Logarithmic Function Matters Policy Gradient Deep Reinforcement Learning

Qi Liu[1], Jingxiang Guo[1], Zhongjian Qiao[2], Pengbin Chen[1], Yanjie Li[1*]

***Abstract*—This paper studies the influence of logarithmic functions in policy gradient deep reinforcement learning (RL) methods, a subject previously under-emphasized in the field. It delves into how different logarithmic bases, specifically *loge*, *log2*, and *log10*, influence the efficiency of policy gradient methods. We analyze the role of these logarithmic functions and their gradients when using Stochastic Gradient Descent as the optimizer in deep RL. Experimental results show that the choice of the logarithmic base influences the policy gradient methods. *loge* and *log2* are generally more effective than *log10* in deep RL algorithms, but *log10* exhibits more stability. Based on analysis and experimental results, we propose the Logarithmic Basis Policy Gradient (LBPG) and Adaptive LBPG algorithms to enhance policy gradient methods. LBPG denotes policy gradient methods with *loge*, *log2*, and *log10* as the logarithmic base, respectively. Furthermore, Adaptive LBPG dynamically selects logarithmic basis among *loge*, *log2*, and *log10* based on the variance-to-mean ratio of returns, to optimize speed and stability. Experimental results show the importance of choosing appropriate logarithmic functions for policy gradient RL methods and highlight Adaptive LBPG's potential to refine learning in deep RL tasks.**

***Index Terms*—Deep reinforcement learning, policy gradient methods, logarithmic basis policy gradient**

I. Introduction

Deep reinforcement learning (RL) has been applied to various challenging domains, such as robotic control [1], [2], multi-agent systems [3], and video games [4]. Deep RL methods can be divided into two classes: value-based methods [4] and policy gradient methods [5]. This paper focuses on policy gradient methods.

A famous policy gradient method is the REINFORCE algorithm [5] that approximates the policy using a network. After that, deep RL methods that approximate both the policy and the value function by using neural networks are called the actor-critic architecture, where 'actor' denotes the policy, and 'critic' denotes the value function [6]. Modern variants include Natural Actor-Critic [7], Advantage Actor-Critic [8], Trust Region Policy Optimization (TRPO) [9], Proximal Policy Optimization (PPO) [10] algorithms, etc.

Literature [11] presents the n-step TD method to balance the return bias-variance trade-off. The n-step TD method can be generalized to the Monte Carlo (MC) and one-step TD methods. One can shift from one to the other smoothly as needed to meet a particular task's demands. Literature [8] proposes the advantage function to reduce the variance of the estimated value function. After that, general advantage estimation (GAE) [12] uses value functions to substantially reduce variance at the cost of some bias, with an exponentially weighted estimator of the advantage function analogous to TD(λ) [11]. Literature [13] proposes the control variate method to reduce the estimation variance of the return and does not change the expected return value. The literature [14] presents a detailed overview of off-policy [11] evaluation in deep RL.

There are two most widely used algorithms in deep RL: Advantage Actor-Critic (A2C) [8], and General Advantage Estimation (GAE) [12]. A2C is one of the most widely used methods in deep RL. After that, GAE [12] uses value functions to substantially reduce the variance of policy gradient estimates at the cost of some bias, with an exponentially weighted estimator of the advantage function analogous to TD(λ) [11].

As discussed above, although policy gradient methods have been studied extensively, most work focuses on the value function to address the bias-variance trade-off. Thus, a research gap remains in the exploration of the logarithmic function's role in policy gradient methods. Distinct from existing methods, we diverge from conventional policy gradient methods by focusing on the unexplored aspect of logarithmic functions within policy gradient methods. We critically examine how various logarithmic bases, particularly when used with the prevalent stochastic gradient descent (SGD) optimizer, impact policy gradient methods. Specifically, we select the logarithmic bases of *loge*, *log2*, and *log10* because of their prevalence in mathematical computations. This paper aims to understand how the choice of these bases and their corresponding gradients impact the performance and stability of policy gradient methods.

The main contributions of this work can be summarized as follows:

- In this paper, we analyze the influence of different logarithmic bases (*loge*, *log2*, and *log10*) on policy gradient methods in policy gradient deep RL methods. Experimental results in classical control benchmark environments illustrate how *loge* and *log2* generally lead to faster learning, whereas *log10* offers enhanced stability. These findings provide practical insights for deep RL researchers in selecting suitable logarithmic functions.
- We propose the adaptive Logarithmic Basis Policy Gradient (LBPG) algorithm that dynamically adapts among

This work was supported by the National Natural Science Foundation of China [61977019, U1813206] and Shenzhen Fundamental Research Program [JCYJ20180507183837726, JCYJ20220818102415033, JSGG20201103093802006]. *(Corresponding author: Yanjie Li, autolyj@hit.edu.cn)*

[1]Qi Liu, Jingxiang Guo, Pengbin Chen, and Yanjie Li are with the Guangdong Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics and School of Mechanical Engineering and Automation, the Harbin Institute of Technology Shenzhen, 518055.

