ORAL: Adaptive Gap Increasing for Advantage Learning via Occam's Razor Principle

Zhe Zhang, Yongle Zhou, Yuyang Long, Jia Zhang, *Member, IEEE*, Juanjuan Weng, Zhetao Li, *Member, IEEE*, Yaozhong Gan, and Xiaoyang Tan

***Abstract*—Benefiting from the gap increasing between the optimal action and its competitors, the advantage learning (AL) operator is more robust to estimation errors in the approximated Q-functions than the Bellman optimality operator in reinforcement learning (RL). However, our analysis reveals that its robustness and larger action gaps come at the cost of a worse performance loss bound, leading to slower convergence of value functions. To address this issue, we present a novel method, named Occam's Razor-based AL (ORAL), which follows Occam's Razor principle and takes the necessity into consideration when increasing the action gap. Specifically, our ORAL can adaptively increase the action gap for different state–action pairs, depending on the proximity of their Q values to the optimal ones. We first propose a naive implementation of ORAL, employing a nonsmooth clipping function to realize the above idea, and then introduce a smooth version of ORAL aimed at achieving more stable learning. Furthermore, our methods can be easily plugged into other AL-based operators and extended to more complex continuous-control tasks. Theoretical analysis supports the feasibility of our approaches, demonstrating their ability to balance the gap increasing with fast convergence. Empirical results further validate its effectiveness, showing significant performance improvements across multiple benchmarks.**

***Index Terms*—Action gap increasing, advantage learning (AL), fast convergence, Occam's Razor principle, reinforcement learning (RL).**

I. Introduction

DEEP reinforcement learning (deep RL) has made great progress in solving complex decision tasks, such as game playing [1], [2], robotic manipulation [3], [4], and healthcare [5], [6]. No matter whether value-based [7], [8] or actor-critic [9], [10] algorithms in deep RL, the Bellman optimality operator [11] plays an important role in policy evaluation due to its contraction property. It has been proven that the sequence of value functions iterated by the Bellman optimality operator will converge to the optimal one, which can induce the optimal policy.

To further improve the robustness of evaluated Q-functions, researchers proposed the advantage learning (AL) operator [12], [13], only imposing an extra scaling advantage function term on the Bellman optimality operator. The AL operator can effectively increase the value difference between optimal actions and others (also called "action gap"), compared with the naive Bellman optimality operator. In common, a larger action gap is beneficial to mitigate the undesirable effects of estimation errors from the approximated value function [14], [15]. In addition to this *gap-increasing* property, the AL operator is also *optimality preserving* and can learn a convergent Q-function that does not change the optimal actions.

Recent research has explored the AL operator from the perspective of implicit regularization [16], [17], elaborating its robustness through error propagation analysis [13], [15]. Building on these desirable properties, several works [18], [19], [20] have successfully integrated the AL operator with various existing RL methods. Unfortunately, most of these approaches ignore the potential negative effects of the gap increasing caused by the advantage term.

In this article, we identify a key issue: while the AL operator can increase action gaps, it may also impede value improvement. Specifically, if the optimal action induced by the approximated value function does not match the true optimal action, the AL operator assigns a negative advantage to the true optimal action, leading to an underestimated target in the next iteration This misalignment between the approximated and true optimal actions persists until the optimal policy is achieved, causing a pessimistic estimation that hinders the value improvement of the true optimal action and slows convergence to the optimal Q value. Both theoretical and empirical evidence presented in this article corroborate this issue. To address this, it is crucial to balance faster convergence with the benefits of larger action gaps by considering whether it is necessary to add the extra advantage term for each state–action pair, avoiding a blind increase in the action gap.

To achieve the above goals, this article proposes an improved AL operator that adaptively increases the action gap, building on a recent conference paper [21] Our major contributions are summarized as follows.

First, we provide an in-depth analysis of the impact of the additional advantage term in the AL operator. Theoretical evidence reveals that due to the mismatch between the induced and true optimal actions, the AL operator can introduce more

Received 20 November 2024; revised 2 September 2025; accepted 21 October 2025. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 21625348 and Grant 21625110 and in part by the National Natural Science Foundation of China under Grant 62506191 and Grant 62476128. *(Corresponding authors: Yaozhong Gan; Xiaoyang Tan; Zhetao Li.)*

Zhe Zhang, Yongle Zhou, Yuyang Long, Jia Zhang, Juanjuan Weng, and Zhetao Li are with the College of Information Science and Technology, Jinan University, Guangzhou 510632, China (e-mail: zhangzhe1012@jnu.edu.cn; zyl1574637723@stu2024.jnu.edu.cn; zzy666@stu2023.jnu.edu.cn; jiazhang@jnu.edu.cn; jjweng@jnu.edu.cn; liztchina@hotmail.com).

Yaozhong Gan is with the Qiyuan Lab, Beijing 100095, China (e-mail: yzgancn@163.com).

Xiaoyang Tan is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: x.tan@nuaa.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3626536, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3626536

errors into the performance loss bound, resulting in slower value convergence. This insight motivates us to avoid blind action gap increasing. We further illustrate this problem and its motivation through a classic toy example.

Second, to achieve adaptive gap control, we propose selectively increasing the action gap for different state–action pairs. The advantage term is retained to pursue a larger action gap only when the Q value of a state–action pair is close to that of the optimal action (i.e., when the action gap is small). Initially, we implement this adaptive control using a simple clipping function and subsequently improve it with a smooth Sigmoid function to seek more stable learning. Our analysis verifies the feasibility of these approaches, and we collectively refer to these methods as Occam's Razor-based AL (ORAL), as our key idea can be summarized as "the advantage term must not be multiplied beyond necessity," echoing Occam's Razor principle.

Finally, to evaluate the broader effectiveness of our approach, we extend our methods to both other AL-based algorithms and more complex continuous-control tasks. Extensive empirical results show that our methods not only balance larger action gaps with faster convergence but also deliver significant performance improvements across various RL benchmarks.

Difference From the Recent Conference Paper: Compared with the recent conference paper [21], this work differs from the previous study in the following aspects.

1) Methodologically, we first identify the instability issues present in the original "clipped AL" method from the previous work [21], using an illustrative toy example. To address these issues, we introduce a smooth version of this adaptive AL method, aiming for more stable learning and better convergence properties. We collectively refer to both approaches as ORAL methods. To comprehensively verify their effectiveness, we not only extend them to other AL-based methods, such as persistent AL (PAL) [12] and MRL [16] but also apply them to more complex continuous-control tasks in conjunction with the normalized advantage function (NAF) technique.
2) Theoretically, we mainly add the convergence analysis of our ORAL methods, focusing on the conditions that enhance the stability of the Q-value function. Furthermore, the analysis results also highlight the advantages of the smooth-version ORAL in improving value convergence, corresponding to our motivation.
3) In terms of experiments, in addition to supplementing the performance comparisons on MinAtar tasks, this article further evaluates the performance of various algorithms on more challenging benchmarks, such as discrete-control Atari tasks [22] and continuous-control Mujoco Locomotion tasks [23]. Moreover, we also conduct additional empirical analyses to verify the effectiveness and feasibility of our ORAL methods.

In what follows, we first summarize the related work in Section II, and then, the introduction to AL and its properties will be discussed in Section III. Section IV analyzes the potential issues of the AL operator and elaborates on our proposed solutions, including a naive ORAL and its smooth version, as well as their extensions. Then, Section V provides the theoretical analysis on the proposed methods and Section VI demonstrates performance comparisons and empirical analysis. Finally, we summarize and conclude this article in Section VII.

II. Related Work

As a representative one among numerous alternatives to the Bellman optimality operator (see [24], [25]), the recently proposed AL operator [12] derives from the residual algorithms in the context of continuous time problem [26] and can be viewed as a more general policy iteration of dynamic policy programming [15]. Follow-up related work either aims to analyze and explain the merits of the AL operator or design better AL-based algorithms.

First, to comprehensively understand the benefit of gap-increasing property owned by the AL operator, many researchers have dived deep into the action gap phenomenon in RL algorithms. Farahmand [14] found and studied this action gap phenomenon in the case of two-action discounted Markov decision processes (MDPs) and proved that a favorable action gap regularity can lead to smaller performance loss. Vieillard et al. [16] established a connection between an implicit KL regularization and the action gap regularity, and the error propagation analysis showed that this regularization could lead to a smoother error accumulation, which is thought of as beneficial to stable learning. Besides, Van Seijen et al. [27] put forward an assumption that a larger variance in the action gap across the state space would hurt the performance of RL methods and provided sufficient empirical evidence to validate this hypothesis.

Second, recent works have focused on proposing improved AL-based algorithms to address various existing issues. These works primarily aim to enhance the original AL from two key perspectives.

One direction involves integrating the AL operator with other RL methods. For example, to stabilize the training procedure, Gan et al. [19] presented the smoothing AL (SAL) method that adds the scaling advantage term into the smooth Bellman optimality operator [28]. Ferret et al. [20] plugged the AL operator into self-imitation learning (SIL) [29], seeking for an optimistic exploration. While by combining AL and Retrace (λ) [30] methods, Kozuno et al. [18] proposed a multistep version of the AL algorithm. Wiltzer et al. [31] defined distributional perspectives on action gaps and advantages that shed light on the superiority-based Distributional RL algorithm, mitigating performance issues in high-frequency value-based RL. Viewing the Munchausen term as a soft advantage term, some researchers [32], [33] introduced the concept of gap increasing to offline RL methods.

Another direction is to seek a more robust gap increasing. In the original paper, Bellemare et al. [12] further proposed the PAL operator to encourage greedy policies that infrequently switch between actions in domains with a high temporal resolution. Conservative valuation iteration (CVI) [34] tried to achieve a soft gap increasing by replacing max operators in AL with softmax ones, which could control the tradeoff between error tolerance and convergence rate. Besides, Munchausen DQN (MDQN) [16] adopted a clipping function on its log-policy term so as to avoid the numerical issue when implementing the soft gap increasing.

In summary, previous research on AL-based methods has primarily focused on the theoretical superiority of the gap-increasing property, integrating and extending this gap-increasing mechanism with other methods to achieve improved performance. However, these works neglected the potential negative implications associated with the gap-increasing property. In contrast, our study investigates the potential adverse effects that the gap-increasing property may have on value

function learning within AL-based methods. This provides a more comprehensive understanding of the advantages and disadvantages of the gap-increasing property and enables us to propose effective solutions.

III. Background

A. Definition

Similar to prior work, we aim to solve within the MDP framework, which is defined as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0\}$. Notably, $\mathcal{S}$ and $\mathcal{A}$ represent the state and action space, respectively. The dynamic transition probability model $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ that satisfies the probability simplex over the state space and the Markov property. The reward function $r : \mathcal{S} \times \mathcal{A} \to [R_{\min}, R_{\max}]$ outputs the immediate reward for each state–action pair, and the discount factor γ is set up for the infinite horizon. $\mu_0 : \mathcal{S} \to [0, 1]$ denotes the probability distribution of the initial state s_0. The RL agent will interact with the environment following the policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and generate the sampling trajectory $\tau = (s_0, a_0, r_0, \ldots, s_t, a_t, r_t, \ldots)$.[1]

1) Bellman Optimality Operator: To evaluate the quality of any policy, it is common to define its state–action value function (i.e., Q-value function) as $Q^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right]$, where $\mathbb{E}_\pi$ represents the expectation over all the trajectories sampled by the policy π. The corresponding state value function is denoted by $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s\right]$. The purpose of the RL algorithms is to find the optimal policy π^*, which can maximize both the value functions over the space of nonstationary and randomized policies Π: $V^* = \sup_{\pi\in\Pi} V^\pi(s), Q^* = \sup_{\pi\in\Pi} Q^\pi(s, a)$. Previous work [35] has shown that there exists a deterministic and stationary optimal policy whose value functions align with the optimal ones: $V^{\pi^*}(s) = V^*(s), Q^{\pi^*}(s, a) = Q^*(s, a)$. The optimal value functions must satisfy $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s, a)$ and the following *Bellman optimality equation*:

$$Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^*(s')\right]. \tag{1}$$

Furthermore, by rewriting (1) as the vector form, the *Bellman optimality operator* $\mathcal{T}$ is denoted by

$$\mathcal{T}Q \triangleq r + \gamma PV \tag{2}$$

where $P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|}, Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, V \in \mathbb{R}^{|\mathcal{S}|}$, and $V(s) \triangleq \max_a Q(s, a)$. $\mathcal{T}$ is also a contraction operator whose fixed point is the optimal Q-value function.

2) AL Operator: Usually, the parameterized Q-function $Q_\theta(s, a)$ suffers from severe approximation errors because of the usage of the approximation function in deep RL. To further improve the robustness to approximation errors, recent work proposed the AL operator as an alternative to learning the Q-function, and its specific definition is as follows:

$$\mathcal{T}_{\mathrm{AL}}Q(s, a) \triangleq \mathcal{T}Q(s, a) + \alpha(Q(s, a) - V(s)) \tag{3}$$

where $A(s, a) = Q(s, a) - V(s) \le 0$ is commonly named the *advantage function* and $\alpha \in [0, 1)$ is the scaling factor. Notably, $\mathcal{T}_{\mathrm{AL}}$ will keep the same bootstrapping target of the induced optimal action $\tilde{a}^* = \arg\max_a Q(s, a)$ with $\mathcal{T}$ due to $A(s, \tilde{a}^*) = 0$. While the induced suboptimal actions $\tilde{a} \in \mathcal{A}/\tilde{a}^*$ own the underestimated targets because of $A(s, \tilde{a}) < 0$. Intuitively, this *nonpositive* advantage function will lead to a larger Q-value difference between the induced optimal and suboptimal actions. Moreover, Section III-B shows that this larger value difference also exists between the truly optimal and suboptimal actions.

