A SMALL-GAIN ANALYSIS OF SINGLE TIMESCALE ACTOR CRITIC

ALEX OLSHEVSKY* AND BAHMAN GHARESIFARD†

Abstract. We consider a version of actor-critic which uses proportional step-sizes and only one critic update with a single sample from the stationary distribution per actor step. We provide an analysis of this method using the small-gain theorem. Specifically, we prove that this method can be used to find a stationary point, and that the resulting sample complexity improves the state of the art for actor-critic methods to $O\left(\mu^{-2}\epsilon^{-2}\right)$ to find an ϵ-approximate stationary point where μ is the condition number associated with the critic.

Key words. Reinforcement Learning, Actor Critic, Nonconvex Optimization.

AMS subject classifications. 93E35, 90-08

1. Introduction.

Since their introduction decades ago in [3], actor-critic methods have emerged as a popular approach to variety of reinforcement learning problems. Actor-only methods, such as policy gradient, seek to optimize a policy from a given class using noisy samples of the gradient. Estimating the gradient of a policy, however, is challenging because it requires knowing the true Q-values corresponding to the policy. While policy gradient methods usually estimate these using sample path rollout(s), actor-critic method adds a second parallel iteration designed to estimate the Q-values. It is sometimes asserted that, for a number of problems, actor-critic methods can outperform policy gradient due to the potentially high variance in sample path rollouts.

Possibly because of this, variants on actor-critic have achieved state-of-the-art performance on a number of benchmarks. For example, [19] performs a comparison of several popular RL methods on a variety Atari games, finding that in the human-start regime an actor-critic variant comes first.

Most previous theoretical analyses of actor critic were based on two variations on the same idea. The first variation is to slow down the actor by giving the actor a step-size that decays slower than the critic (e.g., [16, 17]). With this strategy, we can use a perturbed version of the analysis of Temporal Difference (TD) methods to argue that the critic always has a good approximation of the correct Q-values. With such step-sizes, actor-critic can then be thought of as a gradient method with error.

An alternative approach is to do many critic steps per actor step. This can ensure that the actor always has good approximations of the correct Q-values coming from the critic, and, once again, this allows an analysis of actor-critic as a gradient method with error. It should be emphasized, however, that this method can outperform vanilla policy gradient methods because the variance from this procedure can be much better than than the variance from trajectory rollouts (a point made in [32]).

However, it is not clear that actor-critic really needs either of these approaches, and indeed they are often not used in practice. Either approach is inconvenient, since both involve artificially slowing down the actor, and this results in slower convergence rates.

Indeed, as we explain more formally later in the paper, actor-critic is an attempt

*Department of Electrical and Computer Engineering and Division of System Engineering, Boston University, (alexols@bu.edu, https://sites.bu.edu/aolshevsky).

†Department of Electrical and Computer Engineering, University of California at Los Angeles, (gharesifard@ucla.edu, https://gharesifard.github.io).

to perform approximate iterations of stochastic gradient descent (SGD) on a non-convex objective. The choice of step-size is crucial here. For non-convex SGD, a simple argument shows that a step-size of $1/T^u$ multiplying the gradient at each step translates into a convergence rate of $\max(1/T^u, 1/T^{1-u})$ [1]. Informally speaking, the first term of the maximum comes from analyzing the expected progress towards the optimal solution at each step, while the second comes from the variance of each update. The best choice is $u = 1/2$ which trades off the best between these and clearly any other choice of $u \in [0, 1]$ results in a worse scaling of the maximum above. In actor-critic the same tradeoff reappears: slowing down the actor relative to the best $T^{-1/2}$ step-size decay to make it possible for the critic to "catch up" results in a worse decay with T.

The underlying technical challenge in getting rid of these approaches is that, without two timescales or having the critic take many steps per actor step, one cannot immediately argue that the actor has access to approximately correct Q-values from the critic. As a result, it becomes much more challenging to connect actor-critic to perturbed gradient descent. Our work seeks to address this by showing how a version of the small gain theorem can be used to give an analysis of actor-critic without either of the above approaches.

1.1. Literature Review. Because actor-critic methods are so widely used in reinforcement learning, recent years have seen several efforts to analyze their convergence. The most natural metric is the sample complexity, i.e., the number of samples needed to achieve a certain objective. In this case, the sample complexities are typically given in terms of finding a stationary point, i.e., one where the gradient squared of the actor is at most ϵ, or the running averages of the squared norms of gradients is at most ϵ. When the critic uses a linear approximation that cannot perfectly describe the actor's Q-values, the results below will give sample complexities until an approximate sttionary point, where the quality of the approximation depends on the accuracy that the critic is able to attain. We will use $O(\cdot)$ notation below, and this will be understood to hide logarithmic factors. Moreover, γ denotes the discount factor of the problem (to be formally defined later).

We begin by discussing the closely related works [31, 26, 18, 32]. All four of these papers gave a convergence analysis of actor-critic methods along with an associated sample complexities. The paper [31] studied a "two time scale" version where the actor and critic update using step-sizes that decay at different rates, attaining an $O(\epsilon^{-2.5})$ sample complexity. The paper [26] gave a non-asymptotic analysis of an approximate version of actor-critic where an additional source of error proportional to $O(1-\gamma)$ was introduced in the final outcome. The algorithm also followed the two-timescale approach. This work also attained $O(\epsilon^{-2.5})$ sample complexity. The paper [18] considers several versions of actor-critic methods. While a sample complexity of $O(\epsilon^{-3})$ is obtained for one version of actor-critic, this was improved to $O(\epsilon^{-2.5})$ for an accelerated variation. These sample complexities required many critic steps per actor step.

More recently, [33] was the first paper to attain a sample complexity $O(\epsilon^{-2})$. The algorithm used many critic steps per actor step. To explain the meaning of this result, note first that $O(\epsilon^{-2})$ is the rate one would obtain from gradient descent on a nonconvex function with noisy gradient evaluations. Thus if we did infinitely many critic steps per actor step, and then performed a standard Stochastic Gradient Descent (SGD) analysis, we would immediately obtain an $O(\epsilon^{-2})$ rate. The work [33] attains the same decay rate with ϵ, but with finitely many critic steps per actor step. We

note that the interpretation of some of the results in [33] was questioned in [15] (see Appendix C of that paper), but the critiques do not seem to apply to the actor-critic sample complexity we are discussing here.

While writing this paper, we found the work [8] which is very closely related to the present paper: [8] gives an analysis of single timescale actor-critic with $O(\epsilon^{-2})$ sample complexity. As far as we are aware, [33] and [8] are the only two papers to attain an $O(\epsilon^{-2})$ sample complexity for actor-critic; for these papers, we will thus also describe scaling with the condition number of the critic, to be formally defined in the main body of the paper, and which we denote by μ. In terms of both ϵ and μ (and treating the remaining variables as constants), the scaling in [33] is $O\left(\mu^{-3}\epsilon^{-2}\right)$. The scaling in [8] is a little worse at $O\left(\mu^{-4}\epsilon^{-2}\right)$.

Other related work considered algorithms building on the actor-critic framework. For example, the paper [14] considered a two-time scale version of the "natural" actor-critic method, where additional entropy regularizations are added to the update. Applied to a linearized version of the actor-critic objective, this gave an $O(\epsilon^{-4})$ sample complexity until that linearized objective is within ϵ of its optimal value. The paper [2] considers a variation on actor critic by introducing a target update, with the result being three different time scales for updates. The final sample complexity is $O(\epsilon^{-3})$. A similar sample complexity was given in the off-policy setting in [15]. We also mention [24], which gave an iteration complexity analysis of actor-critic assuming the critic computes an approximation to the correct values per each actor step.

For policy gradient methods, a sample complexity of $O(\epsilon^{-2})$ is known; see e.g., [34] for a modern analysis, and see also [32] for speedup. If more structure about the problem is known and one considers an exact version, one can sometimes obtain a better rate; for example, [12] gives a geometric rate for an exact version of policy gradient when applied to a Linear-Quadratic Regulator (LQR) problem. Likewise, information about about the structure of the optimal policy in general was shown to be useful in [25].

1.2. Our contribution. Our contribution is to use the small-gain theorem to analyze a single timescale actor critic method, i.e., a method which uses proportional step-sizes and a single critic step per actor step. We show that this method works in the same sense that all the previous methods work: it finds a stationary point when the critic's approximation is perfect, and an approximate stationary point otherwise. We obtain a sample complexity of $O\left(\mu^{-2}\epsilon^{-2}\right)$.

This compares favorably to the previous literature. One point of comparison is the now-classic work [7] also gave a single timescale method; however, that method only converged to a neighborhood of a stationary solution even when the critic's approximation is perfect. As mentioned above, the only other analysis we know of single-timescale actor critic is from [8], which gives a worse sample complexity of $O\left(\mu^{-4}\epsilon^{-2}\right)$. The main difference appears to be that where [8] use an explicit Lyapunov analysis, we instead rely on a small-gain analysis, which allows us to improve by a quadratic factor in μ^{-1}. The key step in our approach is a nonlinear version of the small gain theorem which handles the complex dependencies between the actor-to-critic and critic-to-actor gains.

2. Backround: Markov Decision Processes and Actor-Critic. We consider a Markov Decision Process (MDP) with a finite number of states and actions. When action a is taken in state s, we will denote the probability that the next state is s_{next} by $P(s_{\text{next}}|s,a)$, with $c(s,a)$ being the expected cost. A policy is defined as a mapping from the set of states to the set of actions. Defining c_t to be the random

variable representing the cost suffered at step t, our goal is to find a policy minimizing the infinite-horizon objective $E\left[\sum_{t=0}^{\infty}\gamma^t c_t\right]$, where $\gamma \in (0,1)$ is the discount factor.

We will assume that we have parametrized the set of policies by θ with $\pi_\theta(a|s)$ being the probability of choosing action a in state s. We define the matrix P_θ to be the probability transition matrix among state-action pairs when actions are chosen according to π_θ. We define the value of a state s as $J^*_\theta(s) = E_{\theta,s}\left[\sum_{t=0}^{\infty}\gamma^t c_t\right]$, where the subscripts indicate that the starting state is s and the policy followed is π_θ. Similarly, we define the Q-value as $Q^*_\theta(s,a) = E_{\theta,s,a}\left[\sum_{t=0}^{\infty}\gamma^t c_t\right]$, where the subscripts indicate that one starts by taking action a in state s and follows the policy π_θ thereafter. We fix the initial distribution to some η and consider the expectation

$$V^\eta_\theta = \sum_s \eta(s) J^*_\theta(s), \tag{2.1}$$

which is the expectation of the objective starting from distribution η.

We will use $p^\eta_{t,\theta}(s)$ to denote the probability that the state is s at time t if policy π_θ is followed after initializing the state from distribution η. The quantity $p^\eta_{t,\theta}$ will refer to the vector that stacks up $p^\eta_{t,\theta}(s)$ over all states s. Since η will be fixed throughout this paper, we will usually omit it and simply write $p_{t,\theta}(s)$, and likewise, we will simply write V_θ to represent the left-hand side of Eq. (2.1).

The policy gradient theorem [28] allows us to efficiently differentiate V_θ:

$$\frac{d}{d\theta}V_\theta = \sum_{s,t=0,1,\ldots} \gamma^t p_{t,\theta}(s) \sum_a \pi_\theta(a|s) Q^*_\theta(s,a) \frac{d}{d\theta}\log \pi_\theta(a|s). \tag{2.2}$$

This identity allows us to efficiently apply a stochastic version of gradient descent as follows. We generate a random triple (t, s_t, a_t) with probability $(1-\gamma)\gamma^t p_{t,\theta}(s_t)\pi_\theta(a_t|s_t)$ and update as

$$\theta_{t+1} = \theta_t - \beta_t Q^*_{\theta_t}(s_t,a_t)\frac{d}{d\theta}\log \pi_{\theta_t}(a_t|s_t), \tag{2.3}$$

for an appropriately chosen step-size β_t. In the sequel, we will use ν_θ to refer to the probability distribution over pairs (s_t, a_t) induced by this sampling procedure.

