1 Appendix

1.1 Proof

Let's start with some useful lemmas.

Lemma 1.1. *((Kakade & Langford, 2002)) Consider any two policies $\hat{\pi}$ and π, we have*

$$\eta(\pi) - \eta(\hat{\pi}) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho^{\pi}} A^{\hat{\pi}}(s,a).$$

Corollary 1.1. *Consider any two policies $\hat{\pi}$ and π, we have*

- $V^{\pi}(s_0) - V^{\hat{\pi}}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot|s_0)} A^{\hat{\pi}}(s,a)$.
- $Q^{\pi}(s_0,a_0) - Q^{\hat{\pi}}(s_0,a_0) = \frac{\gamma}{1-\gamma}\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot|s_0,a_0)} A^{\hat{\pi}}(s,a)$.

Proof. The first formula is simple, due to $\eta(\pi) = \mathbb{E}_{s_0\sim\rho_0} V^{\pi}(s_0)$.

Let's prove the second formula.

$$\begin{aligned}
&Q^{\pi}(s_0,a_0) - Q^{\hat{\pi}}(s_0,a_0) \\
=&\gamma\mathbb{E}_{s'\sim P(s'|s_0,a_0)}\left[V^{\pi}(s') - V^{\hat{\pi}}(s')\right] \\
=&\frac{\gamma}{1-\gamma}\mathbb{E}_{s'\sim P(s'|s_0,a_0)}\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot|s')} A^{\hat{\pi}}(s,a) \\
=&\frac{\gamma}{1-\gamma}\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot|s_0,a_0)} A^{\hat{\pi}}(s,a).
\end{aligned}$$

□

Lemma 1.2. *((Tomczak et al., 2019)) Consider any two policies $\hat{\pi}$ and π, we have*

$$\eta(\pi) - \eta(\hat{\pi}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho^{\hat{\pi}},a\sim\pi} A^{\hat{\pi}}(s,a) + \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho^{\hat{\pi}}}\left[\frac{\pi(a|s)}{\hat{\pi}(a|s)} - 1\right]\left[Q^{\pi}(s,a) - Q^{\hat{\pi}}(s,a)\right]$$

Lemma 1.3. *Consider a current policy $\hat{\pi}$, and any policies π, we have*

$$\begin{aligned}
&\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot)} A^{\hat{\pi}}(s,a) - \mathbb{E}_{s\sim\rho^{\hat{\pi}},a\sim\pi} A^{\hat{\pi}}(s,a) \\
=&\frac{\gamma}{1-\gamma}\underset{\substack{s,a\sim\rho^{\hat{\pi}}(\cdot) \\ s',a'\sim\rho^{\pi}(\cdot|s,a)}}{\mathrm{E}}\left[\frac{\pi(a|s)}{\hat{\pi}(a|s)} - 1\right] A^{\hat{\pi}}(s',a')
\end{aligned}$$

Proof. From Lemma 1.1 and 1.2, we have

$$\begin{aligned}
&\mathbb{E}_{s,a\sim\rho^{\pi}} A^{\hat{\pi}}(s,a) - \mathbb{E}_{s\sim\rho^{\hat{\pi}},a\sim\pi} A^{\hat{\pi}}(s,a) \\
=&\mathbb{E}_{s,a\sim\rho^{\hat{\pi}}}\left[\frac{\pi(a|s)}{\hat{\pi}(a|s)} - 1\right]\left[Q^{\pi}(s,a) - Q^{\hat{\pi}}(s,a)\right]
\end{aligned}$$

