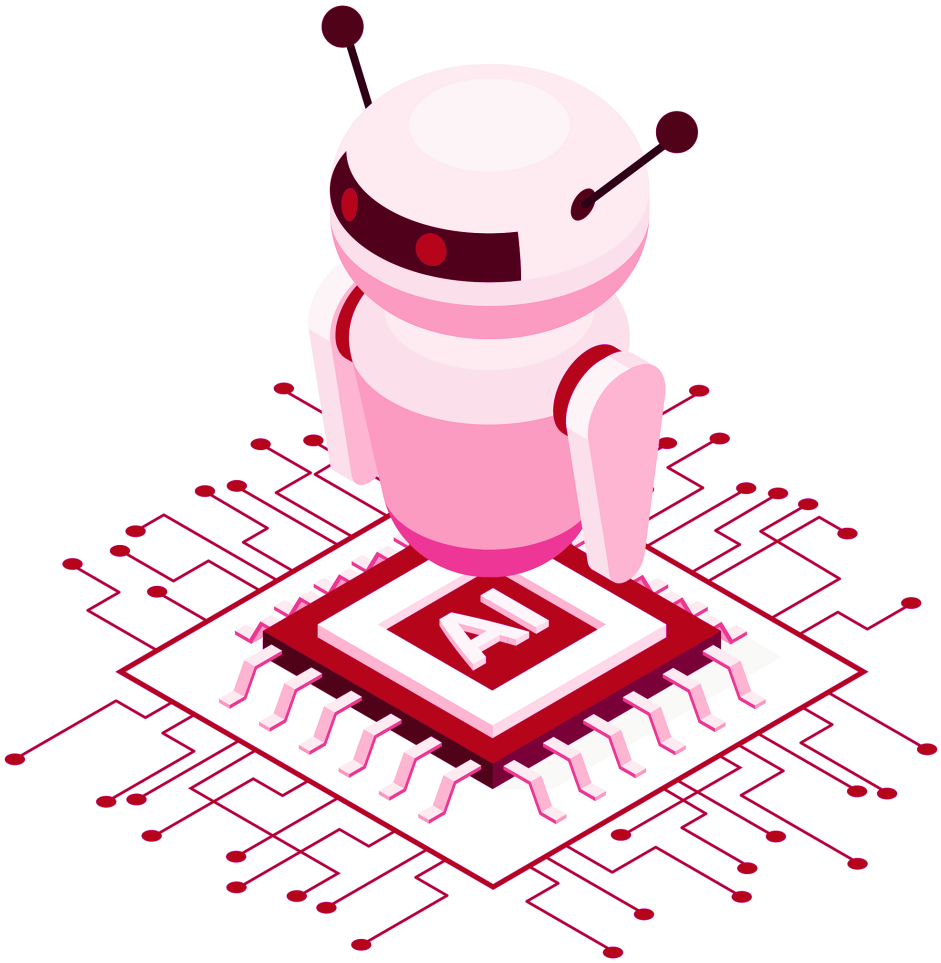

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 8:

Policy-Based Theory

By:

Taha Majlesi

Student ID: 810101504

RIML

Deep RL Course

Spring 2025

Contents

1 Policy Gradient Theorem **1**

1.1 Notations 1

1.2 Proving the Policy Gradient Theorem 1

1.3 Compatible Function Approximation Theorem 3

2 Trust Region Policy Optimization **6**

2.1 Notations and Preliminaries 6

2.2 Monotonic Improvement Guarantee for General Stochastic Policies 10

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:

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

Solution:

We will prove the Policy Gradient Theorem by directly differentiating the objective function $J(\theta) = V^{\pi_\theta}(s_0)$.

Step 1: Express the objective in terms of trajectories

The value function can be written as:

$$J(\theta) = V^{\pi_\theta}(s_0) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] \tag{3}$$

where $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ is a trajectory sampled from policy π_θ.

