An Analytical Update Rule for General Policy Optimization

Hepeng Li [1] **Nicholas Clavette** [1] **Haibo He** [1]

Abstract

We present an analytical policy update rule that is independent of parametric function approximators. The policy update rule is suitable for optimizing general stochastic policies and has a monotonic improvement guarantee. It is derived from a closed-form solution to trust-region optimization using calculus of variation, following a new theoretical result that tightens existing bounds for policy improvement using trust-region methods. The update rule builds a connection between policy search methods and value function methods. Moreover, off-policy reinforcement learning algorithms can be derived from the update rule since it does not need to compute integration over on-policy states. In addition, the update rule extends immediately to cooperative multi-agent systems when policy updates are performed by one agent at a time.

1. Introduction

Policy search methods have gained great popularity in reinforcement learning (RL) for the last decade. As opposed to value function methods, in which the policy is represented implicitly by a greedy action-selection strategy with respect to an estimated value function, policy search methods search directly in the space of policy representations for a good policy. The advantages of policy search methods include being able to learn stochastic policies (Singh et al., 1994), better convergence, and effectiveness in high-dimensional or continuous action spaces. Generally, policy search approaches use function approximators, such as neural networks, to construct a parametric policy. The parametric policy is then optimized using policy gradient (Williams, 1992; Sutton et al., 1999) or derivative-free algorithms (Szita & Lörincz, 2006) by searching in the parameter space.

[1]Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, South Kingstown, RI, USA. Correspondence to: Haibo He <haibohe@uri.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

In this paper, we present an analytical policy update rule that is independent of parametric function approximators. We prove that the update rule has a monotonic improvement guarantee and is suitable for optimizing general stochastic policies with continuous or discrete actions. The update rule provides a new theoretical foundation for policy-based RL, which traditionally restricts the policy search to a family of parametric functions, such as policy gradient (Sutton et al., 1999), deterministic policy gradient (Silver et al., 2014; Lillicrap et al., 2016), actor critic (Konda & Tsitsiklis, 1999; Degris et al., 2012), soft actor-critic (SAC) (Haarnoja et al., 2018a;b), and so on.

Our update rule is derived from a closed-form solution to a trust region method using calculus of variation. Trust-region method is one of the most important tools in RL. The basic idea is to search for an improved policy iteratively in a local area around the current policy, in which the objective function is well-approximated by a manageable surrogate model. A representative trust-region method for RL is trust region policy optimization (TRPO) (Schulman et al., 2015). TRPO introduces a simple and functional surrogate model that can be evaluated using the current best policy and provides an upper bound of the approximation error of the surrogate model. This is particularly useful because by subtracting the bound from the surrogate model we obtain the worst-case performance degradation, or a lower bound, of the true objective. It follows that maximizing the lower bound leads to an improved policy with non-decreasing performance (Schulman et al., 2015).

The theory of TRPO is of significance to policy-based RL for it provides an approach that guarantees to improve the policy monotonically. However, the bound derived in TRPO depends on the maximum KL-Divergence of the current policy π and a proposed policy π' on the entire state space, i.e., $\max_s D_{\mathrm{KL}}[\pi' \| \pi](s)$, which can be extremely large or infinity even if π' and π are close at most states. To address this issue, TRPO heuristically imposes a strict constraint to bound the KL-Divergence at every state, but it is intractable to implement this constraint when the state space is large or continuous. To derive a practical algorithm, an empirical approximation using an expected KL-Divergence, e.g. $\mathbb{E}_{s\sim d^{\pi}}[D_{\mathrm{KL}}[\pi' \| \pi](s)]$, is usually adopted (Schulman et al., 2015; Achiam et al., 2017). Nevertheless, the monotonic improvement property is no longer guaranteed.

In this paper, we prove a new theoretical result on the bound of the surrogate approximation error by relating it to the expected KL-Divergence. This result leads to a more practical lower bound of the objective, which improves previous analysis on this topic in terms of KL-Divergence, such as Schulman et al. (2015); Achiam et al. (2017); Akrour et al. (2018). It also closes the gap between theory and practice in TRPO and the related approaches. Furthermore, this result enables us to derive a closed-form solution for policy optimization. The closed-form solution introduces a very simple policy update rule that guarantees to produce monotonically improving policies.

From an algorithmic viewpoint, the policy update rule enables the development of off-policy algorithms that do not rely on policy gradient (Sutton et al., 1999), which is known to have high variance and low sample efficiency. This is because the policy update rule does not require integrating on-policy distributions over the state space. Thus, we can reuse the past experience obtained from a behavioral policy and circumvent the high variance and sample efficiency issues. In addition, since the policy update rule is analytical, it applies to both parametric and non-parametric policies. However, policy gradient-based approaches are subject to parametric policies.

Furthermore, we prove that the update rule extends immediately to partially observable Markov games with cooperative agents and the monotonic improvement guarantee still holds when updates are performed by one agent at a time.

The contributions of this paper include: (1) a new theoretical result that tightens existing bounds for local policy search using trust-region methods; (2) a closed-form update rule for general stochastic policies with monotonic improvement guarantee; (3) a proof that shows that the policy update rule is extendable to partially observable multi-agent RL problems without compromising the monotonic improvement guarantee.

2. Related Work

The idea of restricting policy search to a local area of the current policy is common in model-free RL. For instance, instead of imposing a hard boundary on the searching area, Kakade & Langford (2002) proposed a conservative update scheme mixing the current policy and a greedy update via a weighted sum. A lower bound on the performance improvement as a function of the weighting coefficient was proven. Following this line of work, Pirotta et al. (2013) proposed two more general lower bounds connecting to the difference between two policies. Then, two conservative update algorithms were developed by maximizing the proposed bounds, respectively. Zhu & Matsubara (2020) proposed a similar bound and a practical algorithm for entropy-regularized RL. While monotonic improvement guarantee is derived in the previous studies, the update scheme cannot apply to non-mixture policies. Schulman et al. (2015) extended this line of work to general stochastic policies and proposed a new bound that connected it to the maximum KL-Divergence between two successive policies on the state space. However, this bound is intractable when the state space is large. Although a tighter bound relating it to an average total variation distance is proposed in Achiam et al. (2017), deriving a closed-form policy update rule from the lower bound is still challenging.

In practice, many approaches use a hard constraint to bound the searching area but they generally lose the monotonic improvement guarantee. Peters et al. (2010) proposed relative entropy policy search (REPS) to restrict the relative entropy between observed data distribution of the state-action pairs and the distribution generated by the new policy. A closed-form update rule in a softmax form was derived using the method of Lagrange multipliers. However, this approach is not straightforwardly extendable to general non-linear policies. To apply nonlinear policies, TRPO (Schulman et al., 2015) and constrained policy optimization (Achiam et al., 2017) approximately constrained the on-policy expected KL-Divergence by using second-order Taylor expansion, which was closely related to natural policy gradient (Kakade, 2001). Extending the work in TRPO, Akrour et al. (2018) provided a monotonic improvement guarantee for bounding the expected KL-divergence, but the result only held for linear-Gaussian policies. Nachum et al. (2017; 2018) presented multi-step softmax consistencies under entropy regularization and adopted a discounted relative entropy trust-region constraint to improve exploration and stability. By relating policy search to probabilistic inference (Levine, 2018), Abdolmaleki et al. (2018) proposed the maximum a posterior policy optimization (MPO) algorithm based on Expectation-Maximization, where the policy update was decomposed into E-step and M-step. A closed-form E-step combined with a maximum-a-posteriori-estimation M-step for Gaussian policies was provided. Although a monotonic improvement guarantee is claimed, the guarantee is for the KL-Divergence regularized objective rather than the true expected return. Besides, suffering from the same issue as in (Peters et al., 2010), the policy update rule needs to determine the optimal Lagrangian multipliers of the dual problem, which requires a costly nonlinear optimization in the inner loop. Different from previous works, Otto et al. (2021) proposed projection-based solutions to impose trust-region constraints on the individual state, which enabled exact guarantees of monotonic improvement. Three closed-form projection layers based on Wasserstein L2 distance, Frobenius norm, and KL-Divergence were proposed to project the updated policy onto trust regions. However, the proposed approach only applies to Gaussian policies.

3. Preliminaries

3.1. Markov Decision Process

A Markov decision process (MDP) is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, \rho_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the transition probability density, $r : \mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$ is the reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the probability density of the initial state s_0, $\gamma \in [0, 1)$ is the discount factor.

Denote a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ ($\to [0, 1]$ for discrete actions) by $\pi(a|s)$, which represents the probability density (or probability mass function) of the action a given the state s. The goal is to find an optimal policy that maximizes the expected discounted return

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] \tag{1}$$

where τ denotes the trajectory $\tau := (s_0, a_0, s_1, \dots)$, and $\tau \sim \pi$ indicates that the distribution over the trajectory depends on $\pi : s_0 \sim \rho_0, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim p(\cdot|s_t, a_t)$. Letting $R(\tau) = \sum_{t=0}^{\infty}\gamma^t r_t$ denote the discounted return of the trajectory τ, we can compactly express the value function as $V_\pi(s) = \mathbb{E}_{\tau\sim\pi}[R(\tau)|s_0 = s]$, the state-action value function as $Q_\pi(s, a) = \mathbb{E}_{\tau\sim\pi}[R(\tau)|s_0 = s, a_0 = a]$, and the advantage function as $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$. We define the discounted state visitation distribution as

$$d^\pi(s) = (1 - \gamma)\sum_{t=0}^{\infty}\gamma^t\rho_t^\pi(s), \tag{2}$$

where $\rho_t^\pi : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is probability density function (PDF) of the state at timestep t given the policy π.

3.2. Partially Observable Markov Game

A Markov game (Littman, 1994) is a game defined on a state space, $\mathcal{S}$, and a collection of action spaces, $\mathcal{A}^1, \dots, \mathcal{A}^N$, one for each agent in the environment. The state transition $s \mapsto s^{'}(s, s^{'} \in \mathcal{S})$ happens following the probability density $P : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \times \mathcal{S} \mapsto \mathbb{R}_{\geq 0}$ when the actions $a = [a^1, \dots, a^N]$, $a^i \in \mathcal{A}^i$, $i \in \{1, \dots, N\}$, are exerted on the environment at state s. Each agent is rewarded based on a local reward function $r^i : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \mapsto [r^i_{\min}, r^i_{\max}]$, which depends on the current state s and the joint action a.

In a partially observable Markov game (POMG), each agent has a local observation of the environment, o^i, which contains incomplete information of the state s. At state s, o^i is observed with a likelihood, $P_o^i : \mathcal{S} \times \mathcal{O}^i \mapsto \mathbb{R}_{\geq 0}$, where $\mathcal{O}^i$ is the observation space of the agent. Each agent acts according to a policy $\pi^i : \mathcal{O}^i \times \mathcal{A}^i \mapsto \mathbb{R}_{\geq 0}$ (or $\mapsto [0, 1]$), which is a probability distribution (or a probability mass function) over the action space $\mathcal{A}^i$ given the observation o^i. We will use the following definitions of the joint policy $\pi(a|s)$ and the joint policy $\pi^{-i}(a^{-i}|s)$ except i:

$$\pi(a|s) = \prod_{i\in\mathcal{N}}\int_{o^i}\pi^i(a^i|o^i)P_o^i(o^i|s)do^i \tag{3}$$

$$\pi^{-i}(a^{-i}|s) = \prod_{j\in\mathcal{N}\setminus\{i\}}\int_{o^j}\pi^j(a^j|o^j)P_o^j(o^j|s)do^j \tag{4}$$

where $\mathcal{N} = \{1, 2, \dots, N\}$ is a set of agent's IDs. The goal of the agents is to learn a set of distributed policies $\{\pi^i(a^i|o^i)|i \in \mathcal{N}\}$ to maximize the expected return

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r_t^1 + \cdots + r_t^N)\right] \tag{5}$$

where $\tau \sim \pi$ indicates that $s_0 \sim \rho_0, o_t^i \sim P_o^i(\cdot|s_t), a_t^i \sim \pi^i(\cdot|o_t^i), s_{t+1} \sim P(\cdot|s_t, a_t^1, ..., a_t^N)$.