The update of Eq. (2.3) requires knowledge of the true Q-values. A policy gradient method will accomplish this with a trajectory rollout procedure. Specifically, a trajectory will be generated starting at state (s_t, a_t) and the Q-value estimated by computing the empirical reward $\sum_t \gamma^t c_t$ on that trajectory. In practice, one typically attempts to save on sample complexity by generating just one trajectory and applying Eq. (2.3) to each state-action pair while generating the empirical Q-value estimate by looking at the sub-trajectory beginning at s_t, a_t.

By contrast, an actor-critic method will run a second update intended to produce estimates of the true Q-values. Since the number of state-action pairs (s,a) is, in general, quite large, one typically attempts to find an approximation of the true Q-values. We will assume that the critic will try to approximate the quantity $Q^*_\theta(s,a)$ linearly as

$$Q^*_\theta(s,a) \approx \sum_i \omega_i \phi_i(s,a),$$

where each $\phi_i(\cdot,\cdot)$ can be thought of as extracting a feature from a state-action pair. We can write this compactly as

$$Q^*_\theta(s,a) \approx \phi(s,a)^T \omega,$$

where the vector $\phi(s,a)$ stacks up all the $\phi_i(s,a)$, and likewise ω stacks up the weights ω_i. Moreover, by stacking up these vectors $\phi_i(s,a)$ as rows into a matrix Φ, this can be written as

$$Q_\theta^* \approx \Phi\omega.$$

Here Q_θ^* is a vector with as many entries as the number of state action pairs.

It is usually assumed that the best weights ω are unknown while a method to compute the features is available. To compute good weights, the critic will perform a TD(0) update by generating the state-action pair s_t', a_t', then generating the next state and action s_t'', a_t'' by following the MDP and the policy π_θ and suffering a cost of c_t' in the process, and finally performing the update:

$$\omega_{t+1} = \omega_t + \alpha_t \left(c_t' + \gamma\phi(s_t'', a_t'')^T\omega_t - \phi(s_t', a_t')^T\omega_t\right)\phi(s_t', a_t'), \tag{2.4}$$

for some appropriately chosen step-size α_t. It is important that s_t', a_t' are sampled from the stationary distribution corresponding to policy π_θ; otherwise, the TD(0) update above may diverge even for fixed θ. When s_t', a_t' are sampled from the stationary distribution, we will denote by μ_θ the corresponding distribution over the four-tuple $(s_t', a_t', s_t'', a_t'')$.

In this paper, we will analyze an algorithm along these lines which we call *single-sample actor critic*. Its pseudocode is given in the algorithm box below. Note that it replicates Eq. (2.3) and Eq. (2.4) except that, in the actor's update of Eq. (2.3) now uses approximation $\phi(s_t,a_t)^T\omega_t$ instead of the true Q-value.

We also add a projection onto the set Ω to the critic update, which we take to be a compact convex set. Below, we discuss how to choose this set to be large enough so that it does not affect what the method is converging to. This is a standard tweak in actor-critic methods: it is a simple way to ensure a-priori that the iterates remain bounded. For example, the earlier papers [31, 14, 26, 11] all add a projection, while [2] explicitly assume that the iterates remain bounded (which we can do as well without affecting any of the results in this work).

Algorithm 2.1 Single sample actor critic

Initialize at arbitrary θ_0, ω_0. **for** $t = 1, 2, \ldots$

Generate a pair (s_t, a_t) from ν_θ and four-tuple $(s_t', a_t', s_t'', a_t'')$ from μ_θ (independently from each other and from all past samples) and update as

$$\theta_{t+1} = \theta_t - \beta_t \left(\phi(s_t,a_t)^T\omega_t\right)\frac{d}{d\theta}\log\pi_\theta(a_t|s_t) \tag{2.5}$$

$$\omega_{t+1} = P_\Omega\left[\omega_t + \alpha_t\left(c_t' + \gamma\phi(s_t'', a_t'')^T\omega_t - \phi(s_t', a_t')^T\omega_t\right)\phi(s_t', a_t')\right] \tag{2.6}$$

We remark that we call our method "single sample actor critic" because it only uses a single sample to update the actor and a single sample to update the critic at each step. However, note that the samples here are not samples from the MDP directly: rather, we need to sample from the stationary distribution for the critic, and from the distribution ν_θ for the actor. Multiple samples from the MDP will be needed to generate these single samples.

Specifically, this can be done by having the critic generate a sufficiently long path from the MDP for mixing to occur and throw away all the samples except the last one. Likewise, the actor can generate a geometric random variable with density proportional to γ^t, sample from it, and generate path from the MDP with length

equal to the sample thus generated. As a consequence, in order to translate any sample complexity obtained for this procedure into a sample complexity in terms of MDP samples, one needs multiply by $(1-\gamma)^{-1}$ plus an upper bound on the the mixing time of the policies.

3. Assumptions. To analyze single-sample actor critic, we will need to make a number of assumptions. The assumptions we will make are quite standard in the literature, and will generally assume that a quantity encountered through the course of the algorithm is bounded. Our first assumption will assume that the costs encountered throughout the algorithm do not blow up. This assumption helps ensure that various quantities appearing throughout the execution of the actor-critic method are naturally bounded.

ASSUMPTION 1 (Bounded Costs). *There exists some finite $C_{\max}$ such that the costs c_t belong to $[-C_{\max}, C_{\max}]$ with probability one.*

Next, recall that the critic will generate samples from the stationary distribution from each policy. We need to ensure that the stationary distribution exists, and additionally, the convergence to it needs to be uniform over all θ, as in the following assumption.

ASSUMPTION 2 (Uniform Mixing). *Let $\rho_{t,\theta}$ be the distribution over state-action pairs after t transitions of following policy π_θ. There is a distribution ρ_θ and constants C and $\lambda \in [0,1)$ such that*

$$||\rho_{t,\theta} - \rho_\theta||_1 \leq C\lambda^t.$$

Note again the uniformity in the right-hand side above in that C and λ do not depend on θ. Next, we will also need some regularity assumption on the parametrization of the policy. It goes without saying that optimization of smooth functions is easier than optimization of non-smooth functions and the following assumption ensures that various quantities appearing throughout the execution of actor-critic are smooth.

ASSUMPTION 3 (Smoothness of policy parametrization). *The function $\pi_\theta(a|s)$ is a thrice continuously differentiable function of θ for all state-action pairs s, a. Further, there exist constants K, K', K'' such that for all θ_i, s, a we have*

$$||\nabla_\theta \log \pi_\theta(a|s)|| \leq K, \quad ||\nabla_\theta \pi_\theta(a|s)|| \leq K', \quad ||\nabla^2 \pi_\theta(a|s)|| \leq K''.$$

Next, observe that if the columns of the matrix Φ are linearly dependent, then we have some redundancy in the features: we can delete some features without altering the subspace that can be represented as a linear combination of feature vectors. It is of course helpful to assume the features are not redundant. If the features were redundant, then the critic might have multiple optimal solutions when approximating the value function corresponding to a policy, which complicates the analysis. Additionally, it is also convenient to normalize the features so that all the feature vectors $\phi(s,a)$ have at most norm one.

ASSUMPTION 4 (Nonredundancy and norm of features). *The matrix Φ is nonsingular and each of its rows has at most unit norm.*

Next, it is natural that the final accuracy of actor-critic depends on how well the critic is able to approximate the true Q-functions encountered in the course of the

algorithms as linear combinations of features. For example, if the featurs are "bad" in that they fail to accurately approximate the Q-values associated with the actor policies, clearly actor-critic will perform poorly. To that end, we have an assumption that provides a measure of this.

ASSUMPTION 5 (Approximability of the true Q-values by the critic). *Define ω_θ to be the limit of temporal difference update of Eq. (2.4) when the policy is fixed to π_θ and the projection is removed*[1]*. Then for some $\delta > 0$,*

$$\sup_\theta E_{s,a\sim\nu_\theta} \ |Q^*_\theta(s,a) - \phi(s,a)^T\omega_\theta| \leq \delta.$$

In words, δ determins how well the true Q-values are approximated by TD learning on the average. If the linear approximation always perfectly describes the true Q-values, then $\delta = 0$. As we will see, δ will appear in our convergence results below, as in the appropriate sense getting a final error in actor-critic that is too small compared to δ is not possible.

Before making our next assumption, we need to introduce some new notation. Observe that it is possible to write (2.6) as

$$\omega_{t+1} = P_\Omega\left((I + \alpha_t A_t)\omega_t - \alpha_t b_t)\right). \tag{3.1}$$

Indeed, inspecting (2.6), we have that

$$A_t = \phi(s'_t, a'_t)(\gamma\phi(s''_t, a''_t) - \phi(s'_t, a'_t))^T, \quad b_t = -c'_t\phi(s'_t, a'_t) \tag{3.2}$$

Further, let us define $A_\theta = E[A_t]$ and $b_\theta = E[b_t]$. We emphasize the dependence on θ here because the expectation is being taken by generating the tuple s'_t, a'_t, s''_t, a''_t from μ_θ. Since these samples are being generated i.i.d., there is no dependence on t. With this notation, our next assumption is as follows.

ASSUMPTION 6 (Exploration). *There exists a $\mu \in (0,1)$ such that*

$$\sup_\theta \sup_{||x||=1} x^T A_\theta x \leq -\frac{\mu}{2} < 0.$$

The assumption is labeled as "exploration" because it holds if the policies π_θ explore all state-action pairs, as explained in Section 3.1 below. It is also a standard assumption: it is made in [31, 26, 24], and is closely related to assumptions which are made in [18, 2].

Finally, we need to make sure our compact set Ω is large enough so that projection onto it does not introduce spurious minima.

ASSUMPTION 7 (Projection Set). *The set Ω is a compact convex set which contains all of the vectors ω_θ in its interior.*

The compactness of this set allows us to ensure that the critic iterates do not grow unbounded. The further assumption that ω_θ is there to make sure that the quality of critic approximation is unaffected by the projection.

Remark 3.1. With one exception, there are no dependency relationships between the assumptions we make. In particular, it is possible for any subset of them to

[1] That this limit exists was proven under Assumptions 1, 2, and 4 in [29].

be satisfied or unsatisfied. The only exception to this is that Assumption 7 cannot even be stated without Assumption 2. Indeed, the matrix A_θ is defined by random sampling from the distribution μ_θ (defined immediately after Eq. (2.4)), which is built from the stationary distribution of the policy π_θ; Assumption 2 says that this stationary distribution is well-defined.

3.1. The Exploration Assumption. We now discuss the form of Assumption 6: specifically, we explain why we need to have a set of policies that have some exploration, and why that implies that Assumption 6 must hold.

Indeed, suppose that a particular action a in state s has a large negative cost, so much so that the best policy must always take it. Any sample-based method needs to select that action at least once before it can incorporate this knowledge. However, glancing at Eq. (2.5) we see this can fail to happen. Indeed, it may be that $\pi_\theta(a|s) = 0$ for a set of θ; then (s, a) will never appear in Eq. (2.5) and the update might always stay within the set of θ where $\pi_\theta(a|s) = 0$.

One way to overcome this is to require that the policies in question have a positive probability of choosing every state action pair. For a finite-state MDP, this immediately implies Assumption 6, as the proposition below explains.

PROPOSITION 3.2. *Suppose Assumption 2 and Assumption 4 hold, and suppose further there exists some* $p_{\min} > 0$ *such that*

$$\inf_{\theta,s,a} \pi_\theta(a|s) \geq p_{\min}.$$

Then Assumption 6 holds.