[2]Zhongjian Qiao is with Tsinghua Shenzhen International Graduate School, Tsinghua University, China.

log2, *loge*, and *log10* in policy gradient based on the variance-to-mean ratio of returns. Experimental results show that Adaptive LBPG can balance learning efficiency and stability under different reward conditions, enabling tailored learning processes that are fine-tuned to environments' unique challenges and dynamics.

II. PRELIMINARIES

Policy gradient methods directly parameterize the policy and optimize the policy parameters by policy gradient [11]. The policy π_θ is usually modeled with a parameterized function for θ. Deep RL aims to learn a policy π_θ that can maximize the expected return $J(\pi_\theta)$:

$$\begin{aligned} J(\pi_\theta) &= \sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) V^{\pi_\theta}(s) \\ &= \sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} Q^{\pi_\theta}(s,a)\nabla_\theta \pi_\theta(a \mid s) \end{aligned} \tag{1}$$

The parameterized policy π_θ can be updated by the gradient [11]:

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &\propto \sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} Q^{\pi_\theta}(s,a)\nabla_\theta \pi_\theta(a \mid s) \\ &= \sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} Q^{\pi_\theta}(s,a)\pi_\theta(a \mid s)\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)} \\ &= \mathbb{E}^{\pi_\theta}\left[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a \mid s)\right] \end{aligned} \tag{2}$$

where $\nabla_\theta J(\pi_\theta)$ denotes the gradients of $J(\pi_\theta)$ with respect to θ. $d^{\pi_\theta}(s)$ represents the state distribution under π_θ. V^{π_θ} denotes the state-value function, Q^{π_θ} denotes the action-value function, and $\mathbb{E}^{\pi_\theta}$ denotes the expectation under policy π_θ. At timestep, t, an action a_t is taken under policy π_θ at state s_t, and thereafter following policy π_θ. The policy gradient theorem (Eq. (2)) lays the theoretical foundation for various policy gradient algorithms [5], [7]–[10]. Policy gradient methods maximize the expected total reward by repeatedly estimating the gradient.

Eq. (2) shows that the policy gradient involves the state-action value term $Q^{\pi_\theta}(s,a)$. Much work on policy gradient methods [8], [11], [12] focuses on the value function part ($Q^{\pi_\theta}(s,a)$) in policy gradient, intending to balance the bias-variance trade-off. There are several different related expressions for the policy gradient, which have the form:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right)\right] \tag{3}$$

where Ψ_t may be one of the following: (1) $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory; (2) $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t; (3) $\sum_{t'=t}^{\infty} r_{t'} - b\left(s_t\right)$: base-lined version of (2); (4) $Q^{\pi_\theta}\left(s_t, a_t\right)$: state-action value function; (5) $r_t + V^{\pi_\theta}\left(s_{t+1}\right) - V^{\pi_\theta}\left(s_t\right)$: temporal-difference (TD) residual; (6) $A^{\pi_\theta}\left(s_t, a_t\right) = Q^{\pi_\theta}\left(s_t, a_t\right) - V^{\pi_\theta}\left(s_t\right)$: advantage function. The latter formulas use the following definitions:

$$V^{\pi_\theta}\left(s_t\right) = \mathbb{E}^{\pi_\theta}_{s_{t:\infty}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \tag{4}$$

$$Q^{\pi_\theta}\left(s_t, a_t\right) = \mathbb{E}^{\pi_\theta}_{\substack{s_{t:\infty} \\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \tag{5}$$

where γ is a discount factor.

Literature [12] proposes γ-just estimator of the advantage function, showing in Eq. (6). The only source of bias in the estimator is the discount factor γ. Note that the discount factor γ can also be seen as a variance reduction technique that reduces the weight of future rewards but introduces some bias.