According to Corollary 1.1, it is easy to get the conclusion. □

Theorem 1.1. *Consider a current policy $\hat{\pi}$, and any policies π, we have*

$$\eta(\pi) = \eta(\hat{\pi}) + \sum_{i=0}^{k-1}\alpha_i L_i(\pi,\hat{\pi}) + \beta_k G_k(\pi,\hat{\pi})$$

where

$$L_i(\pi,\hat{\pi}) = \underset{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ \cdots \\ s_{i-1},a_{i-1}\sim\rho^{\hat{\pi}}(\cdot|s_{i-2},a_{i-2})}}{\mathrm{E}} \prod_{t=0}^{i-1}(r_t-1)l_i(\pi,\hat{\pi}),$$

$$G_k(\pi,\hat{\pi}) = \underset{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ \cdots \\ s_{k-1},a_{k-1}\sim\rho^{\hat{\pi}}(\cdot|s_{k-2},a_{k-2})}}{\mathrm{E}} \prod_{t=0}^{k-1}(r_t-1)g_k(\pi,\hat{\pi}),$$

$$l_i(\pi,\hat{\pi}) = \mathbb{E}_{s_i\sim\rho^{\hat{\pi}}(\cdot|s_{i-1},a_{i-1}),a_i\sim\pi(\cdot|s_i)}A^{\hat{\pi}}(s_i,a_i),$$

$$g_k(\pi,\hat{\pi}) = \mathbb{E}_{s_k,a_k\sim\rho^{\pi}(\cdot|s_{k-1},a_{k-1})}A^{\hat{\pi}}(s_k,a_k),$$

and

$$r_t = \frac{\pi(a_t|s_t)}{\hat{\pi}(a_t|s_t)},\ \alpha_i = \frac{\gamma^i}{(1-\gamma)^{i+1}},\ \beta_k = \frac{\gamma^k}{(1-\gamma)^{k+1}}.$$

Proof. From Lemma 1.3, this formula creates a link between $\mathbb{E}_{s,a\sim\rho^{\pi}(\cdot)}A^{\hat{\pi}}(s,a)$ and $\mathbb{E}_{s',a'\sim\rho^{\pi}(\cdot|s,a)}A^{\hat{\pi}}(s',a')$, resulting in a recursive relationship.

According to Lemma 1.2, and using recursive relationships, defined

$$l_i(\pi,\hat{\pi}) = \mathbb{E}_{s_i\sim\rho^{\hat{\pi}}(\cdot|s_{i-1},a_{i-1}),a_i\sim\pi(\cdot|s_i)}A^{\hat{\pi}}(s_i,a_i),$$

we have

$$
\begin{aligned}
&\eta(\pi)-\eta(\hat{\pi})\\
=&\frac{1}{1-\gamma}\mathbb{E}_{s_0\sim\rho^{\hat{\pi}},a_0\sim\pi}A^{\hat{\pi}}(s_0,a_0) + \frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}[r_0-1]\mathbb{E}_{s_1,a_1\sim\rho^{\pi}(\cdot|s_0,a_0)}A^{\hat{\pi}}(s_1,a_1)\\
=&\frac{1}{1-\gamma}l_0(\pi,\hat{\pi})\\
&+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}[r_0-1]\left(l_1(\pi,\hat{\pi})+\frac{\gamma}{1-\gamma}\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}[r_1-1]\mathbb{E}_{s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}A^{\hat{\pi}}(s_2,a_2)\right)\\
=&\frac{1}{1-\gamma}l_0(\pi,\hat{\pi})+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}[r_0-1]l_1(\pi,\hat{\pi})\\
&+\frac{\gamma^2}{(1-\gamma)^3}\underset{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)\\ s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}}{\mathrm{E}}[r_0-1][r_1-1]A^{\hat{\pi}}(s_2,a_2)\\
&\cdots\\
=&\sum_{i=0}^{k-1}\alpha_i L_i(\pi,\hat{\pi})+\beta_k G_k(\pi,\hat{\pi})
\end{aligned}
$$

□

Corollary 1.2. *According to the definition of G_k, we have*

$$|\beta_k G_k(\pi,\hat{\pi})| \le \frac{\gamma^k}{(1-\gamma)^{k+2}}\epsilon^{k+1}R_{\max},$$

where $\epsilon \triangleq \|\pi-\hat{\pi}\|_1 = \max_s\sum_a|\pi(a|s)-\hat{\pi}(a|s)|$ and $R_{\max}\triangleq\max_{s,a}|R(s,a)|$.

