Hinge Policy Optimization: Reinterpreting PPO-Clip and Attaining Global Optimality

Nai-Chieh Huang¹, Ping-Chun Hsieh¹, Kuo-Hao Ho¹, Hsuan-Yu Yao², Kai-Chun Hu¹, Liang-Chun Ouyang¹, I-Chen Wu¹,²
¹Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
²Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
{naich.cs09, pinghsieh}@nycu.edu.tw

Abstract

Policy optimization is a fundamental principle for designing reinforcement learning algorithms, and one example is the proximal policy optimization algorithm with a clipped surrogate objective (PPO-clip), which has been popularly used in deep reinforcement learning due to its simplicity and effectiveness. Despite its superior empirical performance, PPO-clip has not been justified via theoretical proof up to date. This paper proposes to rethink policy optimization based on the principle of hinge policy optimization (HPO), called to achieve policy improvement by solving a large-margin classification problem with hinge loss, and thereby reinterprets PPO-clip as an instance of HPO. From this perspective, we present HPO algorithms and establish the following convergence guarantees: (i) We first study HPO with direct parameterization and establish almost-sure asymptotic convergence in the tabular setting; (ii) We present the NeuralHPO algorithm, which substantiates the HPO principle in the neural function approximation setting via a regression-based policy update scheme, and thereafter characterize the convergence rate of NeuralHPO. To the best of our knowledge, this is the first-ever that can prove global convergence to an optimal policy and convergence rate for a variant of PPO-clip.

1 Introduction

Policy optimization is a well-known technique for solving reinforcement learning problems, which iteratively updates the parameters to optimize the designated objective. Policy gradient methods were introduced as a technique to directly solve policy optimization problems. The goal is to find an optimal policy that maximizes the total expected reward in interactions with an environment [1–3]. The step size is an important hyperparameter that significantly affects the performance of policy gradient algorithms. Due to the difficulty of selecting a proper step size, Trust Region Policy Optimization (TRPO) [4] was formulated to address this issue. TRPO leverages the trust-region method with the second-order approximation to attain strict policy improvement. In contrast to TRPO, which is computationally expensive, Proximal Policy Optimization (PPO) [5] enforces policy improvement only via first-order derivatives. PPO has two variants, PPO-KL and PPO-clip, which augment the objective with the Kullback-Leibler divergence as a penalty and the clipping of probability ratio, respectively. Both of the above approaches are extraordinary in various environments, and PPO shines in possessing better computational efficiency [6–8].

Due to the empirical successes of these policy optimization algorithms, a plethora of recent works makes important progress in terms of their theoretical guarantees. In particular, [9,10] prove the global convergence result of the policy gradient algorithm under several different settings. Moreover, [11] establishes the convergence rates of the softmax policy gradient in both the standard and the entropy-regularized settings. Moreover, it has been shown that various policy gradient algorithms also enjoy global convergence [13–15]. With regard to TRPO and PPO, [16] leverages the mirror
descent method and establishes the convergence rate of adaptive TRPO under the standard and entropy-regularized setting. Moreover, [17] proves the convergence rate of PPO-KL and TRPO under neural function approximation. By contrast, despite that PPO-clip is a computationally efficient and empirically successful method, the following question about the theory of PPO-clip remains largely open: Does PPO-clip enjoy provable global convergence or have any convergence rate guarantee?

To address this question, we propose to reinterpret and generalize the idea of PPO-clip through the lens of Hinge Policy Optimization (HPO) by connecting state-wise policy improvement with solving a large-margin classification problem. Specifically, we cast the process of policy improvement as training a binary classifier with hinge loss via empirical risk minimization. Interestingly, the popular PPO-clip algorithm can be shown to be an instance of HPO with a specific type of classifier. From this perspective, we are able to prove the almost-sure global convergence of HPO in the tabular setting. Moreover, to establish the convergence rate and consider the more practical setting, we propose NeuralHPO and prove the global convergence under neural function approximation. We also characterize the sufficient conditions of global convergence for the family of NeuralHPO algorithms. To the best of our knowledge, our analysis provides the first global convergence guarantee and the convergence rate for PPO-clip. The main contributions of this paper can be summarized as follows:

- We propose HPO, a policy optimization framework where the policy update is built on state-wise policy improvement via hinge loss. The members of the HPO family share a generic loss function, and their differences lie in the classifier and the choice of the margin. We also show that the widely-used PPO-clip algorithm can be viewed as an instance of this family. With this reinterpretation, we establish the asymptotic convergence of PPO-clip in the tabular setting.
- We propose NeuralHPO, which leverages EMDA and a regression-based policy update scheme for neural networks, and thereby provide an insight that decouples the policy search and policy parameterization. This approach enables a generic analytical framework for the HPO family with neural function approximation. We explicitly characterize the convergence rate of NeuralHPO with sufficient conditions, and the results can be easily adapted to various classifiers. We therefore establish the first global convergence result as well as the $O(1/\sqrt{T})$ convergence rate of PPO-clip and provides an affirmative answer to one critical open question about PPO-clip.

2 Preliminaries

Markov Decision Processes. Consider a discounted Markov Decision Process $(S, A, P, R, \gamma, \mu)$, where $S$ is the state space (possibly infinite), $A$ is a finite action space, $P : S \times A \times S \to [0, 1]$ is the transition dynamic of the environment, $R : S \times A \to [0, R_{\text{max}}]$ is the bounded reward function, $\gamma \in (0, 1)$ is the discount factor, and $\mu$ is the initial state distribution. Given a policy $\pi : S \to \Delta(A)$, where $\Delta(A)$ is the unit simplex over $A$, we define the state-action value function $Q^\pi(s, a)$ as

$$Q^\pi(s, a) := \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s, a_0 = a \right].$$

(1)

Moreover, we define $V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot)}[Q^\pi(s, a)]$ and $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$. Also, we denote $\pi^*$ as an optimal policy that attains the maximum total expected reward, and denote $\pi_0$ as the uniform policy. We introduce $\nu_{\pi}(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s | s_0 \sim \mu, \pi)$ as the discounted state visitation distribution induced by $\pi$ and $\sigma_{\pi}(s, a) = \nu_{\pi}(s) \cdot \pi(a|s)$ as the state-action visitation distribution induced by $\pi$. In addition, we define the distribution $\nu^*$ and $\sigma^*$ as the discounted state visitation distribution and the state-action visitation distribution induced by the optimal policy $\pi^*$, respectively. Moreover, we define $\tilde{\sigma}_{\pi} = \nu_{\pi} \rho \pi_0$ as the state-action distribution induced by interactions with the environment through $\pi$, sampling actions from the uniform policy $\pi_0$. We use $\mathbb{E}_{\nu_{\pi}}[\cdot]$ and $\mathbb{E}_{\sigma_{\pi}}[\cdot]$ as the shorthand notations of $\mathbb{E}_{a \sim \nu_{\pi}}[\cdot]$ and $\mathbb{E}_{(s, a) \sim \sigma_{\pi}}[\cdot]$, respectively.

For the convergence property, we define the total expected reward over the state distribution $\nu^*$ as

$$L(\pi) := \mathbb{E}_{\nu^*}[V^\pi(s)].$$

(2)

Here, a maximizer of (2) is equivalent to the original definition of the optimal policy $\pi^*$. We will prove the global convergence by analyzing the difference in $L$ between our policy and the optimal policy and show that the total expected reward monotonically increases.

Proximal Policy Optimization (PPO). PPO is an empirically successful algorithm that achieves monotonic policy improvement by maximizing a surrogate lower bound of the original objective,
either through Kullback-Leibler penalty (termed PPO-KL) or the clipped probability ratio (termed PPO-clip). PPO-KL and PPO-clip are the two variants of PPO. In this paper, our focus is PPO-clip.

Let \( \rho_{s,a}(\theta) \) denote the probability ratio \( \frac{\pi(\alpha_s|s)}{\pi(\alpha_s|s)^{(0)}} \). PPO-clip avoids large policy update by applying a simple heuristic that clips the probability ratio by the clipping range \( \epsilon \) and thereby removes the incentive for moving \( \rho_{s,a}(\theta) \) away from 1. Specifically, the objective of PPO-clip is

\[
L^{\text{clip}}(\theta) = \mathbb{E}_{\sigma_s} \left[ \min \{ \rho_{s,a}(\theta) A^{\pi^{(0)}}(s,a), \text{clip}(\rho_{s,a}(\theta), 1 - \epsilon, 1 + \epsilon) A^{\pi^{(0)}}(s,a) \} \right].
\]

Neural Networks. We present the notations and assumptions about neural networks. Without loss of generality, suppose \((s,a) \in \mathbb{R}^d \) for every \((s,a) \in S \times A\). We use two-layer neural network as \( \text{NN}(\alpha; m) \) to parameterize our policy \( \pi_\theta \) and \( Q \) function. The parameterized function of \( \text{NN}(\alpha; m) \) is

\[
u_{\alpha}(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \sigma([\alpha]_i^T (s,a)),
\]

where \( m \) is the width, \( \alpha = ([\alpha]_1^T, \ldots, [\alpha]_m^T) \in \mathbb{R}^{md} \) is the input weights where \([\alpha]_i \in \mathbb{R}^d\), \( b_i \in \{-1, 1\} \) are the weights of the output, and \( \sigma(\cdot) \) is the Rectified Linear Unit (ReLU) activation function. The initialization of \( \alpha(0) \) and \( b_i \) is as follows:

\[
b_i \sim \text{Unif}\{1, -1\}, [\alpha(0)]_i \sim \mathcal{N}(0, I_d/d),
\]

where both \( b_i \) and \([\alpha(0)]_i \) are i.i.d. for each \( i \in [m] \). We fixed the \( b_i \) after the initialization, we only train for the weights \( \alpha \). Also, to ensure that the local linearization properties hold, we use the projection to restrict the training weights \( \alpha \) in an \( \ell_2 \)-ball centered at \( \alpha(0) \). We denote the ball as \( B_\alpha = \{ \alpha : \|\alpha - \alpha(0)\|_2 \leq R_\alpha \} \) and define the projection as \( \Pi_{B_\alpha}(\alpha') := \arg \min_{\alpha \in B_\alpha} \|\alpha - \alpha'\|_2 \).

Our analysis of neural networks relies on the following assumptions, which are both commonly used regularity conditions for neural networks and neural tangent kernel (NTK):

Assumption 1 (Q-Value Function Class). For any \( R > 0 \), define a function class

\[
F_{R,m} = \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot 1 \{ [\alpha(0)]^T_i (s,a) > 0 \} \cdot [\alpha]_i^T (s,a) : \|\alpha - \alpha(0)\|_2 \leq R \right\},
\]

where \( b_i \) and \( \alpha(0) \) are initialized as [5]. We assume that \( Q^{\pi}(s,a) \in F_{R,m,q} \) for any policy \( \pi \).

Assumption 2 (Regularity of Stationary Distribution). Given any state-action visitation distribution \( \sigma_\pi \), there exists a upper bounding constant \( c > 0 \) for any weight vector \( z \in \mathbb{R}^d \) and \( \zeta > 0 \), such that \( \mathbb{E}_{\sigma_\pi} [1 \{ z^T (s,a) \} \leq \zeta] \| z \|_2 \leq c \cdot \zeta / \| z \|_2 \) holds almost surely.

Since \( T^\pi Q^\pi \) is still a \( Q \) function, Assumption 1 gives us the closeness of our function class under the Bellman operator \( T^\pi \). Assumption 2 states that the distribution \( \sigma_\pi \) is sufficiently regular, and this is required for the analysis of the neural network error. Moreover, Assumption 2 can be interpreted as a variant of commonly-used concentrability-type assumption, under which the state-action visitation distribution always has an upper bounded distribution. The above regularity conditions are standard in the reinforcement learning literature [17][22] for analyses.

Notations: We use \( \langle a,b \rangle \) and \( a \circ b \) to denote the inner product and the Hadamard product, respectively. We use \( 1 \{ \cdot \} \) to denote the indicator function. Let \( I_d \) denote the \( d \times d \) identity matrix.

3 Hinge Policy Optimization

A Sufficient Condition for State-Wise Policy Improvement. To begin with, we first motivate the HPO framework by stating the following lemma about state-wise policy improvement [23][24].

Lemma 1. Given policies \( \pi_1 \) and \( \pi_2 \), \( V^{\pi_1}(s) \geq V^{\pi_2}(s) \) for all \( s \in S \) if the following holds:

\[
(\pi_1(a|s) - \pi_2(a|s)) A^{\pi_2}(s,a) \geq 0, \forall (s,a) \in S \times A.
\]

(7)

Notably, Lemma 1 offers a useful insight that state-wise policy improvement can be achieved by determining the sign of the advantage of each state-action pair (regardless of its magnitude) and adjusting the action probabilities accordingly. More specifically, we can draw an analogy between (7) in Lemma 1 and the training of a linear classifier: (i) Features: The state-action pair serves as the
feature vector of a training sample; (ii) **Labels**: The sign of \( A^{π_2}(s,a) \) plays the role of a binary label; (iii) **Classifiers**: \( π_1(a|s) − π_2(a|s) \) resembles the prediction of a linear classifier. In the next section, we substantiate this insight and present the HPO framework.

**Connecting PPO-clip and Hinge Loss.** In PPO-clip, the policy stops being updated when the probability ratio is out of the clipping range. This behavior coincides with the large-margin classification where the classifier intends to “push” the predicted label out of a margin [25]. Specifically, the gradient of the clipped objective is indeed the negative of the gradient of hinge loss objective \[ i.e., \]

\[
\frac{\partial}{\partial \theta} \min \{\rho_{sa}(\theta) A^+(s,a), \text{clip}(\rho_{sa}(\theta), 1 - \epsilon, 1 + \epsilon) A^+(s,a)\} \\
= - \frac{\partial}{\partial \theta} \left[ A^+(s,a) \right] \ell(\text{sign}(A^+(s,a)), \rho_{sa}(\theta) − 1, \epsilon),
\]

where \( \ell(y_i, f_θ(x_i), \epsilon) \) is the hinge loss defined as \( \max\{0, \epsilon − y_i f_θ(x_i)\} \), \( \epsilon \) is the margin, \( y_i \in \{-1, 1\} \) the label corresponding to the data \( x_i \), and \( f_θ(x_i) \) serves as the binary classifier. Once \( y_i f_θ(x_i) \) is larger than the margin, \( \ell(y_i, f_θ(x_i), \epsilon) \) will equal zero, which reflects the sample clipping mechanism in PPO-clip. Note that hinge loss has been commonly used for large-margin classification, most notably for support vector machines [26]. From the above, in the tabular settings maximizing the objective in (3) can be rewritten as minimizing the following loss:

\[
L(θ) = \sum_{s∈S} d^t_q(s) \sum_{a∈A} \left( π(a|s) A^+(s,a) \right) \cdot \ell(\text{sign}(A^+(s,a)), \rho_{sa}(\theta) − 1, \epsilon).
\]

In practice, we draw a mini-batch \( D \) of state-action pairs and use the sample average to approximately minimize the loss function in (9).

**Hinge Policy Optimization.** Based on the similarity of PPO-clip objective as that to minimize a hinge loss, we propose a generalized family of algorithms called Hinge Policy Optimization (HPO), where the general form of the loss function for policy improvement is

\[
L_{HPO}(θ) = \frac{1}{|D|} \sum_{(s,a)∈D} \text{weight} \times \ell(\text{label}, \text{classifier}, \text{margin}).
\]

Different combinations of classifiers, margins, and weights lead to different loss functions, and hence represents different algorithms in this new HPO family.

### 4 Tabular Case: HPO with Direct Policy Parameterization

In this section, we study the global convergence of HPO with direct parameterization, i.e., policies are parameterized by \( π(a|s) = θ_{s,a}, \) where \( θ_{s,a} ∈ Δ(A) \) denotes the vector \( θ_s, \) and \( θ ∈ Δ(A)^{|S|} \). We use \( V^{(t)}(s) \) and \( A^{(t)}(s,a) \) as the shorthands for \( V^{π^{(t)}}(s) \) and \( A^{π^{(t)}}(s,a) \), respectively.