$$\begin{aligned} &\mathbb{E}^{\pi_\theta}_{\substack{s_{t:\infty} \\ a_{t:\infty}}}\left[\hat{A}_t\left(s_{0:\infty}, a_{0:\infty}\right)\nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right)\right] \\ &= \mathbb{E}^{\pi_\theta}_{\substack{s_{t:\infty} \\ a_{t:\infty}}}\left[A^{\pi_\theta,\gamma}\left(s_t, a_t\right)\nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right)\right] \end{aligned} \tag{6}$$

where $A^{\pi_\theta,\gamma}\left(s_t, a_t\right) = Q^{\pi_\theta,\gamma}\left(s_t, a_t\right) - V^{\pi_\theta,\gamma}\left(s_t\right)$ is the real discounted advantage function following policy π. For the n-step advantage:

$$\begin{aligned} A_t^{(n)} &= \sum_{i=0}^{n-1} r_{t+i} + \gamma^n V\left(s_{t+n}\right) - V\left(s_t\right) \\ &= r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right) \\ &\quad + \gamma\left(\sum_{i=1}^{n-1} r_{t+i+1} + \gamma^{n-1} V\left(s_{t+n}\right) - V\left(s_{t+1}\right)\right) \\ &= \delta_t + \gamma\left(\sum_{i=0}^{n-2} r_{t+i+1} + \gamma^{n-1} V\left(s_{t+n}\right) - V\left(s_{t+1}\right)\right) \\ &= \sum_{i=0}^{n-1} \gamma^i \delta_{t+i} \end{aligned} \tag{7}$$

Eq. (7) considers δ_t as a shaped reward with $V^{\pi_\theta,\gamma}\left(s_t\right)$ as the potential function. The n-step advantage is equal to the γ discounted sum of these shaped rewards.

Literature [12] defines the generalized advantage estimator GAE(γ, λ) as the exponentially weighted average of all n-step advantages:

$$\begin{aligned} A_t^{\mathrm{GAE}(\gamma,\lambda)} &= (1-\lambda)\left(\sum_{n=0}^{\infty} \lambda^n A_t^{(n+1)}\right) \\ &= (1-\lambda)\left(\sum_{n=0}^{\infty} \lambda^n \sum_{i=0}^{n} \gamma^i \delta_{t+i}\right) \\ &= (1-\lambda)\left(\sum_{i=0}^{\infty} \gamma^i \delta_{t+i} \sum_{n=i}^{\infty} \lambda^n\right) \\ &= (1-\lambda)\left(\sum_{i=0}^{\infty} \gamma^i \delta_{t+i} \lambda^i \sum_{n=0}^{\infty} \lambda^n\right) \\ &= (1-\lambda)\left(\sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i} \frac{1}{1-\lambda}\right) \qquad \text{since } \lambda < 1 \\ &= \sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i} \end{aligned} \tag{8}$$

III. LOGARITHMIC BASIS POLICY GRADIENT

This section studies the influence of different logarithmic bases within policy gradient methods. First, we study how varying the base of the logarithmic function influences the efficacy of policy gradient methods. We analyze these effects from the standpoint of the logarithmic functions and their gradients. Second, we propose the Adaptive LBPG method, which represents a paradigm shift by incorporating a dynamic logarithmic function adjustment based on real-time training metrics.

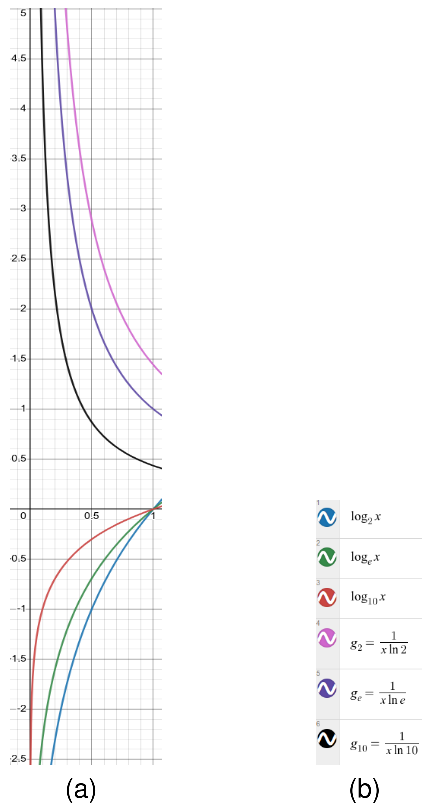

Fig. 1. Logarithmic function and their gradients. (a) The curves of logarithmic function and their gradients. (b) The corresponding curve markings, where x denotes the policy $\pi_\theta\left(a_t \mid s_t\right)$, g_b denotes the gradient of logarithmic function with base b.