Proof. According to the definition of $G_k(\pi,\hat{\pi})$, and defined $\epsilon\triangleq\|\pi-\hat{\pi}\|_1$, we have

$$
\begin{aligned}
|G_k(\pi,\hat{\pi})| &\le \epsilon^k\cdot|\mathbb{E}_{s_k,a_k\sim\rho^{\pi}(\cdot|s_{k-1},a_{k-1})}A^{\hat{\pi}}(s_k,a_k)|\\
&\le \epsilon^k\cdot|\int_a(\pi-\hat{\pi})Q^{\hat{\pi}}(s,a)da|\\
&\le \frac{R_{\max}}{1-\gamma}\epsilon^{k+1}
\end{aligned}
$$

Combining with β_k, we can get this conclusion. $\square$

Corollary 1.3. *Compared with Theorem 2 of the paper (Tang et al., 2020), we give a tighter lower bound.*

Proof. From the paper (Tang et al., 2020), they give the gap between the policy performance of π and the general surrogate object

$$\hat{G}_k = \frac{1}{\gamma(1-\gamma)}\left(1 - \frac{\gamma}{1-\gamma}\epsilon\right)^{-1}\left(\frac{\gamma\epsilon}{1-\gamma}\right)^{K+1} R_{\max}$$

Next, from Corollary 1.2, we will prove that the following inequality holds

$$\frac{\gamma^k}{(1-\gamma)^{k+2}}\epsilon^{k+1}R_{\max} < \hat{G}_k.$$

That is, we need to prove

$$\frac{\gamma^k}{(1-\gamma)^{k+2}}\epsilon^{k+1}R_{\max} < \frac{1}{\gamma(1-\gamma)}\left(1 - \frac{\gamma}{1-\gamma}\epsilon\right)^{-1}\left(\frac{\gamma\epsilon}{1-\gamma}\right)^{K+1} R_{\max}$$

After simplification, we get

$$\frac{1}{1-\gamma} < \frac{1}{1-\gamma-\gamma\epsilon}.$$

The inequality obviously holds. So, we give a tighter lower bound. $\square$

Theorem 1.2. *Consider a current policy $\hat{\pi}$, and any policies π, we have*

$$\eta(\pi) - \eta(\hat{\pi}) \geq \sum_{i=0}^{k-1} \alpha_i \hat{L}_i(\pi, \hat{\pi}) - \hat{C}_k(\pi, \hat{\pi})$$

where

$$\hat{L}_i(\pi, \hat{\pi}) = \mathop{\mathrm{E}}_{\substack{s_0, a_0 \sim \rho^{\hat{\pi}}(\cdot) \\ \cdots \\ s_{i-1}, a_{i-1} \sim \rho^{\hat{\pi}}(\cdot|s_{i-2}, a_{i-2}) \\ s_i, a_i \sim \rho^{\hat{\pi}}(\cdot|s_{i-1}, a_{i-1})}} \prod_{t=0}^{i} r_t A^{\hat{\pi}}(s_i, a_i),$$

$$\hat{C}_k(\pi, \hat{\pi}) = \frac{\gamma R_{\max} I_{k\geq 2}}{(1-\gamma)^2(1-2\gamma)}\left(1 - \frac{\gamma^k}{(1-\gamma)^k}\right)\|\pi - \hat{\pi}\|_1 + \frac{\gamma^k R_{\max}}{(1-\gamma)^{k+2}}\|\pi - \hat{\pi}\|_1^2$$

and $I_{k\geq 2}$ is the indicator function w.r.t. $k \in N$, $\alpha_i = \frac{\gamma^i}{(1-\gamma)^{i+1}}$.