**HPO with Direct Policy Parameterization.** HPO proceeds iteratively. Define the sample-based loss function of HPO for each iteration \( t \) as

\[
\hat{L}^{(t)}(θ) := \frac{1}{|D^{(t)}|} \sum_{(s,a)∈D^{(t)}} W^{(t)}(s,a) \ell(\text{sign}(A^{(t)}(s,a)), h(ρ^{(t)}_{sa}(θ)), ϵ),
\]

where \( D^{(t)} \) denotes the batch of state-action pairs, \( h(ρ^{(t)}_{sa}(θ)) \) denotes the classifier, and \( W^{(t)}(s,a) ∈ (0, W_{\text{max}}] \) denotes the weight in (10) for each \( (s,a) \). Note that by choosing \( W^{(t)}(s,a) = |A^{(t)}(s,a)| \) and \( h(ρ^{(t)}_{sa}(θ)) = ρ^{(t)}_{sa}(θ) − 1 \), HPO would recover the form of the loss of PPO-clip. HPO with direct parameterization is shown in Algorithm [4]. In each iteration, HPO updates the policy by minimizing the loss in (11) via the entropic mirror descent algorithm (EMDA) [27]. While there are alternative ways to minimize the loss \( \hat{L}^{(t)}(θ) \) over \( Δ(A)^{|S|} \) (e.g., the projected subgradient method), we leverage EMDA for the following two reasons: (i) HPO achieves policy improvement by increasing or decreasing the probability of those state-action pairs in \( D^{(t)} \) based on the sign of \( A^{(t)}(s,a) \) as well as properly reallocating the probabilities of those state-action pairs not contained in the batch (to ensure the probability sum is one). Using EMDA allows us to enforce a proper

---

1 Please see Appendix [4] for the detailed comparison of the two objectives.
Algorithm 1: HPO

1. **Output:** Learned policy $\pi^{(\infty)}$;
2. Initialize policy $\pi^{(t)} = \pi^{(0)}$, initial state distribution $\mu$, $\alpha \in (0, 1)$, step size of EMDA $\eta$, number of EMDA iterations $K^{(t)}$;
3. for $t = 0, 1, \cdots$ do
4. Collect a set of trajectories $\tau \in D^{(t)}$ under policy $\pi^{(t)} = \pi^{(0)}$;
5. Find $A^{(t)}$ by a policy evaluation method;
6. Compute $L^{(t)}(\theta)$ based on $A^{(t)}$ and the collected samples in $D^{(t)}$;
7. Update the policy by $\theta^{(t+1)} = \text{EMDA}(L^{(t)}(\theta), \eta, K^{(t)}, D^{(t)}, \theta^{(t)})$
8. end

Algorithm 2: EMDA($L(\theta), \eta, K, D, \theta_{\text{init}}$)

1. **Output:** Learned parameter $\tilde{\theta}$;
2. **Input:** Objective $L(\theta)$, step size $\eta$, number of iteration $K$, dataset $D$, and initial parameter $\theta_{\text{init}}$;
3. Initialize $\theta^{(0)} = \theta_{\text{init}}$, $\tilde{\theta} = \theta_{\text{init}}$;
4. for $k = 0, \cdots, K - 1$ do
5. for each state $s$ in $D$ do
6. Find $\hat{g}_{s,a}^{(k)} = \frac{\partial L(\theta)}{\partial \theta_{s,a}}_{\theta=\hat{\theta}^{(k)}}$ for each $a$;
7. Let $w_{s} = (e^{-\eta g_{s,a}^{(k)}}, \cdots, e^{-\eta g_{s,a}^{(k)}|A|})$;
8. $\tilde{\theta}_{s}^{(k+1)} = \frac{1}{\langle w_{s}, \tilde{\theta}^{(k)} \rangle}(w_{s} \circ \tilde{\theta}_{s}^{(k)})$;
9. end
10. end

Global Convergence of PPO-Clip With Direct Parameterization. We use PPO-clip as a canonical example to present the analysis of HPO algorithms. We first make the following assumptions.

**Assumption 3** (Infinite Visitation to Each State-Action Pair). Each state-action pair $(s, a)$ appears infinitely often in $\{D^{(t)}\}$, i.e., $\lim_{t \to \infty} \sum_{t=0}^{t} 1 \{ (s, a) \in D^{(t)} \} = \infty$, with probability one.

**Assumption 4.** In each iteration, the state-action pairs in $D^{(t)}$ have distinct states.

Assumption 3 resembles the standard infinite-exploration condition commonly used in the temporal-difference methods, such as Sarsa [28]. Assumption 4 is rather mild: (i) This can be met by post-processing the mini-batch of state-action pairs via an additional sub-sampling step; (ii) In most RL problems with discrete actions, the state space is typically much larger than the action space.

**Theorem 1** (Global Convergence of PPO-clip). Under PPO-clip, we have $V^{(t)}(s) \to V^{*}(s)$ as $t \to \infty$, $\forall s \in S$, with probability one.

The proof of Theorem 1 is provided in Appendix B. We highlight the main ideas behind the proof of Theorem 1: (i) State-wise policy improvement: Through the lens of HPO, we show that PPO-clip enjoys state-wise policy improvement in every iteration with the help of the EMDA subroutine. This property greatly facilitates the rest of the convergence analysis. (ii) Quantifying the probabilities of those actions with positive or negative advantages in the limit: By (i), we know the limits of the value functions and the advantage function all exist. Then, we proceed to show that the actions with positive advantage in the limit cannot exist by establishing a contradiction.

The above also manifests how reinterpreting PPO-clip through HPO helps with establishing the convergence guarantee of PPO-clip.

Global Convergence of HPO With Alternative Classifiers. Notably, the HPO update scheme and the convergence analysis of Theorem 1 can be readily extended to various other classifiers. Due to space limitation, the theoretical results of HPO with alternative classifiers are given in Appendix C.

5 Function Approximation: NeuralHPO

5.1 From Tabular HPO to NeuralHPO: Neural Approximate Policy Search

In this section, we build on our policy improvement framework which substantiates the HPO principle in the neural function approximation setting. For high-level intuition, this framework first searches
for the policy in the policy space and then approximates it in the parameter space. In particular, in each iteration, we do a policy search by EMDA as the tabular HPO for finitely many iterations (i.e., $K^{(t)} \leq K$, for all $t$) and obtain the next-step policy $\hat{\pi}_{t+1}$ as the improved target. Recall that the improved target is a direct parameterization policy in the discrete policy space composed of the states of the sample trajectory. With a proper improved target, we then utilize the representation power of neural networks to approximate the improved policy in the parameter space via a regression-based policy update scheme. We will discuss more details in the next section.

**Remark 1.** Most gradient-based policy optimization methods have the policy search and policy improvement entangled because of taking the gradient of the objective directly. In contrast, our framework decouples these two subroutines. The advantage of this framework is that it facilitates the analysis. Although we can directly take a closed-form optimal solution to the policy improvement objective as the ideal improved target, such a closed-form optimal solution of an arbitrary objective function does not always exist. In particular, the loss of PPO-clip is an example that does not have a simple closed-form optimal solution. In view of this, EMDA, which enjoys closed-form updates, makes our framework more analytically flexible. Moreover, as this framework is not tied to properties specific to HPO, we could apply this analytical framework to various policy improvement objectives other than the loss function of HPO.

### 5.2 NeuralHPO

**Parameterization Setting.** We parameterize our policy at each iteration $t$ as an energy-based policy $\pi_{\theta_t}(a|s) \propto \exp\{\tau_t^{-1} f_{\theta_t}(s, a)\}$, where $\tau_t$ is the temperature parameter and $f_{\theta_t}(s, a) = \text{NN}(\theta_t; m_f)$ are the energy functions. Similarly, we parameterize our state-action value function $Q_{\omega_t}(s, a) = \text{NN}(\omega_t; m_Q)$. We define $V_{\omega}(s)$ as the value function obtained by using the Bellman Equation. Moreover, we define $A_{\omega}(s, a) := Q_{\omega}(s, a) - V_{\omega}(s)$.

**Policy Improvement.** According to the framework presented in Section 5.1, we first give the closed-form of the EMDA improved target of our NeuralHPO. The detailed proof is in Appendix D.

**Proposition 1 (EMDA Improved Target).** For the improved target obtained by the EMDA subroutine at the $t$-th iteration, we have

$$\log \hat{\pi}_{t+1}(a|s) \propto C_t(s, a) \cdot A_{\omega_t}(s, a) + \tau_t^{-1} f_{\theta_t}(s, a),$$

where $C_t(s, a) = -\sum_{k=0}^{K^{(t)}-1} \eta g_{k, a}$. 

Recall that the improved target $\hat{\pi}$ is the direct parameterization in the policy space, but our policy $\pi_{\theta}$ is a softmax policy that is proportional to the exponentiated energy function. This explains why we consider the $\log \hat{\pi}_{t+1}(a|s)$ in Proposition 1. Another benefit of using EMDA is that it closely matches the energy-based policies considered in NeuralHPO due to the inherent exponentiated gradient update.

Then, we discuss the details of the neural function approximation of our policy. After obtaining the improved target by Proposition 1, we solve the Mean Squared Error (MSE) subproblem with respect to $\theta$ to approximate the improved target as follows:

$$\mathbb{E}_{\tilde{\sigma}_t}[(f_\theta(s, a) - \tau_{t+1} \cdot (C_t(s, a) \cdot A_{\omega_t}(s, a) + \tau_t^{-1} f_{\theta_t}(s, a))^2].$$

Notice that we consider the state-action distribution $\tilde{\sigma}_t$ which samples the action through a uniform policy $\pi_0$. In this manner, we use more exploratory data to improve our current policy. In particular, we use the SGD to tackle the above subproblem, and the pseudo code is provided in Appendix A.

**Policy Evaluation.** To evaluate $Q$, we use a neural network to approximate the true state-action value function $Q^{\pi_0}$ by solving the Mean Square Bellman Error (MSBE) subproblem. The MSBE subproblem is to minimize the following objective with respect to $\omega$ at each iteration $t$:

$$\mathbb{E}_{\sigma_t}[(Q_\omega(s, a) - |T^{\pi_0} Q_\omega|(s, a))^2],$$

where $T^{\pi_0}$ is the Bellman operator of policy $\pi_{\theta_t}$ such that

$$[T^{\pi_0} Q_\omega](s, a) = \mathbb{E}[r(s, a) + \gamma Q_\omega(s', a')|s' \sim \mathcal{P}(|s, a), a' \sim \pi_{\theta_t}(|s')].$$

The pseudo code of neural TD update for state-action value function $Q_\omega$ is in Appendix A.

It is worth mentioning that NeuralHPO is not a fully on-policy algorithm. Although we interact with the environment by our current policy, for policy improvement, we sample the actions by the uniform policy $\pi_0$. We provide the pseudo code of NeuralHPO as the following Algorithm 5. The pseudo code of Algorithms 4-6 is in Appendix A.
Algorithm 3: NeuralHPO

Input: MDP \((S, A, \mathcal{P}, r, \gamma, \mu)\), Objective function \(L\), EMDA step size \(\eta\), number of EMDA iteration \(K\), number of SGD and TD update iterations \(T_{\text{upd}}\), number of HPO iterations \(T\). the clipping range \(\epsilon\)

1. Initialize the policy \(\pi_0\) as a uniform policy

2. for \(t = 0, \ldots, T - 1\) do
   3. Set temperature parameter \(\tau_{t+1}\)
   4. Sample the tuple \(\{s_i, a_i, s'_i, a'_i\}_i^{T_{\text{upd}}}\), where \((s_i, a_i) \sim \sigma_t, a'_i \sim \pi_0(\cdot | s_i), s'_i \sim \mathcal{P}(\cdot | s_i, a_i)\) and \(a'_i \sim \pi_{\theta_t}(\cdot | s'_i)\)
   5. Use the states of the trajectories with nonzero advantage as the sample batch \(\{s_i\}_i^{T_{\text{upd}}}\) to solve the tabular EMDA as Algorithm 4 and obtain the return policy \(\pi_{t+1}\) and \(C_t\)
   6. Solve for \(Q_{\omega_t} = NN(\omega_t; \theta_t)\) by using TD update as Algorithm 5
   7. Use the Bellman expectation equation to obtain \(V_{\omega_t}\) and the advantage \(A_{\omega_t} = Q_{\omega_t} - V_{\omega_t}\)
   8. Solve for \(f_{\theta_{t+1}} = NN(\theta_{t+1}; m_f)\) by using SGD as Algorithm 6 based on the EMDA result
   9. Update the policy \(\pi_{\theta_{t+1}} \propto \exp\{\tau_{t+1} f_{\theta_{t+1}}\}\)

end

6 Convergence Guarantee of NeuralHPO

In this section, we present the convergence analysis of NeuralHPO. Inspired by the analysis of [17], we analyze the convergence behavior of NeuralHPO based on the NTK technique. Nevertheless, NeuralHPO presents several unique technical challenges in establishing the convergence: (i) Tight coupling between function approximation error and the clipping behavior: The clipping mechanism can be viewed as an indicator function. The function approximation for advantage would significantly influence the value of the indicator function in a highly complex manner. As a result, handling the error between the neural approximated advantage and the true advantage serves as one major challenge in the analysis; (ii) Lack of a closed-form expression of policy update: Due to the clipping function in the HPO objective and the iterative updates in the EMDA subroutine, the new policy does not have a simple closed-form expression. This is one salient difference between the analysis of NeuralHPO and other neural algorithms (cf. [17]); (iii) NTK technique on advantage function: Another technicality is that the advantage function requires the NTK projection and linearization properties to characterize the approximation error. However, since we use the neural network to approximate the state-action value function instead of the advantage function, it requires additional effort to establish the error bound of the advantage function. Throughout this section, we suppose Assumptions 4 and 5 hold.

Given that we need to analyze the error between our approximation and the true function, we further define the improved target under the true advantage function \(A^{\pi_t}\) as \(\tau_{t+1}(a|s) := C_t(s, a) A^{\pi_t}(s, a) + \tau_{t+1}^{-1} f_{\theta_t}(s, a)\), where \(C_t(s, a)\) is the \(C_t(s, a)\) obtained under \(A^{\pi_t}\). Moreover, all the expectations about \(A_c\) throughout the analysis are with respect to the randomness of the neural network initialization. Below we state the convergence rate and the sufficient condition of NeuralHPO, which is also the main theorem of our paper.

Theorem 2 (Convergence Rate of NeuralHPO). If the following sufficient conditions hold,

\begin{align}
(i) & \quad L_C \cdot |A^t(s, a)| \leq \hat{C}_t(s, a) \cdot |A^t(s, a)| \leq U_C \cdot |A^t(s, a)| \quad \text{for all } t, \tag{16} \\
(ii) & \quad L_C = \omega(T^{-1}), U_C = O(T^{-1/2}) \tag{17}
\end{align}

where \(L_C, U_C > 0\) are finite values dependent on \(T\). Then, the policy sequence \(\{\pi_{\theta_t}\}_{t=0}^{T}\) obtained by NeuralHPO satisfies

\begin{equation}
\min_{0 \leq t \leq T} \{L(C^* - L(\pi_{\theta_t}))\} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\epsilon_t + \epsilon_t') + TU_C^2(\sqrt{2}\psi_{t+1}^* + M)}{TL_C(1 - \gamma)}, \tag{18}
\end{equation}

where \(\epsilon_t = C_{\infty}T_{t+1}\phi^2_t \epsilon_t^{1/2} + Y^{1/2}\psi_{t+1}^2 \epsilon_t^{1/2}, \epsilon_t' = |A|C_{\infty}T_{t+1}\epsilon_t^{1/2}, M = 4E_{\psi_t} \max_{a}(Q_{\omega_0}(s, a))^2 + 4R_{\psi_t}^2, \text{ and } Y = 2M + 2R_{\max}(1 - \gamma)^2 - E_t^t\).