Proof. We will rely on the main result of [20], which interprets the mean TD(0) direction as a "gradient splitting" of an appropriately defined function and uses this to derive the following identity:

$$(\omega - \omega_\theta)^T A_\theta(\omega - \omega_\theta) = -(1-\gamma)||\Phi(\omega - \omega_\theta)||_{\mathrm{D}}^2 - \gamma||\Phi(\omega - \omega_\theta)||_{\mathrm{Dir}}^2, \tag{3.3}$$

where, adopting the notation $\rho_\theta(s)$ for the stationary probability of state s under the policy π_θ,

$$\begin{aligned} ||x||_{\mathrm{D}}^2 &= \sum_{s,a} \rho_\theta(s)\pi_\theta(a|s)x(s,a)^2 \\ ||x||_{\mathrm{Dir}}^2 &= \sum_{s,a,s',a'} \rho_\theta(s)\pi_\theta(a|s)P(s'|s,a)\pi_\theta(a'|s')(x(s,a) - x(s',a'))^2 \end{aligned}$$

Note that while the main result of [20] was shown for TD(0), the proof applies immediately to TD(0) on the state action pairs. Now the assumptions clearly imply a uniform bound over all θ on how small $\rho_\theta(s)\pi_\theta(a|s)$ can be, which along with (3.3) yield the result. ∎

The parameter μ may be thought of as the condition number associated with the critic. Indeed, glancing at Eq. (3.3) we see that the left-hand side measures the inner product between the $\omega - \omega_\theta$, the offset from the TD limit, and the TD direction $A(\omega - \omega_\theta)$. The definition of μ bounds this as being below $-(\mu/2)||\omega - \omega_\theta||^2$. Thus μ tells us how well the TD direction of the critic aligns with the vector pointing from ω to the final limit. It is thus not surprising that existing analysis of temporal difference learning under the optimally decaying step-size $1/t$ have factors of μ^{-1} in the final bounds (e.g., [5, 27, 9]).

4. Our Main Result. We will use the short hand

$$\Delta_t = \omega_t - \omega_{\theta_t}$$

to denote the difference between the critic iterate at time t and the ultimate limit of the critic update corresponding to the policy π_{θ_t}. Further, it will be convenient to adopt the shorthand

$$\nabla_t = \nabla V(\theta_t).$$

Since the function $V(\theta)$ is not assumed to be convex, our goal is to argue that actor-critic finds a point such that $\nabla_t \approx 0$. A standard performance measure for this in non-convex optimization is the running average of the gradients, and it is typically argued that this is close to zero. Below, we find it convenient to make a slight modification, taking the average over the last half of the iterations.

THEOREM 4.1. *Suppose Assumptions 1 – 7 hold. We can choose $\alpha_t = 1/\sqrt{t}$ and $\beta_t = c/\sqrt{t}$, where $c > 0$ is an appropriate constant chosen depending on the problem parameters such that the sequence of iterates produced by the Single Sample Actor Critic satisfies*

$$\begin{aligned} \frac{1}{t/2}\sum_{k=t/2}^{t-1} ||\nabla_k||^2 &= O\left(\delta^2 + \frac{\mu^{-1}}{\sqrt{t}}\right) \\ \frac{1}{t/2}\sum_{k=t/2}^{t-1} ||\Delta_k||^2 &= O\left(\delta^2 + \frac{\mu^{-1}}{\sqrt{t}}\right) \end{aligned}$$

where all quantities except δ, μ, t are treated as constants in the $O(\cdot)$-notation.

In particular, if the critic can approximate the Q-values exactly, i.e., $\delta = 0$, we have that the running average of the last half of the gradient norms squared approaches zero. Moreover, this running average is below ϵ after $O\left(\mu^{-2}\epsilon^{-2}\right)$ iterations. As discussed earlier, this sample complexity compares favorably to the scalings with μ and ϵ obtained in the previous literature.

Finally, if a specific point with gradient squared upper bounded by ϵ is sought, a simple trick is to take t large enough so that the right-hand side of Theorem 4.1 is below ϵ, and then take a uniformly random iterate from the last $t/2$ iterates. Then, in expectation the squared norm of the resulting gradient will be at most ϵ.

Remark 4.2. In practice, step-sizes in actor-critic are almost always taken to be small and fixed on both the actor and the critic. In particular, as afar as we are aware, practitioners never choose a step-size on the actor that decays faster to zero than the step-size of the critic. A reader wishing to read about the details of a modern, state-of-the-art implementation of actor-critic methods can refer to [19].

4.1. Key ideas in the proof of Theorem 4.1. We will proceed by applying a version of the small gain theorem to the quantities of the left-hand side of Theorem 4.1, i.e., to the averages of $||\nabla_k||^2$ and $||\Delta_k||^2$. In the "easy" direction, it is clear that if the error on the critic side is small (i.e., if the $||\Delta_k||$s are small), then the error on the actor side is small too; this follows via standard "SGD with errors" type analysis (indeed, we can view the actor as performing SGD with the error coming from the critic approximation). In the more difficult direction, one needs to argue that if the actor is approximately close to stationarity, i.e., the ∇_k are small, then the errors of the critic Δ_k will also be small. If separate bounds conforming to this intuition can

be established, one can attempt to put them together into an unconditional bound on both actor and critic using a small-gain type analysis.

Let us explain the challenges in deriving a bound on the critic error Δ_t in terms of the actor stationarity ∇_t. If the actor simply followed gradient descent on the objective $V(\theta)$, then it would be fairly straightforward that small actor gradients imply small critic errors; this would follow since small gradients ∇_k means the actor moves little in expectation from step to step, so that the critic is approximately doing temporal difference steps. Thus one could simply apply a standard analysis of temporal difference learning with some added perturbations.

Unfortunately, because the actor relies on the critic for its estimates of the Q-values from which it builds its estimate of the gradient direction, it is possible for the actor to be close to stationarity (i.e., ∇_k small) and yet move quite far from step to step (i.e., $\theta_{k+1} - \theta_k$ large). This dependency is what makes it challenging to argue that if the actor is close to stationarity, the critic errors are small. Ultimately, this is what results in a nonlinear relationship between the gains involved and forces us to use a nonlinear small gain theorem for the analysis. We remark that our proof does not rely on the ODE method [6, 10].

5. Proof of our main result. We now begin a proof of our main result. Our first step is to reformulate the actor-critic update in terms of a coupled gradient-liked updates which has more of an optimization flavor. However, we first need to introduce some convenient notation. *From this point on, we assume Assumptions 1 – 7 hold.*

5.1. Notation. For convenience, we will introduce the notation

$$Q_\theta(s,a) = \phi(s,a)^T \omega_\theta,$$

where recall that ω_θ was introduced in the statement of Assumption 5. Informally, $Q_\theta(s,a)$ will denote the "ultimate TD approximation" of the true Q-value $Q^*_\theta(s,a)$: it is what we would get if we fixed θ and let the critic take infinitely many steps.

Moreover, we will use

$$Q_t(s,a) = \phi(s,a)^T \omega_t.$$

Informally, $Q_t(s,a)$ the estimate of the true Q value $Q^*_\theta(s,a)$ which is obtained by the critic after t steps of the actor-critic method. We stack these up into the vector Q_t which has as many entries as the number of state-action pairs. Formally, $Q_t = \Phi\omega_t$.

5.2. Reformulating Actor-Critic. We now give a re-formulation of the underlying problem in terms of two coupled gradient-like updates. We note that this section is not particularly new and such reformulations are common in all previous analysis of actor-critic.

We begin by rewriting Eq. (2.6) as

$$\omega_{t+1} = P_\Omega \left[\omega_t - \alpha_t \left(\widetilde{\nabla}(\theta_t, \omega_t) + w_c(t) \right) \right].$$

Here $w_c(t)$ is a random variable with zero expectation conditional on the entire past trajectory; and $\widetilde{\nabla}(\theta_t, \omega_t)$ is the expected TD direction. More formally, we have that

$$\widetilde{\nabla}(\theta_t, \omega_t) = -A_{\theta_t}\omega_t + b_{\theta_t},$$

where the quantities A_θ, b_θ were defined earlier after Eq. (3.2).

Note that, since ω_θ is defined to be the limiting point of temporal diference learning (which exists due to the results of [29]), and is assumed to lie in the interior of the set Ω, we have that

$$\widetilde{\nabla}(\theta, \omega_\theta) = 0. \tag{5.1}$$

An immediate consequence of this is that

$$\begin{aligned} \widetilde{\nabla}(\theta,\omega)^T(\omega - \omega_\theta) &= (-A_\theta \omega + b_\theta)^T(\omega - \omega_\theta) \\ &= (-A_\theta \omega + b_\theta)^T(\omega - \omega_\theta) - (-A_\theta \omega_\theta + b_\theta)^T(\omega - \omega_\theta) \\ &\geq \frac{\mu}{2}||\omega - \omega_\theta||^2, \end{aligned}$$

where we used Eq. (5.1) while the last inequality used Assumption 6.

We next turn to the actor update. It is immediate that we can rewrite Eq. (2.5) as

$$\theta_{t+1} = \theta_t + \beta_t \left(\sum_{s,k=0,\ldots} \gamma^k p_{k,\theta_t}(s) \sum_a \pi_{\theta_t}(a|s) Q_t(s,a) \frac{d}{d\theta} \log \pi_{\theta_t}(a|s) + w_{\rm a}(t) \right),$$

where $w_{\rm a}(t)$ has zero expectation conditional on the entire past trajectory. Moreover, by the independent sampling by the actor and the critic, we have that conditional on the past trajectory, $w_{\rm a}(t)$ and $w_{\rm c}(t)$ are independent.

Let us further abstract this somewhat by defining

$$\hat{g}(\theta, Q) = \sum_{s,t=0,\ldots} \gamma^t p_{t,\theta}(s) \sum_a \pi(a|s) Q(s,a) \frac{d}{d\theta} \log \pi_\theta(a|s)$$

In terms of this notation, the policy gradient theorem can be reformulated as simply $\nabla V(\theta) = \hat{g}(\theta, Q^*_\theta)$. Further, we can rewrite the actor update as

$$\theta_{t+1} = \theta_t - \beta_t \left(\hat{g}(\theta_t, \Phi\omega_t) + w_{\rm a}(t) \right).$$

We summarize this discussion as the following proposition.

PROPOSITION 5.1 (Reformulation of actor-critic). *The single-sample actor-critic update can be written as*

$$\begin{aligned} \theta_{t+1} &= \theta_t - \beta_t \left(\hat{g}(\theta_t, \Phi\omega_t) + w_{\rm a}(t) \right) \\ \omega_{t+1} &= P_\Omega \left(\omega_t - \alpha_t \left(\widetilde{\nabla}(\theta_t, \omega_t) + w_{\rm c}(t) \right) \right), \end{aligned} \tag{5.2}$$

where if $\mathcal{F}_t$ is the σ-field generated by $w_{\rm a}(0), \ldots, w_{\rm a}(t-1), w_{\rm c}(0), \ldots, w_{\rm c}(t-1)$, then

$$E[w_{\rm a}(t)|\mathcal{F}_t] = E[w_{\rm c}(t)|\mathcal{F}_t] = 0,$$

and further that $w_{\rm a}(t), w_{\rm c}(t)$ are conditionally independent given $\mathcal{F}_t$. Moreover, for all θ, ω,

$$\nabla V(\theta) = \hat{g}(\theta, Q^*_\theta) \tag{5.3}$$

$$\widetilde{\nabla}(\theta,\omega)^T(\omega - \omega_\theta) \geq \frac{\mu}{2}||\omega - \omega_\theta||^2, \tag{5.4}$$

for some strictly positive μ.

We next introduce one more piece of notation. We will find it convenient to define

$$g(\theta,\omega) = \hat{g}(\theta,\Phi\omega), \tag{5.5}$$

so that the actor-update can be written simply as

$$\theta_{t+1} = \theta_t - \beta_t \left(g(\theta_t,\omega_t) + w_{\rm a}(t)\right).$$

Finally, we remark that by definition

$$g(\theta,\omega_\theta) - \hat{g}(\theta,Q_\theta^*) \;=\; \sum_k \gamma^k p_{k,\theta} \sum_a \pi_\theta(a|s)\left(\phi(s,a)^T\omega_\theta - Q_{\theta^*}\right)\frac{d}{d\theta}\log\pi_\theta(a|s)$$

and consequently we can use Assumption 5 to obtain that

$$\begin{aligned} ||g(\theta,\omega_\theta) - \hat{g}(\theta,Q_\theta^*)|| &\leq \sum_k \gamma^k p_{k,\theta}\sum_a \pi_\theta(a|s)\left|\phi(s,a)^T\omega_\theta - Q_{\theta^*}\right| K \\ &\leq \frac{K}{1-\gamma}E_{(s,a)\sim\nu_\theta}|\phi(s,a)^T\omega_\theta - Q_\theta^*(s,a)| \leq \frac{K}{1-\gamma}\delta \end{aligned} \tag{5.6}$$

With all these observations in place, we next turn to the analysis of the dynamics of Proposition 5.1. We have structured this analysis in the following manner:

1. Section 5.3 contains technical results on regularity of the underlying functions, updates, and parameters
2. Section 5.4 contains our convergence rate analysis and is divided to three parts:
   - Section 5.4.1 focuses on the critic update;
   - Section 5.4.2 focuses on the actor update;
   - Section 5.4.3 includes the construction of the nonlinear small-gain theorem that, combining our previous findings, yields our main proof.