Proof. For the definition of $L_i(\pi,\hat{\pi})$, we have

$$
\begin{aligned}
&\eta(\pi)-\eta(\hat{\pi})\\
=&\frac{1}{1-\gamma}\mathbb{E}_{s_0\sim\rho^{\hat{\pi}},a_0\sim\pi}A^{\hat{\pi}}(s_0,a_0)+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}[r_0-1]\mathbb{E}_{s_1,a_1\sim\rho^{\pi}(\cdot|s_0,a_0)}A^{\hat{\pi}}(s_1,a_1)\\
=&\frac{1}{1-\gamma}l_0(\pi,\hat{\pi})-\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}A^{\hat{\pi}}(s_1,a_1)\\
&+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}r_0\left(l_1(\pi,\hat{\pi})+\frac{\gamma}{1-\gamma}\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}[r_1-1]\mathbb{E}_{s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}A^{\hat{\pi}}(s_2,a_2)\right)\\
=&\frac{1}{1-\gamma}l_0(\pi,\hat{\pi})-\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}A^{\hat{\pi}}(s_1,a_1)\\
&+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}r_0l_1(\pi,\hat{\pi})-\frac{\gamma^2}{(1-\gamma)^3}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}r_0\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}\mathbb{E}_{s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}A^{\hat{\pi}}(s_2,a_2)\\
&+\frac{\gamma^2}{(1-\gamma)^3}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}r_0\mathbb{E}_{s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)}r_1\mathbb{E}_{s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}A^{\hat{\pi}}(s_2,a_2)\\
=&\frac{1}{1-\gamma}l_0(\pi,\hat{\pi})+\frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s_0,a_0\sim\rho^{\hat{\pi}}}[r_0-1]l_1(\pi,\hat{\pi})\\
&+\frac{\gamma^2}{(1-\gamma)^3}\mathop{\mathrm{E}}_{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ s_1,a_1\sim\rho^{\hat{\pi}}(\cdot|s_0,a_0)\\ s_2,a_2\sim\rho^{\pi}(\cdot|s_1,a_1)}}[r_0-1][r_1-1]A^{\hat{\pi}}(s_2,a_2)\\
&\ldots\\
=&\sum_{i=0}^{k-1}\alpha_i\hat{L}_i(\pi,\hat{\pi})-\sum_{i=1}^{k-1}\alpha_i\hat{H}_i(\pi,\hat{\pi})+\beta_k\hat{G}_k(\pi,\hat{\pi})
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{L}_i(\pi,\hat{\pi})&=\mathop{\mathrm{E}}_{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ \ldots\\ s_{i-1},a_{i-1}\sim\rho^{\hat{\pi}}(\cdot|s_{i-2},a_{i-2})\\ s_i,a_i\sim\rho^{\pi}(\cdot|s_{i-1},a_{i-1})}}\prod_{t=0}^{i}r_tA^{\hat{\pi}}(s_i,a_i),\\
\hat{H}_i(\pi,\hat{\pi})&=\mathop{\mathrm{E}}_{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ \ldots\\ s_{i-1},a_{i-1}\sim\rho^{\hat{\pi}}(\cdot|s_{i-2},a_{i-2})\\ s_i,a_i\sim\rho^{\pi}(\cdot|s_{i-1},a_{i-1})}}\prod_{t=0}^{i-2}r_tA^{\hat{\pi}}(s_i,a_i),\\
\hat{G}_k(\pi,\hat{\pi})&=\mathop{\mathrm{E}}_{\substack{s_0,a_0\sim\rho^{\hat{\pi}}(\cdot)\\ \ldots\\ s_{i-1},a_{i-1}\sim\rho^{\hat{\pi}}(\cdot|s_{i-2},a_{i-2})\\ s_i,a_i\sim\rho^{\pi}(\cdot|s_{i-1},a_{i-1})}}\prod_{t=0}^{i-2}r_t[r_{i-1}-1]A^{\hat{\pi}}(s_i,a_i),
\end{aligned}
$$

and $\alpha_i=\frac{\gamma^i}{(1-\gamma)^{i+1}}$, $\beta_k=\frac{\gamma^k}{(1-\gamma)^{k+1}}$.

