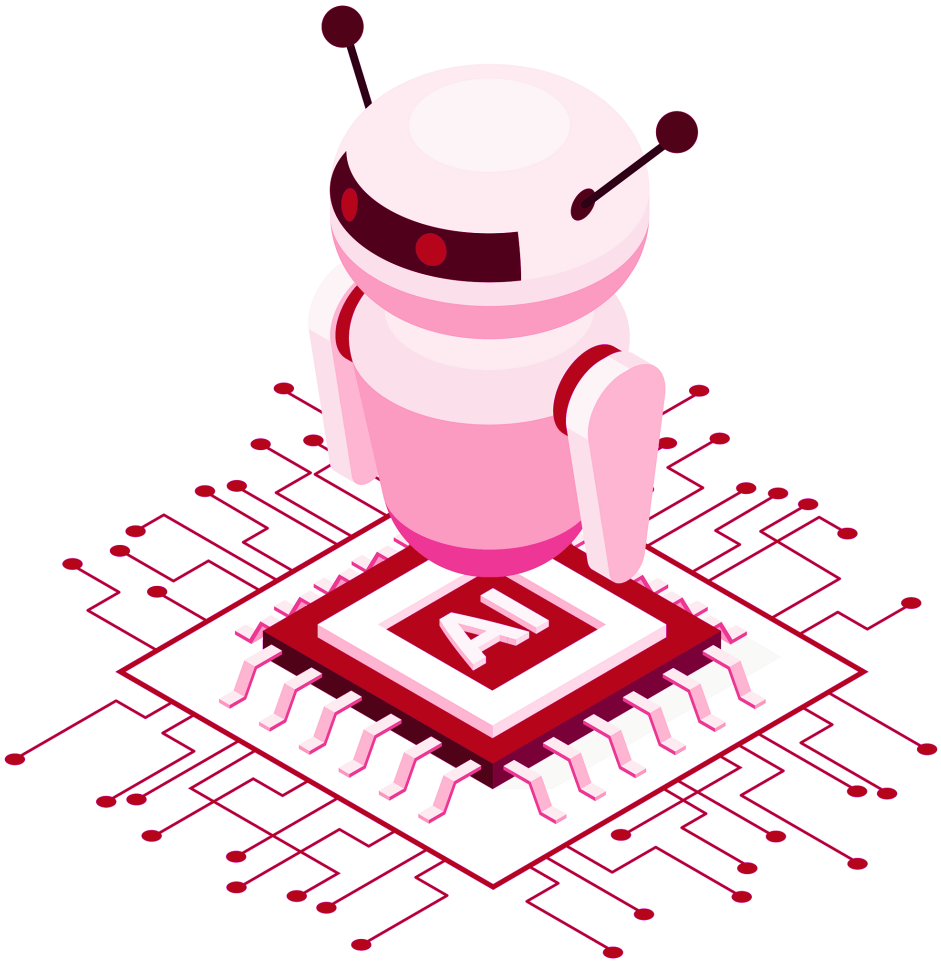

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 8:

Policy-Based Theory

By:

Ayeen Poostforoushan

401105742

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:

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

As we know,

$$V^{\pi_\theta}(s_0) = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t) \;\Big|\; s_0\Big].$$

So

$$\nabla_\theta V^{\pi_\theta}(s_0) = \nabla_\theta \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t) \;\Big|\; s_0\Big].$$

we do the log trick:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t)\; \nabla_\theta \log p_\theta(\tau) \;\Big|\; s_0\Big].$$

also we know that by factorizing the probability of the trajectory, conditioned on s_0, we have:

$$p_\theta(\tau \mid s_0) = \prod_{t=0}^{\infty} \pi_\theta(a_t \mid s_t)\; P(s_{t+1} \mid s_t, a_t).$$

so we have:

$$\nabla_\theta \log p_\theta(\tau \mid s_0) = \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t).$$

therefore:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t)\; \sum_{u=0}^{\infty} \nabla_\theta \log \pi_\theta(a_u \mid s_u) \;\Big|\; s_0\Big].$$

