Average-Reward Reinforcement Learning with Trust Region Methods

Xiaoteng Ma\textsuperscript{1}, Xiaohang Tang\textsuperscript{2}, Li Xia\textsuperscript{3*}, Jun Yang\textsuperscript{1} and Qianchuan Zhao\textsuperscript{1}

\textsuperscript{1}Department of Automation, Tsinghua University
\textsuperscript{2}Department of Statistical Science, University College London
\textsuperscript{3}Business School, Sun Yat-sen University

ma-xt17@mails.tsinghua.edu.cn, xiaohang.tang.20@ucl.ac.uk, xiali5@mail.sysu.edu.cn, yangjun603@mail.tsinghua.edu.cn, zhaqc@tsinghua.edu.cn,

Abstract

Most of reinforcement learning algorithms optimize the discounted criterion which is beneficial to accelerate the convergence and reduce the variance of estimates. Although the discounted criterion is appropriate for certain tasks such as financial related problems, many engineering problems treat future rewards equally and prefer a long-run average criterion. In this paper, we study the reinforcement learning problem with the long-run average criterion. Firstly, we develop a unified trust region theory with discounted and average criteria and derive a novel performance bound within the trust region with the Perturbation Analysis (PA) theory. Secondly, we propose a practical algorithm named Average Policy Optimization (APO), which improves the value estimation with a novel technique named Average Value Constraint. Finally, experiments are conducted in the continuous control environment MuJoCo. In most tasks, APO performs better than the discounted PPO, which demonstrates the effectiveness of our approach. Our work provides a unified framework of the trust region approach including both the discounted and average criteria, which may complement the framework of reinforcement learning beyond the discounted objectives.

1 Introduction

The deep reinforcement learning (DRL) achieves prominent progress in many fields [Silver et al., 2016; Vinyals et al., 2019; Senior et al., 2020]. Most of the modern RL algorithms aim at maximizing $E\left[\sum_{t}^{\infty} \gamma^t r_t\right]$, where $\gamma$ denotes a discount factor strictly less than one. The discount factor undermines the contribution of future rewards on the current value and ignores the long-run gains [Sutton and Barto, 2018], which means that the reward received after $k$ steps in the future is only worth $\gamma^k - 1$ of it to the current state. The existence of discount factor simplifies the theoretical analysis by constructing the contraction mapping and offers a faster and more stable online learning by reducing estimation variances [Kakade, 2001; Marbach and Tsitsiklis, 2003]. However, the discounted criterion is not consistent with the natural metric in many real scenarios, and introduces unnecessary bias of estimates despite lower variance [Thomas, 2014; Schulman et al., 2016]. In empirical implementations, the discount factor serves as a hyper-parameter and the algorithm performance is very sensitive to it [Duan et al., 2016]. Thus, the decision maker should make a compromise between the foresight of the decision and the learning stability.

Different from the discounted criterion, the average criterion directly focuses on the long-run average performance. In the everlasting real-world problems, such as manufacturing, power systems and traffic controls, the average criterion is best suited [Mahadevan, 1996; Dewanto et al., 2020]. Optimizing the average criterion completely removes the discrepancy between the training performance and the evaluating metric due to discounting. Therefore, the average criterion is well explored in the classical MDP literature [Howard, 1960; Puterman, 1994; Cao, 2007]. However, for the data-driven implementation, the average-reward RL algorithms is much less investigated compared with the discounted criterion. It is relatively hard to study the average RL algorithms, since many properties are lost when setting $\gamma = 1$ in the discounted framework. An obvious evidence is that the Bellman operator is no longer a contraction mapping in the average setting [Bertsekas and Tsitsiklis, 1995].

One of the key difficulties in the average-reward RL is to guarantee the values estimated from the sample paths bounded without discounting. Many previous works extended the standard Q-learning to the average setting, such as R-learning [Schwartz, 1993], RVI-learning [Puterman, 1994; Bertsekas et al., 1995], CSV-learning [Yang et al., 2016], and Differential Q-learning [Wan et al., 2021]. In the methods above, a relative value (the average reward estimate, e.g., the reference state value or a constant from prior knowledge) is subtracted from the target value, while they may still cause divergence and lead to unstable training [Dewanto et al., 2020]. Compared with the value-iteration based ones, the policy-iteration based methods, such as policy gradient [Sutton et al., 2000] and Actor-Critic [Konda and Tsitsiklis, 2000], are perhaps more suitable for average reward algorithms due to more stable training and potential for large scale problems. However, they still suffer the instability problem of value estimation in the average setting.

*Corresponding author: Li Xia, Jun Yang
In this paper, we solve the average-reward RL problem by introducing the trust region approach for the average criterion. Based on the Perturbation Analysis (PA) theory [Cao and Chen, 1997; Cao, 2007], we overcome the analysis difficulty and develop a unified trust region theory with discounted and average criteria. Motivated by using the Kemeny’s constant to bound the trust region policy improvement in the average reward case [Zhang and Ross, 2021], we provide a unified performance bound which tightens the previous one [Schulman et al., 2015; Achiam et al., 2017] when $\gamma \rightarrow 1$. The bound also provides a direct evidence that average criterion is preferred than the discounted one for ergodic MDPs. From the view of implementation, we develop a practical algorithm called Average Policy Optimization (APO). We address the difficulty of estimating value with the average criterion and provide a new method named Average Value Constraint to solve the value drifting problem. Our APO algorithm provides a novel implementation for the average DRL algorithm with neural networks as approximators. Finally, our experiments on the continuous control benchmark MuJoCo show that APO can beat the discounted PPO in the metric of average rewards, which further confirms the superiority of our approach.

2 Preliminaries

2.1 Average Reward MDPs

Consider an infinite-horizon ergodic Markov decision process (MDP), defined by the tuple $(S, A, P, r, d_0)$, where $S$ denotes a finite set of states, $A$ denotes a finite set of actions, $P : S \times A \times S \rightarrow [0, 1]$ denotes the transition probability, $r : S \times A \rightarrow \mathbb{R}$ denotes the reward function, $d_0 : S \rightarrow [0, 1]$ denotes the distribution of the initial state $s_0$. Let $\pi : S \times A \rightarrow [0, 1]$ denote a stochastic policy, $\tau$ denote a trajectory $(s_0, a_0, s_1, \ldots)$, and $\pi \sim \tau$ denote that the distribution of trajectories depends on $\pi$: $s_0 \sim d_0, a_t \sim \pi(\cdot | s_t), s_{t+1} \sim P(\cdot | s_t, a_t)$.

With the assumption of ergodicity of MDPs, we aim at maximizing the long-run average performance under $\pi$:

$$\eta_\pi := \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) \right].$$  \hspace{1cm} (1)

We are interested in the steady-state distribution $d_\pi$, which is defined as follows:

$$d_\pi(s) := \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{T-1} P(s_t = s) \right].$$  \hspace{1cm} (2)

Note that since we focus on the contribution of each action within the infinite horizon, the long-run performance is independent of the initial state distribution. Let $V_\pi(s)$ denote the state-value function, $Q_\pi(s, a)$ denote the action-value function, and $A_\pi(s, a)$ denote the advantage function. The value functions are defined by subtracting the long-run average reward $\eta_\pi$ from the current reward to make the functions bounded [Howard, 1960]:

$$V_\pi(s) := \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} (r(s_t, a_t) - \eta_\pi) \mid s_0 = s \right],$$

$$Q_\pi(s, a) := \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} (r(s_t, a_t) - \eta_\pi) \mid s_0 = s, a_0 = a \right],$$

$$A_\pi(s, a) := Q_\pi(s, a) - V_\pi(s).$$  \hspace{1cm} (3)

2.2 Discounted Reward MDPs

With the discounted reward criterion, we focus on the recent rewards by introducing the discount factor $\gamma \in (0, 1)$. We optimize the policy to maximize its normalized expected discounted return:

$$\eta_{\pi, \gamma} := (1 - \gamma) \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 \sim d_0 \right].$$  \hspace{1cm} (4)

The discounted performance $\eta_{\pi, \gamma}$ can be viewed as the average reward under discounted steady-state distribution $d_{\pi, \gamma}$:

$$d_{\pi, \gamma}(s) := (1 - \gamma) \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t P(s_t = s) \mid s_0 \sim d_0 \right].$$  \hspace{1cm} (5)

Different from the average criterion, the policy performance with the discounted criterion is dependent on $d_0$. The discount factor makes a smooth interpolation between $d_0$ and $d_\pi$. To see that, we have $\lim_{\gamma \rightarrow 0} d_{\pi, \gamma} = d_0$ (the bandit case) and $\lim_{\gamma \rightarrow 1} d_{\pi, \gamma} = d_\pi$ (the long-run average reward case).