[1]Note that we may abuse these function notations slightly as their vector forms, please refer to the context.

B. Properties

According to the above definitions, we know that the only difference between $\mathcal{T}$ and $\mathcal{T}_{\mathrm{AL}}$ is the additional scaling advantage function term ("advantage term" for short). The original paper [12] shows that this minor modification makes the AL operator have the following good properties.

Property 1 (Optimality Preserving): An operator $\mathcal{T}'$ is optimality preserving if, $\forall Q_0 \in \mathcal{Q}$ and $s \in \mathcal{S}$, letting $Q_{k+1} := \mathcal{T}'Q_k$

$$\tilde{V}(s) := \lim_{k\to\infty} \max_{a\in\mathcal{A}} Q_k(s, a)$$

exists, is unique, $\tilde{V}(s) = V^*(s)$, and $\forall a \in \mathcal{A}$

$$Q^*(s, a) < V^*(s) \Longrightarrow \limsup_{k\to\infty} Q_k(s, a) < V^*(s).$$

Property 1 indicates that any optimality-preserving operator will not change the suboptimal actions to the optimal ones and make at least one optimal action remain optimal. This also implies that the AL operator can still induce the optimal policy finally. Moreover, considering the concept of "*action gap*" for a Q-function, as defined in the following equation:

$$G(s, a) \triangleq V(s) - Q(s, a) = \max_{a'\in\mathcal{A}} Q(s, a') - Q(s, a) \tag{4}$$

where $G(s, a)$ reflects the Q-value difference between the optimal action and a suboptimal action. The AL operator also exhibits the following "*gap-increasing*" property.

Property 2 (Gap Increasing): Let $\mathcal{M}$ be an MDP. An operator $\mathcal{T}'$ for $\mathcal{M}$ is gap increasing if $\forall Q_0 \in \mathcal{Q}, s \in \mathcal{S}, a \in \mathcal{A}$, letting $Q_{k+1} := \mathcal{T}'Q_k$ and $V_k(s) := \max_{a'\in\mathcal{A}} Q_k(s, a')$

$$\liminf_{k\to\infty} [V_k(s) - Q_k(s, a)] \ge V^*(s) - Q^*(s, a) = G^*(s, a).$$

The gap-increasing property indicates that the AL operator induces a larger action gap compared to that induced by the Bellman optimality operator [i.e., $G^*(s, a)$]. Generally, larger action gaps are more beneficial for the robustness to estimation errors. The later work [13] further proved that the AL operator was also a contraction operator and provided the quantitative result of the action gap.

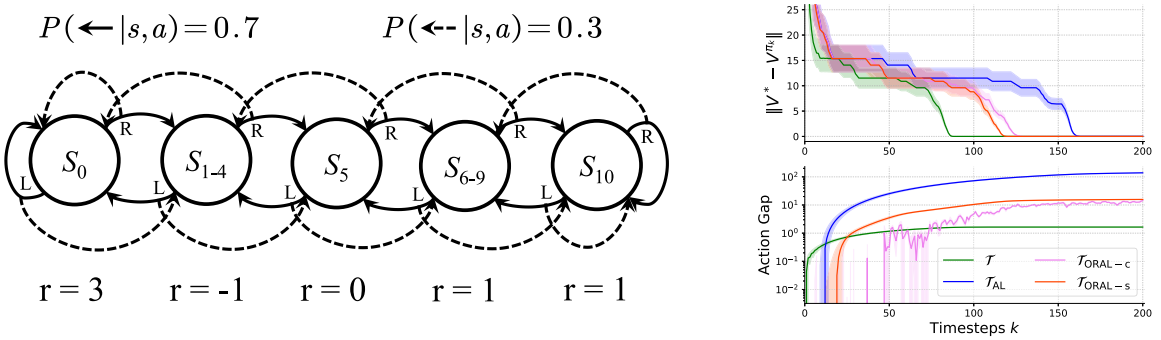
Property 3: Suppose a function $Q_0 \in \mathcal{Q}$ and update it following $\mathcal{T}_{\mathrm{AL}}$, and letting $Q_{k+1} := \mathcal{T}_{\mathrm{AL}}Q_k$, we have:

$$V^*(s) - \lim_{k\to\infty} Q_k(s, a) = \frac{1}{1-\alpha}(V^*(s) - Q^*(s, a)).$$

We can see that the scaling factor $\alpha \in [0, 1)$ used in the AL operator [shown in (3)] determines the magnitude of action gaps, finally achieving $1/(1 - \alpha)$ times action gaps for any state–action pair than the Bellman optimality operator.

IV. Methods

In this section, we prove through detailed analysis that while the AL operator possesses the "gap-increasing" property compared to the Bellman operator, the Bellman operator still achieves faster convergence of the value function. This finding inspires us to design a novel operator that can effectively combine the advantages of both, achieving a balance between larger action gaps and faster value function convergence.



(a) (b)

Fig. 1. Numerical experiments on the common Chainwalk toy example. (a) Detailed illustrations on 11-state Chainwalk MDP. (b) Estimations and comparisons between different learning operators about performance loss (top) and action gaps (bottom). All the results are averaged over three random seeds, with the shaded area corresponding to one standard error.

A. Motivation

1) Performance Loss Bound: As mentioned previously, the additional advantage term in the AL operator helps mitigate the influence of approximation errors. However, this term may also be detrimental to the learning of the value function. For example, when the true optimal action a^* is mistaken for suboptimal actions by the current iterated value function, the AL operator provides an underestimated bootstrapping target for its next iteration because of the negative advantage term ($A(s,a^*)<0$), leading to its slower convergence to the optimal state value $V^*(s)$. To support this observation, we analyzed the performance loss bound of the AL operator as follows (see the Appendix for the complete proof).

Lemma 1: Denote the optimal policy by π^* and the optimal state value function by V^*. $\forall Q_0 \in \mathcal{Q}$ and $\pi \in \Pi$, letting $Q_{k+1}(s,a) = \mathcal{T}_{\mathrm{AL}} Q_k(s,a)$, $V_k(s) = \max_{a\in\mathcal{A}} Q_k(s,a)$ and assuming $\|V^\pi\|_\infty \le V_{\max}$. Define the "*optimality mismatch error*": $\Delta_k^{\pi^*} \in \mathbb{R}^{|\mathcal{S}|}$ whose each entry is represented as: $\Delta_k^{\pi^*}(s) = V_k(s) - Q_k(s,\pi^*(s))$. Then, for the Kth iteration, we have

$$\|V^* - V^{\pi_{K+1}}\|_\infty \le \frac{2\gamma}{1-\gamma}\left[2\gamma^{K+1}V_{\max} + \alpha\sum_{k=0}^{k=K-1}\gamma^{K-k-1}\left\|\Delta_k^{\pi^*}\right\|_\infty\right].$$

Lemma 1 measures the corresponding V-value distance between the induced greedy policy π_{K+1} and optimal policy π^*. Because the induced policy is defined as $\pi_{K+1}(s) = \arg\max_{a\in\mathcal{A}}[r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_K(s')]]$, this bound actually reflects the distance of Kth iteration V_K from V^*. Note that the performance loss bound of the AL operator will accumulate an extra error term $\Delta_k^{\pi^*}$ at each iteration. Recall the definition $\Delta_k^{\pi^*} = V_k(s) - Q_k(s,\pi^*(s))$, and we know that this error will become nonzero if and only if the induced optimal action and true optimal action do not match [i.e., $\pi^*(s) \neq \arg\max_{a\mathcal{A}} Q_k(s,a)$], so we call it "optimality mismatch error." Because this mismatch continues to exist until the agent learns the optimal policy, the AL operator accumulates more errors than the Bellman optimality operator, which implies slower convergence to the optimal value function V^*.

2) Toy Example Illustration: To better understand the advantages and disadvantages of both operators, we illustrate their learning properties using an 11-state Chainwalk toy example. As shown in Fig. 1(a), the agent can move either left or right at each state and would be transitioned to the state in the intended direction with probability 0.7, while to the state in the opposite direction with probability 0.3. At both ends of the chain, attempted movement to outside of the chain results in staying at the ends. In addition, the agent gets 0 reward once reaching the middle state (s_5). If the agent moves to the right side of the chain ($s_6 - s_{10}$), it can get 1 reward; otherwise, it gets -1 reward on the left side of this chain ($s1 - s4$) except 3 reward on the left end (s_0).

According to these settings, the optimal policy is to choose the "L(eft)" action at all states so that the agent can reach and stay in the leftmost state as soon as possible. Fig. 1(b) shows the results of numerical experiments on both performance loss and action gap. We can see that although the AL operator (blue line) can achieve larger action gaps, the Bellman optimality operator (green line) can obtain a faster decrease in performance loss, indicating faster value convergence. These empirical results verify the conclusions in Property 2 and Lemma 1 and suggest that the robustness of the AL operator comes at the cost of slower convergence.

B. Adaptive AL

To balance faster convergence and larger action gaps for the value function, we propose a novel adaptive AL operator in this section. In particular, we adaptively adjust the scaling factor of the advantage term in the original AL operator based on the need to further increase the action gap for each state–action pair. Specifically, the advantage term is amplified only when the Q value of any action $Q(s,a)$ approaches that of optimal actions [i.e., $V(s)$], enhancing robustness by enlarging the action gap. Otherwise, we weaken its effect to promote faster convergence of the value function. We summarize this intuitive idea as "advantage term must not be multiplied beyond necessity," similar to Occam's Razor principle, and name it ORAL. Next, we present two ways to realize this idea.

1) ORAL-Clip: First, we choose a simple clipping function to control the presence or absence of the scaling advantage term, depending on the closeness of $Q(s,a)$ to $V(s)$. We abbreviate this method as "ORAL-clip" and the corresponding operator is defined as follows:

$$\mathcal{T}_{\mathrm{ORAL}-c}\, Q(s,a) \triangleq \mathcal{T}Q(s,a) + \alpha\left(Q(s,a) - V(s)\right)\cdot\mathbb{I}\left[c(s,a)\ge\beta\right]. \quad (5)$$

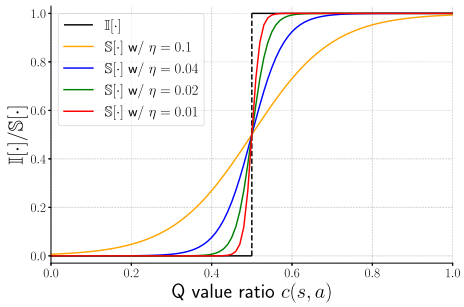


Fig. 2. Comparison between $\mathbb{I}[\cdot]$ and $\mathbb{S}[\cdot]$ with the ratio threshold $\beta = 0.5$.

Note that $c(s,a) = (Q(s,a) - Q_l)/(V(s) - Q_l)$ is the Q-value ratio between the current action and optimal action, where Q_l is a lower bound used to avoid a negative ratio. For simplicity, we set Q_l to the theoretical minimum of the discounted sum of rewards $Q_l = 1/(1-\gamma)R_{\min}$ during practical implementation. The ratio threshold $\beta \in (0,1)$ represents the degree of closeness that determines whether to add the advantage term or not. The indicator function $\mathbb{I}[\cdot]$ plays the role of clipping function.

Furthermore, we can view $\mathcal{T}_{\text{ORAL}-c}$ as a "*two-stage*" piecewise operator and rewrite it in a more direct way

$$\mathcal{T}_{\text{ORAL}-c} = \begin{cases} \mathcal{T}_{\text{AL}}, & \text{if } Q(s,a) - Q_l \geq \beta(V(s) - Q_l) \\ \mathcal{T}, & \text{otherwise.} \end{cases} \tag{6}$$

According to the expression in (6), we can see that $\mathcal{T}_{\text{ORAL}-c}$ keeps the same with $\mathcal{T}_{\text{AL}}$ when the Q value of any action exceeds a certain threshold and approaches that of optimal actions; otherwise, $\mathcal{T}_{\text{ORAL}-c}$ adopts $\mathcal{T}$ for the Q-value iteration. Through this combination $\mathcal{T}$ and $\mathcal{T}_{\text{AL}}$, we aim to mitigate the blind action gap increasing of the AL operator so that $\mathcal{T}_{\text{ORAL}-c}$ can speed up the value convergence while maintaining appropriate action gaps.