5.3. First part of the proof: properties of the underlying functions, updates, and gradients. To analyze gradient descent one typically needs some assumptions on the underlying functions. These tend to involve continuity of various gradients and updates, boundedness and finite variance of noise, and so forth. As we have discussed in the previous section, actor-critic consists of two gradient-like updates; we will thus need to establish similar properties. We will do that in this section, and the main part of the proof itself, which will use these properties, will begin in the following section.

We first show that the the mean direction of the critic update (before projection) is Lipschitz. Formally, we have the following.

LEMMA 5.2 (Lipschitz-ness of critic update). *There exists a constant $L_\nabla < \infty$ such that for all θ,ω_1,ω_2,*

$$||\widetilde{\nabla}(\theta,\omega_1) - \widetilde{\nabla}(\theta,\omega_2)|| \leq L_\nabla||\omega_1-\omega_2||$$

Proof. Indeed, by definition of $\widetilde{\nabla}(\theta,\omega)$, we have that

$$\begin{aligned} ||\widetilde{\nabla}(\theta,\omega_1) - \widetilde{\nabla}(\theta,\omega_2)|| &= ||A_\theta\omega_1 + b_\theta - A_\theta\omega_2 - b_\theta|| \\ &\leq ||A_\theta||\cdot||\omega_1-\omega_2|| \\ &= \left|\left|E[\gamma\phi(s,a)\phi(s',a')^T - \phi(s,a)\phi(s,a)^T]\right|\right|\cdot||\omega_1-\omega_2||. \end{aligned}$$

Now because we assumed in Assumption 4 that each $\phi(s,a)$ satisfies $||\phi(s,a)|| \leq 1$, we have that

$$\left|\left|E[\gamma\phi(s,a)\phi(s',a')^T - \phi(s,a)\phi(s,a)^T]\right|\right| \leq 2.$$

Thus we can take $L_\nabla = 2$. ∎

Next, at some point we will want to use the fact that none of the actor or critic updates ever leaves some compact set; while a proof can be done without this assertion, it certainly simplifies some of the arguments. To that end, our next two lemmas demonstrate that, by construction, the noises $w_{\rm c}(t), w_{\rm a}(t)$ have compact support. This turns out to be an immediate consequence of the projection of the critic update onto the compact set Ω.

LEMMA 5.3 (Bounded support for critic noise). *The support of the random vector $w_{\rm c}(t)$ belongs to some compact set.*

Proof. By definition of $w_{\rm c}(t)$, we have that

$$w_{\rm c}(t) \quad = \quad -(A_t - E[A_t])\omega_t + (b_t - E[b_t])$$

So

$$\begin{aligned} ||w_{\rm c}||_2^2 &\leq 2||A_t - E[A_t]||^2||\omega_t||^2 + 2||b_t - E[b_t]||^2 \\ &\leq 2(2||A_t||^2 + 2||E[A_t]||^2)||\omega_t||^2 + 2(2||b_t||^2 + 2||E[b_t]||^2) \end{aligned}$$

Recall, however, that all costs are in $[-C_{\max}, C_{\max}]$ and that $||\phi(s,a)|| \leq 1$ for all state-action pairs (s,a). We can plug this into Eq. (3.2) to obtain

$$||w_{\rm c}(t)||_2^2 \leq 32||\omega_t||^2 + 8C_{\max}^2$$

But since $\omega_t \in \Omega$, and Ω is a compact set, we obtain that $w_{\rm c}(t)$ indeed has compact support. ∎

LEMMA 5.4 (Bounded support for actor noise). *The support of the random vector $w_{\rm a}(t)$ belongs to a compact set.*

Proof. Indeed, $w_{\rm a}(t)$ is defined as the difference between the sum on the right-hand side of Eq. (2.2) and a single term of it chosen at random so that its expectation is the entire sum. Inspecting Eq. (2.2), it follows that because ω_t lies in the compact set Ω and, by Assumption 3, $||\frac{d}{d\theta}\log\pi_\theta(a|s)|| \leq K$ for all θ, s, a, we have that $w_{\rm a}(t)$ is the difference of two quantities each of which has norm at most

$$\frac{K}{1-\gamma}||\Phi||\sup_{\omega\in\Omega}||\omega||.$$

Thus its support is bounded. ∎

Since we have shown that the vectors $w_{\rm a}(t), w_{\rm c}(t)$ have bounded support, we now introduce the notation

$$E\left[||w_{\rm a}(t)||_2^2 \mid \mathcal{F}_t\right] \leq \sigma_a^2, \quad E\left[||w_{\rm c}(t)||_2^2 \mid \mathcal{F}_t\right] \leq \sigma_c^2.$$

By the above lemmas, we have that $\sigma_{\rm a} < \infty$ and $\sigma_{\rm c} < \infty$. We will also define $\sigma_{\rm a'}^2$ through

$$\sqrt{E\left[||w_{\rm a}(t)||_2^4 \mid \mathcal{F}_t\right]} \leq \sigma_{\rm a'}^2$$

Again, this is well-defined as $w_{\rm a}(t)$ has compact support.

Next, we need to establish that the mean of the critic update is Lipschitz. This is done in the following lemma. We remind the reader that $g(\theta,\omega)$ is defined in Eq. (5.5).

LEMMA 5.5 (Lipschitz-ness of actor update). *For all $\theta, \omega_1, \omega_2$, we have that*

$$||g(\theta, \omega_1) - g(\theta, \omega_2)|| \leq L_g ||\omega_1 - \omega_2||,$$

for some $L_g < \infty$.

Proof. Indeed,

$$g(\theta, \omega_1) - g(\theta, \omega_2) = \sum_{s,t=0,\ldots} \gamma^t p_{t,\theta}(s) \sum_a \pi_\theta(s,a) \phi(s,a)^T (\omega_1 - \omega_2) \frac{d}{d\theta} \log \pi_\theta(a|s)$$

However, a convex combination of a collection of vectors has norm upper bounded by the largest of the norms of these vectors, so that

$$||g(\theta, \omega_1) - g(\theta, \omega_2)|| \leq \frac{K}{1-\gamma} ||\Phi|| \cdot ||\omega_1 - \omega_2||$$

∎

The following is a technical lemma which will be useful for us.

LEMMA 5.6. *Fix two vectors u and v and suppose*

$$q_\theta = u^T (I - \gamma P_\theta)^{-1} v.$$

Then $||\nabla_\theta q_\theta|| \leq L_q$, for some $L_q < \infty$ that does not depend on θ.

Proof. Using the well-known formula for derivative of the matrix inverse (see [30, Eq. (14.46)]),

$$\frac{\partial q_\theta}{\partial \theta_i} = u^T (I - \gamma P_\theta)^{-1} \left[-\gamma \frac{\partial P_\theta}{\partial \theta_i} \right] (I - \gamma P_\theta)^{-1} v \tag{5.7}$$

We next argue that the 2-norm of each term in the product on the right-hand side can be bounded independently of θ. Indeed, $||(I - \gamma P_\theta)x||_\infty \geq (1-\gamma)||x||_\infty$, as $||(I - \gamma P_\theta)^{-1}||_\infty$ is uniformly bounded over θ (and so is the 2-norm). As for the term in brackets, we have that by Assumption 3 the quantities $||\frac{d}{d\theta}\pi_\theta(a|s)||$ are bounded independently of θ, implying any norm one takes of the term in brackets is bounded independently of θ as well. ∎

Next, to derive any kind of guarantee on minimizing V, we need to make some assumptions on this function. We already know it is smooth from the policy gradient theorem, provided that Assumption 3 holds. It is standard to have a bound asserting that its gradient cannot change too quickly.

PROPOSITION 5.7 (Smooth objective). *There exists some $L_V < \infty$ such that the function V_θ has L_V-Lipschitz gradient.*

This follows from Theorem 3 of [23].

We next begin a sequence of lemmas and propositions whose ultimate goal is to show that the quantity ω_θ (the ultimate limit of the TD update when θ is fixed) is a twice differentiable function of θ with a uniform upper bound on its Hessian.

As a first step towards that, we will need to obtain upper bounds on the derivatives of the matrix A_θ (or rather, on the derivative of its inverse). Recall that matrix is obtained by sampling states s_t', a_t', s_t'', a_t'' from the distributed μ_θ (see Algorithm 2.1). Our first proposition shows this distribution changes in a Lipschitz manner as a function of θ. We translate this into a bound on the derivatives of the entries of the matrix A_θ.

LEMMA 5.8 (Uniform boundedness of derivatives and Hessians of A_θ). *There exist constants L'_A, L''_A, L'_b, L''_b not depending on θ such that, for all k, l,*

$$\left|\frac{\partial [A_\theta]_{ij}}{\partial \theta_k}\right| \leq L'_A, \quad \left|\frac{\partial^2 [A_\theta]_{ij}}{\partial \theta_k \theta_l}\right| \leq L''_A, \quad \left|\frac{\partial [b_\theta]_i}{\partial \theta_k}\right| \leq L'_b, \quad \left|\frac{\partial^2 [b_\theta]_i}{\partial \theta_k \theta_l}\right| \leq L'_b.$$

Proof. Let us use the notation $\rho_\theta(s,a)$ for the stationary distribution of state s and action a when following policy π_θ. We have that

$$[A_\theta]_{ij} = \sum_{s',a',s'',a''} \rho_\theta(s',a') P(s''|s',a') \pi_\theta(a''|s'') \phi(s',a') \left(\gamma \phi(s'',a'') - \phi(s',a')\right).$$

There are only two terms in this sum depending on θ. It is thus immediate that the theorem follows if we can bound the absolute values of the quantities

$$\pi_\theta(a|s), \frac{\partial \pi_\theta(a|s)}{\partial \theta_j}, \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta_j \partial \theta_k}$$
$$\rho_\theta(s,a), \frac{\partial \rho_\theta(a|s)}{\partial \theta_j}, \frac{\partial^2 \rho_\theta(a|s)}{\partial \theta_j \partial \theta_k}$$

independently of θ, for all i, j, k.

Of course, both $\pi_\theta(a|s)$ and $\rho_\theta(a|s)$ are upper bounded by one. Now for the derivatives and second derivatives of π_θ, boundedness follows by Assumption 3. Finally, for the stationary distribution ρ_θ, boundedness of all the quantities above follows from [13], specifically from the discussion in the beginning of Section 4 of that paper, when put together with the earlier Remark 6 in the same paper.

Note that here we need to use both our assumptions the uniform upper bound on the first two derivatives of $\pi_\theta(a|s)$ in Assumption 3 as well as the uniform mixing condition of Assumption 2, since both of these are used by [13] in their arguments.

The proof of the upper bound on the derivatives and second derivatives of the entries of b_θ is similar. ∎

We next establish a uniform bound on the how big A_θ^{-1} can get.

