# Coordinate Ascent for Off-Policy RL with Global Convergence Guarantees

**Hsin-En Su**[*1] **Yen-Ju Chen**[*1] **Ping-Chun Hsieh**[1] **Xi Liu**[2]
{littlecrazymouse.cs10,mru.11,pinghsieh}@nycu.edu.tw, xiliu.tamu@gmail.com
[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2]Applied Machine Learning, Meta AI, Menlo Park, CA, USA
[*]Equal Contribution

## Abstract

We revisit the domain of off-policy policy optimization in RL from the perspective of coordinate ascent. One commonly-used approach is to leverage the off-policy policy gradient to optimize a surrogate objective – the total discounted in expectation return of the target policy with respect to the state distribution of the behavior policy. However, this approach has been shown to suffer from the distribution mismatch issue, and therefore significant efforts are needed for correcting this mismatch either via state distribution correction or a counterfactual method. In this paper, we rethink off-policy learning via Coordinate Ascent Policy Optimization (CAPO), an off-policy actor-critic algorithm that decouples policy improvement from the state distribution of the behavior policy without using the policy gradient. This design obviates the need for distribution correction or importance sampling in the policy improvement step of off-policy policy gradient. We establish the global convergence of CAPO with general coordinate selection and then further quantify the convergence rates of several instances of CAPO with popular coordinate selection rules, including the cyclic and the randomized variants of CAPO. We then extend CAPO to neural policies for a more practical implementation. Through experiments, we demonstrate that CAPO provides a competitive approach to RL in practice.

## 1 INTRODUCTION

Policy gradient (PG) has served as one fundamental principle of a plethora of benchmark reinforcement learning algorithms (Degris et al., 2012; Lillicrap et al., 2016; Gu et al., 2017b; Mnih et al., 2016). In addition to the empirical success, PG algorithms have recently been shown to enjoy provably global convergence guarantees in the *on-policy* settings, including the true gradient settings (Agarwal et al., 2019; Bhandari and Russo, 2019; Mei et al., 2020; Cen et al., 2022) and the Monte-Carlo stochastic gradient settings (Liu et al., 2020a; Mei et al., 2021). However, on-policy PG is known to suffer from data inefficiency and lack of exploration due to the tight coupling between the learned target policy and the sampled trajectories. As a result, in many cases, *off-policy* learning is preferred to achieve better exploration with an aim to either increase sample efficiency or address the committal behavior in the on-policy learning scenarios (Mei et al., 2021; Chung et al., 2021). To address this, the off-policy PG theorem (Degris et al., 2012; Imani et al., 2018; Maei, 2018) and the corresponding off-policy actor-critic methods, which are established to optimize a *surrogate objective* defined as the total discounted return of the target policy in expectation with respect to the state distribution of the *behavior policy*, has been proposed and widely adopted to decouple policy learning from trajectory sampling (Wang et al., 2017; Gu et al., 2017a; Chung et al., 2021; Ciosek and Whiteson, 2018; Espeholt et al., 2018).

Despite the better exploration capability, off-policy PG methods are subject to the following fundamental issues: (i) *Correction for distribution mismatch*: The standard off-policy PG methods resort to a surrogate objective, which ignores the mismatch between on-policy and the off-policy state distributions. Notably, it has been shown that such mismatch could lead to sub-optimal policies as well as poor empirical performance (Liu et al., 2020b). As a result, substantial efforts are needed to correct this distribution mismatch (Imani et al., 2018; Liu et al., 2020b; Zhang et al., 2020). (ii) *Fixed behavior policy and importance sampling*: The formulation of off-policy PG presumes the use of a static behavior policy throughout training as it is designed to optimize a surrogate objective with respect to the behavior policy. However, in many cases, we do prefer that the behavior policy varies with the target policy (e.g., epsilon-greedy exploration) as it is widely known that importance

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. 

sampling could lead to significant variance in gradient estimation, especially when the behavior policy substantially deviates from the current policy. As a result, one fundamental research question that we would like to answer is: "*How to achieve off-policy policy optimization with global convergence guarantees, but without the above limitations of off-policy PG?*"

To answer this question, in this paper we take a different approach and propose an alternative off-policy policy optimization framework termed Coordinate Ascent Policy Optimization (CAPO), which revisits the policy optimization problem through the lens of coordinate ascent. Our key insight is that *the distribution mismatch and the fixed behavior policy issues in off-policy PG both result from the tight coupling between the behavior policy and the objective function in policy optimization.* To address this issue, we propose to still adopt the original objective of standard on-policy PG, but from the perspective of *coordinate ascent* with the update coordinates determined by the behavior policy. Through this design, we can completely decouple the objective function from the behavior policy while still enabling off-policy policy updates. Under the canonical tabular softmax parameterization, where each "coordinate" corresponds to a parameter specific to each state-action pair, CAPO iteratively updates the policy by performing coordinate ascent for those state-action pairs in the mini-batch, without resorting to the full gradient information or any gradient estimation. While being a rather simple method in the optimization literature, coordinate ascent and the resulting CAPO enjoy two salient features that appear rather useful in the context of RL:

- With the simple coordinate update, CAPO is capable of improving the policy by following any policy under a mild condition, directly enabling off-policy policy updates with an adaptive behavior policy. This feature addresses the issue of fixed behavior policy.
- Unlike PG, which requires having either full gradient information (the true PG setting) or an unbiased estimate of the gradient (the stochastic PG setting), updating the policy in a coordinate-wise manner allows CAPO to obviate the need for true gradient or unbiasedness while still retaining strict policy improvement in each update. As a result, this feature also obviates the need for distribution correction or importance sampling in the policy update.

To establish the global convergence of CAPO, we need to tackle the following main challenges: (i) In the coordinate descent literature, one common property is that the coordinates selected for the update are either determined according to a deterministic sequence (e.g., cyclic coordinate descent) or drawn independently from some distribution (e.g., randomized block coordinate descent) (Nesterov, 2012). By contrast, given the highly stochastic and non-i.i.d. nature of RL environments, in the general update scheme of CAPO, we impose no assumption on the data collection process, except for the standard condition of infinite visitation to each state-action pair (Singh et al., 2000; Munos et al., 2016). (ii) The function of total discounted expected return is in general non-concave, and the coordinate ascent methods could only converge to a stationary point under the general non-concave functions. Despite the above, we are able to show that the proposed CAPO algorithm attains a globally optimal policy with properly-designed step sizes under the canonical softmax parameterization. (iii) In the optimization literature, it is known that the coordinate ascent methods can typically converge slowly compared to the gradient counterpart. Somewhat surprisingly, we show that CAPO achieves comparable convergence rates as the true on-policy PG (Mei et al., 2020). Through our convergence analysis, we found that this can be attributed to the design of the state-action-dependent variable step sizes.

Built on the above results, we further generalize CAPO to the case of neural policy parameterization for practical implementation. Specifically, Neural CAPO (NCAPO) proceeds by the following two steps: (i) Given a mini-batch of state-action pairs, we leverage the tabular CAPO as a subroutine to obtain a collection of reference action distributions for those states in the mini-batch. (ii) By constructing a loss function (e.g., Kullback-Leibler divergence), we guide the policy network to update its parameters towards the state-wise reference action distributions. Such update can also be interpreted as solving a distributional regression problem.

**Our Contributions.** In this work, we revisit off-policy policy optimization and propose a novel policy-based learning algorithm from the perspective of coordinate ascent. The main contributions can be summarized as follows:

- We propose CAPO, a simple yet practical off-policy actor-critic framework with global convergence, and naturally enables direct off-policy policy updates with more flexible use of adaptive behavior policies, without the need for distribution correction or importance sampling correction to the policy gradient.
- We show that the proposed CAPO converges to a globally optimal policy under tabular softmax parameterization for general coordinate selection rules and further characterize the convergence rates of CAPO under multiple popular variants of coordinate ascent. We then extend the idea of CAPO to learning general neural policies to address practical RL settings.
- Through experiments, we demonstrate that NCAPO achieves comparable or better empirical performance than various popular benchmark methods in the MinAtar environment (Young and Tian, 2019).

**Notations**. Throughout the paper, we use $[n]$ to denote the set of integers $\{1, \cdots, n\}$. For any $x \in \mathbb{R}\backslash\{0\}$, we use $\text{sign}(x)$ to denote $\frac{x}{|x|}$ and set $\text{sign}(0) = 0$. We use $\mathbb{I}\{\cdot\}$ to denote the indicator function.

## 2 PRELIMINARIES

**Markov Decision Processes.** We consider an infinite-horizon Markov decision process (MDP) characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho)$, where (i) $\mathcal{S}$ denotes the state space, (ii) $\mathcal{A}$ denotes a *finite* action space, (iii) $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel determining the transition probability $\mathcal{P}(s'|s, a)$ from each state-action pair $(s, a)$ to a next state $s'$, where $\Delta(\mathcal{S})$ is a probability simplex over $\mathcal{S}$, (iv) $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, (v) $\gamma \in (0, 1)$ is the discount factor, and (vi) $\rho$ is the initial state distribution. In this paper, we consider learning a stationary parametric stochastic policy denoted as $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$, which specifies through a parameter vector $\theta$ the action distribution from a probability simplex $\Delta(\mathcal{A})$ over $\mathcal{A}$ for each state. For a policy $\pi_\theta$, the value function $V^{\pi_\theta} : \mathcal{S} \to \mathbb{R}$ is defined as the sum of discounted expected future rewards obtained by starting from state $s$ and following $\pi_\theta$, i.e.,

$$V^{\pi_\theta}(s) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big|\pi_\theta, s_0 = s\Big], \quad (1)$$

where $t$ represents the timestep of the trajectory $\{(s_t, a_t)\}_{t=0}^{\infty}$ induced by the policy $\pi_\theta$ with the initial state $s_0 = s$. The goal of the learner is to search for a policy that maximizes the following objective function as

$$V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}[V^{\pi_\theta}(s)]. \quad (2)$$

For ease of exposition, we use $\pi^*$ to denote an optimal policy and let $V^*(s)$ be a shorthand notation for $V^{\pi^*}(s)$. Moreover, for any given policy $\pi_\theta$, we define the $Q$-function $Q^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q^{\pi_\theta}(s, a) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big|\pi, s_0 = s, a_0 = a\Big]. \quad (3)$$

We also define the advantage function $A^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$A^{\pi_\theta}(s, a) := Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s), \quad (4)$$

which reflects the relative benefit of taking the action $a$ at state $s$ under policy $\pi_\theta$. Moreover, throughout this paper, we use $m$ as the index of the training iterations and use $\pi_m$ and $\pi_{\theta_m}$ interchangeably to denote the parameterized policy at iteration $m$.

**Policy Gradients.** The policy gradient is a popular policy optimization method that updates the parameterized policy $\pi_\theta$ by applying gradient ascent with respect to an objective function $V^{\pi_\theta}(\mu)$, where $\mu$ is some starting state distribution. The standard stochastic policy gradient theorem states that the policy gradient $\nabla_\theta V^{\pi_\theta}(\mu)$ takes the form as (Sutton et al., 1999)

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \log \pi_\theta(a|s) A^{\pi_\theta}(s, a)\big], \quad (5)$$

where the outer expectation is taken over the *discounted state visitation distribution* under $\mu$ as

$$d_\mu^{\pi_\theta}(s) := \mathbb{E}_{s_0\sim\mu}\Big[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t \,\mathbb{P}\left(s_t = s|s_0, \pi_\theta\right)\Big]. \quad (6)$$

Note that $d_\mu^{\pi_\theta}(s)$ reflects how frequently the learner would visit the state $s$ under $\pi_\theta$.

Regarding PG for off-policy learning, the learner's goal is to learn an optimal policy $\pi^*$ by following a behavior policy. Degris et al. (2012) proposed to optimize the following surrogate objective defined as

$$J^{\pi_\theta}(\beta) := \sum_{s\in\mathcal{S}} \bar{d}^\beta(s) V^{\pi_\theta}(s), \quad (7)$$

where $\beta : \mathcal{S} \to \Delta(\mathcal{A})$ is a *fixed* behavior policy and $\bar{d}^\beta(s)$ is the stationary state distribution under $\beta$ (which is assumed to exist in (Degris et al., 2012)). The resulting off-policy PG enjoys a closed-form expression as

$$\nabla_\theta J^{\pi_\theta}(\beta) = \mathbb{E}_{s\sim\bar{d}^\beta(s)}\Big[\sum_{a\in\mathcal{A}}\Big(\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(a|s) + \pi_\theta(a|s)\nabla_\theta Q^{\pi_\theta}(s, a)\Big)\Big]. \quad (8)$$

Moreover, Degris et al. (2012) showed that one can ignore the term $\pi_\theta(a|s)\nabla_\theta Q^{\pi_\theta}(s, a)$ in (8) under tabular parameterization without introducing any bias and proposed the corresponding Off-Policy Actor-Critic algorithm (Off-PAC)

$$\theta_{m+1} = \theta_m + \eta\cdot\omega_m(s, a)Q^{\pi_m}(s, a)\nabla_\theta \log \pi_{\theta_m}(s, a), \quad (9)$$

where $s$ is drawn from $\bar{d}^\beta$, $a$ is sampled from $\beta(\cdot|s)$, and $\omega_m(s, a) := \frac{\pi_m(a|s)}{\beta(a|s)}$ denotes the importance ratio. Subsequently, the off-policy PG has been generalized by incorporating state-dependent emphatic weightings (Imani et al., 2018) and introducing a counterfactual objective (Zhang et al., 2019).

**Coordinate Ascent.** Coordinate ascent (CA) methods optimize a parameterized objective function $f(\theta) : \mathbb{R}^n \to \mathbb{R}$ by iteratively updating the parameters along coordinate directions or coordinate hyperplanes. Specifically, in the $m$-th iteration, the CA update along the $i_m$-th coordinate is

$$\theta_{m+1} = \theta_m + \eta \cdot [\nabla_\theta f(\theta)]_{i_m} e_{i_m}, \quad (10)$$

where $e_{i_m}$ denotes the one-hot vector of the $i_m$-th coordinate and $\eta$ denotes the step size. The main difference among the CA methods mainly lies in the selection of coordinates for updates. Popular variants of CA methods include:

- **Cyclic CA**: The choice of coordinate proceeds in a predetermined cyclic order (Saha and Tewari, 2013). For example, one possible configuration is $i_m \leftarrow m \bmod n$.
- **Randomized CA**: In each iteration, one coordinate is drawn randomly from some distribution with support $[n]$ (Nesterov, 2012).

Moreover, the CA updates can be extended to the *blockwise* scheme (Tseng, 2001; Beck and Tetruashvili, 2013), where multiple coordinates are selected in each iteration. Despite the simplicity, the CA methods have been widely used in variational inference (Jordan et al., 1999) and large-scale machine learning (Nesterov, 2012) due to its parallelization capability. To the best of our knowledge, CA has remained largely unexplored in the context of policy optimization.

# 3 METHODOLOGY

In this section, we present the proposed CAPO algorithm, which improves the policy through coordinate ascent updates. Throughout this section, we consider the class of tabular softmax policies. Specifically, for each state-action pair $(s, a)$, let $\theta(s, a)$ denote the corresponding parameter. The probability of selecting action $a$ given state $s$ is given by $\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}$.

## 3.1 Coordinate Ascent Policy Optimization

To begin with, we present the general policy update scheme of CAPO. The discussion about the specific instances of CAPO along with their convergence rates will be provided subsequently in Section 3.3. To motivate the policy improvement scheme of CAPO, we first state the following lemma (Agarwal et al., 2019; Mei et al., 2020).

**Lemma 1.** *Under tabular softmax policies, the standard policy gradient with respect to $\theta$ is given by*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a). \tag{11}$$

Based on Lemma 1, we see that the update direction of each coordinate is completely determined by the *sign* of the advantage function. Accordingly, the proposed general CAPO update scheme is as follows: In each update iteration $m$, let $B_m$ denote the mini-batch of state-action pairs sampled by the behavior policy. The batch $B_m$ determines the coordinates of the policy parameter to be updated. Specifically, the policy is updated by

$$\begin{aligned}&\theta_{m+1}(s,a)\\ &= \theta_m(s,a) + \alpha_m(s,a)\mathbb{I}\{(s,a) \in B_m\} \cdot \operatorname{sign}\left(A^{\pi_{\theta_m}}(s,a)\right),\end{aligned} \tag{12}$$

where $\alpha_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ is the function that controls the *magnitude* of the update and plays the role of the learning rate, the term $\operatorname{sign}(A^{\pi_{\theta_m}}(s,a))$ controls the update *direction*, and $B_m$ is the sampled batch of state-action pairs in the $m$-th iteration and determines the *coordinate selection*. Under CAPO, only those parameters associated with the sampled state-action pairs will be updated accordingly, as suggested by (12). Based on this, we could reinterpret $B_m$ as produced by a *coordinate generator*, which could be induced by the behavior policies.

**Remark 1.** Note that under the general CAPO update, the learning rate $\alpha$ is state-action-dependent. This is one salient difference from the learning rates of conventional coordinate ascent methods in the optimization literature (Nesterov, 2012; Saha and Tewari, 2013). As will be shown momentarily in Section 3.2, this design allows CAPO to attain global optimality without statistical assumptions about the samples (i.e., the selected coordinates). On the other hand, while it appears that the update rule in (12) only involves the sign of the advantage function, the magnitude of the advantage $|A(s, a)|$ could also be taken into account if needed through $\alpha(s, a)$, which is also state-action-dependent. As a result, (12) indeed provides a flexible expression that separates the effect of the sign and magnitude of the advantage. Interestingly, as will be shown in the next subsections, we establish that CAPO can achieve global convergence without the knowledge of the magnitude of the advantage. Also, we would like to highlight the fundamental differences between CAPO and PG: (i) Coordinate ascent via $\mathbb{I}\{(s, a) \in B_m\}$ which opens up the use of various coordinate selection rules and accordingly (ii) the required adaptive step size for convergence.

**Remark 2.** Compared to the off-policy PG methods (Degris et al., 2012; Wang et al., 2017; Imani et al., 2018), one salient property of CAPO is that it allows *off-policy learning* through coordinate ascent on the original *on-policy* total expected reward $\mathbb{E}_{s\sim\rho}[V^\pi(s)]$, instead of the *off-policy* total expected reward over the discounted state visitation distribution induced by the behavior policy. On the other hand, regarding the learning of a critic, similar to the off-policy PG methods, CAPO can be integrated with any off-policy policy evaluation algorithm, such as Retrace (Munos et al., 2016) or V-trace (Espeholt et al., 2018).

## 3.2 Asymptotic Global Convergence of CAPO With General Coordinate Selection

In this section we discuss the convergence result of CAPO under softmax parameterization. In the subsequent analysis, we assume that the following Condition 1 is satisfied.

**Condition 1.** $\lim_{M\to\infty} \sum_{m=1}^{M} \mathbb{I}\{(s, a) \in B_m\} \to \infty$

Note that Condition 1 is rather mild as it could be met by exploratory behavior policies (e.g., $\epsilon$-greedy policies) given the off-policy capability of CAPO. Moreover, Condition 1 is similar to the standard condition of infinite visitation required by various RL methods (Singh et al., 2000; Munos et al., 2016). Notably, Condition 1 indicates that under CAPO the coordinates are not required to be selected by following a specific policy, as long as infinite visitation to every state-action pair is satisfied. This feature naturally enables flexible off-policy learning, justifies the use of a replay buffer, and enables the flexibility to decouple policy improvement from value estimation.

We first show that CAPO guarantees strict improvement under tabular softmax parameterization.

**Lemma 2** (Strict Policy Improvement). *Under the CAPO update given by (12), we have $V^{\pi_{m+1}}(s) \geq V^{\pi_m}(s)$, for all $s \in S$, for all $m \in \mathbb{N}$.*

*Proof.* The proof can be found in Appendix A.1. ☐

We proceed to substantiate the benefit of the state-action-dependent learning rate used in the general CAPO update in (12) by showing that CAPO can attain a globally optimal policy with a properly designed learning rate $\alpha(\cdot, \cdot)$.

**Theorem 1.** *Consider a tabular softmax parameterized policy $\pi_\theta$. Under (12) with $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if Condition 1 is satisfied, then we have $V^{\pi_m}(s) \to V^*(s)$ as $m \to \infty$, for all $s \in \mathcal{S}$.*

*Proof Sketch.* The detailed proof can be found in Appendix A.2. To highlight the main ideas of the analysis, we provide a sketch of the proof as follows: (i) Since the expected total reward is bounded above, with the strict policy improvement property of CAPO update (cf. Lemma 2), the sequence of value functions is guaranteed to converge, i.e., the limit of $V^{\pi_m}(s)$ exists. (ii) The proof proceeds by contradiction. We suppose that CAPO converges to a sub-optimal policy, which implies that there exists at least one state-action pair $(s', a')$ such that $A^{(\infty)}(s', a') > 0$ and $\pi_\infty(a''|s'') = 0$ for all state-action pair $(s'', a'')$ satisfying $A^{(\infty)}(s'', a'') > 0$. As a result, this implies that for any $\epsilon > 0$, there must exist a time $M^\epsilon$ such that $\pi_m(a''|s'') < \epsilon, \forall m > M^\epsilon$. (iii) However, under CAPO update, we show that the policy weight shall approach 1 for the state-action pair which has the greatest advantage value, and this leads to a contradiction. ☐

**Remark 3.** The proof of Theorem 1 is inspired by (Agarwal et al., 2019). Nevertheless, the analysis of CAPO presents its own salient challenge: Under true PG, the policy updates in all the iterations can be fully determined once the initial policy and the step size are specified. By contrast, under CAPO, the policy obtained in each iteration depends on the selected coordinates, which can be almost arbitrary under Condition 1. This makes it challenging to establish a contradiction under CAPO, compared to the argument of directly deriving the policy parameters in the limit in true PG (Agarwal et al., 2019). Despite this, we address the challenge by using a novel induction argument based on the action ordering w.r.t. the $Q$ values in the limit.

**Remark 4.** Notably, the condition of the learning rate $\alpha$ in Theorem 1 does not depend on the advantage, but only on the action probability $\pi_\theta(a|s)$. As a result, the CAPO update only requires the sign of the advantage function, without the knowledge of the magnitude of the advantage. Therefore, CAPO can still converge even under a low-fidelity critic that merely learns the sign of the advantage function.

### 3.3 Convergence Rates of CAPO With Specific Coordinate Selection Rules

In this section, we proceed to characterize the convergence rates of CAPO under softmax parameterization and the three specific coordinate generators, namely, Cyclic, Batch, and Randomized CAPO.

- **Cyclic CAPO**: Under Cyclic CAPO, every state action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ will be chosen for policy update by the coordinate generator cyclically. Specifically, Cyclic CAPO sets $|B_m| = 1$ and $\bigcup_{i=1}^{|\mathcal{S}||\mathcal{A}|} B_{m \cdot |\mathcal{S}||\mathcal{A}| + i} = \mathcal{S} \times \mathcal{A}$.
- **Randomized CAPO**: Under Randomized CAPO, in each iteration, one state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is chosen randomly from some coordinate generator distribution $d_{\text{gen}}$ with support $\mathcal{S} \times \mathcal{A}$ for policy update, where $d_{\text{gen}}(s,a) > 0$ for all $(s,a)$. For ease of exposition, we focus on the case of a fixed $d_{\text{gen}}$. Our convergence analysis can be readily extended to the case of time-varying $d_{\text{gen}}$.
- **Batch CAPO**: Under Batch CAPO, we let each batch contain all of the state-action pairs, i.e., $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$, in each iteration. Despite that Batch CAPO may not be a very practical choice, we use this variant to further highlight the difference in convergence rate between CAPO and the true PG.

We proceed to state the convergence rates of the above three instances of CAPO as follows.