Here, we have the \(O((TL_C)^{-1})\) convergence rate of NeuralHPO.
Then, the following holds,

\[
L^{(t)}(\theta) = \mathbb{E}_{\pi^*}[\langle \pi_\theta(\cdot|s), |A^{\pi^*}(s, \cdot)| \circ \ell(\text{sign}(A^{\pi^*}(s, \cdot)), \frac{\pi^*(s, \cdot)}{\pi_\theta(\cdot|s)} - 1, \epsilon)\rangle]
\]

where the above fractions are the Radon–Nikodym Derivatives. Also, we define the concentrability.

Assumption 5

Hence, we provide the \( O \) convergence rate of the suboptimality gap between the optimal policy \( \pi^* \) and our policy \( \pi_\theta \) in the proof of Theorem 2. The detailed proof of Corollary 1 is in Appendix F, where we also provide the convergence rate of NeuralHPO with other classifiers. In the sequel, we present the supporting lemmas and then the proof of Theorem 2. The detailed proofs of all the supporting lemmas are in Appendix E. Throughout this section, we assume \( L_C, U_C \) satisfy the condition in (16), which specifies a lower bound and an upper bound of \( C_t \).

Before we jump into the supporting lemmas, we need the assumption about distribution density for analyses. Several research works [17, 24] have the concentrability assumption, we also have this common regularity condition which is stated as follows:

Assumption 5 (Concentrability Coefficient and Density Ratio). We define the density ratio between the policy-induced distributions and the policies,

\[
\phi_t^* = \mathbb{E}_{\pi^*}\left[\left(\frac{d\pi^*}{d\pi_t} - \frac{d\pi_{\theta_t}}{d\pi_t}\right)^2\right]^{1/2}, \quad \psi_t^* = \mathbb{E}_{\pi^*}\left[\left(\frac{d\pi^*}{d\pi_t} - \frac{d\nu_t}{d\pi_t}\right)^2\right]^{1/2},
\]

(21)

where the above fractions are the Radon–Nikodym Derivatives. Also, we define the concentrability coefficient \( C_{\infty} \), which is the upper bound of the density ratio between the optimal state distribution and any state distribution, i.e. \( \|\nu^*/\nu\|_{\infty} < C_{\infty} \) for any \( \nu \). We assume that these density differences \( \phi_t^*, \psi_t^* \) and the concentrability coefficient \( C_{\infty} \) are bounded.

Lemma 2 (Error Propagation). Let \( \pi_{t+1} \) be the improved target obtained by EMDA with the true advantage. Suppose the policy improvement error satisfies

\[
\mathbb{E}_{\pi^*_t}\left[(f_{\theta_{t+1}}(s, a) - \tau_{t+1} \cdot (C_t(s, a) \cdot A_{\omega_t}(s, a) + \tau_{t}^{-1} f_{\theta_t}(s, a)))^2\right] \leq \epsilon_t + 1,
\]

(22)

and the policy evaluation error satisfies

\[
\mathbb{E}_{\pi_t^*}\left[(A_{\omega_t}(s, a) - A^{\pi^*}(s, a))^2\right] \leq \epsilon_t',
\]

(23)

Then, the following holds,

\[
|\mathbb{E}_{\nu^*}\left[|\log \pi_{t+1}(\cdot|s) - \log \pi_{t+1}(\cdot|s), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s)|\right]| \leq \epsilon_t + \epsilon_{err}
\]

(24)

where \( \epsilon_t = C_{\infty} \tau_{t+1}^{-1} \phi_t^{1/2} + U_C X^{1/2} \psi_t^{1/2} \) and \( \epsilon_{err} = \sqrt{2} U_C \epsilon_{err} \psi_t^* \), and \( X = \left[2/\epsilon_{err}^2(M + (R_{max}/(1 - \gamma))^2 - \epsilon_t')/2\right] \), and \( M = 4\mathbb{E}_{\pi_t}[\max_{a}(Q_{\omega_t}(s, a))]^2 + 4R_f \).

The above \( \epsilon_t, \epsilon_t' \) in (22) and (23) are the policy improvement and policy evaluation error and can always be small by properly choosing the width of the neural networks. Please see Appendix E for more details. We continue to use the conditions in Lemma 2 and present the following.

Lemma 3 (Stepwise Energy Difference).

\[
\mathbb{E}_{\nu^*}\left[|\tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - \tau_{t}^{-1} f_{\theta_t}(s, \cdot)|_{\infty}^2\right] \leq 2\epsilon_t' + 2U_C^2 M,
\]

(25)

where \( \epsilon_t' = |A| \cdot C_{\infty} \tau_{t+1}^{-2} \epsilon_t + M = 4\mathbb{E}_{\nu^*}[\max_{a}(Q_{\omega_t}(s, a))]^2 + 4R_f \).

After showing the lemmas about errors, we give the following lemmas for the suboptimality gap. Roughly speaking, we use the KL divergence difference as the potential function to analyze the suboptimality gap between the optimal policy \( \pi^* \) and our policy \( \pi_{\theta_t} \) in the proof of Theorem 2.
Lemma 4 (Stepwise KL Difference). The KL difference is as follows,
\[
\begin{align*}
KL(\pi^*(s)\|\pi_{\theta_{t+1}}(s)) - KL(\pi^*(s)\|\pi_{\theta_t}(s)) \\
\leq \langle \log \pi_{\theta_{t+1}}(s) - \log \pi_{\theta_{t}}(s), \pi_{\theta_t}(s) - \pi^*(s) \rangle - \langle \bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s) \rangle \\
- \frac{1}{2} \|\pi_{\theta_{t+1}}(s) - \pi_{\theta_t}(s)\|^2_s - \langle \log \pi_{\theta_{t+1}}(s) - \log \pi_{\theta_t}(s), \pi_{\theta_t}(s) - \pi_{\theta_{t+1}}(s) \rangle 
\end{align*}
\]
(26)

Lemma 5 (Performance Difference Using Advantage). Recall that \(\mathcal{L}(\pi) = E_{\nu^*}[\mathcal{V}(\pi)]\). We have
\[
\mathcal{L}(\pi^*) - \mathcal{L}(\pi) = (1 - \gamma)^{-1} \cdot E_{\nu^*}[\langle A^\pi(s,\cdot), \pi^*(s) - \pi(\cdot)\rangle].
\]
(28)

We now have all the supporting lemmas. Here, we provide the proof sketch of Theorem 2. Please refer to Appendix B for the detailed proof.

Proof Sketch of Theorem 2. By taking expectation over \(\nu^*\) on the difference between KL in Lemma 4, the following holds by using the Hölder’s inequality and the fact that \(2xy - x^2 \leq y^2\),
\[
\begin{align*}
E_{\nu^*}[KL(\pi^*(s)\|\pi_{\theta_{t+1}}(s)) - KL(\pi^*(s)\|\pi_{\theta_t}(s))] \\
\leq \varepsilon_t + \varepsilon_{err} - \frac{E_{\nu^*}[(\bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s))]}{2} \\
- \frac{1}{2} \|\pi_{\theta_{t+1}}(s) - \pi_{\theta_t}(s)\|^2_s - \frac{E_{\nu^*}[(\bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s))]}{2} \\
\leq \varepsilon_t + \varepsilon_{err} - \frac{E_{\nu^*}[(\bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s))]}{2} \\
+ \frac{1}{2} \|\pi_{\theta_{t+1}}(s) - \pi_{\theta_t}(s)\|^2_s.
\end{align*}
\]
(29)

By the Lemma 3 and rearranging the terms, we obtain that
\[
\begin{align*}
E_{\nu^*}[\langle \bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s) \rangle] \\
\leq E_{\nu^*}[KL(\pi^*(s)\|\pi_{\theta_t}(s))] - KL(\pi^*(s)\|\pi_{\theta_{t+1}}(s))] + \varepsilon_t + \varepsilon_{err} + \varepsilon_t' + U_C^2 M.
\end{align*}
\]
(32)

By (16), we have \(L_C E_{\nu^*}[(A^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s))] \leq E_{\nu^*}[(\bar{C}_t(s,\cdot) \circ \mathcal{A}^{\theta_t}(s,\cdot), \pi^*(s) - \pi_{\theta_t}(s))]\). By Lemma 5 and taking the telescoping sum of (16) from \(t = 0\) to \(T - 1\), we obtain
\[
\begin{align*}
(1 - \gamma)L_C \sum_{t=0}^{T-1} (\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})) \\
\leq E_{\nu^*}[KL(\pi^*(s)\|\pi_{\theta_t}(s))] - KL(\pi^*(s)\|\pi_{\theta_{T}}(s))] + \sum_{t=0}^{T-1} (\varepsilon_t + \varepsilon_{err} + \varepsilon_t') + MTU_C^2.
\end{align*}
\]
(34)

By the facts that (i) \(E_{\nu^*}[KL(\pi^*(s)\|\pi_{\theta_t}(s))] \leq \log |\mathcal{A}|\), (ii) KL divergence is nonnegative, (iii) \(\sum_{t=0}^{T-1} (\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})) \geq T \cdot \min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\}\), we have
\[
\min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\} \leq \log |\mathcal{A}| + \sum_{t=0}^{T-1} (\varepsilon_t + \varepsilon_t') + T(\varepsilon_{err} + MU_C^2). 
\]
(35)

Since we have \(\varepsilon_{err} = \sqrt{2U_C^2} \varepsilon_{err}\pi_t^*,\) we set \(\varepsilon_{err} = U_C^2\) and thereby obtain the result of (18). \(\square\)

It is worth mentioning that \(\varepsilon_t, \varepsilon_t'\) can be arbitrarily small by setting a large neural network width and a sufficiently large \(T_{\text{ upd}}\) for neural network updates.

Remark 2 (EMDA Step Size and Clipping Mechanism). Notably, according to the original idea of PPO, the KL penalty parameter \(\beta\) of PPO-KL somehow connects to the clipping range \(c\) of PPO-clip. Interestingly, through our analysis, we find that the EMDA step size \(\eta\) effectively plays the role of the KL penalty parameter \(\beta\) in NeuralHPO.

7 Concluding Remarks

In this paper, we rethink the policy optimization methods and propose a classification-based policy update scheme leveraging hinge loss for policy improvement, and thereby propose a new family of algorithms called Hinge Policy Optimization (HPO). This work proves the almost-sure convergence of HPO in the tabular setting and characterizes the convergence rate under neural function approximation. One limitation of this work is that our analysis is built on NTK, which requires a large neural network width. In practice, PPO-clip does not require a large neural network, and one interesting future work is to bridge the gap between the NTK theory and the empirical success of PPO-clip.
References

[1] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[2] 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.

[3] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

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

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

[6] Gang Chen, Yiming Peng, and Mengjie Zhang. An adaptive clipping approach for proximal policy optimization. *arXiv preprint arXiv:1804.06461*, 2018.

[7] Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6672–6679, 2020.

[8] Ju-Seung Byun, Byungmoon Kim, and Huamin Wang. Proximal policy gradient: Ppo with policy gradient. *arXiv preprint arXiv:2010.09933*, 2020.

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

[10] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

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

[12] 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.

[13] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[14] 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, 2020.

[15] Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pages 10772–10782. PMLR, 2021.

[16] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

[17] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32:10565–10576, 2019.

[18] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

[19] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[20] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.
[21] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1): 4809–4874, 2016.

[22] Samuele Tosatto, Matteo Pirotta, Carlo d’Eramo, and Marcello Restelli. Boosted fitted q-iteration. In *International Conference on Machine Learning*, pages 3434–3443. PMLR, 2017.

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

[24] Kai-Chun Hu, Ping-Chun Hsieh, Ting Han Wei, and I-Chen Wu. Rethinking deep policy gradients via state-wise policy improvement. In *"I Can’t Believe It’s Not Better!" NeurIPS 2020 workshop*, 2020.

[25] Chen-Huan Pi, Kai-Chun Hu, Stone Cheng, and I-Chen Wu. Low-level autonomous control and tracking of quadrotor using reinforcement learning. *Control Engineering Practice*, 95:104222, 2020.

[26] Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

[27] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[28] 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.

[29] Michail G Lagoudakis and Ronald Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *International Conference on Machine Learning*, pages 424–431, 2003.

[30] Alessandro Lazaric, Mohammad Ghavamzadeh, and Remi Munos. Analysis of a classification-based policy iteration algorithm. In *International Conference on Machine Learning*, pages 607–614, 2010.

[31] Amir-massoud Farahmand, Doina Precup, André Barreto, and Mohammad Ghavamzadeh. Classification-based approximate policy iteration: Experiments and extended discussions. *arXiv preprint arXiv:1407.0449*, 2014.
Appendix

A Pseudo Code of Algorithms

Remark 3. In the NeuralHPO algorithms, the choices of the EMDA step size \( \eta \) and the temperature parameters \( \{ \tau_i \} \) of the neural networks are important factors to the convergence rate and hence shall be configured properly for different algorithms in the HPO family. As a result, we do not specify the specific choices of \( \eta \) and \( \{ \tau_i \} \) in the following pseudo code of the generic NeuralHPO. Please refer to Corollaries 1-2 in Appendix F for the choices of \( \eta \) and \( \{ \tau_i \} \) for Neural PPO-clip and another variant of NeuralHPO.

Algorithm 4: EMDA

**Input:** Objective function \( L(\theta) \), EMDA step size parameter \( \eta \), number of EMDA iteration \( K \), initial policy \( \pi_{\theta_0} \), sample batch \( \{ s_i \}_{i=1}^{T_{\text{upd}}} \)

1. Initialize \( \theta(0) = \theta_0 \), \( C_i(s,a) = 0 \), for all \( s,a \)
2. for \( k = 0, \ldots, K-1 \) do
   3. for each state \( s \) in the batch do
      4. Find \( g(s,a) = \frac{\partial L(\theta)}{\partial \theta} |_{\theta=\theta(k)} \), for each \( a \)
      5. Let \( w_a = (e^{-\eta g_{s,a}}, \ldots, e^{-\eta g_{s,A}}) \)
      6. \( \hat{\theta}(k+1) = \frac{1}{|w_a|} (w_a \circ \hat{\theta}(k)) \)
      7. \( C_i(s,a) \leftarrow C_i(s,a) - \eta g_{s,a} / A_w(s,a) \), for each \( a \) with \( A_w(s,a) \neq 0 \)
   8. end
9. \( \pi_{\theta_{k+1}} = \hat{\theta}(K) \)
10. **Output:** Return the policy \( \pi_{\theta_{k+1}} \), and \( C_i \)

Algorithm 5: Policy Evaluation via TD

**Input:** \( (S,A,P,r,\gamma) \), initial weights \( b_i \), \( [\omega(0)]_i \ (i \in [m_{\omega}]) \), number of iterations \( T_{\text{upd}} \), sample \( \{ s_i, a_i, s_i', a_i \}_{i=1}^{T_{\text{upd}}} \)

1. Set the step size \( \eta_{\text{upd}} \leftarrow T_{\text{upd}}^{-1/2} \)
2. for \( t = 0, \ldots, T_{\text{upd}} - 1 \) do
   3. \( (s,a,s',a') \leftarrow (s_i, a_i, s_i', a_i') \)
   4. \( \omega(t+1) / 2 \leftarrow \omega(t) - \eta_{\text{upd}} \cdot \gamma Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a') \cdot \nabla_\omega Q_{\omega(t)}(s,a) \)
   5. \( \omega(t+1) \leftarrow \arg \min_{\omega \in b_i^{(R_{Q_i})}} \{ ||\omega - \omega(t+1)||_2 \} \)
6. end
7. Take the average over path \( \omega \leftarrow 1 / T_{\text{upd}} \cdot \sum_{t=0}^{T_{\text{upd}}-1} \omega(t) \)
8. **Output:** \( Q_{\omega} \)

Algorithm 6: Policy Improvement via SGD

**Input:** \( (S,A,P,r,\gamma) \), the current energy function \( f_{\theta_i} \), initial weights \( b_i \), \( [\theta(0)]_i \ (i \in [m_{\theta_i}]) \), number of iterations \( T_{\text{upd}} \), sample \( \{ s_i, a_i \}_{i=1}^{T_{\text{upd}}} \)

1. Set the step size \( \eta_{\text{upd}} \leftarrow T_{\text{upd}}^{-1/2} \)
2. for \( t = 0, \ldots, T_{\text{upd}} - 1 \) do
   3. \( (s,a) \leftarrow (s_i, a_i) \)
   4. \( \theta(t+1) / 2 \leftarrow \theta(t) - \eta_{\text{upd}} \cdot \nabla_\theta \cdot f_{\theta_i}(s,a) - \gamma C_{\omega_i}(s,a) \cdot A_{\omega_i}(s,a) + \tau_{t+1}^{-1} f_{\theta_i}(s,a) \) \( \cdot \nabla_\theta f_{\theta_i}(s,a) \)
   5. \( \theta(t+1) \leftarrow \arg \min_{\theta \in b_i^{(R_{f_i})}} \{ ||\theta - \theta(t+1)||_2 \} \)
6. end
7. Take the average over path \( \theta \leftarrow 1 / T_{\text{upd}} \cdot \sum_{t=0}^{T_{\text{upd}}-1} \theta(t) \)
8. **Output:** \( f_{\theta} \)
B Proof of Theorem 1

B.1 Supporting Lemmas for the Proof of Theorem 1

As described in Section 3, one major component of the proof of Theorem 1 is the state-wise policy improvement property of HPO. For ease of exposition, we introduce the following definition regarding the partial ordering over policies.