LEMMA 5.9 (Uniform boundedness of A_θ^{-1}).

$$\sup_\theta ||A_\theta^{-1}|| < \mu^{-1}.$$

Proof. We prove a more general claim: If $\lambda_{\max} \frac{1}{2}(A + A^T) \leq -\lambda \leq 0$, then $||A^{-1}|| \leq \lambda^{-1}$. By assumption, we have that $\frac{1}{2} x^T (A + A^T) x \leq -\lambda$, for any unit vector $x \in \mathbb{R}^n$. This implies that $x^T y \leq -\lambda$, where $y = Ax$ and x is of unit norm. On the other hand, since $\|x\| = 1$, we have that $-1 \leq \frac{x^T y}{\|y\|}$. Using the fact that $-x^T y \geq \lambda$, we conclude that $\|y\| \geq \lambda$; since $\lambda \geq 0$, we have that $x^T A A^T x \geq \lambda^2$, for any unit vector x. In particular,

$$\lambda_{\min}(AA^T) = \min_{\|x\| \neq 0, \|x\| = 1} x^T A A^T x \geq \lambda^2.$$

Since $\lambda_{\max}((AA^T)^{-1}) = \lambda_{\min}(AA^T)^{-1}$, we conclude that

$$\|A^{-1}\| = \sqrt{\lambda_{\max}((AA^T)^{-1})} \leq \frac{1}{\lambda}.$$

∎

Next, we state as a lemma the fact that the limit point of TD is Lipschitz. This fact is well-known, and is an immediate consequence of the standard upper bound

$$\left|\left|\frac{\partial A(\theta)^{-1}}{\partial \theta_j}\right|\right|_p \leq ||A(\theta)^{-1}||_p^2 \cdot \left|\left|\frac{\partial A(\theta)}{\partial \theta_j}\right|\right|_p \tag{5.8}$$

for the norm of a matrix inverse (see [30, Eq. (14.46)]). For a concrete reference, the next lemma is Proposition 4.4 of [31].

LEMMA 5.10 (Lipschitz continuity of the TD fixed point). *There is a constant L_ω such that*

$$||\omega_{\theta_1} - \omega_{\theta_2}|| \leq L_\omega ||\theta_1 - \theta_2||$$

It turns out we will need a more than this; specifically, we need to argue that the second of the TD limit is independent of θ. This is done in the following lemma.

LEMMA 5.11 (Bounded curvature of the TD fixed point). *There is some quantity λ_i independent of θ such that*

$$\sup_\theta \lambda_{\max}(\nabla^2 w_\theta(i)) \leq \lambda_i.$$

Proof. We differentiate twice the equation $\omega_\theta = A_\theta^{-1} b_\theta$ to get

$$\frac{\partial^2 w_\theta(i)}{\partial \theta_j \partial \theta_l} = \sum_k \frac{\partial^2 [A(\theta)^{-1}]_{ik}}{\partial \theta_j \partial \theta_l} b_k(\theta) + \frac{\partial [A(\theta)^{-1}]_{ik}}{\partial \theta_j} \frac{\partial b_k(\theta)}{\partial \theta_l} + \frac{\partial [A(\theta)^{-1}]_{ik}}{\partial \theta_l} \frac{\partial b_k(\theta)}{\partial \theta_j} + [A(\theta)^{-1}]_{ik} \frac{\partial^2 b_k(\theta)}{\partial^2 \theta_j}$$

We need to upper bound the right-hand side independently of θ. Here we can apply Lemma 5.9 and Lemma 5.8 and Eq. (5.8). We see that the only term not covered by these three sources is $\frac{\partial^2 [A(\theta)^{-1}]_{ik}}{\partial \theta_j \partial \theta_l}$: all the other terms have already been bounded independently of θ.

We now turn to analyzing that term. Taking $A_\theta A_\theta^{-1} = I$, differentiating twice and rearranging we obtain:

$$\frac{\partial^2 A_\theta^{-1}}{\partial \theta_j \partial \theta_l} = -A_\theta^{-1} \left(\frac{\partial^2 A_\theta}{\partial \theta_j \partial \theta_l} A_\theta^{-1} + \frac{\partial A_\theta}{\partial \theta_j} \frac{\partial A_\theta^{-1}}{\partial \theta_l} + \frac{\partial A_\theta}{\partial \theta_l} \frac{\partial A_\theta^{-1}}{\partial \theta_j} \right)$$

∎

The norms of all the terms on the right-hand side are bounded independently of θ as a consequence of Lemmas 5.9 and Lemma 5.8 and Eq. (5.8), and thus we are done.

For convenience, we define

$$\lambda = \sqrt{\sum_i \lambda_i^2}. \tag{5.9}$$

5.4. Second part of the proof: small-gain analysis. We now turn to the main body of the proof itself. Having established that various quantities appearing in our updates are Lipschitz and/or bounded, we will now consider how the critic and actor errors relate to each other. As before, our first step is to introduce (yet more) notation.

We will find it convenient to define

$$\theta_{t+1/2} = \theta_t - \beta_t g(\theta_t, \omega_t),$$

so that $\theta_{t+1/2}$ is a deterministic function of θ_t, ω_t. The next actor iterate θ_{t+1} is then obtained from $\theta_{t+1/2}$ by adding noise:

$$\theta_{t+1} = \theta_{t+1/2} - \beta_t w_{\rm a}(t). \tag{5.10}$$

Let us observe a couple of consequence of these equations. First, we can apply Lemma 5.10 have that

$$||\omega_{\theta_{t+1/2}} - \omega_{\theta_t}||^2 \quad \leq \quad L_\omega^2 ||\theta_{t+1/2} - \theta_t||^2 = L_\omega^2 \beta_t^2 ||g(\theta_t, \omega_t)||^2 \tag{5.11}$$

Similarly, using Lemma 5.10 we have that

$$E||\omega_{\theta_{t+1}} - \omega_{\theta_{t+1/2}}||^2 \quad \leq \quad E L_\omega^2 ||\theta_{t+1} - \theta_{t+1/2}||^2 \leq L_\omega^2 \beta_t^2 \sigma_a^2 \tag{5.12}$$

5.4.1. Analysis of the Critic Update. With the above notation and preliminaries in place, we now turn to the proof itself. Our first step is to obtain a performance error bound on the critic. Of course, this cannot be done in isolation of what happens in the actor. The final bound we will derive in this subsection will bound the critic's performance in terms of the closeness to stationary of the actor.

Our first step is to argue that, without noise, a small enough step starting from ω_t and moving in the direction of $\widetilde{\nabla}(\theta_t, \omega_t)$ reduces the distance to ω_{θ_t}, which, recall, is the ultimate limit of the TD(0) iteration were θ_t to be left fixed. This is stated formally, along with an estimate of the size of the reduction, in the following lemma.

LEMMA 5.12 (Contractivity of the TD part of the update). *If* $\alpha_t \leq \mu/(2L_\nabla^2)$, *then*

$$||\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t)|| \leq (1 - \alpha_t \mu/4)||\Delta_t||.$$

Proof. Indeed,

$$\begin{aligned}
||\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t)||^2 &= ||\omega_t - \omega_{\theta_t}||^2 - 2\alpha_t \widetilde{\nabla}(\theta_t, \omega_t)^T (\omega_t - \omega_{\theta_t}) + \alpha_t^2 ||\widetilde{\nabla}(\theta_t, \omega_t)||^2 \\
&\leq ||\omega_t - \omega_{\theta_t}||_2^2 - 2\alpha_t \frac{\mu}{2} ||\omega_t - \omega_{\theta_t}||_2^2 + \alpha_t^2 ||\widetilde{\nabla}(\theta_t, \omega_t) - \widetilde{\nabla}(\theta_t, \omega_{\theta_t})||^2 \\
&\leq (1 - \alpha_t \mu + L_\nabla^2 \alpha_t^2) ||\omega_t - \omega_{\theta_t}||^2 \\
&\leq (1 - \alpha_t \mu/2) ||\omega_t - \omega_{\theta_t}||_2^2.
\end{aligned}$$

where the first step is just expansion of the quadratic; the second step uses Eq. (5.1) and Eq. (5.4); the third step uses Lemma 5.2; and the final step uses the inequality $\alpha_t \leq \mu/(2L_\nabla^2)$ to get that $L_\nabla^2 \alpha_t^2 \leq \alpha_t \mu/2$. Taking square roots of both sides completes the proof. ∎

Recursion relation for the critic. With this lemma in place, let us attempt to derive a recursion relation that can be used to argue that ω_t tracks ω_{θ_t}, i.e., let us prove an upper bound on the difference between these two quantities.

Indeed, because

$$\omega_{t+1} - \omega_{\theta_{t+1}} \quad = \quad P_\Omega \left(\omega_t - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_{\rm c}(t) \right) - \omega_{\theta_{t+1}}$$

by the non-expansiveness of projection as well by Assumption 7 (which states that $\omega_\theta \in \Omega$ for all θ), we have that

$$||\omega_{t+1} - \omega_{\theta_{t+1}}||^2 \leq ||\omega_t - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_{\rm c}(t) - \omega_{\theta_{t+1}}||^2$$

We will then rewrite this as

$$||\omega_{t+1} - \omega_{\theta_{t+1}}||^2 = \left|\left|\left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right) + \left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right) + \left(\omega_{\theta_t} - \omega_{\theta_{t+1/2}}\right)\right|\right|^2$$

We now take the (conditional) squared expectation of both sides to obtain

$$\begin{aligned} E\left[||\omega_{t+1} - \omega_{\theta_{t+1}}||^2 \mid \mathcal{F}_t\right] &= E\left[\left|\left|\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right|\right|^2 \mid \mathcal{F}_t\right] \\ &+ E\left[\left|\left|\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right) + \left(\omega_{\theta_t} - \omega_{\theta_{t+1/2}}\right)\right|\right|^2 \mid \mathcal{F}_t\right] \\ &+ E\left[\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}} + \omega_{\theta_t} - \omega_{\theta_{t+1/2}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}_t - \alpha_t w_c(t)\right) | \mathcal{F}_t\right] \end{aligned}$$

where, recall, $\mathcal{F}_t$ was defined in Proposition 5.1; informally, it is the history of the process up to time t. We then upper bound this as

$$\begin{aligned} E\left[||\omega_{t+1} - \omega_{\theta_{t+1}}|| \mid \mathcal{F}_t\right]^2 &= (1 - \alpha_t\mu/4)||\omega_t - \omega_{\theta_t}||^2 + \alpha_t^2\sigma_c^2 \\ &+ 2L_\omega^2\beta_t^2\sigma_a^2 + 2L_\omega^2\beta_t^2||g(\theta_t, \omega_t)||_2^2 \\ &+ E\left[\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right) \mid \mathcal{F}_t\right] \\ &+ E\left[\left(\omega_{\theta_t} - \omega_{\theta_{t+1/2}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right) \mid \mathcal{F}_t\right] \end{aligned}$$

where in the first line we used Lemma 5.12, as well as that, conditional on $\mathcal{F}_t$, $w_c(t)$ has zero mean; and in the second line, we used Eq. (5.11) and Eq. (5.12) as well as the inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

We next turn our attention to the fourth line of the above equation. Using Cauchy-Schwarz, Eq. (5.11), and the conditional zero-mean of $w_c(t)$, we obtain

$$\begin{aligned} E\left[||\omega_{t+1} - \omega_{\theta_{t+1}}|| \mid \mathcal{F}_t\right]^2 &\leq (1 - \alpha_t\mu/4)||\omega_t - \omega_{\theta_t}||^2 + \alpha_t^2\sigma_c^2 \\ &+ 2L_\omega^2\beta_t^2\sigma_a^2 + 2L_\omega^2\beta_t^2||g(\theta_t, \omega_t)||_2^2 \\ &+ E\left[\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right) \mid \mathcal{F}_t\right] \\ &+ \beta_t L_\omega||g(\theta_t, \omega_t)|| \;\; ||\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t)||. \end{aligned}$$

Finally, if α_t is small enough to satisfy the conditions of Lemma 5.12, we can modify the last line of this equation as

$$\begin{aligned} E\left[||\omega_{t+1} - \omega_{\theta_{t+1}}|| \mid \mathcal{F}_t\right]^2 &\leq (1 - \alpha_t\mu/4)||\omega_t - \omega_{\theta_t}||^2 + \alpha_t^2\sigma_c^2 \\ &+ 2L_\omega^2\beta_t^2\sigma_a^2 + 2L_\omega^2\beta_t^2||g(\theta_t, \omega_t)||_2^2 \\ &+ E\left[\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t) - \alpha_t w_c(t)\right) \mid \mathcal{F}_t\right] \\ &+ \beta_t L_\omega||g(\theta_t, \omega_t)|| \;\; ||\Delta_t|| \end{aligned} \tag{5.13}$$

Next, we need to obtain an estimate of the size of the inner product in the third line. This is done in our next lemma.