3.3. Trust Region Method

Trust region method is one of the most important techniques for solving policy optimization in a Markov decision process. It works by restricting policy search to a local region around the current best solution, where the objective function is well-approximated by a surrogate model. Specifically, it solves the following optimization:

$$\pi_{k+1} = \arg\max_{\pi'\in\Pi}\tilde{J}(\pi'),\ s.t.\ D(\pi', \pi_k) \leq \delta \tag{6}$$

where $\tilde{J}(\pi')$ is some surrogate model, D is a distance measure, and $\delta > 0$ is the radius of a spherical region, in which we search for an improved policy. A simple and effective choice for the surrogate model is

$$L_{\pi_k}(\pi') = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi'}[A_{\pi_k}(s, a)]. \tag{7}$$

Schulman et al. (2015) prove that the difference between the surrogate model and the true objective is bounded by:

$$\begin{aligned}&\left|J(\pi') - L_{\pi_k}(\pi')\right| \leq C\max_s D_{\mathrm{KL}}[\pi'\|\pi_k](s),\\ &\textit{where } C = \frac{4\gamma\epsilon}{(1-\gamma)^2},\ \epsilon = \max_{s,a}|A_{\pi_k}(s, a)|,\end{aligned} \tag{8}$$

which connects it to the maximum KL-Divergence over the state space, $\max_s D_{\mathrm{KL}}[\pi'\|\pi_k](s)$. By using this bound, we can get the worst-case performance degradation of the true objective:

$$J(\pi') \geq L_{\pi_k}(\pi') - C\max_s D_{\mathrm{KL}}[\pi'\|\pi_k](s), \tag{9}$$

It follows that maximizing the right-hand side of the inequality, which is a lower bound of the true objective function, can lead to guaranteed improvement in the performance. This result has fostered a branch of practical trust-region algorithms (i.e. Schulman et al. (2015; 2017); Achiam et al. (2017); Nachum et al. (2018); Wu et al. (2017)) that approximately optimize the lower bound to improve policies.

4. Analytical Policy Update Rule with Monotonic Improvement Guarantee

Our principle result is an analytical solution for policy optimization based on trust-region methods, following a new bound on the difference between the surrogate model and the true objective. The analytical solution introduces a policy update rule that guarantees monotonic policy improvement and is suitable for general stochastic policies with discrete or continuous actions. Moreover, the update rule extends immediately to cooperative multi-agent systems when updates are performed by one agent at a time.

We first present the new bound on the difference between the surrogate model and the objective in the following theorem.

Theorem 4.1. *For any stochastic policies π', π and discount factor $\gamma \in [0.5, 1)$, the following bound holds:*

$$\left|J(\pi') - L_\pi(\pi')\right| \leq \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi'\|\pi](s)\right],$$
$$\text{where } C_\pi = \frac{\gamma^2 \epsilon}{(1-\gamma)^3},\ \epsilon = \max_{s,a} |A_\pi(s,a)|. \tag{10}$$

Proof. We provide the proof in Appendix A. The proof extends Schulman et al. (2015)'s result using the concept of α-coupling (Levin et al., 2006) and its relationship with total variation distance. However, different from the proof in (Schulman et al., 2015) that uses the maximum α over the state space, we instead use a state-dependent $\alpha(s)$ to represent the coupling between two arbitrary policies given s, which enables us to connect the bound to the expected KL-Divergence. □

The new bound is tighter in terms of KL-Divergence compared with (8) derived from (Schulman et al., 2015). While the improvement in tightness is at a cost of $\gamma/(4(1-\gamma)^2)$, this result directly relates the bound to the expected KL-Divergence $\mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi'\|\pi](s)\right]$, which closes the gap between theory and practice in TRPO and related algorithms. In addition, the new bound improves prior analysis in the literature, such as (Akrour et al., 2018; Achiam et al., 2017), in terms of either KL-Divergence or total variation distance (from $D_{\mathrm{TV}}[\pi'\|\pi]$ to $D^2_{\mathrm{TV}}[\pi'\|\pi]$, see Appendix A). Furthermore, using this result, we can derive a new lower bound of the true objective:

$$J(\pi') \geq L_{\pi_k}(\pi') - \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi'\|\pi](s)\right]. \tag{11}$$

Then, we can improve the policy by maximizing the lower bound. Next, we present a closed-form solution to the maximization of the lower bound, which introduces a simple policy update rule with monotonic improvement guarantee.

Theorem 4.2. *For any stochastic policies $\pi_{\mathrm{new}}, \pi_{\mathrm{old}}$ that are continuously differentiable on the state space $\mathcal{S}$, the inequality, $J(\pi_{\mathrm{new}}) \geq J(\pi_{\mathrm{old}})$, holds when*

$$\pi_{\mathrm{new}} = \pi_{\mathrm{old}} \cdot \frac{e^{\alpha_{\pi_{\mathrm{old}}}}}{\mathbb{E}_{a\sim\pi_{\mathrm{old}}}\left[e^{\alpha_{\pi_{\mathrm{old}}}}\right]}, \tag{12}$$

where $\alpha_{\pi_{\mathrm{old}}} = A_{\pi_{\mathrm{old}}}/C_{\pi_{\mathrm{old}}}$.

Proof. We provide the proof in Appendix B. The proof introduces calculus of variation (Calder, 2020; Kot, 2014) to the policy optimization problem. Based on the assumption of continuously differentiable policies on the state space $\mathcal{S}$, we derive a closed-form solution for general stochastic policies with continuous or discrete actions. In the proof, we show that the closed-form solution is a necessary and sufficient condition for the policy optimization. □

Another interesting result of Theorem 4.2 is that the update rule immediately extends to cooperative multi-agent RL problems while the monotonic improvement guarantee still holds if the agents perform local policy updates in turn. We present this result in the following corollary.

Corollary 4.3. *For any stochastic policies $\pi^i_{\mathrm{new}}, \pi^i_{\mathrm{old}}$ of agent i that are continuously differentiable on the local observation space $\mathcal{O}^i$, and the corresponding joint policies $\pi_{\mathrm{new}}, \pi_{\mathrm{old}}$, the inequality, $J(\pi_{\mathrm{new}}) \geq J(\pi_{\mathrm{old}})$, holds when*

$$\pi^i_{\mathrm{new}} = \pi^i_{\mathrm{old}} \cdot \frac{e^{\alpha_{\pi_{\mathrm{old}}}}}{\mathbb{E}_{a\sim\pi_{\mathrm{old}}}\left[e^{\alpha_{\pi_{\mathrm{old}}}}\right]},$$
$$\pi^{-i}_{\mathrm{new}} = \pi^{-i}_{\mathrm{old}}. \tag{13}$$

where $\pi^{-i}_{\mathrm{new}}, \pi^{-i}_{\mathrm{old}}$ are joint policies of all agents except i.

Proof. Based on Theorem 4.2, we have

$$\pi_{\mathrm{new}}(a|s) = \pi_{\mathrm{old}}(a|s) \frac{e^{\alpha_{\pi_{\mathrm{old}}}}}{\mathbb{E}_{a\sim\pi_{\mathrm{old}}}\left[e^{\alpha_{\pi_{\mathrm{old}}}}\right]} \tag{14}$$

Note that the joint policy can be decomposed as follows:

$$\pi(a|s) = \pi^i(a^i|s)\pi^{-i}(a^{-i}|s), \tag{15}$$

where $\pi^i(a^i|s) = \int_{o^i} \pi^i(a^i|o^i) P^i_o(o^i|s) do^i$. Thus, we can rewrite Eq. (14) as follows:

$$\pi^{-i}_{\mathrm{new}} \int_{\mathcal{O}^i} \pi^i_{\mathrm{new}}(a^i|o^i) P^i_o(o^i|s) do^i$$
$$= \pi^{-i}_{\mathrm{old}} \int_{\mathcal{O}^i} \pi^i_{\mathrm{old}}(a^i|o^i) P^i_o(o^i|s) do^i \cdot \frac{e^{\alpha_{\pi_{\mathrm{old}}}}}{\mathbb{E}_{a\sim\pi_{\mathrm{old}}}\left[e^{\alpha_{\pi_{\mathrm{old}}}}\right]}, \tag{16}$$

when $\pi^{-i}_{\mathrm{new}} = \pi^{-i}_{\mathrm{old}}$, they cancel each other on both sides. Then, simplifying the above equation, the result follows. □

5. Connections with Prior Work

In this section, we connect the proposed policy update rule with some state-of-the-art algorithms and discuss how the update rule can help explain these algorithms from a different perspective.

5.1. TRPO and Proximal Policy Optimization

Note that the exponential factor in (12) can be written as

$$\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{\max_{s,a}|A_{\pi_{\text{old}}}(s,a)|} \cdot \frac{(1-\gamma)^3}{\gamma^2}, \tag{17}$$

where $\gamma \in [0.5, 1)$. The first term on the right-hand side is a normalized advantage and the second term is a positive constant smaller than 1. Letting $[\alpha_{\min}, \alpha_{\max}]$ denote the range of $\alpha_{\pi_{\text{old}}}$, then we have $\alpha_{\min} \leq 0 \leq \alpha_{\max}$, as shown in Figure 1. In addition, since $\alpha_{\pi_{\text{old}}}$ is a random variable given s, we have $e^{\alpha_{\min}} \leq \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}] \leq e^{\alpha_{\max}}$. Then, based on the update rule (12), the ratio of the new policy to the old policy is bounded by

$$\frac{\pi_{\text{new}}}{\pi_{\text{old}}} \in \left[\frac{e^{\alpha_{\min}}}{Z}, \frac{e^{\alpha_{\max}}}{Z}\right] = [1-\epsilon_1, 1+\epsilon_2], \tag{18}$$

where $Z = \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$ and ϵ_1, ϵ_2 are positive numbers ($\epsilon_1 < 1$). Equation (18) indicates that bounding the policy ratio is an effective way to confine the searching area. This help explain the success of the proximal policy optimization (PPO) algorithm (Schulman et al., 2017), which clips the policy ratio by $[1-\epsilon, 1+\epsilon], 0 < \epsilon < 1$.

It is also noted that the policy ratio $\pi_{\text{new}}/\pi_{\text{old}}$ will be greater than 1 if $e^{\alpha_{\pi_{\text{old}}}} > \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$, and vice versa (shown in Figure 1). Note that the exponential term $e^{\alpha_{\pi_{\text{old}}}}$ is monotonically increasing with respect to $A_{\pi_{\text{old}}}(s,a)$, and so is the policy ratio $\pi_{\text{new}}/\pi_{\text{old}}$. Less rigorously, consider the term $\mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$ as an "average" advantage of the policy π_{old}. Then, selecting the action a at state s is encouraged, i.e. $\pi_{\text{new}}(a|s) > \pi_{\text{old}}(a|s)$, if it leads to an advantage that is above average. On the contrary, selecting the action a at state s is discouraged, i.e. $\pi_{\text{new}}(a|s) < \pi_{\text{old}}(a|s)$, if it leads to an advantage that is below average. To what extent the action a is encouraged or discouraged is determined by the value of $A_{\pi_{\text{old}}}(s,a)$. This result matches the TRPO algorithm (Schulman et al., 2015), which maximizes

$$\max_{\pi} \mathbb{E}_{s\sim d^{\pi_{\text{old}}}, a\sim\pi_{\text{old}}}\left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A_{\pi_{old}}(s,a)\right], \tag{19}$$

where $\pi(a|s)$ is increased to gain weights for large advantages and decreased to lose weights for small advantages. Although our update rule suggests that the policy ratio is proportional to an *exponential* advantage, rather than a *linear* advantage as suggested in TRPO and PPO, it is easy to verify that $e^{A_{\text{old}}/C_{\text{old}}} \approx A_{\text{old}}/C_{\text{old}} + 1$ when the policy ratio is bounded around 1.