In the modern DRL literature, the value functions are defined as following:

$$V_{\pi, \gamma}(s) := \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right],$$

$$Q_{\pi, \gamma}(s, a) := \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right],$$

$$A_{\pi, \gamma}(s, a) := Q_{\pi, \gamma}(s, a) - V_{\pi, \gamma}(s).$$  \hspace{1cm} (6)

As the value function is a core concept in RL, people may expect to optimize the long-run performance by letting $\gamma \rightarrow 1$. Unfortunately, many important properties are lost when $\gamma \rightarrow 1$, e.g., value functions will be unbounded which devastate the algorithm performance. Another example is the trust region theory for policy optimization, which we will discuss next.

2.3 Trust Region Method

The trust region methods [Schulman et al., 2015; Schulman et al., 2017; Achiam et al., 2017] are a series of DRL algorithms updating policies with the following approximate policy iteration:

$$\pi_{k+1} = \arg \max_{\pi} \mathbb{E}_{s \sim d_{\pi_k, \gamma}, a \sim \pi} [A_{\pi_k, \gamma}(s, a)]$$  \hspace{1cm} (7)

$$\text{s.t. } D(\pi \parallel \pi_k) \leq \delta,$$

where $D(\pi \parallel \pi_k) = \mathbb{E}_{s \sim d_{\pi_k}} [D(\pi \parallel \pi_k)[s]]$ with $D$ denoting some distance measure and $\delta > 0$ is the step size. The policy set $\{\pi \mid D(\pi \parallel \pi_k) \leq \delta\}$ is called trust region,
in which it is safe to evaluate policies with the samples collected with $\pi_k$. In theoretical analysis, $D$ is often chosen as the total variation divergence defined by $D_{TV}(\pi' \parallel \pi)|s| = \frac{1}{2} \sum_{a \in A} |\pi'(a \mid s) - \pi(a \mid s)|$. In practice, Kullback-Leibler divergence is preferred, which is defined as $D_{KL}(\pi' \parallel \pi) = \sum_{a \in A} \pi'(a \mid s) \log \frac{\pi'(a \mid s)}{\pi(a \mid s)}$. Since $D_{TV}(p \parallel q) \leq \sqrt{D_{KL}(p \parallel q)/2}$, the theoretical analysis is aligned with the practical implementation.

The primary motivation of the trust region method is that optimizing the surrogate objective in the trust region guarantees monotonic performance improvements. When the distance is bounded by $D_{TV}$, the solution of the problem in (7) has the lower performance bound [Achiam et al., 2017]

$$\eta_{\pi_k+1, \gamma} - \eta_{\pi_k, \gamma} \geq \mathbb{E}_{s \sim d_{\pi_k+1, \gamma}, a \sim \pi_k+1} \left[A_{\pi_k, \gamma}(s, a) - \frac{2\gamma \epsilon_{\gamma}}{1 - \gamma} D_{TV}(\pi_k \parallel \pi)\right][s]$$

where $\epsilon_{\gamma} = \max_A |\mathbb{E}_{s \sim d_{\pi_k+1}}[A_{\pi_k+1, \gamma}(s, a)]|$. The advantage of the trust region method majorly comes from two aspects. In each iteration, the samples collected with $\pi_k$ are reused in evaluating any $\pi$ in trust region, which significantly improves the data efficiency compared with the traditional policy gradient methods. The other hand, the performance lower bound avoids collapse after bad updates when neural networks are used as the policy approximator [Duan et al., 2016]. However, existing analysis of the lower performance bound for the trust region method is not applicable when $\gamma \to 1$, which can be verified by $\lim_{\gamma \to 1} \gamma/(1 - \gamma) = \infty$. As the larger discounted factor and average reward setting are important in real problems, more delicate theoretical analysis is needed to understand the behavior of trust region method when $\gamma \to 1$ even $\gamma = 1$.

3 Unified Trust Region Theory with Discounted and Average Criteria

In this section, we propose a unified trust region theory based on the PA theory that extends the existing analysis of performance improvement bound for both the average and discounted criteria. Moreover, we provide a novel policy performance bound for police updating. With the help of this bound, performance monotonicity of policy improvement is guaranteed when $\gamma \to 1$.

3.1 General Formulation

We redefine the discounted value functions in (6) as follows:

$$V_{\pi, \gamma}(s) := \mathbb{E}_{s' \sim p} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s, a_t) - \eta_{\pi}) \mid s_0 = s \right],$$

$$Q_{\pi, \gamma}(s, a) := \mathbb{E}_{s' \sim p} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s, a_t) - \eta_{\pi}) \mid s_0 = s, a_0 = a \right],$$

$$A_{\pi, \gamma}(s, a) := Q_{\pi, \gamma}(s, a) - V_{\pi, \gamma}(s).$$

Although this formulaic approach may be not familiar to DRL researchers, it is commonly used in the MDP literature [Cao, 2007]. The main benefit of this value definition is that it bridges the gap between the discounted and average criteria. With the new definition of value functions, we obtain that $\lim_{\gamma \to 1} V_{\pi, \gamma} = V_{\pi}$ immediately.

Next, we present the performance difference formula for any policies $\pi, \pi'$.

**Lemma 1.** For any policies $\pi, \pi'$, the following equation holds:

$$\eta_{\pi', \gamma} - \eta_{\pi, \gamma} = \mathbb{E}_{s \sim d_{\pi', \gamma}, a \sim \pi'} [A_{\pi, \gamma}(s, a)].$$

This relationship has been shown in many previous studies of trust region method [Kakade and Langford, 2002; Schulman et al., 2015; Achiam et al., 2017]. Actually, (9) can be viewed as a rewriting of the performance difference formula in a sample path form. The performance difference formula is the key result of PA theory [Cao, 2007] and it quantifies the change of the performance of Markov systems corresponding to the change of policies. For the readers unfamiliar with PA, we attach a brief introduction in Appendix A1.

Powered by the performance difference formula, we derive the average performance difference formula at once.

**Corollary 1.** For any policies $\pi, \pi'$, the following equation holds:

$$\eta_{\pi', \gamma} - \eta_{\pi} = \mathbb{E}_{s \sim d_{\pi', \gamma}, a \sim \pi'} [A_{\pi}(s, a)].$$

3.2 Policy Improvement Bound

While the performance difference formula accurately describes the performance difference for two arbitrary policies, it is difficult to use the formula to develop practical algorithms directly. The bottleneck is that to evaluate the performance difference, we need the samples from $d_{\pi', \gamma}$. That is self-contradictory since no trajectory of the new policy is available unless the policy is updated. Thus, instead of optimizing the difference formula directly, the trust region method optimizes the following surrogate objective:

$$L_{\pi, \gamma}(\pi') := \mathbb{E}_{s' \sim d_{\pi', \gamma}, a \sim \pi'} \left[ \frac{\pi'(s, a)}{\pi(s, a)} A_{\pi, \gamma}(s, a) \right],$$

where $d_{\pi', \gamma}$ is replaced by $d_{\pi, \gamma}$ in (9) and the probability of the sampled actions is corrected by importance sampling.

It is natural to query what is the difference between the surrogate objective and the real performance difference. To answer this question, we present the bound as follows.

**Proposition 1.** For any two stochastic policies $\pi, \pi'$, the following bound holds:

$$\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \geq L_{\pi, \gamma}(\pi') - 2\epsilon_{\gamma} D_{TV}(d_{\pi', \gamma} \parallel d_{\pi, \gamma}),$$

$$\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \leq L_{\pi, \gamma}(\pi') + 2\epsilon_{\gamma} D_{TV}(d_{\pi', \gamma} \parallel d_{\pi, \gamma}).$$

The result tells us that if the discounted steady-state distribution of two policies is close enough and the advantage function is bounded, we are safe to use the surrogate objective for policy updating.

Next, we build up the connection between the distance of the discounted steady-state distributions and the distance of policies.
Proposition 2. For any two stochastic policies \( \pi, \pi' \), the following bound holds:
\[
D_{\text{TV}}(d_{\pi'}, \gamma) \leq \xi_\gamma \mathbb{E}_{s \sim d_{\pi}, \gamma}[D_{\text{TV}}(\pi' || \pi)[s]],
\]
where \( \xi_\gamma = \min \{ \frac{\gamma}{1-\gamma}, \frac{\gamma(\kappa-1)}{1-(\gamma-\gamma)\kappa}\} \) and \( \kappa_\pi \) is the Kemeny’s constant of the Markov chain induced by \( \pi \).