Step 2: Differentiate with respect to policy parameters

Taking the gradient:

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] \tag{4}$$

$$= \sum_\tau P(\tau|\theta)\nabla_\theta\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] + \sum_\tau\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right]\nabla_\theta P(\tau|\theta) \tag{5}$$

The first term is zero since rewards don't depend on θ. For the second term:

$$\nabla_\theta P(\tau|\theta) = \nabla_\theta\left[\prod_{t=0}^{\infty}\pi_\theta(a_t|s_t)P(s_{t+1}|s_t, a_t)\right] \tag{6}$$

$$= P(\tau|\theta)\nabla_\theta\left[\sum_{t=0}^{\infty}\log\pi_\theta(a_t|s_t)\right] \tag{7}$$

$$= P(\tau|\theta)\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t) \tag{8}$$

Step 3: Combine and rearrange

Substituting back:

$$\nabla_\theta J(\theta) = \sum_\tau P(\tau|\theta)\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right]\left[\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\right] \tag{9}$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right)\left(\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\right)\right] \tag{10}$$

Step 4: Use the log-derivative trick

We can rewrite this as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\left(\sum_{k=t}^{\infty}\gamma^k r(s_k, a_k)\right)\right] \tag{11}$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\gamma^t\left(\sum_{k=0}^{\infty}\gamma^k r(s_{t+k}, a_{t+k})\right)\right] \tag{12}$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\gamma^t Q^{\pi_\theta}(s_t, a_t)\right] \tag{13}$$

Step 5: Convert to state-action expectation

$$\nabla_\theta J(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim d_{s_0,t}^{\pi_\theta}} \mathbb{E}_{a_t \sim \pi_\theta(.|s_t)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t, a_t) \right] \tag{14}$$

where $d_{s_0,t}^{\pi_\theta}(s) = P(s_t = s|s_0, \pi_\theta)$ is the state visitation probability at time t.

Step 6: Use discounted state visitation distribution

The discounted state visitation distribution is:

$$d_{s_0}^{\pi_\theta}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t d_{s_0,t}^{\pi_\theta}(s) \tag{15}$$

Therefore:

$$\sum_{t=0}^{\infty} \gamma^t d_{s_0,t}^{\pi_\theta}(s) = \frac{d_{s_0}^{\pi_\theta}(s)}{1-\gamma} \tag{16}$$

Step 7: Final result

Substituting:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{\infty} \gamma^t \sum_s d_{s_0,t}^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \tag{17}$$

$$= \sum_s \left(\sum_{t=0}^{\infty} \gamma^t d_{s_0,t}^{\pi_\theta}(s) \right) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \tag{18}$$

$$= \sum_s \frac{d_{s_0}^{\pi_\theta}(s)}{1-\gamma} \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \tag{19}$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] \tag{20}$$

This completes the proof of the Policy Gradient Theorem. □

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.

Solution:

We need to prove that under the two conditions stated in the theorem, the policy gradient using the function approximator $Q_\phi(s,a)$ is exact.

Given conditions:

1. $\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$ (compatibility condition)
2. ϕ minimizes $\epsilon = \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2]$

Goal: Show that

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)] \tag{21}$$

Proof:

From the Policy Gradient Theorem (part a), we have:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \tag{22}$$

We want to show that we can replace $Q^{\pi_\theta}(s,a)$ with $Q_\phi(s,a)$ without changing the gradient.

Step 1: Express the difference

Let $\Delta(s,a) = Q^{\pi_\theta}(s,a) - Q_\phi(s,a)$ be the approximation error. Then:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s)(Q_\phi(s,a) + \Delta(s,a))] \tag{23}$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)] \tag{24}$$

$$+ \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) \Delta(s,a)] \tag{25}$$

Step 2: Show the error term is zero

We need to show that:

$$\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) \Delta(s,a)] = 0 \tag{26}$$

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

Step 3: Use the compatibility condition

From condition 1, we have $\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$.