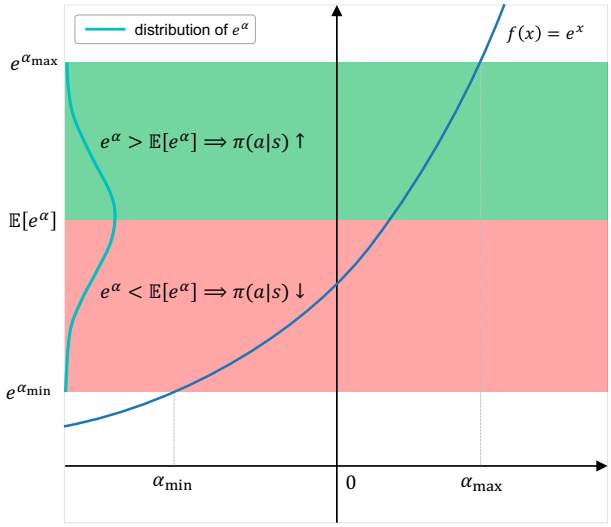

Figure 1. From the policy update rule, we can derive $\pi_{\text{new}}/\pi_{\text{old}} \in [e^{\alpha_{\min}}/Z, e^{\alpha_{\max}}/Z]$, where $Z = \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$ and $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum values of $\alpha_{\pi_{\text{old}}}$, respectively. Since $e^{\alpha_{\min}} \leq \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}] \leq e^{\alpha_{\max}}$, the upper and lower bounds of the policy ratio $\pi_{\text{new}}/\pi_{\text{old}}$ can be expressed as $[1-\epsilon_1, 1+\epsilon_2]$, where $\epsilon_1, \epsilon_2 \geq 0$ and $\epsilon_1 < 1$. This indicates that we can bound the policy ratio to restrict the search area, which has been adopted in PPO and proven effective in practice. In addition, if $e^{\alpha_{\pi_{\text{old}}}} > \mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$, the policy ratio $\pi_{\text{new}}/\pi_{\text{old}}$ will be greater than 1, and vice versa. Note that $e^{\alpha_{\pi_{\text{old}}}}$ is monotonically increasing with respect to $A_{\pi_{\text{old}}}(s,a)$, and so is the ratio $\pi_{\text{new}}/\pi_{\text{old}}$. Less rigorously, consider the term $\mathbb{E}_{a\sim\pi_{\text{old}}}[e^{\alpha_{\pi_{\text{old}}}}]$ as the "average" advantage of the policy π_{old}. Then, selecting the action a at state s is encouraged, i.e. $\pi_{\text{new}}(a|s) > \pi_{\text{old}}(a|s)$, if it leads to an "above average" advantage. On the contrary, selecting the action a at state s is discouraged, i.e. $\pi_{\text{new}}(a|s) < \pi_{\text{old}}(a|s)$, if it leads to a "below average" advantage. To what extent the action a is encouraged or discouraged is determined by the value of $A_{\pi_{\text{old}}}(s,a)$.

5.2. Value-Based Methods and Dynamic Programming

In this section, we provide a different explanation of the policy update rule by considering discrete actions and then connect it to value function methods. By multiplying the numerator and denominator both by $e^{V_{\pi_{\text{old}}}(s)/C_{\pi_{\text{old}}}}$, we can rewrite the update rule as

$$\begin{aligned} \pi_{\text{new}}(a^i|s) &= \frac{\pi_{\text{old}}(a^i|s)\omega^i_{\text{old}}}{\sum_j \pi_{\text{old}}(a^j|s)\omega^j_{\text{old}}}, \\ \omega^i_{\text{old}} &= \exp\{Q_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}\} \end{aligned} \tag{20}$$

As shown in (20), the new policy is a weighted probability mass function of the old policy in a softmax form. The weights are the exponential terms, $\exp\{Q_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}\}$. That indicates actions with larger Q values will get better chance to be selected in the future. In fact, the policy update

rule can be deemed as a stochastic analogy of the ϵ-greedy policy used in value function methods, such as SARSA (Sutton & Barto, 2018).

In addition, we can verify the monotonic improvement guarantee of the policy update rule via dynamic programming. To see this, we will show $V_{\pi_{\text{old}}}(s) \leq V_{\pi_{\text{new}}}(s)$ for all $s \in \mathcal{S}$. Note that

$$\begin{aligned} V_{\pi_{\text{old}}}(s) &= \sum_i \pi_{\text{old}}(a^i|s) Q_{\pi_{\text{old}}}(s, a^i) \\ &\leq \sum_i \frac{\pi_{\text{old}}(a^i|s)\omega^i_{\text{old}}}{\sum_j \pi_{\text{old}}(a^j|s)\omega^j_{\text{old}}} Q_{\pi_{\text{old}}}(s, a^i) \qquad (21) \\ &= \sum_i \pi_{\text{new}}(a^i|s) Q_{\pi_{\text{old}}}(s, a^i). \end{aligned}$$

For brevity, we will use $P^{s'}_{sa} := P(s'|s, a)$. Then, we have

$$\begin{aligned} V_{\pi_{\text{old}}}(s) &\leq \mathbb{E}_{\pi_{\text{new}}}\left[Q_{\pi_{\text{old}}}(s, a)\right] \\ &= \mathbb{E}_{\pi_{\text{new}}}\left[r(s, a) + \gamma \mathbb{E}_{P^{s'}_{sa}}[V_{\pi_{\text{old}}}(s')]\right] \\ &\leq \mathbb{E}_{\pi_{\text{new}}}\left[r(s, a) + \gamma \mathbb{E}_{P^{s'}_{sa}}\left[\mathbb{E}_{\pi_{\text{new}}}[Q_{\pi_{\text{old}}}(s', a')]\right]\right] \\ &\;\;\vdots \\ &\leq \mathbb{E}_{\pi_{\text{new}}}\left[r(s, a) + \gamma \mathbb{E}_{P^{s'}_{sa}}[r(s', a')] + \cdots\right] \\ &= V_{\pi_{\text{new}}}(s). \end{aligned} \qquad (22)$$

Therefore, by applying the update rule (20), we can obtain a sequence of monotonically improving policies and value functions:

$$\pi_0 \to V_{\pi_0} \to \pi_1 \to V_{\pi_1} \to \cdots \to \pi_* \to V_{\pi_*},$$

where $V_{\pi_0}(s) \leq V_{\pi_1}(s) \leq \cdots \leq V_{\pi_*}(s)$ for all $s \in \mathcal{S}$.

5.3. Relative Entropy Policy Search and Maximum a Posterior Policy Optimization

The REPS (Relative Entropy Policy Search) algorithm (Peters et al., 2010) can be obtained as a special case of the update rule by replacing π_{old} with the observed data distribution and the coefficient $C_{\pi_{\text{old}}}$ with the Lagrange multiplier η. However, the REPS algorithm is based on finite MDPs with discrete actions and not extendable to general continuous policies. A similar closed-form update rule has also been derived in the MPO (Maximum a posterior Policy Optimization) algorithm (Abdolmaleki et al., 2018) in its E-step for evaluating a variational policy, which is then used to optimize policy parameters.

Our policy update rule is different from the previous work because it directly expresses the new policy as a closed-form function of the current policy. That means the policy update can be accurately calculated using the current policy without involving policy gradient or policy optimization. Especially, the proposed update rule provides an explicit formula for determining the coefficient $C_{\pi_{\text{old}}}$ and guarantees monotonic improvement on performance. However, the update rules in (Peters et al., 2010; Abdolmaleki et al., 2018) need to numerically determine the optimal Lagrangian multiplier η, which requires a costly nonlinear optimization in the inner loop and no monotonic improvement is guaranteed.

5.4. Soft Actor-Critic

The SAC (Soft Actor-Critic) algorithm (Haarnoja et al., 2018a;b) can also be derived as a special case of the policy update rule. Note that the update rule (12) can be expressed as a Gibbs measure (Boltzmann distribution in case of discrete actions):

$$\begin{aligned} \pi_{\text{new}}(a|s) &= \pi_{\text{old}}(a|s) \frac{e^{A_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{A_{\pi_{\text{old}}}(s,a)}/C_{\pi_{\text{old}}}\right]} \\ &= \pi_{\text{old}}(a|s) \frac{e^{Q_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{Q_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}}}\right]} \qquad (23) \\ &= \frac{1}{Z} \exp\left\{\frac{Q_{\pi_{\text{old}}}(s, a)}{C_{\pi_{\text{old}}}} + \log \pi_{\text{old}}(a|s)\right\}, \end{aligned}$$

where $Z = \mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{Q_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}}}\right]$ is the partition function.

To optimize a policy π, we can minimize the KL-Divergence between π and π_{new}:

$$\min_\pi \; D_{\text{KL}}\left(\pi(\cdot|s) \middle\| \frac{\exp\left(\frac{1}{C_{\pi_{\text{old}}}} \widetilde{Q}_{\pi_{\text{old}}}(s, \cdot)\right)}{Z}\right), \qquad (24)$$

where $\widetilde{Q}_{\pi_{\text{old}}}$ is the soft Q-function:

$$\widetilde{Q}_{\pi_{\text{old}}}(s, a) = Q_{\pi_{\text{old}}}(s, a) + C_{\pi_{\text{old}}} \log \pi_{\text{old}}(a|s). \qquad (25)$$

Replacing $C_{\pi_{\text{old}}}$ with a temperature parameter α, we immediately get the SAC algorithm (Haarnoja et al., 2018b).

A slight difference of the algorithm (24) than SAC is that it minimizes the policy entropy instead of maximizing it. Note that the soft state value function derived from our update rule is given by

$$\widetilde{V}_\pi(s) = \mathbb{E}_{a\sim\pi}[\widetilde{Q}_\pi(s, a)] = V_\pi(s) - C_\pi \mathcal{H}(\pi(\cdot|s)), \qquad (26)$$

where $\mathcal{H}(\pi(\cdot|s))$ is the policy entropy. Since C_π is always positive, the policy entropy is penalized in the soft state value function. Thus, applying (24) will minimize the policy entropy. This is reasonable because the policy distribution should be concentrating more and more on the optimal action as the policy improves monotonically.

The derivation of SAC also verifies that the update rule is essentially off-policy.

6. Limitations and Discussions

6.1. Tightness of the Bound in Terms of γ

The bound in Theorem 4.1 improves prior analysis in terms of KL-Divergence, but not in terms of γ, which could be a limitation of the policy update rule. Compared to the bound in TRPO, the improvement is at a cost of $\gamma/(4(1-\gamma)^2)$. When γ is close to 1, the penalty coefficient $C_{\pi_{\text{old}}}$ for the KL-Divergence can be large, resulting in small step sizes for policy updates. While $C_{\pi_{\text{old}}}$ can be tuned to allow larger step-sizes in practice, a proven bound that is tighter in terms of γ will be an interesting direction for future work.

6.2. Monotonic Guarantee and Function Approximation

The policy update rule is a closed-form solution, so it assumes an exact advantage function and an exact maximum of its absolute value. In large MDPs, these quantities generally need to be estimated by function approximators. The use of function approximatiors will inevitably introduce errors and can undermine the monotonic improvement guarantee. While our goal is to provide the theory, we would like to clarify this to encourage the development of efficient algorithms using function approximation. We also look forward to new RL theories building upon the update rule given its simplicity and wide connections with prior RL approaches.

6.3. Simultaneous Update for Multi-Agent RL

The extension of the update rule to multi-agent RL requires agents to take turns updating their policies. Thus, the learning process could be slow if there are many agents. From Equation (16) we see that the main reason for this requirement is that we need to make sure $\pi^{-i}_{\text{new}} = \pi^{-i}_{\text{old}}$. We believe that relaxing this requirement so as for the agents to update policies simultaneously without jeopardizing the monotonic improvement guarantee is worth studying in the future.