It is easy to prove that the following inequality holds

$$\hat{H}_i(\pi,\hat{\pi})\leq\frac{R_{\max}}{1-\gamma}\|\pi-\hat{\pi}\|_1,\ \hat{G}_k(\pi,\hat{\pi})\leq\frac{R_{\max}}{1-\gamma}\|\pi-\hat{\pi}\|_1^2.$$

Since $\sum_{k-1}^{i=0}\alpha_i=\frac{\gamma}{(1-\gamma)(1-2\gamma)}\left(1-\frac{\gamma^k}{(1-\gamma)^k}\right)$, we have

$$\eta(\pi)-\eta(\hat{\pi})\geq\sum_{i=0}^{k-1}\alpha_i\hat{L}_i(\pi,\hat{\pi})-\frac{\gamma R_{\max}I_{k\geq 2}\|\pi-\hat{\pi}\|_1}{(1-\gamma)^2(1-2\gamma)}\left(1-\frac{\gamma^k}{(1-\gamma)^k}\right)-\frac{\gamma^kR_{\max}}{(1-\gamma)^{k+2}}\|\pi-\hat{\pi}\|_1^2$$

□

Theorem 1.3. *Define two sets*

$$\Psi_1 = \left\{ \mu \mid \alpha_0 \hat{L}_0(\mu, \hat{\pi}) - \hat{C}_1(\mu, \hat{\pi}) > 0 \right\},$$
$$\Psi_2 = \left\{ \mu \mid \alpha_0 \hat{L}_0(\mu, \hat{\pi}) + \alpha_1 \hat{L}_1(\mu, \hat{\pi}) - \hat{C}_2(\mu, \hat{\pi}) > 0 \right\},$$

then we have

$$\Psi_2 \subseteq \Psi_1.$$

Proof. Let $\mu \in \Psi_1$, we have

$$\hat{L}_0(\pi, \hat{\pi}) - \frac{\gamma R_{\max}}{(1-\gamma)^2} \|\mu - \hat{\pi}\|_1^2 > 0 \tag{1}$$

Below, we will show that μ may not be in the set Ψ_2.

For $\hat{L}_1(\pi, \hat{\pi})$, we can get

$$\hat{L}_1(\pi, \hat{\pi}) = \mathop{\mathrm{E}}_{\substack{s_0 \sim \rho^{\hat{\pi}}(\cdot), a_0 \sim \pi(\cdot|s_0) \\ s_1 \sim \rho^{\hat{\pi}}(\cdot|s_0, a_0), a_1 \sim \pi(\cdot|s_1)}} A^{\hat{\pi}}(s_1, a_1) \tag{2}$$

$$= \mathop{\mathrm{E}}_{\substack{s_0 \sim \rho^{\hat{\pi}}(\cdot), a_0 \sim \hat{\pi}(\cdot|s_0) \\ s_1 \sim \rho^{\hat{\pi}}(\cdot|s_0, a_0), a_1 \sim \pi(\cdot|s_1)}} A^{\hat{\pi}}(s_1, a_1) + \left(\mathop{\mathrm{E}}_{\substack{s_0 \sim \rho^{\hat{\pi}}(\cdot), a_0 \sim \pi(\cdot|s_0) \\ s_1 \sim \rho^{\hat{\pi}}(\cdot|s_0, a_0), a_1 \sim \pi(\cdot|s_1)}} - \mathop{\mathrm{E}}_{\substack{s_0 \sim \rho^{\hat{\pi}}(\cdot), a_0 \sim \hat{\pi}(\cdot|s_0) \\ s_1 \sim \rho^{\hat{\pi}}(\cdot|s_0, a_0), a_1 \sim \pi(\cdot|s_1)}} \right) A^{\hat{\pi}}(s_1, a_1) \tag{3}$$

$$\geq \mathop{\mathrm{E}}_{s_1 \sim \rho^{\hat{\pi}}(\cdot), a_1 \sim \pi(\cdot|s_1)} A^{\hat{\pi}}(s_1, a_1) - \frac{R_{\max}}{1-\gamma} \|\pi - \hat{\pi}\|_1^2 \tag{4}$$

The last inequality uses $\mathrm{E}_{s_0 \sim \rho^{\hat{\pi}}(\cdot), a_0 \sim \hat{\pi}(\cdot|s_0)} \rho^{\hat{\pi}}(\cdot|s_0, a_0) = \rho^{\hat{\pi}}(\cdot)$ and Hölder's inequality (Finner, 1992).