**Theorem 2** (Cyclic CAPO). *Consider a tabular softmax policy $\pi_\theta$. Under Cyclic CAPO with $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$ and $|B_m| = 1$, $\bigcup_{i=1}^{|\mathcal{S}||\mathcal{A}|} B_{m \cdot |\mathcal{S}||\mathcal{A}| + i} = \mathcal{S} \times \mathcal{A}$, we have:*

$$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{|\mathcal{S}||\mathcal{A}|}{c} \cdot \frac{1}{m}, \quad \text{for all } m \geq 1 \quad (13)$$

*where* $c = \frac{(1-\gamma)^4}{2} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \min\left\{\frac{\min_s \mu(s)}{2}, \frac{(1-\gamma)}{|\mathcal{S}||\mathcal{A}|}\right\} > 0.$

*Proof Sketch.* The detailed proof and the upper bound of the partial sum can be found in Appendix B. To highlight the main ideas of the analysis, we provide a sketch of the proof as follows: (i) We first write the one-step improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in state visitation distribution, policy weight, and advantage value, and also construct the lower bound of it. (ii) We then construct the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$. (iii) Since the bound in (i) and (ii) both include advantage value, we can connect them and construct the upper bound of the performance difference using one-step improvement of the performance. (iv) Finally, we can get the desired convergence rate by induction. ☐

Notably, it is somewhat surprising that Theorem 2 holds under Cyclic CAPO *without* any further requirement on the specific cyclic ordering. This indicates that Cyclic CAPO is rather flexible in the sense that it provably attains $\mathcal{O}(\frac{1}{m})$ convergence rate under any cyclic ordering or even cyclic orderings that vary across cycles. On the flip side, such a flexible coordinate selection rule also imposes significant challenges on the analysis: (i) While Lemma 2 ensures strict improvement in each iteration, it remains unclear how much improvement each Cyclic CAPO update can actually achieve, especially under an *arbitrary cyclic ordering*. This is one salient difference compared to the analysis of the true PG (Mei et al., 2020). (ii) Moreover, an update along one coordinate can already significantly change the advantage value (and its sign as well) of other state-action pairs. Therefore, it appears possible that there might exist a well-crafted cyclic ordering that leads to only minimal improvement in each coordinate update within a cycle.

Despite the above, we tackle the challenges by arguing that in each cycle, under a properly-designed variable step size $\alpha$, there must exist at least one state-action pair such that the one-step improvement is sufficiently large, regardless of the cyclic ordering. Moreover, by the same proof technique, Theorem 2 can be readily extended to CAPO with almost-cyclic coordinate selection, where the cycle length is greater than $|S||A|$ and each coordinate appears at least once.

We extend the proof technique of Theorem 2 to establish the convergence rates of the Batch and Randomized CAPO.

**Theorem 3** (Batch CAPO). *Consider a tabular softmax policy $\pi_\theta$. Under Batch CAPO with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$ and $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$, we have :*

$$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{1}{c} \cdot \frac{1}{m}, \quad \text{for all } m \geq 1 \tag{14}$$

*where* $c = \frac{(1-\gamma)^4}{|\mathcal{A}|} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \min_s \{\mu(s)\} > 0.$

*Proof.* The proof and the upper bound of the partial sum can be found in Appendix B.2. □

**Theorem 4** (Randomized CAPO). *Consider a tabular softmax policy $\pi_\theta$. Under Randomized CAPO with $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$, we have :*

$$\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}} [V^*(\rho) - V^{\pi_m}(\rho)] \leq \frac{1}{c} \cdot \frac{1}{m}, \quad \text{for all } m \geq 1 \tag{15}$$

*where* $c = \frac{(1-\gamma)^4}{2} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \min_{(s,a)} \{d_{gen}(s,a) \cdot \mu(s)\} > 0$ *and* $d_{gen} : \mathcal{S} \times \mathcal{A} \to (0,1)$, $d_{gen}(s,a) = \mathbb{P}((s,a) \in B_m)$.

*Proof.* The proof and the upper bound of the partial sum can be found in Appendix B.3. □

**Remark 5.** The above three specific instances of CAPO all converge to a globally optimal policy at a rate $\mathcal{O}(\frac{1}{m})$ and attains a better pre-constant than the standard policy gradient (Mei et al., 2020) under tabular softmax parameterization. Moreover, as the CAPO update can be combined with a variety of coordinate selection rules, one interesting future direction is to design coordinate generators that improve over the convergence rates of the above three instances.

# 4 DISCUSSIONS

In this section, we describe the connection between CAPO and the existing policy optimization methods and present additional useful features of CAPO.

**Batch CAPO and True PG.** We use Batch CAPO to highlight the fundamental difference in convergence rate between CAPO and the true PG as they both take all the state-action pairs into account in one policy update. Compared to the rate of true PG (Mei et al., 2020), Batch CAPO removes the dependency on the size of state space $|\mathcal{S}|$ and $\inf_{m\geq 1} \pi_m(a^*|s)^2$, making CAPO more robust to the policy initialization. In true PG, these two terms arise in the construction of the Łojasiewicz inequality, which quantifies the amount of policy improvement with the help of an optimal policy $\pi^*$, which explains why $\inf_{m\geq 1} \pi_m(a^*|s)^2$ appears in the convergence rate. By contrast, in Batch CAPO, we quantify the amount of policy improvement based on the coordinate with the largest advantage, and this proof technique contributes to the improved rate of Batch CAPO compared to true PG. Moreover, we emphasize that this technique is feasible in Batch CAPO but not in true PG mainly due to the properly-designed learning rate of CAPO.

In addition, we can tell from Table 1 that CAPO also removes the dependency on the term $\left\|d_\mu^{\pi^*}/\mu\right\|_\infty^2$ in PG which also arises in the construction of the Łojasiewicz inequality and is environment-dependent due to $d_\mu^{\pi^*}$ By contrast, the term $\min_{s\in\mathcal{S}}\{\mu(s)\}^{-1}$ in CAPO can be configured by the agent, and hence the constant of the rate is less sensitive to the environment.

**Connecting CAPO With Natural Policy Gradient.** The natural policy gradient (NPG) (Kakade, 2001) exploits the landscape of the parameter space and updates the policy by:

$$\theta_{m+1} = \theta_m + \eta \left(F_\rho^{\theta_m}\right)^\dagger \nabla_\theta J^{\pi_\theta}(\rho), \tag{16}$$

where $\eta$ is the step size and $\left(F_\rho^{\theta_m}\right)^\dagger$ is the Moore-Penrose pseudo inverse of the Fisher information matrix $F_\rho^{\theta_m} := \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_m}}, a\sim\pi_{\theta_m}(\cdot|s)}[(\nabla_\theta \log \pi_{\theta_m}(a|s))(\nabla_\theta \log \pi_{\theta_m}(a|s))^\top]$. Moreover, under softmax parameterization, the true NPG update takes the following form (Agarwal et al., 2019):

$$\theta_{m+1} = \theta_m + \frac{\eta}{1-\gamma} A^{\pi_{\theta_m}}, \tag{17}$$

where $A^{\pi_{\theta_m}}$ denotes the $|S||A|$-dimensional vector of all the advantage values of $\pi_{\theta_m}$. It has been shown that the true NPG can attain linear convergence (Mei et al., 2021;

Table 1: A summary of convergence rates under tabular softmax parameterization under different algorithms.

| **Algorithm** | **Convergence Rate** |
|---|---|
| Policy Gradient (Mei et al., 2020) | $V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{16\cdot\lvert\mathcal{S}\rvert}{\inf_{m\geq 1}\pi_m(a^*\vert s)^2\cdot(1-\gamma)^6} \cdot \left\lVert \frac{d_\mu^{\pi^*}}{\mu} \right\rVert_\infty^2 \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \frac{1}{m}$ |
| Cyclic CAPO (Theorem 2) | $V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{2\cdot\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \max\left\{\frac{2}{\min_s \mu(s)}, \frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)}\right\} \cdot \frac{1}{m}$ |
| Batch CAPO (Theorem 3) | $V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{\lvert\mathcal{A}\rvert}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \frac{1}{\min\limits_s\{\mu(s)\}} \cdot \frac{1}{m}$ |
| Randomized CAPO (Theorem 4) | $\mathop{\mathbb{E}}\limits_{(s_m,a_m)\sim d_{\text{gen}}}[V^*(\rho) - V^{\pi_m}(\rho)] \leq \frac{2}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \frac{1}{\min\limits_{(s,a)}\{d_{\text{gen}}(s,a)\cdot\mu(s)\}} \cdot \frac{1}{m}$ |

Khodadadian et al., 2021b). Given the expression in (17), CAPO can be interpreted as adapting NPG to the mini-batch or stochastic settings. That said, compared to true NPG, CAPO only requires the sign of the advantage function, not the magnitude of the advantage. On the other hand, it has recently been shown that some variants of on-policy stochastic NPG could exhibit committal behavior and thereby suffer from convergence to sub-optimal policies (Mei et al., 2021). The analysis of CAPO could also provide useful insights into the design of stochastic NPG methods. Interestingly, in the context of variational inference, a theoretical connection between coordinate ascent and the natural gradient has also been recently discovered (Ji et al., 2021).

**CAPO for Low-Fidelity RL Tasks.** One salient feature of CAPO is that it requires only the sign of the advantage function, instead of the exact advantage value. It has been shown that accurate estimation of the advantage value could be rather challenging under benchmark RL algorithms (Ilyas et al., 2019). As a result, CAPO could serve as a promising candidate solution for RL tasks with low-fidelity or multi-fidelity value estimation (Cutler et al., 2014; Kandasamy et al., 2016; Khairy and Balaprakash, 2022). To illustrate this empirically, we have also conducted an experiment in section 6 to showcase the robustness to the inaccurate critics under CAPO.

**CAPO for On-Policy Learning.** The original motivation of CAPO is to achieve off-policy policy updates without the issues of distribution mismatch and fixed behavior policy. Despite this, the CAPO scheme in (12) can also be used in an *on-policy* manner. Notably, the design of on-policy CAPO is subject to a similar challenge of committal behavior in on-policy stochastic PG and stochastic NPG (Chung et al., 2021; Mei et al., 2021). Specifically: (i) We show that on-policy CAPO with a fixed step size could converge to sub-optimal policies through a multi-armed bandit example similar to that in (Chung et al., 2021). (ii) We design a proper step size for on-policy CAPO and establish asymptotic global convergence. Through a simple bandit experiment, we show that this variant of on-policy CAPO can avoid the committal behavior. Due to space limitation, all the above results are provided in Appendix C.

## 5 PRACTICAL IMPLEMENTATION of CAPO

To address the large state and action spaces of the practical RL problems, we proceed to parameterize the policy for CAPO by a neural network and make use of its powerful representation ability. As presented in Section 4, the coordinate update and variable learning rate are two salient features of CAPO. These features are difficult to preserve if the policy is trained in a completely end-to-end manner. Instead, we take a two-step approach by first leveraging the tabular CAPO to derive target action distributions and then design a loss function that moves the output of the neural network towards the target distribution. Specifically, we designed a neural version of CAPO, called Neural Coordinate Ascent Policy Optimization (NCAPO): Let $f_\theta(s, a)$ denote the output of the policy network parameterized by $\theta$, for each $(s, a)$. In NCAPO, we use neural softmax policies, i.e., $\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))}$.

- Inspired by the tabular CAPO, we compute a target softmax policy $\pi_{\hat{\theta}}(s, a)$ by following the CAPO update (12)
$$\tilde{\theta}(s, a) = f_\theta(s, a) + \alpha(s, a)\mathbb{I}\{(s, a) \in B\}\cdot\text{sign}\left(A^{\pi_\theta}(s, a)\right). \tag{18}$$
The target action distribution is then computed w.r.t. $\tilde{\theta}$ as $\tilde{\pi}(a|s) = \frac{\exp(\tilde{\theta}(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\tilde{\theta}(s,a'))}$.
- Finally, we learn $f_\theta$ by minimizing the NCAPO loss, which is the KL-divergence loss between the current policy and the target policy:
$$\mathcal{L}(\theta) = \sum_{s\in B} D_{\text{KL}}\left(\pi_\theta(\cdot|s) \| \tilde{\pi}(\cdot|s)\right). \tag{19}$$

## 6 EXPERIMENTAL RESULTS

In this section, we empirically evaluate the performance of CAPO on several benchmark RL tasks. We evaluate NCAPO in MinAtar (Young and Tian, 2019), a simplified Arcade Learning Environment (ALE), and consider a variety

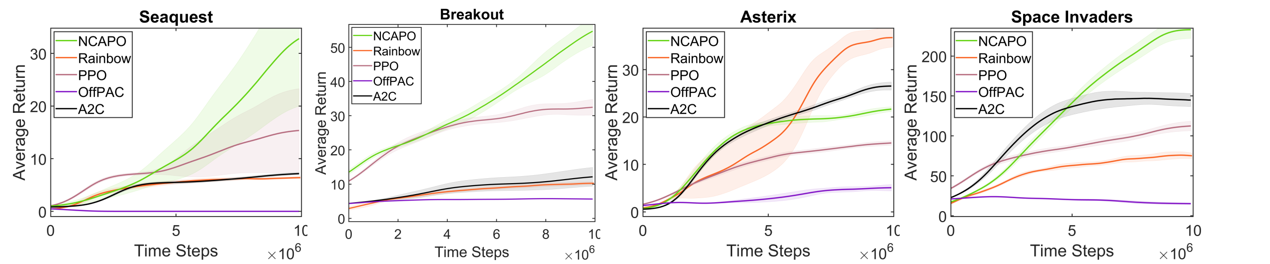


Figure 1: A comparison between the performance of NCAPO and other benchmark methods algorithms in MinAtar. All the results are averaged over 10 random seeds (with the shaded area showing the range of mean $\pm\, 0.5 \cdot$ std).

of environments, including *Seaquest, Breakout, Asterix, and Space Invaders*. Each environment is associated with $10 \times 10 \times n$ binary state representation, which corresponds to the $10 \times 10$ grid and $n$ channels (the value $n$ depends on the game).

**Benchmark Methods.** We select several benchmark methods for comparison, including Rainbow (Hessel et al., 2018; Obando-Ceron and Castro, 2021), PPO (Schulman et al., 2017), Off-PAC (Degris et al., 2012), and Advantage Actor-Critic (A2C) (Mnih et al., 2016), to demonstrate the effectiveness of NCAPO. For Rainbow, we use the code provided by (Obando-Ceron and Castro, 2021) without any change. For the other methods, we use the open-source implementation provided by Stable Baselines3 (Raffin et al., 2019).

**Empirical Evaluation.** The detailed implementation of NCAPO is provided in Appendix G. From Figure 1, we can observe that NCAPO has the best performance in *Seaquest, Breakout, Space Invaders*. We also see that NCAPO is more robust across tasks than PPO and Rainbow. For example, Rainbow performs especially well in *Asterix*, while relatively poorly in *Space Invaders*. PPO performs relatively strong in *Breakout* and *Seaquest*, but converges rather slowly in *Asterix*. Off-PAC with a uniform behavior policy has very little improvement throughout training in all the tasks due to the issue of fixed behavior policy, which could hardly find sufficiently long trajectories with high scores. By contrast, NCAPO outperforms all the benchmark methods in three out of four environments, while being on par with other methods in the remaining environment.

**Sign vs. Magnitude of Advantage.** In addition to being useful in off-policy learning without importance sampling, we can tell from 12 that CAPO does not require the exact magnitude of advantage and can be more robust to inaccurate critics or highly stochastic environments. To illustrate this, we further evaluate NCAPO in MinAtar with noisy rewards (i.e. for $5\%$ of steps a large noise $\mathcal{N}(0, \sigma^2)$ is injected). Figure 2 shows NCAPO is more robust against reward noise, while PPO can barely learn when $\sigma = 10$.

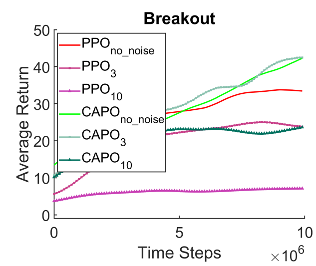


Figure 2: A comparison between the performance of NCAPO and PPO in Breakout under different reward noises ($5\%$ of steps a large noise $\mathcal{N}(0, \sigma^2)$ is injected, subscript: $\sigma$).

## 7 RELATED WORK

**Off-Policy Policy Gradients**. Off-policy learning via PG has been an on-going research topic. Built on the off-policy PG theorem (Degris et al., 2012; Silver et al., 2014; Zhang et al., 2019; Imani et al., 2018), various off-policy actor-critic algorithms have been developed with an aim to achieve more sample-efficient RL (Wang et al., 2017; Gu et al., 2017b; Chung et al., 2021; Ciosek and Whiteson, 2018; Espeholt et al., 2018; Schmitt et al., 2020). In the standard off-policy PG formulation, the main idea lies in the use of a surrogate objective, which is the expected total return with expectation taken over the stationary distribution induced by the behavior policy. While this design avoids the issue of an exponentially-growing importance sampling ratio, it has been shown that this surrogate objective can suffer from convergence to sub-optimal policies due to distribution mismatch, and distribution correction is therefore needed, either via a learned density correction ratio (Liu et al., 2020b) or emphatic weighting (Maei, 2018; Zhang et al., 2019, 2020). On the other hand, off-policy actor-critic based on NPG has been recently shown to achieve provable sample complexity guarantees in both tabular (Khodadadian et al., 2021a) and linear function approximation setting (Chen and Maguluri, 2022; Chen et al., 2022). Another line of research is on characterizing the convergence of off-policy actor-critic methods in the *offline* setting, where the learner is given only a fixed dataset of samples (Xu et al., 2021; Huang and Jiang, 2022).

Some recent attempts propose to enable off-policy learning beyond the use of policy gradient. For example, (Laroche and Tachet des Combes, 2021) extends the on-policy PG to an off-policy policy update by generalizing the role of the discounted state visitation distribution. (Laroche and Des Combes, 2022) proposes to use the gradient of the cross-entropy loss with respect to the action with maximum Q. Both approaches are shown to attain similar convergence rates as the on-policy true PG. Different from all the above, CAPO serves as the first attempt to address off-policy policy optimization through the lens of coordinate ascent, without using the policy gradient.

**Exploiting the Sign of Advantage Function.** As pointed out in Section 2, the sign of the advantage function (or temporal difference (TD) residual as a surrogate) can serve as an indicator of policy improvement. For example, (Van Hasselt and Wiering, 2007) proposed Actor Critic Learning Automaton (ACLA), which is designed to reinforce only those state-action pairs with positive TD residual and ignore those pairs with non-positive TD residual. The idea of ACLA is later extended by (Zimmer et al., 2016) to Neural Fitted Actor Critic (NFAC) , which learns neural policies for continuous control, and penalized version of NFAC for improved empirical performance (Zimmer and Weng, 2019). On the other hand, (Tessler et al., 2019) proposes generative actor critic (GAC), a distributional policy optimization approach that leverages the actions with positive advantage to construct a target distribution. By contrast, CAPO takes the first step towards understanding the use of coordinate ascent with convergence guarantees for off-policy RL.

## 8 CONCLUSION

We propose CAPO, which takes the first step towards addressing off-policy policy optimization by exploring the use of coordinate ascent in RL. Through CAPO, we enable off-policy learning without the need for importance sampling or distribution correction. We show that the general CAPO can attain asymptotic global convergence and establish the convergence rates of CAPO with several popular coordinate selection rules. Moreover, through experiments, we show that the neural implementation of CAPO can serve as a competitive solution compared to the benchmark RL methods and thereby demonstrates the future potential of CAPO.

### Acknowledgements

This material is based upon work partially supported by the National Science and Technology Council (NSTC), Taiwan under Contract No. 110-2628-E-A49-014 and Contract No. 111-2628-E-A49-019, and based upon work partially supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

## References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv:1908.00261*, 2019.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv:1906.01786*, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214, 2022.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

Wesley Chung, Valentin Thomas, Marlos C Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *International Conference on Machine Learning*, pages 1999–2009, 2021.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Mark Cutler, Thomas J Walsh, and Jonathan P How. Reinforcement learning with multi-fidelity simulators. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3888–3895, 2014.

Thomas Degris, Martha White, and Richard S Sutton. Off-Policy Actor-Critic. In *International Conference on Machine Learning*, pages 179–186, 2012.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. IMPALA: Scalable Distributed Deep-RL With Importance Weighted Actor-Learner Architectures. In *International Conference on Machine Learning*, pages 1407–1416, 2018.

S Gu, T Lillicrap, Z Ghahramani, RE Turner, B Schölkopf, and S Levine. Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning. *Advances in Neural Information Processing Systems*, 2017:3847–3856, 2017a.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E Turner, and Sergey Levine. Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic. In *International Conference on Learning Representations*, 2017b.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jiawei Huang and Nan Jiang. On the convergence rate of off-policy policy optimization methods with density-ratio correction. In *International Conference on Artificial Intelligence and Statistics*, pages 2658–2705, 2022.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2019.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. *Advances in Neural Information Processing Systems*, 31, 2018.

Geng Ji, Debora Sujono, and Erik B Sudderth. Marginalized stochastic natural gradients for black-box variational inference. In *International Conference on Machine Learning*, pages 4870–4881, 2021.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233, 1999.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in Neural Information Processing Systems*, 29, 2016.

Sami Khairy and Prasanna Balaprakash. Multifidelity reinforcement learning with control variates. *arXiv:2206.05165*, 2022.

Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning*, pages 5420–5431, 2021a.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *IEEE Conference on Decision and Control (CDC)*, pages 3794–3799, 2021b.

Romain Laroche and Remi Tachet Des Combes. Beyond the policy gradient theorem for efficient policy updates in actor-critic algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 5658–5688, 2022.

Romain Laroche and Remi Tachet des Combes. Dr Jekyll & Mr Hyde: the strange case of off-policy policy updates. *Advances in Neural Information Processing Systems*, 34: 24442–24454, 2021.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020a.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with stationary distribution correction. In *Uncertainty in Artificial Intelligence*, pages 1180–1190, 2020b.

Hamid Reza Maei. Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv:1802.07842*, 2018.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016.

Remi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and Efficient Off-Policy Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29:1054–1062, 2016.

Yu Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Johan S Obando-Ceron and Pablo Samuel Castro. Revisiting rainbow: Promoting more insightful and inclusive deep reinforcement learning research. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2021.

Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. Stable baselines3. https://github.com/DLR-RM/stable-baselines3, 2019.

Ankan Saha and Ambuj Tewari. On the Nonasymptotic Convergence of Cyclic Coordinate Descent Methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013.

Simon Schmitt, Matteo Hessel, and Karen Simonyan. Off-Policy Actor-Critic With Shared Experience Replay. In *International Conference on Machine Learning*, pages 8545–8554, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395, 2014.

Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine learning*, 38(3):287–308, 2000.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Chen Tessler, Guy Tennenholtz, and Shie Mannor. Distributional Policy Optimization: An Alternative Approach for Continuous Control. *Advances in Neural Information Processing Systems*, 32:1352–1362, 2019.

Paul Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Hado Van Hasselt and Marco A Wiering. Reinforcement Learning in Continuous Action Spaces. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279, 2007.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2017.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pages 11581–11591, 2021.

Kenny Young and Tian Tian. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments. *arXiv:1903.03176*, 2019.

Shangtong Zhang, Wendelin Boehmer, and Shimon Whiteson. Generalized off-policy actor-critic. *Advances in Neural Information Processing Systems*, 32, 2019.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pages 11204–11213, 2020.

Matthieu Zimmer and Paul Weng. Exploiting the sign of the advantage function to learn deterministic policies in continuous domains. In *International Joint Conferences on Artificial Intelligence*, 2019.

Matthieu Zimmer, Yann Boniface, and Alain Dutech. Neural Fitted Actor-Critic. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)*, 2016.

# Appendix

## A PROOFS OF THE THEORETICAL RESULTS IN SECTION 3.2

### A.1 Proof of Lemma 2

**Lemma 3** (Performance Difference Lemma in (Kakade and Langford, 2002))**.** *For each state $s_0$, the difference in the value of $s_0$ between two policies $\pi$ and $\pi'$ can be characterized as:*

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right] \tag{20}$$

Now we are ready to prove Lemma 2. For ease of exposition, we restate Lemma 2 as follows.