Definition 1 (Partial ordering over policies). Let π1 and π2 be two policies. Then, π1 ⪰ π2, called π1 improves upon π2, if and only if V π1 (s) ≥ V π2 (s), ∀s ∈ S. Moreover, we say π1 > π2, called π1 strictly improves upon π2, if and only if π1 ⪰ π2 and there exists at least one state s such that V π1 (s) > V π2 (s).

Lemma 6 (Sufficient condition of state-wise policy improvement). Given any two policies π1 and π2, we have π1 ⪰ π2 if the following condition holds:

\[ \sum_{a \in A} \pi_1(a|s) A^{t+1}(s,a) \geq 0, \forall s \in S. \]  \hfill (36)

Proof of Lemma 6 This is a direct result of the performance difference lemma [23]. □

Next, we present two critical properties that hold under HPO for every sample path.

Lemma 7 (Strict improvement and strict positivity of policy under HPO with direct tabular parameterization). In any iteration t, suppose π(t) is strictly positive in all state-action pairs, i.e., π(t)(a|s) > 0, for all (s, a). Under HPO in Algorithm 1, π(t+1) satisfies that (i) π(t+1) > π(t) and (ii) π(t+1)(a|s) > 0, for all (s, a).

Proof of Lemma 7 Consider the t-th iteration of HPO (cf. Algorithm 1) and the corresponding update from π(t) to π(t+1). Regarding (ii), recall from Algorithm 1 that K(t) denotes the number of iterations undergone by the EMDA subroutine for the update from π(t) to π(t+1) and that K(t) is designed to be finite. Therefore, it is easy to verify that π(t+1)(a|s) > 0 for all (s, a) by the exponentiated gradient update scheme of EMDA and the strict positivity of π(t).

Next, for ease of exposition, for each k ∈ {0, 1, · · · , K(t)} and for each state-action pair (s, a), let  \( \theta_{s,a}^{(k)} \) denote the policy parameter after k EMDA iterations. Regarding (i), recall that we define  \( g_{s,a}^{(k)} := \frac{\partial C(\theta)}{\partial \theta_{s,a}} \bigg|_{\theta = \theta_{s,a}^{(k)}} \). Let  \( w_s^{(k)} := (\exp(-\eta g_{s,a}^{(k)}), \cdots, \exp(-\eta g_{s,a}^{(k)})) \). Note that as the weights in the loss function of HPO only affects the effective step sizes of EMDA, we simply set the weights of HPO-AM to be one, without loss of generality. By EMDA in Algorithm 2 for every (s, a) ∈ D(t), we have

\[ \pi^{(t+1)}(a|s) = \frac{\prod_{k=0}^{K(t)-1} \exp(-\eta g_{s,a}^{(k)})}{\prod_{k=0}^{K(t)-1} w_s^{(k)}, \theta_{s,a}^{(k)}} \cdot \pi^{(t)}(a|s). \]  \hfill (37)

Note that  \( g_{s,a}^{(k)} \) can be written as

\[ g_{s,a}^{(k)} = \begin{cases} \frac{1}{\pi^{(t)}(a|s)} \text{sign}(A^{(t)}(s,a)) & \text{if } \left( \frac{\hat{\theta}_{s,a}^{(k)}}{\pi^{(t)}(a|s)} - 1 \right) \text{sign}(A^{(t)}(s,a)) < \epsilon, (s,a) \in D^{(t)} \text{, otherwise} \end{cases} \]  \hfill (38)

By (37)–(38), it is easy to verify that for those (s, a) ∈ D(t) with positive advantage, we must have  \( \prod_{k=0}^{K(t)-1} \exp(-\eta g_{s,a}^{(k)}) > 1. \) Similarly, for those (s, a) ∈ D(t) with negative advantage, we have  \( \prod_{k=0}^{K(t)-1} \exp(-\eta g_{s,a}^{(k)}) < 1. \) Now we are ready to check the condition of strict policy improvement given by Lemma 6. For each s ∈ S, we have

\[ \sum_{a \in A} \pi^{(t+1)}(a|s) A^{(t)}(a|s) = \frac{1}{\prod_{k=0}^{K(t)-1} w_s^{(k)}, \theta_{s,a}^{(k)}} \sum_{a \in A} \left( \prod_{k=0}^{K(t)-1} \exp(-\eta g_{s,a}^{(k)}) \right) \pi^{(t)}(a|s) A^{(t)}(a|s) > 0. \]  \hfill (39)

Hence, we conclude that under HPO, the strict state-wise policy improvement property indeed holds, i.e., π(t+1) > π(t). □
Note that Lemma 7 implies that the limits \( V^{(\infty)}(s) \), \( Q^{(\infty)}(s, a) \), \( A^{(\infty)}(s, a) \) exist for every sample path. By the strict policy improvement shown in Lemma 7, we know that the sequence of state values is point-wise monotone increasing, i.e., \( V^{(\ell+1)}(s) \geq V^{(\ell)}(s) \), \( \forall s \in S \). Moreover, by the bounded reward and the discounted setting, we have \( \frac{-R_{\text{max}}}{1-\gamma} \leq V^{(\ell)}(s) \leq \frac{R_{\text{max}}}{1-\gamma} \). The above monotone increasing property and boundedness imply convergence, i.e., \( V^{(\ell)}(s) \to V^{(\infty)}(s) \), for each sample path. Similarly, we also know that \( Q^{(\ell)}(s, a) \to Q^{(\infty)}(s, a) \), and thus \( A^{(\ell)}(s, a) \to A^{(\infty)}(s, a) \). As a result, we can define the three sets \( I^+_s \), \( I^0_s \) and \( I^-_s \) as
\[
I^+_s := \{ a \in A | A^{(\infty)}(s, a) > 0 \}, \\
I^0_s := \{ a \in A | A^{(\infty)}(s, a) = 0 \}, \\
I^-_s := \{ a \in A | A^{(\infty)}(s, a) < 0 \}.
\]
Note that for each sample path, the sets \( I^+_s \), \( I^0_s \) and \( I^-_s \) are well-defined, based on the limit \( A^{(\infty)}(s, a) \).

**Lemma 8.** Conditioned on the event that each state-action pair occurs infinitely often in \( \{D^{(t)}\} \), under HPO, if \( I^+_s \) is not an empty set, then we have \( \sum_{a \in I^-_s} \pi^{(t)}(a|s) \to 0 \), as \( t \to \infty \).

**Proof of Lemma 8.** We discuss each state separately as it is sufficient to show that for each state \( s \), given some fixed \( a' \in I^+_s \), for any \( a'' \in I^-_s \), we have \( \pi^{(t)}(a''|s) \to 0 \), as \( t \to \infty \). For ease of exposition, we reuse some of the notations from the proof of Lemma 7. Recall that we let \( K^{(t)} \) denote the number of iterations undergone by the EMDA subroutine for the update from \( \pi^{(t)} \) to \( \pi^{(t+1)} \), and \( K^{(t)} \) is designed to be finite. For each \( k \in \{0, 1, \cdots, K^{(t)}\} \) and for each state-action pair \( (s, a) \), let \( \theta_{s,a}^{(k)} \) denote the policy parameter after \( k \) EMDA iterations. Recall from Algorithm 2 that \( g_{s,a}^{(0)} := \frac{\partial L_{\theta}^{(0)}(\theta = \theta_{s,a})}{\partial \theta_{s,a}} \) and \( w_{s,a}^{(k)} := (e^{-\eta g_{s,a}^{(0)}}, \cdots, e^{-\eta g_{s,a}^{(k)}}) \). Define \( \Delta_s := \min_{a \in I^+_s \cup I^-_s} |A^{(\infty)}(s, a)| > 0 \) (and here \( \Delta_s \) is a random variable as \( A^{(\infty)}(s, a) \) is defined with respect to each sample path). By the definition of \( I^+_s \), \( I^-_s \) and \( \Delta_s \), we know that for each sample path, there must exist finite \( T_+ \) and \( T_- \) such that: (i) for every \( a \in I^+_s \), \( A^{(t)}(s, a) \geq \frac{\Delta_s}{2} \), for all \( t > T_+ \), and (ii) for every \( a \in I^-_s \), \( A^{(t)}(s, a) \leq -\frac{\Delta_s}{2} \), for all \( t > T_- \). Under Assumption 4, at each HPO iteration \( t \) with \( t > \max\{T_+, T_-\} \), there are three possible cases regarding the state-action pairs \( (s, a') \) and \( (s, a'') \):

- **Case 1:** \((s, a') \in D^{(t)} \), \((s, a'') \notin D^{(t)} \).
  By the EMDA subroutine and (37), we have
  \[
  \frac{\pi^{(t+1)}(a''|s)}{\pi^{(t+1)}(a'|s)} = \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \prod_{k=0}^{K^{(t)}-1} \exp(\eta g_{s,a'}^{(k)}) \leq \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \exp(-\eta),
  \]
  where the last inequality holds by (38), \( a' \in I^+_s \), and \( \pi^{(t)}(a'|s) \leq 1 \).

- **Case 2:** \((s, a') \notin D^{(t)} \), \((s, a'') \in D^{(t)} \).
  By the EMDA subroutine, we have \( -g_{s,a'}^{(0)} < 0 \) and \( -g_{s,a''}^{(k)} \leq 0 \) for all \( k \in \{1, \cdots, K^{(t)}\} \). Therefore, we have
  \[
  \frac{\pi^{(t+1)}(a''|s)}{\pi^{(t+1)}(a'|s)} < \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)}.
  \]

- **Case 3:** \((s, a') \notin D^{(t)} \), \((s, a'') \notin D^{(t)} \).
  Under EMDA, as neither \((s, a')\) nor \((s, a'')\) is in \( \notin D^{(t)} \), the action probability ratio between these two actions remains unchanged (despite that the values of \( \pi^{(t)}(a''|s) \) and \( \pi^{(t+1)}(a''|s) \) can still change if there is an action \( a''' \) such that \( a''' \neq a' \), \( a''' \neq a'' \), and \( (s, a''') \in D^{(t)} \), i.e.,
  \[
  \frac{\pi^{(t+1)}(a''|s)}{\pi^{(t+1)}(a'|s)} = \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)}.
  \]

Conditioned on the event that each state-action pair occurs infinitely often in \( \{D^{(t)}\} \), we know Case 1 and 3 must occur infinitely often. By (43)-(45), we conclude that \( \pi^{(t)}(a''|s) \to 0 \), as \( t \to \infty \), for every \( a'' \in I^-_s \).

\qed
Lemma 9. Conditioned on the event that each state-action pair occurs infinitely often in \( \{D^{(t)}\} \), under HPO, if \( I^+_s \) is not an empty set, then there exists some constant \( c > 0 \) such that \( \sum_{a \in I^-_s} \pi^{(t)}(a|s) \geq c \), for infinitely many \( t \).

Proof of Lemma 9. For each \((s, a)\), define \( T_{s,a} := \{ t : (s, a) \in D^{(t)} \} \) to be the index set that collects the time indices at which \((s, a)\) is contained in the mini-batch. Given that each state-action pair occurs infinitely often, we know \( T_{s,a} \) is a (countably) infinite set.

For ease of exposition, define a positive constant \( \chi \) as

\[
\chi := \frac{e \cdot \eta}{e \cdot \eta + 1} < 1.
\]

Define \( \Delta := \min_{a \in I^+_s} A^{(\infty)}(s, a) > 0 \) (and here \( \Delta \) is a random variable as \( A^{(\infty)}(s, a) \) is defined with respect to each sample path). By the definition of \( I^+_s \) and \( \Delta \), we know that there must exist a finite \( T^{(\ast)} \) such that for every \( a \in I^+_s \), \( A^{(t)}(s, a) \geq \frac{\Delta}{e} \), for all \( t > T^{(\ast)} \). Similarly, by the definition of \( I^0_s \), there must exist a finite \( T^{(0)} \) such that for every \( a \in I^0_s \), \( |A^{(t)}(s, a)| \leq \frac{\Delta}{e} \), for all \( t > T^{(0)} \). We also define \( T^{(\ast)} := \max\{T^{(\ast)}, T^{(0)}\} \).

We reuse some of the notations from the proof of Lemma 7. Recall that we let \( K^{(t)} \) denote the number of iterations undergone by the EMDA subroutine for the update from \( \pi^{(t)} \) to \( \pi^{(t+1)} \), and \( K^{(t)} \) is a finite positive integer. For ease of exposition, for each \( k \in \{0, 1, \cdots, K^{(t)}\} \) and for each state-action pair \((s, a)\), let \( \theta^{(k)}(s, a) \) denote the policy parameter after \( k \) EMDA iterations. Recall that we define \( g_s^{(k)} := \frac{\partial L^{(0)}}{\partial g_s} |_{\theta = \theta^{(0)}} \) and \( u_s^{(k)} := (e^{-u_s^{(k)}}, \cdots, e^{-u_s^{(k)}}) \). If \( I^+_s \) is not an empty set, then we can select an arbitrary action \( a' \in I^+_s \). For any \( t \) with \( t > T^{(\ast)} \) and \( t \in T_{s,a} \), by (37) we have

\[
\pi^{(t+1)}(a'|s) = \prod_{k=0}^{K^{(t)}-1} \frac{\prod_{k=0}^{K^{(t)}-1} \exp(-\eta g_s^{(k)} \cdot \pi^{(t)}(a'|s))}{\prod_{k=0}^{K^{(t)}-1} \exp(-\eta g_s^{(k)})} \cdot \pi^{(t)}(a'|s)
\]

(47)

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(-\eta g_s^{(0)})}{\pi^{(t)}(a'|s) \exp(-\eta g_s^{(0)}) + 1}
\]

(48)

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(\eta \pi^{(t)}(a'|s))}{\pi^{(t)}(a'|s) \exp(\eta \pi^{(t)}(a'|s)) + 1}
\]

(49)

\[
\geq \frac{e \cdot \eta}{e \cdot \eta + 1} = \chi,
\]

(50)

where (48) holds due to the fact that \( \frac{\eta}{\pi^{(t)}(a'|s)} \) is non-decreasing with \( k \) under Assumption 4. (49) follows from (38) and that \( a' \in I^+_s \), and (50) holds by the fact that \( q(z) = z \cdot \exp(\eta / z) \) has a unique minimizer at \( z = \eta \) with minimum value \( e \cdot \eta \). For all \( t \) satisfies \( (t-1) \in T_{s,a} \) and \( t > T^{(\ast)} \), we have

\[
\sum_{a \in I^+_s} \pi^{(t)}(a|s) \geq \sum_{a \in I^+_s} \pi^{(t)}(a|s) A^{(t)}(s, a) + \sum_{a \in I^0_s} \pi^{(t)}(a|s) A^{(t)}(s, a)
\]

\[
\max_{a \in I^-_s} |A^{(t)}(s, a)|
\]

(51)

\[
\geq \chi \frac{3\Delta}{4} - 1 \cdot \chi \frac{\Delta}{4}
\]

(52)

\[
= \chi \frac{\Delta}{2R_{\max} - 1 / \gamma}
\]

(53)

where (51) follows from that \( \sum_{a \in A} \pi^{(t)}(a|s) = 0 \) and \( A^{(t)}(s, a) < 0 \) for all \( a \in I^-_s \), and (52) follows from the definition of \( T^{(\ast)}, T^{(0)} \) as well as the boundedness of rewards. Since \( T_{s,a} \) is a countably infinite set, we know \( \sum_{a \in I^-_s} \pi^{(t)}(a|s) \geq \chi \frac{\Delta}{2R_{\max}} \), for infinitely many \( t \).
B.2 Proof of Theorem 1

Now we are ready to show Theorem 1. For ease of exposition, we restate Theorem 1 as follows.