Combining with $\hat{L}_0(\pi, \hat{\pi})$ and $\hat{C}_2(\pi, \hat{\pi})$, we have

$$\begin{aligned}
&\hat{L}_0(\pi, \hat{\pi}) + \frac{\gamma}{1-\gamma} \hat{L}_1(\pi, \hat{\pi}) - \frac{\gamma R_{\max}}{(1-\gamma)^2} \|\pi - \hat{\pi}\|_1 - \frac{\gamma^2 R_{\max}}{(1-\gamma)^3} \|\pi - \hat{\pi}\|_1^2 \\
\geq & \hat{L}_0(\pi, \hat{\pi}) + \frac{\gamma}{1-\gamma} \left(\mathop{\mathrm{E}}_{s_1 \sim \rho^{\hat{\pi}}(\cdot), a_1 \sim \pi(\cdot|s_1)} A^{\hat{\pi}}(s_1, a_1) - \frac{R_{\max}}{1-\gamma} \|\pi - \hat{\pi}\|_1^2 \right) - \frac{\gamma R_{\max}}{(1-\gamma)^2} \|\pi - \hat{\pi}\|_1 - \frac{\gamma^2 R_{\max}}{(1-\gamma)^3} \|\pi - \hat{\pi}\|_1^2 \\
= & \frac{1}{1-\gamma} \hat{L}_0(\pi, \hat{\pi}) - \frac{\gamma R_{\max}}{(1-\gamma)^3} \|\pi - \hat{\pi}\|_1^2 - \frac{\gamma R_{\max}}{(1-\gamma)^2} \|\pi - \hat{\pi}\|_1
\end{aligned}$$

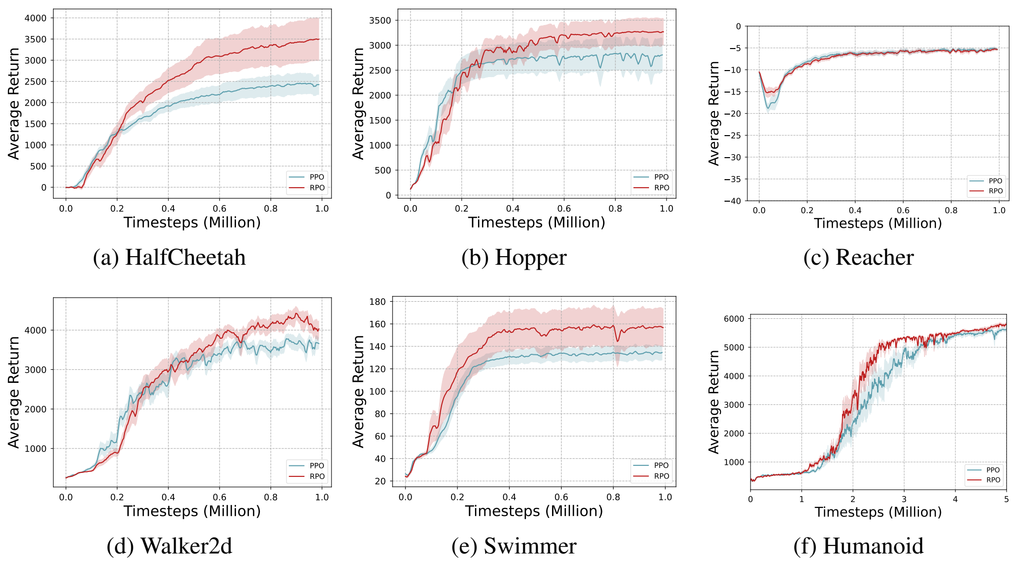
Combining with the inequality (1), we have

$$\hat{L}_0(\mu, \hat{\pi}) + \frac{\gamma}{1-\gamma} \hat{L}_1(\mu, \hat{\pi}) - \frac{\gamma R_{\max}}{(1-\gamma)^2} \|\mu - \hat{\pi}\|_1 - \frac{\gamma^2 R_{\max}}{(1-\gamma)^3} \|\mu - \hat{\pi}\|_1^2 \geq -\frac{\gamma R_{\max}}{(1-\gamma)^2} \|\mu - \hat{\pi}\|_1$$

From the above inequality, it shows that μ may not be in set Ψ_2. So, we have $\Psi_2 \subseteq \Psi_1$. □