The above conclusion is the key contribution of this paper, which reveals that the difference of discounted steady-state distribution is bounded with the distance of the policies multiplied with a constant factor \( \xi_\gamma \). This factor \( \xi_\gamma \) is a time constant that reflects how policy changes affect long-term behaviors of Markov chains. When \( \gamma \) is not very large, the long-run effect of the current decision gradually decreases, allowing decision makers to focus on a few steps after the decision and ignore the impact afterwards. However, if \( \gamma \) is large enough, we cannot use discounts to suppress the long-term impact of the policy anymore. Fortunately, when \( \gamma \rightarrow 1 \), we find that it is still able to bound the long-run impact by introducing Kemeny’s constant [Kemeny and Snell, 1960], which means the average time of returning to the steady-state distribution starting from an arbitrary state. The idea of using Kemeny’s constant to bound the distance of steady-state distributions is first introduced by Zhang and Ross [Zhang and Ross, 2021], and we extend it to a more general case with both the average and discounted criteria.

We further explore the theoretical result of \( \xi_\gamma \) by observing following phenomena.

Remark 1. For the average reward criterion, we have \( \xi = \lim_{\gamma \rightarrow 1} \xi_\gamma = \kappa_\pi - 1 \), which is consistent with [Zhang and Ross, 2021].

Remark 2. The \( \xi_\gamma \) is bounded by a \( \gamma \)-independent constant:
\[
\xi_\gamma \leq 2(\kappa_\pi - 1), \forall \gamma \in [0, 1].
\]
In particular, the maximum is not taken at \( \gamma = 1 \) but
\[
\arg \max_\gamma \xi_\gamma = 1 - \frac{1}{2\kappa_\pi - 1}.
\]

Remark 2 says that we are always safe to implement the trust region approach whatever discount factor we have. We may not intuitively believe that a smaller \( \gamma \) always leads to a tighter performance bound. When \( \gamma \) is larger than (15), the lower bound decreases as the \( \gamma \) increases and finally becomes \( \kappa_\pi - 1 \) at \( \gamma = 1 \). It supports that when the problem focuses more on long-term performance, the average criterion is better than a large discount factor in the trust region method, since the former is able to obtain a more accurate estimation.

Combining the preceding two bounds together, we conclude the following theorem.

Theorem 1. For any two stochastic policies \( \pi, \pi' \), the following bound holds:
\[
\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \geq \xi_\gamma \mathbb{E}_{s \sim d_{\pi}, \gamma}[D_{\text{TV}}(\pi' || \pi)[s]],
\]
\[
\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \leq \xi_\gamma \mathbb{E}_{s \sim d_{\pi}, \gamma}[D_{\text{TV}}(\pi' || \pi)[s]],
\]
In particular, the bounds hold with the average criterion:
\[
\eta_{\pi', \gamma} - \eta_\pi \geq \xi \mathbb{E}_{s \sim d_{\pi}, \gamma}[D_{\text{TV}}(\pi' || \pi)[s]],
\]
\[
\eta_{\pi', \gamma} - \eta_\pi \leq \xi \mathbb{E}_{s \sim d_{\pi}, \gamma}[D_{\text{TV}}(\pi' || \pi)[s]].
\]

4 Average Policy Optimization

Based on the unified trust region theory, we extend the current discounted trust region algorithms to the average reward criterion. We address the value drifting problem raised in the average setting, and present a technique named Average Value Constraint for better value estimation. By approximating the policy and value function with neutral networks, we develop an algorithm based on the Actor-Critic framework, which is named Average Policy Optimization.

4.1 Value Estimation

In the most DRL algorithms, the value function is approximated by a value network \( V_\phi(s) \) with \( \phi \) as the parameters, which is updated by minimizing the following loss with stochastic gradient descent (SGD) method
\[
\min_\phi J_V(\phi) = \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{1}{2} (\hat{V}(s_n) - V_\phi(s_n))^2 \right],
\]
where \( n \) indexes a batch of states sampled with \( \pi \) and \( \hat{V}(s_n) \) is the target value for \( s_n \).

Under the discounted criterion, \( \hat{V}(s_n) \) can be calculated directly using Monte Carlo (MC) method with \( V_{\text{MC}}(s_n) = \sum_{t=0}^{\infty} \gamma^t r_{n+t} \), or using the value of next step to bootstrap (TD): \( \hat{V}(s_n) = r_n + \gamma V_\phi(s_{n+1}) \) [Sutton and Barto, 2018], where the latter is preferred in most cases for low-variance estimates. In the average setting, we adopt the update rule of R-learning [Schwartz, 1993] to evaluate the target value with \( V_\phi(s_n) = r_n - \hat{V} + \gamma V_\phi(s_{n+1}) \), where \( \hat{V} \) is the estimate of \( V_\phi \). To smooth the value of \( \hat{V} \) over batches, we update \( \hat{V} \) by a moving average method: \( \hat{V} \leftarrow (1 - \alpha) \hat{V} + \alpha \frac{1}{n} \sum_{t=0}^{\infty} r_n \), where \( \alpha \) is the step size.

Let \( \phi' \) denote the updated parameters:
\[
\phi' = \phi + \beta (V_\phi(s) - V_\phi(s)) \nabla V_\phi(s),
\]
where \( \beta \) is the learning rate. As the value network is updated by SGD, noise in the estimate is inevitable. Suppose that there is a bias \( \epsilon_\phi \) between the value function and the true value
\[ V_\phi(s) = V_\phi^2(s) + \epsilon_V, \forall s \in S. \] The real updated parameters based on the \( V_\phi \) is
\[ \hat{\phi'} = \phi + \beta(\bar{V}_\phi(s) - V_\phi(s))\nabla V_\phi^2(s). \tag{18} \]
As \( \beta \) is sufficiently small, the post-update value function can be well-approximated by linearizing around \( \phi \) using Taylor’s expansion:
\[
\begin{align*}
V_\phi^2(s) & \approx V_\phi(s) + \beta(\bar{V}_\phi(s) - V_\phi(s))\nabla V_\phi^2(s) + \frac{\beta^2}{2}\nabla V_\phi^2(s)\nabla V_\phi^2(s)^T.
\end{align*}
\]

Subtracting the first equation from the second one, we have
\[
V_\phi^2(s) - V_\phi^2(s) \approx (1 - \beta(1 - \gamma))\nabla V_\phi^2(s)^T \epsilon_V. \tag{19}
\]
For \( \gamma < 1 \), the bias between the real values and approximated values is gradually eliminated during training. However, this feedback mechanism disappears for \( \gamma = 1 \) as \( V_\phi^2(s) - V_\phi^2(s) \approx \epsilon_V \). That means the approximated values are drifting away from the true values during training due to noises. We call it the value drifting problem in the average reward setting.

The analysis above explains why the value estimation in average setting is inaccurate in the data-driven mode. Next we try to correct the estimation by taking advantage of the following property.

**Proposition 3.** \( \mathbb{E}_{s \sim d_\pi} [V_\gamma(s)] = 0. \)

It tells us that the ideal estimates of values defined by (8) should have zero mean. We rewrite the problem in (16) by adding this condition as a constraint:
\[
\begin{align*}
\min_{\phi} J_V(\phi) &= \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{1}{2} \bar{V}(s_n) - V_\phi(s_n) \right]^2, \\
\text{s.t. } \frac{1}{N} \sum_{n=1}^{N} [V_\phi(s_n)] &= 0. \tag{20}
\end{align*}
\]
By constructing the Lagrangian function of (20), we have
\[
\begin{align*}
\min_{\phi} L_V(\phi, \nu) &= \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{1}{2} \bar{V}(s_n) - V_\phi(s_n) \right]^2 - \frac{1}{2} \nu b^2, \\
\text{where } b := \frac{1}{N} \sum_{n=1}^{N} [V_\phi(s_n)] \text{ and } \nu \text{ is the Lagrangian multiplier.}
\end{align*}
\]
Taking the gradient of \( L_V(\phi, \nu) \), we obtain
\[
\nabla L_V(\phi, \nu) = \frac{1}{N} \sum_{n=1}^{N} \left[ (\bar{V}(s_n) - V_\phi(s_n))\nabla \phi V_\phi(s_n) - \nu b \sum_{n=1}^{N} \nabla \phi [V_\phi(s_n)] \right].
\]
\[
\begin{align*}
\nabla L_V(\phi, \nu) &= \frac{1}{N} \sum_{n=1}^{N} \left[ (\bar{V}(s_n) - V_\phi(s_n))\nabla \phi V_\phi(s_n) - \nu b \nabla \phi [V_\phi(s_n)] \right].
\end{align*}
\]
The gradient of \( J_V(\phi) \) is equivalent to the following problem
\[
\begin{align*}
\min_{\phi} J_V(\phi) &= \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{1}{2} (\bar{V}(s_n) - V_\phi(s_n))^2 \right]. \tag{21}
\end{align*}
\]
where \( \bar{V}(s_n) := \bar{V}(s_n) - \nu b. \) Compared with the discounted variant, we explicitly control the bias in value estimation. We name this method as Average Value Constraint.

