RESPONSES TO REVIEWERS' COMMENTS

Anonymous authors
Paper under double-blind review

Attached please find a revision of paper with manuscript ID Paper397, entitled "Deep Reinforcement Learning With Adaptive Combined Critics". We are grateful to all four reviewers for their careful reading of our paper and many helpful suggestions. We have revised the manuscript taking into account these suggestions and will summarize below each of the specific issues raised in the reviews (in *italics*) along with our response (in blue). All the numbers of equations are updated as the revised version in Replies.

To Reviewer 1

(1) First, the theorems are incorrect or meaningless. Theorem 1 says the proposed algorithm AD3 (i.e. adaptive delayed deep deterministic policy gradient) converges. But the proof is for the tabular case and it becomes trivial. Furthermore, the proof itself seems to be incorrect. There should be restrictions on the function μ as it would affect what type of operator it is. Theorem 2 is simply incorrect. The inequality (13) does not imply policy improvement at all. I guess the authors want to say the policy under the policy parameter of the i+1th step has a higher expected return. But the notations are not saying that. The proof itself uses approximation signs arbitrarily, and it is unclear what the meaning is.

Reply: We have redone the proof of Theorem 1 in red color. The old Theorem 1 was not for the tabular case, subscripts represented different update iterations (time steps).

I did not know what restrictions on the function μ and what type of operator it affects, could you please clarify so I can learn it?

In Theorem 2, we use the clear notation of asymptotically expected policy improvement, and eq. (12) shows it. We use approximation signs to show the average is statistically equal to the expectation. Because our algorithms uses neural networks, all proofs are only meaningful statistically, which means we cannot solve problems strictly by the expectation of objective function.

(2) Second, experiments lack details and it is unclear what message they are trying to convey (e.x. Fig 2). How were those hyperparameters chosen? How many runs? What is the standard deviation across runs?

Reply: All figures have been explained in this paper. Message from Fig. 2 is shown in the third paragraph of subsection CONTINUOUS MAZE. You can also search for the key word Figure 2 and Fig. 2 to find it. We list detailed hyperparameters in the revised version in Table 1. The number of runs are shown as x-axis of figures. Also, 800 thousand iterations is written in the second paragraph of Subsection ROBOT ARM. We will also reproduce figures as standard deviation version.

(3) Third, many of the statements and notations are not rigorous.

Reply: We redefined some statements and notations in red colors. Others have been rigorously and clearly denoted.

(4) The authors never clearly state the objective function they are using. Is it the same as DDPG? And, it never makes a clear distinction between the objective used for updating the critic parameter and the actor parameter.

Reply: The objective functions are clearly denoted in this paper. You can search the key word "objective" to quickly find them. The overall objective is eq. (5), which is separated into several sub-objective functions. Specifically, objective functions for updating critics ω and Ω are eqs. (3) and (4), objective function for updating actor θ is eq. (6), objective function for updating the weight factor λ is eq. (7). By the way, we have clear definition of λ just below eq. (5). We have just defined them once they show up for the first time.

Besides, we make clear that the goal (expected return) is approximated by the combined critic value in added Lemma 1.

(5) The definition of $J(\theta)$ is confusing, what is the underlying distribution of s (first paragraph on page 3)? In eq (6) and (7), the definition of $J(\theta)$ changes.

Reply: The definition of $J(\theta)$ (first paragraph on page 3) is the expected return. More importantly, it is the objective function to update θ defined in DDPG. This is Section Background. The context after this section begins our work. Our definition for the objective function to update θ still follows the notation $J(\theta)$, but its expression changes into eq. (6).

Overall, the notation used to represent the objective function to update θ never changes throughout the paper, but its expression changes according to different algorithms. By the way, θ is the actor network parameter.

(6) "In actor-critic methods, the policy is deterministic and commonly parameterized as the actor network." This is inaccurate. The policy does not have to be deterministic in actor-critic methods.

Reply: This is indeed inaccurate, we have merged it into another sentence in red color.

(7) "However, updating two Q networks according to the same target estimate will make them less independent, which will further negatively affect the training efficiency." Why?

Reply: The critics are updated based on MSE between the critic and the target value. So if the target value is shared the same, then two critics are updated towards the same value, which can also be seen in the proof for Theorem 1 of Fujimoto et al. (2018).