LEMMA 5.13. *Suppose $\alpha_t \leq \mu/(2L_\nabla^2)$. Then*

$$E\left[\left(\omega_{\theta_{t+1/2}} - \omega_{\theta_{t+1}}\right)^T \left(\omega_t - \omega_{\theta_t} - \alpha_t \widetilde{\nabla}(\theta_t, \omega_t)\right) \mid \mathcal{F}_t\right] \leq \beta_t^2 \frac{\lambda}{2}\sigma_{\grave{a}'}^2||\Delta_t||$$

Proof. Let us define I_t to be the left-hand side of the above equation. Consider the quantity $\omega_\theta(i)$, i.e., the i'th entry of the vector ω_θ. Recall that by Lemma 5.11,

this is a twice continuously differentiable function of θ with an available upper bound of λ_i on the norm of the Hessian at any θ.

We can thus use a second order expansion to write

$$\omega_{\theta_{t+1}}(i) = \omega_{\theta_{t+1/2}}(i) + \nabla\omega_{\theta_{t+1/2}}(i)^T(\beta_t w_{\rm a}(t)) + \frac{1}{2}\beta_t^2 w_{\rm a}(t)^T\nabla^2\omega_{\theta_i'}(i)w_{\rm a}(t),$$

where the gradient and Hessian are taking with respect to θ and θ_i' is some point on the line segment connecting $\theta_{t+1/2}$ and θ_{t+1}. Consequently,

$$\begin{aligned} I_t &= E\left[\sum_i \left(\beta_t\nabla\omega_{\theta_{t+1/2}}(i)^T w_{\rm a}(t) - \beta_t^2 w_{\rm a}(t)^T\frac{1}{2}\nabla^2\omega_{\theta_i'}(i)w_{\rm a}(t)\right)\right. \\ &\qquad \left.\left(\omega_t(s,a) - \omega_{\theta_t}(s,a) - \alpha_t\widetilde{\nabla}(\theta_t,\omega_t)(i)\right) \mid \mathcal{F}_t\right], \end{aligned}$$

where note that we are abusing notation somewhat, as the θ' is actually different in each term of the sum.

Since $E[w_{\rm a}(t)|\mathcal{F}_t] = 0$ and all the quantities $w_{\rm a}(t)$ multiplies above are deterministic conditional on $\mathcal{F}_t$, we obtain that

$$I_t = \beta_t^2 E\left[\sum_i \left(-w_{\rm a}(t)^T\frac{1}{2}\nabla^2\omega_{\theta_i'}(i)w_{\rm a}(t)\right)\left(\omega_t(i) - \omega_{\theta_t}(i) - \alpha_t\widetilde{\nabla}(\theta_t,\omega_t)(i)\right) \mid \mathcal{F}_t\right].$$

We apply Cauchy-Schwarz and concavity of square root to obtain

$$I_t \le \beta_t^2\sqrt{E\left[\sum_i\left(w_{\rm a}(t)^T\frac{1}{2}\nabla^2\omega_{\theta_i'}(i)w_{\rm a}(t)\right)^2 \mid \mathcal{F}_t\right]}\sqrt{\sum_i(\omega_t(i) - \omega_{\theta_t}(i) - \alpha_t\widetilde{\nabla}(\theta_t,\omega_t)(i))^2}$$

By Lemma 5.11, we have that $w_{\rm a}(t)\nabla^2\omega_{\theta_i'}(i)w_{\rm a}(t) \le \lambda_i||w_a(t)||^2$, and plugging this in above and using Lemma 5.12 completes the proof. ∎

Having obtained this lemma, we can now go back to Eq. (5.13) and use it to bound the third line. This argument leads to the following lemma, whose punchline is that an average of the quantities $||\Delta_t||^2$ satisfies an upper bound which we will later show to be quite small.

LEMMA 5.14 (Bound on the critic error). *Suppose that $\alpha_t \in (0,1), \beta_t$ are non-increasing sequences satisfying $\alpha_{t/2} \le c_\alpha\alpha_t$ and $\beta_{t/2} \le c_\beta\beta_t$ for all t and for some constants c_α, c_β. Suppose further that for all $t \ge T/2$ we have that $24c_\beta\beta_t L_\omega \le 1$ and $\alpha_t \le \mu/(2L_\nabla^2)$ and*

$$6L_\omega^2\frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}L_g^2 + \frac{c_\beta\beta_T}{\alpha_T}\frac{2L_\omega}{\mu} + \frac{c_\beta\beta_T}{\alpha_T}\frac{4}{\mu}L_\omega L_g \le \frac{1}{2}. \tag{5.14}$$

Then

$$\begin{aligned} \frac{1}{T/2}\sum_{t=T/2}^{T-1} E\left[||\Delta_t||\right]^2 \quad \le \quad & \frac{1}{\alpha_T}\frac{16}{\mu T}||\omega_{T/2} - \omega_{\theta_{T/2}}||^2 + c_\alpha^2\alpha_T\frac{8}{\mu}\sigma_c^2 \\ & + 2L_\omega^2\frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}\sigma_a^2 + 6L_\omega^2\frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}\delta^2 \\ & + 8\frac{c_\beta^2\beta_T^2\sigma_{\rm a'}^2(\lambda/2) + c_\beta\beta_T L_\omega\delta}{\alpha_T\mu}\sqrt{\frac{1}{T/2}\sum_{t=T/2}^{T-1}E||\Delta_t||^2} \\ & + \frac{c_\beta\beta_T}{\alpha_T}\frac{8L_\omega}{\mu}\frac{1}{T/2}\sum_{t=T/2}^{T-1}E||\nabla_t||^2 \end{aligned}$$

Proof. As mentioned above, our first step is to plug Lemma 5.13 into Eq. (5.13). We thus obtain:

$$\begin{aligned} E\left[||\omega_{t+1}-\omega_{\theta_{t+1}}|| \mid \mathcal{F}_t\right]^2 &= (1-\alpha_t\mu/4)||\omega_t-\omega_{\theta_t}||^2+\alpha_t^2\sigma_c^2 \\ &\quad +2L_\omega^2\beta_t^2\sigma_a^2+2L_\omega^2\beta_t^2||g(\theta_t,\omega_t)||^2 \\ &\quad +\beta_t^2\frac{\lambda}{2}\sigma_{a'}^2||\Delta_t|| \\ &\quad +\beta_t L_\omega||g(\theta_t,\omega_t)||\ ||\Delta_t||. \end{aligned} \tag{5.15}$$

We next bound the norm $||g(\theta_t,\omega_t)||$ which appears twice in the above equation. Indeed, we have that

$$\begin{aligned} ||g(\theta_t,\omega_t)|| &= ||g(\theta_t,\omega_{\theta_t})+g(\theta_t,\omega_t)-g(\theta_t,\omega_{\theta_t})|| \\ &\leq ||g(\theta_t,\omega_{\theta_t})||+L_g||\omega_t-\omega_{\theta_t}|| \\ &= ||\hat{g}(\theta_t,Q_{\theta_t}^*)+g(\theta_t,\omega_{\theta_t})-\hat{g}(\theta_t,Q_{\theta_t}^*)||+L_g||\Delta_t|| \\ &\leq ||\nabla_t||+\bar{\delta}+L_g||\Delta_t|| \end{aligned}$$

Here the first line simply adds and subtracts the same quantity; the second line uses Lemma 5.5; the third line adds and subtracts the same quantity; the fourth line uses Eq. (5.6) and we define

$$\bar{\delta}=\frac{K}{1-\gamma}\delta.$$

Finally, the last line simply uses our notation $\nabla_t=\nabla V(\theta_t)$.

Next, using the inequality $(a+b)^2\leq 3(a^2+b^2+c^2)$ we thus have

$$||g(\theta_t,\omega_t)||^2\leq 3||\nabla_t||^2+3\bar{\delta}^2+3L_g^2||\Delta_t||^2.$$

Plugging this into Eq. (5.15), we obtain

$$\begin{aligned} E\left[||\omega_{t+1}-\omega_{\theta_{t+1}}|| \mid \mathcal{F}_t\right]^2 &= (1-\alpha_t\mu/4)||\omega_t-\omega_{\theta_t}||^2+\alpha_t^2\sigma_c^2 \\ &\quad +2L_\omega^2\beta_t^2\sigma_a^2+6L_\omega^2\beta_t^2||\nabla_t||^2+6L_\omega^2\beta_t^2\bar{\delta}^2+6L_\omega^2\beta_t^2L_g^2||\Delta_t||^2 \\ &\quad +\beta_t^2\frac{\lambda}{2}\sigma_{a'}^2||\Delta_t|| \\ &\quad +\beta_t L_\omega\left(||\nabla_t||+\bar{\delta}+L_g||\Delta_t||\right)||\Delta_t|| \end{aligned} \tag{5.16}$$

Using the inequality $ab\leq(1/2)(a^2+b^2)$, taking expectations, and combining both terms containing $||\Delta_t||$, we obtain

$$\begin{aligned} E\left[||\omega_{t+1}-\omega_{\theta_{t+1}}||\right]^2 &= (1-\alpha_t\mu/4)E\left[||\omega_t-\omega_{\theta_t}||^2\right]+\alpha_t^2\sigma_c^2 \\ &\quad +2L_\omega^2\beta_t^2\sigma_a^2+6L_\omega^2\beta_t^2E\left[||\nabla_t||^2\right]+6L_\omega^2\beta_t^2\bar{\delta}^2+6L_\omega^2\beta_t^2L_g^2E\left[||\Delta_t||^2\right] \\ &\quad +\left(\beta_t^2\frac{\lambda}{2}\sigma_{a'}^2+\beta_tL_\omega\bar{\delta}\right)E\left[||\Delta_t||\right] \\ &\quad +\beta_tL_\omega\frac{E\left[||\Delta_t||^2\right]+E\left[||\nabla_t||^2\right]}{2}+\beta_tL_\omega L_gE\left[||\Delta_t||^2\right] \end{aligned} \tag{5.17}$$

Our next step is to sum this up over $t=T/2,\ldots,T-1$. To do this without excessive calculation we will use the following rough bound. Observe that if we have the recursion relation

$$y_{t+1}\leq(1-\zeta)y_t+e_t$$

for some $\zeta \in (0,1)$, then we have the rough bound

$$\sum_{t=a}^{b} y_t \leq \frac{y_a}{\zeta} + \sum_{t=a}^{b} \frac{e_t}{\zeta}.$$

Let us apply this to Eq. (5.17) with $\zeta = \alpha_T \mu/4$. We can do this because the sequence α_t is assumed to be in $(0,1)$ and nonincreasing and $\alpha_t \mu \leq 1$ since $\alpha_t \leq \mu/(2L_\nabla^2)$ by assumption. Whenever the bound α_t/α_T appears, we can upper bound it by c_α, and likewise with β_t/β_T. Proceeding this way and applying Cauchy-Schwarz and Jensen to the fourth line, we obtain

$$\begin{aligned}
\sum_{t=T/2}^{T-1} E\left[||\omega_t - \omega_{\theta_t}||\right]^2 \quad \leq \quad & \frac{1}{\alpha_T}\frac{4}{\mu}||\omega_{T/2} - \omega_{\theta_{T/2}}||^2 + c_\alpha^2 \alpha_T \frac{T}{2}\frac{4}{\mu}\sigma_c^2 \\
& + 2L_\omega^2 \frac{c_\beta^2 \beta_T^2}{\alpha_T}\frac{T}{2}\frac{4}{\mu}\sigma_a^2 + 6L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}\sum_{t=T/2}^{T-1} E||\nabla_t||^2 + 6L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{T}{2}\frac{4}{\mu}\bar{\delta}^2 \\
& + 6L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}L_g^2 \sum_{t=T/2}^{T-1} E||\Delta_t||^2 \\
& + 4\frac{c_\beta^2\beta_T^2\sigma_{a'}^2(\lambda/2) + c_\beta\beta_T L_\omega \bar{\delta}}{\alpha_T \mu}\sqrt{T/2}\sqrt{E\sum_{t=T/2}^{T-1}||\Delta_t||^2} \\
& + \frac{c_\beta\beta_T}{\alpha_T}\frac{2L_\omega}{\mu}\sum_{t=T/2}^{T-1} E||\nabla_t||^2 + \frac{c_\beta\beta_T}{\alpha_T}\frac{2L_\omega}{\mu}\sum_{t=T/2}^{T-1} E||\Delta_t||^2 \\
& + \frac{c_\beta\beta_T}{\alpha_T}\frac{4}{\mu}L_g L_\omega \sum_{t=T/2}^{T-1} E||\Delta_t||_2^2
\end{aligned}$$