**Theorem (Global Convergence of PPO-clip).** Under PPO-clip, we have \( V^{(t)}(s) \rightarrow V^*(s) \) as \( t \rightarrow \infty \), \( \forall s \in S \), with probability one.

**Proof.** We establish that \( \pi^{(t)} \) converges to an optimal policy by showing that \( I^*_s \) is an empty set for all \( s \). Under Assumption 3, the analysis below is presumed to be conditioned on the event that each state-action pair occurs infinitely often in \( \{ D^{(t)} \} \). The proof proceeds by contradiction as follows: Suppose \( I^*_s \) is non-empty. From Lemma 8, we have that \( \sum_{a \in I^*_s} \pi^{(t)}(a|s) \rightarrow 0, \) as \( t \rightarrow \infty \). However, Lemma 9 suggests that there exists some constant \( c > 0 \) such that \( \sum_{a \in I^*_s} \pi^{(t)}(a|s) \geq c \) infinitely often. This leads to a contraction, and thus completes the proof.

\[ \square \]

C Global Convergence of Tabular HPO With Alternative Classifiers

**Theorem 3.** Theorem 1 also holds under the following algorithms: (i) HPO with the classifier \( \log(\pi_0(a|s)) - \log(\pi(a|s)) \) (termed HPO-log); (ii) HPO with the classifier \( \sqrt{p_{s,a}(\theta)} - 1 \) (termed HPO-root).

**Proof of Theorem 3.** We show that Theorem 1 can be extended to these two alternative classifiers by following the proof procedure of Theorem 1. Specifically, we extend the supporting lemmas (cf. Lemma 7, Lemma 8, and Lemma 9) as follows:

- To extend Lemma 7 to the alternative classifiers, we can reuse (37) and rewrite (54) for each classifier. That is, for HPO-log, we have
  \[
  g^{(k)}_{s,a} = \begin{cases} 
  \frac{-1}{\theta_{s,a}^{(k)}} \text{sign}(A^{(t)}(s,a)), & \text{if } \log \left( \frac{\theta_{s,a}^{(t)}}{\pi^{(t)}(a|s)} \right) \text{sign}(A^{(t)}(s,a)) < \epsilon, (s,a) \in D^{(t)} \\
  0, & \text{otherwise}
  \end{cases}
  \]

  (54)

  On the other hand, for HPO-root, we have
  \[
  g^{(k)}_{s,a} = \begin{cases} 
  \frac{-1}{2\theta_{s,a}^{(k)} \pi^{(t)}(a|s)}} \text{sign}(A^{(t)}(s,a)), & \text{if } \left( \sqrt{\frac{\theta_{s,a}^{(t)}}{\pi^{(t)}(a|s)}} - 1 \right) \text{sign}(A^{(t)}(s,a)) < \epsilon, (s,a) \in D^{(t)} \\
  0, & \text{otherwise}
  \end{cases}
  \]

  (55)

  As the sign of \( g^{(k)}_{s,a} \) depends only on the sign of the advantage, it is easy to verify that (39) still goes through and hence the sufficient condition of Lemma 8 is satisfied under these two alternative classifiers. Moreover, by using the same argument of EMDA as that in Lemma 7, it is easy to verify that \( \pi^{(t+1)}(a|s) > 0 \) for all \( (s,a) \).

- Regarding Lemma 8, we can extend this result again by considering the three cases as in Lemma 8. For Case 1, given the \( g^{(k)}_{s,a} \) in (54) and (55), we have: For HPO-log,
  \[
  \frac{\pi^{(t+1)}(a''|s)}{\pi^{(t+1)}(a'|s)} = \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \prod_{k=0}^{K^{(t)}-1} \exp(\eta g^{(k)}_{s,a'}) \leq \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \frac{\exp(-\eta)}{<1}. 
  \]

  Similarly, for HPO-root, we have
  \[
  \frac{\pi^{(t+1)}(a''|s)}{\pi^{(t+1)}(a'|s)} = \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \prod_{k=0}^{K^{(t)}-1} \exp(\eta g^{(k)}_{s,a'}) \leq \frac{\pi^{(t)}(a''|s)}{\pi^{(t)}(a'|s)} \cdot \frac{\exp(-\eta)}{<1}. 
  \]

  (56)

  (57)

  Moreover, it is easy to verify that the arguments in Case 2 and Case 3 still hold under these two alternative classifiers. Hence, Lemma 8 still holds.

16
• Regarding Lemma 9, we can reuse all the setup and slightly revise (47)-(50) for the two alternative classifiers: For HPO-log, by (54), we have

\[
\pi^{(t+1)}(a'|s) = \prod_{k=0}^{K(t)-1} \exp(-\eta g^{(k)}_{s,a'}) / \prod_{k=0}^{K(t)-1} \langle w^k_s, \theta^k_s \rangle \cdot \pi^{(t)}(a'|s)
\]

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(-\eta g^{(0)}_{s,a'})}{\pi^{(t)}(a'|s) \exp(-\eta g^{(0)}_{s,a'}) + 1}
\]

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(\eta / \pi^{(t)}(a'|s))}{\pi^{(t)}(a'|s) \exp(\eta / \pi^{(t)}(a'|s)) + 1}
\]

\[
\geq e \cdot \eta + 1.
\]

Similarly, for HPO-root, by (55), we have

\[
\pi^{(t+1)}(a'|s) = \prod_{k=0}^{K(t)-1} \exp(-\eta g^{(k)}_{s,a'}) / \prod_{k=0}^{K(t)-1} \langle w^k_s, \theta^k_s \rangle \cdot \pi^{(t)}(a'|s)
\]

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(-\eta g^{(0)}_{s,a'})}{\pi^{(t)}(a'|s) \exp(-\eta g^{(0)}_{s,a'}) + 1}
\]

\[
\geq \frac{\pi^{(t)}(a'|s) \exp(\eta / 2 \pi^{(t)}(a'|s))}{\pi^{(t)}(a'|s) \exp(\eta / 2 \pi^{(t)}(a'|s)) + 1}
\]

\[
\geq e \cdot \eta + 1.
\]

Accordingly, (51)-(53) still go through and hence Lemma 9 indeed holds under HPO-log and HPO-root.

In summary, since all the supporting lemmas hold for these alternative classifiers, we complete this part of the proof by obtaining a contradiction similar to that in Theorem 1.

As a case study, the convergence behavior of three variants of HPO with direct parameterization (with a constant margin \(e = 0.1\)) in a 4 \times 4 gridworld environment is provided as follows. Under all the three variants, the policy is trained for 40000 episodes under 10 random seeds. The source code can be found in an anonymous repository.

![Figure 1: L1-norm between the policies obtained by HPO and the optimal policy.](https://anonymous.4open.science/r/NeurIPS2022-2A8C)
D Proof of Proposition 1

We expand the closed-form of the log of the EMDA improved target,

\[
\log \pi_{t+1}(a|s) = \log \left( \prod_{k=0}^{K_{t}-1} \frac{\exp(-ng_{s,a}^{(k)})}{\langle w_s, \hat{g}^{(k)} \rangle} \cdot \pi_{\theta_t}(a|s) \right)
\]

(66)

\[
= \sum_{k=0}^{K_{t}-1} -ng_{s,a}^{(k)} - \sum_{k=0}^{K_{t}-1} \log(\langle w_s, \hat{g}^{(k)} \rangle) + \log \pi_{\theta_t}(a|s)
\]

(67)

\[
= \sum_{k=0}^{K_{t}-1} -ng_{s,a}^{(k)} - \sum_{k=0}^{K_{t}-1} \log(\langle w_s, \hat{g}^{(k)} \rangle) + \tau_{t}^{-1} f_{\theta_t}(s,a) - \log(Z_t(s))
\]

(68)

\[
\propto C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_{t}^{-1} f_{\theta_t}(s,a).
\]

(69)

where \(Z_t(s)\) is the normalizing factor of the policy at step \(t\). Since both the \(\sum_{k=0}^{K_{t}-1} \log(\langle w_s, \hat{g}^{(k)} \rangle)\) and \(\log(Z_t(s))\) are state-dependent, we can cancel it under softmax policy. We obtain \(C_t(s,a)\) from Algorithm 4 and complete the proof.

\( \square \)

E Proof of the Supporting Lemmas for Theorem 2

E.1 Additional Supporting Lemmas

Throughout this section, we slightly abuse the notation that we use \(E_{\text{init}}[\cdot]\) to denote the expectation over the initialization of neural networks. Also, we assume that Assumption 1 and 2 hold in the following proofs.

**Lemma 10** (Policy Evaluation Error). The output \(A_{\omega_t} = Q_{\omega_t} - V_{\omega_t}\) of Algorithm 5 and Bellman expectation equation satisfies

\[
E_{\text{init},\pi_t}(A_{\omega_t}(s,a) - A_{\tau^{\pi_t}}(s,a))^2 = O(R^2_{Q} T^{1/2}_{\text{upd}} + R^{5/2}_{Q} m_Q^{-1/4} + R^{3}_{Q} m_Q^{-1/2}).
\]

(70)

To prove Lemma 10, we start by stating a bound on the error of the estimated state-action value function.

**Lemma 11** (Theorem 4.6 in [17]). The output \(Q_{\omega_t}\) of Algorithm 5 satisfies

\[
E_{\text{init},\pi_t}(Q_{\omega_t}(s,a) - Q_{\pi_t}(s,a))^2 = O(R^2_{Q} T^{1/2}_{\text{upd}} + R^{5/2}_{Q} m_Q^{-1/4} + R^{3}_{Q} m_Q^{-1/2}).
\]

(71)

**Proof of Lemma 10** We are ready to show the policy evaluation error of the advantage function. First, we find the bound of \(|A_{\omega_t}(s,a) - A_{\tau^{\pi_t}}(s,a)|\). We have

\[
|A_{\omega_t}(s,a) - A_{\tau^{\pi_t}}(s,a)| = |Q_{\omega_t}(s,a) - V_{\omega_t}(s) - Q_{\pi_t}(s,a) + V_{\pi_t}(s)|
\]

(72)

\[
= |Q_{\omega_t}(s,a) - Q_{\pi_t}(s,a) + \sum_{a'} \pi_{\theta_t}(a'|s) \cdot (Q_{\pi_t}(s,a') - Q_{\omega_t}(s,a'))|
\]

(73)

\[
\leq |Q_{\omega_t}(s,a) - Q_{\pi_t}(s,a) + E_{a' \sim \pi_t}[Q_{\pi_t}(s,a') - Q_{\omega_t}(s,a')]| + |E_{a' \sim \pi_t}[Q_{\pi_t}(s,a') - Q_{\omega_t}(s,a')]|.
\]

(74)

Then, we can derive the bound of \((A_{\tau^{\pi_t}}(s,a) - A_{\omega_t}(s,a))^2\) as follows,

\[
(A_{\tau^{\pi_t}}(s,a) - A_{\omega_t}(s,a))^2 \leq 2(Q_{\tau^{\pi_t}}(s,a) - Q_{\omega_t}(s,a))^2 + 2|\mathbb{E}_{a' \sim \pi_t}[Q_{\tau^{\pi_t}}(s,a') - Q_{\omega_t}(s,a')]|^2
\]

(76)

\[
\leq 2(Q_{\tau^{\pi_t}}(s,a) - Q_{\omega_t}(s,a))^2 + 2\mathbb{E}_{a' \sim \pi_t}[Q_{\tau^{\pi_t}}(s,a') - Q_{\omega_t}(s,a')]^2
\]

(77)
where (77) holds by Jensen’s inequality. By taking the expectation of (76)-(77) over the state-action distribution \( \sigma_t \), we have

\[
\mathbb{E}_{\sigma_t}[(A^{\pi_\tau}(s, a) - A_{\omega_i}(s, a))^2] 
\leq 2 \mathbb{E}_{\sigma_t}[(Q^{\pi_\tau}(s, a) - Q_{\omega_i}(s, a))^2] + 2 \mathbb{E}_{\sigma_t}[\mathbb{E}_{\sigma_{t'-\tau}}[(Q^{\pi_\tau}(s, a') - Q_{\omega_i}(s, a'))^2]] 
= 4 \mathbb{E}_{\sigma_t}[(Q^{\pi_\tau}(s, a) - Q_{\omega_i}(s, a))^2],
\]

where the last equality in (80) is obtained by the actions are directly sampled by \( \pi_{\theta_t} \), so we can ignore it in the latter term. Last, we leverage Lemma 11 to obtain the result of Lemma 10. \( \square \)

**Lemma 12** (Policy Improvement Error). The output \( f_\theta \) of Algorithm 6 satisfies

\[
\mathbb{E}_{init, \sigma_t}[(f_\theta(s, a) - \tau_{t+1} \cdot (C_t(s, a) \cdot A_{\omega_i}(s, a) + \tau_t^{-1} f_{\theta_t}(s, a)))^2] 
= O(R_1^2 T_{udp}^{-1/2} + R_f^5 m_j^{-1/4} + R_j^3 m_j^{-1/2}),
\]

To prove Lemma 12, we first state the following useful result originally proposed by [17].

**Theorem 4** (17). Meta-Algorithm of Neural Networks. Consider a meta-algorithm with the following update:

\[
\begin{align*}
\alpha(t + 1) &\leftarrow \prod_{B_\alpha}(\alpha(t + 1/2)) = \text{arg min}_{\alpha \in B_\alpha} \| \alpha - \alpha(t + 1/2) \|_2, \\
\alpha(t + 1/2) &\leftarrow \alpha(t) - \eta_{udp} \cdot (u_{\alpha(t)}(s, a) - v(s, a) - \mu \cdot u_{\alpha(t)}(s', a')) \cdot \nabla_\alpha u_{\alpha(t)}(s, a),
\end{align*}
\]

where \( \mu \in [0, 1) \) is a constant, \( (s, a, s', a') \) is sampled from some stationary distribution \( d, u_{\alpha} \) is parameterized as a two-layer neural network \( NN(\alpha; m) \), and \( v(s, a) \) satisfies

\[
\mathbb{E}_d[(v(s, a))^2] \leq \bar{v}_1 \cdot \mathbb{E}_d[(u_{\alpha(0)}(s, a))^2] + \bar{v}_2 \cdot R_u^2 + \bar{v}_3,
\]

for some constants \( \bar{v}_1, \bar{v}_2, \bar{v}_3 \geq 0 \). We define the update operator \( T u(s, a) = \mathbb{E}[v(s, a) + \mu \cdot u(s', a')|s' \sim P(\cdot|s), a' \sim \pi(\cdot|s)] \), and define \( \alpha^* \) as the approximate stationary point (cf. (D.18) in [17]), which inherently have the property \( u_{\alpha^*} = \prod_{t=n}^{T} T u_{\alpha^*} \), where \( u_{\alpha^*} \) is the linearization of \( u \) at \( \alpha^* \). Suppose we run the above meta-algorithm in (82)-(83) for \( T \) iterations with \( T \geq 64/(1-\mu)^2 \) and set the step size \( \eta_{udp} = T^{-1/2} \). Then, we have

\[
\begin{align*}
\mathbb{E}_{init, \sigma_t}[(u_{\alpha^*}(s, a) - u_{\alpha_t}(s, a))^2] &\leq O(R_u^2 T_{udp}^{-1/2} + R_f^5 m_j^{-1/4} + R_j^3 m_j^{-1/2}), \\
\mathbb{E}_{init, \sigma_t}[(u_{\alpha_t}(s, a) - u_{\alpha^*}(s, a))^2] &\leq O(R_u^3 m_j^{-1/2}),
\end{align*}
\]

where \( \bar{\alpha} := 1/T \cdot (\sum_{t=0}^{T-1} \alpha(t)) \) and \( \alpha^* \) is a parameter in \( B_\alpha \).