2) ORAL-Sigmoid: Despite its simplicity, $\mathcal{T}_{\text{ORAL}-c}$ suffers from sudden shifts in the applied operator near the ratio threshold owing to the nonsmooth nature of the indicator function. Such frequent operator switching can easily cause instability learning in value functions. For example, Fig. 1(b) illustrates the changes in the action gap over the iteration process. We can see that $\mathcal{T}_{\text{ORAL}-c}$ (pink line) results in greater fluctuations compared to both $\mathcal{T}$ (green line) and $\mathcal{T}_{\text{AL}}$ (blue line), indicating the instability in value function learning. To address this issue, we propose an alternative method, named "ORAL-Sigmoid" that approximates the indicator function in (5) with a smooth Sigmoid function. Its specific definition is as follows:

$$\begin{aligned} &\mathcal{T}_{\text{ORAL}-s}\, Q(s,a) \\ &\quad \triangleq \mathcal{T} Q(s,a) + \alpha\left(Q(s,a) - V(s)\right) \cdot \mathbb{S}\left[c(s,a), \beta, \eta\right] \end{aligned} \tag{7}$$

where the Sigmoid function is denoted by $\mathbb{S}[x,\beta,\eta] = 1/(1 + e^{-(x-\beta)/\eta})$, which includes an extra temperature parameter $\eta \in (0, +\infty)$. Fig. 2 shows the effects of different temperature parameters. We can see that this critical parameter controls the degree of the smooth approximation of the indicator function. We choose the Sigmoid function as an alternative to the indicator function mainly because: 1) the Sigmoid function is a more generable indicator function form, and $\mathcal{T}_{\text{ORAL}-c}$ is a special case of $\mathcal{T}_{\text{ORAL}-s}$ where $\eta \to 0$; 2) the monotonic increasing nature of the Sigmoid function aligns with the motivation behind our Occam's Razor principle, which states that the advantage term is amplified only with a larger Q-value ratio $c(s,a)$; and 3) the range of the Sigmoid function [i.e., $\mathbb{S}[\cdot] \in (0,1)$] satisfies the conditions specified in Theorem 1 ([12]), guaranteeing the optimality-preserving and gap-increasing properties of $\mathcal{T}_{\text{ORAL}-s}$.

3) Operator Comparisons: We also compare the proposed methods in the case of the Chainwalk toy example. As shown in Fig. 1(b), both "ORAL-clip" (pink line) and "ORAL-Sigmoid" (red line) can converge to the optimal value function faster than the original AL operator while maintaining larger action gaps than the Bellman optimality operator. These results demonstrate that our proposed methods can achieve a balance between large action gaps and fast convergence.

Notably, when comparing the curves of action gaps, Fig. 1(b) also shows that "ORAL-Sigmoid" (red line) can increase the action gap more steadily than "ORAL-clip" (pink line) does, which verifies the stable value function benefited from the smooth Sigmoid function used in (7).

C. Extensions to Other AL-Based Methods

We have proposed some general designs on adaptive AL in Section IV-B, and these methods can be easily plugged into other off-the-shelf AL-based algorithms. Next, we mainly introduce two extensions to our proposals.

1) Persistent AL: The infrequent switch between actions is believed to benefit the domain of high temporal resolution, so Bellemare et al. [12] further proposed the PAL operator whose definition is as follows:

$$\mathcal{T}_{\text{PAL}} Q(s,a) \triangleq \max\left\{\mathcal{T}_{\text{AL}} Q(s,a), r(s,a) + \gamma \mathbb{E}_{s'}\left[Q\left(s', a\right)\right]\right\}. \tag{8}$$

According to the definition, if a higher bootstrapped target was achieved with the persistence of the same action a at the next state s', $\mathcal{T}_{\text{PAL}}$ will choose it as the iterated Q value; otherwise, $\mathcal{T}_{\text{PAL}}$ will keep consistent with $\mathcal{T}_{\text{AL}}$. By substituting the component of the AL operator in (8) with our proposed ORAL operators, we can directly define the "*PAL-clip*" and "*PAL-Sigmoid*" methods

$$\begin{aligned} &\mathcal{T}_{\text{PAL}-\text{c/s}} Q(s,a) \\ &\quad \triangleq \max\left\{\mathcal{T}_{\text{ORAL}-\text{c/s}} Q(s,a), r(s,a) + \gamma \mathbb{E}_{s'}\left[Q\left(s', a\right)\right]\right\}. \end{aligned} \tag{9}$$

2) Munchausen RL: Within the framework of entropy-regularized MDP, the only modification by the Munchausen RL (MRL) [16] is to reshape the immediate reward with the scaled log-policy term, and its operator can be written as

$$\begin{aligned} \mathcal{T}_{\text{MRL}} Q(s,a) \triangleq\ & r(s,a) + \alpha\left[\tau \ln \pi(a|s)\right]_{l_0}^{0} \\ & + \gamma \mathbb{E}_{\substack{s' \sim P(\cdot|s,a) \\ a' \sim \pi(\cdot|s')}}\left[Q\left(s', a'\right) - \tau \ln \pi\left(a'\middle|s'\right)\right] \end{aligned} \tag{10}$$

where τ is the temperature parameter scaling the entropy and $\alpha \in [0,1]$ is a scaling factor of the shaping reward $[\tau \ln \pi(a|s)]_{l_0}^{0}$, where $[\cdot]_y^x$ is the clipping function to ensure that the log-policy term is within a proper range $[l_0, 0]$. According to the closed solution in entropy-regularized MDP, the policy update satisfies: $\pi(a|s) = (\exp Q(s,a)/\tau)/\left(\sum_{a'} \exp Q(s,a')/\tau\right)$. Then, the additional scaled log-policy term in (10) is equivalent to $\alpha\left(Q(s,a) - \tau \ln \sum_{a'} \exp Q(s,a')/\tau\right)$, i.e., a scaled and soft advantage function term.

Analogously, we can apply the clipping and Sigmoid functions to the scaled log-policy term and name the extended

methods as "MRL-clip" and "MRL-Sigmoid," respectively. The reshaped reward in (10) will be modified as

$$\begin{aligned}\textbf{MRL-clip:}\quad & \alpha\left[\tau\ln\pi\left(a|s\right)\right]_{l_0}^{0}\cdot\mathbb{I}\left[c\left(s,a\right)\geq\beta\right]\\ \textbf{MRL-Sigmoid:}\quad & \alpha\left[\tau\ln\pi\left(a|s\right)\right]_{l_0}^{0}\cdot\mathbb{S}\left[c\left(s,a\right),\beta,\eta\right].\end{aligned}$$

D. Extension to Continuous-Action Tasks

Notably, the ORAL methods and their extensions discussed above are naturally suited for discrete-control tasks and are not directly applicable to continuous-control tasks. The primary challenge is how to take the maximum of Q values in a continuous-action space, i.e., computing the V value $V(s)=\max_{a\sim\mathcal{A}}Q(s,a)$.

To tackle this issue, we utilize the NAF [36] technique whose main idea is to represent the Q-function $Q(s,a)$ by two separate value functions, $V(s)$ and $A(s,a)$. Specifically, the advantage function is parameterized as a quadratic function of nonlinear features of the state, leading to an easy and analytical maximum of Q values, $\arg\max_a Q(s,a)$

$$\begin{aligned}Q\left(s,a\right)&=A\left(s,a|\theta_A\right)+V\left(s|\theta_V\right)\\ A\left(s,a|\theta_A\right)&=-\frac{1}{2}\left(a-\mu\left(s|\theta_\mu\right)\right)^{\top}P\left(s|\theta_P\right)\left(a-\mu\left(s\middle|\theta_\mu\right)\right)\end{aligned}$$

where $P(s|\theta_P)$ is a state-dependent, positive-definite square matrix to ensure a nonpositive advantage function, $A(s,a|\theta_A)\leq 0$. Furthermore, this square matrix can be achieved by a parameterized lower triangular matrix, $P(s|\theta_P)\triangleq L(s|\theta_P)L(s|\theta_P)^{\top}$. The policy function $\mu(a|\theta_\mu)$ outputs a deterministic action that can be considered as the maximizer of the $Q(s,a)$ and $A(s,a)$. Following these designs, we can achieve the AL operator in continuous-action tasks by utilizing this explicit advantage function directly:

$$\begin{aligned}&\min_{\theta_V,\theta_A}\mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[\left(Q\left(s,a|\theta_V,\theta_P,\theta_\mu\right)-y_{\mathrm{AL}}\right)^2\right]\\ &y_{\mathrm{AL}}=r\left(s,a\right)+\alpha A\left(s,a\middle|\theta_A^{-}\right)+\gamma V\left(s'\middle|\theta_V^{-}\right)\end{aligned}\tag{11}$$

where θ and θ^{-} represent the parameters of the updated network and target network, respectively. Analogously, we can obtain the updated targets of our "ORAL-clip" and "ORAL-Sigmoid" methods as follows:

$$\begin{aligned}&y_{\mathrm{ORAL-c}}\\ &\quad=r\left(s,a\right)+\alpha A\left(s,a\middle|\theta_A^{-}\right)\cdot\mathbb{I}\left[c\left(s,a\right)\geq\beta\right]+\gamma V\left(s'\middle|\theta_V^{-}\right)\end{aligned}\tag{12}$$

and

$$\begin{aligned}&y_{\mathrm{ORAL-s}}\\ &\quad=r\left(s,a\right)+\alpha A\left(s,a\middle|\theta_A^{-}\right)\cdot\mathbb{S}\left[c\left(s,a\right),\beta,\eta\right]+\gamma V\left(s'\middle|\theta_V^{-}\right).\end{aligned}\tag{13}$$

V. Theoretical Analysis

In this section, we first demonstrate the comparative relationship of quantitative action gaps between different learning operators. Second, we give the convergence analysis of our "ORAL-clip" and "ORAL-Sigmoid" algorithms and show the advantage of the smooth-version ORAL.

A. Optimality Preserving and Gap Increasing

Bellemare et al. [12] state the conditions that the family of operators with these two properties needs to satisfy in their paper, which is summarized as the following theorem.

Theorem 1 [12]: Let $\mathcal{T}$ be the Bellman optimality operator defined by (2). Let $\mathcal{T}'$ be an operator with the property that there exists an $\alpha\in[0,1)$ such that for all $Q\in\mathcal{Q}$, $s\in\mathcal{S}$, $a\in\mathcal{A}$, and letting $V(s)=\max_a Q(s,a)$.

1) $\mathcal{T}'Q(s,a)\leq\mathcal{T}Q(s,a)$.
2) $\mathcal{T}'Q(s,a)\geq\mathcal{T}Q(s,a)+\alpha[Q(s,a)-V(s)]$.

Then, $\mathcal{T}'$ is both *optimality preserving* and *gap increasing*.

According to the definitions in (5) and (7), it is obvious that both $\alpha\cdot\mathbb{I}[c(s,a)\geq\beta]$ and $\alpha\cdot\mathbb{S}[c(s,a),\beta,\eta]$ will lie in the range of $[0,1)$. Thus, we can easily deduce the following corollary.

Corollary 1: Both "ORAL-clip" operator $\mathcal{T}_{\mathrm{ORAL-c}}$ and "ORAL-Sigmoid" operator $\mathcal{T}_{\mathrm{ORAL-s}}$ satisfy the conditions stated in Theorem 1 and then keep *optimality preserving* and *gap increasing*.

Corollary 1 suggests that our methods can still learn the optimal policy and retain the robustness to estimation errors. Furthermore, we attempt to compare the action gaps induced by different operators and first define the action gap of any state–action pair as

$$G\left(s,a\right)=\liminf_{k\to\infty}\left[V_k\left(s\right)-Q_k\left(s,a\right)\right]\tag{14}$$

where $\{Q_k\}_{k=0}^{\infty}$ and $\{V_k\}_{k=0}^{\infty}$ represent the (action) value function sequence iterated by any learning operator and $V_k(s)=\max_a Q_k(s,a)$. Then, we have the following results (see the Appendix for the proof).

Lemma 2: For $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, letting $G^*(s,a), G_{\mathrm{AL}}(s,a)$ and $G_{\mathrm{ORAL}}(s,a)$ represent the action gap induced by $\mathcal{T},\mathcal{T}_{\mathrm{AL}}$ and $\mathcal{T}_{\mathrm{ORAL}}$,[2] respectively, then these action gaps satisfy

$$G^*\left(s,a\right)\leq G_{\mathrm{ORAL}}\left(s,a\right)\leq G_{\mathrm{AL}}\left(s,a\right).$$

Lemma 2 demonstrates that action gaps induced by our methods are between those of the Bellman optimality operator and the AL operator. This is intuitive because $\mathcal{T}_{\mathrm{ORAL}}$ is essentially a composite operator between $\mathcal{T}$ and $\mathcal{T}_{\mathrm{AL}}$ as shown in (6). Also, the theoretical analysis is consistent with the empirical results in Fig. 1(b), that is, our proposed operator can achieve a balance between larger action gaps and faster convergence.

B. Convergence Analysis

In addition to optimality-preserving and gap-increasing properties, we also focus on the convergence property of our methods. We analyze their convergence conditions in this section and provide all the proofs in the Appendix.

As mentioned before, the V-value sequences obtained by both $\mathcal{T}_{\mathrm{ORAL-c}}$ and $\mathcal{T}_{\mathrm{ORAL-s}}$ will converge to the optimal V value, V^*, due to their optimality-preserving property. Therefore, whether the action gaps converge or not equally reflects the convergence property of the Q-value sequence. For convenience, we mainly focus on the convergence of the iterated action gap, whose definition is as follows:

$$G_k\left(s,a\right)\triangleq V^*\left(s\right)-Q_k\left(s,a\right)\quad\forall s,a\in\mathcal{S}\times\mathcal{A}.\tag{15}$$

[2]This comparison result will hold true for both $\mathcal{T}_{\mathrm{ORAL-c}}$ and $\mathcal{T}_{\mathrm{ORAL-s}}$, so we use $\mathcal{T}_{\mathrm{ORAL}}$ represent them uniformly.

We first conduct an in-depth analysis on the convergence property of $\mathcal{T}_{\text{ORAL-c}}$ operator, whose convergence condition is claimed in Theorem 2 (see the Appendix for the proof).