7. Conclusion

We have presented a closed-form update rule for general stochastic policy optimization with monotonic improvement guarantee. A new theoretical result has been provided by relating the lower bound of the performance to an expected KL-Divergence, which closes the gap between theory and practice in the literature. Based on the theoretical result, calculus of variation has been introduced to derive the policy update rule. Furthermore, we have proved that the policy update rule is extendable to cooperative multi-agent RL when agents take turns performing policy updates. Since the proposed update rule is analytical, we hope that it serves as a stepping stone for future work on novel RL theories and principled RL algorithms using parametric or non-parametric policies.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. ECCS 1917275.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a posteriori policy optimisation. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR'18, Vancouver, Canada, Apr 30 - May 3 2018.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 22–31, Sydney, NSW, Australia, 2017. JMLR.org.

Akrour, R., Abdolmaleki, A., Abdulsamad, H., Peters, J., and Neumann, G. Model-free trajectory-based policy optimization with monotonic improvement. *Journal of Machine Learning Research*, 19(1):565–589, Jan. 2018. ISSN 1532-4435.

Calder, J. *The Calculus of Variations*. 2020. URL http://www-users.math.umn.edu/~jwcalder/8385F19/CalculusOfVariations.pdf.

Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. In *Proceedings of the 29th International Coference on Machine Learning*, ICML'12, pp. 179–186, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018a.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. Soft actor-critic algorithms and applications, 2018b. URL https://arxiv.org/abs/1812.05905.

Kakade, S. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 1531–1538, Cambridge, MA, USA, 2001. MIT Press.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Kot, M. *A First Course in the Calculus of Variations*. American Mathematical Society, 2014.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. American Mathematical Society, 2006.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *Proceedings of the 4th International Conference on Learning Representations*, ICLR'16, San Juan, Puerto Rico, May 2-4 2016.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, ICML'94, pp. 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2772–2782, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Trust-PCL: An off-policy trust region method for continuous control. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR'18, Vancouver, Canada, Apr 30 - May 3 2018.

Otto, F., Becker, P., Ngo, V. A., Ziesche, H. C. M., and Neumann, G. Differentiable trust region layers for deep reinforcement learning. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR'21, May 3-7 2021.

Peters, J., Mülling, K., and Altün, Y. Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pp. 1607–1612, Atlanta, Georgia, 2010. AAAI Press.

Pirotta, M., Restelli, M., Pecorino, A., and Calandriello, D. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 307–315, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1889–1897, Lille, France, 2015. JMLR.org.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–387–I–395, Beijing, China, 2014. JMLR.org.

Singh, S. P., Jaakkola, T. S., and Jordan, M. I. Learning without state-estimation in partially observable markovian decision processes. In *Proceedings of the 11th International Conference on Machine Learning*, ICML'94, pp. 284–292, New Brunswick, NJ, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

Szita, I. and Lörincz, A. Learning tetris using the noisy cross-entropy method. *Neural Computation*, 18(12):2936–2941, 2006. doi: 10.1162/neco.2006.18.12.2936.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696.

Wu, Y., Mansimov, E., Liao, S., Grosse, R., and Ba, J. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5285–5294, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Zhu, L. and Matsubara, T. Ensuring monotonic policy improvement in entropy-regularized value-based reinforcement learning, 2020. URL https://arxiv.org/abs/2008.10806.

A. Proof of Policy Performance Bound

This proof uses techniques from the proof of Lemma 3. in (Schulman et al., 2015), exploiting them to derive a new bound that relates to an average divergence between policies, π', π. An informal overview is as follows. First, using Lemma 1. in (Schulman et al., 2015), the gap between the surrogate and the objective is decomposed into the difference of two expected advantages over the policies π', π. Then, we use the *coupling* technique to measure the coincidence of two trajectories resulted from π', π before an arbitrary timestep t. Finally, we constrain the gap to an average KL-Divergence using Pinsker's inequality.

Definition A.1 (Notations)**.** We consider a Markov decision process with a continuous state space. The following definitions and notations will be used.

1. Probability density function (PDF) of the state at timestep t given the policy π:

$$\rho_t^\pi(s) = PDF(s_t = s|\pi).$$

Note that $\rho_0^\pi(s) = \rho_0(s)$ is the PDF of the initial state s_0, which is independent of π.

2. Discounted state visitation PDF:

$$\begin{aligned} d^\pi(s) &= (1-\gamma)\big[\rho_0^\pi(s) + \gamma\rho_1^\pi(s) + \gamma^2\rho_2^\pi(s) + \cdots\big] \\ &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho_t^\pi(s). \end{aligned} \tag{27}$$

3. One-step state transition density given the policy π:

$$\nu_\pi(s'|s) = \int_{\mathcal{A}} p(s'|s,a)\pi(a|s)da. \tag{28}$$

4. t-step state transition density given the policy π (the Chapman Kolmogorov equation):

$$\nu_\pi^t(s'|s) = \int_{\mathcal{S}} \nu_\pi^m(s'|\tilde{s})\nu_\pi^{t-m}(\tilde{s}|s)d\tilde{s}, \tag{29}$$

where $0 \le m \le t$, and $\nu_\pi^0(s'|s)$ is a Dirac delta distribution:

$$\nu_\pi^0(s'|s) = \begin{cases} \infty, \text{ if } s' = s, \\ 0, \text{ otherwise.} \end{cases} \tag{30}$$

Note that $\nu_\pi^0(s'|s)$ is independent of the policy π, and thus $\nu_\pi^0(s'|s) = \nu_{\pi'}^0(s'|s)$.

5. Discounted state transition PDF given the policy π:

$$\begin{aligned} \mu_\pi(s'|s) =& (1-\gamma)\Big[\nu_\pi^0(s'|s) + \gamma\nu_\pi^1(s'|s) + \gamma^2\nu_\pi^2(s'|s) + \cdots\Big] \\ =& (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\nu_\pi^t(s'|s). \end{aligned} \tag{31}$$

Then, the discounted visitation PDF can be written as

$$d^\pi(s') = \int_{\mathcal{S}} \rho_0(s)\mu_\pi(s'|s)ds. \tag{32}$$

6. Surrogate model:

$$L_\pi(\pi') = J(\pi) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi, a\sim\pi'}[A_\pi(s,a)]$$

7. Function spaces: For an open set $U \subset \mathbb{R}^d$, we define

$$C^k(U) := \{\text{Functions } u : U \to \mathbb{R} \text{ that are } k\text{-times continuously differentiable function on } U\}. \tag{33}$$

We start by introducing the definition of α-*coupled policies* from the Definition 1 in (Schulman et al., 2015) with some changes.

Definition A.2 (α-coupled policies)**.** A coupling of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν (Levin et al., 2006).

The policies $\pi'(a'|s)$ and $\pi(a|s)$ are called α-coupled if they define a coupling of (π', π) such that

$$P(a' \neq a|s) \leq \alpha(s). \tag{34}$$

Numerically, α-coupling means that the actions a' and a given state s match with probability of at least $1 - \alpha(s)$ when their samples are drawn using the same seed.

The technique of coupling is useful because it relates two policies to their total variation distance. According to the lemma 4.7 in (Levin et al., 2006), for policies π' and π, there exists a coupling that satisfies

$$D_{\mathrm{TV}}[\pi'||\pi](s) = \inf\{P(a' \neq a|s), a' \text{ and } a \text{ is a coupling of } \pi' \text{ and } \pi\}. \tag{35}$$

where $D_{\mathrm{TV}}[\pi'||\pi](s)$ represents the total variation distance between policies π' and π given the state s. This means that $D_{\mathrm{TV}}[\pi'||\pi](s)$ is the infimum of the probability $P(a' \neq a|s)$, and therefore we can select $\alpha(s)$ to be $D_{\mathrm{TV}}[\pi'||\pi](s)$.

Note that our definition of $\alpha(s)$, depending on the state s, is different from the definition in (Schulman et al., 2015), which is the maximum over the state space, i.e. $\max_{s \in \mathcal{S}} \alpha(s)$.

Next, we present a lemma from (Kakade & Langford, 2002) and (Schulman et al., 2015) that shows that the performance difference between two arbitrary policies can be expressed as an expected advantage of one policy over a trajectory resulted from the other.

Lemma A.3. *Given two policies π', π, we have*

$$J(\pi') = J(\pi) + \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi'}, a \sim \pi'} \left[A_\pi(s, a) \right]. \tag{36}$$

Proof. Note that $A_\pi(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s, a) + \gamma V_\pi(s') - V_\pi(s) \right]$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{s \sim d^{\pi'}, a \sim \pi'} \left[A_\pi(s, a) \right] \\
&= \mathbb{E}_{s \sim d^{\pi'}, a \sim \pi', s' \sim P} \left[r(s, a) + \gamma V_\pi(s') - V_\pi(s) \right] \\
&= (1-\gamma) \mathbb{E}_{s_t \sim \rho_t^{\pi'}, a_t \sim \pi', s_{t+1} \sim P} \left[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t) \Big) \right] \\
&= (1-\gamma) \mathbb{E}_{s_t \sim \rho_t^{\pi'}, a_t \sim \pi'} \left[-V_\pi(s_0) + \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \\
&= (1-\gamma) \left(-\mathbb{E}_{s_0 \sim \rho_0} \left[V_\pi(s_0) \right] + \mathbb{E}_{s_t \sim \rho_t^{\pi'}, a_t \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \right) \\
&= (1-\gamma) \left[-J(\pi) + J(\pi') \right]
\end{aligned} \tag{37}$$

Rearranging it, the result follows. □

Lemma A.4. *Given two stochastic policies π', π and their discounted state transition PDFs, $\mu_{\pi'}(s'|s), \mu_\pi(s'|s)$, the following inequality holds:*

$$\int_{\mathcal{S}} \left| \mu_{\pi'}(s'|s) - \mu_\pi(s'|s) \right| ds' \leq \frac{2\gamma^2}{1-\gamma} \int_{\mathcal{S}} \mu_\pi(s'|s) D_{\mathrm{TV}}[\pi'||\pi](s') ds'. \tag{38}$$

Proof. First note that

$$\begin{aligned}
&\gamma \int_{\mathcal{S}} \nu_\pi(s'|\bar{s})\mu_\pi(\bar{s}|s)d\bar{s} \\
&= \gamma \int_{\mathcal{S}} \nu_\pi(s'|\bar{s})(1-\gamma)\Big[\nu_\pi^0(\bar{s}|s) + \gamma\nu_\pi^1(\bar{s}|s) + \gamma^2\nu_\pi^2(\bar{s}|s) + \cdots\Big]d\bar{s} \\
&= (1-\gamma)\Big[\gamma\nu_\pi^1(s'|s) + \gamma^2\nu_\pi^2(s'|s) + \gamma^3\nu_\pi^3(s'|s) + \cdots\Big] \\
&= \mu_\pi(s'|s) - (1-\gamma)\nu_\pi^0(s'|s).
\end{aligned} \tag{39}$$