**Algorithm 1 Average Policy Optimization**

**Input:** \( \alpha, \beta, \lambda, \nu, \epsilon, K, N, M \)

1. Initialize \( \theta, \phi \) randomly.
2. Set \( \eta = 0, b = 0 \).
3. for \( k = 1, 2, \cdots, K \) do
   4. Run policy \( \pi_{\hat{\theta}_k} \) for \( N \) times to collect \( \{s_n, a_n, r(s_n, a_n), s_{n+1}\} \), \( n = 1, \cdots, N \).
   5. Update \( \eta \leftarrow (1 - \alpha)\eta + \alpha \frac{1}{N} \sum_{n=1}^{N} r(s_n, a_n) \).
   6. Update \( b \leftarrow (1 - \alpha)b + \alpha \frac{1}{N} \sum_{n=1}^{N} V_\phi(s_n) \).
   7. Compute \( \delta_t \) with (23) at all timesteps.
   8. Compute \( A(s_n, a_n) \) with (24) at all timesteps.
   9. Update the \( \theta \) with \( J_\pi(\theta) \) in (22) for \( M \) epochs.
10. Update the \( \phi \) with \( J_V(\phi) \) in (21) for \( M \) epochs.
11. end for

### 4.2 Policy Optimization

Many algorithms [Schulman et al., 2015; Schulman et al., 2017; Wu et al., 2017] have been proposed to approximately solve the discounted problem in (7) with a policy network. Here we choose PPO [Schulman et al., 2017] as the base to develop our average reward algorithm, and point out that the policy updating is similar with other trust region methods. Instead of optimizing the original objective with the constraint, we optimize the surrogate loss as follows:
\[
\begin{align*}
J_\pi(\theta) &= \frac{1}{N} \sum_{n=1}^{N} \left[ \min \left( \omega(\theta)A_n, \text{clip}(\omega(\theta), 1 - \epsilon, 1 + \epsilon)A_n \right) \right], \tag{22}
\end{align*}
\]
where \( \theta \) is the parameters of the policy network, \( \omega(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_\pi(a_t | s_t)} \) is the importance sampling ratio and \( A_n \) is a shorthand of advantage estimation \( A(s_n, a_n) \). To better balance the variance and bias of policy gradients, we further modify the discounted GAE [Schulman et al., 2016] to an average variant. Define the average TD residual as
\[
\delta_t = r(s_t, a_t) - \hat{\eta} + V_\phi(s_{t+1}) - V_\phi(s_t). \tag{23}
\]
Then the advantage estimator is
\[
\hat{A}(s_n, a_n) = \sum_{t=0}^{\infty} \lambda^t \delta_{n+t}, \tag{24}
\]
where \( \lambda \) provides the trade-off between bias and variance.

### 5 Related Work

Previous studies of average-reward RL mainly focus on Q-learning type algorithms in the tabular setting [Schwartz, 1993; Abounadi et al., 2001; Yang et al., 2016; Wan et al., 2021]. For the policy gradient methods [Baxter and Bartlett, 2001; Kakade, 2001], they often use a discount factor approaching one to simulate the average case instead of optimizing the average objective directly. A rapid developing line of average-reward RL with function approximations [Wei et al., 2021; Zhang et al., 2021; Hao et al., 2021] contributes valuable insights under the linear function assumption. However, the average reward setting has not yet caught too much attention of the DRL community.
Concurrently, there are new advances in the trust region approach for average-reward RL [Zhang and Ross, 2021; Dai and Gluzman, 2021]. Zhang and Ross [Zhang and Ross, 2021] first introduced the Kemeny’s constant in the bound of average reward trust region RL, which provided an interesting and important way to bound the performance distance in average scenarios. Dai and Gluzman [Dai and Gluzman, 2020; Dai and Gluzman, 2021] presented a general bound which depends on the discount factor continuously and studied the application of average-reward RL in a queueing network. Compared with these works, we give a general bound with the discounted and average criteria by using the Kemeny’s constant. Besides, we discover the property of the value drifting in average reward value estimation and propose a novel technique called Average Value Constraint to fix it, which we believe is essential to make average DRL algorithms practical in real-world scenarios.

6 Experiments

We conducted a series of experiments to evaluate APO, which are used to answer two questions:

• Compared with the discounted PPO, does APO have better performance in environments suitable with average criterion?

• What are the factors that affect the performance of APO? Specifically, does APO benefit from the Average Value Constraint?

We choose the continuous control benchmark MuJoCo [Todorov et al., 2012] with the OpenAI Gym [Brockman et al., 2016]. The MuJoCo tasks are designed to control the robot to finish certain tasks such as running as fast as possible without falling. It is natural to use the average criterion for some tasks in MuJoCo, such as HalfCheetah in which the robot keeps running and is rewarded by its speed. Moreover, we also evaluate APO in the other experiments with terminal states to examine the generalization ability of APO beyond the theoretical assumptions.

Both APO and PPO are implemented based on a modular RL package in PyTorch named rlpyt [Stooke and Abbeel, 2019]. While the performance of PPO is largely dependent on the code-optimizations [Andrychowicz et al., 2020], we do not consider any tricks apart from the vanilla PPO for controlling variables to justify true impact factors. For each task, we run the algorithm with 5 random seeds for 3 million steps and do the evaluation every 2000 steps. In the evaluation, we run 10 episodes without exploration by setting the standard deviation of policy as zero. All the hyper-parameter combinations we consider are grid searched, which are showed in Appendix B. The computing infrastructure for running experiments is a server with 2 AMD EPYC 7702 64-Core Processor CPUs and 8 Nvidia GeForce RTX 2080 Ti GPUs. The time consumed of each experiment varies from 1.2h to 2h on average according to the complexity of the task.

6.1 Comparison to Discounted PPO

We compare our APO with the standard discounted PPO with the different discounted factors, and select the combination of hyperparameters which achieves the best average performance. The results in Figure 1 show that APO surpasses all the variants of PPO with different discount factors. We observe that the performances of PPO vary drastically with the change of γ. Meanwhile, The relationship between performance and γ variation is not monotonic. For example, γ = 0.99 achieves a sound result in Ant, while the performance with γ = 0.999 is terrible. This confirms that the performance with the average criterion cannot be optimized by letting γ approach 1 in the discounted framework. The experimental results appear to be compatible with the results obtained by Zhang and Ross [Zhang et al., 2021].

The environments shown in Figure 1 are suitable for average setting, as they have no terminal states (Swimmer and HalfCheetah) or hard to fall even with a random policy (Ant). It is natural that APO beats the discounted PPO in these tasks. The audience may be curious about the performances of APO in other tasks, such as Hopper, Walker2d and Humanoid. We also evaluate APO in them and show the results in Appendix C. As we expected, APO achieves sound results with the competent PPO but does not beat the best in Hopper and Walker. However, when we change the metric from average episode return to average reward, we find that APO is better than discounted PPO, especially in Humanoid. It shows that APO achieves higher speed but ignores the terminal reward as it focuses on the average performance under the steady-state distribution. It should be more suitable to model the safety requirements as constraints, which is easy to be extended from

Figure 1: Compare APO and PPO with different discount factors in MuJoCo tasks. Results for more tasks can be found in Appendix C.
APO [Achiam et al., 2017].

6.2 Ablation Study on Average Value Constraint

We further evaluate the proposed method Average Value Constraint for the value estimation in average setting. We fix the penalty coefficient $\nu$ at different levels to see the changes of average values during training. The performances are improved by 66.14% with the Average Value Constraint. As shown in Figure 2, a larger $\nu$ produces a tighter constraint on the average value. Without the average value constraint ($\nu = 0$), the average values fluctuate dramatically during training. Since only the relative value of different states is meaningful for the decision, the bias of average value is detrimental to the policy gradient estimation. As the average value constraint is derived from the Lagrangian method, we can also optimize $\nu$ adaptively during training [Stooke et al., 2020].

7 Conclusion

In this paper, we study the trust region approach with the average-reward RL. Based on the PA theory, we prove a lower performance bound with the average criterion and tighten the previous discounted bound with a large discount factor. With the novel lower bound, the monotonic policy improvement of the average trust region method is guaranteed. Furthermore, we analyze the value estimation in the average setting and propose a new technique named Average Value Constraint to stabilize the value estimation during training. In empirical experiments, our proposed algorithm APO achieves better performance than most PPO variants with different discount factors in MuJoCo, demonstrating the average criterion’s superiority in many engineering problems. In summary, this work fills the gap in the theory of DRL between the average and discount criteria and provides a practical algorithm with a more appropriate criterion for many practical problems beyond the current benchmarks.