Theorem 2: Suppose that $\{G_k\}_{k=0}^{\infty}$ is the sequence iterated by $\mathcal{T}_{\text{ORAL-c}}$. When $Q_l \le \min_{s,a}(Q^* - V^*)/(1-\alpha)$, $\forall s, a \in \mathcal{S} \times \mathcal{A}$

$$\lim_{k\to\infty} G_k(s,a) = \begin{cases} V^*(s) - Q^*(s,a), & \text{if} \beta \in \left(\max_{s,a} \dfrac{Q^* - Q_l}{V^* - Q_l}, 1\right) \\ \dfrac{(V^*(s) - Q^*(s,a))}{1-\alpha}, & \\ & \text{if } \beta \in \left(0, \min_{s,a} \dfrac{Q^* - \alpha V^* - (1-\alpha) Q_l}{(1-\alpha)(V^* - Q_l)}\right]. \end{cases}$$

As stated in Theorem 2, the Q value learned by $\mathcal{T}_{\text{ORAL-c}}$ can converge to the fixed point of either the Bellman optimality operator $\mathcal{T}$ or the AL operator $\mathcal{T}_{\text{AL}}$ in some cases. Except for these conditions, $\mathcal{T}_{\text{ORAL-c}}$ will lead to a Q-value fluctuation between the results from $\mathcal{T}$ and $\mathcal{T}_{\text{ORAL-c}}$. This conclusion makes sense due to the nature of the piecewise operator of the $\mathcal{T}_{\text{ORAL-c}}$. In addition, we further concern the convergence conditions for a more general formulation shown in Theorem 3 (see the Appendix for the proof).

Theorem 3: Define $f_k(s,a) \doteq f(G_k(s,a))$, where $f(\cdot) : \mathbb{R} \to [0,1]$ is continuous and differentiable with respect to $G_k(s,a), \forall k \in \mathbb{N}, (s,a) \in \mathcal{S} \times \mathcal{A}$. Let $\alpha \in [0,1)$ and $\mathcal{T}'$ be an operator that satisfies $\mathcal{T}'Q(s,a) = \mathcal{T}Q(s,a) + \alpha(Q(s,a) - V(s)) \cdot f(s,a)$; then, the sequence $\{G_k\}_{k=0}^{\infty}$ iterated by $\mathcal{T}'$ will be convergent if $f(\cdot)$ satisfies

$$\left\| \frac{d(f_k \cdot G_k)}{dG_k} \right\|_{\infty} \le \frac{1}{\alpha} \quad \forall k \in \mathbb{N}.$$

Notably, Theorem 3 provides a sufficient condition for the convergence of value functions learned by a general adaptive AL operator. From this, we can derive the following insights.

Remark 1: According to Theorem 3, if $f_k(s,a) = \mathbb{S}[c(s,a), \beta, \eta]$, the convergence condition can be represented as: $\|1-(G_k \cdot f_k \cdot (1-f_k))/(\eta \cdot (V^* - Q_l))\|_{\infty} \le 1/\alpha$. Note that a small enough η may amplify $(G_k \cdot f_k(1-f_k))/(V_k - Q_l)$ largely, causing it to deviate a lot from 1 and easily violate this convergence condition. Thus, our $\mathcal{T}_{\text{ORAL-s}}$ with a proper η is more likely to achieve the value convergence compared to our $\mathcal{T}_{\text{ORAL-c}}$ with $\eta \to 0$.

VI. Experiments

In this section, we first introduce the experiment settings and the details about implementation and evaluation. Then, we will analyze and compare the specific experimental results, so as to verify the effectiveness of our methods.

A. Experimental Settings

1) Benchmarks: For the discrete-control tasks, we evaluate all the methods in the popular *MinAtar* [37] and *Atari* [22] benchmarks. The former includes several simplified Atari games that not only retain the general mechanics of specific Atari games but also simplify the representational complexity to focus more on the behavioral challenges. As for the complex Atari games, we choose to compare the performance across a subset of 15 original Atari tasks. For the continuous-control tasks, we mainly implement the experiments in some *Mujoco* locomotion tasks [23] for a sufficient verification of the effectiveness of our methods.

2) Baselines: In addition to the AL-based algorithms (AL, PAL, and MRL), we also compare our methods with some classical algorithms for comprehensive comparison, including: a) *DDQN* [38]: a specific adaptation to the DQN algorithm with double Q-learning; b) *SoftDQN*: a discrete-action version of soft action-critic [39]; c) *Ensemble-based DQN*: a class of special DQN algorithms utilizing ensemble Q-functions [40], e.g., *EnsembleDQN*, *AveragedDQN*, and *MaxminDQN*; d) *MeDQN* [41]: a memory-efficient DQN algorithm by learning new knowledge and consolidating old knowledge; e) *VEB* [42]: a value evolutionary-based RL that focuses on the integration of evolutionary algorithms with value-based RL; and f) *EBO* [43]: special operators that map a distribution over Q values to the pushforward of regular Bellman operators with additive noise.

3) Implementation: All the agents are modeled as a two-layer perceptron, and we train them according to the settings in the reference paper [40]. For these non-AL-based baselines, we mainly fine-tune their learning rate from a candidate set $\{1e-3, 3e-3, 1e-4, 3e-4, 1e-5\}$, and except for that the number of networks for all ensemble-based baselines is selected from $\{6, 7\}$. For the naive AL-based methods, we choose the recommended scaling factor $\alpha = 0.9$ for AL and PAL agents and set the temperature parameter $\tau = 0.03$ for the SoftDQN agent. The above two parameters are also chosen as the Munchausen scaling term and the entropy temperature term in the MRL agent. As for our presented methods, we keep the same scaling factor $\alpha = 0.9$ and fine-tune the ratio threshold β from the candidate set $\{0.5, 0.6, 0.7, 0.8\}$ and the temperature parameter $1/\eta$ from $\{20, 25, 30\}$ for these "*-Sigmoid*" methods while fixing $\beta = 0.8$ for those "*-clip*" methods. Besides, we set the lower bound as $Q_l = 1/(1-\gamma)R_{\min}$, i.e., the least discounted sum of rewards for an infinite-length trajectory. More implementation details can be found in our Supplementary Materials.

B. Algorithm Evaluation

1) Performance Comparison: First, we compare the DQN-normalized scores between the chosen baselines and our methods. The DQN-normalized score is defined as $s-r/(|b-r|)$, where s represents the score of the compared algorithm, b is the score of the baseline, and r is the score of a random policy. Table I presents all the quantitative results on the MinAtar tasks, with bold and underlined fonts indicating the best and second-best results, respectively. Fig. 3 includes the learning curves of our ORAL methods and other chosen baselines.

As shown in Table I, our methods achieve the best or second-best performance across nearly all the tasks, with the exception of the Asterix task. Notably, even on the Asterix task, these AL-based algorithms equipped with our Occam's Razor mechanism can demonstrate significant improvements compared to their naive counterparts, regardless of whether the clipping or Sigmoid function is employed to achieve the adaptive gap increasing. The last column illustrates the average ranking of these methods across all the tasks, from which we can see that the top-3 methods in the average ranking all belong to our adaptive methods achieved by the Sigmoid function, i.e., PAL-Sigmoid, ORAL-Sigmoid, and MRL-Sigmoid.

Besides, Fig. 3 shows that our *"-clip"* and *"-Sigmoid"* methods (red and blue lines) generally exhibit better performance and higher sample efficiency than other baselines on

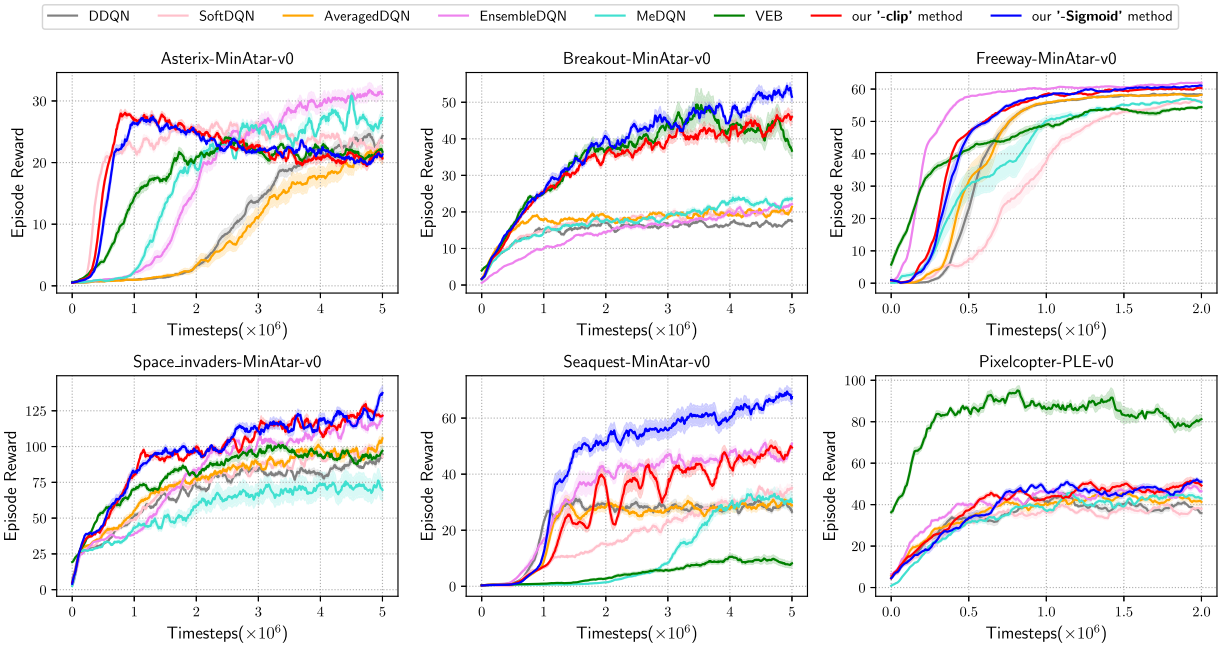


Fig. 3. Learning curves of our methods (taking the best "-clip" and "-Sigmoid" methods) and some chosen baselines on the MinAtar benchmarks. All the results are averaged over five random seeds, with the shade area corresponding to one standard error.

TABLE I

DQN-NORMALIZED SCORE COMPARISONS BETWEEN DIFFERENT ALGORITHMS, WHICH MEANS THE PERCENTAGE (%) OF PERFORMANCE IMPROVEMENT THAN THE DQN BASELINE. ALL RESULTS ARE AVERAGED OVER FIVE SEEDS, AND ONE STANDARD DEVIATION IS INCLUDED IN THE PARENTHESES

Algorithms	Asterix	Breakout	Freeway	Space-Invaders	Seaquest	Pixelcopter	Avg. Rank
DDQN	40.25 (±0.16)	-10.06 (±0.12)	-0.49 (±0.01)	0.08 (±0.10)	36.98 (±0.35)	13.61 (±0.14)	12
SoftDQN	36.59 (±0.20)	5.28 (±0.17)	-3.96 (±0.02)	-0.89 (±0.11)	74.56 (±0.25)	19.52 (±0.20)	11.5
AveragedDQN	21.74 (±0.32)	9.47 (±0.17)	-0.73 (±0.01)	11.28 (±0.16)	49.15 (±0.39)	29.30 (±0.30)	11.5
MaxminDQN	18.01 (±0.30)	-52.06 (±0.07)	-75.18 (±0.43)	-33.25 (±0.06)	-100.00(±0.00)	46.81(±0.20)	13.67
EnsembleDQN	**82.08 (±0.35)**	14.46 (±0.26)	**5.62 (±0.01)**	28.93 (±0.18)	153.96 (±0.32)	42.01 (±0.12)	5.33
MeDQN	59.79 (±0.42)	23.64 (±0.27)	-4.46 (±0.02)	-25.78 (±0.37)	55.47 (±0.65)	31.61 (±0.05)	10.83
VEB	24.03 (±0.20)	91.22 (±0.58)	-5.88 (±0.01)	4.17 (±0.21)	-58.33 (±0.29)	**133.69 (±0.19)**	9
EBO	-97.99 (±0.03)	-70.27 (±0.03)	-72.26 (±0.03)	-78.66 (±0.03)	-69.61 (±0.07)	-88.07 (±0.02)	16.67
AL	-1.6 (±0.14)	-4.94 (±0.17)	2.52 (±0.01)	23.02 (±0.16)	62.75 (±0.64)	39.77 (±0.17)	11.17
ORAL-clip	6.9 (±0.17)	92.34 (±0.27)	2.1 (±0.02)	27.88 (±0.18)	100.39 (±0.36)	44.79 (±0.10)	8.17
ORAL-Sigmoid	23.38 (±0.26)	79.97 (±0.30)	3.84 (±0.01)	**47.08 (±0.25)**	232.56 (±0.77)	49.81 (±0.12)	4
PAL	9.05 (±0.17)	111.69 (±0.28)	3.53 (±0.01)	22.75 (±0.10)	90.02 (±0.31)	44.01 (±0.21)	8
PAL-clip	4.37 (±0.09)	136.86 (±0.39)	2.95 (±0.01)	27.51 (±0.21)	109.43 (±0.14)	51.05 (±0.26)	6.67
PAL-Sigmoid	13.13 (±0.20)	**172.74 (±0.54)**	4.14 (±0.01)	40.71 (±0.20)	182.67 (±0.67)	55.84 (±0.25)	**3.5**
MRL	12.46 (±0.16)	63.06 (±0.29)	2.03 (±0.01)	20.88 (±0.22)	136.89 (±0.41)	31.22 (±0.21)	10
MRL-clip	21.68 (±0.29)	96.11 (±0.36)	2.45 (±0.02)	29.95 (±0.15)	151.92 (±0.45)	44.37 (±0.28)	6.17
MRL-Sigmoid	21.15 (±0.14)	84.19 (±0.27)	3.73 (±0.01)	36.28 (±0.10)	**244.23 (±0.80)**	49.12 (±0.29)	4.83

most tasks except for the "Asterix" and "Pixelcopter" tasks. Notably, we observed that, unlike non-AL-based methods, our methods exhibit a rapid performance increase in the early stages, which gradually slows down until convergence on the "Asterix" task. We hypothesize that this is due to the dynamic nature of the "Asterix" task itself. In particular, its difficulty is periodically increased by increasing the speed and spawn rate of enemies and treasure. AL-based methods, while able to quickly learn more robust decisions by increasing the action gap early on, also tend to make the agent overconfident in its early optimal decision, making it more difficult to update that optimal decision as the environment changes and eventually leading to suboptimal decisions. Regarding the "Pixelcopter" task, although our method does not outperform the VEB method combined with evolutionary algorithms (EAs), it still achieves superior performance compared to other DQN-based baselines while avoiding the complex computations associated with EAs.