Then, we have

$$\begin{aligned}
&\gamma \iint_{\mathcal{S}\times\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\left[\nu_{\pi'}(\tilde{s}|\bar{s}) - \nu_\pi(\tilde{s}|\bar{s})\right]\mu_\pi(\bar{s}|s)d\tilde{s}d\bar{s} \\
&= \int_{\mathcal{S}} \left(\gamma \int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\nu_{\pi'}(\tilde{s}|\bar{s})d\tilde{s}\right)\mu_\pi(\bar{s}|s)d\bar{s} - \int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\left(\gamma\int_{\mathcal{S}} \nu_\pi(\tilde{s}|\bar{s})\mu_\pi(\bar{s}|s)d\bar{s}\right)d\tilde{s} \\
&= \int_{\mathcal{S}} \left[\mu_{\pi'}(s'|\bar{s}) - (1-\gamma)\nu_{\pi'}^0(s'|\bar{s})\right]\mu_\pi(\bar{s}|s)d\bar{s} - \int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\left[\mu_\pi(\tilde{s}|s) - (1-\gamma)\nu_\pi^0(\tilde{s}|s)\right]d\tilde{s} \\
&= \cancel{\int_{\mathcal{S}} \mu_{\pi'}(s'|\bar{s})\mu_\pi(\bar{s}|s)d\bar{s}} - (1-\gamma)\int_{\mathcal{S}} \nu_{\pi'}^0(s'|\bar{s})\mu_\pi(\bar{s}|s)d\bar{s} - \cancel{\int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\mu_\pi(\tilde{s}|s)d\tilde{s}} + (1-\gamma)\int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\nu_\pi^0(\tilde{s}|s)d\tilde{s} \\
&= (1-\gamma)\left[\int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\nu_\pi^0(\tilde{s}|s)d\tilde{s} - \int_{\mathcal{S}} \nu_{\pi'}^0(s'|\bar{s})\mu_\pi(\bar{s}|s)d\bar{s}\right] \quad (\text{Note that } \nu_\pi^0(\tilde{s}|s) = \nu_{\pi'}^0(\tilde{s}|s)) \\
&= \frac{1-\gamma}{\gamma}\left[\Big(\mu_{\pi'}(s'|s) - (1-\gamma)\nu_{\pi'}^0(s'|s)\Big) - \Big(\mu_\pi(s'|s) - (1-\gamma)\nu_\pi^0(s'|s)\Big)\right] \\
&= \frac{1-\gamma}{\gamma}\left[\mu_{\pi'}(s'|s) - \mu_\pi(s'|s)\right].
\end{aligned} \tag{40}$$

Rearranging the equation, we have

$$\mu_{\pi'}(s'|s) - \mu_\pi(s'|s) = \frac{\gamma^2}{1-\gamma}\iint_{\mathcal{S}\times\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})\left[\nu_{\pi'}(\tilde{s}|\bar{s}) - \nu_\pi(\tilde{s}|\bar{s})\right]\mu_\pi(\bar{s}|s)d\tilde{s}d\bar{s}. \tag{41}$$

Recalling the definition of one-step state transition density in Equation (28), we have

$$\nu_{\pi'}(\tilde{s}|\bar{s}) - \nu_\pi(\tilde{s}|\bar{s}) = \int_{\mathcal{A}} p(\tilde{s}|\bar{s},\bar{a})\left[\pi'(\bar{a}|\bar{s}) - \pi(\bar{a}|\bar{s})\right]d\bar{a}. \tag{42}$$

Then, we have

$$\begin{aligned}
&\int_{\mathcal{S}} |\mu_{\pi'}(s'|s) - \mu_\pi(s'|s)|ds' \\
&\le \frac{\gamma^2}{1-\gamma}\iiiint_{\mathcal{S}\times\mathcal{S}\times\mathcal{S}\times\mathcal{A}} \mu_{\pi'}(s'|\tilde{s})p(\tilde{s}|\bar{s},\bar{a})|\pi'(\bar{a}|\bar{s}) - \pi(\bar{a}|\bar{s})|\mu_\pi(\bar{s}|s)ds'd\tilde{s}d\bar{s}d\bar{a} \\
&= \frac{\gamma^2}{1-\gamma}\int_{\mathcal{S}} \mu_{\pi'}(s'|\tilde{s})ds' \iiint_{\mathcal{S}\times\mathcal{S}\times\mathcal{A}} p(\tilde{s}|\bar{s},\bar{a})|\pi'(\bar{a}|\bar{s}) - \pi(\bar{a}|\bar{s})|\mu_\pi(\bar{s}|s)d\tilde{s}d\bar{s}d\bar{a} \\
&= \frac{\gamma^2}{1-\gamma}\int_{\mathcal{S}} p(\tilde{s}|\bar{s},\bar{a})d\tilde{s} \iint_{\mathcal{S}\times\mathcal{A}} |\pi'(\bar{a}|\bar{s}) - \pi(\bar{a}|\bar{s})|\mu_\pi(\bar{s}|s)d\bar{s}d\bar{a} \\
&= \frac{\gamma^2}{1-\gamma}\int_{\mathcal{S}} \mu_\pi(\bar{s}|s)\int_{\mathcal{A}} |\pi'(\bar{a}|\bar{s}) - \pi(\bar{a}|\bar{s})|d\bar{a}d\bar{s} \\
&= \frac{2\gamma^2}{1-\gamma}\int_{\mathcal{S}} \mu_\pi(\bar{s}|s)D_{\mathrm{TV}}[\pi'||\pi](\bar{s})d\bar{s}.
\end{aligned} \tag{43}$$

Replacing all $\bar{s}$ with s', the result follows. □

Lemma A.5. *Let $\overline{\alpha}_t$ and γ be any real numbers within $\overline{\alpha}_t \in [0,1], \forall t \in \mathbb{N}$ and $\gamma \in [0.5, 1)$, the following inequality holds:*

$$(1-\gamma)^2 \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t \overline{\alpha}_{0t} \leq \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2. \tag{44}$$

where $\overline{\alpha}_{0t} = 1 - \prod_{i=0}^{t}(1-\overline{\alpha}_i)$.

Proof. First note that $\overline{\alpha}_{0t}$ can be expressed as

$$\overline{\alpha}_{0t} = \overline{\alpha}_t + (1-\overline{\alpha}_t)\overline{\alpha}_{t-1} + (1-\overline{\alpha}_t)(1-\overline{\alpha}_{t-1})\overline{\alpha}_{t-2} + \cdots + \prod_{i=1}^{t}(1-\overline{\alpha}_i)\overline{\alpha}_0, \tag{45}$$

or in a recursive form:

$$\overline{\alpha}_{0t} = \overline{\alpha}_t + (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1}, \tag{46}$$

where $\overline{\alpha}_{00} = \overline{\alpha}_0$. Then, we have

$$\begin{aligned}
&(1-\gamma)^2 \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t \overline{\alpha}_{0t} \\
=& (1-\gamma)^2 \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 + (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \\
=& \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 - (2\gamma-\gamma^2) \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 + (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \\
=& \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 - \left[(2\gamma-\gamma^2) \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 - (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \right]
\end{aligned} \tag{47}$$

For the inequality (44) to hold, we only need to prove that the subtrahend on the rightest-hand side of (47) is greater than 0. Note that

$$\begin{aligned}
&(2\gamma-\gamma^2) \sum_{t=0}^{\infty} \gamma^t \overline{\alpha}_t^2 - (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \\
=& \sum_{t=0}^{\infty} \gamma^t \Big[\gamma + \gamma(1-\gamma) \Big] \overline{\alpha}_t^2 - (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \\
=& \sum_{t=0}^{\infty} \gamma^t \left[(1-\gamma) \sum_{i=1}^{\infty} \gamma^i + (1-\gamma)^2 \sum_{i=1}^{\infty} \gamma^i \right] \overline{\alpha}_t^2 - (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \\
=& (1-\gamma)^2 \sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{\infty} \gamma^i \left[\frac{1}{1-\gamma} + 1 \right] \overline{\alpha}_t^2 - (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \overline{\alpha}_t (1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \quad \left(\frac{1}{1-\gamma} = 1 + \gamma + \gamma^2 + \cdots \right) \\
=& \gamma(1-\gamma)\overline{\alpha}_0^2 + (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \left[\sum_{n=0}^{t-1} (t-n)\overline{\alpha}_n^2 + \frac{\gamma}{1-\gamma}\overline{\alpha}_t^2 - \overline{\alpha}_t(1-\overline{\alpha}_t)\overline{\alpha}_{0t-1} \right] \\
\geq& \gamma(1-\gamma)^2\overline{\alpha}_0^2 + (1-\gamma)^2 \sum_{t=1}^{\infty} \gamma^t \left[\sum_{n=0}^{t-1} (t-n)\overline{\alpha}_n^2 + \overline{\alpha}_t^2 - \overline{\alpha}_t\overline{\alpha}_{0t-1} \right]. \quad (\text{since } \gamma > 1-\gamma \text{ when } \gamma \geq 0.5)
\end{aligned} \tag{48}$$

In the expanded form, the rightest-hand side of (48) can be expressed as

$$\begin{aligned}
&\gamma(1-\gamma)^2\overline{\alpha}_0^2 + (1-\gamma)^2\sum_{t=1}^{\infty}\gamma^t\left[\sum_{n=0}^{t-1}(t-n)\overline{\alpha}_n^2 + \overline{\alpha}_t^2 - \overline{\alpha}_t\overline{\alpha}_{0t-1}\right] \\
=\ &\gamma(1-\gamma)^2\Big[\overline{\alpha}_0^2 + \overline{\alpha}_0^2 + \overline{\alpha}_1^2 - \overline{\alpha}_1\overline{\alpha}_0\Big] + \\
&\gamma^2(1-\gamma)^2\Big[2\overline{\alpha}_0^2 + \overline{\alpha}_1^2 + \overline{\alpha}_2^2 - \overline{\alpha}_2\overline{\alpha}_{01}\Big] + \\
&\gamma^3(1-\gamma)^2\Big[3\overline{\alpha}_0^2 + 2\overline{\alpha}_1^2 + \overline{\alpha}_2^2 + \overline{\alpha}_3^2 - \overline{\alpha}_3\overline{\alpha}_{02}\Big] + \\
&\gamma^4(1-\gamma)^2\Big[4\overline{\alpha}_0^2 + 3\overline{\alpha}_1^2 + 2\overline{\alpha}_2^2 + \overline{\alpha}_3^2 + \overline{\alpha}_4^2 - \overline{\alpha}_4\overline{\alpha}_{03}\Big] + \\
&\cdots
\end{aligned} \tag{49}$$

Since

$$\gamma^t(1-\gamma)^2 = \gamma^t(1-\gamma)^2(1-\gamma+\gamma) = \gamma^t(1-\gamma)^3 + \gamma^{t+1}(1-\gamma)^2, \tag{50}$$

Equation (49) can be rewritten as

$$\begin{aligned}
&\gamma(1-\gamma)^3\Big[2\overline{\alpha}_0^2 + \overline{\alpha}_1^2 - \overline{\alpha}_1\overline{\alpha}_0\Big] + \\
&\gamma^2(1-\gamma)^3\Big[4\overline{\alpha}_0^2 + 2\overline{\alpha}_1^2 + \overline{\alpha}_2^2 - \overline{\alpha}_1\overline{\alpha}_0 - \overline{\alpha}_2\overline{\alpha}_{01}\Big] + \\
&\gamma^3(1-\gamma)^3\Big[7\overline{\alpha}_0^2 + 4\overline{\alpha}_1^2 + 2\overline{\alpha}_2^2 + \overline{\alpha}_3^2 - \overline{\alpha}_1\overline{\alpha}_0 - \overline{\alpha}_2\overline{\alpha}_{01} - \overline{\alpha}_3\overline{\alpha}_{02}\Big] + \\
&\gamma^4(1-\gamma)^3\Big[11\overline{\alpha}_0^2 + 7\overline{\alpha}_1^2 + 4\overline{\alpha}_2^2 + 2\overline{\alpha}_3^2 + \overline{\alpha}_4^2 - \overline{\alpha}_1\overline{\alpha}_0 - \overline{\alpha}_2\overline{\alpha}_{01} - \overline{\alpha}_3\overline{\alpha}_{02} - \overline{\alpha}_4\overline{\alpha}_{03}\Big] + \\
&\cdots \\
=\ &(1-\gamma)^3\sum_{t=1}^{\infty}\gamma^t\left[a_t\overline{\alpha}_0^2 + \sum_{i=1}^{t}\Big(a_{t-i}\overline{\alpha}_i^2 - \overline{\alpha}_i\overline{\alpha}_{0i-1}\Big)\right] \\
=\ &(1-\gamma)^3\sum_{t=1}^{\infty}\gamma^t H_t
\end{aligned} \tag{51}$$

where

$$a_t = 1 + \sum_{j=0}^{t} j, \tag{52}$$

and

$$H_t = a_t\overline{\alpha}_0^2 + \sum_{i=1}^{t}\Big(a_{t-i}\overline{\alpha}_i^2 - \overline{\alpha}_i\overline{\alpha}_{0i-1}\Big), \tag{53}$$