**Lemma.** *Under the CAPO update given by (12), we have $V^{\pi_{m+1}}(s) \geq V^{\pi_m}(s)$, for all $s \in S$, for all $m \in \mathbb{N}$.*

*Proof of Lemma 2.* Note that by the definition of $A(s,a)$, we have

$$\sum_{a\in\mathcal{A}} \pi_m(a|s)A^m(s,a) = 0, \quad \forall s \in \mathcal{S} \tag{21}$$

To simplify notation, let $Z_m(s) := \sum_{a\in\mathcal{A}} \exp(\theta_m(s,a))$. Then, $\pi_m(a|s)$ and $\pi_{m+1}(a|s)$ can be simplified as:

$$\pi_m(a|s) = \frac{\exp(\theta_m(s,a))}{Z_m(s)}, \quad \pi_{m+1}(a|s) = \frac{\exp(\theta_{m+1}(s,a))}{Z_{m+1}(s)}. \tag{22}$$

By Lemma 3, in order to show that $V^{\pi_{m+1}}(s) \geq V^{\pi_m}(s), \forall s \in S$, it is sufficient to show that

$$\sum_{a\in\mathcal{A}} \pi_{m+1}(a|s)A^m(s,a) > 0, \quad \forall s \in \mathcal{S}. \tag{23}$$

For ease of notation, we define $B_m(s) := \{a|(s,a) \in B_m\}$. To establish (22), we have that for all $s \in \mathcal{S}$,

$$\sum_{a\in\mathcal{A}} \pi_{m+1}(a|s)A^m(s,a) = \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{m+1}(s,a))}{Z_{m+1}(s)}A^m(s,a) \tag{24}$$

$$= \frac{Z_m(s)}{Z_{m+1}(s)}\sum_{a\in\mathcal{A}} \frac{\exp(\theta_{m+1}(s,a))}{Z_m(s)}A^m(s,a) \tag{25}$$

$$= \frac{Z_m(s)}{Z_{m+1}(s)}\left[\sum_{a\in B_m(s)} \frac{\exp(\theta_{m+1}(s,a))}{Z_m(s)}A^m(s,a) + \sum_{a\notin B_m(s)} \frac{\exp(\theta_m(s,a))}{Z_m(s)}A^m(s,a)\right] \tag{26}$$

$$> \frac{Z_m(s)}{Z_{m+1}(s)}\left[\sum_{a\in B_m(s)} \frac{\exp(\theta_m(s,a))}{Z_m(s)}A^m(s,a) + \sum_{a\notin B_m(s)} \frac{\exp(\theta_m(s,a))}{Z_m(s)}A^m(s,a)\right] \tag{27}$$

$$= \frac{Z_m(s)}{Z_{m+1}(s)}\sum_{a\in\mathcal{A}} \pi_m(a|s)A^m(s,a) \tag{28}$$

$$= 0, \tag{29}$$

where (27) holds by the CAPO update given by (12). □

### A.2 Proof of Theorem 1

Since $\{V^m\}$ is bounded above and enjoys strict improvement by Lemma 2. By the monotone convergence theorem, the limit of $\{V^m\}$ is guaranteed to exist. Similarly, we know that the limit of $\{Q^m\}$ also exists. We use $V^{(\infty)}(s)$ and $Q^{(\infty)}(s,a)$ to denote the limits of $\{V^{(m)}(s)\}$ and $\{Q^{(m)}(s,a)\}$, respectively. We also define $A^{(\infty)}(s,a) := Q^{(\infty)}(s,a) - V^{(\infty)}(s)$. Our concern is whether the corresponding policy $\pi_\infty$ is optimal. Inspired by (Agarwal et al., 2019), we first define the following three sets as

$$I_0^s := \left\{a | Q^{(\infty)}(s,a) = V^{(\infty)}(s)\right\}, \tag{30}$$

$$I_+^s := \left\{a | Q^{(\infty)}(s,a) > V^{(\infty)}(s)\right\}, \tag{31}$$

$$I_-^s := \left\{a | Q^{(\infty)}(s,a) < V^{(\infty)}(s)\right\}. \tag{32}$$

By definition, $V^{(\infty)}$ is optimal if and only if $I_+^s$ is empty, for all $s$. We prove by contradiction that $V^{(\infty)}$ is optimal by showing $I_+^s = \emptyset$.

**Main steps of the proof**. The proof procedure can be summarized as follows:

- Step 1: We first assume $V^{(\infty)}$ is not optimal so that by definition $\exists s \in \mathcal{S}, I_+^s \neq \emptyset$.
- Step 2: We then show in Lemma 5, $\forall s \in \mathcal{S}$, actions $a_- \in I_-^s$ have zero weights in policy (i.e. $\pi_\infty(a_-|s) = 0, \forall a_- \in I_-^s$).
- Step 3: Since the actions in $I_-^s$ have zero probability, by (21), this directly implies Lemma 6: $\forall I_+^s \neq \emptyset$, $a \in I_+^s$ must also have zero probability (i.e. $\pi_\infty(a_+|s) = 0, \forall a_+ \in I_+^s$).
- Step 4: Moreover, under CAPO, in the sequel we can show Claim 1, which states that as long as Condition 1 is satisfied, there must exist one action $a_+ \in I_+^s$ such that $\lim_{m\to\infty} \pi_m(a_+|s) = 1$. This contradicts the assumption that $\exists s \in \mathcal{S}, I_+^s \neq \emptyset$, proving that $I_+^s = \emptyset, \forall s \in \mathcal{S}$.

**Lemma 4.** *Under CAPO, there exists $M_1$ such that for all $m > M_1, s \in \mathcal{S}, a \in \mathcal{A}$, we have :*

$$A^{(m)}(s,a) < -\frac{\Delta}{4}, \quad \textit{for } a \in I_-^s, \tag{33}$$

$$A^{(m)}(s,a) > \frac{\Delta}{4}, \quad \textit{for } a \in I_+^s, \tag{34}$$

*where* $\Delta := \min_{\{s,a | A^{(\infty)}(s,a) \neq 0\}} \left|A^{(\infty)}(s,a)\right|$.

*Proof of Lemma 4.* Given the strict policy improvement property of CAPO in Lemma 2, this can be shown by applying Lemma C.4 in (Agarwal et al., 2019). □

**Lemma 5.** *Under CAPO, $\pi_\infty(a_-|s) = 0, \forall s \in \mathcal{S}, a_- \in I_-^s$.*

*Proof of Lemma 5.* Lemma 4 shows that for all $m > M_1$, the sign of $A^{(m)}(s,a)$ is fixed. Moreover, we know that under CAPO update, $\theta_m(s,a_-)$ is non-increasing, $\forall a_- \in I_-^s, \forall m > M_1$. Similarly, $\forall a_+ \in I_+^s, m > M_1$, $\theta_m(s,a_+)$ is non-decreasing. By Condition 1, all the state-action pairs with negative advantage are guaranteed to be sampled for infinitely many times as $m \to \infty$. Under the CAPO update in (12), we have

$$\theta_{m+1}(s,a_-) - \theta_m(a_-|s) \leq -\log\left(\frac{1}{\pi_m(a_-|s)}\right) < 0. \tag{35}$$

Given the infinite visitation, we know that $\lim_{m\to\infty} \theta_m(s,a_-) = -\infty$. □

We now show in Lemma 6 that Lemma 5 implies $\sum_{a_+ \in I_+^s} \pi_\infty(a_+|s) = 0$.

**Lemma 6.** *If $I_+^s \neq \emptyset$ is true, then Lemma 5 implies $\sum_{a_+ \in I_+^s} \pi_\infty(a_+|s) = 0$.*

*Proof of Lemma 6.* Recall from (21) that $\sum_{a\in\mathcal{A}} \pi_m(a|s) A^m(s,a) = 0, \forall s \in \mathcal{S}, m > 0$. By definition, $\sum_{a_0 \in I_0^s} \pi_\infty(a_0|s) A^\infty(s,a_0) = 0$, which directly implies that

$$\sum_{a_+ \in I_+^s} \pi_\infty(a_+|s) A^\infty(s,a_+) = \sum_{a\in\mathcal{A}} \pi_\infty(a|s) A^\infty(s,a) - \sum_{a_0 \in I_0^s} \pi_\infty(a_0|s) A^\infty(s,a) - \sum_{a_- \in I_-^s} \pi_\infty(a_-|s) A^\infty(s,a) \tag{36}$$

$$= 0 - 0 - 0 = 0, \tag{37}$$

where the second equality holds by Lemma 5. Since $A^\infty(s,a_+) > 0$ and $\pi_\infty(a_+|s) \geq 0$, we have $\sum_{a_+ \in I_+^s} \pi_\infty(a_+|s) = 0$. This completes the proof of Lemma 6. □

In Lemma 6, we have that if $I_+^s \neq \emptyset$ is true, then $\pi_m(a_+|s) \rightarrow 0$ as $m \rightarrow \infty$. To establish contradiction, we proceed to show in the following Claim 1 that there must exist one action $a \in I_+^s$ such that $\lim_{m\rightarrow\infty} \pi_m(a|s) = 1$, which contradicts Lemma 6 and hence implies the desired result that $I_+^s = \emptyset$.

If $I_+^s \neq \emptyset$ is true, then there exist $K$ such that $\forall m > K, s \in \mathcal{S}$, we have:

$$Q^m(s,a^+) > Q^m(s,a^0) > Q^m(s,a^-), \quad \text{for all } a^+ \in I_+^s, a^0 \in I_0^s, a^- \in I_-^s. \tag{38}$$

Without loss of generality, assume that the order of $Q^m$, $\forall m > K$, can be written as

$$Q^m(s,\tilde{a}^+) > Q^m(s,a_1) > Q^m(s,a_2) > \cdots > Q^m(s,a_{|\mathcal{A}|-1}), \quad \text{provided that } I_+^s \neq \emptyset, \tag{39}$$

where $\tilde{a}^+ := \text{argmax}_{a^+ \in I_+^s} Q^{(\infty)}(s,a^+)$. Note that we simplify the case above by considering "strictly greater than" instead of "greater than or equal to", but the simplification can be relaxed with a little extra work.

**Claim 1.** *If $I_+^s \neq \emptyset$ is true, then there must exist one action $a_+ \in I_+^s$ such that $\lim_{m\rightarrow\infty} \pi_m(a_+|s) = 1$ under (12) with $\alpha_m(s,a) \geq \log \frac{1}{\pi_m(a|s)}$.*

To establish Claim 1, we show that if $I_+^s \neq \emptyset$, then $\lim_{m\rightarrow\infty} \pi_m(a|s) = 0$ for all $a \neq \tilde{a}^+$ by induction. For ease of exposition, we first present the following propositions.

**Proposition 1.** *For any $m \geq 1, s \in \mathcal{S}, a \in \mathcal{A}$, if $A^m(s,a) \leq 0$ and $\exists\ a' \neq a$, $a' \in B_m(s)$, satisfying $A^m(s,a') > 0$, then $\pi_{m+1}(a|s) \leq \frac{1}{2}$, regardless of whether $a \in B_m(s)$ or not.*

*Proof of Proposition 1.*

Since $A^m(s,a) \leq 0$, we have $\text{sign}(A^m(s,a)) \cdot \alpha_m(s,a) \leq 0$. As a result, we have:

$$\begin{cases} \pi_{m+1}(a|s) = \frac{\exp(\theta_m(s,a)+\text{sign}(A^m(s,a))\cdot\alpha_m(s,a))}{Z_{m+1}(s)} \leq \frac{\exp(\theta_m(s,a))}{Z_{m+1}(s)} \leq \frac{Z_m(s)}{Z_{m+1}(s)} \\ \pi_{m+1}(a'|s) = \frac{\exp(\theta_m(s,a')+\alpha_m(s,a'))}{Z_{m+1}(s)} \geq \frac{\exp(\theta_m(s,a')+\log(\frac{1}{\pi_m(a'|s)}))}{Z_{m+1}(s)} = \frac{Z_m(s)}{Z_{m+1}(s)} \end{cases} \tag{40}$$

Hence, we have $\pi_{m+1}(a'|s) \geq \pi_{m+1}(a|s)$. Since $\pi_{m+1}(a'|s) + \pi_{m+1}(a|s) \leq 1$, we get $\pi_{m+1}(a|s) \leq \frac{1}{2}$.

□

**Proposition 2.** *For any $s \in \mathcal{S}, a \in \mathcal{A} \setminus \{\tilde{a}^+\}$, if $\exists\ T \in \mathbb{N}$ such that $\forall m > T$, $A^m(s,a) \leq 0$, then $\exists\ n \in \mathbb{N}$, $\bar{K} \in \mathbb{N}$ such that $A^{m+n+1}(s,a) < 0$, $\forall\ m > \bar{K}$.*

*Proof of Proposition 2.*

By Condition 1 and $I_+^s \neq \emptyset$, there exist some finite $n \in \mathbb{N}$ such that $\exists\ a' \neq a$, $a' \in B_{m+n}(s)$, satisfying $A^{m+n+1}(s,a') > 0$. Then, by Proposition 1, we have

$$\pi_{m+n+1}(a|s) \leq \frac{1}{2}, \quad \forall m \geq T. \tag{41}$$

Hence, we have

$$V^{m+n+1}(s) = \sum_{a\in\mathcal{A}} \pi_{m+n+1}(a|s) \cdot Q^{m+n+1}(s,a) \geq \frac{1}{2} \cdot \left(Q^{m+n+1}(s,a) + Q^{m+n+1}(s,a')\right) \tag{42}$$

$$\text{where } a' = \underset{\substack{a''\in\mathcal{A} \\ Q^\infty(s,a'')>Q^\infty(s,a)}}{\text{argmin}} Q^\infty(s,a'') \tag{43}$$

Moreover, by the ordering of $Q^m$ and that $\lim_{m\rightarrow\infty} Q^m(s,a) = Q^\infty(s,a)$, for $\epsilon = \frac{1}{4} \cdot (Q^\infty(s,a') - Q^\infty(s,a)) > 0$, $\exists\ \bar{T}$ such that for all $m > \bar{T}$:

$$\begin{cases} Q^m(s,a) \in (Q^\infty(s,a) - \epsilon, Q^\infty(s,a) + \epsilon) \\ Q^m(s,a') \in (Q^\infty(s,a') - \epsilon, Q^\infty(s,a') + \epsilon) \end{cases} \tag{44}$$

Finally, we have that for all $m > \max\left\{T, \bar{T}\right\}$:

$$V^{m+n+1}(s) \geq \frac{1}{2} \cdot \left(Q^{m+n+1}(s,a) + Q^{m+n+1}(s,a')\right) \tag{45}$$

$$> \frac{1}{2} \cdot \left((Q^{\infty}(s,a)) + (Q^{\infty}(s,a'))\right) \tag{46}$$

$$> \frac{1}{2} \cdot \left(Q^{\infty}(s,a) + Q^{\infty}(s,a')\right) - \epsilon \tag{47}$$

$$= Q^{\infty}(s,a) + \epsilon > Q^{m+n+1}(s,a). \tag{48}$$

The above is equivalent to $A^{m+n+1}(s,a) < 0 \;\forall\; m > \bar{K}$, where $\bar{K} = \max\left\{T, \bar{T}\right\}$. □

**Proposition 3.** *If $V^m(s) \in \left(Q^m(s, a_{|\mathcal{A}|-k}), Q^m(s, a_{|\mathcal{A}|-(k+2)})\right)$, then $\exists\; T' \in \mathbb{N}$ such that for all $m > T'$:*

$$\frac{\sum_{\substack{a\in\mathcal{A}\\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)}{\sum_{\substack{a\in\mathcal{A}\\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)} \geq \frac{Q^m(s,a_{|\mathcal{A}|-(k+1)}) - Q^m(s,a_{|\mathcal{A}|-1})}{Q^m(s,a_{|\mathcal{A}|-(k+2)}) - Q^m(s,a_{|\mathcal{A}|-(k+1)})} \tag{49}$$

$$\textit{provided } V^m(s) \in \left(Q^m(s,a_{|\mathcal{A}|-k}), Q^m(s,a_{|\mathcal{A}|-(k+2)})\right) \tag{50}$$

*Proof of Proposition 3.*

Since $V^m(s) \in \left(Q^m(s, a_{|\mathcal{A}|-k}), Q^m(s, a_{|\mathcal{A}|-(k+2)})\right)$, we have $A^m(s, a_{|\mathcal{A}|-j}) < 0, \;\; \forall\; j = 1, 2, \ldots, k$. By Condition 1, there exists some finite $n \in \mathbb{N}$ such that $\bar{a}^+ \in \mathcal{B}_{m+n}(s)$ for some $\bar{a}^+ \in \left\{\tilde{a}^+, a_1, a_2, \ldots, a_{|\mathcal{A}|-(k+2)}\right\}$.
Hence, we have that for all $\bar{a}^- \in \left\{a_{|\mathcal{A}|-k}, a_{|\mathcal{A}|-(k-1)}, \ldots, a_{|\mathcal{A}|-1}\right\}$,

$$\frac{\pi_{m+n+1}(\bar{a}^+|s)}{\pi_{m+n+1}(\bar{a}^-|s)} \geq \frac{\frac{e^{\theta_{m+n}(s,\bar{a}^+)+\log\frac{1}{\pi_{m+n}(\bar{a}^+|s)}}}{Z_{m+n+1}(s)}}{\frac{e^{\theta_{m+n}(s,\bar{a}^-)}}{Z_{m+n+1}(s)}} = \frac{Z_{m+n}(s)}{e^{\theta_{m+n}(s,\bar{a}^-)}} = \frac{1}{\pi_{m+n}(\bar{a}^-|s)}. \tag{51}$$

Since $\lim_{m\to\infty} \pi_m(s,a) = 0$, we have $\forall z \in \mathbb{Z}, \exists\; T \in \mathbb{N}$ such that $\frac{\pi_{m+n+1}(\bar{a}^+|s)}{\pi_{m+n+1}(\bar{a}^-|s)} \geq z, \;\; \forall m > T$. For $m > K$, we have $Q^m(s, a_{|\mathcal{A}|-(k+2)}) - Q^m(s, a_{|\mathcal{A}|-(k+1)}) > 0$. Hence, by simply choosing $z = \frac{1}{\mathcal{A}} \cdot \frac{Q^m(s,a_{|\mathcal{A}|-(k+1)})-Q^m(s,a_{|\mathcal{A}|-1})}{Q^m(s,a_{|\mathcal{A}|-(k+2)})-Q^m(s,a_{|\mathcal{A}|-(k+1)})}$ and taking the summation of the ratio over $\bar{a}^+$ and $\bar{a}^-$, we can reach the desired result with $T' = \max\{K, T\}$. □

Now, we are ready to prove Claim 1 by an induction argument.

*Proof of Claim 1.*

- Show that if $I^s_+ \neq \emptyset$, then $\lim_{m\to\infty} \pi_m(a_{|\mathcal{A}|-1}|s) = 0$:
  By the ordering of $Q^m$, we have:

$$V^m(s) = \sum_{a\in\mathcal{A}} \pi_m(a|s) \cdot Q^m(s,a) \geq 1 \cdot Q^m(s, a_{|\mathcal{A}|-1}), \quad \forall m > K\ . \tag{52}$$

  Hence, for all $m > K$, we have:

$$A^m(s, a_{|\mathcal{A}|-1}) = Q^m(s, a_{|\mathcal{A}|-1}) - V^m(s) \leq Q^m(s, a_{|\mathcal{A}|-1}) - Q^m(s, a_{|\mathcal{A}|-1}) = 0. \tag{53}$$

  Therefore, by Proposition 2, we have $\exists\; n_{|\mathcal{A}|-1} \in \mathbb{N}$, $K_{|\mathcal{A}|-1} \in \mathbb{N}$ such that:

$$A^{m+n_{|\mathcal{A}|-1}+1}(s, a_{|\mathcal{A}|-1}) < 0, \quad \forall m > K_{|\mathcal{A}|-1}\ . \tag{54}$$

  Moreover,

$$\operatorname{sign}(A^m(s, a_{|\mathcal{A}|-1})) \cdot \alpha_m\left(s, a_{|\mathcal{A}|-1}\right) < 0, \quad \forall m > K_{|\mathcal{A}|-1}. \tag{55}$$

  With the monotone-decreasing property and the infinite visitation condition, it is guaranteed that $\lim_{m\to\infty} \theta_m(s, a_{|\mathcal{A}|-1}) = -\infty$. Hence, we have $\lim_{m\to\infty} \pi_m(s, a_{|\mathcal{A}|-1}) = 0$.

- Suppose that $\lim_{m\to\infty} \pi_m(a_{|\mathcal{A}|-1}|s) = \lim_{m\to\infty} \pi_m(a_{|\mathcal{A}|-2}|s) = \cdots = \lim_{m\to\infty} \pi_m(a_{|\mathcal{A}|-k}|s) = 0$, where $k \in [1, (|\mathcal{A}|-2)]$. Then we would like to derive $\lim_{m\to\infty} \pi_m(a_{|\mathcal{A}|-(k+1)}|s)$:
  By the above assumption, we have:

$$\lim_{m\to\infty} \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) = 0 \tag{56}$$

By Proposition 3, $\exists\ K'_{|\mathcal{A}|-(k+1)} \in \mathbb{N}$ such that $\forall\ m > K'_{|\mathcal{A}|-(k+1)}$, we can establish the ratio between the summation of policy weight of the policy worse than $a_{|\mathcal{A}|-(k+1)}$ and the policy better than $a_{|\mathcal{A}|-(k+1)}$:

$$\frac{\sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)}{\sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)} \geq \frac{Q^m(s,a_{|\mathcal{A}|-(k+1)}) - Q^m(s,a_{|\mathcal{A}|-1})}{Q^m(s,a_{|\mathcal{A}|-(k+2)}) - Q^m(s,a_{|\mathcal{A}|-(k+1)})} \tag{57}$$

$$\text{provided } V^m(s) \in \big(Q^m(s,a_{|\mathcal{A}|-k}), Q^m(s,a_{|\mathcal{A}|-(k+2)})\big) \tag{58}$$

And by the ordering of $Q^m$, we have:

$$V^m(s) = \sum_{a\in\mathcal{A}} \pi_m(a|s)\cdot Q^m(s,a) \tag{59}$$

$$= \Bigg[ \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)\cdot Q^m(s,a) + Q^m(s,a_{|\mathcal{A}|-(k+1)})\cdot \pi_m(a_{|\mathcal{A}|-(k+1)}|s) \tag{60}$$

$$+ \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s)\cdot Q^m(s,a) \Bigg] \tag{61}$$

$$\geq \Bigg[ Q^m(s,a_{|\mathcal{A}|-(k+2)})\cdot \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) + Q^m(s,a_{|\mathcal{A}|-(k+1)})\cdot \pi_m(a_{|\mathcal{A}|-(k+1)}|s) \tag{62}$$

$$+ Q^m(s,a_{|\mathcal{A}|-1})\cdot \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) \Bigg], \quad \forall m > K\ . \tag{63}$$

Hence, for all $m > K'_{|\mathcal{A}|-(k+1)}$, we have:

$$A^m(s,a_{|\mathcal{A}|-(k+1)}) = Q^m(s,a_{|\mathcal{A}|-(k+1)}) - V^m(s) \tag{64}$$

$$\leq Q^m(s,a_{|\mathcal{A}|-(k+1)}) - \Bigg[ Q^m(s,a_{|\mathcal{A}|-(k+2)})\cdot \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) \tag{65}$$

$$+ Q^m(s,a_{|\mathcal{A}|-(k+1)})\cdot \pi_m(a_{|\mathcal{A}|-(k+1)}|s) \tag{66}$$

$$+ Q^m(s,a_{|\mathcal{A}|-1})\cdot \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)<Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) \Bigg] \tag{67}$$

$$\tag{68}$$

$$= \big(Q^m(s,a_{|\mathcal{A}|-(k+1)}) - Q^m(s,a_{|\mathcal{A}|-(k+2)})\big)\cdot \sum_{\substack{a\in\mathcal{A} \\ Q^{\infty}(s,a)>Q^{\infty}(s,a_{|\mathcal{A}|-(k+1)})}} \pi_m(a|s) \tag{69}$$

$$+\left(Q^m(s,a_{|\mathcal{A}|-(k+1)})-Q^m(s,a_{|\mathcal{A}|-1})\right)\cdot\sum_{\substack{a\in\mathcal{A}\\ Q^\infty(s,a)<Q^\infty(s,a_{|\mathcal{A}|-(k+1)})}}\pi_m(a|s) \tag{70}$$

$$\leq\left(Q^m(s,a_{|\mathcal{A}|-(k+1)})-Q^m(s,a_{|\mathcal{A}|-(k+2)})\right)\cdot\sum_{\substack{a\in\mathcal{A}\\ Q^\infty(s,a)>Q^\infty(s,a_{|\mathcal{A}|-(k+1)})}}\pi_m(a|s) \tag{71}$$

$$+\left(Q^m(s,a_{|\mathcal{A}|-(k+1)})-Q^m(s,a_{|\mathcal{A}|-1})\right)\cdot\frac{Q^m(s,a_{|\mathcal{A}|-(k+2)})-Q^m(s,a_{|\mathcal{A}|-(k+1)})}{Q^m(s,a_{|\mathcal{A}|-(k+1)})-Q^m(s,a_{|\mathcal{A}|-1})} \tag{72}$$

$$\cdot\sum_{\substack{a\in\mathcal{A}\\ Q^\infty(s,a)>Q^\infty(s,a_{|\mathcal{A}|-(k+1)})}}\pi_m(a|s) \tag{73}$$

$$=0 \tag{74}$$

By Proposition 2, we have $\exists\ n_{|\mathcal{A}|-(k+1)}\in\mathbb{N}$, $K_{|\mathcal{A}|-(k+1)}\in\mathbb{N}$ such that:

$$A^{m+n_{|\mathcal{A}|-(k+1)}+1}(s,a_{|\mathcal{A}|-(k+1)})<0,\quad \forall m>K_{|\mathcal{A}|-(k+1)} \tag{75}$$

Moreover,

$$\text{sign}(A^m(s,a_{|\mathcal{A}|-1}))\cdot\alpha_m\left(s,a_{|\mathcal{A}|-(k+1)}\right)<0,\quad \forall m>K_{|\mathcal{A}|-(k+1)} \tag{76}$$

With the monotone-decreasing property and the infinite visitation, it is guaranteed that $\lim_{m\to\infty}\theta_m(s,a_{|\mathcal{A}|-(k+1)})=-\infty$. Hence we have $\lim_{m\to\infty}\pi_m(s,a_{|\mathcal{A}|-(k+1)})=0$.