1.2 Additional experimental results

To verify the effectiveness of the proposed RPO method, we select six continuous control tasks from the MuJoCo environments (Todorov et al., 2012) in OpenAI Gym (Brockman et al., 2016). We conduct all the experiments mainly based on the code from (Queeney et al., 2021). The test procedures are averaged over ten test episodes across ten independent runs. The same neural network architecture is used for all methods The policy network is a Gaussian distribution, and the output of the state-value network is a scalar value. The mean action of the policy network and state-value network are a multi-layer perceptron with hidden layer fixed to [64, 64] and tanh activation (Henderson et al., 2018). The standard deviation of the policy network is parameterized separately



(a) HalfCheetah (b) Hopper (c) Reacher

(d) Walker2d (e) Swimmer (f) Humanoid

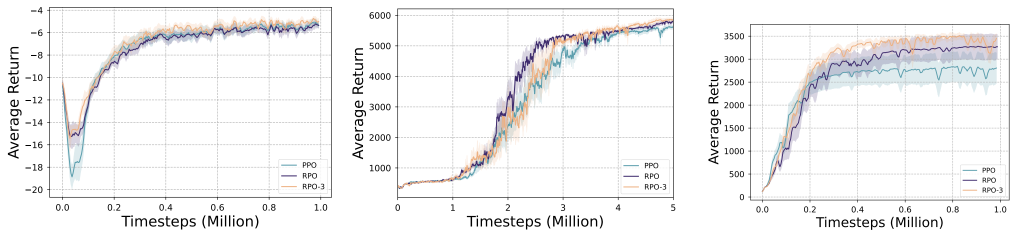
Figure 1: Learning curves on the Gym environments. Performance of RPO vs. PPO.

Table 1: Hyperparameters for RPO on Mujoco tasks.

Hyperparameter	Value
Discount rate γ	0.995
GAE parameter	0.97
Minibatches per epoch	32
Epochs per update	10
Optimizer	Adam
Learning rate ϕ	3e-4
Minimum batch size (n)	2048
ϵ	0.2
ϵ_1	0.1
weighting parameter β	0.3

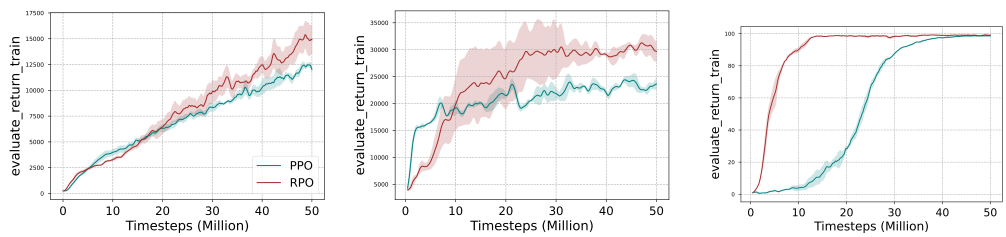
(**?**). For the experimental parameters, we use the default parameters from (Dhariwal et al., 2017; Henderson et al., 2018), for example, the discount factor is $\gamma = 0.995$, and we use the Adam optimizer (Kingma & Ba, 2015) throughout the training progress. For PPO, the clipping parameter is $\epsilon^{\text{PPO}} = 0.2$, and the batch size is $B = 2048$. For GePPO, the clipping parameter is $\epsilon^{\text{PPO}} = 0.1$, and the batch size of each policy is $B = 1024$. For TRPO and off-policy TRPO (OTRPO), the bound of trust region is $\delta = 0.01$, and the batch size of each policy is $B = 1024$.

To verify the effectiveness of the proposed RPO method in discrete environments, we randomly selected twelve Atari games for our experiments and the code is based on (Zhang, 2018). We run our experiments across three seeds with fair evaluation metrics. We use the same hyperparameters $\epsilon_1 = 0.1$ and $\beta = 3.0$ in all environments and do not fine-tune them.