Let us observe that the coefficient of $\sum_{t=T/2}^{T-1}||\Delta_t||^2$ on the right-hand side is

$$6L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{8}{\mu}L_g^2 + \frac{c_\beta\beta_T}{\alpha_T}\frac{2L_\omega}{\mu} + \frac{c_\beta\beta_T}{\alpha_T}\frac{4}{\mu}L_\omega L_g \leq \frac{1}{2},$$

where the inequality is by assumption in Eq. (5.14). We thus have

$$\begin{aligned}
\frac{1}{2}\sum_{t=T/2}^{T-1} E\left[||\omega_t - \omega_{\theta_t}||\right]^2 \quad \leq \quad & \frac{1}{\alpha_T}\frac{4}{\mu}||\omega_{T/2} - \omega_{\theta_{T/2}}||^2 + c_\alpha^2 \alpha_T \frac{T}{2}\frac{4}{\mu}\sigma_c^2 \\
& + 2L_\omega^2 \frac{c_\beta^2 \beta_T^2}{\alpha_T}\frac{T}{2}\frac{4}{\mu}\sigma_a^2 + 6L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{T}{2}\frac{4}{\mu}\bar{\delta}^2 \\
& + 4\frac{c_\beta^2\beta_T^2\sigma_{a'}^2(\lambda/2) + c_\beta\beta_T L_\omega \bar{\delta}}{\alpha_T \mu}\sqrt{T/2}\sqrt{E\sum_{t=T/2}^{T-1}||\Delta_t||^2} \\
& + \frac{c_\beta\beta_T}{\alpha_T}\frac{4L_\omega}{\mu}\sum_{t=T/2}^{T-1} E||\nabla_t||^2,
\end{aligned}$$

where we also combined two different terms that multiply $\sum_{t=T/2}^{T-1} E||\nabla_t||^2$ using the assumption that $24 c_\beta \beta_T L_\omega \leq 1$. Finally, dividing by $T/4$ we obtain the statement of the lemma.

5.4.2. Analysis of the actor update. We have just given an analysis of the performance of the critic, and that analysis had an error term corresponding to the actor stationarity, i.e., to the quantities $||\nabla_t||$. We now give an analysis of the reverse, i.e., a performance bound on the actor in terms of the critic error.

LEMMA 5.15 (Bound on the actor error). *Suppose $\beta_t \leq 1/(6L_V)$ for all $t \geq T/2$. Then*

$$\frac{1}{T/2}\sum_{t=T/2}^{T-1} E||\nabla_t||^2 \leq 8\left(\frac{V(\theta_{T/2})-V(\theta_T)}{\beta_T T} + c_\beta L_g^2 \frac{1}{T/2}\sum_{t=T/2}^{T-1} E||\Delta_t||^2 + c_\beta \bar{\delta}^2 + c_\beta^2 \beta_T \frac{3}{4} L_V \sigma_{\rm a}^2\right)$$

Proof. We will attempt to analyze the actor update as an inexact gradient descent. Our first step is to upper bound the difference between the expected actor update direction $g(\theta_t, \omega_t)$ and the true gradient $\nabla V(\theta_t) = \hat{g}(\theta_t, Q^*_{\theta_t})$. To do this, we argue as

$$\begin{aligned} ||g(\theta_t,\omega_t) - \hat{g}(\theta_t, Q^*_{\theta_t})|| &\leq ||g(\theta_t,\omega_t) - g(\theta_t,\omega_{\theta_t})|| + ||g(\theta_t,\omega_{\theta_t}) - \hat{g}(\theta_t, Q^*_{\theta_t})|| \\ &\leq L_g||\Delta_t|| + \bar{\delta}, \end{aligned}$$

where the bound on the first term comes from Lemma 5.5 while the bound on the second term comes from Eq. (5.6) where, recall that $\bar{\delta} = (K/(1-\gamma))\delta$. Consequently, we can write the actor update as

$$\theta_{t+1} = \theta_t - \beta_t \nabla V(\theta_t) - \beta_t d_t - \beta_t w_{\rm a}(t), \tag{5.18}$$

where

$$||d_t|| \leq L_g||\Delta_t|| + \bar{\delta}, \tag{5.19}$$

By Proposition 5.7, the function $V(\theta)$ has L-Lipschitz gradient. By the "descent lemma" (e.g., Theorem 4.22 in [4]) it satisfies

$$V(\theta_{t+1}) \leq V(\theta_t) + \nabla V(\theta_t)^T(\theta_{t+1}-\theta_t) + \frac{L_V}{2}||\theta_{t+1}-\theta_t||^2.$$

Plugging this into Eq. (5.18), taking expectations, and using the inequality

$$(\sum_{i=1}^k |a_i|)^2 \leq k\sum_{i=1}^k a_i^2,$$

we obtain

$$\begin{aligned} EV(\theta_{t+1}) &\leq EV(\theta_t) - (\beta_t/2)E||\nabla V(\theta_t)||^2 + \frac{\beta_t}{2}E||d_t||^2 \\ &\quad +3\frac{L_V\beta_t^2}{2}\left(E||\nabla V(\theta_t)||^2 + E||d_t||^2 + \sigma_a^2\right) \end{aligned}$$

Now as a consequence of our assumptions, we have that $(3/2)L_V\beta_t^2 \leq \beta_t/4$ so that we have

$$EV(\theta_{t+1}) \leq EV(\theta_t) - (\beta_t/4)E||\nabla V(\theta_t)||^2 + \beta_t E||d_t||^2 + \frac{3}{2}L_V\beta_t^2\sigma_{\rm a}^2.$$

Now Eq. (5.19) allows us to rewrite this as

$$EV(\theta_{t+1}) \leq EV(\theta_t) - (\beta_t/4)E||\nabla_t||_2^2 + 2\beta_t L_g^2 E||\Delta_t||^2 + 2\beta_t\bar{\delta}^2 + \frac{3}{2}L_V\beta_t^2\sigma_{\rm a}^2.$$

We re-arrange this as

$$(\beta_t/4)E||\nabla_t||_2^2 \leq EV(\theta_t) - EV(\theta_{t+1}) + 2\beta_t L_g^2 E||\Delta_t||^2 + 2\beta_t\bar{\delta}^2 + \frac{3}{2}L_V\beta_t^2\sigma_{\rm a}^2.$$

Summing this over $t = T/2, \ldots, T-1$ we obtain

$$\sum_{t=T/2}^{T-1}\frac{\beta_t}{4}E||\nabla_t||^2 \leq EV(\theta_{T/2}) - EV(\theta_T) + 2L_g^2\sum_{t=T/2}^{T-1}\beta_t E||\Delta_t||^2 + 2\bar{\delta}^2\sum_{t=T/2}^{T-1}\beta_t + \frac{3}{2}L_V\sigma_{\rm a}^2\sum_{t=T/2}^{T-1}\beta_t^2$$

Dividing by $(\beta_T/4)(T/2)$ concludes the proof. ∎

5.4.3. A nonlinear small-gain theorem. We have just finished deriving two relations, one analyzing how the actor error affects the critic error (Lemma 5.14), and the other one analyzing how critic error affects actor error (Lemma 5.15). We now need to put these two bounds together. This is done using the following nonlinear version of the small gain theorem.

LEMMA 5.16 (Small-gain bound). *Suppose $x_1, x_2, \ldots$ and $z_1, z_2, \ldots$, are sequences of vectors and $||\cdot||_T$ is a semi-norm on the first T element of a vector sequence. If there exist nonnegative numbers a, b, c, d, e such that*

$$||x||_T \leq a + b\sqrt{||x||_T} + c||z||_T, \tag{5.20}$$

$$||z||_T \leq d + e||x||_T \tag{5.21}$$

and $2ce < 1$, then for all T,

$$||x||_T \leq \frac{2a + b^2 + 2cd}{1 - 2ce}$$

Note that this differs from the usual small-gain theorem due to the square root in the first equation. This difference is what forces us to put an extra factor of two into the gain bound. In words, what this lemma says is that the introduction of a square root into one of the small gain bounds adds a term to the final bound which is quadratic in the variable b multiplying the square root. The proof is given next.

Proof. Let us substitute $||x||_T = y^2$ where y is also nonnegative. Then

$$y^2 \leq a + by + c||z||_T$$

or

$$y^2 - by - (a + c||z||_T) \leq 0$$

The set where a quadratic with a leading coefficient of 1 is nonpositive is always the interval between its roots. This we can upper bound y by the largest root of this quadratic:

$$y \leq \frac{b + \sqrt{b^2 + 4(a + c||z||_T)}}{2}.$$

Squaring both sides and using the inequality $(u+v)^2 \leq 2(u^2+v^2)$,

$$y^2 \leq \frac{2b^2 + 2(b^2 + 4(a + c||z||_T))}{4},$$

and simplifying and using $||x||_T = y^2$, we obtain.

$$||x||_T \leq 2a + 2c||z||_T + b^2. \tag{5.22}$$

We now plug Eq. (5.21) into this to obtain that

$$||x||_T \leq 2a + b^2 + 2c(d + e||x||_T),$$

so that

$$||x||_T \leq \frac{2a + b^2 + 2cd}{1 - 2ce}$$

□

We now use this nonlinear small-gain theorem we have derived to put together our bounds on actor and critic. Specifically, we will simply combine Lemma 5.14 giving us a performance bound on the critic with Lemma 5.15 on the actor. Inspecting these two lemmas, we see that their form is exactly the form one needs to apply the small gain theorem above.

Proof of Theorem 4.1 (our main result). To apply the small gain theorem just derived, we will plug in $x_t = \Delta_t$ and $y_t = \nabla_t$. Naturally, the semi-norm we have derived will be $||u||_T = \frac{1}{T/2}\sum_{t=T/2}^{T-1} u_t^2$. Inspecting Lemma 5.14 and Lemma 5.15, we can pattern match the quantities a, b, c, d, e to the quantities from those equations, and obtain

$$\begin{aligned}\frac{1}{T/2}\sum_{t=T/2}^{T-1} E\left[||\Delta_t||\right]^2 \quad \leq \quad & \frac{1}{\alpha_T}\frac{64}{\mu T}||\omega_{T/2} - \omega_{\theta_{T/2}}||^2 + c_\alpha^2 \alpha_T \frac{32}{\mu}\sigma_c^2 \\ & +4L_\omega^2 \frac{c_\beta^2 \beta_T^2}{\alpha_T}\frac{16}{\mu}\sigma_a^2 + 12L_\omega^2 \frac{c_\beta^2\beta_T^2}{\alpha_T}\frac{16}{\mu}\delta^2 \\ & +2\frac{2c_\beta^4\beta_T^4\sigma_{a'}^4(\lambda/2)^2 + 2c_\beta^2\beta_T^2 L_\omega^2\delta^2}{\alpha_T^2\mu^2} \\ & +\frac{c_\beta\beta_T}{\alpha_T}\frac{16L_\omega}{\mu}16\left(\frac{V(\theta_{T/2}) - V(\theta_T)}{\beta_T T} + c_\beta\bar{\delta}^2 + c_\beta^2\beta_T\frac{3}{4}L_V\sigma_a^2\right)\end{aligned} \tag{5.23}$$

□

For this to be a valid step, we need to make sure the gain condition of $1-2ec > 1/2$ is satisfied. For that, we need to choose $\beta_T = c'\alpha_T$ for a small enough c'. Then the small gain condition $2ec < 1/2$ is equivalent to $c' \leq \frac{\mu}{8L_\omega}\frac{1}{4c_\beta^2 L_g^2}$.