References

[Abounadi et al., 2001] Jinane Abounadi, Dimitri Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. SIAM Journal on Control and Optimization, 40(3):681–698, 2001.

[Achiam et al., 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In International Conference on Machine Learning, pages 22–31, 2017.

[Andrychowicz et al., 2020] Marcin Andrychowicz, Anton Raichuk, Piotr Statczuk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Husserot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters in on-policy reinforcement learning? a large-scale empirical study. arXiv preprint arXiv:2006.05990, 2020.

[Baxter and Bartlett, 2001] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research, 15:319–350, 2001.

[Bertsekas and Tsitsiklis, 1995] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In IEEE Conference on Decision and Control, pages 560–564. IEEE, 1995.

[Bertsekas et al., 1995] Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. Dynamic programming and optimal control, volume II. Athena scientific Belmont, 1995.

[Brockman et al., 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. arXiv preprint arXiv:1606.01540, 2016.

[Cao and Chen, 1997] Xi-Ren Cao and Han-Fu Chen. Perturbation realization, potentials, and sensitivity analysis of markov processes. IEEE Transactions on Automatic Control, 42(10):1382–1393, 1997.

[Cao, 2007] Xi-Ren Cao. Stochastic Learning and Optimization: A Sensitivity-Based Approach. Springer, 2007.

[Dai and Gluzman, 2020] Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. arXiv preprint arXiv:2008.01644, 2020.

[Dai and Gluzman, 2021] Jim G Dai and Mark Gluzman. Refined policy improvement bounds for mdps. In Workshop on Reinforcement Learning Theory, International Conference on Machine Learning, 2021.

[Dewanto et al., 2020] Vektor Dewanto, George Dunn, Ali Eshragh, Marcus Gallagher, and Fred Roosta. Average-reward model-free reinforcement learning: a systematic review and literature mapping. arXiv preprint arXiv:2010.08920, 2020.

[Duan et al., 2016] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In International Conference on Machine Learning, pages 1329–1338, 2016.

[Hao et al., 2021] Botao Hao, Nevena Lazic, Yasin Abbassi-Yadkori, Pooria Joulani, and Csaba Szepesvari. Adaptive approximate policy iteration. In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 523–531. PMLR, 2021.

[Howard, 1960] Ronald A. Howard. Dynamic programming and markov processes. Mathematical Gazette, 3(358):120, 1960.

[Kakade and Langford, 2002] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In International Conference on Machine Learning, volume 2, pages 267–274, 2002.

[Kakade, 2001] Sham Kakade. Optimizing average reward using discounted rewards. In International Conference on Computational Learning Theory, pages 605–615. Springer, 2001.

[Kemeny and Snell, 1960] John G Kemeny and J Laurie Snell. Finite markov chains. Van Nostrand, New Jersey, 1960.
[Konda and Tsitsiklis, 2000] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In Advances in neural information processing systems, pages 1008–1014, 2000.

[Mahadevan, 1996] Sridhar Mahadevan. Average reward reinforcement learning: foundations, algorithms, and empirical results. Machine Learning, 22(1):159–195, 1996.

[Marbach and Tsitsiklis, 2003] Peter Marbach and John N Tsitsiklis. Approximate gradient methods in policy-space optimization of markov reward processes. Discrete Event Dynamic Systems, 13(1-2):111–148, 2003.

[Puterman, 1994] Martin L. Puterman. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 1994.

[Schulman et al., 2015] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897, 2015.

[Schulman et al., 2016] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In International Conference on Learning Representations, 2016.

[Schulman et al., 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[Schwartz, 1993] Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In International Conference on Machine Learning, pages 298–305, 1993.

[Senior et al., 2020] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, 2020.

[Silver et al., 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

[Stooke and Abbeel, 2019] Adam Stooke and Pieter Abbeel. rlpyt: A research code base for deep reinforcement learning in pytorch. arXiv preprint arXiv:1909.01500, 2019.

[Stooke et al., 2020] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In International Conference on Machine Learning, pages 9133–9143. PMLR, 2020.

[Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT press, 2018.

[Sutton et al., 2000] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063, 2000.

[Thomas, 2014] Philip Thomas. Bias in natural actor-critic algorithms. In International conference on machine learning, pages 441–448, 2014.

[Todorov et al., 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026–5033, 2012.

[Vinyals et al., 2019] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michael Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782):350–354, 2019.

[Wan et al., 2021] Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward markov decision processes. In International Conference on Machine Learning, pages 10653–10662. PMLR, 2021.

[Wei et al., 2021] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In International Conference on Artificial Intelligence and Statistics, pages 3007–3015. PMLR, 2021.

[Wu et al., 2017] Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In Advances in neural information processing systems, pages 5279–5288, 2017.

[Yang et al., 2016] Shangdong Yang, Yang Gao, Bo An, Hao Wang, and Xingguo Chen. Efficient average reward reinforcement learning using constant shifting values. In AAAI Conference on Artificial Intelligence, pages 2258–2264, 2016.

[Zhang and Ross, 2021] Yiming Zhang and Keith W Ross. On-policy deep reinforcement learning for the average-reward criterion. In International conference on machine learning, 2021.

[Zhang et al., 2021] Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. arXiv preprint arXiv:2101.02808, 2021.
A Proofs
A.1 Preliminaries
Perturbation Analysis
Before we present our new result of the performance bound for trust region method, we first review the theory of Perturbation Analysis [Cao and Chen, 1997; Cao, 2007], based on which we give our proofs. Our theory analysis is limited to the finite ergodic Markov chain, which means that from any state it is possible to reach every other state.

With the discounted reward criterion, we are interested in the discounted reward \( \eta_{\pi,\gamma} \), which is defined as

\[
\eta_{\pi,\gamma} = (1 - \gamma) \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \, | \, s_0 \sim d_0 \right].
\]  (25)

We further define the discounted steady-state distribution of the state \( s \) as

\[
d_{\pi,\gamma}(s) = (1 - \gamma) \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t P(s_t = s) \, | \, s_0 \sim d_0 \right].
\]  (26)

In the matrix form, (25) and (26) are

\[
\eta_{\pi,\gamma} = (1 - \gamma)d_0 \sum_{t=0}^{\infty} (\gamma P_\pi)^t r_\pi = (1 - \gamma)d_0 (I - \gamma P_\pi)^{-1} r_\pi = d_{\pi,\gamma} r_\pi,
\]  (27)

\[
d_{\pi,\gamma} = (1 - \gamma)d_0 \sum_{t=0}^{\infty} (\gamma P_\pi)^t = (1 - \gamma)d_0 (I - \gamma P_\pi)^{-1},
\]  (28)

where \( d_0 \) is the initial state distribution, \( P_\pi \) is the transition matrix induced by \( \pi \), and \( r_\pi \) is the reward vector of \( \pi \) with components \( r_\pi(s) = \mathbb{E}_{a \sim \pi, s' \sim p_\pi} r(s, a, s') \). The core concept of perturbation analysis is the performance difference formula:

\[
\eta_{\pi',\gamma} - \eta_{\pi,\gamma} = d_{\pi',\gamma} (r_\pi' - r_\pi + \gamma (P_\pi' - P_\pi) V_{\pi,\gamma}),
\]  (29)

where \( V_{\pi,\gamma} \) is the discounted value function for policy \( \pi \) (which is called discounted potential in the perturbation analysis). It is known that \( V_{\pi,\gamma} \) satisfies the discounted Poisson equation:

\[
(I - \gamma P_\pi + \gamma e d_\pi) V_{\pi,\gamma} = r_\pi,
\]  (30)

where \( d_\pi \) is the steady-state distribution and \( e \) is the unit vector. In (30), the inverse of left matrix \((I - \gamma P_\pi + \gamma e d_\pi)^{-1}\) is called the discounted fundamental matrix, which is denoted as \( Z_{\pi,\gamma} \) in this paper. There are some important properties of \( Z_{\pi,\gamma} \):

\[
Z_{\pi,\gamma} e = e,
\]  (31)

\[
d_\pi Z_{\pi,\gamma} = d_\pi,
\]  (32)

\[
Z_{\pi,\gamma} (I - \gamma P_\pi) = I - \gamma e d_\pi.
\]  (33)

Since all the applications of the value function in optimization depend only on the differences of the components of \( V_{\pi,\gamma} \), we can replace \( V_{\pi,\gamma} \) with \( V_{\pi,\gamma} + ce \), where \( c \) could be any constant. In this paper, we choose to calculate \( V_{\pi,\gamma} \) as

\[
V_{\pi,\gamma}(s) = \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t (r_t - \eta_\pi) \, | \, s_0 = s \right],
\]  (34)

which satisfies \( \mathbb{E}_{s \sim d_\pi} [V_{\pi,\gamma}(s)] = 0 \) (we will give out the proof later).