Since ϕ minimizes the MSE (condition 2), we have:

$$\nabla_\phi \epsilon = 0 \tag{27}$$

Computing the gradient:

$$\begin{aligned}
\nabla_\phi \epsilon &= \nabla_\phi \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2] && (28)\\
&= \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\phi (Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2] && (29)\\
&= \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[2(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))\nabla_\phi Q_\phi(s,a)] && (30)\\
&= 2\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\Delta(s,a)\nabla_\phi Q_\phi(s,a)] && (31)
\end{aligned}$$

Since $\nabla_\phi \epsilon = 0$ and $\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s)$ (compatibility condition), we have:

$$\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\Delta(s,a)\nabla_\theta \log \pi_\theta(a|s)] = 0 \tag{32}$$

Step 4: Conclusion

Therefore, the error term in Step 1 is zero, and we have:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(.|s)}[\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)] \tag{33}$$

This proves that the policy gradient using the function approximator $Q_\phi(s,a)$ is exact under the stated conditions. □

Interpretation:

The compatibility condition ensures that the critic's gradient direction matches the actor's score function direction. The MSE minimization condition ensures that the critic provides unbiased estimates. Together, these conditions guarantee that the approximate policy gradient equals the true policy gradient.

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:

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

a) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0, a_0, \ldots \sim \pi'}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{34}$$

Solution:

We need to prove that the difference in expected discounted reward between two policies π' and π can be expressed in terms of the advantage function $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$.

Step 1: Express $\eta(\pi')$ in terms of trajectories

$$\eta(\pi') = \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right] \tag{35}$$

$$= \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \tag{36}$$

Step 2: Add and subtract $\eta(\pi)$

$$\eta(\pi') = \eta(\pi) + \eta(\pi') - \eta(\pi) \tag{37}$$

$$= \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] - \mathbb{E}_{s_0,a_0,\ldots\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \tag{38}$$

Step 3: Use the definition of advantage function

For any trajectory $(s_0, a_0, s_1, a_1, \ldots)$:

$$\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) = \sum_{t=0}^{\infty}\gamma^t[Q_\pi(s_t,a_t) - \gamma V_\pi(s_{t+1})] \tag{39}$$

$$= \sum_{t=0}^{\infty}\gamma^t Q_\pi(s_t,a_t) - \sum_{t=0}^{\infty}\gamma^{t+1}V_\pi(s_{t+1}) \tag{40}$$

$$= \sum_{t=0}^{\infty}\gamma^t Q_\pi(s_t,a_t) - \sum_{t=1}^{\infty}\gamma^t V_\pi(s_t) \tag{41}$$

$$= Q_\pi(s_0,a_0) + \sum_{t=1}^{\infty}\gamma^t[Q_\pi(s_t,a_t) - V_\pi(s_t)] \tag{42}$$

$$= V_\pi(s_0) + A_\pi(s_0,a_0) + \sum_{t=1}^{\infty}\gamma^t A_\pi(s_t,a_t) \tag{43}$$

$$= V_\pi(s_0) + \sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t) \tag{44}$$

Step 4: Apply to both policies

For policy π':

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] = \mathbb{E}_{s_0\sim\rho_0}[V_\pi(s_0)] + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \tag{45}$$

For policy π:

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] = \mathbb{E}_{s_0\sim\rho_0}[V_\pi(s_0)] + \mathbb{E}_{s_0,a_0,\ldots\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \tag{46}$$

Step 5: Use the fact that $\mathbb{E}_{a\sim\pi(.|s)}[A_\pi(s,a)] = 0$

For any state s:

$$\mathbb{E}_{a\sim\pi(.|s)}[A_\pi(s,a)] = \mathbb{E}_{a\sim\pi(.|s)}[Q_\pi(s,a) - V_\pi(s)] \tag{47}$$
$$= \mathbb{E}_{a\sim\pi(.|s)}[Q_\pi(s,a)] - V_\pi(s) \tag{48}$$
$$= V_\pi(s) - V_\pi(s) = 0 \tag{49}$$

Therefore:

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] = 0 \tag{50}$$

Step 6: Final result

Substituting back into Step 2:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0\sim\rho_0}[V_\pi(s_0)] + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \tag{51}$$
$$- \mathbb{E}_{s_0\sim\rho_0}[V_\pi(s_0)] - 0 \tag{52}$$
$$= \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \tag{53}$$

This completes the proof. □

Equation 34 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:

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

b) Prove the following identity:

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s)\sum_a \pi'(a|s)A_\pi(s,a) \tag{54}$$

Solution:

We need to prove that the expected discounted reward of policy π' can be expressed in terms of the advantage function of policy π and the state visitation distribution of policy π'.