Finally we complete the induction and so we conclude that $\forall a\neq\tilde{a^+}$, $\lim_{m\to\infty}\pi_m(s,a)=0$, which is equivalent to $\lim_{m\to\infty}\pi_m(s,\tilde{a^+})=1$. This completes the proof of Claim 1.

□

Now we are ready to put everything together and prove Theorem 1. For ease of exposition, we restate Theorem 1 as follows.

**Theorem.** *Consider a tabular softmax parameterized policy $\pi_\theta$, under (12) with $\alpha_m(s,a)\geq\log(\frac{1}{\pi_{\theta_m}(a|s)})$, if Condition 1 is satisfied, then we have $V^{\pi_m}(s)\to V^*(s)$ as $m\to\infty$, for all $s\in\mathcal{S}$.*

*Proof of Theorem 1.* In Claim 1, we have that if $I_+^s\neq\emptyset$ is true, then there must exist one action $a\in I_+^s$ such that $\lim_{m\to\infty}\pi_m(a|s)=1$. This leads to the contradiction with Lemma 6, and finally we get the desired result that $I_+^s=\emptyset$, implying that $V^{(\infty)}$ is optimal. □

# B PROOFS OF THE CONVERGENCE RATES OF CAPO IN SECTION 3.3

**Lemma 7.** $|A^m(s,a)|\leq\frac{1}{1-\gamma}\cdot(1-\pi_m(a|s))$, *for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.*

*Proof of Lemma 7.*

If $A^m(s,a)>0$ :

$$|A^m(s,a)|=Q^{\pi_m}(s,a)-V^{\pi_m}(s) \tag{77}$$
$$=Q^{\pi_m}(s,a)-\sum_{a'\in\mathcal{A}}\pi_m(a'|s)\cdot Q^{\pi_m}(s,a') \tag{78}$$
$$\leq Q^{\pi_m}(s,a)-\pi_m(a|s)\cdot Q^{\pi_m}(s,a) \tag{79}$$
$$=Q^{\pi_m}(s,a)\cdot(1-\pi_m(a|s)) \tag{80}$$
$$\leq\frac{1}{1-\gamma}\cdot(1-\pi_m(a|s)) \tag{81}$$

If $A^m(s,a) \leq 0$ :

$$|A^m(s,a)| = V^{\pi_m}(s) - Q^{\pi_m}(s,a) \tag{82}$$

$$= \sum_{a' \in \mathcal{A}} \pi_m(a'|s) \cdot Q^{\pi_m}(s,a') - Q^{\pi_m}(s,a) \tag{83}$$

$$= \sum_{a' \neq a} \pi_m(a'|s) \cdot Q^{\pi_m}(s,a') - (1 - \pi_m(a|s)) \cdot Q^{\pi_m}(s,a) \tag{84}$$

$$\leq \sum_{a' \neq a} \pi_m(a'|s) \cdot Q^{\pi_m}(s,a') \tag{85}$$

$$\leq \frac{1}{1-\gamma} \cdot \sum_{a' \neq a} \pi_m(a'|s) \tag{86}$$

$$= \frac{1}{1-\gamma} \cdot (1 - \pi_m(a|s)) \tag{87}$$

□

**Lemma 8.** $(V^*(s) - V^{\pi_m}(s))^2 \leq \left(\frac{1}{1-\gamma} \cdot A^m(\tilde{s_m}, \tilde{a_m})\right)^2$*, for all* $m \geq 1$ *where* $(\tilde{s_m}, \tilde{a_m}) = \underset{(s,a) \in \mathcal{S} \times \mathcal{A}}{\operatorname{argmax}} A^m(s,a)$.

*Proof of Lemma 8.*

$$(V^*(s) - V^{\pi_m}(s))^2 = \left(\frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi^*}(s') \sum_{a' \in \mathcal{A}} \pi^*(a'|s') \cdot A^m(s',a')\right)^2 \tag{88}$$

$$\leq \left(\frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi^*}(s') \cdot \max_{a' \in \mathcal{A}} A^m(s',a')\right)^2 \tag{89}$$

$$\leq \left(\frac{1}{1-\gamma} \cdot \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} A^m(s',a')\right)^2 \tag{90}$$

$$= \left(\frac{1}{1-\gamma} \cdot A^m(\tilde{s_m}, \tilde{a_m})\right)^2 \tag{91}$$

The first equation holds by Lemma 3.
The first and the second inequality hold since the value inside the quadratic term is non-negative.

□

**Lemma 9.** $V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot (V^*(\mu) - V^{\pi_m}(\mu))$

*Proof of Lemma 9.*

$$V^*(\rho) - V^{\pi_m}(\rho) = \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_\rho^{\pi^*}(s) \sum_{a \in \mathcal{A}} \pi^*(a|s) \cdot A^m(s,a) \tag{92}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_\mu^{\pi^*}(s) \cdot \frac{d_\rho^{\pi^*}(s)}{d_\mu^{\pi^*}(s)} \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^m(s,a) \tag{93}$$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi^*}} \right\|_\infty \cdot \sum_{s \in \mathcal{S}} d_\mu^{\pi^*}(s) \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^m(s,a) \tag{94}$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \sum_{s \in \mathcal{S}} d_\mu^{\pi^*}(s) \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^m(s,a) \tag{95}$$

$$= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot (V^*(\mu) - V^{\pi_m}(\mu)) \tag{96}$$

The first and the last equation holds by the performance difference lemma in Lemma 3.
The first and second inequality holds since the value inside the summation is non-negative.

□

**Lemma 10.** $d^\pi_\mu(s) \geq (1-\gamma)\cdot\mu(s),$ *for any* $\pi, s \in \mathcal{S}$ *where* $\mu(s)$ *is some starting state distribution of the MDP.*

*Proof of Lemma 10.*

$$\begin{aligned} d^\pi_\mu(s) &= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[d^\pi_\mu(s)\right] && (97)\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{P}(s_t = s \mid s_0, \pi)\right] && (98)\\ &\geq \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\mathbb{P}(s_0 = s \mid s_0, \pi)\right] && (99)\\ &= (1-\gamma)\cdot\mu(s) && (100)\end{aligned}$$

The first equation holds by the performance difference lemma in Lemma 3.
The second and the third equation hold since the value inside quadratic term is non-negative.

□

**Lemma 11.** *Given* $\delta_{m+1} \leq \delta_m - c\cdot\delta_m^2$ *where* $\delta_m \leq \frac{1}{1-\gamma}$ *for all* $m \geq 1$ *and* $c \leq \frac{1-\gamma}{2}$*, then* $\delta_m \leq \frac{1}{c}\cdot\frac{1}{m}$ *and* $\sum_{m=1}^{M}\delta_m \leq \min\left\{\sqrt{\frac{M}{c\cdot(1-\gamma)}}, \frac{\log M+1}{c}\right\}$ *for all* $m \geq 1$.

*Proof of Lemma 11.*

We prove this lemma by induction. For $m \leq 2$, $\delta_m \leq \frac{1}{c}\cdot\frac{1}{m}$ directly holds since $c \leq \frac{1-\gamma}{2}$ and $\delta_m \leq \frac{1}{1-\gamma}$.
Let $f_t(x) = x - c\cdot x^2 = -c(x-\frac{1}{2c})^2 + \frac{1}{4c}$. Then $f_t(x)$ is monotonically increasing in $[0, \frac{1}{2c}]$. And so we have :

$$\begin{aligned} \delta_{m+1} &\leq f_t(\delta_m) && (101)\\ &\leq f_t(\frac{1}{c}\cdot\frac{1}{m}) && (102)\\ &= \frac{1}{c}\cdot(\frac{1}{m} - \frac{1}{m^2}) && (103)\\ &\leq \frac{1}{c}\cdot\frac{1}{m+1} && (104)\end{aligned}$$

and by summing up $\delta_m$, we have :

$$\begin{aligned} \sum_{m=1}^{M}\delta_m &\leq \sum_{m=1}^{M}\frac{1}{c}\cdot\frac{1}{m} && (105)\\ &= \frac{1}{c}\cdot\sum_{m=1}^{M}\frac{1}{m} && (106)\\ &\leq \frac{1}{c}\cdot(\ln M + 1) && (107)\end{aligned}$$

On the other hand, we also have :

$$\sum_{m=1}^{M} \delta_m^2 \leq \frac{1}{c} \cdot \sum_{m=1}^{M} (\delta_m - \delta_{m+1}) \tag{108}$$

$$\leq \frac{1}{c} \cdot (\delta_1 - \delta_{M+1}) \tag{109}$$

$$\leq \frac{1}{c} \cdot \frac{1}{1-\gamma} \tag{110}$$

Therefore, by Cauchy-Schwarz,

$$\sum_{m=1}^{M} \delta_m \leq \sqrt{M} \cdot \sqrt{\sum_{m=1}^{M} \delta_m^2} \tag{111}$$

$$\leq \sqrt{M} \cdot \sqrt{\frac{1}{c} \cdot \frac{1}{1-\gamma}} \tag{112}$$

$$= \sqrt{\frac{M}{c \cdot (1-\gamma)}} \tag{113}$$

□

**Lemma 12.** *Under the CAPO update (12) with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s_m, a_m)\}$ and $A^m(s_m, a_m) > 0$, then the policy weight difference $\pi_{m+1}(a|s) - \pi_m(a|s)$ can be written as :*

$$\pi_{m+1}(a|s) - \pi_m(a|s) = \begin{cases} \frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)} & \text{, if } s = s_m, a = a_m \\ -\frac{1-\pi_m(a_m|s_m)}{2-\pi_m(a_m|s_m)} \cdot \pi_m(a|s) & \text{, if } s = s_m, a \neq a_m \\ 0 & \text{, else} \end{cases} \tag{114}$$

*Proof of Lemma 12.*

For $s = s_m, a = a_m$:

$$\pi_{m+1}(a_m|s_m) - \pi_m(a_m|s_m) = \frac{e^{\theta_{m+1}(s_m,a_m)}}{\sum\limits_{a \in \mathcal{A}} e^{\theta_{m+1}(s_m,a)}} - \pi_m(a_m|s_m) \tag{115}$$

$$= \frac{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m))}}{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m))} + \sum\limits_{a \neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{116}$$

$$= \frac{e^{\theta_m(s_m,a_m)+ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(a_m)}})}}{e^{\theta_m(s_m,a_m)+ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(s_m,a_m)}})} + \sum\limits_{a \neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{117}$$

$$= \frac{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)}}{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)} + \sum\limits_{a \neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{118}$$

$$= \frac{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)}}{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)} + (\frac{1}{\pi_m(a_m|s_m)} - 1) \cdot e^{\theta_m}(s_m,a_m)} - \pi_m(a_m|s_m) \tag{119}$$

$$= \frac{\frac{1}{\pi_m(a_m|s_m)}}{\frac{2}{\pi_m(a_m|s_m)} - 1} - \pi_m(a_m|s_m) \tag{120}$$

$$= \frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)} \tag{121}$$

For $s = s_m, a \neq a_m$:

$$\pi_{m+1}(a|s_m) - \pi_m(a|s_m) = \frac{e^{\theta_{m+1}(s_m,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s_m,a)}} - \pi_m(a|s_m) \tag{122}$$

$$= \frac{e^{\theta_m(s_m,a)}}{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m)} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{123}$$

$$= \frac{e^{\theta_m(s_m,a)}}{e^{\theta_m(s_m,a_m)+ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(s_m,a_m)}})} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{124}$$

$$= \frac{e^{\theta_m(s_m,a)}}{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{125}$$

$$= \frac{e^{\theta_m(s_m,a)}}{\frac{e^{\theta_m(s_m,a_m)}}{\pi_m(a_m|s_m)} + (\frac{1}{\pi_m(a_m|s_m)} - 1)\cdot e^{\theta_m}(s_m,a_m)} - \pi_m(a|s_m) \tag{126}$$

$$= \left(\frac{e^{\theta_m(s_m,a)}}{(\frac{2}{\pi_m(a_m|s_m)} - 1)\cdot e^{\theta_m}(s_m,a_m)} \div \pi_m(a|s_m) - 1\right)\cdot \pi_m(a|s_m) \tag{127}$$

$$= \left(\frac{\frac{1}{\pi_m(a_m|s_m)}}{(\frac{2}{\pi_m(a_m|s_m)} - 1)} - 1\right)\cdot \pi_m(a|s_m) \tag{128}$$

$$= -\frac{1 - \pi_m(a_m|s_m)}{2 - \pi_m(a_m|s_m)}\cdot \pi_m(a|s) \tag{129}$$

□

**Lemma 13.** *Under the CAPO update (12) with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s_m, a_m)\}$ and $A^m(s_m,a_m) < 0$, then the policy weight difference $\pi_{m+1}(a|s) - \pi_m(a|s)$ can be written as :*

$$\pi_{m+1}(a|s) - \pi_m(a|s) = \begin{cases} \frac{-\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m)+1} & \text{, if } s = s_m, a = a_m \\ \frac{\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m)+1}\cdot \pi_m(a|s) & \text{, if } s = s_m, a \neq a_m \\ 0 & \text{, else} \end{cases} \tag{130}$$

*Proof of Lemma 13.*

For $s = s_m, a = a_m$ :

$$\pi_{m+1}(a_m|s_m) - \pi_m(a_m|s_m) = \frac{e^{\theta_{m+1}(s_m,a_m)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s_m,a)}} - \pi_m(a_m|s_m) \tag{131}$$

$$= \frac{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m)}}{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m)} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{132}$$

$$= \frac{e^{\theta_m(s_m,a_m)-ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(a_m)}})}}{e^{\theta_m(s_m,a_m)-ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(s_m,a_m)}})} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{133}$$

$$= \frac{e^{\theta_m}(s_m,a_m)\cdot \pi_m(a_m|s_m)}{e^{\theta_m}(s_m,a_m)\cdot \pi_m(a_m|s_m) + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a_m|s_m) \tag{134}$$

$$= \frac{e^{\theta_m}(s_m,a_m)\cdot \pi_m(a_m|s_m)}{e^{\theta_m}(s_m,a_m)\cdot \pi_m(a_m|s_m) + (\frac{1}{\pi_m(a_m|s_m)} - 1)\cdot e^{\theta_m}(s_m,a_m)} - \pi_m(a_m|s_m) \tag{135}$$

$$= \frac{\pi_m(a_m|s_m)}{\pi_m(a_m|s_m) - 1 + \frac{1}{\pi_m(a_m|s_m)}} - \pi_m(a_m|s_m) \tag{136}$$

$$= \frac{-\pi_m(a_m|s_m) \cdot (1 - \pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m) + 1} \tag{137}$$

For $s = s_m, a \neq a_m$ :

$$\pi_{m+1}(a|s_m) - \pi_m(a|s_m) = \frac{e^{\theta_{m+1}(s_m,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s_m,a)}} - \pi_m(a|s_m) \tag{138}$$

$$= \frac{e^{\theta_m(s_m,a)}}{e^{\theta_m(s_m,a_m)+ln(\frac{1}{\pi_m(a_m|s_m)})\cdot sign(A^m(s_m,a_m)} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{139}$$

$$= \frac{e^{\theta_m(s_m,a)}}{e^{\theta_m(s_m,a_m)-ln(\frac{\sum_a e^{\theta_m(a)}}{e^{\theta_m(s_m,a_m)}})} + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{140}$$

$$= \frac{e^{\theta_m(s_m,a)}}{\pi_m(a_m|s_m) \cdot e^{\theta_m}(s_m,a_m) + \sum\limits_{a\neq a_m} e^{\theta_m(s_m,a)}} - \pi_m(a|s_m) \tag{141}$$

$$= \frac{e^{\theta_m(s_m,a)}}{\pi_m(a_m|s_m) \cdot e^{\theta_m}(s_m,a_m) + (\frac{1}{\pi_m(a_m|s_m)} - 1) \cdot e^{\theta_m}(s_m,a_m)} - \pi_m(a|s_m) \tag{142}$$

$$= \left( \frac{e^{\theta_m(s_m,a)}}{(\pi_m(a_m|s_m) - 1 + \frac{1}{\pi_m(a_m|s_m)}) \cdot e^{\theta_m}(s_m,a_m)} \div \pi_m(a|s_m) - 1 \right) \cdot \pi_m(a|s_m) \tag{143}$$

$$= \left( \frac{\frac{1}{\pi_m(a_m|s_m)}}{(\pi_m(a_m|s_m) - 1 + \frac{1}{\pi_m(a_m|s_m)})} - 1 \right) \cdot \pi_m(a|s_m) \tag{144}$$

$$= \frac{\pi_m(a_m|s_m) \cdot (1 - \pi_m(a_m|s_m))}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m) + 1} \cdot \pi_m(a|s) \tag{145}$$

□

**Lemma 14.** *Under the CAPO update (12) with $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s_m, a_m)\}$ and $A^m(s_m,a_m) > 0$, then the policy weight difference $\pi_{m+1}(a|s) - \pi_m(a|s)$ can be written as :*

$$\pi_{m+1}(a|s) - \pi_m(a|s) = \begin{cases} W^+ & \text{, if } s = s_m, a = a_m \\ -W^+ \cdot \frac{\pi_m(a|s)}{1-\pi_m(a_m|s_m)} & \text{, if } s = s_m, a \neq a_m \\ 0 & \text{, else} \end{cases} \tag{146}$$

$$\text{where } (1 - \pi_m(a_m|s_m)) \geq W^+ \geq \frac{(1 - \pi_m(a_m|s_m))^2}{2 - \pi_m(a_m|s_m)} \tag{147}$$

*Proof of Lemma 14.*

By Lemma 12, we have $W^+ = \frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)}$ under $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$. Since $\pi_{m+1}(a|s)$ is proportional to the learning rate $\alpha_m(s,a)$, we establish the lower bound of $W^+$ directly. The upper bound of $W^+$ is constructed by the maximum value of improvement.
Also, for $s = s_m, a \neq a_m$, we have:

$$\frac{\pi_{m+1}(a|s)}{\pi_m(a|s)} = \frac{\frac{e^{\theta_m(s,a)}}{Z_m(s)}}{\frac{e^{\theta_{m+1}(s,a)}}{Z_{m+1}(s)}} = \frac{\frac{e^{\theta_m(s,a)}}{Z_m(s)}}{\frac{e^{\theta_m(s,a)}}{Z_{m+1}(s)}} = \frac{Z_m(s)}{Z_{m+1}(s)} \tag{148}$$

Since $\sum_{a\neq a_m}(\pi_{m+1}(a|s)-\pi_m(a|s)) = -W^+$, we have:

$$\sum_{a\neq a_m}(\pi_{m+1}(a|s)-\pi_m(a|s)) = \sum_{a\neq a_m}\left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot\pi_m(a|s) = \left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot(1-\pi_m(a_m|s)) = -W^+ \tag{149}$$

Hence, for $s = s_m, a\neq a_m$, we get:

$$\pi_{m+1}(a|s)-\pi_m(a|s) = \frac{Z_m(s)\cdot\pi_m(a|s)}{Z_{m+1}(s)}-\pi_m(a|s) = \left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot\pi_m(a|s) = \frac{-W^+}{1-\pi_m(a_m|s)}\cdot\pi_m(a|s) \tag{150}$$

□

**Lemma 15.** *Under the CAPO update (12) with $\alpha_m(s,a)\geq\log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s_m,a_m)\}$ and $A^m(s_m,a_m) < 0$, then the policy weight difference $\pi_{m+1}(a|s)-\pi_m(a|s)$ can be written as :*

$$\pi_{m+1}(a|s)-\pi_m(a|s) = \begin{cases}-W^- & \text{, if } s=s_m, a=a_m\\ W^-\cdot\frac{\pi_m(a|s)}{1-\pi_m(a_m|s_m)} & \text{, if } s=s_m, a\neq a_m\\ 0 & \text{, else}\end{cases}$$

$$\text{where } \pi_m(a_m|s_m)\geq W^-\geq\frac{\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2-\pi_m(a_m|s_m)+1}$$

*Proof of Lemma 15.*

By Lemma 13, we have $W^- = \frac{\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2-\pi_m(a_m|s_m)+1}$ under $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$. Since $\pi_{m+1}(a|s)$ is proportional to the learning rate $\alpha_m(s,a)$, we establish the lower bound of $W^-$ directly. The upper bound of $W^-$ is constructed by the maximum value of improvement.
Also, for $s = s_m, a\neq a_m$, we have:

$$\frac{\pi_{m+1}(a|s)}{\pi_m(a|s)} = \frac{\frac{e^{\theta_m(s,a)}}{Z_m(s)}}{\frac{e^{\theta_{m+1}(s,a)}}{Z_{m+1}(s)}} = \frac{\frac{e^{\theta_m(s,a)}}{Z_m(s)}}{\frac{e^{\theta_m(s,a)}}{Z_{m+1}(s)}} = \frac{Z_m(s)}{Z_{m+1}(s)} \tag{151}$$

Moreover, since $\sum_{a\neq a_m}(\pi_{m+1}(a|s)-\pi_m(a|s)) = W^-$, we have:

$$\sum_{a\neq a_m}(\pi_{m+1}(a|s)-\pi_m(a|s)) = \sum_{a\neq a_m}\left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot\pi_m(a|s) = \left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot(1-\pi_m(a_m|s)) = W^- \tag{152}$$

Hence, for $s = s_m, a\neq a_m$, we get:

$$\pi_{m+1}(a|s)-\pi_m(a|s) = \frac{Z_m(s)\cdot\pi_m(a|s)}{Z_{m+1}(s)}-\pi_m(a|s) = \left(\frac{Z_m(s)}{Z_{m+1}(s)}-1\right)\cdot\pi_m(a|s) = \frac{W^-}{1-\pi_m(a_m|s)}\cdot\pi_m(a|s) \tag{153}$$

□

**Lemma 16.** *Under the CAPO update (12) with $\alpha_m(s,a)\geq\log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s_m,a_m)\}$ then the improvement of the performance $V^{\pi_{m+1}}(s)-V^{\pi_m}(s)$ can be written as :*

$$V^{\pi_{m+1}}(s)-V^{\pi_m}(s) = \begin{cases}\frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma}\cdot\frac{W^+}{1-\pi_m(a_m|s_m)}\cdot A^m(s_m,a_m) & \text{, if } A^m(s_m,a_m)>0\\ \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma}\cdot\frac{W^-}{1-\pi_m(a_m|s_m)}\cdot(-A^m(s_m,a_m)) & \text{, if } A^m(s_m,a_m)<0\end{cases} \tag{154}$$

$$\text{where }\begin{cases}(1-\pi_m(a_m|s_m))\geq W^+\geq\frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)}\\ \pi_m(a_m|s_m)\geq W^-\geq\frac{\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2-\pi_m(a_m|s_m)+1}\end{cases} \tag{155}$$

*and it can also be lower bounded by :*

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(s_m, a_m)^2 & \text{, if } A^m(s_m, a_m) > 0 \\ d_s^{\pi_{m+1}}(s_m) \cdot \pi_m(a_m|s_m) \cdot A^m(s_m, a_m)^2 & \text{, if } A^m(s_m, a_m) < 0 \end{cases} \tag{156}$$

*Proof of Lemma 16.*

If $A^m(s, a) > 0$, then :

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_s^{\pi_{m+1}}(s) \sum_{a \in \mathcal{A}} \pi_{m+1}(a|s) \cdot A^m(s, a) \tag{157}$$

$$= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_s^{\pi_{m+1}}(s) \sum_{a \in \mathcal{A}} (\pi_{m+1}(a|s) - \pi_m(a|s)) \cdot A^m(s, a) \tag{158}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \sum_{a \in \mathcal{A}} (\pi_{m+1}(a|s_m) - \pi_m(a|s_m)) \cdot A^m(s_m, a) \tag{159}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ W^+ \cdot A^m(s_m, a_m) - \sum_{a \neq a_m} \frac{W^+}{1 - \pi_m(a_m|s_m)} \cdot \pi_m(a|s_m) \cdot A^m(s_m, a) \right] \tag{160}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ W^+ \cdot A^m(s_m, a_m) - \frac{W^+}{1 - \pi_m(a_m|s_m)} \cdot \sum_{a \neq a_m} \pi_m(a|s_m) \cdot A^m(s_m, a) \right] \tag{161}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ W^+ \cdot A^m(s_m, a_m) + \frac{W^+}{1 - \pi_m(a_m|s_m)} \cdot \pi_m(a_m|s_m) \cdot A^m(s_m, a_m) \right] \tag{162}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^+}{1 - \pi_m(a_m|s_m)} \cdot A^m(s_m, a_m) \tag{163}$$

$$\geq \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(s_m, a_m)^2 \tag{164}$$

The first equation holds by the performance difference lemma in Lemma 3.
The second equation holds by the definition of $A(s, a)$.
The third equation holds since $\pi_{m+1}(a|s) = \pi_m(a|s), \quad \forall s \neq s_m$.
The fourth equation holds by the difference of the updated policy weight that we have shown in Lemma 12 and Lemma 14.
The last inequality holds by the bound of $A(s, a)$ in Lemma 7.