Next, we prove $H_t \geq 0$ for all $t \in \mathbb{N}^+$ by using convex optimization. Decompose H_t into:

$$\begin{aligned}
H_t &= a_t\overline{\alpha}_0^2 + \sum_{i=1}^{t-1}\Big(a_{t-i}\overline{\alpha}_i^2 - \overline{\alpha}_i\overline{\alpha}_{0i-1}\Big) + a_0\overline{\alpha}_t{}^2 - \overline{\alpha}_t\overline{\alpha}_{0t-1} \\
&= h_{t-1} + a_0\overline{\alpha}_t{}^2 - \overline{\alpha}_t\overline{\alpha}_{0t-1}
\end{aligned} \tag{54}$$

Taking the partial derivative of H_t with respect to $\overline{\alpha}_t$ and setting it to be zero, H_t attains its minimum value, i.e.,

$$H_t \geq h_{t-1} - \frac{1}{4a_0}\overline{\alpha}_{0t-1}^2. \tag{55}$$

Denoting $b_1 = 1/4a_0$ and decomposing h_{t-1}, we get

$$\begin{aligned}
H_t &\geq h_{t-2} + a_1\overline{\alpha}_{t-1}^2 - \overline{\alpha}_{t-1}\overline{\alpha}_{0t-2} - b_1\overline{\alpha}_{0t-1}^2 \\
&\geq h_{t-2} + a_1\overline{\alpha}_{t-1}^2 - \overline{\alpha}_{t-1}\overline{\alpha}_{0t-2} - b_1[\overline{\alpha}_{t-1} + \overline{\alpha}_{0t-2}]^2 \\
&= h_{t-2} + (a_1 - b_1)\overline{\alpha}_{t-1}^2 - (2b_1+1)\overline{\alpha}_{t-1}\overline{\alpha}_{0t-2} - b_1\overline{\alpha}_{0t-2}^2
\end{aligned} \tag{56}$$

Again, taking the partial derivative with respect to $\overline{\alpha}_{t-1}$ and setting it to be zero, we get

$$H_t \geq h_{t-2} - \frac{b_1 + a_1 b_1 + 1/4}{a_1 - b_1}\overline{\alpha}_{0t-2}^2. \tag{57}$$

Recursively, as long as $b_i \geq 0$ and $a_i - b_i \geq 0$ hold for all $i \leq t \in \mathbb{N}^+$, we can repeatedly apply the previous procedure and get

$$H_t \geq h_{t-1} - b_1\overline{\alpha}_{0t-1}^2 \geq h_{t-2} - b_2\overline{\alpha}_{0t-2}^2 \geq \cdots \geq a_t\overline{\alpha}_0^2 - b_t\overline{\alpha}_0^2 \geq 0, \tag{58}$$

where

$$b_{i+1} = \frac{b_i + a_i b_i + 1/4}{a_i - b_i},\ i = 0, \ldots, t-1, \tag{59}$$

$b_0 = 0$ and a_i is defined in Equation (52). Next, we prove $b_i \geq 0$ and $a_i - b_i \geq 0$ for all $i \leq t \in \mathbb{N}^+$.

First, it is easy to manually verify that $b_i \geq 0$ and $a_i - b_i \geq 0$ when $i < 15$. In addition, for $i = 15$, we can verify that the following inequalities holds

$$a_i - b_i \geq 4(b_i + \frac{1}{2})^2, \quad b_i \leq \frac{1}{4}i, \tag{60}$$

since $a_{15} = 121$ and $b_{15} \approx 3.6945$.

Next, we prove the inequalities in (60) holds for $i > 15$. Note that

$$\begin{aligned} b_{i+1} - b_i &= \frac{b_i + a_i b_i + 1/4}{a_i - b_i} - b_i = \frac{(b_i + \frac{1}{2})^2}{a_i - b_i}, \\ a_{i+1} - a_i &= i + 1. \end{aligned} \tag{61}$$

Therefore, we have

$$b_{i+1} = b_i + \frac{(b_i + \frac{1}{2})^2}{a_i - b_i} \leq b_i + \frac{1}{4} \leq \frac{1}{4}(i+1), \tag{62}$$

and

$$\begin{aligned} a_{i+1} - b_{i+1} &= i + 1 + a_i - b_i - \frac{(b_i + \frac{1}{2})^2}{a_i - b_i} \\ &\geq i + 1 + 4(b_i + \frac{1}{2})^2 - \frac{1}{4} \\ &\geq i + 1 + 4\Big(b_{i+1} + \frac{1}{2} - \frac{1}{4}\Big)^2 - \frac{1}{4} \\ &= 4(b_{i+1} + \frac{1}{2})^2 - 2b_{i+1} + i \\ &\geq 4(b_{i+1} + \frac{1}{2})^2 - \frac{1}{2}(i+1) + i \quad (i > 15) \\ &\geq 4(b_{i+1} + \frac{1}{2})^2. \end{aligned} \tag{63}$$

Based on Equations (60), (62) and (63), we can prove that $b_i \geq 0$ and $a_i - b_i \geq 0$ hold for $i \geq 15$ using mathematical induction. Combining the fact that they also hold for $i < 15$, we have $b_i \geq 0$ and $a_i - b_i \geq 0$ for all $i \leq t \in \mathbb{N}^+$. As a result, the inequality (58) holds, i.e., $H_t \geq 0$, which concludes the proof. □

Theorem A.6. *For any stochastic policies π', π and discount factor $\gamma \in [0.5, 1)$, the following bound holds:*

$$\begin{aligned} &|J(\pi') - L_\pi(\pi')| \leq \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi' \| \pi](s)\right], \\ &\textit{where}\ \ C_\pi = \frac{\gamma^2 \epsilon}{(1-\gamma)^3},\ \epsilon = \max_{s,a} |A_\pi(s,a)|. \end{aligned} \tag{64}$$

Proof. Define $\overline{A}(s)$ to be the expected advantage of π' over π at state s:

$$\overline{A}(s) = \mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A_\pi(s,a)\right] \tag{65}$$

Then, Lemma A.3 can be rewritten as follows:

$$J(\pi') = J(\pi) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}\left[\overline{A}(s)\right] = J(\pi) + \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right]. \tag{66}$$

Note that the surrogate model can be written as

$$L_\pi(\pi') = J(\pi) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}}\left[\overline{A}(s)\right] = J(\pi) + \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]. \tag{67}$$

Then, the difference between the surrogate $L_\pi(\pi')$ and the true objective $J(\pi')$ can be written as

$$\begin{aligned} J(\pi') - L_\pi(\pi') &= \sum_{t=0}^{\infty}\gamma^t\Big[\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right] - \mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]\Big] \\ &= \sum_{t=1}^{\infty}\gamma^t\Big[\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right] - \mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]\Big]. \quad (\text{since } p_0^{\pi'} = p_0^{\pi} = \rho_0) \end{aligned} \tag{68}$$

Next, we split the proof into three parts. (1) By using the coupling technique, we decompose the difference terms in (68), i.e. $\mathbb{E}_{s_t\sim p_t^{\pi'}}\left[\overline{A}(s_t)\right] - \mathbb{E}_{s_t\sim p_t^{\pi}}\left[\overline{A}(s_t)\right]$, to derive an equivalent expression. (2) Based on the result from the first part, we use Lemma A.4 to derive an upper bound of $|J(\pi') - L_\pi(\pi')|$, which depends on a bunch of state-dependent total variation distances of π', π. (3) We relate the bound derived from the second part to the expected KL-Divergence between π', π.

i) The first part of the proof is given as follows.

We will use techniques from the proof of Lemma 3. in (Schulman et al., 2015) to measure the coincidence of two trajectories resulted from π', π before an arbitrary timestep t. Let n_t denote the number of times that $a_i' \neq a_i|s_i$ at state s_i for $i < t$. For instance, $n_t = 0$ means the trajectories τ, τ' completely match before timestep t, i.e., $a_i' = a_i|s_i$ for all $i < t$.

The expected advantage at state s_t on the trajectory $\tau' \sim \pi'$ decomposes as follows:

$$\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right] = P(n_t = 0)\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_t=0}\left[\overline{A}(s_t)\right] + P(n_t > 0)\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_t>0}\left[\overline{A}(s_t)\right]. \tag{69}$$

The expected advantage on the trajectory $\tau \sim \pi$ decomposes similarly:

$$\mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right] = P(n_t = 0)\mathbb{E}_{s_t\sim\rho_t^{\pi}|n_t=0}\left[\overline{A}(s_t)\right] + P(n_t > 0)\mathbb{E}_{s_t\sim\rho_t^{\pi}|n_t>0}\left[\overline{A}(s_t)\right]. \tag{70}$$

Subtracting Equation (70) from (69), we get

$$\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right] - \mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right] = P(n_t > 0)\Big(\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_t>0}\left[\overline{A}(s_t)\right] - \mathbb{E}_{s_t\sim\rho_t^{\pi}|n_t>0}\left[\overline{A}(s_t)\right]\Big), \tag{71}$$

because $\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_t=0}\left[\overline{A}(s_t)\right] = \mathbb{E}_{s_t\sim\rho_t^{\pi}|n_t=0}\left[\overline{A}(s_t)\right]$ when $n_t = 0$.

Note that

$$n_t > 0 \Rightarrow \begin{cases} n_{t-1} = 0 \text{ and } a_{t-1}' \neq a_{t-1}|s_{t-1} \text{ for every } s_{t-1}, \text{ or} \\ n_{t-1} > 0, \end{cases} \tag{72}$$

so we have

$$P(n_t > 0) = P(n_{t-1} = 0)\cdot\mathbb{E}_{s_{t-1}\sim\rho_{t-1}^{\pi}}\left[P(a_{t-1}' \neq a_{t-1}|s_{t-1})\right] + P(n_{t-1} > 0). \tag{73}$$

In a recursive form, it can be expressed as:

$$P(n_t > 0) = \sum_{i=0}^{t-1}P(n_i = 0)\mathbb{E}_{s_i\sim\rho_i^{\pi}}\left[P(a_i' \neq a_i|s_i)\right] \tag{74}$$

Substituting (74) into (71), we get

$$\begin{aligned}
&\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right]-\mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]\\
&=\sum_{i=0}^{t-1}P(n_i=0)\mathbb{E}_{s_i\sim\rho_i^{\pi}}\left[P(a_i'\neq a_i|s_i)\right]\Big(\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_i=0}\left[\overline{A}(s_t)\right]-\mathbb{E}_{s_t\sim\rho_t^{\pi}|n_i=0}\left[\overline{A}(s_t)\right]\Big).
\end{aligned}\tag{75}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{s_t\sim\rho_t^{\pi'}|n_i=0}\left[\overline{A}(s_t)\right]-\mathbb{E}_{s_t\sim\rho_t^{\pi}|n_i=0}\left[\overline{A}(s_t)\right]\\
&=\int_{\mathcal{S}}\left[\rho_t^{\pi'}(s_t)-\rho_t^{\pi}(s_t)\right]_{n_i=0}\overline{A}(s_t)ds_t\\
&=\int_{\mathcal{S}}\Big(\int_{\mathcal{S}}\left[\rho_i^{\pi'}(s_i)\nu_{\pi'}^{t-i}(s_t|s_i)-\rho_i^{\pi}(s_i)\nu_{\pi}^{t-i}(s_t|s_i)\right]_{n_i=0}ds_i\Big)\overline{A}(s_t)ds_t\\
&=\iint_{\mathcal{S}\times\mathcal{S}}\rho_i^{\pi}(s_i)\left[\nu_{\pi'}^{t-i}(s_t|s_i)-\nu_{\pi}^{t-i}(s_t|s_i)\right]\overline{A}(s_t)ds_ids_t.\ \text{(since }\rho_i^{\pi'}=\rho_i^{\pi}\text{ when }n_i=0)\\
&=\iint_{\mathcal{S}\times\mathcal{S}}\rho_i^{\pi}(s_i)\delta^{t-i}(s_t|s_i)\overline{A}(s_t)ds_ids_t\quad(\text{denote }\delta^{t-i}(s_t|s_i)=\nu_{\pi'}^{t-i}(s_t|s_i)-\nu_{\pi}^{t-i}(s_t|s_i))\\
&=\iint_{\mathcal{S}\times\mathcal{S}}\rho_i^{\pi}(s)\delta^{t-i}(s'|s)\overline{A}(s')dsds'
\end{aligned}\tag{76}$$

Substituting (76) into (75), we get

$$\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right]-\mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]=\sum_{i=0}^{t-1}P(n_i=0)\mathbb{E}_{s_i\sim\rho_i^{\pi}}\left[P(a_i'\neq a_i|s_i)\right]\iint_{\mathcal{S}\times\mathcal{S}}\rho_i^{\pi}(s)\delta^{t-i}(s'|s)\overline{A}(s')dsds'.\tag{77}$$

For notational simplicity, we denote

$$P_{n_i=0}\overline{P}_{a_i'\neq a_i}:=P(n_i=0)\mathbb{E}_{s_i\sim\rho_i^{\pi}}\left[P(a_i'\neq a_i|s_i)\right].\tag{78}$$

Then, Equation (77) can be expressed as

$$\mathbb{E}_{s_t\sim\rho_t^{\pi'}}\left[\overline{A}(s_t)\right]-\mathbb{E}_{s_t\sim\rho_t^{\pi}}\left[\overline{A}(s_t)\right]=\sum_{i=0}^{t-1}P_{n_i=0}\overline{P}_{a_i'\neq a_i}\iint_{\mathcal{S}\times\mathcal{S}}\rho_i^{\pi}(s)\delta^{t-i}(s'|s)\overline{A}(s')dsds'.\tag{79}$$

ii) The second part of the proof is given as follows.