Step 1: Start from equation 34

From part (a), we have:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \tag{55}$$

Step 2: Expand the expectation

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] = \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}[A_\pi(s_t,a_t)] \tag{56}$$

Step 3: Express in terms of state-action probabilities

For each time step t:

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi'}[A_\pi(s_t,a_t)] = \sum_{s_t} P(s_t = s|s_0,\pi')\sum_{a_t}\pi'(a_t|s_t)A_\pi(s_t,a_t) \tag{57}$$

$$= \sum_{s} P(s_t = s|s_0,\pi')\sum_{a}\pi'(a|s)A_\pi(s,a) \tag{58}$$

Step 4: Sum over all time steps

$$\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s_0,a_0,\ldots\sim\pi'}[A_\pi(s_t,a_t)] = \sum_{t=0}^{\infty}\gamma^t\sum_{s} P(s_t = s|s_0,\pi')\sum_{a}\pi'(a|s)A_\pi(s,a) \tag{59}$$

$$= \sum_{s}\left(\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi')\right)\sum_{a}\pi'(a|s)A_\pi(s,a) \tag{60}$$

Step 5: Use the definition of $\rho_{\pi'}(s)$

The unnormalized discounted visitation frequency is:

$$\rho_{\pi'}(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots = \sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi') \tag{61}$$

Therefore:

$$\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi') = \rho_{\pi'}(s) \tag{62}$$

Step 6: Final result

Substituting back:

$$\mathbb{E}_{s_0,a_0,\ldots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] = \sum_{s}\rho_{\pi'}(s)\sum_{a}\pi'(a|s)A_\pi(s,a) \tag{63}$$

Therefore:

$$\eta(\pi') = \eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{64}$$

This completes the proof. □

Interpretation:

This equation shows that the improvement in expected discounted reward when switching from policy π to policy π' depends on: 1. The advantage function $A_\pi(s,a)$ (how much better action a is than the average in state s under policy π) 2. The state visitation distribution $\rho_{\pi'}(s)$ (how often policy π' visits each state) 3. The action probabilities $\pi'(a|s)$ (how policy π' chooses actions in each state)

The key insight is that we can evaluate the improvement of a new policy π' using the advantage function of the current policy π, as long as we account for the different state visitation patterns.

Equation 54 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:

$$L_\pi(\pi') = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \pi'(a|s) A_\pi(s,a) \tag{65}$$

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 65 as our objective function, showing that equation 65 is good enough in our case.

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 54 and 65 as follows:

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi'}[\sum_{t=0}^{\infty}\gamma^t\bar{A}(s_t)] \tag{66}$$

$$L_\pi(\pi') = \eta(\pi) + \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t\bar{A}(s_t)] \tag{67}$$

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,*

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

Solution:

We need to prove that for α-coupled policies, the expected advantage $\bar{A}(s) = \mathbb{E}_{a\sim\pi'(.|s)}[A_\pi(s,a)]$ is bounded.

Step 1: Express $\bar{A}(s)$ in terms of the coupling

Since (π,π') is an α-coupled policy pair, there exists a joint distribution $(a,a')|s$ such that $P(a\neq a'|s)\leq\alpha$ for all s.

Let $a\sim\pi(.|s)$ and $a'\sim\pi'(.|s)$ be the marginal distributions.