**Proof of Lemma 12** Now we are ready to prove Lemma 12 as follows. To begin with, (82)-(83) match the policy improvement update of NeuralHPO if we put \( u(s, a) = f(s, a), v(s, a) = \tau_{t+1} (C_t(s, a) \cdot A_{\omega_i}(s, a) + \tau_t^{-1} f_{\theta_t}(s, a)), \mu = 0, d = \bar{\sigma}_t, \) and \( R_u = R_f \). For \( \mathbb{E}_{\bar{\sigma}_t}[(v(s, a))^2] \), we have

\[
\begin{align*}
\mathbb{E}_{\bar{\sigma}_t}[(v(s, a))^2] &\leq 2 \tau_{t+1}^2 (U_C^2 \cdot \mathbb{E}_{\bar{\sigma}_t}[(A_{\omega_i}(s, a))^2] + \tau_t^{-2} \mathbb{E}_{\bar{\sigma}_t}[(f_{\theta_t}(s, a))^2]) \\
&\leq 20 \mathbb{E}_{\bar{\sigma}_t}[(f_{\theta_0}(s, a))^2] + 20 R_f^2.
\end{align*}
\]

Here, since \( C_t \) and \( \bar{C}_t \) are dependent only on the EDMA step size \( \eta \) and the indicator function that depends on the sign of the advantage (either under the true advantage \( A^{\pi^*} \) or the approximated advantage \( A_{\omega_i} \)), one can always find one common upper bound \( U_C \) for both \( C_t \) and \( \bar{C}_t \). In particular, as shown in Corollary 1, we set \( U_C = \sum_{k=0}^{K-1} \eta \) for PPO-clip, which is independent from the advantage function. The inequality in (88) holds by the condition that \( \tau_{t+1}^2 (U_C^2 + \tau_t^{-2}) \leq 1, (a + b)^2 \leq 2a^2 + 2b^2, \mathbb{E}_{\bar{\sigma}_t}[(A_{\omega_i}(s, a))^2] \leq 4 \mathbb{E}_{\bar{\sigma}_t}[(Q_{\omega_i}(s, a))^2], \) and \( \mathbb{E}_{\bar{\sigma}_t}[(u_{\alpha_t}(s, a))^2] \leq 2 \mathbb{E}_{\bar{\sigma}_t}[(u_{\alpha_0}(s, a))^2] + 2 R_f^2 \) which holds by using the Lipschitz property of neural networks where \( u_{\alpha} = f_{\theta_t}, A_{\omega} \). The condition \( \tau_{t+1}^2 (U_C^2 + \tau_t^{-2}) \leq 1 \) can be satisfied by configuring proper \( \{\tau_t\} \), as described momentarily in Appendix 1. We also use that \( \mathbb{E}_{\bar{\sigma}_t}[Q_{\omega_0}] = \mathbb{E}_{\bar{\sigma}_t}[f_{\theta_0}] \) because they share the same initialization. Thus, we have \( v_1 = v_2 = 20 \) and \( v_3 = 0 \) in (84).
Due to that $\theta^*$ is the approximate stationary point, we have $f_{\theta^*}^0 = \mathbb{E}_{d \sim \tau_1(C_t \circ A_{\omega_t} + \tau_t^{-1} f_{\theta_t})}$. Thus,
\[
f_{\theta^*}^0 = \arg \min_{f \in \mathcal{F}_{\tau_1,m_f}} \| f - \tau_1(C_t \circ A_{\omega_t} + \tau_t^{-1} f_{\theta_t}) \|_{2,\sigma_t},
\]
where $\| \cdot \|_{2,\sigma_t} = \mathbb{E}_{\epsilon_1 \sim \sigma_t} [\| \cdot \|_2]^{1/2}$ is the $\sigma_t$-weighted $L_2$-norm. Then, by the fact that $\tau_{t+1}(C_t(s,a) \cdot A^\omega_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a)) \in \mathcal{F}_{\tau_1,m_f}$ and that $A^\omega_{\omega_t}(s,a) = Q^\omega_{\omega_t}(s,a) - \sum_{a \in A} \pi(a|s)Q^\omega_{\omega_t}(s,a)$, we obtain
\[
\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - \tau_{t+1}(C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a))^2] \leq \mathbb{E}_{\epsilon_1 \sim \sigma_t} [(\tau_{t+1}(C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a)) - (\tau_{t+1}(C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a)))^2]
\]
\[
\leq 2\tau_{t+1} U^\omega_{\epsilon_1}(s,a) = \sum_{a' \in A} \pi(a'|s)Q^\omega_{\omega_t}(s,a') - (Q_{\omega_t}(s,a) - \sum_{a' \in A} \pi(a'|s)Q_{\omega_t}(s,a'))^2
\]
\[
+ 2\tau_{t+1} \tau_t^{-2} \mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - f_{\theta_t}(s,a))^2] \leq 8\tau_{t+1} U^\omega_{\epsilon_1}(s,a) - Q_{\omega_t}(s,a) + 2\tau_{t+1} \tau_t^{-2} \mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - f_{\theta_t}(s,a))^2] = O(R^3 m_f^{-1/2}).
\]
We obtain (93) as the same reason in (76)-(80) in the proof of Lemma 10. The terms in (93) are both the designated form as the (86), we leverage the (86) in Theorem 4 and obtain the result in (94). Last, we bound the error of our policy improvement, we have
\[
\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - \tau_{t+1} \cdot (C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a)))^2]
\]
\[
\leq 2\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - f_{\theta_t}(s,a))^2] + 2\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(f_{\theta}^0(s,a) - \tau_{t+1}(C_t(s,a) \cdot A_{\omega_t}(s,a) + \tau_t^{-1} f_{\theta_t}(s,a)))^2]
\]
\[
= O(R^2 T_{\text{upd}}^{-1/2} + R^5 m_f^{-1/4} + R^3 m_f^{-1/2}),
\]
where (96) is bounded as $O(R^2 T_{\text{upd}}^{-1/2} + R^5 m_f^{-1/4} + R^3 m_f^{-1/2})$ by (85) of Theorem 4 and (97) is bounded as $O(R^3 m_f^{-1/2})$ by the derivation of (94). Thus, we obtain (98) and complete the proof.

**Lemma 13 (Error Probability of Advantage).** Given the policy $\pi_{\theta_t}$, the probability of the event that the advantage error is greater than $\epsilon_{err}$ can be bounded as
\[
\mathbb{P}(|A_{\omega_t}(s,a) - A_{\pi_{\theta_t}}(s,a)| > \epsilon_{err}) \leq \frac{\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(A_{\omega_t}(s,a) - A_{\pi_{\theta_t}}(s,a))^2]}{\epsilon_{err}^2}.
\]

**Proof of Lemma 13** By applying Markov’s inequality, we have
\[
\mathbb{P}(|A_{\omega_t}(s,a) - A_{\pi_{\theta_t}}(s,a)| > \epsilon_{err}) = \mathbb{P}(|A_{\omega_t}(s,a) - A_{\pi_{\theta_t}}(s,a)|^2 > \epsilon_{err}^2) \leq \frac{\mathbb{E}_{\epsilon_1 \sim \sigma_t} [(A_{\omega_t}(s,a) - A_{\pi_{\theta_t}}(s,a))^2]}{\epsilon_{err}^2}.
\]
Notice that the randomness of the above event in (99) comes from the state-action visitation distribution $\sigma_t$ and the initialization of the neural networks.

**E.2 Proof of Lemma 2**

For ease of exposition, we restate Lemma 2 as follows. In the following, we slightly abuse the notations $\mathbb{E}_{\epsilon_1}, \mathbb{E}_{\sigma_t}$, and $\mathbb{E}_{\nu_{\epsilon_t}}$ to denote the expectations (over the respective distribution) conditioned on the policy $\pi_{\theta_t}$.
Lemma (Error Propagation). Let $\pi_{t+1}$ be the improved target obtained by EMDA with the true advantage. Suppose the policy improvement error satisfies

$$\mathbb{E}_{\tau_0}[(f_{\theta_{t+1}}(s, a) - \tau_{t+1} \cdot (C_t(s, a) \cdot A_{\omega_t}(s, a) + \tau^{-1}_t f_\theta(s, a)))^2] \leq \varepsilon_{t+1},$$  

(102)

and the policy evaluation error satisfies

$$\mathbb{E}_{\tau_0}[(A_{\omega_t}(s, a) - A^{\pi_{t+1}}(s, a))^2] \leq \varepsilon_t',$$

(103)

Then, the following holds,

$$\mathbb{E}_{\nu^*}[(\log \pi_{t+1}(\cdot|s) - \log \pi_{t+1}(\cdot|s), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))] \leq \varepsilon_t + \varepsilon_{err}$$

(104)

where $\varepsilon_t = C_\infty \tau_{t+1} g_{t+1}^{1/2} + U_C X^{1/2} \psi_t^{1/2}$ and $\varepsilon_{err} = \sqrt{2U_C \varepsilon_{err} \psi_t}$, and $X = [(2/\varepsilon_{err})/(M + (R_{\text{max}}/(1-\gamma))^2 - \varepsilon_t/2)]$, and $M' = 4\mathbb{E}_{\nu^t}[\max_a(Q_{\omega_t}(s, a))^2] + 4R_\infty^2$.

Remark 4. Notice that $\varepsilon_{t+1}$ in (102) and $\varepsilon_t'$ in (103) can be controlled by the width of neural networks and the number of iteration for each SGD and TD updates based on Lemma 10 and 12. Therefore, $\varepsilon_t$ could be made sufficiently small per our requirement.

Proof of Lemma. For ease of exposition, let us first fix a policy $\pi_{\theta_t}$. Through the analysis, we will show that one can derive an upper bound (in the form of (104)) that holds regardless of the policy $\pi_{\theta_t}$. Recall that $C_t(s, a) = - \sum_{k=0}^{k_{\text{max}}(t)} g_{t+1}^{(k)}$, where $g_{t+1}$ is obtained in the EMDA subroutine and depends on the sign of the estimated advantage $A_{\omega_t}$. Similarly, we define $C_t(s, a)$ as the counterpart of $C_t(s, a)$ by replacing $A_{\omega_t}$ with the true advantage $A^{\pi_{t+1}}$. We first simplify $(\log \pi_{\theta_{t+1}}(\cdot|s) - \log \pi_{t+1}(\cdot|s), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))$. The normalizing factor $Z$ of the policies $\pi_{\theta_{t+1}}$ and $\pi_{t+1}$ is state-dependent, and the inner product between any state-dependent function and the policy difference $\pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s)$ is always zero. Thus, we have

$$\langle \log \pi_{\theta_{t+1}}(\cdot|s) - \log \pi_{t+1}(\cdot|s), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle$$

(105)

$$= \langle \tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - \langle C_t(s, \cdot) \circ A^{\pi_{t+1}}(s, \cdot) + \tau^{-1}_t f_{\theta_t}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle.$$  

(106)

Then, we decompose the above equation into two terms: (i) the error in the policy improvement and (ii) the error between the true advantage and the approximated advantage, i.e.,

$$\langle \tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - (C_t(s, \cdot) A^{\pi_{t+1}}(s, \cdot) + \tau^{-1}_t f_{\theta_t}(s, \cdot)), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle$$

(107)

$$= \langle \tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - \langle C_t(s, \cdot) \circ A_{\omega_t}(s, \cdot) + \tau^{-1}_t f_{\theta_t}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle$$

(108)

$$+ \langle C_t(s, \cdot) \circ A_{\omega_t}(s, \cdot) - \langle C_t(s, \cdot) \circ A^{\pi_{t+1}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle.$$

(109)

We first bound the expectation of (i) over $\nu^*$ as follows.

$$\mathbb{E}_{\nu^*}[\langle \tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - (C_t(s, \cdot) A^{\pi_{t+1}}(s, \cdot) + \tau^{-1}_t f_{\theta_t}(s, \cdot)), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle]$$

(110)

$$= \int_{S \times A} \langle \tau_{t+1}^{-1} f_{\theta_{t+1}}(s, \cdot) - (C_t(s, \cdot) \circ A_{\omega_t}(s, \cdot) + \tau^{-1}_t f_{\theta_t}(s, \cdot)), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s) \rangle \cdot \nu^*(s) \text{d}s$$

(111)

$$\leq C_\infty \mathbb{E}_{\tilde{\tau}_t} \left( \| (\tau_{t+1}^{-1} f_{\theta_{t+1}}(s, a) - (C_t(s, a) A_{\omega_t}(s, a) + \tau^{-1}_t f_{\theta_t}(s, a))) \| \right)^{1/2} \cdot \mathbb{E}_{\tilde{\tau}_t} \left[ \left| \frac{d \pi^*}{d \pi_{\theta_t}} - \frac{d \pi_{\theta_t}}{d \pi_{\theta_t}} \right|^2 \right]^{1/2} \cdot \mathbb{E}_{\tilde{\tau}_t} \left[ \left| \frac{d \pi^*}{d \pi_{\theta_t}} - \frac{d \pi_{\theta_t}}{d \pi_{\theta_t}} \right|^2 \right]^{1/2}$$

(112)

$$\leq C_\infty \tau_{t+1}^{-1} \varepsilon_{t+1},$$

(113)

where (112) follows from the definition of $\tilde{\tau}_t$, (113) is obtained by Cauchy-Schwarz inequality and Assumption 5 and the last inequality in (114) holds by the condition in (22) and that $\| \nu^*/\nu^* \|_\infty < C_\infty$. 


Similarly, we consider the expectation of (ii) over \(\nu^*\) as follows.

\[
\mathbb{E}_{\nu^*}[\langle C_t(s, \cdot) \circ A_{\omega_t}(s, \cdot) - \tilde{C}_t(s, \cdot) \circ A^*_{\omega_t}(s, \cdot), \pi^* (|s) - \pi_{\theta_t} (|s)\rangle]
\]

where (119) holds by the Cauchy-Schwarz inequality. Next, we bound for the term.