If $A^m(s, a) < 0$, then :

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_s^{\pi_{m+1}}(s) \sum_{a \in \mathcal{A}} \pi_{m+1}(a|s) \cdot A^m(s, a) \tag{165}$$

$$= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_s^{\pi_{m+1}}(s) \sum_{a \in \mathcal{A}} (\pi_{m+1}(a|s) - \pi_m(a|s)) \cdot A^m(s, a) \tag{166}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \sum_{a \in \mathcal{A}} (\pi_{m+1}(a|s_m) - \pi_m(a|s_m)) \cdot A^m(s_m, a) \tag{167}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ -W^- \cdot A^m(s_m, a_m) + \sum_{a \neq a_m} \frac{W^-}{1 - \pi_m(a_m|s_m)} \cdot \pi_m(a|s_m) \cdot A^m(s_m, a) \right] \tag{168}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ -W^- \cdot A^m(s_m, a_m) + \frac{W^-}{1 - \pi_m(a_m|s_m)} \cdot \sum_{a \neq a_m} \pi_m(a|s_m) \cdot A^m(s_m, a) \right] \tag{169}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \left[ -W^- \cdot A^m(s_m, a_m) - \frac{W^-}{1-\pi_m(a_m|s_m)} \cdot \pi_m(a_m|s_m) \cdot A^m(s_m, a_m) \right] \tag{170}$$

$$= \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^-}{1-\pi_m(a_m|s_m)} \cdot (-A^m(s_m, a_m)) \tag{171}$$

$$\geq d_s^{\pi_{m+1}}(s_m) \cdot \pi_m(a_m|s_m) \cdot A^m(s_m, a_m)^2 \tag{172}$$

The first equation holds by the performance difference lemma in Lemma 3.
The second equation holds by the definition of $A(s, a)$.
The third equation holds since $\pi_{m+1}(a|s) = \pi_m(a|s), \quad \forall s \neq s_m$.
The fourth equation holds by the difference of the updated policy weight that we have shown in Lemma 13 and Lemma 15.
The last inequality holds by the bound of $A(s, a)$ in Lemma 7.

□

### B.1 Convergence Rate of Cyclic CAPO

For ease of exposition, we restate Theorem 2 as follows.

**Theorem.** *Consider a tabular softmax parameterized policy $\pi_\theta$. Under Cyclic CAPO with $\alpha_m(s, a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$ and $|B_m| = 1$, $\bigcup_{i=1}^{|\mathcal{S}||\mathcal{A}|} B_{m \cdot |\mathcal{S}||\mathcal{A}|+i} = \mathcal{S} \times \mathcal{A}$, we have :*

$$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{|\mathcal{S}||\mathcal{A}|}{c} \cdot \frac{1}{m}, \quad \textit{for all } m \geq 1 \tag{173}$$

$$\sum_{m=1}^{M} V^*(\rho) - V^{\pi_m}(\rho) \leq |\mathcal{S}||\mathcal{A}| \cdot \min \left\{ \sqrt{\frac{M}{c \cdot (1-\gamma)}}, \frac{\log M + 1}{c} \right\}, \quad \textit{for all } m \geq 1 \tag{174}$$

*where* $c = \frac{(1-\gamma)^4}{2} \cdot \left\| \frac{1}{\mu} \right\|_\infty^{-1} \cdot \min \left\{ \frac{\min_s \mu(s)}{2}, \frac{(1-\gamma)}{|\mathcal{S}||\mathcal{A}|} \right\} > 0.$

*Proof of Theorem 2.*

The proof can be summarized as:

1. We first write the improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in state visitation distribution, policy weight, and advantage value in Lemma 16, and also construct the lower bound of it.
2. We then construct the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$ using $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s)$.
3. Finally, we can show the desired result inductively by Lemma 11.

By Lemma 16, we have for all $m \geq 1$:

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^+}{1-\pi_m(a_m|s_m)} \cdot A^m(s_m, a_m) & , \text{if } A^m(s_m, a_m) > 0 \\ \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^-}{1-\pi_m(a_m|s_m)} \cdot (-A^m(s_m, a_m)) & , \text{if } A^m(s_m, a_m) < 0 \end{cases} \tag{175}$$

$$\text{where } \begin{cases} (1-\pi_m(a_m|s_m)) \geq W^+ \geq \frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)} \\ \pi_m(a_m|s_m) \geq W^- \geq \frac{\pi_m(a_m|s_m) \cdot (1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m) + 1} \end{cases} \tag{176}$$

and it can also be lower bounded by:

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(s_m, a_m)^2 & , \text{if } A^m(s_m, a_m) > 0 \\ d_s^{\pi_{m+1}}(s_m) \cdot \pi_m(a_m|s_m) \cdot A^m(s_m, a_m)^2 & , \text{if } A^m(s_m, a_m) < 0 \end{cases} \tag{177}$$

Now, we're going to construct the upper bound of the performance difference $V^*(s)-V^{\pi_m}(s)$ using $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s)-V^{\pi_m}(s)$. Note that by Lemma 8, there exists $(\tilde{s_m}, \tilde{a_m})$ such that $(V^*(s) - V^{\pi_m}(s))^2 \leq \left(\frac{1}{1-\gamma} \cdot A^m(\tilde{s_m}, \tilde{a_m})\right)^2$ for all $m \geq 1$.

Hence, if we construct the upper bound of $A^m(\tilde{s_m}, \tilde{a_m})^2$ using $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s)$, which is the improvement of the performance during the whole cycle, then we can get the the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$ using $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s)$ for all $m \equiv 0 \pmod{|\mathcal{S}||\mathcal{A}|}$.

Without loss of generality, Assume we update $(\tilde{s_m}, \tilde{a_m})$ at episode $(m + T)$, where $T \in [0, |\mathcal{S}||\mathcal{A}|) \bigcap \mathbb{N}$, $m \equiv 0 \pmod{|\mathcal{S}||\mathcal{A}|}$. We discuss two possible cases as follows:

- Case 1: $V^{\pi_{m+T}}(s) - V^{\pi_m}(s) \geq A^m(\tilde{s_m}, \tilde{a_m})$:

$$A^m(\tilde{s_m}, \tilde{a_m})^2 \leq \left(V^{\pi_{m+T}}(s) - V^{\pi_m}(s)\right)^2 \tag{178}$$

$$= \left(\sum_{k=m}^{m+T-1} (V^{\pi_{k+1}}(s) - V^{\pi_k}(s))\right)^2 \tag{179}$$

$$= \left(\sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} (V^{\pi_{k+1}}(s) - V^{\pi_k}(s)) + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} (V^{\pi_{k+1}}(s) - V^{\pi_k}(s))\right)^2 \tag{180}$$

$$= \left(\sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} \frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k+}}{1-\pi_k(a_k|s_k)} \cdot A^k(s_k, a_k)\right. \tag{181}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} \frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k-}}{1-\pi_k(a_k|s_k)} \cdot (-A^k(s_k, a_k))\right)^2 \tag{182}$$

$$\leq T \cdot \left(\sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} \left(\frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k+}}{1-\pi_k(a_k|s_k)} \cdot A^k(s_k, a_k)\right)^2\right. \tag{183}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} \left(\frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k-}}{1-\pi_k(a_k|s_k)} \cdot A^k(s_k, a_k)\right)^2\right) \tag{184}$$

$$= T \cdot \left(\sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} \left(\frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k+}}{1-\pi_k(a_k|s_k)}\right)^2 \cdot A^k(s_k, a_k)^2\right. \tag{185}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} \left(\frac{d_s^{\pi_{k+1}}(s_k)}{1-\gamma} \cdot \frac{W^{k-}}{1-\pi_k(a_k|s_k)}\right)^2 \cdot A^k(s_k, a_k)^2\right) \tag{186}$$

$$= T \cdot \left( \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} \left( \frac{d_s^{\pi_{k+1}}(s_k) \cdot W^{k+}}{1-\gamma} \right)^2 \cdot \frac{|A^k(s_k,a_k)|}{1-\pi_k(a_k|s_k)} \cdot \frac{|A^k(s_k,a_k)|}{1-\pi_k(a_k|s_k)} \right. \tag{187}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} \left( \frac{d_s^{\pi_{k+1}}(s_k) \cdot W^{k-}}{1-\gamma} \right)^2 \cdot \frac{|A^k(s_k,a_k)|}{1-\pi_k(a_k|s_k)} \cdot \frac{|A^k(s_k,a_k)|}{1-\pi_k(a_k|s_k)} \right) \tag{188}$$

$$\leq T \cdot \left( \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} \left( \frac{d_s^{\pi_{k+1}}(s_k) \cdot W^{k+}}{1-\gamma} \right)^2 \cdot \frac{1}{1-\gamma} \cdot \frac{1-\gamma}{d_s^{\pi_{k+1}}(s_k) \cdot W^{k+}} \cdot (V^{\pi_{k+1}}(s) - V^{\pi_k}(s)) \right. \tag{189}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} \left( \frac{d_s^{\pi_{k+1}}(s_k) \cdot W^{k-}}{1-\gamma} \right)^2 \cdot \frac{1}{1-\gamma} \cdot \frac{1-\gamma}{d_s^{\pi_{k+1}}(s_k) \cdot W^{k-}} \cdot (V^{\pi_{k+1}}(s) - V^{\pi_k}(s)) \right) \tag{190}$$

$$= \frac{T}{(1-\gamma)^2} \cdot \left( \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)>0}} d_s^{\pi_{k+1}}(s_k) \cdot W^{k+} \cdot (V^{\pi_{k+1}}(s) - V^{\pi_k}(s)) \right. \tag{191}$$

$$\left. + \sum_{\substack{k\in[m,m+T-1]\\ A^k(s_k,a_k)<0}} d_s^{\pi_{k+1}}(s_k) \cdot W^{k-} \cdot (V^{\pi_{k+1}}(s) - V^{\pi_k}(s)) \right) \tag{192}$$

$$\leq \frac{T}{(1-\gamma)^2} \cdot \left( \sum_{k=m}^{m+T-1} V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \right) \tag{193}$$

$$\cdot \max_{k\in[m,m+T-1]} \left\{ \mathbb{I}\left\{ A^k(s_k,a_k) > 0 \right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot W^{k+} \right. \tag{194}$$

$$\left. + \mathbb{I}\left\{ A^k(s_k,a_k) < 0 \right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot W^{k-} \right\} \tag{195}$$

$$\leq c_m \cdot \frac{T}{(1-\gamma)^2} \cdot (V^{\pi_{m+T}}(s) - V^{\pi_m}(s)) \tag{196}$$

$$\leq c_m \cdot \frac{T}{(1-\gamma)^2} \cdot (V^{\pi_{m+T+1}}(s) - V^{\pi_m}(s)) \tag{197}$$

$$\leq 2 \cdot \max \left\{ \frac{2}{d_s^{\pi_{m+T+1}}(s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2} \right\} \cdot (V^{\pi_{m+T+1}}(s) - V^{\pi_m}(s)) \tag{198}$$

where $c_m = \max_{k\in[m,m+T-1]} \{c_{k1}, c_{k2}\} \in [0,1]$
and $c_{k1} = \mathbb{I}\left\{ A^k(s_k,a_k) > 0 \right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot W^{k+}$, $c_{k2} = \mathbb{I}\left\{ A^k(s_k,a_k) < 0 \right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot W^{k-}$.

The third equation holds by Lemma 16.
The second inequality holds by Cauchy-Schwarz.
The third inequality holds by Lemma 7 and Lemma 16.

- Case 2: $V^{\pi_{m+T}}(s) - V^{\pi_m}(s) < A^m(\tilde{s_m}, \tilde{a_m})$:

$$
\begin{aligned}
A^m(\tilde{s_m}, \tilde{a_m})^2 &= ((Q^{\pi_m}(\tilde{s_m}, \tilde{a_m}) - V^{\pi_{m+T}}(s)) + (V^{\pi_{m+T}}(s) - V^{\pi_m}(s)))^2 && (199)\\
&\leq ((Q^{\pi_{m+T}}(\tilde{s_m}, \tilde{a_m}) - V^{\pi_{m+T}}(s)) + (V^{\pi_{m+T}}(s) - V^{\pi_m}(s)))^2 && (200)\\
&= \left(A^{m+T}(\tilde{s_m}, \tilde{a_m}) + (V^{\pi_{m+T}}(s) - V^{\pi_m}(s))\right)^2 && (201)\\
&\leq \left(A^{m+T}(\tilde{s_m}, \tilde{a_m})^2 + (V^{\pi_{m+T}}(s) - V^{\pi_m}(s))^2\right) \cdot (1^2 + 1^2) && (202)\\
&\leq 2 \cdot \left( \frac{2}{d_s^{\pi_{m+T+1}}(s_{m+T})} \cdot (V^{\pi_{m+T+1}}(s) - V^{\pi_{m+T}}(s)) \right. && (203)\\
&\qquad \left. + c_m \cdot \frac{T}{(1-\gamma)^2} \cdot (V^{\pi_{m+T}}(s) - V^{\pi_m}(s)) \right) && (204)\\
&\leq 2 \cdot \max\left\{ \frac{2}{d_s^{\pi_{m+T+1}}(s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2} \right\} \cdot (V^{\pi_{m+T+1}}(s) - V^{\pi_m}(s)) && (205)
\end{aligned}
$$

The first inequality holds by the strict improvement of $V^\pi(s)$ A.1, leading to the strict improvement of $Q^\pi(s, a)$.
The second inequality holds by Cauchy-Schwarz.
The third inequality holds by the result of Case 1 and Lemma 16

Hence, in both case we get:

$$
V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s) \geq V^{\pi_{m+T+1}}(s) - V^{\pi_m}(s) \geq \frac{1}{2} \cdot \frac{1}{\max\left\{ \frac{2}{d_s^{\pi_{m+T+1}}(s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2} \right\}} \cdot A^m(\tilde{s_m}, \tilde{a_m})^2 \quad (206)
$$

for all $m \equiv 0 \pmod{|\mathcal{S}||\mathcal{A}|}$.

Combining Lemma 8, we can construct the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$ using $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s)$:

$$
\begin{aligned}
V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s) &\geq \frac{(1-\gamma)^2}{2} \cdot \frac{1}{\max\left\{ \frac{2}{d_s^{\pi_{m+T+1}}(s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2} \right\}} \cdot (V^*(s) - V^{\pi_m}(s))^2 && (207)\\
&= \frac{(1-\gamma)^2}{2} \cdot \min\left\{ \frac{d_s^{\pi_{m+T+1}}(s_{m+T})}{2}, \frac{(1-\gamma)^2}{c_m \cdot T} \right\} \cdot (V^*(s) - V^{\pi_m}(s))^2 && (208)
\end{aligned}
$$

and if we consider the whole initial state distribution, $\mu$, we have :

$$
\begin{aligned}
V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(\mu) - V^{\pi_m}(\mu) &\geq \frac{(1-\gamma)^2}{2} \cdot \frac{1}{\max\left\{ \frac{2}{d_\mu^{\pi_{m+T+1}}(s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2} \right\}} \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 && (209)\\
&= \frac{(1-\gamma)^2}{2} \cdot \min\left\{ \frac{d_\mu^{\pi_{m+T+1}}(s_{m+T})}{2}, \frac{(1-\gamma)^2}{c_m \cdot T} \right\} \cdot (V^*(s) - V^{\pi_m}(s))^2 && (210)\\
&\geq \frac{(1-\gamma)^2}{2} \cdot \min\left\{ \frac{(1-\gamma) \cdot \min_s \mu(s)}{2}, \frac{(1-\gamma)^2}{|\mathcal{S}||\mathcal{A}|} \right\} \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 && (211)\\
&\geq \underbrace{\frac{(1-\gamma)^3}{2} \cdot \min\left\{ \frac{\min_s \mu(s)}{2}, \frac{(1-\gamma)}{|\mathcal{S}||\mathcal{A}|} \right\}}_{:=c'>0} \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 && (212)
\end{aligned}
$$

The second inequality holds since $d_\mu^\pi(s) \geq (1-\gamma) \cdot \mu(s)$ 10.

And since $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(\mu) - V^{\pi_m}(\mu) = (V^{\pi^*}(\mu) - V^{\pi_m}(\mu)) - (V^{\pi^*}(\mu) - V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(\mu))$, by rearranging the inequality above, we have :

$$
\delta_{m+|\mathcal{S}||\mathcal{A}|} \leq \delta_m - c' \cdot \delta_m^2 \quad \text{where } \delta_m = V^{\pi^*}(\mu) - V^{\pi_m}(\mu) \text{ for all } m \equiv 0 \pmod{|\mathcal{S}||\mathcal{A}|} \quad (213)
$$

Then, we can get the following result by induction 11 :

$$V^*(\mu) - V^{\pi_m}(\mu) \le \frac{1}{c'} \cdot \frac{1}{\max\left\{\left\lfloor \frac{m}{|\mathcal{S}||\mathcal{A}|} \right\rfloor, 1\right\}} \le \frac{1}{c'} \cdot \min\left\{\frac{|\mathcal{S}||\mathcal{A}|}{m}, 1\right\} \le \frac{|\mathcal{S}||\mathcal{A}|}{c'} \cdot \frac{1}{m}, \quad \text{for all } m \ge 1 \tag{214}$$

$$\sum_{m=1}^{M} V^*(\mu) - V^{\pi_m}(\mu) \le |\mathcal{S}||\mathcal{A}| \cdot \min\left\{\sqrt{\frac{M}{c' \cdot (1-\gamma)}}, \frac{\log M + 1}{c'}\right\}, \quad \text{for all } m \ge 1 \tag{215}$$

where $c' = \frac{(1-\gamma)^3}{2} \cdot \min\left\{\frac{\min_s \mu(s)}{2}, \frac{(1-\gamma)}{|\mathcal{S}||\mathcal{A}|}\right\} > 0$.

Finally, we get the desired result by Lemma 9:

$$V^*(\rho) - V^{\pi_m}(\rho) \le \frac{1}{1-\gamma} \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot (V^*(\mu) - V^{\pi_m}(\mu)) \le \frac{|\mathcal{S}||\mathcal{A}|}{c} \cdot \frac{1}{m}, \quad \text{for all } m \ge 1 \tag{216}$$

$$\sum_{m=1}^{M} V^*(\rho) - V^{\pi_m}(\rho) \le |\mathcal{S}||\mathcal{A}| \cdot \min\left\{\sqrt{\frac{M}{c \cdot (1-\gamma)}}, \frac{\log M + 1}{c}\right\}, \quad \text{for all } m \ge 1 \tag{217}$$

where $c = \frac{(1-\gamma)^4}{2} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \min\left\{\frac{\min_s \mu(s)}{2}, \frac{(1-\gamma)}{|\mathcal{S}||\mathcal{A}|}\right\} > 0$. □

### B.2 Convergence Rate of Batch CAPO

For ease of exposition, we restate Theorem 3 as follows.

**Theorem.** *Consider a tabular softmax parameterized policy $\pi_\theta$. Under Batch CAPO with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$ and $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$, we have :*

$$V^*(\rho) - V^{\pi_m}(\rho) \le \frac{1}{c} \cdot \frac{1}{m}, \quad \textit{for all } m \ge 1 \tag{218}$$

$$\sum_{m=1}^{M} V^*(\rho) - V^{\pi_m}(\rho) \le \min\left\{\sqrt{\frac{M}{c \cdot (1-\gamma)}}, \frac{\log M + 1}{c}\right\}, \quad \textit{for all } m \ge 1 \tag{219}$$

$$\tag{220}$$

*where $c = \frac{(1-\gamma)^4}{|\mathcal{A}|} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \min_s \{\mu(s)\} > 0$.*

*Proof of Theorem 3.*

The proof can be summarized as follows:

1. We first construct the lower bound of the improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in state visitation distribution, number of actions, and advantage value in Lemma 16.
2. We then construct the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$ using $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$.
3. Finally, we can show the desired result inductively 11.

**Lemma 17.** *Under (12) with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$, then the updated policy weight $\pi_{m+1}(a|s)$ can be written as :*

$$\pi_{m+1}(a|s) = \begin{cases} \frac{1}{|s_m^+| + \sum\limits_{A^m(s,a)=0} \pi_m(a|s) + \sum\limits_{A^m(s,a)<0} \pi_m(a|s)^2} & \textit{, if } A^m(s,a) > 0 \\ \frac{\pi_m(a)}{|s_m^+| + \sum\limits_{A^m(s,a)=0} \pi_m(a|s) + \sum\limits_{A^m(s,a)<0} \pi_m(a|s)^2} & \textit{, if } A^m(s,a) = 0 \\ \frac{\pi_m(a)^2}{|s_m^+| + \sum\limits_{A^m(s,a)=0} \pi_m(a|s) + \sum\limits_{A^m(s,a)<0} \pi_m(a|s)^2} & \textit{, if } A^m(s,a) < 0 \end{cases}$$

*where $s_m^+ := \{a \in \mathcal{S} \mid A^m(s,a) > 0\}$*

*Proof of Lemma 17.*

For $A^m(s,a) > 0:$

$$\pi_{m+1}(a|s) = \frac{e^{\theta_m(s,a)+\ln\frac{1}{\pi_m(a|s)}}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \frac{e^{\theta_m(s,a)+\ln\frac{\sum_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{e^{\theta_m(s,a)}}}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} \tag{221}$$

For $A^m(s,a) = 0:$

$$\pi_{m+1}(a|s) = \frac{e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \frac{e^{\theta_m(s,a)} \cdot \sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)} \sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \pi_m(a|s) \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} \tag{222}$$

For $A^m(s,a) < 0:$

$$\pi_{m+1}(a|s) = \frac{e^{\theta_m(s,a)-\ln\frac{1}{\pi_m(a|s)}}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \frac{e^{2\cdot\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)} \sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \pi_m(a|s)^2 \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} \tag{223}$$

Moreover, since $\sum\limits_{a\in\mathcal{A}} \pi_{m+1}(a|s) = 1$, we have:

$$\sum_{A^m(s,a)>0} \pi_{m+1}(a|s) + \sum_{A^m(s,a)=0} \pi_{m+1}(a|s) + \sum_{A^m(s,a)<0} \pi_{m+1}(a|s) \tag{224}$$

$$= |s_m^+| \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} + \sum_{A^m(s,a)=0} \pi_m(a|s) \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} + \sum_{A^m(s,a)<0} \pi_m(a|s)^2 \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} \tag{225}$$

$$= \left( |s_m^+| + \sum_{A^m(s,a)=0} \pi_m(a|s) + \sum_{A^m(s,a)<0} \pi_m(a|s)^2 \right) \cdot \frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = 1, \tag{226}$$

where $s_m^+ := \{a \in \mathcal{S} \mid A^m(s,a) > 0\}$

Hence, we get:

$$\frac{\sum\limits_{a\in\mathcal{A}} e^{\theta_m(s,a)}}{\sum\limits_{a\in\mathcal{A}} e^{\theta_{m+1}(s,a)}} = \frac{1}{|s_m^+| + \sum\limits_{A^m(s,a)=0} \pi_m(a|s) + \sum\limits_{A^m(s,a)<0} \pi_m(a|s)^2}. \tag{227}$$

Finally, we get the desired result by substitution.