With the power of performance difference formula, the relationship of (9) is attained at once by observing that

\[
[r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi) V_{\pi,\gamma}] (s) = \mathbb{E}_{a \sim \pi', s' \sim p} [r(s, a, s') - \eta_\pi + \gamma V_{\pi,\gamma}(s')] - \mathbb{E}_{a \sim \pi, s' \sim p} [r(s, a, s') - \eta_\pi + \gamma V_{\pi,\gamma}(s')]
\]  (35)

\[
= \mathbb{E}_{a \sim \pi', s' \sim p} [r(s, a, s') - \eta_\pi + \gamma V_{\pi,\gamma}(s')] - V_{\pi,\gamma}(s)
\]  (36)

\[
= \mathbb{E}_{a \sim \pi'} [A^\pi(s', a)].
\]  (37)

Moreover, when \( \gamma \to 1 \), it is easy to get the limitations of above equations with the average reward criterion:

- Average reward: \( \lim_{\gamma \to 1} \eta_{\pi,\gamma} = \eta_\pi = d_\pi r_\pi \).
- Steady-state distribution: \( \lim_{\gamma \to 1} d_{\pi,\gamma} = d_\pi \).
- Average value function: \( \lim_{\gamma \to 1} V_{\pi,\gamma} = V_\pi \).
- Fundamental matrix: \( \lim_{\gamma \to 1} Z_{\pi,\gamma} = Z_\pi = (I - P_\pi + e d_\pi)^{-1} \).
- Performance difference formula: \( \eta_{\pi'} - \eta_\pi = d_\pi (r_{\pi'} - r_\pi + (P_{\pi'} - P_\pi) V_\pi) \).

Thus, powered by the PA, we are able to extend the analysis of trust region with the discounted reward criterion into the average reward criterion.
The mean first passage matrix and Kemeny’s constant

Next, we review the concepts of the mean first passage matrix [Kemeny and Snell, 1960]. If an ergodic Markov chain induced by \( \pi \) is started in state \( s \), the expected number of steps to reach state \( s' \) for the first time is called the mean first passage time from \( s \) to \( s' \), denoted by \( M_\pi(s, s') \). The matrix \( M_\pi \) satisfies the following equation:

\[
M_\pi = P_\pi(M_\pi - D_\pi) + E,
\]

where \( E \) is a square matrix with all elements being 1 and \( D_\pi = (M_\pi)_{dg} \) satisfying that \( D_\pi(s, s) = 1/d_\pi(s) \) and \( D_\pi(s, s') = 0 \) for \( s \neq s' \). The subscript ‘dg’ for some square matrix \( N \) refers to a diagonal matrix whose elements are the diagonals of \( N \). The mean first passage matrix \( M_\pi \) is determined from the fundamental matrix \( Z_\pi = (I - P_\pi + ed_\pi)^{-1} \) by

\[
M_\pi = (I - Z_\pi + E(Z_\pi)_{dg}) D_\pi.
\]

Kemeny [1960] show that the mean time starting from a given state \( s \) to the equilibrium distribution is a constant independent of \( s \):

\[
\kappa_\pi = \sum_{s'} d_\pi(s') M_\pi(s, s'),
\]

where the constant \( \kappa_\pi \) is called Kemeny’s constant. This constant is an invariant of MDP in the planning horizon.

A.2 Main Results

**Lemma A.1.** The matrix \( I - (1 - \gamma)M_\pi D_\pi^{-1} \) is invertible when \( (1 - \gamma)\kappa_\pi < 1 \).

**Proof.** First we will show that \( A = \frac{1}{\kappa_\pi} M_\pi D_\pi^{-1} \) is a stochastic matrix. By saying a stochastic matrix, we means that all its elements are non-negative and it satisfies \( Ae = e \). By the definition of \( M_\pi \) and \( D_\pi \), we already know \( A \) is a non-negative matrix. Substituting (39) into (A.1) we have

\[
Ae = \frac{1}{\kappa_\pi} M_\pi D_\pi^{-1} e = \frac{1}{\kappa_\pi} (I - Z_\pi + E(Z_\pi)_{dg}) e = \frac{1}{\kappa_\pi} (e - e + \kappa_\pi e) = e.
\]

With the property of stochastic matrix, we know \( \rho(A) = 1 \), where \( \rho \) is the spectral radius of a square matrix. Further, we have

\[
\rho \left( (1 - \gamma)M_\pi D_\pi^{-1} \right) < \rho \left( \frac{1}{\kappa_\pi} M_\pi D_\pi^{-1} \right) = \rho(A) = 1.
\]

Again, with the property of stochastic matrix, we know that \( I - (1 - \gamma)M_\pi D_\pi^{-1} \) is invertible. \( \square \)

**Lemma A.2.** If \( (1 - \gamma)\kappa_\pi < 1 \), \( Z_{\pi,\gamma} = [I - M_\pi D_\pi^{-1} + E(Z_{\pi,\gamma})_{dg} - (1 - \gamma)E(Z_{\pi,\gamma} M_\pi)_{dg} D_\pi^{-1} ] (I - (1 - \gamma)M_\pi D_\pi^{-1})^{-1} \).

**Proof.** Multiply \( \gamma \) on both sides of (38) and rearranging items as follows:

\[
(1 - \gamma P_\pi)(M_\pi - D_\pi) = \gamma E + (1 - \gamma)M_\pi - D_\pi.
\]

Left-multiplying \( Z_{\pi,\gamma} \), we have

\[
Z_{\pi,\gamma}(1 - \gamma P_\pi)(M_\pi - D_\pi) = \gamma Z_{\pi,\gamma} E + (1 - \gamma)Z_{\pi,\gamma} M_\pi - Z_{\pi,\gamma} D_\pi = \gamma E + (1 - \gamma)Z_{\pi,\gamma} M_\pi - Z_{\pi,\gamma} D_\pi,
\]

where the equality comes from (32). Substituting (33) into the left hand of above equation, we have

\[
Z_{\pi,\gamma}(1 - \gamma P_\pi)(M_\pi - D_\pi) = (I - \gamma ed_\pi)(M_\pi - D_\pi)
\]

\[
= M_\pi - \gamma ed_\pi M_\pi - D_\pi + \gamma ed_\pi D_\pi
\]

\[
= M_\pi - \gamma ed_\pi M_\pi - D_\pi + \gamma E.
\]

Thus we obtain

\[
M_\pi - \gamma ed_\pi M_\pi - D_\pi = (1 - \gamma)Z_{\pi,\gamma} M_\pi - Z_{\pi,\gamma} D_\pi.
\]

When \( s = s' \), we have the following relationship with the fact \( M_\pi(s, s) = D_\pi(s, s) \):

\[
-\gamma[d_\pi M_\pi](s) = (1 - \gamma)[Z_{\pi,\gamma} M_\pi](s, s) - Z_{\pi,\gamma}(s, s)/d_\pi(s).
\]

Rewriting the above equation into matrix form, we get

\[
-\gamma ed_\pi M_\pi = (1 - \gamma)E(Z_{\pi,\gamma} M_\pi)_{dg} - E(Z_{\pi,\gamma})_{dg} D_\pi.
\]

Substituting (51) into (49), we obtain

\[
M_\pi - D_\pi + (1 - \gamma)E(Z_{\pi,\gamma} M_\pi)_{dg} - E(Z_{\pi,\gamma})_{dg} D_\pi = (1 - \gamma)Z_{\pi,\gamma} M_\pi - Z_{\pi,\gamma} D_\pi.
\]

Right-multiplying both sides by \( D_\pi^{-1} \) gives us

\[
Z_{\pi,\gamma} (I - (1 - \gamma)M_\pi D_\pi^{-1}) = I - M_\pi D_\pi^{-1} + E(Z_{\pi,\gamma})_{dg} - (1 - \gamma)E(Z_{\pi,\gamma} M_\pi)_{dg} D_\pi^{-1}.
\]

Finally, with the fact that \( I - (1 - \gamma)M_\pi D_\pi^{-1} \) is invertible by Lemma A.1, we finish the proof. \( \square \)
Lemma A.3.\quad d_{\pi',\gamma} - d_{\pi,\gamma} = \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi)Z_{\pi',\gamma}.