(8) "Although the underestimation bias accompanying the minimization operator is far preferable ..., it indeed negatively affect the policy improvement at every iteration and further brings fluctuation on algorithm convergence." Why does it "indeed affect the policy improvement"?

Reply: At every iteration, if the lower critic is always chosen, it induces underestimation. The underestimation will harm policy improvement. References about underestimation are shown in the 3rd paragraph in Introduction.

(9) Important detail missing in the background. The authors should cite the DDPG or the DPG papers at least when describing the updating rule (2). In the original work, there is also a slower moving network for the actor used for computing bootstrap target to update the critic. For a paper concerning removing overestimation bias, it is a very important detail and should be clearly stated.

Reply: This is our mistake, citation for eq. (2) has been added.

The "soft" target updates (slower moving network) was cited below Lemma 1 or above eq. (10).

(10) Last, those important derivations (i.e. those from eq (8)) do not have justification.

Reply: I think there is no need to deduce soft updates in eq. (10). In DDPG and TD3, there are no derivations for soft updates. The soft updates are more like empirical techniques. We will see eq. (10) is just variants of soft updates.

To Reviewer 2

(1) I have three concerns due to which I am currently recommending rejection: First, the smaller concern is that the paper is currently difficult to understand, second, my major concern is that I am not sure about it's correctness, due to equation (9). Lastly, the experimental evaluation is insufficient.

Reply: The first and last comment are a bit subjective. So I explain eq. (8) in the revised version ((9) in old version).

It is meaningless to judge eq. (8) itself because it is just a definition. The purpose of eq. (8) is to ensure policy improvement by eq. (7). The correctness of its purpose is shown in Proof of Theorem 2. We rigorously show the policy improvement statistically, i.e., expected policy improvement. In continuous control, only expected policy improvement is meaningful.

We will reproduce experiments.

(2) Regarding understanding: The paper currently has a significant number of grammatical mistakes and unusual or wrong word choices. While this often is without consequence, unfortunately, at times it does impact the clarity and creates uncertainty about the argument the authors are trying to make. However, overall I believe I was able to understand the algorithm the authors are proposing, so this is not my major concern.

Reply: Thanks for this suggestion, we will improve.

(3) Regarding correctness of the algorithm: My main issue is with equation (9), i.e. with how lambda (the trade-off parameter in the convex combination between the two q-value functions) is updated. I believe equation (9) should almost always be negative because the minimum of two values is always smaller or equal than any convex combination of both, since lambda is restricted to 0 lambda 1. Which means that based on equation (11), lambda is actually optimized to maximize the value, i.e. it encourages overestimation instead of preventing it.

Reply: Eq. (8) is the comparison between the current lower Q-value (or minimized combined Q-value) with the combined value of the last iteration. It is not necessarily negative. If the policy is improving, it will be positive.

So if the statistical policy improvement is threatened, we prevent underestimation. Otherwise, we prevent overestimation.

(4) One caveat might be that equation (9) actually compares Q-values from before and after each update. It is not explained why this is done, but I believe that should change the sign of (9) in only very rare cases as I'd expect Q to move only very little in each update.

Reply: Just above eq. (8) in the revised version, we clarify that we provide the guarantee of policy improvement by eq. (7). This is exactly the explanation why this is done.

I also expect Q to move little per update. Based on this fact, the lower value (or minimized combined value) after each update has moderate probability to be smaller than the combined value before the update. This may cause some oscillation without the adjustment of eq. (8) (I think this is the truth by instinct).

(5) Lastly, λ is a global variable, whereas the minimum operation of TD3 is local, i.e. performed independently in each state. I believe this is a big difference in the proposed algorithms which is not discussed at all.

Reply: λ is shown as a global multi-dimension variable in the context. However, it can also be formulated as a state-dependent function approximation. And all theorems and proofs can apply to it when λ is state-dependent. We use both global network and state-dependent network for λ in our programming and the state-dependent case is better in performance. We can only have this discussion but we didt because this part would definitely exceed 8-pages limit. We have added this discussion part in Conclusion in red color.

(6) Regarding experimental evaluation: As the authors propose a novel, general purpose algorithm which aims to improve on TD3, it is important to evaluate it on a wide range of benchmarks, including the typical, widely used mujoco environments on which most current continuous control algorithms have been evaluated and compared.