□

**Lemma 18.** *Under (12) with $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$, if $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ then the improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ can be bounded by :*

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \frac{1}{|\mathcal{A}|} \cdot \sum_{s'\in\mathcal{S}} d_s^{\pi_m}(s') \sum_{a\in s_m'^+} A^m(s',a)^2 \tag{228}$$

*where $s_m'^+ := \{a \in \mathcal{S} \mid A^m(s',a) > 0\}$*

*Proof of Lemma 18.*

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \sum_{a \in \mathcal{A}} \pi_{m+1}(a|s') \cdot A^m(s', a) \tag{229}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \frac{1}{|s_m^+| + \sum_{A^m(s,a)=0} \pi_m(a|s) + \sum_{A^m(s,a)<0} \pi_m(a|s)^2} \tag{230}$$

$$\cdot \left( \sum_{a \in s_m'^+} A^m(s', a) + \sum_{a \notin s_m'^+} \pi_m(a|s')^2 \cdot A^m(s', a) \right) \tag{231}$$

$$\geq \frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \frac{1}{|s_m^+| + \sum_{A^m(s,a)=0} \pi_m(a|s) + \sum_{A^m(s,a)<0} \pi_m(a|s)^2} \tag{232}$$

$$\cdot \left( \sum_{a \in s_m'^+} A^m(s', a) + \sum_{a \notin s_m'^+} \pi_m(a|s') \cdot A^m(s', a) \right) \tag{233}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \frac{1}{|s_m^+| + \sum_{A^m(s,a)=0} \pi_m(a|s) + \sum_{A^m(s,a)<0} \pi_m(a|s)^2} \tag{234}$$

$$\cdot \left( \sum_{a \in s_m'^+} (1 - \pi_m(a|s')) \cdot A^m(s', a) \right) \tag{235}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{|\mathcal{A}|} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \left( \sum_{a \in s_m'^+} (1 - \pi_m(a|s')) \cdot A^m(s', a) \right) \tag{236}$$

$$\geq \frac{1}{|\mathcal{A}|} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \sum_{a \in s_m'^+} A^m(s', a)^2 \tag{237}$$

The first equation holds by the performance difference lemma in Lemma 3.
The second equation holds by Lemma 13.
The third equation holds by the definition of $A(s, a)$.
The last inequality holds by the bound of $A(s, a)$ in Lemma 7.

□

Hence, combining Lemma 18 and Lemma 8, we can construct the upper bound of the performance difference $V^*(s) - V^{\pi_m}(s)$ using $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ :

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \frac{1}{|\mathcal{A}|} \cdot \sum_{s' \in \mathcal{S}} d_s^{\pi_{m+1}}(s') \cdot \sum_{a \in s_m'^+} A^m(s', a)^2 \tag{238}$$

$$\geq \frac{1}{|\mathcal{A}|} \cdot d_s^{\pi_{m+1}}(\tilde{s_m}) \cdot A^m(\tilde{s_m}, \tilde{a_m})^2 \tag{239}$$

$$= \frac{1}{|\mathcal{A}|} \cdot d_s^{\pi_{m+1}}(\tilde{s_m}) \cdot (1-\gamma)^2 \cdot (\frac{1}{1-\gamma})^2 \cdot A^m(\tilde{s_m}, \tilde{a_m})^2 \tag{240}$$

$$\geq \frac{1}{|\mathcal{A}|} \cdot d_s^{\pi_{m+1}}(\tilde{s_m}) \cdot (1-\gamma)^2 \cdot (V^*(s) - V^{\pi_m}(s))^2 \tag{241}$$

Moreover, if we consider the whole starting state distribution $\mu$, we have :

$$V^{\pi_{m+1}}(\mu) - V^{\pi_m}(\mu) \geq \frac{1}{|\mathcal{A}|} \cdot d_{\mu}^{\pi_{m+1}}(\tilde{s_m}) \cdot (1-\gamma)^2 \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 \tag{242}$$

$$\geq \frac{1}{|\mathcal{A}|} \cdot \mu(\tilde{s_m}) \cdot (1-\gamma)^3 \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 \tag{243}$$

$$\geq \underbrace{\frac{(1-\gamma)^3}{|\mathcal{A}|} \cdot \min_{s' \in \mathcal{S}} \{\mu(s')\}}_{:=c'>0} \cdot (V^*(\mu) - V^{\pi_m}(\mu))^2 \tag{244}$$

The second inequality holds since $d_{\mu}^{\pi}(s) \geq (1-\gamma) \cdot \mu(s)$ in Lemma 10.

Since $V^{\pi_{m+1}}(\mu) - V^{\pi_m}(\mu) = (V^{\pi^*}(\mu) - V^{\pi_m}(\mu)) - (V^{\pi^*}(\mu) - V^{\pi_{m+1}}(\mu))$, by rearranging the inequality above, we have :

$$\delta_{m+1} \leq \delta_m - c' \cdot \delta_m^2 \quad \text{where } \delta_m = V^{\pi^*}(\mu) - V^{\pi_m}(\mu) \tag{245}$$

Then, we can get the following result by induction based on Lemma 11 :

$$V^*(\mu) - V^{\pi_m}(\mu) \leq \frac{1}{c'} \cdot \frac{1}{m}, \quad \text{for all } m \geq 1 \tag{246}$$

$$\sum_{m=1}^{M} V^*(\mu) - V^{\pi_m}(\mu) \leq \min\left\{\sqrt{\frac{M}{c \cdot (1-\gamma)}}, \frac{\log M + 1}{c'}\right\}, \quad \text{for all } m \geq 1 \tag{247}$$

$$\tag{248}$$

where $c' = \frac{(1-\gamma)^3}{|\mathcal{A}|} \cdot \underset{s}{min}\{\mu(s)\} > 0$.

Finally, we get the desired result by Lemma 9:

$$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot (V^*(\mu) - V^{\pi_m}(\mu)) \leq \frac{1}{c} \cdot \frac{1}{m}, \quad \text{for all } m \geq 1 \tag{249}$$

$$\sum_{m=1}^{M} V^*(\rho) - V^{\pi_m}(\rho) \leq \min\left\{\sqrt{\frac{M}{c \cdot (1-\gamma)}}, \frac{\log M + 1}{c}\right\}, \quad \text{for all } m \geq 1 \tag{250}$$

where $c = \frac{(1-\gamma)^4}{|\mathcal{A}|} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \underset{s}{min}\{\mu(s)\} > 0$.

**Remark 6.** In Theorem 3, we choose the learning rate $\alpha_m(s,a)$ to be exactly $\log(\frac{1}{\pi_{\theta_m}(a|s)})$ instead of greater than or equal to $\log(\frac{1}{\pi_{\theta_m}(a|s)})$. The reason is that under $\alpha_m(s,a) = \log(\frac{1}{\pi_{\theta_m}(a|s)})$, we can guarantee that all state-action pair with positive advantage value can get the same amount of the policy weight with each other actions in the same state after every update 17. This property directly leads to the result of Lemma 18 that the one-step improvement $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ can be quantified using the summation of all positive advantage value $\sum_{a \in s_m'^+} A^m(s',a)^2$, and hence it guarantees that one of the $A^m(s',a)^2$ will connect the one-step improvement with the performance difference $V^{\pi^*}(s) - V^{\pi_m}(s)$. This property also prevents some extreme cases where one of the learning rates of the state-action pairs with extremely tiny but positive advantage value dominates the updated policy weight, i.e., $\pi_{m+1}(a_m|s_m) \to 1$, leading to tiny one-step improvement.

□

### B.3 Convergence Rate of Randomized CAPO

For ease of exposition, we restate Theorem 4 as follows.

**Theorem.** *Consider a tabular softmax parameterized policy $\pi_\theta$, under (12) with $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$ and $|B_m| = 1$, if Condition 1 is satisfied, then we have :*

$$\underset{(s_m,a_m)\sim d_{gen}}{\mathbb{E}}[V^*(\rho) - V^{\pi_m}(\rho)] \leq \frac{1}{c} \cdot \frac{1}{m}, \quad \textit{for all } m \geq 1 \tag{251}$$

$$\sum_{m=1}^{M} \underset{(s_m,a_m)\sim d_{gen}}{\mathbb{E}}[V^*(\rho) - V^{\pi_m}(\rho)] \leq \min\left\{\sqrt{\frac{M}{c\cdot(1-\gamma)}}, \frac{\log M + 1}{c}\right\}, \quad \textit{for all } m \geq 1 \tag{252}$$

$$\tag{253}$$

*where* $c = \frac{(1-\gamma)^4}{2} \cdot \left\|\frac{1}{\mu}\right\|_\infty^{-1} \cdot \underset{(s,a)}{min}\{d_{gen}(s,a)\cdot\mu(s)\} > 0$ *and* $d_{gen} : \mathcal{S}\times\mathcal{A} \to (0,1)$, $d_{gen}(s,a) = \mathbb{P}((s,a)\in B_m)$.

*Proof of Theorem 4.*

The proof can be summarized as:

1. We first write the improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in state visitation distribution, policy weight, and advantage value in Lemma 16, and also construct the lower bound of it. Note that the result is the same as Appendix B.1.
2. We then write the improvement of the performance $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in probability form condition on $(s_m, a_m)$.
3. By taking expectation of the probability form, we get the upper bound of the expected performance difference $\underset{(s_m,a_m)\sim d_{gen}}{\mathbb{E}}[V^*(\mu) - V^{\pi_m}(\mu)]$ using $\underset{(s_m,a_m)\sim d_{gen}}{\mathbb{E}}[V^{\pi_{m+1}}(\mu) - V^{\pi_m}(\mu)]$.
4. Finally, we can show the desired result by induction based on Lemma 11.

By Lemma 16, we have for all $m \geq 1$:

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^+}{1-\pi_m(a_m|s_m)} \cdot A^m(s_m,a_m) & \text{, if } A^m(s_m,a_m) > 0 \\ \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{W^-}{1-\pi_m(a_m|s_m)} \cdot (-A^m(s_m,a_m)) & \text{, if } A^m(s_m,a_m) < 0 \end{cases} \tag{254}$$

$$\text{where } \begin{cases} (1-\pi_m(a_m|s_m)) \geq W^+ \geq \frac{(1-\pi_m(a_m|s_m))^2}{2-\pi_m(a_m|s_m)} \\ \pi_m(a_m|s_m) \geq W^- \geq \frac{\pi_m(a_m|s_m)\cdot(1-\pi_m(a_m|s_m))^2}{\pi_m(a_m|s_m)^2 - \pi_m(a_m|s_m)+1} \end{cases} \tag{255}$$

and it can also be lower bounded by:

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(s_m,a_m)^2 & \text{, if } A^m(s_m,a_m) > 0 \\ d_s^{\pi_{m+1}}(s_m)\cdot\pi_m(a_m|s_m)\cdot A^m(s_m,a_m)^2 & \text{, if } A^m(s_m,a_m) < 0 \end{cases} \tag{256}$$

Hence, considering the randomness of the generator, it will choose (s,a) with probability $d_{gen}(s,a)$ to update in each episode $m$. Then we can rewrite Lemma 16 in probability form :

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \geq \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(\tilde{s_m},\tilde{a_m})^2 & \text{, if } A^m(s_m,a_m) > 0\text{, w.p. } d_{gen}(\tilde{s_m},\tilde{a_m}) \\ \frac{d_s^{\pi_{m+1}}(s_m)}{2} \cdot A^m(s_m,a_m)^2 & \text{, if } A^m(s_m,a_m) > 0\text{, w.p. } d_{gen}(s,a) \\ d_s^{\pi_{m+1}}(s_m)\cdot\pi_m(a_m|s_m)\cdot A^m(s_m,a_m)^2 & \text{, if } A^m(s_m,a_m) < 0\text{, w.p. } d_{gen}(s,a) \end{cases} \tag{257}$$

Then, by taking expectation, we have :

$$\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^{\pi_{m+1}}(s) - V^{\pi_m}(s)] = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} d_{gen}(s',a')\cdot[V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \mid (s_m,a_m)=(s',a')] \tag{258}$$

$$\geq d_{gen}(\tilde{s_m},\tilde{a_m})\cdot[V^{\pi_{m+1}}(s) - V^{\pi_m}(s) \mid (s_m,a_m)=(\tilde{s_m},\tilde{a_m})] \tag{259}$$

$$\geq d_{gen}(\tilde{s_m},\tilde{a_m})\cdot\frac{d_s^{\pi_{m+1}}(s_m)}{2}\cdot A^m(\tilde{s_m},\tilde{a_m})^2 \tag{260}$$

$$\geq d_{gen}(\tilde{s_m},\tilde{a_m})\cdot\frac{d_s^{\pi_{m+1}}(s_m)}{2}\cdot(1-\gamma)^2\cdot\left(V^*(s) - V^{\pi_m}(s)\right)^2 \tag{261}$$

$$= d_{gen}(\tilde{s_m},\tilde{a_m})\cdot\frac{d_s^{\pi_{m+1}}(s_m)}{2}\cdot(1-\gamma)^2\cdot\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(s) - V^{\pi_m}(s)]^2 \tag{262}$$

The third inequality holds by Lemma 8.
The last equation holds since the performance difference at episode $m$ is independent of $(s_m, a_m)$, which is the state action pair chosen at episode $m$.

If we consider the whole starting state distribution $\mu$, we have:

$$\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^{\pi_{m+1}}(\mu) - V^{\pi_m}(\mu)] \geq d_{gen}(\tilde{s_m},\tilde{a_m})\cdot\frac{d_\mu^{\pi_{m+1}}(s_m)}{2}\cdot(1-\gamma)^2\cdot\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu) - V^{\pi_m}(\mu)]^2 \tag{263}$$

$$\geq d_{gen}(\tilde{s_m},\tilde{a_m})\cdot\frac{\mu(s_m)}{2}\cdot(1-\gamma)^3\cdot\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu) - V^{\pi_m}(\mu)]^2 \tag{264}$$

$$\geq \underbrace{\min_{(s',a')\in\mathcal{S}\times\mathcal{A}}\{d_{gen}(s',a')\cdot\mu(s')\}\cdot\frac{(1-\gamma)^3}{2}}_{:=c'>0}\cdot\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu) - V^{\pi_m}(\mu)]^2 \tag{265}$$

The second inequality holds since $d_\mu^\pi \geq (1-\gamma)\cdot\mu(s)$ by Lemma 10.

Since $\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^{\pi_{m+1}}(\mu) - V^{\pi_m}(\mu)] = \mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}\left[V^{\pi^*}(\mu) - V^{\pi_m}(\mu)\right] - \mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}\left[V^{\pi^*}(\mu) - V^{\pi_{m+1}}(\mu)\right]$, by rearranging the inequality above, we have:

$$\delta_{m+1} \leq \delta_m - c'\cdot\delta_m^2 \quad \text{where } \delta_m = \mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}\left[V^{\pi^*}(\mu) - V^{\pi_m}(\mu)\right] \tag{266}$$

Then, we can get the following result by Lemma 11 :

$$\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu) - V^{\pi_m}(\mu)] \leq \frac{1}{c'}\cdot\frac{1}{m}, \quad \text{for all } m\geq 1 \tag{267}$$

$$\sum_{m=1}^{M}\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu) - V^{\pi_m}(\mu)] \leq \min\left\{\sqrt{\frac{M}{c'\cdot(1-\gamma)}}, \frac{\log M + 1}{c'}\right\}, \quad \text{for all } m\geq 1 \tag{268}$$

(269)

where $c' = \frac{(1-\gamma)^3}{2}\cdot\min_{(s,a)}\{d_{gen}(s,a)\cdot\mu(s)\} > 0$.

Finally, we get the desired result by Lemma 9:

$$\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\rho)-V^{\pi_m}(\rho)] \le \frac{1}{1-\gamma}\cdot\left\|\frac{1}{\mu}\right\|_\infty \cdot \mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu)-V^{\pi_m}(\mu)]) \le \frac{1}{c}\cdot\frac{1}{m}, \quad \text{for all } m\ge 1 \tag{270}$$

$$\sum_{m=1}^{M}\mathop{\mathbb{E}}_{(s_m,a_m)\sim d_{gen}}[V^*(\mu)-V^{\pi_m}(\mu)] \le \min\left\{\sqrt{\frac{M}{c\cdot(1-\gamma)}}, \frac{\log M+1}{c}\right\}, \quad \text{for all } m\ge 1 \tag{271}$$

where $c=\frac{(1-\gamma)^4}{2}\cdot\left\|\frac{1}{\mu}\right\|_\infty^{-1}\cdot \mathop{min}_{(s,a)}\{d_{gen}(s,a)\cdot\mu(s)\}>0$.

□

# C ON-POLICY CAPO WITH GLOBAL CONVERGENCE

The main focus and motivation for CAPO is on off-policy RL. Despite this, we show that it is also possible to apply CAPO to on-policy learning. While on-policy learning is a fairly natural RL setting, one fundamental issue with on-policy learning is the *committal issue*, which was recently discovered by (Chung et al., 2021; Mei et al., 2021). In this section, we show that CAPO could tackle the committal issue with the help of variable learning rates. Consider on-policy CAPO with state-action dependent learning rate:

$$\theta_{m+1}(s,a)=\theta_m(s,a)+\alpha^{(m)}(s,a)\cdot\operatorname{sign}(A^{(m)}(s,a))\cdot\mathbb{I}\{a=a_m\}, \tag{272}$$

where $N^{(k)}(s,a)=\sum_{m=0}^{k}\mathbb{I}\{(s,a)\in\mathcal{B}_m\}$ and $\alpha^{(m)}(s,a)$ is given by:

$$\alpha^{(m)}(s,a)=\begin{cases}\log\left(\frac{1}{\pi^{(m)}(a|s)}\right), & \text{if } A^{(m)}(s,a)\le 0\\ \log\left(\frac{\beta}{1-\beta}\cdot\frac{1}{\pi^{(m)}(a|s)}\right), & \text{if } A^{(m)}(s,a)>0 \text{ and } \pi^{(m)}(a|s)<\beta\\ \zeta\log\left(\frac{N^{(m)}(s,a)+1}{N^{(k)}(s,a)}\right), & \text{otherwise}\end{cases} \tag{273}$$

**Remark 7.** *Since on-policy exploration does not necessarily achieve infinite visitation, a more careful design of step size is needed. We include on-CAPO to demonstrate the novelty of using variable learning rate as opposed to fixed learning rate and show that on-CAPO can still achieve global convergence in Theorem 5 below.*

### C.1 Global Convergence of On-Policy CAPO

Recall that in the on-policy setting, we choose the step size of CAPO as

$$\alpha^{(k)}(\pi^{(k)}(a|s))=\begin{cases}\log\left(\frac{1}{\pi^{(k)}(a|s)}\right), & \text{if } A^{(k)}(s,a)\le 0\\ \log\left(\frac{\beta}{1-\beta}\cdot\frac{1}{\pi^{(k)}(a|s)}\right), & \text{if } A^{(k)}(s,a)>0 \text{ and } \pi^{(k)}(a|s)<\beta\\ \zeta\log\left(\frac{N^{(k)}(s,a)+1}{N^{(k)}(s,a)}\right), & \text{otherwise}\end{cases} \tag{274}$$

**Theorem 5.** *Under on-policy CAPO with $0<\beta\le\frac{1}{|\mathcal{A}|+1}$ and $0<\zeta\le\frac{1}{|\mathcal{A}|}$, we have $V_k(s)\to V^*(s)$ as $k\to\infty$, for all $s\in\mathcal{S}$, almost surely.*

To prove this result, we start by introducing multiple supporting lemmas.

**Lemma 19** (A Lower Bound of Action Probability)**.** *Under on-policy CAPO, in any iteration $k$, if an action $a$ that satisfies $\pi^{(k)}(a|s)<\beta$ and $A^{(k)}(s,a)>0$ is selected for policy update, then we have $\pi^{(k+1)}(a|s)>\beta$.*

*Proof of Lemma 19.* By the on-policy CAPO update in (274), we know that if the selected action $a$ satisfies $\pi^{(k)}(a|s)<\beta$

and $A^{(k)}(s,a) > 0$, we have

$$\theta_{s,a}^{(k+1)} = \theta_{s,a}^{(k)} + \log\big(\frac{\beta}{1-\beta}\cdot\frac{1}{\pi^{(k)}(a|s)}\big) \tag{275}$$

$$= \theta_{s,a}^{(k)} + \log\Big(\frac{\beta}{1-\beta}\cdot\frac{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'}^{(k)})}{\exp(\theta_{s,a}^{(k)})}\Big) \tag{276}$$

$$= \log\Big(\frac{\beta}{1-\beta}\cdot\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'}^{(k)})\Big). \tag{277}$$

Therefore, by the softmax policy parameterization, we have

$$\pi^{(k+1)}(a|s) = \frac{\frac{\beta}{1-\beta}\cdot\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'}^{(k)})}{\frac{\beta}{1-\beta}\cdot\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'}^{(k)}) + \sum_{a''\in\mathcal{A},a''\neq a}\exp(\theta_{s,a''}^{(k)})} \tag{278}$$

$$= \frac{\frac{\beta}{1-\beta}}{\frac{\beta}{1-\beta} + (1-\pi^{(k)}(a|s))} > \beta. \tag{279}$$

□

As we consider tabular policy parameterization, we could discuss the convergence behavior of each state separately. For ease of exposition, we first fix a state $s\in\mathcal{S}$ and analyze the convergence regarding the policy at state $s$. Define the following events:

$$E_0 := \Big\{\omega : I_s^+(\omega) \neq \varnothing\Big\}, \tag{280}$$

$$E_1 := \Big\{\omega : \lim_{k\to\infty}\pi_{s,a}^{(k)}(\omega) = 0, \forall a\in I_s^-(\omega)\Big\}, \tag{281}$$

$$E_{1,1} := \Big\{\omega : \exists a\in I_s^- \text{ with } N_{s,a}^{(\infty)}(\omega) = \infty\Big\}, \tag{282}$$

$$E_{1,2} := \Big\{\omega : \exists a\in I_s^+ \text{ with } N_{s,a}^{(\infty)}(\omega) = \infty\Big\}, \tag{283}$$

$$E_{1,3} := \Big\{\omega : \exists a'\in I_s^0(\omega) \text{ with } N_{s,a'}^{(\infty)}(\omega) = \infty\Big\}. \tag{284}$$

Since there shall always exist at least one action $a\in\mathcal{A}$ with $N_{s,a}^{(\infty)} = \infty$ for each sample path, then we have $E_{1,1}\cup E_{1,2}\cup E_{1,3} = \Omega$. Therefore, we can rewrite the event $E_1^c$ as $E_1^c = (E_1^c\cap E_{1,1})\cup(E_1^c\cap E_{1,2})\cup(E_1^c\cap E_{1,3})$. By the union bound, we have

$$\mathbb{P}(E_1^c|E_0) \leq \sum_{i=1}^{3}\mathbb{P}(E_1^c\cap E_{1,i}|E_0). \tag{285}$$

**Lemma 20.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_1^c\cap E_{1,1}|E_0) = 0$.*

*Proof of Lemma 20.* Under on-policy CAPO and the condition that $E_0$ happens, for each $\omega$, there exists an action $a'\in I_s^+(\omega)$ and some finite constant $B_0$ such that $\theta^{(k)}(s,a') \geq B_0$, for all sufficiently large $k \geq T_{s,a'}^+(\omega)$. On the other hand, for each $a''\in I_s^-(\omega)$, we know that $\theta^{(k)}(s,a'')$ is non-increasing for all $k\geq T_{s,a''}^-$. Therefore, $\pi^{(k)}(s,a'') \leq \frac{\exp\left(\theta^{\left(T_{s,a''}^-\right)}(s,a'')\right)}{\exp\left(\theta^{\left(T_{s,a''}^-\right)}(s,a'')\right)+\exp(B_0)}$, for all $k\geq\max\{T_{s,a'}^+, T_{s,a''}^-\}$. As a result, we know if $(s,a'')$ is contained in $\mathcal{B}^{(k)}$ with $k\geq\max\{T_{s,a'}^+, T_{s,a''}^-\}$, under CAPO, we must have

$$\theta_{s,a''}^{(k+1)} - \theta_{s,a''}^{(k)} \leq -\log\left(\frac{\exp\left(\theta^{(T_{s,a''}^-)}(s,a'')\right)+\exp(B_0)}{\exp\left(\theta^{(T_{s,a''}^-)}(s,a'')\right)}\right). \tag{286}$$

Therefore, for each $\omega\in E_0$ and for each $a''\in I_s^-(\omega)$, if $N_{s,a''}^{(\infty)}(\omega)=\infty$, then we have $\theta_{s,a''}^{(k)}(\omega)\to-\infty$ as $k\to\infty$. This implies that $\mathbb{P}(E_1^c\cap E_{1,1}|E_0) = 0$. □

**Lemma 21.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_1^c \cap E_{1,2}|E_0) = 0$.*

*Proof of Lemma 21.* By Lemma 20, we have $\mathbb{P}(E_1^c \cap E_{1,2}|E_0) = \mathbb{P}(E_1^c \cap E_{1,1}^c \cap E_{1,2}|E_0)$. Let $a \in I_s^+$ be an action with $N_{s,a}^{(\infty)}(\omega) = \infty$, and suppose $N_{s,a'}^{(\infty)}$ are finite for all $a' \in I_s^-$ (which also implies that $\theta_{s,a'}^{(k)}$ are finite for all $k \in \mathbb{N}$). Let $\{k_m\}_{m=1}^{\infty}$ be the sequence of iteration indices where $(s, a)$ in included in the batch. Now we discuss two possible cases as follows:

- Case 1: $\pi^{(k_m)}(a|s) \to 1$ as $m \to \infty$

  Conditioning on $E_0$, both $I_s^+$ and $I_s^-$ are non-empty. Since $\theta_{s,a'}^{(k)}$ is finite for each $a' \in I_s^-$, we know that $\pi^{(k_m)}(a|s) \to 1$ implies that
  $$\theta_{s,a}^{(k_m)} \to \infty, \text{ as } m \to \infty. \tag{287}$$
  Moreover, under CAPO, as $\theta_{s,a}^{(k_m)}$ shall be increasing for all sufficiently large $m$ (given that $a \in I_s^+$), we know (287) implies that $\theta_{s,a}^{(k)} \to \infty,$ as $k \to \infty$. Therefore, we have $\lim_{k\to\infty} \pi^{(k)}(a'|s) = 0$, for all $a' \in I_s^-$.