Step 2: Use the coupling to bound the difference

$$\begin{aligned}
\bar{A}(s) &= \mathbb{E}_{a'\sim\pi'(.|s)}[A_\pi(s,a')] &(68)\\
&= \mathbb{E}_{(a,a')\sim\text{coupling}}[A_\pi(s,a')] &(69)\\
&= \mathbb{E}_{(a,a')\sim\text{coupling}}[A_\pi(s,a') - A_\pi(s,a) + A_\pi(s,a)] &(70)\\
&= \mathbb{E}_{(a,a')\sim\text{coupling}}[A_\pi(s,a') - A_\pi(s,a)] + \mathbb{E}_{a\sim\pi(.|s)}[A_\pi(s,a)] &(71)
\end{aligned}$$

Step 3: Use the fact that $\mathbb{E}_{a\sim\pi(.|s)}[A_\pi(s,a)] = 0$

For any policy π and state s:

$$\mathbb{E}_{a\sim\pi(.|s)}[A_\pi(s,a)] = \mathbb{E}_{a\sim\pi(.|s)}[Q_\pi(s,a) - V_\pi(s)] \tag{72}$$
$$= \mathbb{E}_{a\sim\pi(.|s)}[Q_\pi(s,a)] - V_\pi(s) \tag{73}$$
$$= V_\pi(s) - V_\pi(s) = 0 \tag{74}$$

Therefore:

$$\bar{A}(s) = \mathbb{E}_{(a,a')\sim\text{coupling}}[A_\pi(s,a') - A_\pi(s,a)] \tag{75}$$

Step 4: Bound the difference using the coupling property

$$|\bar{A}(s)| = \left|\mathbb{E}_{(a,a')\sim\text{coupling}}[A_\pi(s,a') - A_\pi(s,a)]\right| \tag{76}$$
$$\leq \mathbb{E}_{(a,a')\sim\text{coupling}}[|A_\pi(s,a') - A_\pi(s,a)|] \tag{77}$$

Since $P(a \neq a'|s) \leq \alpha$, we can split the expectation:

$$\mathbb{E}_{(a,a')\sim\text{coupling}}[|A_\pi(s,a') - A_\pi(s,a)|] = P(a = a'|s) \cdot 0 + P(a \neq a'|s) \cdot \mathbb{E}[|A_\pi(s,a') - A_\pi(s,a)||a \neq a'] \tag{78}$$
$$\leq \alpha \cdot 2 \max_{s,a} |A_\pi(s,a)| \tag{79}$$

The factor of 2 comes from the fact that $|A_\pi(s,a') - A_\pi(s,a)| \leq |A_\pi(s,a')| + |A_\pi(s,a)| \leq 2\max_{s,a}|A_\pi(s,a)|$.

Step 5: Final result

Therefore:

$$|\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)| \tag{80}$$

This completes the proof. □

d) Prove the following lemma:

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

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

Solution:

We need to prove that the difference in expected advantage between trajectories sampled from π' and π is bounded.

Step 1: Express the difference

$$|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]| = \left|\sum_{s_t} P(s_t = s|s_0,\pi')\bar{A}(s) - \sum_{s_t} P(s_t = s|s_0,\pi)\bar{A}(s)\right| \tag{81}$$
$$= \left|\sum_{s} [P(s_t = s|s_0,\pi') - P(s_t = s|s_0,\pi)]\bar{A}(s)\right| \tag{82}$$

Step 2: Use the bound from part (c)

From part (c), we have $|\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)|$. Therefore:

$$\left|\sum_s [P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)]\bar{A}(s)\right| \leq \sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \cdot |\bar{A}(s)| \tag{83}$$

$$\leq 2\alpha \max_{s,a} |A_\pi(s,a)| \sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \tag{84}$$

Step 3: Bound the total variation distance

The sum $\sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)|$ is exactly twice the total variation distance between the state distributions at time t.

For α-coupled policies, we can show that:

$$\sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \leq 2(1 - (1-\alpha)^t) \tag{85}$$

Step 4: Proof of the total variation bound

We prove this by induction on t.

Base case ($t = 0$): At time 0, both policies start from the same initial state distribution, so the total variation distance is 0.