Reply: Yes, I know you are right. But 6 continuous control environments are used, including Maze, Robot arm, Pendulum, Acrobat, Mountain car and Cartpole in this paper, which is strong enough to show effectiveness of the proposed algorithm.

The reason we did not use mujoco is because it cannot support synchronous and parallel training, which is applied in UAD3.

(7) Other, minor, points: In equation (5): What is the min over? What is the max over? Also, the Q functions have the wrong argument signature.

Below equation (5): I don't understand what xi and xi' are.

Reply: $\min$ is over λ, as shown under $\min$, to mitigate the overestimation bias. And $\max$ is over policy to maximize return. And eq. (5) has been more clear in revised version.

They are concerned about state space of MDP. ξ' is the transient state space. And ξ is the state space.

(8) Very minor point for Equation (6) and (7): Personally I don't like the introduction of the¯version of lambda and would find (1-λ) clearer in the equation.

Reply: Yes, you are right. We have revised.

To Reviewer 3

(1) The authors propose a deterministic policy-gradient algorithm that extends the TD3 algorithm (Fujimoto 2018). The main claim is that it reduces overestimation issues in a more effective way. Two Q-critics are maintained with separate parameters, but updated using the same transitions. Then a convex combination of these critics is used in the deterministic policy gradient update. The mixture parameter is learned on a slower time-scale to minimize this convex combination over states (instead of taking the minimum of the 2 critics per batch as in TD3). Another contribution in the paper is the Unbiased variant of the algorithm (UAD3), which addresses the off-policy nature of the replay mechanism of the AD3 algorithm described above. My understanding is that this is simply a version of the algorithm that does not use any replay mechanism and samples the state iid from the on-policy distribution, so it isnt a novel idea in itself.

Reply: The unbiased framework is not a novel idea in our work. But UAD3 is a new algorithm proposed in this paper. So we claim that UAD3 is an important part of our work.

(2) There are two theorems given to justify the algorithm choices, but I want to question their validity. The first one says that AD3 converges asymptotically, but no formal statement of what this means is given and the proof for it in the appendix only states broad facts about stochastic approximation, but nothing specific that applies to the AD3 algorithm. Theorem 2 is misleading in another way, it says that AD3 has "the property of asymptotical expected policy improvement", but it only really says that the critic value will be increasing, not that the actual policy value is increasing (and so an actual policy improvement step). Moreover, the proof contains some approximation steps which are not justified.

Reply: Proof of Theorem 1 has been redone in red color.

Concerning Theorem 2, I think the policy improvement means the policy choice should make the goal (expected return, or statistically, the critic value) increase. In our algorithm, the combined critic value approximates the goal (expected return). Theorem 2 tells that after the updates of actor, critic and weight factor, the goal is improved statistically, i.e., expected policy improvement.

Approximation steps are used to show the average is statistically equal to the expectation. We have clarified in the paper in red color. Because our algorithms uses neural networks, all proofs are only meaningful statistically, which means we cannot solve problems strictly by the expectation of objective function.

(3) The approach is tested in two simple continuous control environments (maze + reacher task). (Are these using the full state as input?). There the proposed approaches perform better it seems than the baselines (TD3 and DDPG), but there isnt any analysis to understand whether that was due to better critics - why not plot the estimated and true returns during learning to see whether AD3 indeed does better than the other critic update strategies? The experimental section is missing details to make these results reproducible and interpretable, for example what network architecture was used for the policy and critic? All learning curves have rather strange oscillation patterns. Is that an artifact of the smoothing used? How many seeds were used to obtain each learning curve?

Reply: Firstly, 6 continuous control environments are used, including Maze, Robot arm, Pendulum, Acrobat, Mountain car and Cartpole.

Whether to use the full state space as input depends on the observation of specific environment, for example, the classical control experiments start from small regions around the origins as inputs, and all comparisons are fair.

The better performance is due to the underestimation mitigation (expected policy improvement), which is included in Theorem 2. Besides, adaptive λ provides a superior way to reduce overestimation.

The accurate estimate of true discounted cumulative return needs large amount of data, so we just plot the average Q-value in this revision in the Appendix due to the time limit.

We have added details about hyperparameters in Table 1 and network architecture in Appendix for reproducible results.

Oscillation is because of the log compression of X-axis. We address the data using log x-scale to see more information before convergence. It is polynomial fitting of scatterers. Random seeds are used.