Given the widespread use of the advantage function in deep RL methods, as detailed in works like [8] and [10], This paper's investigation is grounded in analyzing the policy gradient with the advantage function:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} A^{\pi_\theta}\left(s_t, a_t\right) \nabla_\theta \log_b \pi_\theta\left(a_t \mid s_t\right)\right] \quad (9)$$

In Eq (9), the subscript b in $\log_b$ denotes the base of the logarithmic function. Conventional policy gradient methods predominantly use the natural logarithm base $e \approx 2.718$ (i.e., $b = e$). This paper seeks to answer a pivotal question: *How do different logarithmic bases in the policy gradient formula (as shown in Eq. (9)) affect overall performance?* To answer this question, we scrutinize three commonly used bases in mathematical analysis—*loge*, *log2*, and *log10*, and examine their gradients. The gradient of the logarithmic function is:

$$\nabla_\theta \log_b \pi_\theta\left(a_t \mid s_t\right) = \frac{1}{\pi_\theta\left(a_t \mid s_t\right)\ln(b)} \nabla_\theta \pi_\theta\left(a_t \mid s_t\right) \quad (10)$$

Then, Eq. (9) can be written as:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} A^{\pi_\theta}\left(s_t, a_t\right) \frac{\nabla_\theta \pi_\theta\left(a_t \mid s_t\right)}{\pi_\theta\left(a_t \mid s_t\right)\ln(b)}\right] \quad (11)$$

Eq. (10) shows that although the probability of $\pi_\theta\left(a_t \mid s_t\right)$ is equal in $\log_e \pi_\theta\left(a_t \mid s_t\right), \log_2 \pi_\theta\left(a_t \mid s_t\right), \log_{10} \pi_\theta\left(a_t \mid s_t\right)$, there exist a constant $\ln(b)$ in $\nabla_\theta \log_b \pi_\theta\left(a_t \mid s_t\right)$. This leads to a different learning rate for these three bases.

Fig. 1 shows these logarithmic functions with different bases and the gradients of these logarithmic functions. Analyzing Fig. 1, we can find that although the gradients of $log_e \pi_\theta\left(a_t \mid s_t\right)$, $log_2 \pi_\theta\left(a_t \mid s_t\right)$, and $log_{10} \pi_\theta\left(a_t \mid s_t\right)$ have the same varying tendency, their values vary greatly. This undoubtedly affects the policy gradients and the final performances of the policy after multi-step updates.

Furthermore, although $log_e x$, $log_2 x$, and $log_{10} x$ have the same varying tendency, the function values vary greatly for the same x. In the vast majority cases, the probability of a good action of $\pi(a_{good} \mid s)$ is greater than the difference in the probability of the action $\pi(a_{bad} \mid s)$. because $\pi(a_{good} \mid s) > \pi(a_{bad} \mid s)$, by above knowable, $log\pi(a_{good} \mid s) < log\pi(a_{bad} \mid s)$, and while $\pi(a_{good} \mid s)$is close to 1, $log\pi(a_{good} \mid s)$is very small, close to 0. For a good action, because $\pi(a_{good} \mid s)$, so $log\pi(a_{good} \mid s)$is very small, this caused it to calculate the gradient is small, does this will seriously affect the contribution to the gradient update of the sample. For bad action, because $\pi(a_{good} \mid s)$is small, so $log\pi(a_{good} \mid s)$ is greater than good action $log\pi(a_{good} \mid s)$. As a result, the contribution of poor action samples to gradient updating is greater than that of good action samples. According to the above problem, the policy gradient formula of $log\pi(a \mid s)$ does improve.

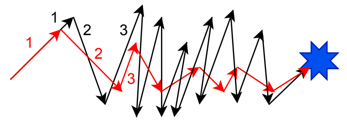

Fig. 2. Schematic diagram of multi-step gradient updates with different bases

Importantly, the probability of $\pi_\theta\left(a_t \mid s_t\right)$ is actually not equal in $\log_e \pi_\theta\left(a_t \mid s_t\right), \log_2 \pi_\theta\left(a_t \mid s_t\right), \log_{10} \pi_\theta\left(a_t \mid s_t\right)$. Fig. 2 shows a schematic diagram of multi-step gradient updates with different bases. For example, let the red and black lines represent the update trajectories of $\log_e \pi_\theta\left(a_t \mid s_t\right)$ and $\log_{10} \pi_\theta\left(a_t \mid s_t\right)$, respectively. Fig. 2 show that, even if the directions of the first gradient steps in $\log_e \pi_\theta\left(a_t \mid s_t\right)$ and $\log_{10} \pi_\theta\left(a_t \mid s_t\right)$ are the same, there exists a constant update size due to the logarithmic function gradient with different bases. Then, in the second update process, the vector sums of the gradients in two consecutive steps are different. Thus, after many gradient updates, the policy parameter of $\log_e \pi_\theta\left(a_t \mid s_t\right)$ and $\log_{10} \pi_\theta\left(a_t \mid s_t\right)$ are also different. Moreover, as shown in Eq. (10), the policy parameter differences superposes with the influence of $ln(b)$ will lead to a non-negligible influence on policy gradient.