As we studied before, we know that in the double sum, each $\nabla_\theta \log \pi(a_u|s_u)$ only affects rewards from time u onward, so

$$\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) \;\times\; \sum_{u=0}^{\infty} \nabla_\theta \log \pi(a_u|s_u) = \sum_{u=0}^{\infty} \nabla_\theta \log \pi(a_u|s_u) \;\times\; \sum_{k=0}^{\infty} \gamma^{u+k}\, r(s_{u+k}, a_{u+k}).$$

showing the equivalence and then to Q^{π_θ}:

$$\sum_{k=0}^{\infty} \gamma^k\, r(s_{u+k}, a_{u+k}) = Q^{\pi_\theta}(s_u, a_u).$$

now we open the expectation with summation, collect each (s, a) and factor out $Q^{\pi_\theta}(s, a)$, then note that each (s, a) appears with weight $(1-\gamma)\gamma^u$ under the expectation, giving the discounted visitation. Here's the proof:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \mathbb{E}\Big[\sum_{u=0}^{\infty} \gamma^u\, \nabla_\theta \log \pi_\theta(a_u \mid s_u)\, Q^{\pi_\theta}(s_u, a_u) \;\Big|\; s_0\Big].$$

open the expectation with the sum:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_{u=0}^{\infty} \gamma^u\, \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a_u \mid s_u)\, Q^{\pi_\theta}(s_u, a_u) \;\Big|\; s_0\Big].$$

now write that inner expectation as a sum over all (s, a):

$$\mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a_u \mid s_u)\, Q^{\pi_\theta}(s_u, a_u) \;\Big|\; s_0\Big] = \sum_s \sum_a \Pr(s_u = s,\, a_u = a \mid s_0)\, \nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s, a).$$

factor out the $Q^{\pi_\theta}(s, a)$ since it does not depend on u:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \sum_s \sum_a \nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s, a) \sum_{u=0}^{\infty} \gamma^u\, \Pr(s_u = s,\, a_u = a \mid s_0).$$

but $\Pr(s_u = s,\, a_u = a \mid s_0) = \Pr(s_u = s \mid s_0)\, \pi_\theta(a \mid s)$. So

$$\sum_{u=0}^{\infty} \gamma^u\, \Pr(s_u = s,\, a_u = a \mid s_0) = \pi_\theta(a \mid s) \sum_{u=0}^{\infty} \gamma^u \Pr(s_u = s \mid s_0).$$

define the discounted visitation, $d_{s_0}^{\pi_\theta}(s) = (1-\gamma)\sum_{u=0}^{\infty}\gamma^u \Pr(s_u = s \mid s_0)$, so $\sum_{u=0}^{\infty}\gamma^u \Pr(s_u = s \mid s_0) = \frac{d_{s_0}^{\pi_\theta}(s)}{1-\gamma}$. plug back in:

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

recognize the double sum as an expectation:

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

And now we have reached the required form 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.

We begin with the exact policy gradient:

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

Since ϕ minimizes $\epsilon = \mathbb{E}_{s,a}[(Q^{\pi_\theta}(s,a) - Q_\phi(s,a))^2]$, its gradient on ϕ becomes zero:

$$0 = \nabla_\phi \epsilon = \mathbb{E}_{s,a}\big[2\,(Q_\phi(s,a) - Q^{\pi_\theta}(s,a))\,\nabla_\phi Q_\phi(s,a)\big].$$

By the equation we assumed, $\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a \mid s)$, this becomes

$$\mathbb{E}_{s,a}\big[(Q_\phi(s,a) - Q^{\pi_\theta}(s,a))\,\nabla_\theta \log \pi_\theta(a \mid s)\big] = 0.$$

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

Rearrange:

$$\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a)\big] = \mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, Q_\phi(s,a)\big].$$

So we have:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, Q_\phi(s,a)\big],$$

and the policy gradient using the learned critic is exact.

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:

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

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:

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

a) Prove the following identity:

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

We start from

$$\eta(\pi') = \mathbb{E}_{s_0\sim\rho_0}\left[V^{\pi'}(s_0)\right] = \mathbb{E}_{\tau\sim\pi'}\left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t)\right],$$

Since ρ_0 is the same initial-state distr. under any π', we replaced the outer s_0-expectation by the full trajectory expectation under.

We can replace expectation over s_0 distribution with the whole trajectory distribution over any policy π' because the s_0 distribution is still the same and the other parts of the distribution can simply be marginalized out.