(4) Other comments: The objective to decide how to mix the critics could be better motivated. Making sure that the critic increase may not be the best criterion to ensure stability and reduce over/underestimation for example.

Reply: Thanks for this suggestion. There might be better combined critics extended from this work.

About the critic increase, see **Reply (2)**. Policy improvement means the policy choice should make the goal (combined critic value in this paper) increase.

(5) Minor things: "In actor-critic methods, the policy is deterministic" Actor-critic methods are actually more commonly described in a stochastic policy setting I believe. This work seems to be based on the deterministic policy gradient paper which is not cited.

Reply: We have revised this sentence into another sentence in red color. We have cited DDPG and TD3 in Introduction. DDPG is also cited above eq. (2) and above eq. (10).

(6) Citation style should use parentheses around names in most cases in the text.

Reply: Thanks, we have revised.

(7) Gattami 2019 is not the right reference for the Bellman equations and derived RL updates.

Reply: We have deleted this reference.

(8) Lots of equations are redundant and add unnecessary complexity. For example Eq 3 and Eq 4 are the same equations with different parameters, so why not call the two parameters w_1 and w_2 and write the equation once with w_i for $i \in 1, 2$ The hat on the λ in Eq 6,7 are hard to see because they are attached to the left bracket.

adopts $\rightarrow$ adopt

Transition slots $\rightarrow$ tuples?

Minus distance/reward $\rightarrow$ negative

Reply: We have simplified this paper. Concerning eqs. (3) and (4), using subscripts 1,2 will cause confusion in proofs because the subscripts are used to represent different iterations.

To Reviewer 4

(1) Although the algorithms are described clearly, I do not fully understand why the proposed method mitigates the overestimation bias issues. The authors claim that applying the clip variant of DDQN to the update procedure of the actor is more reasonable, but the experimental results did not support the authors' claim. Is it possible to show value estimates by TD3 and the proposed methods?

Reply: In overall objective in Eq. (5), the $\min$ operator is to mitigate the overestimation bias by tuning lambda. Eq. (6) shows the actor update to maximize return, and eq. (7) updates lambda to mitigate the overestimation as well as ensure policy improvement (mitigate underestimation).

We have plotted the average Q-value, i.e., the approximation of average discounted cumulative reward in the Appendix.

Since the overestimation will negatively affect the reward return performance, so we show it through average return in experimental results.

(2) I agree that the two timescale approach may suffer from a saddle point problem when used to solve Equation (5). However, I am not sure whether the two-step separation method can avoid the saddle point problem. Does AD3 outperform the two-timescale approach?

Reply: Saddle point problem does not exist in the two-step separation method, because λ is the same timescale as actor and critic. Take DDPG for example, the update of actor and critic are also separated but in the same timescale. We use a network to adaptively train λ together with the training of actor and critics per iteration.

By instinct, two-timescale is worse. We did not try the two-timescale approach, but it is reported to be a bad choice in the literature.

(3) The framework of unbiased DRL (UDRL) is interesting, but it is not the main contribution of the paper. So, a comparison of UAD3 with DDPG and TD3 is not fair. Is it possible to apply UDRL to DDPG and TD3? Since Zhang and Huang (2020) proposed Unbiased DDPG, I think the comparison of UAD3 with UDDPG helps understand how AD3 contributes to learning efficiency.

Reply: UAD3 in this paper is a variant of AD3 with the help of UDRL. Although UDRL is others' contribution, UAD3 is part of this work, so the results of UAD3 are plotted. We have reproduced the results of UDDPG in Classical control experiments.

(4) The authors compare the proposed methods with DDPG and TD3. However, DDPG does not consider the overestimation issue. It would be better to compare the proposed method with Averaged-DQN (Anschel et al., ICML 2017) and MaxMin Q-learning (Lan et al., ICLR 2020).

Reply: AD3 is proposed for continuous control, so algorithms for discrete DRL (variants of DQN) may not be proper baselines.

(5) Is Equation (5) correct? I think $-\lambda Q_\pi^2$ should be $+\lambda Q_\pi^2$. In addition, Q_π^1 and Q_π^2 are not defined explicitly. Overall, I think the idea of the paper is novel, but further improvements should be added to increase the score of the paper.

Reply: Yes, this is my mistake, thanks.

Q_π^1 and Q_π^2 were two critics or Q functions. We have revised it to make it clear.

REFERENCES

Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. 80:1587–1596, 2018.