Inductive step: Assume the bound holds for time $t-1$. At time t:

$$P(s_t = s|s_0, \pi') = \sum_{s_{t-1}} P(s_{t-1} = s'|s_0, \pi') \sum_{a'} \pi'(a'|s') P(s|s', a') \tag{86}$$

$$P(s_t = s|s_0, \pi) = \sum_{s_{t-1}} P(s_{t-1} = s'|s_0, \pi) \sum_{a} \pi(a|s') P(s|s', a) \tag{87}$$

The difference can be bounded by considering the coupling: with probability $(1-\alpha)$, the actions are the same, and with probability α, they differ. This gives:

$$|P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \leq \alpha \cdot \text{bound from } t-1 + \alpha \cdot \text{maximum possible difference} \tag{88}$$

By the inductive hypothesis and careful analysis of the coupling, we get:

$$\sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \leq 2(1 - (1-\alpha)^t) \tag{89}$$

Step 5: Final result

Substituting back:

$$|\mathbb{E}_{s_t \sim \pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t \sim \pi}[\bar{A}(s_t)]| \leq 2\alpha \max_{s,a} |A_\pi(s,a)| \cdot 2(1 - (1-\alpha)^t) \tag{90}$$

$$= 4\alpha(1 - (1-\alpha)^t) \max_{s,a} |A_\pi(s,a)| \tag{91}$$

This completes the proof. □

e) Prove the following lemma:

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

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

Solution:

We need to prove that the difference between the true objective $\eta(\pi')$ and the local approximation $L_\pi(\pi')$ is bounded for α-coupled policies.

Step 1: Express the difference

From equations 54 and 65:

$$\eta(\pi') - L_\pi(\pi') = \left(\eta(\pi) + \sum_s \rho_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a)\right) - \left(\eta(\pi) + \sum_s \rho_\pi(s) \sum_a \pi'(a|s) A_\pi(s,a)\right) \tag{92}$$

$$= \sum_s [\rho_{\pi'}(s) - \rho_\pi(s)] \sum_a \pi'(a|s) A_\pi(s,a) \tag{93}$$

Step 2: Use the definition of $\bar{A}(s)$

Since $\bar{A}(s) = \sum_a \pi'(a|s) A_\pi(s,a)$:

$$\eta(\pi') - L_\pi(\pi') = \sum_s [\rho_{\pi'}(s) - \rho_\pi(s)] \bar{A}(s) \tag{94}$$

Step 3: Bound using previous lemmas

From part (c), we have $|\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)| = 2\alpha\epsilon$.

Therefore:

$$|\eta(\pi') - L_\pi(\pi')| \leq \sum_s |\rho_{\pi'}(s) - \rho_\pi(s)| \cdot |\bar{A}(s)| \tag{95}$$

$$\leq 2\alpha\epsilon \sum_s |\rho_{\pi'}(s) - \rho_\pi(s)| \tag{96}$$

Step 4: Bound the difference in visitation frequencies

The difference in visitation frequencies can be expressed as:

$$\rho_{\pi'}(s) - \rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t [P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)] \tag{97}$$

Therefore:

$$\sum_s |\rho_{\pi'}(s) - \rho_\pi(s)| = \sum_s \left| \sum_{t=0}^{\infty} \gamma^t [P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)] \right| \tag{98}$$

$$\leq \sum_s \sum_{t=0}^{\infty} \gamma^t |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \tag{99}$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \tag{100}$$

Step 5: Use the bound from part (d)

From part (d), we have:

$$\sum_s |P(s_t = s|s_0, \pi') - P(s_t = s|s_0, \pi)| \leq 2(1 - (1 - \alpha)^t) \tag{101}$$