We can conclude that: *The logarithmic function with different bases in the policy gradient has different performances. Thus, the logarithmic function matters in policy gradient deep RL.* Notably, while $loge$ and $log2$ generally facilitate faster learning, $log10$ exhibits superior stability. These insights provide valuable guidance for deep RL practitioners in tailoring log functions to specific challenges.

Based on the above insight, we propose the Logarithmic Basis Policy Gradient (LBPG) algorithm, which means the logarithmic function with different bases in the policy gradient. LBPG can be applied to various policy gradient algorithms. We write the general expression of the policy gradient with LBPG:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log_b \pi_\theta(a_t \mid s_t)\right] \tag{12}$$

where $b = [e, 2, 10]$. Specifically, when Ψ_t denotes different values, we get:

- $\Psi_t = \sum_{t'=t}^{\infty} r_{t'}$: LB-REINFORCE (LBREINFORCE).
- $\Psi_t = Q^{\pi_\theta}(s_t, a_t)$ approximated by neural network: LB-Actor-Critic (LBAC).
- $\Psi_t = A^{\pi_\theta}(s_t, a_t)$: LB-Advantage Actor-critic (LBA2C).
- $\Psi_t = A_t^{\mathrm{GAE}(\gamma,\lambda)}$: LB-GAE (LBGAE)

LBPG algorithm is summarized in Algorithm 1.

Furthermore, we propose the Adaptive LBPG algorithm to capitalize on the insights regarding the distinct impacts of various logarithmic bases on policy gradient RL performance. Adaptive LBPG algorithm firstly calculates returns $\mathcal{R}$ using rewards, then computes the return's mean $\mu_\mathcal{R}$ and the return's variance $\sigma^2_\mathcal{R}$. We define return's variance-mean ratio $\rho_\mathcal{R}$:

$$\rho_\mathcal{R} = \frac{\sigma^2_\mathcal{R}}{\mu_\mathcal{R}} \tag{13}$$

Adaptive LBPG algorithm determines the log base b with segmentation function, comparing $\rho_\mathcal{R}$ with low threshold τ_{low}, mid threshold τ_{mid} and high threshold τ_{high}:

$$b = \begin{cases} 2, & \text{if } \rho_\mathcal{R} < \tau_{low}, \\ e, & \text{if } \tau_{mid} \le \rho_\mathcal{R} < \tau_{high}, \\ 10, & \text{otherwise.} \end{cases} \tag{14}$$

A high $\rho_\mathcal{R}$ means a big return's relative variance, representing that the policy indicate instability. Thus, we need to choose a lower log base to stable the learning process. For example, Adaptive LBPG may lean towards a lower base such as $log2$ in environments with lower return's variance-mean ratio, signaling more stable learning conditions. In contrast, a low $\rho_\mathcal{R}$ requires a higher base like $log10$ that would increase the learning speed.

Adaptive LBPG dynamically shifts between logarithmic bases, specifically e, 2, and 10 in response to changing conditions and performance metrics observed during training. This mechanism enables LBPG to fine-tune the policy gradient optimization, making it highly suitable for the diverse and fluctuating requirements of different RL environments. Algorithm 2 summarizes the Adaptive LBPG algorithm.

Adaptive LBPG introduces a higher level of flexibility and efficiency for policy gradient methods, improving their ability to deal with the diverse and ever-changing reward landscapes that exist in various policy gradient methods optimization. As a result, Adaptive LBPG may expand the limits of what can be achieved with conventional policy gradient methods.

Algorithm 1 LBPG: Log-Bases Policy Gradient

Parameters: total training epoch L; learning rate α, β

```
1:  Initialize policy network parameters $\theta$ and value function network parameters $\omega$
2:  for $l = 1, \dots, L$ do
3:      while not done do
4:          Observe state $s_t$, get action $a_t$ from the policy network
5:          Receive $s_{t+1}$ and reward $r_t$ from the environment
6:          Append $(s_t, a_t, r_t, s_{t+1})$ to the transition-sample list
7:      end while
8:      Update value network (by TD residual, advantage value, etc.): For example, TD residual: $\delta_\omega = \frac{1}{M}\sum_m \nabla_\omega \mathbb{E}^{\pi_\theta}[(r_t + Q(s_{t+1}, a) - Q(s_t, a_t))]^2$ $\omega \leftarrow \omega + \beta_t \delta_\omega$
9:      Update the policy network: $\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log_b \pi_\theta(a_t \mid s_t)\right]$ with $b = [2, e, 10]$, where $\Psi_t$ the advantage function and value function, etc. $\theta \leftarrow \theta + \beta \nabla_\theta J(\pi_\theta)$
10: end for
```

Output: policy and critic parameters θ, ω

IV. Experiments and Analysis

This section conducted a series of experiments to validate the efficiency of the proposed methods. The test environments are the widely recognized control benchmark environments in OpenAI Gym [15]. We applied LBPG and Adaptive LBPG to several prominent policy gradient deep RL algorithms, including REINFORCE, Actor-Critic, A2C, and GAE to verify the efficiency of the proposed methods.