To further validate the effectiveness of our proposed methods, we compare the normalized improvement of our "ORAL-clip" and "ORAL-Sigmoid" methods to the naive AL algorithm on the more complex benchmark. We conduct these comparisons mainly on a subset of Atari tasks, and similar to but slightly different from the baseline-normalized score, the normalized improvement is defined as $s - b/(|b - r|)$. Fig. 4 reports the per-game improvement after 1M training steps. As

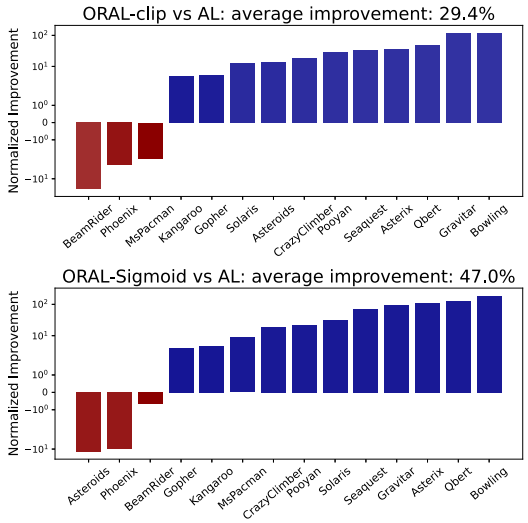


Fig. 4. Normalized improvement of "ORAL-clip" (top) and "ORAL-Sigmoid" (bottom) methods to the original AL algorithm on a subset of Atari tasks.

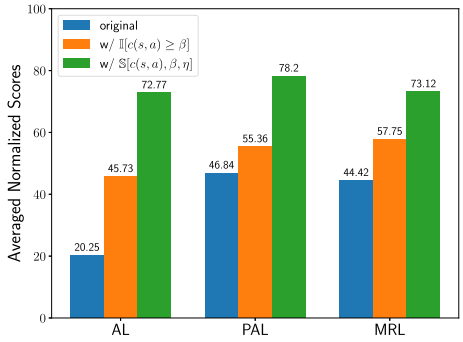


Fig. 5. Performance comparison on the MinAtar benchmark for several AL-based methods *w/* or *w/o* our proposed Occam's Razor mechanism.

shown, both our "ORAL-clip" and "ORAL-Sigmoid" methods achieve performance improvements in the majority of Atari games (11 out of 14 games) compared to the naive AL algorithm. More precisely, "ORAL-clip" demonstrates a 29.4% improvement averaged over the entire subset of tasks, while "ORAL-Sigmoid" achieves a larger 47.0% improvement. For more learning curves, please refer to our Supplementary Materials.

These empirical results from both the MinAtar and Atari benchmarks highlight that our adaptive AL methods, grounded in the principle of Occam's Razor, significantly enhance the agent's performance. Furthermore, our ORAL method, implemented through a smooth Sigmoid function, typically yields superior performance improvements alongside more stable learning.

2) Extensions to Other AL-Based Methods: As mentioned in Section IV-C, we attempt to extend our proposals to some other AL-based methods, including both PAL and MRL algorithms. Performance results about these naive AL-based methods and their variants combined with our Occam's Razor mechanism are summarized in Table I (please refer to our Supplementary Materials for the learning curves). Specifically, the adaptive AL variants based on the clipping function (with "*-clip*" methods) can achieve performance improvements on almost all tasks (except for a few cases on Asterix and Freeway tasks). On the other hand, those variants that adopt a Sigmoid function to achieve the adaptive gap increasing can outperform their naive versions across all the tasks, showing more effectiveness of the smooth and adaptive gap increasing.

For a more intuitive comparison, Fig. 5 illustrates the quantitative results of performance enhancements among various AL-based algorithms and their corresponding variants that adhere to our Occam's Razor principle. We can see that either a clipping function or a Sigmoid function is chosen to achieve the adaptive AL; they can achieve a remarkable performance improvement compared to their naive AL-based algorithm. Especially for the methods equipped with the Sigmoid function, they can achieve over 70.0 normalized scores averaged on the MinAtar tasks, which greatly exceed the performance level of their naive version (around 20.0 ~ 45.0 scores). Notably, our adaptive mechanism can achieve greater effectiveness when applied to the AL algorithm than when applied to PAL or MRL algorithms. We believe that this is attributed to the max operator used in PAL (8) and the clipping function on the log-policy term used in MRL (10), they can also alleviate the blindness to some extent when implementing gap increasing, although not their original intention.

TABLE II

AVERAGED SCORE COMPARISON BETWEEN OUR ENHANCED METHODS AND THE NAIVE AL METHOD ON SEVERAL MUJOCO TASKS AFTER 500K TRAINING STEPS

Alg. \ Task	Pusher	Reacher	Inverted-Pendulum	InvertedDouble-Pendulum
NAF	-35.71 (± 0.22)	-5.77 (± 0.08)	841.04 (± 83.83)	148.51 (± 3.62)
NAF + AL	-34.24 (± 0.24)	-5.92 (± 0.16)	818.42 (± 95.12)	153.75 (± 2.25)
NAF + **ORAL-clip**	-33.71 (± 0.43)	-5.32 (± 0.10)	837.96 (± 79.00)	155.31 (± 3.71)
NAF + **ORAL-Sigmoid**	-34.15 (± 0.26)	-5.59 (± 0.06)	985.53 (± 13.21)	153.47 (± 4.03)
SAC	**-24.04** (± 0.07)	**-3.56** (± 0.02)	**1000.00** (± 0.00)	**9150.43** (± 190.10)
PPO	-61.31 (± 5.56)	-7.15 (± 0.70)	917.53 (± 18.33)	6036.81 (± 338.18)

3) Extensions to Continuous-Control Tasks: We further conduct performance comparisons between NAF, NAF + AL, NAF + ORAL, and standard baselines (SAC [39] and PPO [10]) on several continuous-control Mujoco tasks and Table II summarizes all performance scores averaged over five seeds.

As shown in Table II, our ORAL-Sigmoid method equipped with the NAF technique can enhance the scores significantly compared to the original NAF method. However, with the support of the same NAF method, our ORAL-Sigmoid method can only attain a competitive or slight improvement compared to the naive AL algorithm on some tasks (e.g., "Pusher" and "InvertedDoublePendulum"). Instead, the simpler ORAL-clip method achieves more outstanding improvements over all tasks. This suggests that our ORAL-Sigmoid method requires

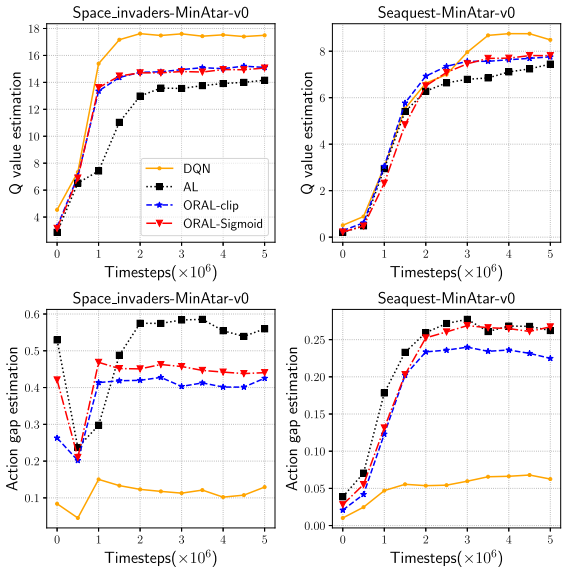


Fig. 6. Estimation Comparisons of Q value and action gap between different algorithms during the learning process.

more careful fine-tuning of the extra temperature parameter in continuous-action tasks.

Meanwhile, limited by the expressive power of the NAF technique, our method achieves suboptimal results compared to the state-of-the-art method (SAC), while compared to another classical baseline (PPO), our method demonstrates significant performance improvements across all tasks except the "InvertedDoublePendulum" task, validating the effectiveness of our Occam's Razor mechanism in the continuous-control scenes and also inspiring us to seek more expressive ways to extend this mechanism in the future.

C. Property Analysis

1) Balance Between Gap Increasing and Value Convergence: Besides better task performance, we attempt to verify the effectiveness of our proposed adaptive AL methods from the perspective of some vital algorithmic properties. First, we compared the changes in action gaps and the Q values of different algorithms during the learning process. Fig. 6 illustrates the comparisons of estimations of Q value and action gap on two MinAtar tasks (Space-Invaders and Sesuest), where the curves represent the averaged results over five seeds.

In general, as shown in Fig. 6, the proposed ORAL method can obtain the intermediate action gaps (bottom subfigures), which are larger than the ones in the DQN algorithm but smaller than those in the AL algorithm. This result is also consistent with our theoretical analysis in Lemma 2. While comparing the Q-value estimations (top subfigures), although a slower convergence of Q values than the DQN algorithm does, our ORAL methods obviously enhance the convergence speed relative to the AL method, which implies that our Occam's Razor mechanism does promote faster value convergence by avoiding blind action gap increasing.

Specifically, when we focus on the difference between ORAL-clip and ORAL-Sigmoid methods, we find that ORAL-Sigmoid has larger action gaps than ORAL-clip with a similar level of Q-value convergence. This reveals another potential advantage of ORAL-Sigmoid besides the stable learning: being able to achieve both larger action gaps and faster value convergence than the ORAL-clip method, simultaneously.

2) Effects on Action Gaps: In this section, we mainly focus on the effects of our methods on the action gap. Fig. 7(a) depicts the changes in the action gaps for our ORAL-clip method with different combinations of α and β parameters. All numeric results are averaged over the last five evaluations (approximate convergence stage). As shown in Fig. 7(a), we can observe that the action gaps basically become larger with the increase of the scaling factor α and the decrease of ratio threshold β. This is intuitive because α determines the strength of the nonpositive advantage term in (3), which is used to increase the action gap. Thus, α exhibits a positive correlation with the magnitude of action gaps. In contrast, a smaller β means more possibilities for applying the AL operator to update the Q-function according to the definition in (6), leading to larger action gaps.

Notably, the change in action gaps caused by the change in α (vertical axes of the heatmap) is much higher than the change in action gaps caused by the change in β (horizontal axes of the heatmap). Therefore, the adjustment of α can be considered as a coarse-grained tune of action gaps, whereas β can achieve a more fine-grained tune on action gaps.

Compared to our ORAL-clip method, our ORAL-Sigmoid method substitutes the nonsmooth clipping function with a Sigmoid function, aiming to achieve more stable learning of the Q-value function. Previous empirical results have shown its superior performance to the ORAL-clip method, and we further validate whether it can stabilize Q-function learning. Fig. 7(b) illustrates the box plot of the last 100 estimations of the action gaps for both the ORAL-clip and ORAL-Sigmoid methods. We can see that our ORAL-clip method is prone to lead to a larger fluctuation of action gaps than our ORAL-Sigmoid method, which means less stability of Q-function during the learning process. This also aligns with the results of the toy example shown in Fig. 1(b), verifying the effectiveness of our smooth-version ORAL.

3) Sensitivity on β and η: As defined in (7), both β and η are critical to the final performance of the ORAL-Sigmoid method. We further compare the performance difference between various options on both hyperparameters and show the results in Fig. 7(c). We can observe that although the performance of our ORAL-Sigmoid exhibits minor differences under various candidate hyperparameters, the overall trend of performance improvement remains consistent. More importantly, all choices can lead to a significant enhancement in performance compared to the original AL method. Therefore, within a rational set of candidate parameters selected for fine-tuning, our ORAL-Sigmoid method is less sensitive to the hyperparameter choices.

VII. Conclusion

Robustness has been extensively studied in RL [44], [45], where it plays an important role in real-world control tasks. AL is a representative gap-increasing algorithm that enhances the robustness of value functions against approximation errors, but current research often overlooks the cost of other properties behind the robustness. This article is dedicated to a more balanced AL operator by an adaptive gap-increasing mechanism.

First, our performance loss analysis demonstrates the negative impact of the gap-increasing nature of the AL operator on value function convergence, which justifies our adaptive gap-increasing motivation.