Proof.\quad By the definition of $d_{\pi',\gamma}, d_{\pi,\gamma}$, we have

\begin{align*}
d_{\pi',\gamma} - d_{\pi,\gamma} &= (1 - \gamma)d_0(I - \gamma P_{\pi'})^{-1} - (1 - \gamma)d_0(I - \gamma P_\pi)^{-1} \\
&= (1 - \gamma)d_0[(I - \gamma P_{\pi'})^{-1} - (I - \gamma P_\pi)^{-1}] \\
&= (1 - \gamma)d_0[(I - \gamma P_{\pi'})^{-1} - (I - \gamma P_{\pi'})^{-1}(I - \gamma P_{\pi'})^{-1}] \\
&= (1 - \gamma)d_0(I - \gamma P_{\pi'})^{-1}\gamma(P_{\pi'} - P_\pi)(I - \gamma P_{\pi'})^{-1} \\
&= \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi)(I - \gamma P_{\pi'})^{-1}. \quad (58)
\end{align*}

With the fact that $P_{\pi'} e = e$, we obtain

\[(I - \gamma P_{\pi'})^{-1} e = \sum_t (P_{\pi'})^t e = \sum_t (P_{\pi'})^t e = \sum_t (\gamma)^t e = \frac{1}{1 - \gamma} e.\]

Then, by right-multiplying $I - \gamma P_{\pi'} + \gamma ed_{\pi'}$ to $d_{\pi',\gamma} - d_{\pi,\gamma}$, we have

\[
(d_{\pi',\gamma} - d_{\pi,\gamma})(I - \gamma P_{\pi'} + \gamma ed_{\pi'}) = \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi)(I - \gamma P_{\pi'})^{-1}(I - \gamma P_{\pi'} + \gamma ed_{\pi'})
= \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi) \left( I - \frac{\gamma}{1 - \gamma} ed_{\pi'} \right)
= \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi).
\]

Thus, by the definition of $Z_{\pi',\gamma}$, we complete the proof:

\[d_{\pi',\gamma} - d_{\pi,\gamma} = \gamma d_{\pi,\gamma}(P_{\pi'} - P_\pi)Z_{\pi',\gamma}, \quad (59)\]

Proposition 1.\quad For any two stochastic policies $\pi, \pi'$, the following bound holds:

\[
\eta_{\pi',\gamma} - \eta_{\pi,\gamma} \geq L_{\pi,\gamma}(\pi') - 2\epsilon_{\gamma} D_{TV}(d_{\pi',\gamma} \parallel d_{\pi,\gamma}),
\quad (60)
\]

\[
\eta_{\pi',\gamma} - \eta_{\pi,\gamma} \leq L_{\pi,\gamma}(\pi') + 2\epsilon_{\gamma} D_{TV}(d_{\pi',\gamma} \parallel d_{\pi,\gamma}), \quad (61)
\]

where $\epsilon_{\gamma} = \max_s |E_{a \sim \pi'}[A_{\pi,\gamma}(s, a)]|$.\]

Proof.\quad Rewrite $L_{\pi,\gamma}(\pi')$ in the matrix form:

\[L_{\pi,\gamma}(\pi') = d_{\pi,\gamma}[r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma}]. \quad (62)\]

Subtracting the above equation from (29), we obtain

\[
(\eta_{\pi',\gamma} - \eta_{\pi,\gamma}) - L_{\pi,\gamma}(\pi') = d_{\pi',\gamma}[r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma}] - d_{\pi,\gamma}[r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma}]
= (d_{\pi',\gamma} - d_{\pi,\gamma})[r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma}] \quad (63)
\]

Hölder’s inequality tells us that

\[
| (\eta_{\pi',\gamma} - \eta_{\pi,\gamma}) - L_{\pi,\gamma}(\pi') | \leq \| d_{\pi',\gamma} - d_{\pi,\gamma} \|_1 \| r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma} \|_\infty
= 2\epsilon_{\gamma} D_{TV}(d_{\pi',\gamma} \parallel d_{\pi,\gamma}), \quad (66)
\]

where $\epsilon_{\gamma} = \| r_{\pi'} - r_\pi + \gamma (P_{\pi'} - P_\pi)V_{\pi,\gamma} \|_\infty = \max_s |E_{a \sim \pi'}[A_{\pi,\gamma}(s, a)]|$ following by (37).\]

Proposition 2.\quad For any two stochastic policies $\pi, \pi'$, the following bound holds:

\[D_{TV}(d_{\pi',\gamma} \parallel d_{\pi,\gamma}) \leq \xi_{\gamma} \mathbb{E}_{s \sim d_{\pi,\gamma}}[D_{TV}(\pi' \parallel \pi)[s]], \quad (67)\]

where $\xi_{\gamma} = \min \left\{ \frac{\gamma}{1 - \gamma}, \frac{\gamma(\kappa_{\pi'} - 1)}{1 - (1 - \gamma)\kappa_{\pi'}} \right\}$ and $\kappa_{\pi'}$ is the Kemeny’s constant of the Markov chain induced by $\pi'$.\]

Proof.\quad The total variation divergence between $d_{\pi',\gamma}, d_{\pi,\gamma}$ is alignend with 1-norm of their difference:

\[D_{TV}(d_{\pi',\gamma} \parallel d_{\pi,\gamma}) = \frac{1}{2} \sum_s |d_{\pi',\gamma}(s) - d_{\pi,\gamma}(s)| = \frac{1}{2} \| d_{\pi',\gamma} - d_{\pi,\gamma} \|_1. \quad (68)\]
Assume that \( \frac{\gamma - 1}{\kappa_{\pi'}} < \gamma \leq 1 \). Putting the results of Lemma A.2 and Lemma A.3 together leads to

\[
d_{\pi', \gamma} - d_{\pi, \gamma} = \gamma d_{\pi, \gamma}(P_{\pi'} - P_{\pi})Z_{\pi', \gamma}
\]

\[
= \gamma d_{\pi, \gamma}(P_{\pi'} - P_{\pi}) \left[ I - M_{\pi'}D_{\pi'}^{-1} + E(Z_{\pi', \gamma}d_{\pi'} - (1 - \gamma)E(Z_{\pi', \gamma}M_{\pi'}d_{\pi'})D_{\pi'}^{-1} \right] (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}
\]

\[
= \gamma d_{\pi, \gamma}(P_{\pi'} - P_{\pi}) \left( I - M_{\pi'}D_{\pi'}^{-1} \right) (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1},
\]

(69)

where the last equality comes from the fact that \( P_{\pi'}E = P_{\pi}E = E \). Comparing the previous result (see Lemma 3 in [Zhang and Ross, 2021]), we obtain an extra term \( \gamma (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1} \), which is negligible when \( \gamma = 1 \) but meaningful with large discount factors.

Thus, we have

\[
\|d_{\pi', \gamma} - d_{\pi, \gamma}\|_1 = \|\gamma d_{\pi, \gamma}(P_{\pi'} - P_{\pi}) \left( I - M_{\pi'}D_{\pi'}^{-1} \right) (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_1
\]

\[
\leq \gamma \|d_{\pi, \gamma}(P_{\pi'} - P_{\pi})\|_1 \| (I - M_{\pi'}D_{\pi'}^{-1}) (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_\infty
\]

\[
\leq \gamma \|d_{\pi, \gamma}(P_{\pi'} - P_{\pi})\|_1 \| (I - M_{\pi'}D_{\pi'}^{-1}) \|_\infty \| (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_\infty,
\]

(72)

where the first inequality follows from the Hölder’s inequality. Next, let us bound the items one by one.

\[
\|d_{\pi, \gamma}(P_{\pi'} - P_{\pi})\|_1 = \sum_{s'} \left| \sum_{s} \left( \sum_{a} \left( P(s' | s, a) \pi'(a | s) - P(s' | s, a) \pi(a | s) \right) \right) d_{\pi, \gamma}(s) \right|
\]

\[
\leq \sum_{s', s} \sum_{a} P(s' | s, a)(\pi'(a | s) - \pi(a | s)) d_{\pi, \gamma}(s)
\]

\[
\leq \sum_{s, s', a} P(s' | s, a)|\pi'(a | s) - \pi(a | s)| d_{\pi, \gamma}(s)
\]

\[
\leq \sum_{s, a} |\pi'(a | s) - \pi(a | s)| d_{\pi, \gamma}(s)
\]

\[
= 2 \mathbb{E}_{s \sim d_{\pi, \gamma}} [D_{TV}(\pi' || \pi)[s]],
\]

(75)

\[
\| (I - M_{\pi'}D_{\pi'}^{-1}) \|_\infty = \sum_{s} \left| \sum_{s'} M(s, s')d_{\pi'}(s) - 1 \right| = \kappa_{\pi'} - 1.
\]