Substituting (79) into (68), we get

$$\begin{aligned}
&J(\pi') - L_\pi(\pi') \\
&= \sum_{t=1}^{\infty} \gamma^t \sum_{i=0}^{t-1} P_{n_i=0} \overline{P}_{a_i' \neq a_i} \iint_{\mathcal{S}\times\mathcal{S}} \rho_i^\pi(s) \delta^{t-i}(s'|s) \overline{A}(s') ds ds' \\
&= \left(P_{n_0=0} \overline{P}_{a_0' \neq a_0} \iint_{\mathcal{S}\times\mathcal{S}} \gamma \rho_0^\pi(s) \delta^1(s'|s) \overline{A}(s') ds ds' \right) + \\
&\quad \left(P_{n_0=0} \overline{P}_{a_0' \neq a_0} \iint_{\mathcal{S}\times\mathcal{S}} \gamma^2 \rho_0^\pi(s) \delta^2(s'|s) \overline{A}(s') ds ds' + P_{n_1=0} \overline{P}_{a_1' \neq a_1} \iint_{\mathcal{S}\times\mathcal{S}} \gamma^2 \rho_1^\pi(s) \delta^1(s'|s) \overline{A}(s') ds ds' \right) + \\
&\quad \vdots \\
&= P_{n_0=0} \overline{P}_{a_0' \neq a_0} \iint_{\mathcal{S}\times\mathcal{S}} \rho_0^\pi(s) \big[\gamma \delta^1(s'|s) + \gamma^2 \delta^2(s'|s) + \cdots \big] \overline{A}(s') ds ds' + \\
&\quad P_{n_1=0} \overline{P}_{a_1' \neq a_1} \iint_{\mathcal{S}\times\mathcal{S}} \gamma \rho_1^\pi(s) \big[\gamma \delta^1(s'|s) + \gamma^2 \delta^2(s'|s) + \cdots \big] \overline{A}(s') ds ds' + \\
&\quad \cdots \\
&= \frac{1}{1-\gamma} \sum_{t=0}^{\infty} P_{n_t=0} \overline{P}_{a_t' \neq a_t} \iint_{\mathcal{S}\times\mathcal{S}} \gamma^t \rho_t^\pi(s) \big[\mu_{\pi'}(s'|s) - \mu_\pi(s'|s) \big] \overline{A}(s') ds ds'. \quad \text{(See the definition of } \mu_\pi \text{ in (31).)}
\end{aligned} \tag{80}$$

Taking absolute values on both sides and applying Hölder's inequality, we get

$$\begin{aligned}
|J(\pi') - L_\pi(\pi')| &\leq \frac{1}{1-\gamma} \sum_{t=0}^{\infty} P_{n_t=0} \overline{P}_{a_t' \neq a_t} \iint_{\mathcal{S}\times\mathcal{S}} \gamma^t \rho_t^\pi(s) \big| \big[\mu_{\pi'}(s'|s) - \mu_\pi(s'|s) \big] \overline{A}(s') \big| ds' ds \\
&\leq \frac{1}{1-\gamma} \sum_{t=0}^{\infty} P_{n_t=0} \overline{P}_{a_t' \neq a_t} \int_{\mathcal{S}} \gamma^t \rho_t^\pi(s) \int_{\mathcal{S}} \big| \mu_{\pi'}(s'|s) - \mu_\pi(s'|s) \big| ds' ds \cdot \max_{s',a'} |A_\pi(s',a')|
\end{aligned} \tag{81}$$

Applying Lemma A.4, we have

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2 \epsilon}{(1-\gamma)^2} \sum_{t=0}^{\infty} P_{n_t=0} \overline{P}_{a_t' \neq a_t} \iint_{\mathcal{S}\times\mathcal{S}} \gamma^t \rho_t^\pi(s) \mu_\pi(s'|s) D_{\mathrm{TV}}[\pi'||\pi](s') ds ds'. \tag{82}$$

Note that the integral part in the above inequality can be expressed as

$$\begin{aligned}
&\iint_{\mathcal{S}\times\mathcal{S}} \gamma^t \rho_t^\pi(s) \mu_\pi(s'|s) D_{\mathrm{TV}}[\pi'||\pi](s') ds ds' \\
&= \int_{\mathcal{S}} \Big(\int_{\mathcal{S}} \gamma^t \rho_t^\pi(s) \mu_\pi(s'|s) ds \Big) D_{\mathrm{TV}}[\pi'||\pi](s') ds' \quad \text{(See the definition of } \mu_\pi \text{ in (31).)} \\
&= \int_{\mathcal{S}} \Big(d^\pi(s') - (1-\gamma) \sum_{i=0}^{t-1} \gamma^i \rho_i^\pi(s') \Big) D_{\mathrm{TV}}[\pi'||\pi](s') ds' \quad \text{(for all } t > 0) \\
&= \mathbb{E}_{s' \sim d^\pi}[D_{\mathrm{TV}}[\pi'||\pi](s')] - (1-\gamma) \sum_{i=0}^{t-1} \gamma^i \mathbb{E}_{s' \sim \rho_i^\pi}[D_{\mathrm{TV}}[\pi'||\pi](s')] \quad \text{(for all } t > 0) \\
&= \mathbb{E}_{s \sim d^\pi}[D_{\mathrm{TV}}[\pi'||\pi](s)] - (1-\gamma) \sum_{i=0}^{t-1} \gamma^i \mathbb{E}_{s \sim \rho_i^\pi}[D_{\mathrm{TV}}[\pi'||\pi](s)] \quad \text{(for all } t > 0)
\end{aligned} \tag{83}$$

In the following, we will replace all total variations $D_{\mathrm{TV}}[\pi'||\pi](s)$ with $\alpha(s)$ (see Definition A.2) and use the following notations for simplicity:

$$\overline{\alpha} := \mathbb{E}_{s \sim d^\pi}[\alpha(s)], \ \overline{\alpha}_i := \mathbb{E}_{s \sim \rho_i^\pi}[\alpha(s)]. \tag{84}$$

Plugging (83) into (82), we have

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2\epsilon}{(1-\gamma)^2}\left[\sum_{t=0}^{\infty} P_{n_t=0}\overline{P}_{a'_t\neq a_t}\cdot\overline{\alpha} - (1-\gamma)\sum_{t=1}^{\infty}\sum_{i=0}^{t-1} P_{n_t=0}\overline{P}_{a'_t\neq a_t}\cdot\gamma^i\overline{\alpha}_i\right] \tag{85}$$

Using Equations (74) and (78), we have

$$\sum_{t=1}^{k-1} P_{n_t=0}\overline{P}_{a'_t\neq a_t} = P[n_k > 0] \to 1 \ \text{ when } k\to\infty. \tag{86}$$

Therefore, the first term in the parentheses on the rightest-hand side of (85) is just

$$\sum_{t=1}^{\infty} P_{n_t=0}\overline{P}_{a'_t\neq a_t}\cdot\overline{\alpha} = \overline{\alpha}. \tag{87}$$

The second term in the parentheses on the rightest-hand side of (85) can be expressed as

$$\begin{aligned}
&(1-\gamma)\sum_{t=1}^{\infty}\sum_{i=0}^{t-1} P_{n_t=0}\overline{P}_{a'_t\neq a_t}\cdot\gamma^i\overline{\alpha}_i \\
&= (1-\gamma)\Big(P_{n_1=0}\overline{P}_{a'_1\neq a_1}\cdot\overline{\alpha}_0\Big) + \\
&\quad (1-\gamma)\Big(P_{n_2=0}\overline{P}_{a'_2\neq a_2}\cdot\overline{\alpha}_0 + P_{n_2=0}\overline{P}_{a'_2\neq a_2}\cdot\gamma\overline{\alpha}_1\Big) + \\
&\quad (1-\gamma)\Big(P_{n_3=0}\overline{P}_{a'_3\neq a_3}\cdot\overline{\alpha}_0 + P_{n_3=0}\overline{P}_{a'_3\neq a_3}\cdot\gamma\overline{\alpha}_1\Big) + P_{n_3=0}\overline{P}_{a'_3\neq a_3}\cdot\gamma^2\overline{\alpha}_2\Big) + \\
&\quad \vdots \\
&= (1-\gamma)\Big(\overline{\alpha}_0\sum_{t=1}^{\infty} P_{n_t=0}\overline{P}_{a'_t\neq a_t} + \gamma\overline{\alpha}_1\sum_{t=2}^{\infty} P_{n_t=0}\overline{P}_{a'_t\neq a_t} + \gamma^2\overline{\alpha}_2\sum_{t=3}^{\infty} P_{n_t=0}\overline{P}_{a'_t\neq a_t} + \cdots\Big) \\
&= (1-\gamma)\Big(\overline{\alpha}_0[1-P(n_1>0)] + \gamma\overline{\alpha}_1[1-P(n_2>0)] + \gamma^2\overline{\alpha}_2[1-P(n_3>0)] + \cdots\Big) \\
&= (1-\gamma)\Big(\sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t - \sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t P(n_{t+1}>0)\Big) \\
&= \overline{\alpha} - (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t P(n_{t+1}>0)
\end{aligned} \tag{88}$$

Substituting (87) and (88) into (85), we get

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2\epsilon}{1-\gamma}\sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t P(n_{t+1}>0). \tag{89}$$

iii) The third part of the proof is given as follows.