Likewise,

$$\eta(\pi) = \mathbb{E}_{s_0 \sim \rho_0}\big[V^{\pi}(s_0)\big] = \mathbb{E}_{\tau \sim \pi'}\big[V^{\pi}(s_0)\big].$$

But

$$V^{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t \, V^{\pi}(s_t) \; - \; \sum_{t=1}^{\infty} \gamma^t \, V^{\pi}(s_t),$$

so

$$\eta(\pi) = \mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t \, V^{\pi}(s_t) \; - \; \sum_{t=1}^{\infty} \gamma^t \, V^{\pi}(s_t)\Big].$$

Subtracting from $\eta(\pi')$ gives

$$\begin{aligned}
\eta(\pi') - \eta(\pi) &= \mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t \, r(s_t) - \big(\sum_{t=0}^{\infty} \gamma^t \, V^{\pi}(s_t) - \sum_{t=1}^{\infty} \gamma^t \, V^{\pi}(s_t)\big)\Big] \\
&= \mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t) + \gamma \, V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\big)\Big] \\
&= \mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t \, A_{\pi}(s_t, a_t)\Big].
\end{aligned}$$

Rearranging gives $\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau \sim \pi'}[\sum_{t=0}^{\infty} \gamma^t A_{\pi}(s_t, a_t)]$, as claimed.

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:

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

b) Prove the following identity:

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

Starting from

$$\eta(\pi') = \eta(\pi) + \mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t A_{\pi}(s_t, a_t)\Big],$$

we open the trajectory expectation into an expectation over states and then actions:

$$\mathbb{E}_{\tau \sim \pi'}\Big[\sum_{t=0}^{\infty} \gamma^t A_{\pi}(s_t, a_t)\Big] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim \pi'} \, \mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\big[A_{\pi}(s_t, a_t)\big].$$

Then we swap the order of summation over t and the sum over states:

$$\sum_{t=0}^{\infty} \gamma^t \sum_{s} \Pr(s_t = s \mid \pi') \, \mathbb{E}_{a \sim \pi'(\cdot|s)}[A_{\pi}(s, a)] = \sum_{s} \Big(\sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \mid \pi')\Big)\Big(\sum_{a} \pi'(a \mid s) A_{\pi}(s, a)\Big),$$

Notice that $\sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \mid \pi') = \rho_{\pi'}(s)$ by definition. So we finally have:

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

which is exactly (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.

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,*

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

We have

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

and also as we know that Value function is the expected of Q function over the policy distribution, the expected value

$$0 = \mathbb{E}_{a\sim\pi(\cdot|s)}\big[A_\pi(s, a)\big] = \sum_a \pi(a \mid s)\, A_\pi(s, a).$$

Subtracting gives

$$\bar{A}(s) = \sum_a \big[\pi'(a \mid s) - \pi(a \mid s)\big]\, A_\pi(s, a).$$

Taking abs of the equation and pulling out the maximum advantage:

$$\big|\bar{A}(s)\big| \leq \max_a \big|A_\pi(s, a)\big| \sum_a \big|\pi'(a \mid s) - \pi(a \mid s)\big|.$$

By definition of total variation distance,

$$\sum_a \big|\pi'(a \mid s) - \pi(a \mid s)\big| = 2\, D_{\mathrm{TV}}\big(\pi(\cdot \mid s),\, \pi'(\cdot \mid s)\big) \leq 2\,\alpha.$$

So we have the desired result:

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

d) Prove the following lemma:

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

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

We start from

$$\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| = \left|\sum_s \big(P_{\pi'}(s_t=s) - P_\pi(s_t=s)\big)\,\bar{A}(s)\right|.$$

Pulling out $\max_{s,a}|\bar{A}(s)|$ gives

$$\left|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| \le 2\alpha \max_{s,a}|\bar{A}(s)| \sum_s \left|P_{\pi'}(s_t=s) - P_\pi(s_t=s)\right|.$$

Now decompose $P_{\pi'}(s_t = s)$ into the "no-mistake" and "at-least-one-mistake" cases. Let α be the per-step probability of π' choosing a different action than π. Then

$$P_{\pi'}(s_t=s) = (1-\alpha)^t\, P_{\text{noMistake}}(s_t=s) + \big[1-(1-\alpha)^t\big]\, P_{\text{mistake}}(s_t=s).$$