Therefore:

$$\begin{aligned}
\sum_s |\rho_{\pi'}(s) - \rho_\pi(s)| &\leq \sum_{t=0}^{\infty} \gamma^t \cdot 2(1 - (1 - \alpha)^t) && (102)\\
&= 2\sum_{t=0}^{\infty} \gamma^t (1 - (1 - \alpha)^t) && (103)\\
&= 2\left(\sum_{t=0}^{\infty} \gamma^t - \sum_{t=0}^{\infty} \gamma^t (1 - \alpha)^t\right) && (104)\\
&= 2\left(\frac{1}{1-\gamma} - \frac{1}{1 - \gamma(1-\alpha)}\right) && (105)\\
&= 2\left(\frac{1 - \gamma(1-\alpha) - (1-\gamma)}{(1-\gamma)(1-\gamma(1-\alpha))}\right) && (106)\\
&= 2\left(\frac{\gamma\alpha}{(1-\gamma)(1-\gamma(1-\alpha))}\right) && (107)\\
&\leq 2\left(\frac{\gamma\alpha}{(1-\gamma)^2}\right) && (108)\\
&= \frac{2\gamma\alpha}{(1-\gamma)^2} && (109)
\end{aligned}$$

The inequality in the last step uses the fact that $1 - \gamma(1 - \alpha) \geq 1 - \gamma$ for $\alpha \geq 0$.

Step 6: Final result

Substituting back:

$$\begin{aligned}
|\eta(\pi') - L_\pi(\pi')| &\leq 2\alpha\epsilon \cdot \frac{2\gamma\alpha}{(1-\gamma)^2} && (110)\\
&= \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} && (111)
\end{aligned}$$

This completes the proof. □

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-coupled policy pair (π, π') with appropriate marginals.[2]

Solution:

We need to prove theorem 2.1:

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi, \pi')$$

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

where $\epsilon = \max_{s,a} |A_\pi(s,a)|$.

Step 1: Use the hint to construct an α-coupled policy pair

From the hint, if $D_{TV}^{\max}(\pi, \pi') \leq \alpha$, then we can define an α-coupled policy pair (π, π') with appropriate marginals.

Let $\alpha = D_{TV}^{\max}(\pi, \pi')$. Then (π, π') is an α-coupled policy pair.

Step 2: Apply lemma from part (e)

From part (e), we have:

$$|\eta(\pi') - L_\pi(\pi')| \leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} \tag{112}$$

This gives us:

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} \tag{113}$$

Step 3: Relate α to KL divergence

We need to relate $\alpha = D_{TV}^{\max}(\pi, \pi')$ to $D_{KL}^{\max}(\pi, \pi')$.

From the given inequality $D_{TV}(p||q)^2 \leq D_{KL}(p||q)$, we have:

$$D_{TV}^{\max}(\pi, \pi')^2 \leq D_{KL}^{\max}(\pi, \pi') \tag{114}$$

Therefore:

$$\alpha^2 = D_{TV}^{\max}(\pi, \pi')^2 \leq D_{KL}^{\max}(\pi, \pi') \tag{115}$$

Step 4: Substitute and complete the proof

Substituting $\alpha^2 \leq D_{KL}^{\max}(\pi, \pi')$ into the bound from Step 2:

$$\eta(\pi') \geq L_\pi(\pi') - \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} \tag{116}$$

$$\geq L_\pi(\pi') - \frac{4D_{KL}^{\max}(\pi, \pi')\gamma\epsilon}{(1-\gamma)^2} \tag{117}$$

This completes the proof of theorem 2.1. □

Interpretation:

This theorem provides a lower bound on the true performance improvement $\eta(\pi')$ in terms of: 1. The local approximation $L_\pi(\pi')$ (which is easier to optimize) 2. A penalty term proportional to the KL divergence between policies

The penalty term $\frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi, \pi')$ ensures that: - If the policies are very different ($D_{KL}^{\max}$ is large), the bound becomes loose - If the policies are similar ($D_{KL}^{\max}$ is small), the bound is tight - The bound becomes tighter as $\gamma \to 0$ (shorter horizon) or $\epsilon \to 0$ (smaller advantages)

This justifies the TRPO algorithm's approach of constraining policy updates to maintain the trust region property.

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:

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