Experimental results in Fig. 3 demonstrate the influence of varying logarithmic bases when used in conjunction with the SGD optimizer. Fig. 3 shows that *loge* and *log2*, facilitate faster learning in most of these algorithms compared to *log10*. However, *log10* exhibits superior stability, suggesting its suitability in scenarios where training stability is paramount.

The experimental results provide valuable insights for deep RL practitioners who need to select the appropriate logarithmic function based on their specific requirements. Depending on the focus of the application, either the *log2* or *log10* bases can be the preferred choice. This adaptability can lead to improved performance in RL scenarios with fluctuating dynamics and reward structures.

Building on the initial findings, we further explore the learning dynamics using LBPG. Fig. 4 illustrates the relationship between average return and the standard deviation-to-mean ratio of batch returns in REINFORCE, AC, A2C, and GAE algorithms. Figs. 4 (a)-(h) show that when the return variance increases, meanwhile the return decreases. An intuitive idea is that we should decrease the learning rate to enhance stability

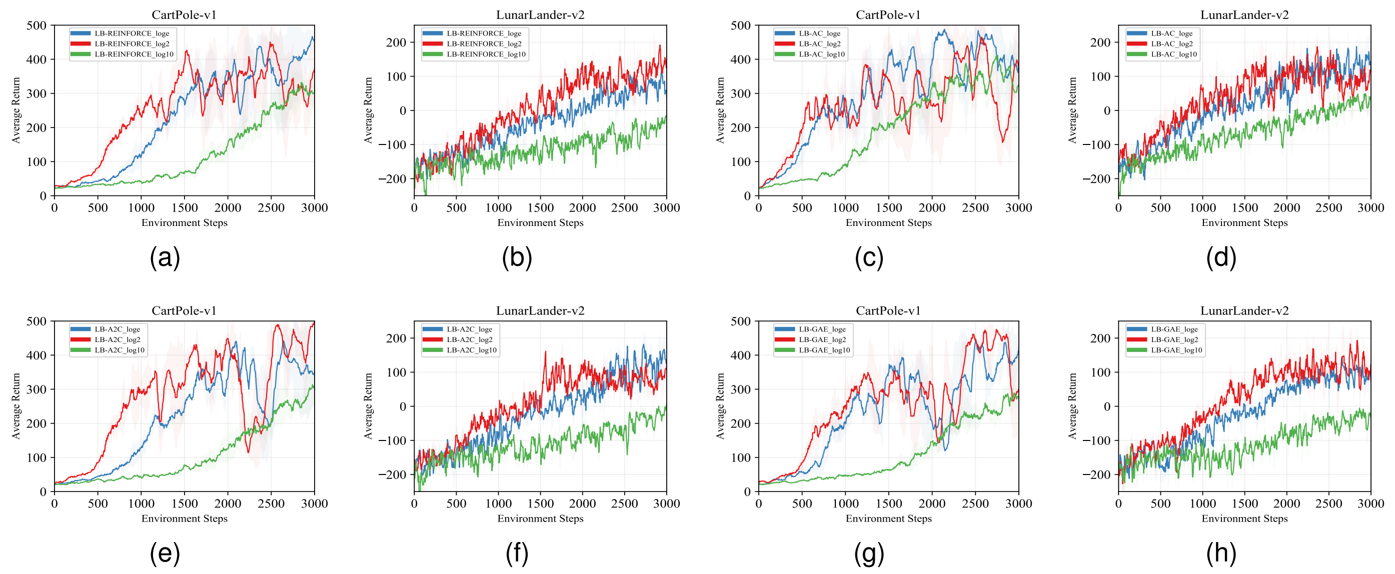

Fig. 3. Performances of LBREINFORCE, LBAC, LBA2C, and LBGAE on CartPole-v1 and LunarLander-v2. (a) LBREINFORCE-CartPole. (b) LBREINFORCE-LunarLander. (c) LBAC-CartPole. (d) LBAC-LunarLander. (e) LBA2C-CartPole. (f) LBA2C-LunarLander. (g) LBGAE-CartPole. (h) LBGAE-LunarLander.