- Case 2: $\pi^{(k_m)}(a|s) \nrightarrow 1$ as $m \to \infty$: Since $A^{(k_m)}(s, a)$ shall be positive for all sufficiently large $m$ (given that $a \in I_s^+$), we know: (i) If $\pi^{(k_m)}(a|s) \geq \beta$, we have $\theta_{s,a}^{(k_m+1)} - \theta_{s,a}^{(k_m)} \geq \zeta \log(\frac{N^{(k_m)}(s,a)+1}{N^{(k_m)}(s,a)}) = \zeta \log(\frac{m+1}{m})$; (ii) Otherwise, if $\pi^{(k_m)}(a|s) < \beta$, we shall have $\theta_{s,a}^{(k_m+1)} - \theta_{s,a}^{(k_m)} \geq \log(\frac{1}{1-\beta}) > \zeta \log(\frac{m+1}{m})$, for all sufficiently large $m$. This implies that $\theta^{(k_m)}(s, a) \to \infty$ as $m \to \infty$. As $\theta_{s,a}^{(k_m)}$ shall be increasing for all sufficiently large $m$ (given that $a \in I_s^+$), we also have $\theta_{s,a}^{(k)} \to \infty$, as $k \to \infty$. As $\theta_{s,a'}^{(k)}$ remains finite for all $a' \in I_s^-$, we therefore have that $\lim_{k\to\infty} \pi^{(k)}(a'|s) = 0$, for all $a' \in I_s^-$.

□

**Lemma 22.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_1^c \cap E_{1,3}|E_0) = 0$.*

*Proof of Lemma 22.* By Lemma 20 and Lemma 21, we have $\mathbb{P}(E_1^c \cap E_{1,3}|E_0) = \mathbb{P}(E_1^c \cap E_{1,2}^c \cap E_{1,1}^c \cap E_{1,3}|E_0)$. Under $E_{1,1}^c \cap E_{1,2}^c$, we know that any action in $I_s^+ \cup I_s^-$ can appear in $\mathcal{B}^{(k)}$ only for finitely many times. This implies that there exists $T_0 \in \mathbb{N}$ such that $\mathcal{B}^{(k)}$ contains only actions in $I_s^0$, for all $k \geq T_0$. In order for the above to happen, we must have $\sum_{a \in I_s^0} \pi^{(k)}(a|s) \to 1$, as $k \to \infty$ (otherwise there would exist some $\epsilon > 0$ such that $\sum_{a \in I_s^0} \pi^{(k)}(a|s) \leq 1 - \epsilon$ for infinitely many $k$). This implies that $\lim_{k\to\infty} \pi^{(k)}(a'|s) = 0$, for any $a' \in I_s^-$. Hence, $\mathbb{P}(E_1^c \cap E_{1,2}^c \cap E_{1,1}^c \cap E_{1,3}|E_0) = 0$. □

**Lemma 23.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_1|E_0) = 1$.*

*Proof of Lemma 23.* By (285), Lemma 20, Lemma 21, and Lemma 22, we know $\mathbb{P}(E_1^c|E_0) = 0$. □

Before we proceed, we define the following events:

$$E_2 := \Big\{\omega : \lim_{k\to\infty} \pi_{s,a}^{(k)}(\omega) = 0, \forall a \in I_s^+(\omega)\Big\}, \tag{288}$$

$$E_3 := \Big\{\omega : \exists a \in I_s^+(\omega) \text{ with } N_{s,a}^{(\infty)}(\omega) = \infty\Big\} \tag{289}$$

**Lemma 24.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_2|E_0) = 1$.*

*Proof.* This is a direct result of Lemma 23. □

**Lemma 25.** *Under on-policy CAPO and the condition that $\mathbb{P}(E_0) > 0$, we have $\mathbb{P}(E_2 \cap E_3|E_0) = 0$.*

*Proof.* Under the event $E_2$, we know that for each action $a \in I_s^+$, for any $\epsilon > 0$, there exists some $T_{a,\epsilon}$ such that $\pi_{s,a}^{(k)} < \epsilon$ for all $k \geq T_{a,\epsilon}$. On the other hand, by Lemma 19, under $E_3$, we know that $\pi_{s,a}^{(k)} > \beta$ infinitely often. Hence, we know $\mathbb{P}(E_2 \cap E_3|E_0) = 0$. □

Note that by Lemma 24 and Lemma 25, we have $\mathbb{P}(E_2 \cap E_3^c | E_0) = 1$.

The main idea of the proof of Theorem 5 is to establish a contradiction by showing that under $E_0$, $E_3^c$ cannot happen with probability one. Let us explicitly write down the event $E_3^c$ as follows:

$$E_3^c := \big\{\omega : \exists \tau(\omega) < \infty \text{ such that } \mathcal{B}^{(k)} \subseteq I_s^0 \cup I_s^-, \forall k \geq \tau(\omega)\big\}. \tag{290}$$

Define

$$\theta_{s,\max}^{(k)} := \max_{a \in \mathcal{A}} \theta_{s,a}^{(k)}. \tag{291}$$

**Lemma 26.** *For any $t \in \mathbb{N}$ and any $K \in \mathbb{N}$, we have*

$$\theta_{s,\max}^{(t+K)} - \theta_{s,\max}^{(t)} \leq \log(K+1), \tag{292}$$

*for every sample path.*

*Proof of Lemma 26.* We consider the changes of $\theta_{s,a}$ of each action separately:

- $\theta_{s,a}^{(k)} < \theta_{s,\max}^{(k)}$, and $\pi^{(k)}(a|s) < \beta$: For such an action $a$, we have

$$\theta_{s,a}^{(k+1)} \leq \theta_{s,a}^{(k)} + \log\Big(\frac{\beta}{(1-\beta)\cdot \pi^{(k)}(a|s)}\Big) \tag{293}$$

$$\leq \log\Big(\frac{\beta}{(1-\beta)} \cdot |\mathcal{A}| \exp(\theta_{s,\max}^{(k)})\Big), \tag{294}$$

$$\leq \log\Big(\frac{\frac{1}{|\mathcal{A}|+1}}{(1-\frac{1}{|\mathcal{A}|+1})} \cdot |\mathcal{A}| \exp(\theta_{s,\max}^{(k)})\Big) \tag{295}$$

$$= \theta_{s,\max}^{(k)}, \tag{296}$$

  where (293) holds by the design of on-policy CAPO, (294) follows from the softmax policy parameterization, and (295) follows from the definition of $\theta_{s,\max}^{(k)}$ and the condition of $\beta$. Note that (293) would be an equality if $A^{(k)}(s,a) > 0$. As a result, this change cannot lead to an increase in $\theta_{s,\max}^{(k)}$.

- $\theta_{s,a}^{(k)} < \theta_{s,\max}^{(k)}$, and $\pi^{(k)}(a|s) \geq \beta$: For such an action $a$, we have

$$\theta_{s,a}^{(k+1)} \leq \theta_{s,a}^{(k)} + \zeta \log\Big(\frac{N_{s,a}^{(k)}+1}{N_{s,a}^{(k)}}\Big), \tag{297}$$

  where (297) holds by the design of on-policy CAPO and would be an equality if $A^{(k)}(s,a) > 0$.

- $\theta_{s,a}^{(k)} = \theta_{s,\max}^{(k)}$: Similarly, we have

$$\theta_{s,a}^{(k+1)} \leq \theta_{s,a}^{(k)} + \zeta \log\Big(\frac{N_{s,a}^{(k)}+1}{N_{s,a}^{(k)}}\Big), \tag{298}$$

  where (298) holds by the design of on-policy CAPO and would be an equality if $A^{(k)}(s,a) > 0$.

Based on the above discussion, we thereby know

$$\theta_{s,\max}^{(k+1)} - \theta_{s,\max}^{(k)} \leq \zeta \sum_{a \in \mathcal{A}} \log\Big(\frac{N_{s,a}^{(k+1)}}{N_{s,a}^{(k)}}\Big), \quad \forall k. \tag{299}$$

Therefore, for any $t \in \mathbb{N}$, the maximum possible increase in $\theta^{(k)}_{s,\max}$ between the $t$-th and the $(t+K)$-th iterations shall be upper bounded as

$$\theta^{(t+K)}_{s,\max} - \theta^{(t)}_{s,\max} \leq \sum_{k=t}^{t+K-1} \zeta \sum_{a \in \mathcal{A}} \log\Big(\frac{N^{(k+1)}_{s,a}}{N^{(k)}_{s,a}}\Big) \tag{300}$$

$$\leq \zeta \cdot \sum_{a \in \mathcal{A}} \log\Big(\frac{N^{(t)}_{s,a} + K}{N^{(t)}_{s,a}}\Big) \tag{301}$$

$$\leq \log(K+1), \tag{302}$$

where (300) follows directly from (299), (300) is obtained by interchanging the summation operators, and (301) holds by the condition that $\zeta \leq \frac{1}{|\mathcal{A}|}$. Hence, we know $\theta^{(t+K)}_{s,\max} - \theta^{(t)}_{s,\max} \leq \log(K+1)$. □

For any fixed action set $I^{\dagger}_s \subset \mathcal{A}$, define

$$E_4(I^{\dagger}_s) := \big\{\omega : \text{ For every } a \in I^{\dagger}_s, N^{(\infty)}(s,a) < \infty\big\}. \tag{303}$$

**Lemma 27.** *For any $I^{\dagger}_s \subset \mathcal{A}$, we have $\mathbb{P}(E_4(I^{\dagger}_s)) = 0$.*

*Proof of Lemma 27.* For a given action set $I^{\dagger}_s \subset \mathcal{A}$, define a sequence of events as follows: For each $n \in \mathbb{N}$,

$$E_{4,n}(I^{\dagger}_s) := \big\{\omega : \text{ For every } a \in I^{\dagger}_s, (s,a) \notin \mathcal{B}^{(k)}, \forall k \geq n\big\}. \tag{304}$$

$\{E_{4,n}(I^{\dagger}_s)\}_{n=1}^{\infty}$ form an increasing sequence of events, i.e., $E_{4,1}(I^{\dagger}_s) \subseteq E_{4,2}(I^{\dagger}_s) \cdots \subseteq E_{4,n}(I^{\dagger}_s) \subseteq E_{4,n+1}(I^{\dagger}_s) \cdots$. Moreover, we have $E_4(I^{\dagger}_s) = \bigcup_{n=1}^{\infty} E_{4,n}(I^{\dagger}_s)$. By the continuity of probability, we have

$$\mathbb{P}(E_4(I^{\dagger}_s)) = \mathbb{P}(\lim_{n\to\infty} E_{4,n}(I^{\dagger}_s)) = \lim_{n\to\infty} \mathbb{P}(E_{4,n}(I^{\dagger}_s)). \tag{305}$$

Next, we proceed to evaluate $\mathbb{P}(E_{4,n}(I^{\dagger}_s))$.

$$\log\big(\mathbb{P}(E_{4,n}(I^{\dagger}_s))\big) \leq \log\Bigg(\prod_{k\geq n} \frac{\sum_{a' \in I^0_s \cup I^-_s} \exp(\theta^{(k)}_{s,a'})}{\sum_{a' \in I^0_s \cup I^-_s} \exp(\theta^{(k)}_{s,a'}) + \sum_{a \in I^+_s} \exp(\theta^{(k)}_{s,a})}\Bigg) \tag{306}$$

$$\leq \log\Bigg(\prod_{k\geq n} \frac{|\mathcal{A}| \exp(\theta^{(k)}_{s,\max})}{|\mathcal{A}| \exp(\theta^{(k)}_{s,\max}) + \sum_{a \in I^+_s} \exp(\theta^{(n)}_{s,a})}\Bigg) \tag{307}$$

$$\leq \log\Bigg(\prod_{m\geq 1} \frac{|\mathcal{A}| \exp\big(\theta^{(n)}_{s,\max} + \log(m+1)\big)}{|\mathcal{A}| \exp\big(\theta^{(n)}_{s,\max} + \log(m+1)\big) + \sum_{a \in I^+_s} \exp(\theta^{(n)}_{s,a})}\Bigg) \tag{308}$$

$$\leq \sum_{m\geq 1} \log\Bigg(1 - \frac{\sum_{a \in I^+_s} \exp(\theta^{(n)}_{s,a})}{|\mathcal{A}|(m+1)\exp(\theta^{(n)}_{s,\max})}\Bigg) = -\infty, \tag{309}$$

where (306) holds by the softmax policy parameterization, (307) holds by the definition of $\theta^{(k)}_{s,\max}$ and $E_{4,n}(I^{\dagger}_s)$, and (307) follows directly from Lemma 26. Equivalently, we have $\mathbb{P}(E_{4,n}(I^{\dagger}_s)) = 0$, for all $n \in \mathbb{N}$. By (305), we conclude that $\mathbb{P}(E_4(I^{\dagger}_s)) = 0$. □

Now we are ready to prove Theorem 5.

*Proof of Theorem 5.* Recall that the main idea is to establish a contradiction by showing that conditioning on $E_0$, $E^c_3$ cannot happen with probability one. Note that by Lemma 24 and Lemma 25, we have $\mathbb{P}(E_2 \cap E^c_3 | E_0) = 1$. However, by Lemma 27, we know that for any fixed action set $I^{\dagger}_s \subset \mathcal{A}$, the event that the actions in $I^{\dagger}_s$ are selected for policy updates for only finitely many times must happen with probability zero. This contradicts the result in Lemma 25. Therefore, we shall have $\mathbb{P}(E_0) = 0$. □

### C.2 On-Policy CAPO with Fixed Learning Rate

One interesting question is whether on-policy CAPO can be applied with a fixed learning rate. Through a simple single state bandit example, we show that without the help of variable learning rate, on-policy CAPO with fixed learning rate will stuck in local optimum with positive probability. Therefore, this fact further motivates the use of variable learning rate in CAPO. We provide the detailed discussion in Appendix D.

## D SUB-OPTIMALITY OF ON-POLICY CAPO DUE TO IMPROPER STEP SIZES

In this section, we construct a toy example to further showcase how the proposed CAPO benefits from the properly-designed step sizes in Algorithm 1. We consider a deterministic $K$-armed bandit with a single state and an action set $[K]$ and a softmax policy $\pi_\theta : [K] \rightarrow [0, 1]$, the reward vector $r \in \mathbb{R}^K$ is the reward corresponding to each action. This setting is the same as the one in Section 2 of (Mei et al., 2021), except that we do not have the assumption of positive rewards such that $r(a) \in [0, 1), \forall a \in [K]$, the reward can be any real number such that $r \in \mathbb{R}^K$. Our goal here is to find the optimal policy $\pi^*$ that maximize the expected total reward. Since there is only one single state, the objective function can by written as:

$$J(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot)}[r(a)]. \tag{310}$$

The on-policy CAPO with fixed learning rate updates the policy parameters by:

$$\theta_{m+1}(s, a) = \theta_m(s, a) + \eta \cdot \text{sign}(A(s, a)) \cdot \mathbb{I}\{a = a_m\} \tag{311}$$

where $\eta$ is a constant representing the fixed learning rate.

To demonstrate that on-policy CAPO with fixed learning rate can get stuck in a sub-optimal policy, we consider a simple three-armed bandit where $K = 3$ (i.e. a single state with 3 actions). We set $r = [1, 0.99, -1]$. Then we have:

**Theorem 6.** *Given a uniform initial policy $\pi_1$ such that $\pi_1(a) = \frac{1}{K}, \forall a \in [K]$, under the policy update of (311), we have $\mathbb{P}(\pi_\infty(a_2) = 1) > 0$.*

The idea is that with $\pi_1(a_1) = \pi_1(a_3)$ and $r(a_1) = -r(a_3)$, when we only sample $a_2$ in the first $t$ steps, $A_m(a_2) > 0, \forall m \leq t$. Thus, $\pi_m(a_2)$ shall be strictly improving, and the probability of sampling $a_2$ will increase accordingly, thus causing a vicious cycle.

Theorem 6 shows that the naive fixed learning rate is insufficient. In the next section, we will show that with a properly chosen variable learning rate, on-policy CAPO can guarantee global convergence. Empirical results can be found in Appendix E.

*Proof of Theorem 6.* Inspired by the proof in (Mei et al., 2021) (Theorem 3, second part), we also consider the event $\mathcal{E}_t$ such that $a_2$ is chosen in the first $t$ time steps. We will show that there exists some sequence $b_s$ such that $\mathbb{P}(\mathcal{E}_t) \geq \prod_{s=1}^{t} b_s > 0$.

The first part argument is the same as (Mei et al., 2021), we restate the argument for completeness: Let $\mathcal{B}_m = \{a_m = a_2\}$ be the event that $a_2$ is sampled at time $m$. Define the event $\mathcal{E}_t = \mathcal{B}_1 \cap \cdots \cap \mathcal{B}_t$ be the event that $a_2$ is chosen in the first $t$ time steps. Since $\{\mathcal{E}_t\}_{t\geq 1}$ is a nested sequence, we have $\lim_{t\to\infty} \mathbb{P}(\mathcal{E}_t) = \mathbb{P}(\mathcal{E})$ by monotone convergence theorem. Following equation (197) and equation (198) in (Mei et al., 2021), we will show that a suitable choice of $b_t$ under the On-policy CAPO with fixed learning rate is:

$$b_t = \exp\left\{-\frac{\sum_{a\neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2)\}} \cdot \frac{\exp\{\eta\}}{\eta}\right\}. \tag{312}$$

**Lemma 28.** $\pi_m(a_1) = \pi_m(a_3), \forall 1 \leq m \leq t.$

*Proof of Lemma 28.* Under uniform initialization $\theta_1(a_1) = \theta_1(a_3)$, since only $a_2$ is sampled in the first $t$ steps, we have $\forall 1 \leq m \leq t$:

$$\pi_m(a_1) = \frac{\exp(\theta_m(a_1))}{\sum_a \exp(\theta_m(a))} \tag{313}$$

$$= \frac{\exp(\theta_1(a_1))}{\sum_a \exp(\theta_m(a))} = \frac{\exp(\theta_1(a_3))}{\sum_a \exp(\theta_m(a))} \tag{314}$$

$$= \pi_m(a_3). \tag{315}$$

□

**Lemma 29.** *For all* $1 \leq m \leq t$*, we have* $A_m(a_2) \geq 0$.

*Proof of Lemma 29.* Note that under the CAPO update (311), we have

$$A_m(a_2) = r(a_2) - \sum_a \pi_m(a) \cdot r(a) \tag{316}$$

$$= (1 - \pi_m(a_2))r(a_2) - \sum_{a \neq a_2} \pi_m(a) \cdot r(a) \tag{317}$$

$$= (1 - \pi_m(a_2))r(a_2) - \sum_{a \neq a_2} \pi_m(a) \cdot r(a) \tag{318}$$

$$= (1 - \pi_m(a_2))r(a_2) \geq 0, \tag{319}$$

where the last equation comes from Lemma 28 and $r(a_1) = -1 \cdot r(a_3)$.

□

**Lemma 30.** $\theta_t(a_2) = \theta_1(a_2) + \eta \cdot (t - 1)$.

*Proof of Lemma 30.* By Lemma 29 and (311), we have:

$$\theta_t(a_2) = \theta_1(a_2) + \eta \cdot \sum_{s=1}^{t-1} \text{sign}(A_s(a_2)) \cdot \mathbb{I}\{a_2 = a_s\} \tag{320}$$

$$= \theta_1(a_2) + \eta \cdot \sum_{s=1}^{t-1} 1 \tag{321}$$

$$= \theta_1(a_2) + \eta \cdot (t - 1). \tag{322}$$

(323)

□

**Lemma 31.** *For all* $x \in (0, 1)$*, we have:*

$$1 - x \geq \exp\left\{\frac{-x}{1-x}\right\} \tag{324}$$

*Proof of Lemma 31.* This is a direct result of Lemma 14 in (Mei et al., 2021). Here we also include the proof for completeness.

$$1 - x = \exp\{\log(1 - x)\} \tag{325}$$

$$\geq \exp\left\{1 - e^{-\log(1-x)}\right\} \quad \left(y \geq 1 - e^{-y}\right) \tag{326}$$

$$= \exp\left\{\frac{-1}{1/x - 1}\right\} \tag{327}$$

$$= \exp\left\{\frac{-x}{1-x}\right\}. \tag{328}$$

Then, we can plug in $x$ as $\frac{a}{b}$ for some $a < b$ to obtain a more useful form of this lemma as follows:

$$1 - \frac{a}{b} \geq \exp\left\{\frac{-a}{b-a}\right\}. \tag{329}$$

□

**Lemma 32.** $\pi_t(a_2) \geq \exp\left\{\frac{-\sum_{a \neq a_2} \exp\{\theta_t(a)\}}{\exp\{\theta_t(a_2)\}}\right\}$.

*Proof of Lemma 32.*

$$\pi_t(a_2) = 1 - \sum_{a \neq a_2} \pi_t(a) \tag{330}$$

$$= 1 - \frac{\sum_{a \neq a_2} \exp\{\theta_t(a)\}}{\exp\{\theta_t(a_2)\} + \sum_{a \neq a_2} \exp\{\theta_t(a)\}} \tag{331}$$

$$\geq \exp\left\{\frac{-\sum_{a \neq a_2} \exp\{\theta_t(a)\}}{\exp\{\theta_t(a_2)\}}\right\}, \tag{332}$$

where the last inequality uses (329). ☐

Finally, we have

$$\prod_{t=1}^{\infty} \pi_t(a_2) \geq \prod_{t=1}^{\infty} \exp\left\{\frac{-\sum_{a \neq a_2} \exp\{\theta_t(a)\}}{\exp\{\theta_t(a_2)\}}\right\} \tag{333}$$

$$= \prod_{t=1}^{\infty} \exp\left\{\frac{-\sum_{a \neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2) + \eta \cdot (t-1)\}}\right\} \tag{334}$$

$$= \exp\left\{\sum_{t=1}^{\infty} \frac{-\sum_{a \neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2) + \eta \cdot (t-1)\}}\right\} \tag{335}$$

$$= \exp\left\{-\frac{\sum_{a \neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2)\}} \cdot \exp\{\eta\} \cdot \sum_{t=1}^{\infty} \frac{1}{\exp\{\eta \cdot t\}}\right\} \tag{336}$$

$$\geq \exp\left\{-\frac{\sum_{a \neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2)\}} \cdot \exp\{\eta\} \cdot \int_{t=0}^{\infty} \frac{1}{\exp\{\eta \cdot t\}}\right\} \tag{337}$$

$$= \exp\left\{-\frac{\sum_{a \neq a_2} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a_2)\}} \cdot \frac{\exp\{\eta\}}{\eta}\right\} \tag{338}$$

$$= \Omega(1), \tag{339}$$

where the last line comes from the fact that $\sum_{a \neq a_2} \exp\{\theta_1(a)\} \in \Theta(1)$, $\exp\{\theta_1(a_2)\} \in \Theta(1)$ and $\frac{\exp\{\eta\}}{\eta} \in \Theta(1)$. ☐

# E A CLOSER LOOK AT THE LEARNING RATE

Unlike most RL algorithms, CAPO leverages variable learning rate that is state action dependent, instead of a fixed learning rate. In this section, we provide some insights into why this design is preferred under CAPO from both theoretical and empirical perspectives.