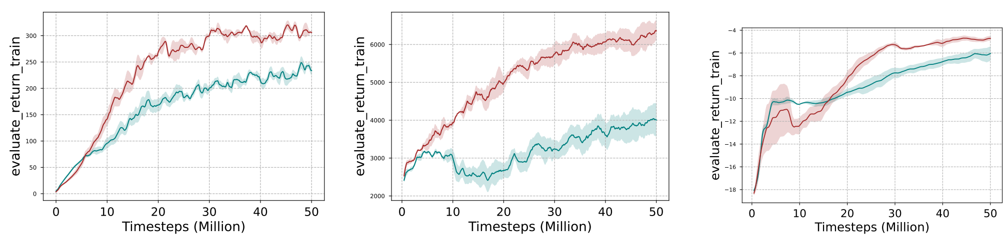


(a) Reacher (b) Humanoid (c) Hopper

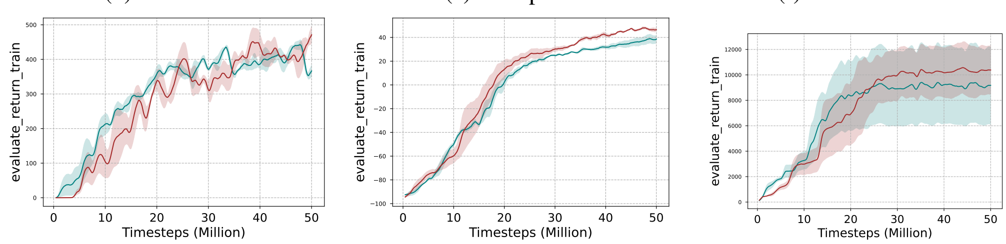
Figure 2: The performance of RPO vs. RPO-3 in three other environments.



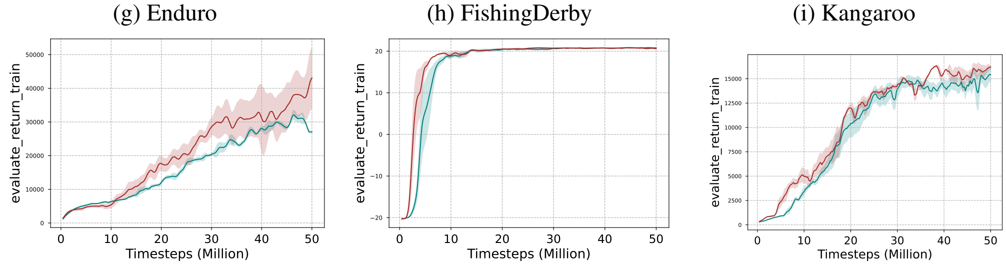
(a) Asterix (b) BattleZone (c) Boxing



(d) Breakout (e) Centipede (f) DoubleDunk



(g) Enduro (h) FishingDerby (i) Kangaroo



(j) Phoenix (k) Pong (l) Qbert

Figure 3: Learning curves on the Atari environments. Performance of RPO vs. PPO.

References

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. `https://github.com/openai/baselines`, 2017.

Helmut Finner. A generalization of holder's inequality and some probability inequalities. *The Annals of Probability*, 20(4):1893–1901, 1992.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI*, pp. 3207–3214. AAAI Press, 2018.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Machine Learning, Proceedings of the Nineteenth International Conference, ICML*, pp. 267–274, 2002.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. URL `http://arxiv.org/abs/1412.6980`.

James Queeney, Yannis Paschalidis, and Christos G. Cassandras. Generalized proximal policy optimization with sample reuse. In *Advances in Neural Information Processing Systems 34, NeurIPS*, pp. 11909–11919, 2021.

Yunhao Tang, Michal Valko, and Rémi Munos. Taylor expansion policy optimization. In *International Conference on Machine Learning*, pp. 9397–9406. PMLR, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 5026–5033. IEEE, 2012.

Marcin B Tomczak, Dongho Kim, Peter Vrancx, and Kee-Eung Kim. Policy optimization through approximate importance sampling. *arXiv preprint arXiv:1910.03857*, 2019.

Shangtong Zhang. Modularized implementation of deep rl algorithms in pytorch. `https://github.com/ShangtongZhang/DeepRL`, 2018.