Recall that \(E\) holds by the fact that \(A\) as the random variable \(A_{\omega_t}(s, a)\), whose randomness results from the state-action pairs sampled from \(\pi_t\) and the initialization of neural networks, and using \(A^*_{\omega_t}\) as the random variable \((s, A)\), whose randomness comes from the state \(s\) and the action pairs sampled from \(\pi_t\). To establish the bound of \(\mathbb{E}[D]\), we consider two different cases for \(\mathbb{E}[D]\): one is that the error is greater than \(\epsilon_{err}\), and the other is that the error is less than or equal to \(\epsilon_{err}\). Specifically,

\[
\mathbb{E}[D] = \mathbb{E}[D | |A_{\omega_t} - A^*_{\omega_t}| > \epsilon_{err}] \cdot \mathbb{P}(|A_{\omega_t} - A^*_{\omega_t}| > \epsilon_{err}) + \mathbb{E}[D | |A_{\omega_t} - A^*_{\omega_t}| \leq \epsilon_{err}] \cdot \mathbb{P}(|A_{\omega_t} - A^*_{\omega_t}| \leq \epsilon_{err})
\]

Then, we upper bound the two terms in (120) separately. Regarding the first term in (120) we have

\[
\mathbb{E}[D | |A_{\omega_t} - A^*_{\omega_t}| > \epsilon_{err}] \cdot \mathbb{P}(|A_{\omega_t} - A^*_{\omega_t}| > \epsilon_{err})
\]

where (119) holds by that \((a + b)^2 \leq 2a^2 + 2b^2\). Next, regarding the second term in (120) we further consider two cases based on whether the absolute value of \(A^*_{\omega_t}\) is greater than \(\epsilon_{err}\) or not. Specifically,

\[
\mathbb{E}[D | |A_{\omega_t} - A^*_{\omega_t}| \leq \epsilon_{err}]
\]

where (122) holds by the fact that we fix a policy \(\pi_{\theta_t}\) as described in the beginning of Appendix E.2 and hence \(A^*_{\omega_t}\) is determined, (122) holds by that the indicator function is no larger than 1, the first term in (124) holds by the fact that \(A_{\omega_t}\) and \(A^*_{\omega_t}\) have the same sign and hence \(C_t\) is equal to \(\tilde{C}_t\), and the second term in (124) follows from that \((a + b)^2 \leq 2a^2 + 2b^2\). Then, by combining the above terms, we have

\[
\mathbb{E}[D] \leq 2U_2^2(\mathbb{E}_{\nu^*} ||A_{\omega_t}(s, a) - A^*_{\omega_t}(s, a)||^2) + (A^*_{\omega_t})^2 \cdot \mathbb{P}(|A_{\omega_t} - A^*_{\omega_t}| > \epsilon_{err})
\]

Recall that \(c' = \mathbb{E}[(A_{\omega_t}(s, a) - A^*_{\omega_t}(s, a))^2]\). As we could choose an \(\epsilon_{err}\) small enough and use the neural network power to make \(c'\) is also small by Lemma 10 such that we have \(2U_2^2(\mathbb{E}_{\nu^*} ||A_{\omega_t}(s, \cdot)||^2) + (A^*_{\omega_t})^2 > U_2^2c' + 2U_2^2\epsilon_{err}\), then by Lemma 13 we have

\[
\mathbb{E}[D] \leq 2U_2^2(\mathbb{E}_{\nu^*} ||A_{\omega_t}(s, \cdot)||^2) + (A^*_{\omega_t})^2 \cdot \frac{c'}{2\epsilon_{err}} + U_2^2c' + 2U_2^2\epsilon_{err} \cdot (1 - \frac{c'}{\epsilon_{err}})
\]
Rearranging the terms in (127), we have
\[
\mathbb{E}[D] \leq \epsilon_t U_C^2 \cdot \left[ \frac{2}{\epsilon_{err}^2} (M' + (A_{\text{max}}^n)^2 - \epsilon_t/2) - 1 \right] + 2U_C^2 \cdot \epsilon_{err}^2 \tag{128}
\]
\[
\leq \epsilon_t U_C^2 \cdot \left[ \frac{2}{\epsilon_{err}^2} (M' + (A_{\text{max}}^n)^2 - \epsilon_t/2) \right] + 2U_C^2 \cdot \epsilon_{err}^2 \tag{129}
\]
where \( M' := 4E_{\nu_t} [\max_a (Q_{\omega_t}(s, a))] + 4R_t^2 \). By introducing the notation \( X = [(2/\epsilon_{err}^2) (M' + (A_{\text{max}}^n)^2 - \epsilon_t/2)] \) and combining all the above results, we have
\[
\left| \mathbb{E}_{\nu_t} \left[ \left( \log \pi_{t+1} (|s|) - \log \pi_{t+1} (|s|), \pi^* (|s|) - \pi_{\theta_t} (|s|) \right) \right] \right| \leq C_\infty t_{t+1} \epsilon_{t+1} \phi_t^f + (\epsilon_t^2 U_C^2 X + 2U_C^2 \epsilon_{err}^2)^{1/2} \psi_t^f \tag{130}
\]
\[
\leq \epsilon_{t+1} \sqrt{C_\infty t_{t+1} \phi_t^f} + \epsilon_t^{1/2} U_C \cdot X \cdot \phi_t^{1/2} \psi_t^* + 2U_C \epsilon_{err} \psi_t^*. \tag{132}
\]
where (132) follows from the inequality \( \sqrt{a + b} \leq \sqrt{a} + \sqrt{b} \) and that \( \epsilon_t = \epsilon_{t+1} \sqrt{C_\infty t_{t+1} \phi_t^f} + \epsilon_t^{1/2} U_C \cdot X \cdot \phi_t^{1/2} \psi_t^* \) and \( \epsilon_{err} = \sqrt{2U_C \epsilon_{err} \psi_t^*} \). The proof is complete. \( \square \)

### E.3 Proof of Lemma \[3\]

For ease of exposition, we restate Lemma \[3\] as follows.  

**Lemma (Stepwise Energy \( \epsilon_\infty \)-Difference).**

\[
\mathbb{E}_{\nu_t} \left[ \left| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) \right|_{\infty}^2 \right] \leq 2\epsilon_t' + 2U_C^2 M, \tag{133}
\]

where \( \epsilon_t' = |A| \cdot C_\infty t_{t+1} \epsilon_{t+1} \) and \( M = 4E_{\nu_t} [\max_a (Q_{\omega_t}(s, a))] + 4R_t^2 \).

**Remark 5.** As described in Remark \[4\] \( \epsilon_{t+1} \) can be sufficiently small due to Lemma \[12\]. Similarly, \( \epsilon_t' \) can also be made arbitrarily small.

**Proof of Lemma \[5\]** We first find an explicit bound for \( \| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) \|_{\infty}^2 \). Note that
\[
\| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) \|_{\infty}^2 \leq 2\| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) - C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \|_{\infty}^2 \tag{134}
\]
\[
+ 2\| C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \|_{\infty}^2.
\]
Next, we consider the expectation of (134) over \( \nu^* \): For the first term in (134), we have
\[
\mathbb{E}_{\nu_t} \left[ \left| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) - C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right]
\]
\[
= \int_S \| \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, \cdot) - \tau_t^{-1} f_{\theta_t} (s, \cdot) - C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \|_{\infty}^2 \nu^* (s) ds \tag{135}
\]
\[
= \int_{S \times A} \pi_0 (a|s) \cdot \left( \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, a) - \tau_t^{-1} f_{\theta_t} (s, a) \right) \cdot \left( \tau_{t+1}^{-1} f_{\theta_{t+1}} (s, a) - \tau_t^{-1} f_{\theta_t} (s, a) \right) \cdot C_t (s, a) \cdot A_{\omega_t} (s, a) d\sigma_t (s, a) \tag{136}
\]
\[
< |A| \cdot C_\infty \tau_{t+1}^2 \epsilon_{t+1}. \tag{138}
\]
where (138) holds by the condition in (23), the definition of the concentrability coefficient, and the fact that \( \pi_0 \) is a uniform policy. Furthermore, we bound \( \mathbb{E}_{\nu_t} \left[ \left| C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right] \), we have
\[
\mathbb{E}_{\nu_t} \left[ \left| C_t (s, \cdot) \circ A_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right] \leq U_C^2 \cdot \mathbb{E}_{\nu_t} \left[ \left| A_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right] \tag{139}
\]
\[
= U_C^2 \cdot \mathbb{E}_{\nu_t} \left[ \left| Q_{\omega_t} (s, \cdot) - \sum_a \pi_{\theta_t} (a|s) Q_{\omega_t} (s, a) \right|_{\infty}^2 \right] \tag{140}
\]
\[
= U_C^2 \cdot \mathbb{E}_{\nu_t} \left[ \left| Q_{\omega_t} (s, \cdot) - \mathbb{E}_{\theta \sim \pi_{\theta_t}} \left[ Q_{\omega_t} (s, \cdot) \right] \right|_{\infty}^2 \right] \tag{141}
\]
\[
\leq 2U_C^2 \mathbb{E}_{\nu_t} \left[ \left| Q_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right] + 2U_C^2 \mathbb{E}_{\nu_t} \left[ \left( \mathbb{E}_{\theta \sim \pi_{\theta_t}} \left[ Q_{\omega_t} (s, a) \right] \right) \right] \tag{142}
\]
\[
\leq 2U_C^2 \mathbb{E}_{\nu_t} \left[ \left| Q_{\omega_t} (s, \cdot) \right|_{\infty}^2 \right] + 2U_C^2 \mathbb{E}_{\nu_t} \left[ \left( \mathbb{E}_{\theta \sim \pi_{\theta_t}} \left[ Q_{\omega_t} (s, a) \right] \right) \right] \tag{143}
\]
\[
\leq 4U_C^2 \cdot \mathbb{E}_{\nu_t} \left[ \left( \mathbb{E}_{\alpha \sim \max_a (Q_{\omega_t} (s, a))} \right) \right] \right] \right] \tag{144}
\]
\[
\leq 4U_C^2 \cdot \left[ \mathbb{E}_{\nu_t} \left[ \left( \max_a (Q_{\omega_t} (s, a)) \right)^2 \right] + R_t^2 \right], \tag{145}
\]
where (143) holds by using Jensen’s inequality and leveraging the $\ell_\infty$-norm instead of the expectation $E_{\tau_t \sim \pi_{\theta_t} \cdot [\cdot]}$, and the last inequality in (145) holds by the 1-Lipschitz property of neural networks with respect to the weights. By setting $\epsilon_t' = |A|^2 C_{\infty} t_{t+1} \epsilon_t$ and $M = 4 E_\nu \cdot [\max_a (Q_{\omega_t} (s, a))^2] + 4 R^2$, we complete the proof of Lemma 4.

### E.4 Proof of the Lemma 4

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

**Lemma (Stepwise KL Difference).** The KL difference is as follows,

\[
KL(\pi^*(\cdot | s)\|\pi_{\theta_{t+1}}(\cdot | s)) = KL(\pi^*(\cdot | s)\|\pi_{\theta_t}(\cdot | s)) - \langle \log \pi_{\theta_{t+1}}(\cdot | s) - \log \pi_{\theta_t}(\cdot | s), \pi_{\theta_t}(\cdot | s) - \pi^*(\cdot | s) \rangle - \langle \nabla_{\theta} C_t(s, \cdot) \circ A \pi_{\theta_t}(\cdot | s), \pi_{\theta_t}(\cdot | s) - \pi^*(\cdot | s) \rangle
\]

(146)

Then, by Pinsker’s inequality, we have

\[
KL(\pi^*(\cdot | s)\|\pi_{\theta_{t+1}}(\cdot | s)) - KL(\pi^*(\cdot | s)\|\pi_{\theta_t}(\cdot | s)) = \left( \log \frac{\pi_{\theta_t}(\cdot | s)}{\pi_{\theta_{t+1}}(\cdot | s)} \right) \pi_{\theta_{t+1}}(\cdot | s)
\]

(147)

**Proof of Lemma 4** We directly expand the one-step KL divergence difference as

\[
KL(\pi^*(\cdot | s)\|\pi_{\theta_{t+1}}(\cdot | s)) - KL(\pi^*(\cdot | s)\|\pi_{\theta_t}(\cdot | s)) = \left( \log \frac{\pi_{\theta_t}(\cdot | s)}{\pi_{\theta_{t+1}}(\cdot | s)} \right) \pi_{\theta_{t+1}}(\cdot | s) + KL(\pi_{\theta_t}(\cdot | s) || \pi_{\theta_{t+1}}(\cdot | s))
\]

(148)

Next, we apply Pinsker’s inequality to get

\[
KL(\pi^*(\cdot | s)\|\pi_{\theta_{t+1}}(\cdot | s)) - KL(\pi^*(\cdot | s)\|\pi_{\theta_t}(\cdot | s)) = \left( \log \frac{\pi_{\theta_t}(\cdot | s)}{\pi_{\theta_{t+1}}(\cdot | s)} \right) \pi_{\theta_{t+1}}(\cdot | s) + KL(\pi_{\theta_t}(\cdot | s) || \pi_{\theta_{t+1}}(\cdot | s))
\]

(149)

Then, by Pinsker’s inequality, we have

\[
KL(\pi^*(\cdot | s)\|\pi_{\theta_{t+1}}(\cdot | s)) - KL(\pi^*(\cdot | s)\|\pi_{\theta_t}(\cdot | s)) = \left( \log \frac{\pi_{\theta_t}(\cdot | s)}{\pi_{\theta_{t+1}}(\cdot | s)} \right) \pi_{\theta_{t+1}}(\cdot | s) + KL(\pi_{\theta_t}(\cdot | s) || \pi_{\theta_{t+1}}(\cdot | s))
\]

(151)

### E.5 Proof of Lemma 5

For ease of exposition, we restate Lemma 5 as follows.

**Lemma (Performance Difference Using Advantage).** Recall that $L(\pi) = E_\nu \cdot [V^\pi(s)]$. We have

\[
L(\pi^*) - L(\pi) = (1 - \gamma)^{-1} \cdot E_\nu \cdot [\langle A^\pi(s, \cdot), \pi^*(\cdot | s) - \pi(\cdot | s) \rangle] = (1 - \gamma)^{-1} \cdot E_\nu \cdot [\langle Q^\pi(s, \cdot), \pi^*(\cdot | s) - \pi(\cdot | s) \rangle].
\]

(154)

Before proving Lemma 5, we first state the following property.

**Lemma 14** We have

\[
L(\pi^*) - L(\pi) = (1 - \gamma)^{-1} \cdot E_\nu \cdot [\langle Q^\pi(s, \cdot), \pi^*(\cdot | s) - \pi(\cdot | s) \rangle].
\]

(155)
Proof of Lemma: As the value function $V^\pi(\cdot)$ is state-dependent, we have
\[
\mathbb{E}_{\nu^*}[E_{\nu^*}[V^\pi(s), \pi^*(\cdot|s) - \pi(\cdot|s)]] = \mathbb{E}_{\nu^*}[V^\pi(s) \cdot \sum_{a \in A} (\pi^*(a|s) - \pi(a|s))]
\]
(156)
\[
= \mathbb{E}_{\nu^*} \left[ V^\pi(s) \cdot \left( \sum_{a \in A} \pi^*(a|s) - \sum_{a \in A} \pi(a|s) \right) \right] = 0. \tag{157}
\]
Therefore, by (157) and Lemma 14, we have
\[
\mathcal{L}(\pi^*) - \mathcal{L}(\pi) = (1 - \gamma)^{-1} \cdot \mathbb{E}_{\nu^*}[Q^\pi(s, \cdot) - V^\pi(s), \pi^*(\cdot|s) - \pi(\cdot|s)]
\]
(158)
\[
= (1 - \gamma)^{-1} \cdot \mathbb{E}_{\nu^*}[(A^\pi(s, \cdot), \pi^*(\cdot|s) - \pi(\cdot|s))]. \tag{159}
\]
\]

E.6 Proof of Theorem 2

By taking expectation of the KL difference in Lemma 3 over $\nu^*$, we obtain
\[
\mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_{t+1}}(\cdot|s)) - KL(\pi(\cdot|s)||\pi_{\theta_1}(\cdot|s))]
\]
(160)
\[
\leq \varepsilon_t + \varepsilon_{err} - \mathbb{E}_{\nu^*}[(\tilde{C}_t(s, \cdot) \circ A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))] - \frac{1}{2} \mathbb{E}_{\nu^*}[||\pi_{\theta_{t+1}}(\cdot|s) - \pi_{\theta_t}(\cdot|s)||^2_\infty]
\]
(161)
\[
\leq \mathbb{E}_{\nu^*}[(\tilde{C}_t(s, \cdot) \circ A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))] - \frac{1}{2} \mathbb{E}_{\nu^*}[||\pi_{\theta_{t+1}}(\cdot|s) - \pi_{\theta_t}(\cdot|s)||^2_\infty]
\]
(162)
\[
\leq \mathbb{E}_{\nu^*}[(\tilde{C}_t(s, \cdot) \circ A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s)] + \mathbb{E}_{\nu^*}[||\pi_{\theta_{t+1}}(s, \cdot) - \pi_{\theta_t}(s, \cdot)||_\infty \cdot ||\pi_{\theta_{t+1}}(\cdot|s) - \pi_{\theta_t}(\cdot|s)||_\infty]
\]
(163)
\[
\leq \frac{1}{2} \mathbb{E}_{\nu^*}[||\pi_{\theta_{t+1}}(s, \cdot) - \pi_{\theta_t}(s, \cdot)||^2_\infty],
\]
where the first inequality follows from Lemma 3 and Lemma 2, the second inequality holds by the Hölder's inequality, and the last inequality holds by the fact that $2xy - x^2 \leq y^2$ and merging the last two terms.