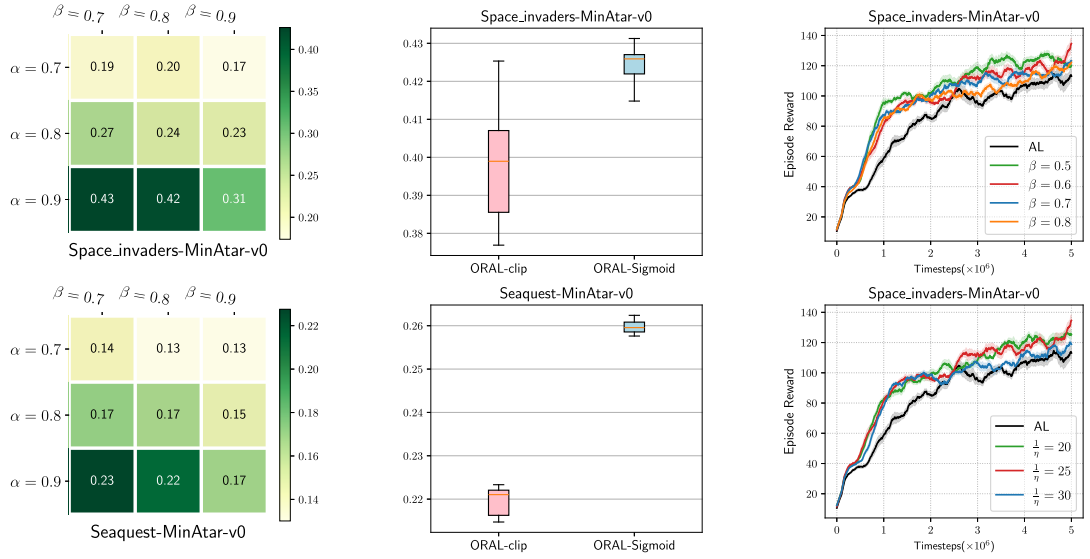


Fig. 7. (a) Effects of α and β on action gaps in ORAL-clip method. (b) Distributions of action gap estimations of the last 100 evaluations. (c) Performance comparison between different β and η values.

Second, inspired by the Occam's Razor principle, we consider adaptively increasing the action gap based on the proximity of the Q value of any action to that of optimal actions. Furthermore, the proposed ORAL solutions realize the adaptive adjustment of the scaling advantage term through the indicator and the Sigmoid function. Comparisons in a toy example and property analysis demonstrate that our methods can achieve a balance between larger action gaps and faster convergence of value functions, verifying the rationality of our designed methods.

Finally, we further extended our methods to other AL-based algorithms, such as PAL and MRL, and proposed to integrate our ORAL methods with the NAF technique, thereby overcoming the application limitations in more complex continuous-control tasks. Empirical results demonstrated the ability of our ORAL methods to balance larger action gaps with faster convergence, as well as their effectiveness in achieving significant performance improvements across various RL benchmarks.

For future work, a promising direction is to explore a learning-based adaptive AL operator, incorporating approaches such as meta-learning or adversarial learning, rather than relying on heuristic methods that require handcrafted adaptive mechanisms and cumbersome fine-tunes.

Appendix A
Proof of Lemma 1

Proof: First, we derive the difference between Bellman operator with respect to the optimal policy π^*, i.e., $\mathcal{T}_{\pi^*}$, and the AL operator $\mathcal{T}_{\mathrm{AL}}$ with the same state value function

$$\begin{aligned}
&\mathcal{T}_{\pi^*}V_k(s) - \mathcal{T}_{\mathrm{AL}}V_k(s)\\
&= r(s,\pi^*(s)) + \gamma\mathbb{E}_{s'\sim P(s'|s,\pi^*(s))}\left[V_k(s')\right]\\
&\quad - \max_a\left[\mathcal{T}Q_k(s,a) - \alpha\left(V_k(s) - Q_k(s,a)\right)\right]\\
&\le r(s,\pi^*(s)) + \gamma\mathbb{E}_{s'\sim P(s'|s,\pi^*(s))}\left[V_k(s')\right]\\
&\quad - \mathcal{T}Q_k(s,\pi^*(s)) + \alpha\left(V_k(s) - Q_k(s,\pi^*(s))\right)\\
&= r(s,\pi^*(s)) + \gamma\mathbb{E}_{s'\sim P(s'|s,\pi^*(s))}\left[V_k(s')\right]\\
&\quad - \left[r(s,\pi^*(s)) + \gamma\mathbb{E}_{s'\sim P(s'|s,\pi^*(s))}\left[V_k(s')\right]\right]\\
&\quad + \alpha\left(V_k(s) - Q_k(s,\pi^*(s))\right)\\
&= \alpha\left(V_k(s) - Q_k(s,\pi^*(s))\right). \qquad (16)
\end{aligned}$$

We define $\Delta_k^{\pi^*}(s) = V_k(s) - Q_k(s,\pi^*(s))$, and let $\alpha\Delta_k^{\pi^*} = \mathcal{T}_{\pi^*}V_k - \mathcal{T}_{\mathrm{AL}}V_k$ represent the corresponding vector of length $|\mathcal{S}|$ where each entry is $\alpha\Delta_k^{\pi^*}(s)$. Then, we can derive the error bound for $V^* - V_{k+1}$

$$\begin{aligned}
&V^* - V_{k+1}\\
&\quad = \mathcal{T}_{\pi^*}V^* - \mathcal{T}_{\pi^*}V_k + \mathcal{T}_{\pi^*}V_k - \mathcal{T}_{\mathrm{AL}}V_k\\
&\quad \le \gamma P_{\pi^*}(V^* - V_k) + \alpha\Delta_k^{\pi^*} \qquad (17)
\end{aligned}$$

where P_{π^*} is the probability transition matrix of size $|\mathcal{S}|\times|\mathcal{S}|$, i.e., $P_{\pi^*}(s,s') = P(s'|s,\pi^*(s))$. By recursively applying (17) for K times, we have

$$\begin{aligned}
&V^* - V_K\\
&\quad \le \gamma P_{\pi^*}(V^* - V_{K-1}) + \alpha\Delta_{K-1}^{\pi^*}\\
&\quad \le (\gamma P_{\pi^*})^2(V^* - V_{K-2}) + \gamma P_{\pi^*}\left(\alpha\Delta_{K-2}^{\pi^*}\right) + \alpha\Delta_{K-1}^{\pi^*}\\
&\quad \cdots\\
&\quad \le (\gamma P_{\pi^*})^K(V^* - V_0) + \alpha\sum_{k=0}^{k=K-1}(\gamma P_{\pi^*})^{K-k-1}\Delta_k^{\pi^*}. \qquad (18)
\end{aligned}$$

Let us denote the greedy policy with respect to V_K, i.e., π_{K+1} by $\pi_{K+1}(s) \triangleq \max_a[r(s,a) + \gamma\mathbb{E}_{s'\sim P(s'|s,a)}[V_K(s')]]$, and $V^{\pi_{K+1}}$ denote the fixed point of $\mathcal{T}_{\pi_{K+1}}$, i.e., $V^{\pi_{K+1}} = \mathcal{T}_{\pi_{K+1}}V^{\pi_{K+1}}$. Then, we derive $V^* - V^{\pi_{K+1}}$ by following the similar process provided in ([14]):

$$\begin{aligned}
&V^* - V^{\pi_{K+1}}\\
&\quad = \mathcal{T}_{\pi^*}V^* - \mathcal{T}_{\pi^*}V_K + \mathcal{T}_{\pi^*}V_K - \mathcal{T}V_K + \mathcal{T}V_K - \mathcal{T}_{\pi_{K+1}}V^{\pi_{K+1}}\\
&\quad \le \gamma P_{\pi^*}(V^* - V_K) + \mathcal{T}_{\pi_{K+1}}V_K - \mathcal{T}_{\pi_{K+1}}V^{\pi_{K+1}}\\
&\quad = \gamma P_{\pi^*}(V^* - V_K) + \gamma P_{\pi_{K+1}}(V_K - V^{\pi_{K+1}}) \qquad (19)
\end{aligned}$$

where (19) is due to the fact that $\mathcal{T}V_K \geq \mathcal{T}_{\pi^*}V_K$ and utilizes the definition of induced greedy policy

$$\begin{aligned}
&\mathcal{T}V_K(s) \\
&= \max_{a\in\mathcal{A}}\left[r(s,a)+\gamma\mathbb{E}_{s'\sim P(s'|s,a)}\left[V_K\left(s'\right)\right]\right] \\
&= r\left(s,\pi_{K+1}(s)\right)+\gamma\mathbb{E}_{s'\sim P\left(s'|s,\pi_{K+1}(s)\right)}\left[V_K\left(s'\right)\right] \\
&= \mathcal{T}_{\pi_{K+1}}V_K(s). \qquad (20)
\end{aligned}$$

Furthermore, we can rewrite (19) as

$$V^*-V^{\pi_{K+1}} \leq \gamma\left(\mathbf{I}-\gamma P_{\pi_{K+1}}\right)^{-1}\left(P_{\pi^*}-P_{\pi_{K+1}}\right)\left(V^*-V_K\right). \quad (21)$$

Then, plugging (18) into (21) and taking $\|\cdot\|_\infty$ on both sides

$$\begin{aligned}
&\|V^*-V^{\pi_{K+1}}\|_\infty \\
&\leq \gamma\left\|\left(\mathbf{I}-\gamma P_{\pi_{K+1}}\right)^{-1}\right\|_\infty\cdot\left\|P_{\pi^*}-P_{\pi_{K+1}}\right\|_\infty\cdot\|V^*-V_K\|_\infty \\
&\leq \frac{2\gamma}{1-\gamma}\left[\left\|(\gamma P_{\pi^*})^K\right\|_\infty\cdot\|V^*-V_0\|_\infty\right. \\
&\qquad\left.+\alpha\sum_{k=0}^{k=K-1}\left\|(\gamma P_{\pi^*})^{K-k-1}\right\|_\infty\cdot\left\|\Delta_k^{\pi^*}\right\|_\infty\right] \\
&= \frac{2\gamma}{1-\gamma}\left[2\gamma^K V_{\max}+\alpha\sum_{k=0}^{k=K-1}\gamma^{K-k-1}\left\|\Delta_k^{\pi^*}\right\|_\infty\right]. \qquad (22)
\end{aligned}$$

■

Appendix B
Proof of Lemma 2

For convenience, we first prove the following lemma.

Lemma 3: $\forall Q_1, Q_2$ that satisfy $V^*(s) = \max_a Q_1(s,a) = \max_a Q_2(s,a)$ for all $s \in \mathcal{S}$, we have the following.

1) If $Q_1 \geq Q_2$, then $\mathcal{T}_{\mathrm{AL}}Q_1 \geq \mathcal{T}_{\mathrm{AL}}Q_2$.
2) If $V^*(s) = Q_1(s,\pi^*(s))$, then $\forall n \in \mathbb{N}^+, V^*(s) = \max_a(\mathcal{T}_{\mathrm{AL}})^n Q_1(s,a)$, and $\mathcal{T}(\mathcal{T}_{\mathrm{AL}})^n Q_1 = Q^*$.

Proof: According to the definition of $\mathcal{T}_{\mathrm{AL}}$, we have

$$\begin{aligned}
&\mathcal{T}_{\mathrm{AL}}Q_1(s,a)-\mathcal{T}_{\mathrm{AL}}Q_2(s,a) \\
&= \mathcal{T}Q_1(s,a)+\alpha\left(Q_1(s,a)-V^*(s)\right) \\
&\quad -\mathcal{T}Q_2(s,a)-\alpha\left(Q_2(s,a)-V^*(s)\right) \\
&= \mathbb{E}_{s'}\left[\max_{a'}Q_1\left(s',a'\right)-\max_{a'}Q_2\left(s',a'\right)\right] \\
&\quad +\alpha\left(Q_1(s,a)-Q_2(s,a)\right) \\
&= \alpha\left(Q_1(s,a)-Q_2(s,a)\right).
\end{aligned}$$

Thus, if $Q_1 \geq Q_2$, we have $\mathcal{T}_{\mathrm{AL}}Q_{k_1} \geq \mathcal{T}_{\mathrm{AL}}Q_{k_2}$, and the first property is verified. We also have that

$$\begin{aligned}
&\max_a \mathcal{T}_{\mathrm{AL}}Q_1(s,a) \\
&= \max_a\left[r(s,a)+\mathbb{E}\left[V^*\left(s'\right)\right]+\alpha\left(Q_1(s,a)-V^*(s)\right)\right] \\
&= \max_a\left[Q^*(s,a)+\alpha\left(Q_1(s,a)-V^*(s)\right)\right] \qquad (23)
\end{aligned}$$

based on the conditions

$$\max_a Q_1(s,a) = Q_1(s,\pi^*(s)) = V^*(s). \qquad (24)$$

Note that $\max_a Q^*(s,a)$ and $\max_a Q_1(s,a)$ have the same maximizer $\pi^*(s)$ and maximum $V^*(s)$, so (23) satisfies

$$\max_a \mathcal{T}_{\mathrm{AL}}Q_1(s,a) = \mathcal{T}_{\mathrm{AL}}Q_1(s,\pi^*(s)) = V^*(s).$$

Thus, $\mathcal{T}_{\mathrm{AL}}Q_1(s,a)$ also satisfies the condition in (24). By applying (23) and (24) repeatedly, we can obtain the final result

$$V^*(s) = \max_a\left(\mathcal{T}_{\mathrm{AL}}\right)^n Q_1(s,a). \qquad (25)$$

Benefiting from the above result, we can further obtain

$$\begin{aligned}
\mathcal{T}\left(\mathcal{T}_{\mathrm{AL}}\right)^n Q_1(s,a) &= r(s,a)+\gamma\mathbb{E}_{s'}\left[\max_{a'}\left(\mathcal{T}_{\mathrm{AL}}\right)^n Q_1\left(s',a'\right)\right] \\
&= r(s,a)+\gamma\mathbb{E}_{s'}\left[V^*(s)\right] \\
&= Q^*(s,a).
\end{aligned}$$

■

Next, we will prove Lemma 2.