We can now do the final step-size selection step. First, we can choose $\alpha_t = 1/\sqrt{t}$ and $\beta_t = c'/\sqrt{t}$, where c' will be chosen to make this step-size satisfy all the assumptions we have imposed throughout. Let us now recall what those are.

First, we have assumed $\alpha_t \leq \mu/(2L_\nabla^2)$ and $\beta_t \leq 1/(6L_V)$ and $24c_\beta\beta_t L_\omega \leq 1$. However, these need to hold for $t \geq T/2$. Thus any step-size choices α_t, β_t which decrease to zero will eventually satisfy these.

However, we have two equations that have imposed relationships between the step-sizes α_t and β_t: these are $c' \leq \frac{\mu}{8L_\omega}\frac{1}{4c_\beta^2 L_g^2}$. and Eq. (5.14). Considering Eq. (5.14), for large enough T, the first term of that equation will be negligible compared to the second and third, and so to satisfy it we need to choose $c' = O\left(\max\left(\frac{\mu}{c_\beta L_\omega}, \frac{\mu}{L_\omega L_g c_\beta}\right)\right)$. We conclude there is a particular choice of c depending on $c_\beta, \mu, L, L_\omega, L_g$ that makes all of the step-size assumptions satisfied. Further, treating all variables except μ as constant, we have that $c' = O(\mu)$.

Let us next examine Eq. (5.23) to see what convergence rate we can obtain in that scenario. We immediately see that every term except for the $O(\bar{\delta}^2)$ terms, which gets multiplied by constant factors, decays like $1/\sqrt{T}$ or faster (we use that $||\omega_{T/2} - \omega_{\theta_{T/2}}||$ is bounded independently of T, because these terms are in the compact set Ω, and

likewise $||V_{\theta_{T/2}} - V_{\theta_T}||$ is bounded independently of T as the value of any policy lies in $C_{\max}/(1-\gamma)$). Thus the right-hand side of Eq. (5.23) is $O(\bar{\delta}^2 + \mu^{-1}/\sqrt{T})$. We then plug this into Lemma 5.15 to obtain the same bound for the average of the last $T/2$ quantities $||\nabla_t||^2$.

Remark 5.17. We briefly revisit the issue of time-scale separation in the context of our proof above. Recall that, in the Introduction, we discussed how slowing down the actor relative to the critic would result in worse convergence rate. Let us see how this issue appears in the context of our small-gain proof.

Indeed, note that, immediately after Eq. (5.23), we choose α_t and β_t both to be proportional to $1/\sqrt{t}$. To see why we make this choice, observe Eq. (5.23) contains terms that scale as $\alpha_T, (\alpha_T T)^{-1}$, which force the choice $\alpha_T \sim 1/\sqrt{T}$ to get the best decay rate of $T^{-1/2}$ that one can get from Eq. (5.23). Likewise, Lemma 5.15 contains terms that scale as $\beta_T, (\beta_T T)^{-1}$, forcing us to make a similar choice for β_t (fortunately, this pair of choices results in all other terms of those bounds decaying at $T^{-1/2}$ or faster).

Thus, consistent with our earlier discussion, we see that a two-timescale approach which artificially slows down the actor by choosing α_t to decay at an asymptotically faster rate than β_t would result in a worse final convergence rate because at least one of the four terms $\alpha_T, (\alpha_T T)^{-1}, \beta_T, (\beta_T T)^{-1}$ would decay slower than $T^{-1/2}$.

Remark 5.18. It is usually possible to take a small gain argument and convert it to an explicit Lyapunov function, and it is natural to wonder if one could do the same here. Ideally, one would write down an explicit Lyapunov function V which converges to an approximate stationary point of actor-critic at a rate of $\mu/\sqrt{t}$ as in our main result. We would conjecture that this is indeed possible. The main difficulties in establishing this lie in the nonlinear nature of the small-gain theorem we use here and the step-size-dependent discrete nature of the updates.

Remark 5.19. The small gain theorem is extremely widely used in control. The core idea of it is applicable to other settings where convergence of interconnected of dynamical systems is under consideration. Such couplings are naturally present in the context of actor-critic.

The above application of small gain theorem is slightly unusual, since the "typical" application in control requires a careful bound on the input-output gain of a dynamical system, usually involving some kind of frequency analysis. Because the recursions here are relatively simpler from the dynamical point of view, frequency analysis turns out to be unnecessary, and we can attempt to directly bound the gains from critic-error to actor-error and vice versa. The argument presented here is in the spirit of the previous work of one of the authors in [22].

6. Conclusion. We showed how to analyze single timescale actor critic using a nonlinear version of the small gain theorem, resulting in an improved sample complexity of $O\left(\mu^{-2}\epsilon^{-2}\right)$. Our work suggests a number of future directions.

First, we have assumed that the critic uses a linear approximation. Generalizing these results to even some tractable classes of nonlinear approximations is likely to be challenging. However, there are a number of theoretical results suggesting that neural networks in the "large width" regime exhibit nonlinearities which are bounded (in the sense of the gradient having a small Lipschitz constant; for a linear function, the Lipschitz constant of the gradient is zero), and consequently may be analyzable using the small-gain approach we have used here.

Second, one may be able to improve the sample complexity still further. However, we believe that such an improvement would most likely require a more complex algorithm, perhaps using momentum and acceleration, and at the very least going beyond the sort of classic actor-critic update we have studied. Moreover, because accelerated methods are less robust to to errors, the analysis of such an accelerated scheme could be challenging.

Third, there are a number of popular variations of actor critic, such a "natural actor critic" obtained by adding regularization terms. In many applications, such methods are known to significantly outperform actor critic. It may be possible to modify our analysis to give a rigorous analysis of those variations. As discussed in our literature review section, as of writing the best rate for natural actor critic is $O(\epsilon^{-3})$.

Finally, it would be interesting to specialize these rates further to problems of epidemic stabilization, where there is some structure in the underlying RL problem that can be exploited to obtain faster rates, e.g., using the model from [21].

REFERENCES

[1] *Large scale machine learning and optimization*. https://papail.io/teaching/901/scribe_09.pdf. Accessed: 2022-08-02.

[2] A. Barakat, P. Bianchi, and J. Lehmann, *Analysis of a target-based actor-critic algorithm with linear function approximation*, arXiv preprint arXiv:2106.07472, (2021).

[3] A. G. Barto, R. S. Sutton, and C. W. Anderson, *Neuronlike adaptive elements that can solve difficult learning control problems*, IEEE transactions on systems, man, and cybernetics, (1983), pp. 834–846.

[4] A. Beck, *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*, SIAM, 2014.

[5] J. Bhandari, D. Russo, and R. Singal, *A finite time analysis of temporal difference learning with linear function approximation*, in Conference on learning theory, PMLR, 2018, pp. 1691–1692.

[6] V. S. Borkar and S. P. Meyn, *The ode method for convergence of stochastic approximation and reinforcement learning*, SIAM Journal on Control and Optimization, 38 (2000), pp. 447–469.

[7] D. D. Castro and R. Meir, *A convergent online single time scale actor critic algorithm*, The Journal of Machine Learning Research, 11 (2010), pp. 367–410.

[8] T. Chen, Y. Sun, and W. Yin, *Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems*, Advances in Neural Information Processing Systems, 34 (2021).

[9] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, *Finite sample analyses for td (0) with function approximation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[10] A. M. Devraj, A. Bušić, and S. Meyn, *Fundamental design principles for reinforcement learning algorithms*, in Handbook of Reinforcement Learning and Control, Springer, 2021, pp. 75–137.

[11] Z. Fu, Z. Yang, and Z. Wang, *Single-timescale actor-critic provably finds globally optimal policy*, in International Conference on Learning Representations, 2020.

[12] B. Hambly, R. Xu, and H. Yang, *Policy gradient methods for the noisy linear quadratic regulator over a finite horizon*, SIAM Journal on Control and Optimization, 59 (2021), pp. 3359–3391.

[13] B. Heidergott and A. Hordijk, *Taylor series expansions for stationary markov chains*, Advances in Applied Probability, 35 (2003), pp. 1046–1070.

[14] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, *A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic*, arXiv preprint arXiv:2007.05170, (2020).

[15] S. Khodadadian, Z. Chen, and S. T. Maguluri, *Finite-sample analysis of off-policy natural actor-critic algorithm*, in International Conference on Machine Learning, PMLR, 2021, pp. 5420–5431.

[16] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for markov decision*

processes, SIAM Journal on control and Optimization, 38 (1999), pp. 94–123.
[17] V. R. Konda and J. N. Tsitsiklis, *On actor-critic algorithms*, SIAM journal on Control and Optimization, 42 (2003), pp. 1143–1166.
[18] H. Kumar, A. Koppel, and A. Ribeiro, *On the sample complexity of actor-critic method for reinforcement learning with function approximation*, arXiv preprint arXiv:1910.08412, (2019).
[19] A. Lazaridis, A. Fachantidis, and I. Vlahavas, *Deep reinforcement learning: A state-of-the-art walkthrough*, Journal of Artificial Intelligence Research, 69 (2020), pp. 1421–1471.
[20] R. Liu and A. Olshevsky, *Temporal difference learning as gradient splitting*, in International Conference on Machine Learning, PMLR, 2021, pp. 6905–6913.
[21] Q. Ma, Y.-Y. Liu, and A. Olshevsky, *Optimal lockdown for pandemic control*, arXiv preprint arXiv:2010.12923, (2020).
[22] A. Nedic, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization, 27 (2017), pp. 2597–2633.
[23] M. Pirotta, M. Restelli, and L. Bascetta, *Policy gradient in lipschitz markov decision processes*, Machine Learning, 100 (2015), pp. 255–283.
[24] S. Qiu, Z. Yang, J. Ye, and Z. Wang, *On finite-time convergence of actor-critic algorithm*, IEEE Journal on Selected Areas in Information Theory, 2 (2021), pp. 652–664.
[25] A. Roy, V. S. Borkar, A. Karandikar, and P. Chaporkar, *Online reinforcement learning of optimal threshold policies for markov decision processes*, IEEE Transactions on Automatic Control, (2021).
[26] H. Shen, K. Zhang, M. Hong, and T. Chen, *Asynchronous advantage actor critic: non-asymptotic analysis and linear speedup*, arXiv preprint arXiv:2012.15511, (2020).
[27] R. Srikant and L. Ying, *Finite-time error bounds for linear stochastic approximation andtd learning*, in Conference on Learning Theory, PMLR, 2019, pp. 2803–2830.
[28] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, in Advances in neural information processing systems, 2000, pp. 1057–1063.
[29] J. Tsitsiklis and B. Van Roy, *Analysis of temporal-diffference learning with function approximation*, Advances in neural information processing systems, 9 (1996).
[30] H. D. Vinod, *Hands-on matrix algebra using R: Active and motivated learning with applications*, World Scientific Publishing Company, 2011.
[31] Y. F. Wu, W. Zhang, P. Xu, and Q. Gu, *A finite-time analysis of two time-scale actor-critic methods*, Advances in Neural Information Processing Systems, 33 (2020), pp. 17617–17628.
[32] P. Xu, F. Gao, and Q. Gu, *An improved convergence analysis of stochastic variance-reduced policy gradient*, in Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 541–551.
[33] T. Xu, Z. Wang, and Y. Liang, *Improving sample complexity bounds for (natural) actor-critic algorithms*, Advances in Neural Information Processing Systems, 33 (2020), pp. 4358–4369.
[34] K. Zhang, A. Koppel, H. Zhu, and T. Basar, *Global convergence of policy gradient methods to (almost) locally optimal policies*, SIAM Journal on Control and Optimization, 58 (2020), pp. 3586–3612.