Then, by Lemma 3 and rearranging the terms, we obtain that
\[
\mathbb{E}_{\nu^*}[(\tilde{C}_t(s, \cdot) \circ A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))]
\]
\[
\leq \mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_0}(\cdot|s)) - KL(\pi^*(\cdot|s)||\pi_{\theta_{t+1}}(\cdot|s))]
\]
\[
+ \varepsilon_t + \varepsilon_{err} + \varepsilon_t' + U_C^2. \tag{164}
\]

By the first condition of (16), we have $L_C \mathbb{E}_{\nu^*}[(A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))] \leq \mathbb{E}_{\nu^*}[(\tilde{C}_t(s, \cdot) \circ A^{\pi_{\theta}}(s, \cdot), \pi^*(\cdot|s) - \pi_{\theta_t}(\cdot|s))].$ By obtaining the performance difference via Lemma 5, we have
\[
(1 - \gamma) L_C (\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t}))
\]
\[
\leq \mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_0}(\cdot|s)) - KL(\pi^*(\cdot|s)||\pi_{\theta_{t+1}}(\cdot|s))]
\]
\[
+ \varepsilon_t + \varepsilon_{err} + \varepsilon_t' + U_C^2. \tag{165}
\]

Then, by taking the telescoping sum of (165) from $t = 0$ to $T - 1$, we have
\[
(1 - \gamma) L_C \sum_{t=0}^{T-1}(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t}))
\]
\[
\leq \mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_0}(\cdot|s))] - \mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_T}(\cdot|s))] + \sum_{t=0}^{T-1}(\varepsilon_t + \varepsilon_{err} + \varepsilon_t') + T U_C^2 M. \tag{166}
\]

By the facts that (i) $\mathbb{E}_{\nu^*}[KL(\pi^*(\cdot|s)||\pi_{\theta_0}(\cdot|s))] \leq \log |A|$, (ii) KL divergence is nonnegative, (iii) $\sum_{t=0}^{T-1}(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})) \geq T \cdot \min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\}$, we have
\[
\min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\} \leq \frac{\log |A| + \sum_{t=0}^{T-1}(\varepsilon_t + \varepsilon_t') + T(\varepsilon_{err} + MU_C^2)}{TL_C(1 - \gamma)}. \tag{168}
\]
Since we have $\varepsilon_{\text{err}} = \sqrt{2U_C\varepsilon_{\text{err}}\psi_t^*}$ and the condition of (17), we know that if we set $\varepsilon_{\text{err}} = U_C$ and $T$ to be sufficiently large, $\varepsilon_{\text{err}}$ shall be sufficiently small and hence satisfy the condition required by (127). Thus, by plugging $\varepsilon_{\text{err}} = U_C$ into (168), we have $\varepsilon_{\text{err}} = \sqrt{2U_C^2\psi_t^*}$ and $\varepsilon_t = \epsilon_{t+1}^{1/2}C_{\infty}\tau_{t+1}^{-1}\phi_t^* + \epsilon_{t}^{1/2}U_C \left[\frac{(2/U_C)(M + (A_{\text{max}})^2 - \epsilon_t/2)}{1/2}\right]^{1/2}\psi_t^* = \epsilon_{t+1}^{1/2}C_{\infty}\tau_{t+1}^{-1}\phi_t^* + \epsilon_{t}^{1/2}U_CY^{1/2}\psi_t^*$, where $Y = 2M + 2(R_{\max}/(1-\gamma))^2 - \epsilon_t$. Finally, we have

$$\min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\varepsilon_t + \varepsilon_t') + TU_C^2(2\psi_t^* + M)}{TL_C(1-\gamma)}. \tag{169}$$

By the condition (17), $U_C^2$ can always cancel out $T$ in the numerator of (169). Moreover, in the denominator of (169), $L_C = \omega(T^{-1})$ is large enough to obtain convergence. Hence, the convergence rate is $O((TL_C)^{-1})$, and we complete the proof. \[\square\]

**Remark 6.** As mentioned in Remark 3, the choices of $\eta$ and $\{\tau_t\}$ would affect the convergence rate and need to be configured properly for different algorithms in the NeuralHPO family. As will be shown in Appendix F, this fact can be further explained through the bounds $U_C$ and $L_C$ obtained in (175) and (181).

### F. Additional Corollaries and Proofs

#### F.1 Proof of Corollary 1

For ease of exposition, we restate the corollary as follows.

**Corollary (Global Convergence of PPO-clip with Convergence Rate).** Consider NeuralHPO with the PPO-clip classifier $\rho_{s,a}(\theta) - 1$ and the objective function in each iteration $t$ as

$$L^{(t)}(\theta) = E_{\pi}(\cdot|s) | A^{\pi_s}(s,\cdot)| \circ \ell(\text{sign}(A^{\pi_s}(s,\cdot)), \frac{\pi_{\theta}(\cdot|s)}{\rho_{\theta}(\cdot|s)} - 1) \tag{170}$$

Recall that $K$ is the maximum number of EMDA iterations. By setting $\eta = 1/\sqrt{T}$ and $\tau_t = \sqrt{T}/(Kt)$, we have

$$\min_{0 \leq t \leq T} \{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t})\} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\varepsilon_t + \varepsilon_t') + K^2(2\psi_t^* + M)}{\sqrt{T}(1-\gamma)}. \tag{171}$$

Hence, we provide the $O(1/\sqrt{T})$ convergence rate of PPO-clip.

**Proof of Corollary 1.** We find the lower and upper bounds $L_C, U_C$ for PPO-clip. We first consider the derivative $g_{s,a}$ of the objective with the true advantage function $A^{\pi_s}$.

$$g_{s,a} = \frac{\partial L^{(t)}(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}_{s,a}} = -A^{\pi_s}(s,a) \cdot \mathbf{1} \left\{ \left( \frac{\hat{\theta}_{s,a}}{\rho_{\theta}(\cdot|s)} - 1 \right) \cdot \text{sign}(A^{\pi_s}(s,a)) < \epsilon \right\}. \tag{172}$$

Then, we check the sufficient conditions (16) and (17). Recall that $K$ is the maximum number of EMDA iteration for each $t$. We sum up the gradients with $\eta$ and rearrange the terms into $\tilde{C}_t(s,a)$. Then, we have the upper bound as

$$\tilde{C}_t(s,a) \cdot |A^{\pi_s}(s,a)| \leq \sum_{k=0}^{K-1} \eta \cdot |A^{\pi_s}(s,a)| \leq K\eta \cdot |A^{\pi_s}(s,a)|. \tag{173}$$

Regarding the lower bound, as we know that under PPO-clip, the first step of EMDA shall always make an update, and hence we have

$$\eta \cdot |A^{\pi_s}(s,a)| \leq \tilde{C}_t(s,a) \cdot |A^{\pi_s}(s,a)|. \tag{174}$$

Lastly, by setting $\eta = 1/\sqrt{T}$ and selecting the temperature as $\tau_t = \sqrt{T}/(Kt)$ to satisfy the condition $\tau_{t+1}(U_C^2 + \tau_t^{-2}) \leq 1$ that we use in (88), we obtain

$$\omega(T^{-1}) = T^{-1/2}|A^{\pi_s}(s,a)| \leq \tilde{C}_t(s,a) \cdot |A^{\pi_s}(s,a)| \leq KT^{-1/2} \cdot |A^{\pi_s}(s,a)| = O(T^{-1/2}). \tag{175}$$
We have checked the sufficient conditions of Theorem 2. Thus, we obtain,
\[
\min_{0 \leq t \leq T} \{ \mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t}) \} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\epsilon_t + \epsilon'_t) + K^2(\sqrt{2}\psi^*_t + M)}{\sqrt{T(1-\gamma)}}.
\]
(176)
We complete the proof and obtain the \(O(1/\sqrt{T})\) convergence rate.

F.2 Convergence Rate of NeuralHPO With an Alternative Classifier

**Corollary 2** (Global Convergence of HPO-sub with Convergence Rate). By using the result in Theorem 2, we consider NeuralHPO with the subtraction classifier \(\pi_{\theta}(\cdot|s) - \pi_{\theta}(\cdot|s)\) and the objective function in each iteration \(t\) is
\[
L^{(t)}(\pi_{\theta}) = \mathbb{E}_{\sigma_t}[[A^{\pi_{\theta}}(s, a)] \cdot \ell(sgn(A^{\pi_{\theta}}(s, a)), \pi_{\theta}(a|s) - \pi_{\theta}(a|s), \epsilon)].
\]
(177)
By setting \(\eta = 1/\sqrt{T}\) and \(\tau_t = \sqrt{T/(Kt)}\), we have
\[
\min_{0 \leq t \leq T} \{ \mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t}) \} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\epsilon_t + \epsilon'_t) + K^2(\sqrt{2}\psi^*_t + M)}{\sqrt{T(1-\gamma)}}.
\]
(178)

**Proof of Corollary 2.** Similar to Corollary [1] we derive the gradient of our objective with the true advantage function \(A^{\pi_{\theta}}(s, a)\). Specifically, we have
\[
g_{s,a} = \frac{\partial L^{(t)}(\pi_{\theta})}{\partial \theta^t_{s,a}} = -A^{\pi_{\theta}}(s, a) \cdot 1 \left\{ \left( \theta^t_{s,a} - \pi_{\theta}(a|s) \right) \cdot sgn(A^{\pi_{\theta}}(s, a)) < \epsilon \right\}.
\]
(179)
Thus, similar to [1] we have
\[
\eta \cdot |A^{\pi_{\theta}}(s, a)| \leq C_{t}(s, a) \cdot |A^{\pi_{\theta}}(s, a)| \leq K \eta \cdot |A^{\pi_{\theta}}(s, a)|.
\]
(180)
We also set \(\eta = 1/\sqrt{T}\) and pick \(\tau_t = \sqrt{T/(Kt)}\) to satisfy the condition \(\tau_t^2 T^{1/2} A_{\tau_t}^2 + \tau_t^{-2} \leq 1\) that we use in [88]. Accordingly, we obtain
\[
\omega(T^{-1}) = T^{-1/2}|A^{\pi_{\theta}}(s, a)| \leq C_{t}(s, a) \cdot |A^{\pi_{\theta}}(s, a)| \leq KT^{-1/2} \cdot |A^{\pi_{\theta}}(s, a)| = O(T^{-1/2}).
\]
(181)
We have checked the sufficient condition of Theorem 2. Therefore, by plugging in \(L_C\) and \(U_C\), we obtain
\[
\min_{0 \leq t \leq T} \{ \mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_t}) \} \leq \frac{\log |A| + \sum_{t=0}^{T-1} (\epsilon_t + \epsilon'_t) + K^2(\sqrt{2}\psi^*_t + M)}{\sqrt{T(1-\gamma)}}.
\]
(182)
We complete the proof and obtain the \(O(1/\sqrt{T})\) convergence rate.

G Supplementary Related Works

**RL as Classification.** Regarding the general idea of casting RL as a classification problem, it has been investigated by the existing literature [29–31], which view the one-step greedy update (e.g., in Q-learning) as a binary classification problem. However, a major difference is the labeling: classification-based approximate policy iteration labels the action with the largest Q value as positive; HPO labels the actions with positive advantage as positive. Despite the high-level resemblance, our paper is fundamentally different from the prior works [29–31] as our paper is meant to study the theoretical foundation of PPOclip, from the perspective of HPO.

**H Comparison of the Clipped Objective in PPO-Clip and the Hinge Loss Objective of HPO**

Recall that the original objective of PPO-clip is
\[
L^{clip}(\theta) = \mathbb{E}_{s \sim \rho_{\pi_{\theta}}, a \sim \pi_{\cdot|s}} \left[ \min\{\rho_{s,a}(\theta)A^+(s,a), \text{clip}(\rho_{s,a}(\theta), 1 - \epsilon, 1 + \epsilon)A^+(s,a)\} \right],
\]
(183)
where $\rho_{s,a}(\theta) = \frac{\pi(a|s)}{\pi(a|s)}$. In practice, $L^{\text{clip}}(\theta)$ is approximated by the sample average as

$$L^{\text{clip}}(\theta) \approx \hat{L}^{\text{clip}}(\theta) = \frac{1}{|D|} \sum_{(s,a) \in D} \min\{\rho_{s,a}(\theta) A^\pi(s,a), \text{clip}(\rho_{s,a}(\theta), 1 - \epsilon, 1 + \epsilon) A^\pi(s,a)\}$$

$$= \frac{1}{|D|} \sum_{(s,a) \in D} [A^\pi(s,a)] \cdot \min\{\rho_{s,a}(\theta) \text{sign}(A^\pi(s,a)), \text{clip}(\rho_{s,a}(\theta), 1 - \epsilon, 1 + \epsilon) \text{sign}(A^\pi(s,a))\}.$$ \hspace{1cm} (184)

Note that $H^{\text{clip}}_{s,a}(\theta)$ can be further written as

$$H^{\text{clip}}_{s,a}(\theta) = \begin{cases} 1 + \epsilon & \text{if } A^\pi(s,a) > 0 \text{ and } \rho_{s,a}(\theta) \geq 1 + \epsilon \\ \rho_{s,a}(\theta) & \text{if } A^\pi(s,a) > 0 \text{ and } \rho_{s,a}(\theta) < 1 + \epsilon \\ -\rho_{s,a}(\theta) & \text{if } A^\pi(s,a) < 0 \text{ and } \rho_{s,a}(\theta) > 1 - \epsilon \\ -(1 - \epsilon) & \text{if } A^\pi(s,a) < 0 \text{ and } \rho_{s,a}(\theta) \leq 1 - \epsilon \\ 0 & \text{otherwise} \end{cases}.$$ \hspace{1cm} (185)

Recall that the loss function of HPO takes the form as

$$L(\theta) \approx \hat{L}(\theta) = \frac{1}{|D|} \sum_{(s,a) \in D} [A^\pi(s,a)] \cdot \max\{0, \epsilon - (\rho_{s,a}(\theta) - 1) \text{sign}(A^\pi(s,a))\}.$$ \hspace{1cm} (186)

Similarly, $H_{s,a}(\theta)$ can be further written as

$$H_{s,a}(\theta) = \begin{cases} 0 & \text{if } A^\pi(s,a) > 0 \text{ and } \rho_{s,a}(\theta) \geq 1 + \epsilon \\ \rho_{s,a}(\theta) - (1 - \epsilon) & \text{if } A^\pi(s,a) > 0 \text{ and } \rho_{s,a}(\theta) < 1 + \epsilon \\ -\rho_{s,a}(\theta) + (1 + \epsilon) & \text{if } A^\pi(s,a) < 0 \text{ and } \rho_{s,a}(\theta) > 1 - \epsilon \\ 0 & \text{if } A^\pi(s,a) < 0 \text{ and } \rho_{s,a}(\theta) \leq 1 - \epsilon \\ \epsilon & \text{otherwise} \end{cases}.$$ \hspace{1cm} (187)

Therefore, it is easy to verify that $\hat{L}^{\text{clip}}(\theta)$ and $-\hat{L}(\theta)$ only differ by a constant with respect to $\theta$. This also implies that $\nabla_{\theta} \hat{L}^{\text{clip}}(\theta) = -\nabla_{\theta} \hat{L}(\theta)$.