On the "no-mistake" event, π' and π produce identical state-distributions, so

$$P_{\text{noMistake}}(s_t=s) = P_\pi(s_t=s).$$

Therefore

$$P_{\pi'}(s_t=s) - P_\pi(s_t=s) = \big[1-(1-\alpha)^t\big]\big(P_{\text{mistake}}(s_t=s) - P_\pi(s_t=s)\big).$$

Taking absolute values and summing over s,

$$\sum_s \left|P_{\pi'}(s_t=s) - P_\pi(s_t=s)\right| \;=\; \big[1-(1-\alpha)^t\big] \sum_s \left|P_{\text{mistake}}(s_t=s) - P_\pi(s_t=s)\right|.$$

Since any two probability distributions have total variation at most 2,

$$\sum_s \left|P_{\text{mistake}}(s_t=s) - P_\pi(s_t=s)\right| \le 2.$$

Then we have:

$$\sum_s \left|P_{\pi'}(s_t=s) - P_\pi(s_t=s)\right| \le 2\big[1-(1-\alpha)^t\big].$$

So we finally have:

$$|\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]| \le 2\alpha \max_{s,a}|\bar{A}(s)| \;\times\; 2\big[1-(1-\alpha)^t\big] = 4\alpha\big(1-(1-\alpha)^t\big) \max_{s,a}|\bar{A}(s)|$$

as required.

e) Prove the following lemma:

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

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

We have

$$\left|\eta(\pi') - L_\pi(\pi')\right| = \left|\sum_{t=0}^{\infty} \gamma^t \big(\mathbb{E}_{s_t\sim\pi'}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\big)\right|.$$

Apply the bound from part (d):

$$\left|\mathbb{E}_{\pi'}[\bar{A}(s_t)] - \mathbb{E}_\pi[\bar{A}(s_t)]\right| \le 4\alpha\big(1-(1-\alpha)^t\big)\,\epsilon.$$

Thus

$$\begin{aligned}
\big|\eta(\pi') - L_\pi(\pi')\big| &\le \sum_{t=0}^{\infty} \gamma^t \, 4\alpha\big(1-(1-\alpha)^t\big)\,\epsilon \\
&= 4\alpha\,\epsilon \sum_{t=0}^{\infty} \big(\gamma^t - \gamma^t(1-\alpha)^t\big) \\
&= 4\alpha\,\epsilon\Big(\sum_{t=0}^{\infty}\gamma^t \;-\; \sum_{t=0}^{\infty}(\gamma(1-\alpha))^t\Big) \\
&= 4\alpha\,\epsilon\Big(\tfrac{1}{1-\gamma} \;-\; \tfrac{1}{1-\gamma(1-\alpha)}\Big) \\
&= 4\alpha\,\epsilon\,\frac{\gamma\alpha}{(1-\gamma)\big(1-\gamma(1-\alpha)\big)} \\
&= 4\alpha^2\gamma\,\epsilon\,\frac{1}{(1-\gamma)\big(1-\gamma+\gamma\alpha\big)}.
\end{aligned}$$

Since $0 \le \alpha \le 1$ implies $1-\gamma+\gamma\alpha \ge 1-\gamma$, we finally have the needed inequality:

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

f) Prove theorem 2.1. Hint: Use the fact that if we have two policies π and π' such that $D_{TV}^{\max}(\pi,\pi') \le \alpha$, then we can define an α-couples policy pair (π,π') with appropriate marginals.[2]

Let $\alpha = D_{TV}^{\max}(\pi,\pi')$. By the hint, there exists an α-coupled policy pair (π,π'). From Lemma (e) we have

$$\big|\eta(\pi') - L_\pi(\pi')\big| \;\le\; \frac{4\,\alpha^2\,\gamma\,\epsilon}{(1-\gamma)^2}.$$

Since for any two distributions p, q, $D_{TV}(p\|q)^2 \le D_{KL}(p\|q)$, it follows that $\alpha^2 \le D_{KL}^{\max}(\pi,\pi')$. Therefore

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

which immediately gives the claimed bound:

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

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:

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

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

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:

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