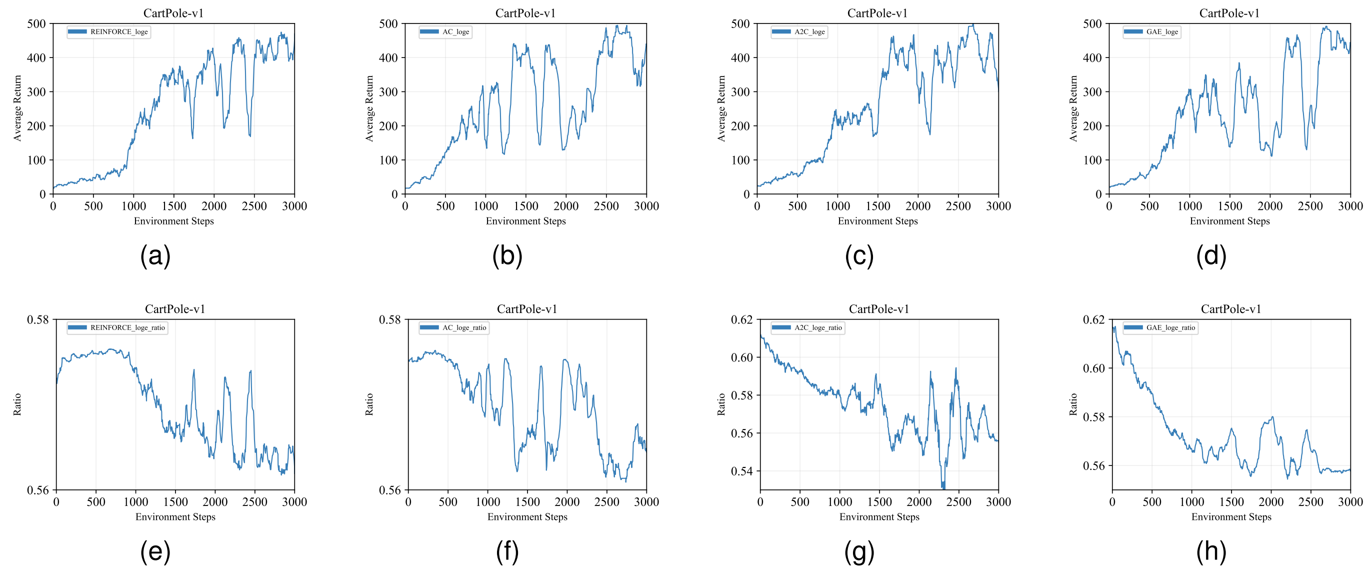

Fig. 4. Relationship of return and batch return variance on CartPole-v1. (a) (b) (c) (d) return of REINFORCE, AC, A2C, GAE. (e) (f) (g) (h) return std-mean ratio of REINFORCE, AC, A2C, and GAE

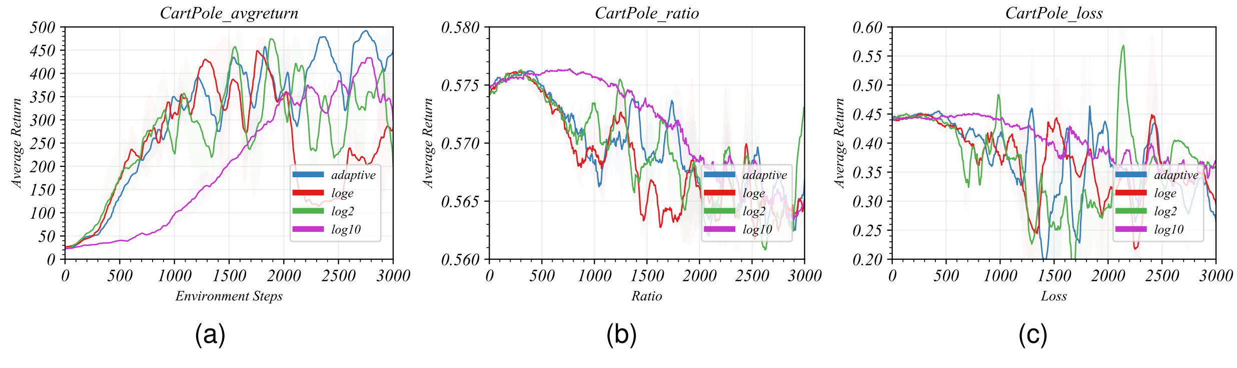

Fig. 5. Comparative analysis of return, loss, and ratio in the CartPole-v1 environment for various methods. (a) Returns using adaptive, log2, log10, and loge methods. (b) Ratio evaluations using adaptive, log2, log10, and loge methods. (c) Loss values for adaptive, log2, log10, and loge.