Proof: Define $\{\tilde{Q}_k\}_{k=0}^\infty$ and $\{\hat{Q}_k\}_{k=0}^\infty$ as the Q-value function sequences obtained by $\mathcal{T}_{\mathrm{ORAL}}$ and $\mathcal{T}_{\mathrm{AL}}$, respectively. The corresponding V-value function sequences are defined as $\{\tilde{V}_k\}_{k=0}^\infty$ and $\{\hat{V}_k\}_{k=0}^\infty$ by following $V(s) = \max_a Q(s,a)$. According to Corollary 1, we know that $\mathcal{T}_{\mathrm{ORAL}}$ and $\mathcal{T}_{\mathrm{AL}}$ are optimality preserving, and thus, both the V-value function sequences will converge to the optimal value function, i.e., $\lim_{k\to\infty}\tilde{V}_k(s) = \lim_{k\to\infty}\hat{V}_k(s) = V^*(s)$, and there is at least one optimal action remains optimal, and all suboptimal actions remain suboptimal, so the induced greedy policy with respect to $\tilde{Q}_\infty(s,a)$ must belong to the set of optimal policies Π^*, and we denote it by π^*. Thus, π^* satisfies

$$V^*(s) = \max_a \tilde{Q}_\infty(s,a) = \tilde{Q}_\infty(s,\pi^*(s)). \qquad (26)$$

Now, we use mathematical induction to prove the following relationship:

$$\left(\mathcal{T}_{\mathrm{ORAL}}\right)^n\tilde{Q}_\infty(s,a) \geq \left(\mathcal{T}_{\mathrm{AL}}\right)^n\tilde{Q}_\infty(s,a). \qquad (27)$$

Specifically, the following conditions hold.

1) When $n = 1$, according to the definitions of $\mathcal{T}_{\mathrm{ORAL}}$ and $\mathcal{T}_{\mathrm{AL}}$, we have
$$\mathcal{T}_{\mathrm{ORAL}}\tilde{Q}_\infty(s,a) \geq \mathcal{T}_{\mathrm{AL}}\tilde{Q}_\infty(s,a). \qquad (28)$$
2) When $n = k$, if $(\mathcal{T}_{\mathrm{ORAL}})^k\tilde{Q}_\infty(s,a) \geq (\mathcal{T}_{\mathrm{AL}})^k\tilde{Q}_\infty(s,a)$ exists, according to (26) and Lemma 3, we know that
$$V^*(s) = \max_a\left(\mathcal{T}_{\mathrm{AL}}\right)^k\tilde{Q}_\infty(s,a). \qquad (29)$$

Furthermore, we have

$$\max_a\left(\mathcal{T}_{\mathrm{ORAL}}\right)^k\tilde{Q}_\infty(s,a) = \max_a\tilde{Q}_{\infty+k}(s,a) = V^*(s). \qquad (30)$$

Based on (29) and (30) and Lemma 3, we can get

$$\mathcal{T}_{\mathrm{AL}}\left(\mathcal{T}_{\mathrm{ORAL}}\right)^k\tilde{Q}_\infty(s,a) \geq \mathcal{T}_{\mathrm{AL}}\left(\mathcal{T}_{\mathrm{AL}}\right)^k\tilde{Q}_\infty(s,a). \qquad (31)$$

Recalling the definitions of $\mathcal{T}_{\mathrm{ORAL}}$ and $\mathcal{T}_{\mathrm{AL}}$ again, we have $\mathcal{T}_{\mathrm{ORAL}}(\mathcal{T}_{\mathrm{ORAL}})^k\tilde{Q}_\infty(s,a) \geq \mathcal{T}_{\mathrm{AL}}(\mathcal{T}_{\mathrm{ORAL}})^k\tilde{Q}_\infty(s,a)$, and together with (31), we derive the conclusion

$$\begin{aligned}
\left(\mathcal{T}_{\mathrm{ORAL}}\right)^{k+1}\tilde{Q}_\infty(s,a) &\geq \mathcal{T}_{\mathrm{AL}}\left(\mathcal{T}_{\mathrm{ORAL}}\right)^k\tilde{Q}_\infty(s,a) \\
&\geq \left(\mathcal{T}_{\mathrm{AL}}\right)^{k+1}\tilde{Q}_\infty(s,a). \qquad (32)
\end{aligned}$$

According to the relationship in (27) induced by mathematical induction, we have the following relation:

$$\begin{aligned}
&\limsup_{n\to\infty}\left(\mathcal{T}_{\mathrm{clipAL}}\right)^n\tilde{Q}_\infty(s,a) \\
&\geq \limsup_{n\to\infty}\left(\mathcal{T}_{\mathrm{AL}}\right)^n\tilde{Q}_\infty(s,a) \\
&\Rightarrow \limsup_{n\to\infty}\tilde{Q}_n(s,a) \geq \limsup_{n\to\infty}\hat{Q}_n(s,a)
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow -\liminf_{n\to\infty} \tilde{Q}_n(s,a) \le -\liminf_{n\to\infty} \hat{Q}_n(s,a)\\
&\Rightarrow V^*(s) - \liminf_{n\to\infty} \tilde{Q}_n(s,a) \le V^*(s) - \liminf_{n\to\infty} \hat{Q}_n(s,a)\\
&\Rightarrow \liminf_{n\to\infty} \left[\tilde{V}_n(s) - \tilde{Q}_n(s,a)\right] \le \liminf_{n\to\infty} \left[\hat{V}_n(s) - \hat{Q}_n(s,a)\right]\\
&\Rightarrow G_{\mathrm{ORAL}}(s,a) \le G_{\mathrm{AL}}(s,a). \qquad (33)
\end{aligned}$$

Thanks to the optimality preserving, we get the results in Lemma 2: $G^*(s,a) \le G_{\mathrm{ORAL}}(s,a) \le G_{\mathrm{AL}}(s,a)$.■

Appendix C
Proof of Theorem 2

Proof: Due to the nature of a piecewise operator, if we want $\mathcal{T}_{\mathrm{ORAL-c}}$ to be a convergent operator, it must keep consistent with either $\mathcal{T}$ or $\mathcal{T}_{\mathrm{AL}}$ when the number of iterations approaches to infinity. A reasonable assumption is that starting from $\tilde{Q}_\infty$ defined in (26), $\mathcal{T}_{\mathrm{ORAL-c}}$ no longer switches back and forth between $\mathcal{T}$ or $\mathcal{T}_{\mathrm{AL}}$.

If $\forall s \in \mathcal{S}, a \in \mathcal{A}, n \in \mathbb{N}^+$, $(\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) = (\mathcal{T})^n \tilde{Q}_\infty(s,a)$, we need

$$\begin{aligned}
r_n(s,a) &\triangleq \frac{(\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) - Q_l}{\max_a (\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) - Q_l}\\
&= \frac{(\mathcal{T})^n \tilde{Q}_\infty(s,a) - Q_l}{V^*(s) - Q_l} < \beta. \qquad (34)
\end{aligned}$$

Note that $\forall n \in \mathbb{N}^+$, we have $(\mathcal{T})^n \tilde{Q}_\infty(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')] = Q^*(s,a)$, so $\mathcal{T}_{\mathrm{ORAL-c}}$ has the same convergent action gaps as that of $\mathcal{T}$, i.e., $\lim_{k\to\infty} G_k(s,a) = V^*(s) - Q^*(s,a)$, the ratio threshold β should satisfy

$$\max_{s,a} \frac{Q^*(s,a) - Q_l}{V^*(s) - Q_l} < \beta < 1.$$

On the other hand, if $\forall s \in \mathcal{S}, a \in \mathcal{A}, n \in \mathbb{N}^+$, $(\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) = (\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty(s,a)$, we need

$$\begin{aligned}
r_n(s,a) &\triangleq \frac{(\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) - Q_l}{\max_a (\mathcal{T}_{\mathrm{ORAL-c}})^n \tilde{Q}_\infty(s,a) - Q_l}\\
&= \frac{(\mathcal{T})^n \tilde{Q}_\infty(s,a) - Q_l}{V^*(s) - Q_l} \ge \beta. \qquad (35)
\end{aligned}$$

Let us see the sequence $\{(\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty\}_{n=1}^\infty$, when $n = 1$

$$\begin{aligned}
\mathcal{T}_{\mathrm{AL}} \tilde{Q}_\infty &= \mathcal{T} \tilde{Q}_\infty + \alpha \left(\tilde{Q}_\infty - \max_a \tilde{Q}_\infty\right)\\
&= Q^* + \alpha \left(\tilde{Q}_\infty - V^*\right)
\end{aligned}$$

and when $n = 2$, combining with the second property in Lemma 3, we can obtain

$$\begin{aligned}
&(\mathcal{T}_{\mathrm{AL}})^2 \tilde{Q}_\infty\\
&= \mathcal{T}\left(\mathcal{T}_{\mathrm{AL}} \tilde{Q}_\infty\right) + \alpha \left(Q^* + \alpha\left(\tilde{Q}_\infty - V^*\right) - \max_a \mathcal{T}_{\mathrm{AL}} \tilde{Q}_\infty\right)\\
&= Q^* + \alpha(Q^* - V^*) + \alpha^2 \left(\tilde{Q}_\infty - V^*\right)
\end{aligned}$$

then $\forall n \in \mathbb{N}^+$, the recursive result can be rewritten as

$$(\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty = V^* + \sum_{t=0}^{t=n-1} \alpha^t (Q^* - V^*) + \alpha^n \left(\tilde{Q}_\infty - V^*\right).$$

Next, we have that

$$\begin{aligned}
&(\mathcal{T}_{\mathrm{AL}})^{n+1} \tilde{Q}_\infty - (\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty\\
&= \alpha^n (Q^* - V^*) + \alpha^{n+1} \left(\tilde{Q}_\infty - V^*\right) - \alpha^n \left(\tilde{Q}_\infty - V^*\right)\\
&= \alpha^n \left(Q^* - \alpha V^* - (1-\alpha) \tilde{Q}_\infty\right)\\
&\overset{(a)}{=} \alpha^n (1-\alpha) \left(\lim_{k\to\infty} \hat{Q}_k - \tilde{Q}_\infty\right) \overset{(b)}{\le} 0 \qquad (36)
\end{aligned}$$

where the equivalence relation (a) in the above formula is built on Property 3, i.e., $\lim_{k\to\infty} \hat{Q}_k = (Q^* - \alpha V^*)/(1-\alpha)$, and the inequality relation (b) is based on (33).

Thus, we know that $\{(\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty\}_{n=1}^\infty$ is a decreasing sequence that converges to $(Q^* - \alpha V^*)/(1-\alpha)$. Then, if $\mathcal{T}_{\mathrm{ORAL-c}}$ has the same convergent action gaps as that of $\mathcal{T}_{\mathrm{AL}}$, ie., $\lim_{k\to\infty} G_k(s,a) = (V^*(s) - Q^*(s,a))/(1-\alpha)$, β should satisfy

$$\begin{aligned}
c \le \min_{s,a,n} r_n(s,a) &= \min_{s,a,n} \frac{(\mathcal{T}_{\mathrm{AL}})^n \tilde{Q}_\infty(s,a) - Q_l}{V^*(s) - Q_l}\\
&= \min_{s,a} \frac{Q^* - \alpha V^* - (1-\alpha) Q_l}{(1-\alpha)(V^* - Q_l)}. \qquad (37)
\end{aligned}$$

■

Appendix D
Proof of Theorem 3

Proof: Recalling the definition $\mathcal{T}'Q(s,a) = \mathcal{T}Q(s,a) + \alpha(Q(s,a) - V(s)) \cdot f(s,a)$, where $f(\cdot) : \mathbb{R} \to [0,1]$, we know that $\mathcal{T}'$ satisfies the conditions described in Theorem 2, i.e., $\mathcal{T}'$ is optimality preserving. Thus, we can assume that

$$\max_{a\in\mathcal{A}} Q_K(s,a) = Q_K(s,a^*) = Q^*(s,a^*) = V^*(s) \qquad (38)$$

after the Kth iteration. Then, we have the following results:

$$\begin{aligned}
Q_{K+1} &= \mathcal{T}' Q_K(s,a)\\
&= Q^*(s,a) + \alpha (Q_K(s,a) - V^*(s)) \cdot f_K(s,a). \qquad (39)
\end{aligned}$$

The result in (39) is based on the fact that

$$\begin{aligned}
\mathcal{T}Q_k(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'} \left[\max_{a'\in\mathcal{A}} Q_K(s',a')\right]\\
&= r(s,a) + \gamma \mathbb{E}_{s'} \left[V^*(s')\right] = Q^*(s,a).
\end{aligned}$$