## E.1 Variable Learning Rate v.s. Fixed Learning Rate

In Lemma 16, we quantify the one-step improvement $V^{\pi_{m+1}}(s) - V^{\pi_m}(s)$ in terms of state visitation distribution, policy weight, and advantage value under learning rate $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$. Now, we provide the one-step improvement *under fixed learning rate*, $\alpha \in \mathbb{R}$, $\alpha > 0$:

$$V^{\pi_{m+1}}(s) - V^{\pi_m}(s) = \begin{cases} \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{(e^{\alpha}-1)\cdot\pi_m(a_m|s_m)}{(e^{\alpha}-1)\cdot\pi_m(a_m|s_m)+1} \cdot A^m(s_m,a_m) & \text{, if } A^m(s_m,a_m) > 0 \\ \frac{d_s^{\pi_{m+1}}(s_m)}{1-\gamma} \cdot \frac{(1-e^{-\alpha})\cdot\pi_m(a_m|s_m)}{(e^{-\alpha}-1)\cdot\pi_m(a_m|s_m)+1} \cdot (-A^m(s_m,a_m)) & \text{, if } A^m(s_m,a_m) < 0 \end{cases} \tag{340}$$

$$\text{where } \alpha \in \mathbb{R}, \alpha > 0 \tag{341}$$

Note that the result above can be obtained by using the same technique in Lemma 12, Lemma 13 and Lemma 16 by substituting the learning rate.

Compared to the one-step improvement under the variable learning rate, the one-step improvement under the fixed learning rate would be tiny as the updated action's policy weight $\pi_m(a_m|s_m) \to 0$. This property makes it difficult for an action that has positive advantage value but small policy weight to contribute enough to overall improvement, i.e., for those actions, the improvement of the policy weight $\pi_{m+1}(a_m|s_m) - \pi_m(a_m|s_m) \to 0$ under some improper fixed learning rate, leading to small one-step improvement.

Now, to provide some further insights into the possible disadvantage of a fixed learning rate, we revisit the proof of the convergence rate of Cyclic CAPO in Appendix B.1. By combining the one-step improvement above, the result from Case 1 and Case 2 under the fixed learning rate, $\alpha \in \mathbb{R}$, $\alpha > 0$ can be rewritten as:

$$V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s) \tag{342}$$

$$\geq \frac{(1-\gamma)^2}{2} \cdot \frac{1}{\max\left\{\frac{(1-\pi_{m+T}(a_{m+T}|s_{m+T}))\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})+1}{(1-\gamma)\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})}, \frac{c_m \cdot T}{(1-\gamma)^2}\right\}} \cdot \left(V^*(s) - V^{\pi_m}(s)\right)^2 \tag{343}$$

$$= \frac{(1-\gamma)^2}{2} \cdot \min\left\{\frac{(1-\gamma)\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})}{(1-\pi_{m+T}(a_{m+T}|s_{m+T}))\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})+1}, \frac{(1-\gamma)^2}{c_m \cdot T}\right\} \tag{344}$$

$$\cdot \left(V^*(s) - V^{\pi_m}(s)\right)^2 \tag{345}$$

where $c_m = \max_{k\in[m,m+T-1]}\{c_{k1}, c_{k2}\} \in [0,1]$

and $c_{k1} = \mathbb{I}\left\{A^k(s_k,a_k) > 0\right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot \frac{(e^{\alpha}-1)\cdot\pi_k(a_k|s_k)\cdot(1-\pi_k(a_k|s_k))}{(e^{\alpha}-1)\cdot\pi_k(a_k|s_k)+1}$, $c_{k2} = \mathbb{I}\left\{A^k(s_k,a_k) < 0\right\} \cdot d_s^{\pi_{k+1}}(s_k) \cdot \frac{(1-e^{-\alpha})\cdot\pi_k(a_k|s_k)\cdot(1-\pi_k(a_k|s_k))}{(e^{-\alpha}-1)\cdot\pi_{m+T}(a_m|s_m)+1}$.

Note that the first term $\frac{(1-\gamma)\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})}{(1-\pi_{m+T}(a_{m+T}|s_{m+T}))\cdot(e^{\alpha}-1)\cdot\pi_{m+T}(a_{m+T}|s_{m+T})+1}$ in the "min" operator is derived from Case 2 and the second term $\frac{(1-\gamma)^2}{c_m \cdot T}$ is derived from Case 1. Once we cannot guarantee that Case 1 provide enough amount of improvement, we must show that we can get the rest of the required improvement in Case 2. However, we can find that there is a term $\pi_{m+T}(a_{m+T}|s_{m+T})$ in the numerator of the first term in the "min" operator, which is provided by Case 2, implying that the multi-step improvement $V^{\pi_{m+|\mathcal{S}||\mathcal{A}|}}(s) - V^{\pi_m}(s)$ might also be tiny when the improvement provided by Case 1 is insufficient and the policy weight $\pi_{m+T}(a_{m+T}|s_{m+T}) \to 0$ in Case 2.

Accordingly, we highlight the importance of the choice of the learning rate, especially when the visitation frequency of the coordinate generator is extremely unbalanced (e.g. sampling the optimal action every $(|\mathcal{S}||\mathcal{A}|)^{1000}$ epoch) or the approximated advantage value is oscillating between positive and negative during the update. The design of the variable learning rate $\alpha_m(s,a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$ somehow tackles the difficulty of the insufficient one-step improvement by providing larger step size to the action with tiny policy weight, solving the problem of small improvement of the policy weight. Therefore, we can conclude that under this design of the learning rate, the one-step improvement is more steady with the policy weight of the action chosen for policy update.

### E.2 Demonstrating the Effect of Learning Rate in a Simple Bandit Environment

In this section, we present the comparison in terms of the empirical convergence behavior of On-policy CAPO and Off-policy CAPO. Specifically, we evaluate the following four algorithms: (i) On-Policy CAPO with state-action-dependent learning rate (cf. (274)), (ii) On-Policy CAPO with fixed learning rate (311), (iii) Off-Policy CAPO with state-action-dependent learning rate (cf. (12)), (iv) Off-Policy CAPO with fixed learning rate.

We consider the multi-armed bandit as in Appendix D with $K = 4$, and $r = [10, 9.9, 9.9, 0]$. To further demonstrate the ability of CAPO in escaping from the sub-optimal policies, instead of considering the uniform initial policy where $\pi_1(a) = \frac{1}{K}, \forall a \in [K]$, we initialize the policy to a policy that already prefers the sub-optimal actions $(a_2, a_3)$ such that $\theta_1 = [0, 3, 3, 0]$ and $\pi_1 \approx [0.0237, 0.4762, 0.4762, 0.0237]$ under the softmax parameterization. For each algorithm, we run the experiments under 100 random seeds. For all the variants of CAPO, we set $|B_m| = 1$.

In Figure 3, On-policy CAPO with fixed learning rate can get stuck in a sub-optimal policy due to the skewed policy initialization that leads to insufficient visitation to each action, and this serves an example for demonstrating the effect described in Theorem 6. On the other hand, on-policy CAPO with state-action dependent learning rate always converges to the global optimum despite the extremely skewed policy initialization. This corroborates the importance of variable learning rate for on-policy CAPO. Without such design, the policies failed to escape from a sub-optimal policy under all the random seeds.

Next, we look at the result of off-policy CAPO: We noticed that off-policy CAPO with fixed learning rate is able to identify the optimal action. However, Off-policy CAPO with fixed learning rate learns much more slowly than its variable learning rate counterpart (notice that the x-axis (Iteration) in each graph is scaled differently for better visualization). Also, we notice that the different choices of fixed learning rate have direct impact on the learning speed, and this introduces a hyperparameter that is dependent on the MDP. On the other hand, $\alpha_m(s, a)$ can be used as a general learning rate for different cases (For example, in Appendix F where a different environment Chain is introduced, learning rate for off-policy Actor Critic has to be tuned while $\alpha_m(s, a)$ can be used as the go-to learning rate.)

Finally, we also validate theory regarding the monotonic improvement of CAPO algorithms in Figure 3. We can tell that off-policy CAPO with $\alpha_m(s, a)$ enjoys monotonic improvement under all the random seeds.

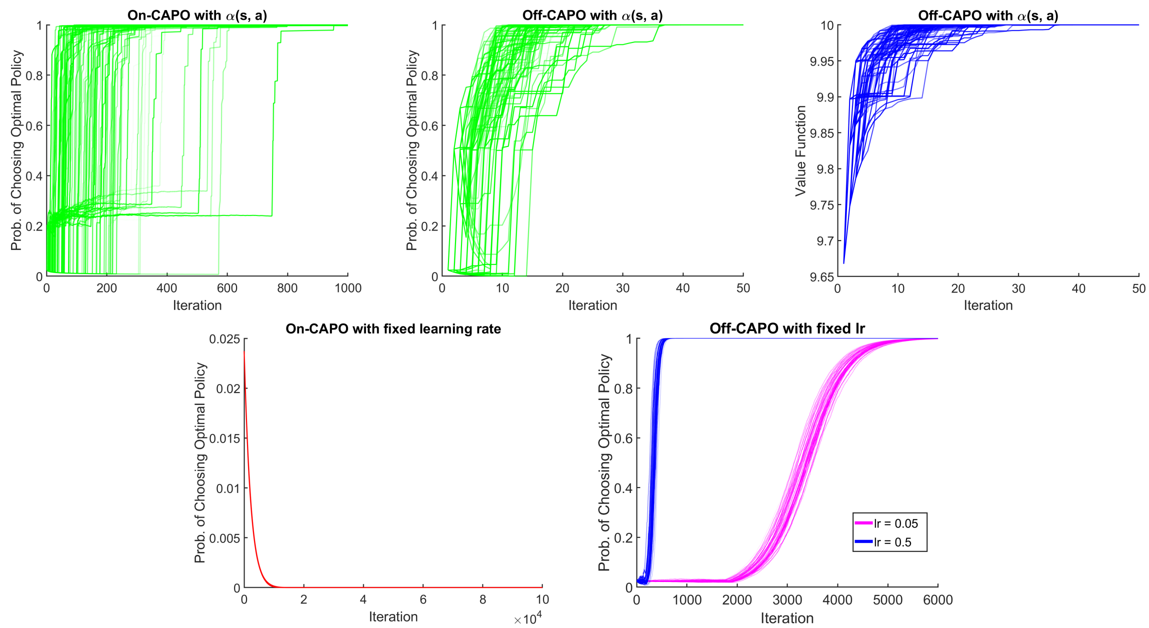


Figure 3: The probability weight and value function of the trained policies on the optimal action at different iterations.

### E.3 Unbounded Learning Rate

In CAPO update, unlike most of the optimization algorithm, we only constructed the lower bound of learning rate $\alpha_m(s, a) \geq \log(\frac{1}{\pi_{\theta_m}(a|s)})$ instead of constructing the upper bound of it. This can be explained by an argument similar to policy iteration (PI). Under PI, the new policy $\pi_{m+1}$ is updated by greedy one-step improvement $\pi_{m+1}(s) = \text{argmax}_{a \in \mathcal{A}}\{Q^{\pi_m}(s, a)\}$, which corresponds to an infinitely large stepsize under softmax policies, and still can achieve monotonic improvement and global optimum. The update of CAPO has a similar flavor, but with two salient differences: (i) CAPO does not take greedy actions (ii) and needs to address various coordinate selection rules and coordinate orderings, which makes the convergence analysis highly non-trivial.

## F ADDITIONAL EXPERIMENTS

### F.1 Exploration Capability Provided by a Coordinate Generator in CAPO

In this subsection, we demonstrate empirically the exploration capability provided by the coordinate generator in CAPO.

The Chain environment is visualized in Figure 4. This environment is meant to evaluate the agent's ability to resist the temptation of immediate reward and look for the better long-term return. We compare the performance of Batch CAPO, Cyclic CAPO and Off-policy Actor Critic on Chain with $N = 10$, and the result can be found in Figure 5. To eliminate the factor of critic estimation, true value of the value function is used during training. All the agents are trained for 1000 iterations with learning rate $= 0.001$. The policies are represented by a neural network with a single hidden layer (with hidden layer size 256). Both Cyclic CAPO and Off-policy Actor Critic is trained with a batch size of 16 and a replay buffer size of 100. As Batch CAPO shall take all the $\mathcal{SA}$-pairs into account by design, the effective batch size of Batch CAPO is equal to $\mathcal{S} \times \mathcal{A}$. Unlike the CAPO methods, Off-policy Actor Critic presumes the use of a fixed behavior policy. As a result, similar to the experimental setup of various prior works (e.g., (Liu et al., 2020b)), we use a uniform behavior policy for Off-policy Actor Critic. The use of a fixed behavior policy makes it difficult to identify an optimal policy, and this highlights the benefit of a coordinate generator in terms of exploration.

From Figure 5 we can see that it is difficult for Off-policy Actor Critic to escape from a sub-optimal policy, despite that the true value of the value function is provided. Since both Cyclic CAPO and Batch CAPO satisfy Condition 1, using such coordinate selection rules provides sufficient exploration for CAPO to identify the optimal policy. This feature can be particularly useful when the reward is sparse and the trajectory is long.

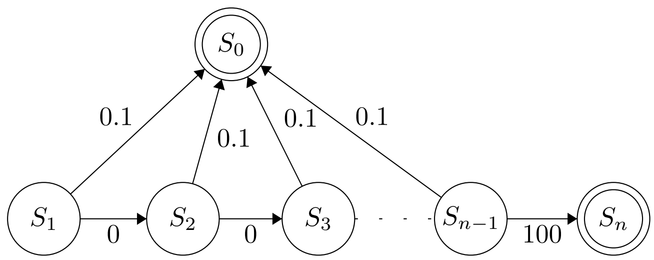


Figure 4: The Chain environment has a total of $n+1$ states, and the agent always starts at state 1. The agent has two actions to choose from at every state, either receive a reward of 0.1 and terminate immediately, or move one state to the right. While moving right will receive no reward in most states, the transition from $S_{n-1}$ to $S_n$ would induce a huge reward of 100. A well-performing policy should prefer the delayed reward of 100 over the immediate reward of 0.1.

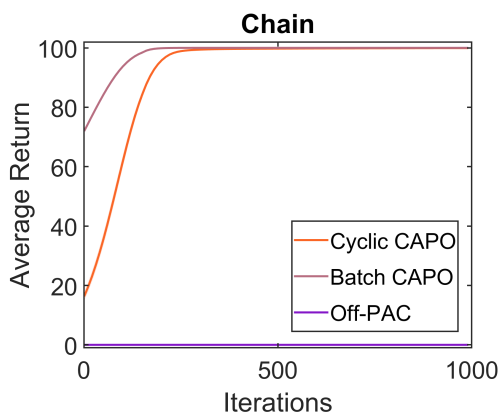


Figure 5: Comparison between Cyclic CAPO, Batch CAPO and off-policy Actor Critic, where the result is the average over 30 runs. We can see that despite the true value function is given and the optimal reward is much larger than the immediate reward (100 v.s. 0.1), Off-policy Actor Critic still suffers from a sub-optimal policy.

### F.2 Stochastic Environments

We validate the theory in a stochastic variant of *Chain* in Appendix F.2. To include stochasticity to *Chain*, when moving right in *Chain*, we modify the stride length to be uniformly random between 0 and 3. In Figure 6, we see that the results are consistent and CAPO performs well under stochasticity and converges faster.

### F.3 Validating theory

Also, since atari games are not the ideal field of application of PG methods, we validate the theory under a relatively simple non-atari environment, Gridworld (the goal is located at the bottom-right corner with a reward of 100, the agent moves with a cost of $-1$). In Figure 7, we see that the results are consistent and CAPO converges faster.

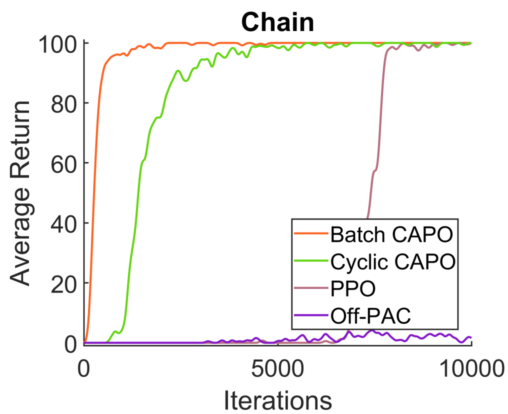


Figure 6: Comparison between Cyclic CAPO, Batch CAPO, PPO and Off-policy Actor Critic, where the result is the average over 30 runs. We can see that CAPO still performs well under stochasticity and converges faster.

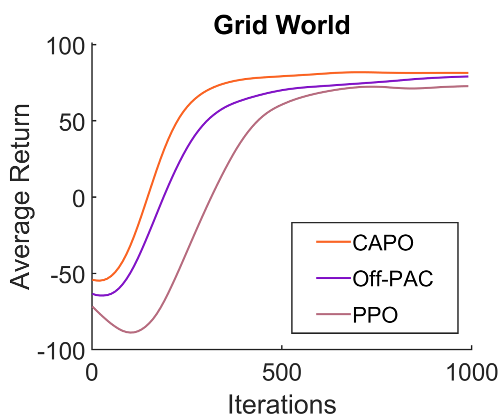


Figure 7: Comparison between CAPO, PPO and Off-policy Actor Critic. We can see that CAPO still performs well under stochasticity and converges faster.

# G DETAILED CONFIGURATION OF EXPERIMENTS

## G.1 Implementation Detail

Algorithm 3 shows the pseudo code of NCAPO. In order to demonstrate the off-policy capability of NCAPO, we use simple $\epsilon$-greedy with initial exploration $\epsilon_{start}$ and decayed exploration for off-policy exploration and estimates $A(s,a)$ with Retrace Munos et al. (2016). NCAPO uses four simple 2-layer feed-forward neural networks, a behavior network ($\theta^b$), a target network ($\theta$), a critic network ($\theta^Q$) and a target critic network ($\theta^{Q'}$). In each episode, $N_{\text{rollouts}}$ rollouts are collected, and each rollout $r = [(s_t, a_t), ..., (s_{t+l}, a_{t+l})]$ has length of $l$. Note that instead of storing a single $(s,a)$-pair in the replay buffer $R$, we store the entire rollout of length $l$ into $R$ to better compute $Q_{\text{retrace}}$. Due to the limited representation capability of floating-point numbers, during the CAPO update, the term $\log \frac{1}{\pi}$ can grow unbounded as $\pi \rightarrow 0$. To address this, we clip the term so that $\alpha(s,a) = \min(\log \frac{1}{\pi(a|s)}, \text{clip})$. As the target networks have demonstrated the ability to stabilize training, the target networks are used and updated by polyak average update with coefficient $\tau_\theta$ and $\tau_Q$. The experiment is conducted on a computational node equipped with Xeon Platinum 8160M CPU with a total of 40 cores. Off-PAC shares a similar code base as NCAPO (Note that both of them estimates $A(s,a)$ with Retrace.), the major difference is that the use of a fixed behavior policy. We choose such behavior policy to be a uniform policy.

## G.2 Hyperparameters

We use the hyperparameters for Atari games of stable-baseline3 for (Raffin et al., 2019) for PPO and A2C, and the exact same hyperparameter from (Obando-Ceron and Castro, 2021) for Rainbow. The hyperparameters are listed in Table 2.

Table 2: Hyperparameters for CAPO and OffPAC

| HYPERPARAMETERS | CAPO | PPO | A2C | OFFPAC |
|---|---|---|---|---|
| BATCH SIZE | 32 | 16 | - | 32 |
| LEARNING RATE | 5E-4 | 2.5E-4 | 7E-4 | 5E-4 |
| EXPLORATION FRACTION | 10% | 0 | 0 | 10% |
| INITIAL EXPLORATION RATE* | 0.3 | 0 | 0 | 0 |
| FINAL EXPLORATION RATE | 0.05 | 0 | 0 | 0 |
| CRITIC LOSS COEFFICIENT* | 1 | 0.38 | 0.25 | 1 |
| MAX GRADIENT NORM | 0.8 | 0.5 | 0.5 | 0.8 |
| GRADIENT STEPS | 30 | 1 | 1 | 30 |
| TRAIN FREQUENCY | (64, STEPS) | (256, STEPS) | - | (64, STEPS) |
| $\tau_Q$ | 0.05 | - | - | 0.05 |
| $\tau_\theta$ | 1 | - | - | 1 |
| GAMMA | 0.99 | 0.98 | 0.99 | 0.99 |
| REPLAY BUFFER | 6400 | - | - | 6400 |
| CLIP VALUE | 50 | - | - | - |
| ENTROPY COEF | 0 | - | 4.04E-6 | 0 |

* For *Asterix*, the critic loss coefficient is 0.25 and the initial exploration rate is 0.8.

# H PSEUDO CODE OF THE PROPOSED ALGORITHMS

**Algorithm 1** Coordinate Ascent Policy Optimization

1: Initialize policy $\pi_\theta$, $\theta \in \mathcal{S} \times \mathcal{A}$
2: **for** $m = 1, \cdots, M$ **do**
3: Generate $|\mathcal{B}|$ state-action pairs $((s_0, a_0), ..., (s_{|\mathcal{B}|}, a_{|\mathcal{B}|}))$ from some coordinate selection rule satisfying Condition 1.
4: **for** $i = 1, \cdots, |\mathcal{B}|$ **do**
5: $\theta_{m+1}(s_i, a_i) \leftarrow \theta_m(s_i, a_i) + \alpha_m(s_i, a_i)\,\mathrm{sign}\,(A^m(s_i, a_i))$
6: **end for**
7: **end for**

**Algorithm 2** Neural Coordinate Ascent Policy Optimization

1: Initialize actor network $f_\theta$, where policy is parameterized as $\pi_\theta(a|s) = \frac{f_\theta(s,a)}{\sum_{a' \in \mathcal{A}} f_\theta(s,a')}$
2: **for** $m = 1, \cdots, M$ **do**
3: Generate state-action pairs $\big((s_0, a_0), ..., (s_{|\mathcal{B}|}, a_{|\mathcal{B}|})\big)$ from some coordinate selection rule satisfying Condition 1.
4: Evaluate Advantage $A^{\pi_m}$ with arbitrary policy evaluation algorithm.
5: Compute target $\hat{\theta}$ by (18).
6: Compute target policy $\hat{\pi}$ by taking softmax over $\hat{\theta}$.
7: Update the policy network with NCAPO loss:
8: $\nabla_\theta L = \nabla_\theta D_{KL}\left(\pi_{f_{\theta_m}} \| \hat{\pi}\right)$
9: **end for**

**Algorithm 3** Neural Coordinate Ascent Policy Optimization with Replay Buffer

Initialize behavior network $f(s, a \mid \theta^b)$, critic $Q(s, a \mid \theta^Q)$
Initialize Replay Buffer $R$,
Initialize target networks $f(s, a \mid \theta) \leftarrow f(s, a \mid \theta^b)$, $Q(s, a|\theta^{Q'}) \leftarrow Q(s, a|\theta^Q)$
**for** episode $m = 1, \cdots, M$ **do**
  Generate behavior policy and target policy by computing softmax $\pi_\theta(a \mid s) = \frac{e^{f(s,a|\theta)}}{\sum_{a' \in \mathcal{A}} e^{f(s,a'|\theta)}}$.
  Collect $N_{rollouts}$ rollouts with length $l$ by following $\pi_{\theta_b}$ with decayed $\epsilon$-greedy, store rollouts to $R$.
  Replace old rollouts if $len(R) > R_{max}$ .
  **for** gradient steps $= 1, ..., \mathcal{G}$ **do**
    Sample rollout $r$ from $R$.
    Compute $Q_{retrace}(s, a)$ for $(s, a) \in r$
    $\theta^Q_{Loss} \leftarrow \sum_{(s,a)\in r} (Q_{retrace}(s, a) - Q_{\theta^Q}(s, a))^2$
    $\theta_{Loss} \leftarrow D_{KL}\left(\pi_m(\cdot \mid s) \mid \pi_{\hat{\theta}}(\cdot \mid s)\right.$
    Update $Q(s, a \mid \theta^Q)$ with gradient $\nabla_{\theta^Q} \theta^Q_{Loss}$
    Update $f(s, a \mid \theta^b)$ with gradient $\nabla_\theta \theta_{Loss}$
  **end for**
  Update target networks:
  $\theta^{Q'} \leftarrow \tau_Q \theta^Q + (1 - \tau_Q)\theta^{Q'}$
  $\theta \leftarrow \tau_\theta \theta^b + (1 - \tau_\theta)\theta$
**end for**