Algorithm 2 Adaptive LBPG: Adaptive Log-Bases Policy Gradient

Parameters: total training epochs L; learning rates α, β; low threshold τ_{low}, mid threshold τ_{mid}, high threshold τ_{high}

```
1: Initialize policy network parameters $\theta$ and value function network parameters $\omega$
2: for $l = 1, \ldots, L$ do
3:   Initialize episode reward to 0
4:   Reset environment and get initial state $s_0$
5:   while not done do
6:     Observe state $s_t$, get action $a_t$ and value estimate $v_t$ from the policy network
7:     Execute action $a_t$, receive reward $r_t$ and new state $s_{t+1}$ from the environment
8:     Store $(s_t, a_t, r_t, s_{t+1})$ in the transition sample list
9:     Update episode reward
10:  end while
11:  Calculate returns $\mathcal{R}$ using rewards
12:  Compute return's mean $\mu_{\mathcal{R}}$ and return's variance $\sigma^2_{\mathcal{R}}$
13:  Calculate returns ratio $\rho_{\mathcal{R}} = \frac{\sigma^2_{\mathcal{R}}}{\mu_{\mathcal{R}}}$
14:  if $\rho_{\mathcal{R}} < \tau_{low}$ then
15:    $b = 2$
16:  else if $\tau_{mid} \leq \rho_{\mathcal{R}} < \tau_{high}$ then
17:    $b = e$
18:  else
19:    $b = 10$
20:  end if
21:  Update value network:
22:  $\delta_\omega = \frac{1}{M} \sum_m \nabla_\omega \mathbb{E}^{\pi_\theta} [(r_t + Q(s_{t+1}, a) - Q(s_t, a_t))]^2$
23:  $\omega \leftarrow \omega + \beta \delta_\omega$
24:  Update the policy network with selected log base $b$:
25:  $\nabla_\theta J(\pi_\theta) = \mathbb{E}^{\pi_\theta} \left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log_b \pi_\theta (a_t \mid s_t)\right]$
26:  $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
27: end for
```

Output: Optimized policy and critic parameters θ, ω

in training. A small learning rate can achieve this; thus $log10$ can be chosen.

The extended experiments in Fig. 5 emphasize the Adaptive LBPG method's adaptive nature. For the thresholds, We set $\tau_{low} = 0.1$, $\tau_{mid} = 0.3$ and $\tau_{high} = 0.57$. Adaptive LBPG optimizes the learning rate by dynamically adjusting the logarithmic base between *log2*, *loge*, and *log10* based on the comparison variance-to-mean ratio of returns and thresholds. This adaptability is evident in Fig. 5, where Adaptive LBPG outperforms other methods in terms of return, stability, and loss. Figs. 5 (a)-(c) show the average return, ratio, and loss in the CartPole-v1 environment.

Limitation: It is noted that while LBPG influences policy gradient methods with the SGD optimizer, its influence is not as pronounced when using the Adam [16] optimizer. This discrepancy may be attributed to the momentum characteristics inherent in the Adam optimizer. Such insights highlight the importance of selecting the appropriate optimizer in conjunction with LBPG to maximize its efficacy in various RL scenarios.

V. Conclusions

This paper studied the influence of logarithmic functions with different bases (*loge*, *log2*, and *log10*) in policy gradient deep RL methods. We analyzed the role of these logarithmic functions and their gradients when using SGD as the optimizer in deep RL. We proposed the LBPG and Adaptive LBPG algorithms to enhance policy gradient methods. Experimental results showed that the choice of the logarithmic base influences policy gradient methods. *loge* and *log2* are generally more effective than *log10* in deep RL algorithms, but *log10* exhibits more stability. Furthermore, experimental results showed the importance of choosing appropriate logarithmic functions for policy gradient RL methods and highlighted Adaptive LBPG's potential to refine learning in deep RL tasks. For future work, we will study to apply Adaptive LBPG to multi-agent RL problems.

References

[1] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1861–1870.

[2] Q. Liu, Y. Li, S. Chen, K. Lin, X. Shi, and Y. Lou, "Distributional reinforcement learning with epistemic and aleatoric uncertainty estimation," *Information Sciences*, vol. 644, p. 119217, 2023.

[3] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[5] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229–256, 1992.

[6] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[7] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, no. 7-9, pp. 1180–1190, 2008.

[8] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1928–1937.

[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1889–1897.

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

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

[12] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *International Conference on Learning Representations*, 2016.

[13] N. Vlassis, A. Chandrashekar, F. Amat, and N. Kallus, "Control variates for slate off-policy evaluation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3667–3679, 2021.

[14] M. Uehara, C. Shi, and N. Kallus, "A review of off-policy evaluation in reinforcement learning," *arXiv preprint arXiv:2212.06355*, 2022.

[15] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.