Recall that n_t denote the number of times that $a'_i \neq a_i|s_i$ at state s_i for $i<t$, and $n_t = 0$ means that $a'_i = a_i|s_i$ for all $i<t$. Based on Definition A.2 (α-*coupled policy*), we have $P(a'_i = a_i|s_i) \geq 1-\alpha(s_i)$ for every s_i. Thus,

$$\begin{aligned}
P(n_{t+1}>0) &= 1 - P(n_{t+1}=0) \\
&= 1 - \prod_{i=0}^{t}\mathbb{E}_{s_i\sim\rho_i^\pi}\left[P(a'_i = a_i|s_i)\right] \\
&\leq 1 - \prod_{i=0}^{t}(1-\overline{\alpha}_i)
\end{aligned} \tag{90}$$

Substituting (90) and (84) into (89), we have

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2\epsilon}{1-\gamma}\sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t\Big(1-\prod_{i=0}^{t}(1-\overline{\alpha}_i)\Big). \tag{91}$$

Using Lemma A.5, the inequality (91) can be further simplified as

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2\epsilon}{(1-\gamma)^3}\sum_{t=0}^{\infty}\gamma^t\overline{\alpha}_t^2. \tag{92}$$

Replacing $\overline{\alpha}_t$ with $\mathbb{E}_{s\sim\rho_t^\pi}[D_{\mathrm{TV}}[\pi'||\pi](s)]$ and applying $\mathbb{E}^2[X] \leq \mathbb{E}[X^2]$, we get

$$|J(\pi') - L_\pi(\pi')| \leq \frac{2\gamma^2\epsilon}{(1-\gamma)^3}\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s\sim\rho_t^\pi}[D_{\mathrm{TV}}^2[\pi'||\pi](s)]. \tag{93}$$

Last, applying Pinsker's inequality, $2D_{\mathrm{TV}}^2[\pi'||\pi](s) \leq D_{\mathrm{KL}}[\pi'||\pi](s)$, the result follows. □

B. Proof of Analytical Policy Update Rule with Monotonic Improvement Guarantee

This proof uses *calculus of variations* to derive an analytical solution for trust region policy update.

Theorem B.1. *For any stochastic policies $\pi_{\mathrm{new}}, \pi_{\mathrm{old}}$ that are continuously differentiable on the state space $\mathcal{S}$, the inequality, $J(\pi_{\mathrm{new}}) \geq J(\pi_{\mathrm{old}})$, holds when*

$$\pi_{\mathrm{new}} = \pi_{\mathrm{old}} \cdot \frac{e^{\alpha_{\pi_{\mathrm{old}}}}}{\mathbb{E}_{a\sim\pi_{\mathrm{old}}}[e^{\alpha_{\pi_{\mathrm{old}}}}]}, \text{ where } \alpha_{\pi_{\mathrm{old}}} = \frac{A_{\pi_{\mathrm{old}}}}{C_{\pi_{\mathrm{old}}}}. \tag{94}$$

Proof. With Theorem 4.1, we can get a lower bound of the objective function $J(\pi')$ when approximating around π_{old}:

$$J(\pi') \geq L_{\pi_{\mathrm{old}}}(\pi') - \frac{1}{1-\gamma}C_{\pi_{\mathrm{old}}}\mathbb{E}_{s\sim d^{\pi_{\mathrm{old}}}}\left[D_{\mathrm{KL}}[\pi'\|\pi_{\mathrm{old}}](s)\right]. \tag{95}$$

It follows that maximizing the lower bound will give us a new policy that is not worse than π_{old}. To see this, let $I(\pi')$ denote the lower bound and π_{new} denote its maximum solution:

$$I(\pi') = L_{\pi_{\mathrm{old}}}(\pi') - \frac{1}{1-\gamma}C_{\pi_{\mathrm{old}}}\mathbb{E}_{s\sim d^{\pi_{\mathrm{old}}}}\left[D_{\mathrm{KL}}[\pi'\|\pi_{\mathrm{old}}](s)\right] \tag{96}$$

$$\pi_{\mathrm{new}} = \arg\max_{\pi'} I(\pi') \tag{97}$$

where $L_{\pi_{\mathrm{old}}}(\pi')$ is the surrogate model. Then, we have

$$J(\pi_{\mathrm{new}}) \geq I(\pi_{\mathrm{new}}) \geq I(\pi_{\mathrm{old}}) = J(\pi_{\mathrm{old}}).$$

Next, we prove that the expression of π_{new} in (94) is a necessary and sufficient condition for the optimal solution of the problem in (97).

B.1. Continuous action space

We will use *calculus of variation* to derive the analytical expression for π_{new}. Let $\pi' \in C^1(U)$ be functions defined on $U \doteq \mathcal{S}\times\mathcal{A}$. Note that the lower bound $I(\pi')$ can be rewritten as follows:

$$I(\pi') = J(\pi_{\mathrm{old}}) + \frac{1}{1-\gamma}\iint_{\mathcal{S}\times\mathcal{A}} d^{\pi_{\mathrm{old}}}(s)\left[\pi'(a|s)A_{\pi_{\mathrm{old}}}(s,a) - C_{\pi_{\mathrm{old}}}\pi'(a|s)\log\frac{\pi'(a|s)}{\pi_{\mathrm{old}}(a|s)}\right]dsda. \tag{98}$$

Note that the policy π' should be a probability distribution, which means that it integrates to 1. To ensure that, we add the following constraint:

$$H(\pi') = \frac{1}{1-\gamma}\int_{\mathcal{S}} d^{\pi_{\mathrm{old}}}(s)\Big[\int_{\mathcal{A}}\pi'(a|s)da - 1\Big]ds = 0. \tag{99}$$

Now, consider all functions in Equations (98) and (99) as variables in function spaces, and define

$$F(s,a,\pi') = d^{\pi_{\text{old}}}\left(\pi' A_{\pi_{\text{old}}} - C_{\pi_{\text{old}}}\pi'\log\pi' + C_{\pi_{\text{old}}}\pi'\log\pi_{\text{old}}\right). \tag{100}$$

$$G(s,a,\pi') = d^{\pi_{\text{old}}}\pi' - d^{\pi_{\text{old}}}\pi_{\text{old}}. \tag{101}$$

Based on Euler-Lagrange equation (Calder, 2020), there must exist a real number λ such that the optimal policy π^* satisfies

$$\nabla_{\pi'}F(s,a,\pi') - \lambda\nabla_{\pi'}G(s,a,\pi') = 0, \tag{102}$$

where λ is the Lagrange multiplier. Solving Equation (102), we have

$$d^{\pi_{\text{old}}}\left(A_{\pi_{\text{old}}} - C_{\pi_{\text{old}}}\log\pi^* - C_{\pi_{\text{old}}} + C_{\pi_{\text{old}}}\log\pi_{\text{old}} - \lambda\right) = 0, \tag{103}$$

and

$$\pi_{\text{new}} = \pi_{\text{old}}\cdot\exp\left\{\frac{A_{\pi_{\text{old}}}}{C_{\pi_{\text{old}}}} - 1 - \frac{\lambda}{C_{\pi_{\text{old}}}}\right\}. \tag{104}$$

Since π' integrates to 1, we have

$$\begin{aligned}\int_{\mathcal{A}}\pi_{\text{new}}(a|s)da &= \int_{\mathcal{A}}\pi_{\text{old}}(a|s)\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}} - 1 - \frac{\lambda}{C_{\pi_{\text{old}}}}\right\}da\\ &= e^{-1-\lambda/C_{\pi_{\text{old}}}}\int_{\mathcal{A}}\pi_{\text{old}}(a|s)\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}}\right\}da\\ &= 1\end{aligned} \tag{105}$$

Rearranging it, we get

$$\int_{\mathcal{A}}\pi_{\text{old}}(a|s)\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}}\right\}da = e^{1+\lambda/C_{\pi_{\text{old}}}}. \tag{106}$$

Taking logarithm on both sides and rearranging it, we get

$$\lambda = C_{\pi_{\text{old}}}\log\int_{\mathcal{A}}\pi_{\text{old}}(a|s)\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}}\right\}da - C_{\pi_{\text{old}}}. \tag{107}$$

Substituting (107) into (104), we get

$$\pi_{\text{new}}(a|s) = \pi_{\text{old}}\cdot\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}} - \log\int_{\mathcal{A}}\pi_{\text{old}}(a|s)\exp\left\{\frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}}\right\}da\right\}. \tag{108}$$

Denote $\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{C_{\pi_{\text{old}}}}$. Then, the optimal policy can be simplified as

$$\pi_{\text{new}} = \pi_{\text{old}}\cdot\frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]}. \tag{109}$$

Until now, we have proved the sufficient condition. Next, we prove that the policy π_{new} in Eq. (109) is also the necessary condition for the optimal solution to the maximization of $I(\pi')$.

Consider weak variations $\epsilon\eta$ such that $\pi' = \pi_{\text{new}} + \epsilon\eta$, where $\eta \in C^1(\overline{U})$ and ϵ is a real number. The second variation can be expressed as,

$$\begin{aligned}\delta^2 I &= \frac{\epsilon^2}{1-\gamma}\iint_{\mathcal{S}\times\mathcal{A}}\nabla^2_{\pi'\pi'}F(s,a,\pi')\eta^2 dsda\\ &= \frac{\epsilon^2}{1-\gamma}\iint_{\mathcal{S}\times\mathcal{A}} -\frac{d^{\pi_{\text{old}}}C_{\pi_{\text{old}}}}{\pi'}\eta^2 dsda\\ &\le 0 \quad \text{(for all weak variations } \eta)\end{aligned} \tag{110}$$

because $C_{\pi_{\text{old}}} \ge 0$, and $d^{\pi_{\text{old}}}, \pi'$ are probability distributions and thus always greater or equal to 0. Based on second-variation condition (Kot, 2014), the functional $I(\pi')$ reaches a maximum at π_{new}.

B.2. Discrete action space

For discrete actions, the functionals $I(\pi')$ and $H(\pi')$ can be rewritten as follows:

$$I(\pi') = J(\pi_{\text{old}}) + \frac{1}{1-\gamma}\int_{\mathcal{S}} d^{\pi_{\text{old}}}(s)\sum_{i=1}^{k}\left[\pi'(a_i|s)A_{\pi_{\text{old}}}(s,a_i) - C_{\pi_{\text{old}}}\pi'(a_i|s)\log\frac{\pi'(a_i|s)}{\pi_{\text{old}}(a_i|s)}\right]ds, \tag{111}$$

$$H(\pi') = \frac{1}{1-\gamma}\int_{\mathcal{S}} d^{\pi_{\text{old}}}(s)\Big[\sum_{i=1}^{k}\pi'(a_i|s) - 1\Big]ds. \tag{112}$$

Now, consider the policy as a vector of functions, $\pi' = [\pi'_1, \pi'_2, \ldots, \pi'_k]$, where $\pi'_i = \pi'(a_i|s) \in C^1(\mathcal{S})$ is a function defined on $\mathcal{S}$ given the action a_i. Then, we can define the Lagrange functions by

$$F(s,a,\pi') = d^{\pi_{\text{old}}}\sum_{i=1}^{k}\Big(\pi'_i A_{\pi_{\text{old}}} - C_{\pi_{\text{old}}}\pi'_i\log\pi'_i + C_{\pi_{\text{old}}}\pi'_i\log\pi_{i,\text{old}}\Big). \tag{113}$$

$$G(s,a,\pi') = \sum_{i=1}^{k}\pi'_i - 1. \tag{114}$$

The Euler-Lagrange Equation (102) becomes

$$\nabla_{\pi'_i}F(s,a,\pi') - \lambda\nabla_{\pi'_i}G(s,a,\pi') = 0,\ \forall i \in \{1,\ldots,k\}. \tag{115}$$

Solving the Euler-Lagrange Equations (115), we get

$$\pi^*_i = \pi_{i,\text{old}}\cdot\exp\Big\{\frac{A_{\pi_{\text{old}}}}{C_{\pi_{\text{old}}}} - 1 - \frac{\lambda}{C_{\pi_{\text{old}}}}\Big\},\ \forall i \in \{1,\ldots,k\}. \tag{116}$$

Note that π^*_i should satisfy

$$\sum_{i=1}^{k}\pi^*_i = 1. \tag{117}$$

Then, we can calculate the Lagrange multiplier λ:

$$\lambda = C_{\pi_{\text{old}}}\log\sum_{i=1}^{k}\pi_{i,\text{old}}(a_i|s)\exp\Big\{\frac{A_{\pi_{\text{old}}}(s,a_i)}{C_{\pi_{\text{old}}}}\Big\} - C_{\pi_{\text{old}}}. \tag{118}$$

Substituting (118) into (116) and use the vector form, we get

$$\pi^* = \pi_{\text{new}} = \pi_{\text{old}}\cdot\frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]}. \tag{119}$$

Use the same method as in Equation (110), we can prove that the second-variation condition is satisfied. □