(76)

\[
\| (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1} \|_\infty = \sum_{t=0}^{\infty} \left| \sum_{s'} (1 - \gamma)M_{\pi'}D_{\pi'}^{-1} \right|_t
\]

\[
\leq \sum_{t=0}^{\infty} \left| \sum_{s'} (1 - \gamma)M_{\pi'}D_{\pi'}^{-1} \right|_t
\]

\[
= \sum_{t=0}^{\infty} ((1 - \gamma)\kappa_{\pi'})^t
\]

\[
= \frac{1}{(1 - \gamma)\kappa_{\pi'}}.
\]

(80)

Combining the bounds above, we conclude that:

\[
\|d_{\pi', \gamma} - d_{\pi, \gamma}\|_1 = \|\gamma d_{\pi, \gamma}(P_{\pi'} - P_{\pi}) \left( I - M_{\pi'}D_{\pi'}^{-1} \right) (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_1
\]

\[
\leq \gamma \|d_{\pi, \gamma}(P_{\pi'} - P_{\pi})\|_1 \| (I - M_{\pi'}D_{\pi'}^{-1}) (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_\infty
\]

\[
\leq \gamma \|d_{\pi, \gamma}(P_{\pi'} - P_{\pi})\|_1 \| (I - M_{\pi'}D_{\pi'}^{-1}) \|_\infty \| (I - (1 - \gamma)M_{\pi'}D_{\pi'}^{-1})^{-1}\|_\infty
\]

\[
\leq \frac{\gamma (\kappa_{\pi'} - 1)}{1 - (1 - \gamma)\kappa_{\pi'}} \mathbb{E}_{s \sim d_{\pi, \gamma}} [D_{TV}(\pi' || \pi)[s]],
\]

(85)
Similar with the above analysis, when \( \gamma < 1 \), we conclude the other part of the bound from (58)
\[
\|d_{\pi', \gamma} - d_{\pi, \gamma}\|_1 = \|d_{\pi, \gamma}(P_\pi - P_{\pi'}) (I - \gamma P_{\pi'})^{-1}\|_1 \\
\leq \gamma \|d_{\pi, \gamma}(P_\pi - P_{\pi'})\|_1 \| (I - \gamma P_{\pi'})^{-1}\|_\infty \\
\leq \frac{\gamma}{1 - \gamma} \| \xi_{s \sim d_{\pi, \gamma}} [D_{TV}(\pi' \parallel \pi)][s] \|,
\]
where
\[
\| (I - \gamma P_{\pi'})^{-1}\|_\infty = \| \sum_{t=0}^\infty (\gamma P_{\pi'})^t \|_\infty \leq \sum_{t=0}^\infty \gamma^t \| P_{\pi'} \|^t \leq \sum_{t=0}^\infty \gamma^t = \frac{1}{1 - \gamma}.
\]
Thus we recover the bound in [Achiam et al., 2017]. Note that (92) is only meaningful for \( \gamma < 1 \) and (88) is only meaningful for \( \kappa_{\pi'}^{-1} - \gamma \leq 1 \). When \( \gamma < \kappa_{\pi'}^{-1} - \gamma \), we have
\[
\frac{\gamma (\kappa_{\pi'} - 1)}{1 - (1 - \gamma) \kappa_{\pi'}} = \frac{\gamma (\kappa_{\pi'} - 1)}{1 - (1 - \gamma) \kappa_{\pi'} - 1} = \frac{\gamma}{1 - \gamma} \left( \frac{\kappa_{\pi'} - 1}{\kappa_{\pi'} - 1/1 - \gamma} \right) > \frac{\gamma}{1 - \gamma}.
\]
Thus we combine the bounds with \( \xi_\gamma = \min \left\{ \frac{1}{1 - \gamma} \left| \frac{\gamma (\kappa_{\pi'} - 1)}{1 - (1 - \gamma) \kappa_{\pi'}} \right| \right\} \) in a unified form.

**Theorem 1.** For any two stochastic policies \( \pi, \pi' \), the following bound holds:
\[
\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \geq L_{\pi, \gamma}(\pi') - 2\epsilon \gamma \xi_\gamma \| \xi_{s \sim d_{\pi, \gamma}} [D_{TV}(\pi' \parallel \pi)][s] \|,
\]
\[
\eta_{\pi', \gamma} - \eta_{\pi, \gamma} \leq L_{\pi, \gamma}(\pi') + 2\epsilon \gamma \xi_\gamma \| \xi_{s \sim d_{\pi, \gamma}} [D_{TV}(\pi' \parallel \pi)][s] \|.
\]
In particular, the bounds hold with the average criterion:
\[
\eta_{\pi'} - \eta_{\pi} \geq L_{\pi}(\pi') - 2\epsilon \xi \| \xi_{s \sim d_{\pi}} [D_{TV}(\pi' \parallel \pi)][s] \|,
\]
\[
\eta_{\pi'} - \eta_{\pi} \leq L_{\pi}(\pi') + 2\epsilon \xi \| \xi_{s \sim d_{\pi}} [D_{TV}(\pi' \parallel \pi)][s] \|.
\]

**Proof.** The theorem is the combination of two above propositions.

**Proposition 3.** \( \mathbb{E}_{s \sim d_{\pi}} [V_{\pi, \gamma}(s)] = 0. \)

**Proof.** Revisit the definition of \( V_{\pi, \gamma} \) in matrix form,
\[
V_{\pi, \gamma} = \sum_{t=0}^\infty (\gamma P_{\pi})^t (r_{\pi} - ed_{\pi} r_{\pi}) = r_{\pi} - ed_{\pi} r_{\pi} + \sum_{t=1}^\infty (\gamma^t (P_{\pi} - ed_{\pi})^t r_{\pi}) = \left[ (I - \gamma P_{\pi} + \gamma ed_{\pi})^{-1} - ed_{\pi} \right] r_{\pi}.
\]
With the fact that \( d_{\pi} P_{\pi} = d_{\pi} \) and \( d_{\pi} e = 1 \), we have
\[
\mathbb{E}_{s \sim d_{\pi}} V_{\pi, \gamma}(s) = d_{\pi} V_{\pi, \gamma} = d_{\pi} \left[ (I - \gamma P_{\pi} + \gamma ed_{\pi})^{-1} - ed_{\pi} \right] r_{\pi} = d_{\pi} \left[ I + \sum_{t=1}^\infty (\gamma^t (P_{\pi} - ed_{\pi})^t) - ed_{\pi} \right] r_{\pi} = d_{\pi} \left[ I + \sum_{t=1}^\infty (\gamma^t (P_{\pi} - ed_{\pi})) - ed_{\pi} \right] r_{\pi} = \left[ d_{\pi} + \sum_{t=1}^\infty (\gamma^t (d_{\pi} - d_{\pi})) - d_{\pi} \right] r_{\pi} = 0.
\]
B Hyperparameters of APO and PPO

| Hyperparameter                  | Value                        |
|---------------------------------|------------------------------|
| **Shared**                      |                              |
| Network learning rate $\beta$   | $3e^{-4}$                    |
| Network hidden sizes            | [64, 64]                     |
| Activation function             | Tanh                         |
| Optimizer                       | Adam                         |
| Batch size                      | 256                          |
| Gradient Clipping               | 10                           |
| Clipping parameter $\varepsilon$| 0.2                          |
| Optimization Epochs $M$         | 10                           |
| GAE parameter $\lambda$        | [0.8, 0.9, 0.95, 0.99]       |
| **PPO**                         |                              |
| Discount factor $\gamma$       | [0.9, 0.95, 0.99, 0.999]     |
| **APO**                         |                              |
| Step size $\alpha$              | [0.03, 0.1, 0.3]             |
| Average Value Constraint Coefficient $\nu$ | [0, 0.03, 0.1, 0.3, 1.0]   |

Table 1: Hyperparameters Sheet

C Additional Results

We benchmark our method on the MuJoCo continuous control tasks. The training curves of average episode return and average reward are shown in Figure 3 and Figure 4 respectively. The environments Swimmer and HalfCheetah do not have terminal states, so the treads of curves are the same in the two metrics. In these tasks, APO has the best performance in both the episode return and average reward. A similar result also exists in Ant, in which the multi-legged robot is safe from falling. In the other tasks with the unsafe states, APO still has better performances in terms of average reward than discounted PPO, especially in Humanoid. There is no evidence that the discounted criterion is better than the average one in problems with unsafe states though APO does not beat the best PPO in terms of average episode return in Hopper and Walker. As we mentioned in Section 6, the safety problem in MuJoCo is more suitable to be formulated as a constrained problem optimizing the average reward, which is a promising direction for future research.
Figure 3: Additional comparison of APO and PPO for the average return in MuJoCo tasks.

Figure 4: Additional comparison of APO and PPO for the average reward in MuJoCo tasks.