Note that the first and second terms in (39) have the same maximizer a^*, and thus,

$$\begin{aligned}
&\mathcal{T}Q_{K+1}(s,a)\\
&= r(s,a) + \gamma \mathbb{E}_{s'} \left[\max_{a'\in\mathcal{A}} Q_{K+1}(s',a')\right]\\
&= r(s,a) + \gamma \mathbb{E}_{s'} \Bigg[Q^*(s',a^*)\\
&\qquad + \underbrace{\alpha \left(Q_K(s',a^*) - V^*(s')\right) \cdot f_K(s',a^*)}_{=0 \text{ as } (38)}\Bigg]\\
&= r(s,a) + \gamma \mathbb{E}_{s'} \left[V^*(s')\right] = Q^*(s,a).
\end{aligned}$$

Thus, we can get

$$\begin{aligned}
&Q_{K+2}(s,a)\\
&= \mathcal{T}' Q_{K+1}(s,a)\\
&= \mathcal{T}Q_{K+1}(s,a) + \alpha \left(Q_{K+1}(s,a) - V^*(s)\right) \cdot f_{K+1}(s,a)\\
&= Q^*(s,a) + \alpha (Q^*(s,a) - V^*(s)) f_{K+1}(s,a)\\
&\quad + \alpha^2 (Q_K(s,a) - V^*(s)) \cdot \left(f_{K+1} \cdot f_K\right)(s,a). \qquad (40)
\end{aligned}$$

All the three terms in (40) share the same maximizer a^*, and then, $\mathcal{T}Q_{K+2}(s,a) = Q^*(s,a)$. Thus, we can repeat the above process and get the result after $K+n$ iteration $(n \ge 2)$

$$Q_{K+n}(s,a)$$

$$
\begin{aligned}
&= Q^*(s,a) \\
&\quad + \left(Q^*(s,a) - V^*(s)\right) \sum_{i=1}^{i=n-1} \alpha^i \cdot \left(f_{K+n-1} \cdots f_{K+n-i}\right)(s,a) \\
&\quad + \left(Q_K(s,a) - V^{(s)}\right) \cdot \alpha^n \left(f_{K+n-1} \cdots f_K\right)(s,a). \qquad (41)
\end{aligned}
$$

Subtracting $V^*(s)$ from both sides of (41), we have

$$
\begin{aligned}
&-G_{K+n}(s,a) \\
&= Q_{K+n}(s,a) - V^*(s) \\
&= Q^*(s,a) - V^*(s) \\
&\quad + \left(Q^*(s,a) - V^*(s)\right) \\
&\quad \times \sum_{i=1}^{i=n-1} \alpha^i \cdot \left(f_{K+n-1} \cdots f_{K+n-i}\right)(s,a) \\
&\quad + \left(Q_K(s,a) - V^{(s)}\right) \cdot \alpha^n \cdot \left(f_{K+n-1} \cdots f_K\right)(s,a) \\
&= Q^*(s,a) - V^*(s) + \alpha \left(Q_{K+n-1}(s,a) - V^*(s)\right) \\
&\quad \cdot f_{K+n-1}(s,a) \\
&= Q^*(s,a) - V^*(s) - \alpha f_{K+n-1}(s,a) \cdot G_{K+n-1}(s,a). \qquad (42)
\end{aligned}
$$

According to (42), we can define a new operator for the iterated action gap $G_k(s,a)$

$$
\hat{\mathcal{T}} G_k(s,a) = \alpha f_k(s,a) \cdot G_k(s,a) + V^*(s) - Q^*(s,a). \qquad (43)
$$

For any two iterated action gap vector $G, G' \in \mathbb{R}^{|\mathcal{S}\|\mathcal{A}|}$

$$
\begin{aligned}
\left\|\hat{\mathcal{T}} G - \hat{\mathcal{T}} G'\right\|_\infty &= \alpha \left\| f \cdot G - f' \cdot G' \right\|_\infty \\
&\le \alpha \left\| \frac{f \cdot G - f' \cdot G'}{G - G'} \right\|_\infty \cdot \left\| G - G' \right\|_\infty. \qquad (44)
\end{aligned}
$$

When $\hat{\mathcal{T}}$ is contractive, it should satisfy

$$
\left\|\hat{\mathcal{T}} G - \hat{\mathcal{T}} G'\right\|_\infty = \alpha \left\| f \cdot G - f' \cdot G' \right\|_\infty \le \left\| G - G' \right\|_\infty.
$$

Combining with (44), we can get this sufficient condition

$$
\begin{aligned}
&\left\|\hat{\mathcal{T}} G - \hat{\mathcal{T}} G'\right\|_\infty \\
&\quad = \alpha \left\| f \cdot G - f' \cdot G' \right\|_\infty \le \left\| G - G' \right\|_\infty \\
&\quad \Rightarrow \alpha \left\| \frac{f \cdot G - f' \cdot G'}{G - G'} \right\|_\infty \cdot \left\| G - G' \right\|_\infty \le \left\| G - G' \right\|_\infty \\
&\quad \Rightarrow \left\| \frac{f \cdot G - f' \cdot G'}{G - G'} \right\|_\infty = \left\| \frac{d(f \cdot G)}{dG} \right\|_\infty \le \frac{1}{\alpha}.
\end{aligned}
$$

■

REFERENCES

[1] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] J. Schrittwieser et al., "Mastering Atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, Dec. 2020.

[3] H. Zeng, P. Zhang, F. Li, C. Lin, and J. Zhou, "AHEGC: Adaptive hindsight experience replay with goal-amended curiosity module for robot control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16602–16615, Nov. 2024.

[4] R. Chai, H. Niu, J. Carrasco, F. Arvin, H. Yin, and B. Lennox, "Design and experimental validation of deep reinforcement learning-based fast trajectory planning and control for mobile robot in unknown environment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5778–5792, Apr. 2024.

[5] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–36, Jan. 2023.

[6] J. Yang et al., "Deep reinforcement learning for multi-class imbalanced training: Applications in healthcare," *Mach. Learn.*, vol. 113, no. 5, pp. 2655–2674, May 2024.

[7] R. Pina, V. D. Silva, J. Hook, and A. Kondoz, "Residual Q-networks for value function factorizing in multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1534–1544, Feb. 2024.

[8] X. Yao, C. Wen, Y. Wang, and X. Tan, "SMIX(λ): Enhancing centralized value functions for cooperative multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 52–63, Jan. 2023.

[9] Z. Zhou, G. Liu, and M. Zhou, "A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14370–14381, Oct. 2024.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[12] M. G. Bellemare, G. Ostrovski, A. Guez, P. S. Thomas, and R. Munos, "Increasing the action gap: New operators for reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1476–1483.

[13] T. Kozuno, E. Uchibe, and K. Doya, "Unifying value iteration, advantage learning, and dynamic policy programming," 2017, *arXiv:1710.10866*.

[14] A. Farahmand, "Action-gap phenomenon in reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 172–180.

[15] M. G. Azar, V. Gómez, and H. J. Kappen, "Dynamic policy programming," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 3207–3245, 2012.

[16] N. Vieillard, O. Pietquin, and M. Geist, "Munchausen reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 4235–4246.

[17] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist, "Leverage the average: An analysis of KL regularization in reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12163–12174.

[18] T. Kozuno, D. Han, and K. Doya, "Gap-increasing policy evaluation for efficient and noise-tolerant reinforcement learning," 2019, *arXiv:1906.07586*.

[19] Y. Gan, Z. Zhang, and X. Tan, "Smoothing advantage learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 6, pp. 6657–6664.

[20] J. Ferret, O. Pietquin, and M. Geist, "Self-imitation advantage learning," in *Proc. 20th Int. Conf. Auto. Agents Multiagent Syst.*, 2020, pp. 501–509.

[21] Z. Zhang, Y. Gan, and X. Tan, "Robust action gap increasing with clipped advantage learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 9145–9152.

[22] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *J. Artif. Intell. Res.*, vol. 47, pp. 253–279, Jun. 2013.

[23] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5026–5033.

[24] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos, "Speedy Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2411–2419.

[25] D. P. Bertsekas and H. Yu, "Q-learning and enhanced policy iteration in discounted dynamic programming," *Math. Oper. Res.*, vol. 37, no. 1, pp. 66–94, Feb. 2012.

[26] L. C. Baird, "Reinforcement learning through gradient descent," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1999.

[27] H. V. Seijen, M. Fatemi, and A. Tavakoli, "Using a logarithmic mapping to enable lower discount factors in reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14134–14144.

[28] E. Smirnova and E. Dohmatob, "On the convergence of smooth regularized approximate value iteration schemes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6540–6550.

[29] J. Oh, Y. Guo, S. Singh, and H. Lee, "Self-imitation learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3878–3887.

[30] R. Munos, T. Stepleton, A. Harutyunyan, and M. G. Bellemare, "Safe and efficient off-policy reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1054–1062.

[31] H. Wiltzer, M. Bellemare, D. Meger, P. Shafto, and Y. Jhaveri, "Action gaps and advantages in continuous-time distributional reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 47815–47848.

[32] H.-Y. Liu, B. Balaji, R. Gupta, and D. Hong, "Offline reinforcement learning with munchausen regularization," in *Proc. Offline Reinforcement Learn. Workshop Neural Inf. Process. Syst.*, 2021, pp. 1–6.

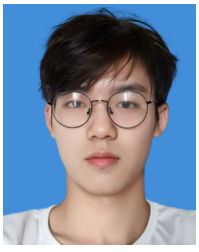
[33] A. Brunnbauer, J. Lemmel, Z. Babaiee, S. Neubauer, and R. Grosu, "Scalable offline reinforcement learning for mean field games," 2024, *arXiv:2410.17898*.
[34] T. Kozuno, E. Uchibe, and K. Doya, "Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2019, pp. 2995–3003.
[35] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," Dept. CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep. 32, 2019.
[36] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep Q-learning with model-based acceleration," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2829–2838.
[37] K. Young and T. Tian, "MinAtar: An Atari-inspired testbed for thorough and reproducible reinforcement learning experiments," 2019, *arXiv:1903.03176*.
[38] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 30th Conf. Artif. Intell. (AAAI)*, Mar. 2016, pp. 2094–2100.
[39] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
[40] Q. Lan, Y. Pan, A. Fyshe, and M. White, "Maxmin Q-learning: Controlling the estimation bias of Q-learning," in *Proc. 8th Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, Apr. 2020.
[41] Q. Lan, Y. Pan, J. Luo, and A. R. Mahmood, "Memory-efficient reinforcement learning with value-based knowledge consolidation," *Trans. Mach. Learn. Res.*, vol. 2022, pp. 1–23, May 2022.
[42] P. R. Van der Vaart, M. T. J. Spaan, and N. Yorke-Smith, "Epistemic Bellman operators," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 20, pp. 20973–20981.
[43] P. Li, H. Jianye, H. Tang, Y. Zheng, and F. Barez, "Value-evolutionary-based reinforcement learning," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 27875–27889.
[44] R. Chen and I. C. Paschalidis, "Distributionally robust learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 4, 2020, pp. 1–243.
[45] K. P. Badrinath and D. Kalathil, "Robust reinforcement learning using least squares policy iteration with provable performance guarantees," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 511–520.

Zhe Zhang received the B.S. degree in transportation engineering, the M.S. degree in vehicle operation engineering, and the Ph.D. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016, 2019, and 2023, respectively.

He is currently a Lecturer at the College of Information Science and Technology, Jinan University, Guangzhou, China. His research focuses on deep reinforcement learning and computer vision.

Yongle Zhou received the B.S. degree in software engineering from East China Jiaotong University, Nanchang, China, in 2024. He is currently pursuing the M.S. degree with the College of Information Science and Technology, Jinan University, Guangzhou, China.

His research interests include reinforcement learning.



Yuyang Long is currently pursuing the B.S. degree in software engineering with Jinan University, Guangzhou, China.

His research interests include reinforcement learning.

Jia Zhang (Member, IEEE) received the Ph.D. degree from the Department of Artificial Intelligence, Xiamen University, Xiamen, China, in 2020.

He is currently a Lecturer at the College of Information Science and Technology, Jinan University, Guangzhou, China. His research interests include machine learning, data mining, and human–computer interaction.

Juanjuan Weng received the B.Sc. degree in computer science and technology from Henan University of Science and Technology, Luoyang, China, in 2019, and the Ph.D. degree from the Department of Artificial Intelligence, Xiamen University, Xiamen, China, in 2024.

She is an Assistant Professor at Jinan University, Guangzhou, China. Her research interests include computer vision, adversarial attacks, and adversarial defense.

Zhetao Li (Member, IEEE) received the B.Eng. degree in electrical information engineering from Xiangtan University, Xiangtan, China, in 2002, the M.Eng. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2005, and the Ph.D. degree in computer application technology from Hunan University, Changsha, China, in 2010.

He is a Professor with the College of Information Science and Technology, Jinan University, Guangzhou, China. His research interests include AI-Modelnet, cloud computing, and mobile edge computing.

Dr. Li is a member of CCF.

Yaozhong Gan received the Ph.D. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2023.

He is currently an Assistant Researcher at Qiyuan Lab, Beijing, China. He has published several papers in conferences, such as ICML, AAAI, ICLR, and ICCV. His research focuses on deep reinforcement learning and large model applications.

Xiaoyang Tan was a Post-Doctoral Researcher with the LEAR Team, INRIAR Rhone-Alpes, Grenoble, France, from 2006 to 2007. He is currently a Professor at the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has authored or co-authored over 50 conference papers and journal articles. His research interests include deep learning, reinforcement learning, and Bayesian learning.

Prof. Tan and his colleagues were awarded the IEEE Signal Processing Society Best Paper in 2015.