ON ACTOR-CRITIC ALGORITHMS[‡]

VIJAY R. KONDA AND JOHN N. TSITSIKLIS[§]

Abstract. In this paper, we propose and analyze a class of actor-critic algorithms. These are two-time-scale algorithms in which the critic uses temporal difference (TD) learning with a linearly parameterized approximation architecture, and the actor is updated in an approximate gradient direction, based on information provided by the critic. We show that the features for the critic should ideally span a subspace prescribed by the choice of parameterization of the actor. We study actor-critic algorithms for Markov decision processes with Polish state and action spaces. We state and prove two results regarding their convergence.

1. Introduction. Many problems in finance, communication networks, operations research, etc., can be formulated as dynamic programming (DP) problems. However, the dimension of the state space in these formulations is often too large for the problem to be tractable. Moreover, the underlying dynamics are seldom known and are often difficult to identify. Reinforcement Learning (RL) and Neuro-Dynamic Programming (NDP) [5, 19] methods try to overcome these difficulties by combining simulation-based learning and compact representations of policies and value functions. The vast majority of these methods falls into one of the following two categories:

(a) Actor-only methods work with a parameterized family of policies. The gradient of the performance, with respect to the actor parameters, is directly estimated by simulation, and the parameters are updated in a direction of improvement [8, 10, 16, 23]. A possible drawback of such methods is that the gradient estimators may have a large variance. Furthermore, as the policy changes, a new gradient is estimated independently of past estimates. Hence, there is no "learning," in the sense of accumulation and consolidation of older information.

(b) Critic-only methods rely exclusively on value function approximation and aim at learning an approximate solution to the Bellman equation, which will then hopefully prescribe a near-optimal policy. Such methods are indirect in the sense that they do not try to optimize directly over a policy space. A method of this type may succeed in constructing a "good" approximation of the value function, yet lack reliable guarantees in terms of near-optimality of the resulting policy.

Actor-critic methods [2] aim at combining the strong points of actor-only and critic-only methods. The critic uses an approximation architecture and simulation to learn a value function, which is then used to update the actor's policy parameters in a direction of performance improvement. Such methods, as long as they are gradient-based, may have desirable convergence properties, in contrast to critic-only methods for which convergence is guaranteed in rather limited settings. They also hold the promise of delivering faster convergence (due to variance reduction), when compared to actor-only methods. On the other hand, theoretical understanding of actor-critic methods has been limited to the case of lookup table representations of policies and value functions [12].

[‡]This research was partially supported by the NSF under contract ECS-9873451 and by the AFOSR under contract F49620-99-1-0320. A preliminary version of this paper was presented at the 1999 NIPS .

[§]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge MA 02139, USA.

In this paper, we propose some actor-critic algorithms in which the critic uses linearly parameterized approximations of the value function, and provide a convergence proof. The algorithms are based on the following important observation: since the number of parameters that the actor has to update is relatively small (compared to the number of states), the critic need not attempt to compute or approximate the exact value function, which is a high-dimensional object. In fact, we show that the critic should ideally compute a certain "projection" of the value function onto a low-dimensional subspace spanned by a set of "basis functions," that are *completely determined* by the parameterization of the actor. This key insight was also derived in simultaneous and independent work [20], that also included a discussion of certain actor-critic algorithms.

The outline of the paper is as follows. In Section 2, we state a formula for the gradient of the average cost in a Markov decision process with finite state and action space. We provide a new interpretation of this formula, and use it in Section 3 to derive our algorithms. In Section 4, we consider Markov decision processes and the gradient of the average cost in much greater generality and describe the algorithms in this more general setting. In Sections 5 and 6, we provide an analysis of the asymptotic behavior of the critic and actor, respectively. The Appendix contains a general result concerning the tracking ability of linear stochastic iterations, which is used in Section 5.

2. Markov decision processes and parameterized families of randomized stationary policies. Consider a Markov decision process with finite state space $\mathbb{X}$, and finite action space $\mathbb{U}$. Let $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ be a given one-stage cost function. Let $p(y|x, u)$ denote the probability that the next state is y, given that the current state is x and the current action is u. A *randomized stationary policy* (RSP) is a mapping μ that assigns to each state x a probability distribution over the action space $\mathbb{U}$. We consider a set of randomized stationary policies $\{\mu_\theta; \theta \in \mathbb{R}^n\}$, parameterized in terms of a vector θ. For each pair $(x, u) \in \mathbb{X} \times \mathbb{U}$, $\mu_\theta(u|x)$ denotes the probability of taking action u when the state x is encountered, under the policy corresponding to θ. Hereafter, we will not distinguish between the parameter of an RSP and the RSP itself. Therefore, whenever we refer to an "RSP θ", we mean the RSP corresponding to parameter vector θ. Note that under any RSP, the sequence of states $\{X_k\}$ and the sequence of state-action pairs $\{X_k, U_k\}$ of the Markov decision process form Markov chains with state spaces $\mathbb{X}$ and $\mathbb{X} \times \mathbb{U}$, respectively. We make the following assumption about the family of policies.

ASSUMPTION 2.1. *(Finite Case)*

(a) *For every $x \in \mathbb{X}$, $u \in \mathbb{U}$, and $\theta \in \mathbb{R}^n$, we have $\mu_\theta(u|x) > 0$.*
(b) *For every $(x, u) \in \mathbb{X} \times \mathbb{U}$, the mapping $\theta \mapsto \mu_\theta(u|x)$ is twice differentiable. Furthermore, the $\mathbb{R}^n$-valued function $\theta \to \nabla \ln \mu_\theta(u|x)$ is bounded, and has a bounded first derivative, for any fixed x and u.*[*]
(c) *For every $\theta \in \mathbb{R}^n$, the Markov chains $\{X_k\}$ and $\{X_k, U_k\}$ are irreducible and aperiodic, with stationary probabilities $\pi_\theta(x)$ and $\eta_\theta(x, u) = \pi_\theta(x)\mu_\theta(u|x)$, respectively, under the RSP θ.*
(d) *There is a positive integer N, state $x^* \in \mathbb{X}$, and $\epsilon_0 > 0$ such that for all*

[*]Throughout the paper, ∇ will stand for the gradient with respect to the vector θ.

$\theta_1, \ldots, \theta_N \in \mathbb{R}^n$,

$$\sum_{k=1}^{N} [P(\theta_1) \cdots P(\theta_k)]_{xx^*} \geq \epsilon_0, \qquad \forall x \in \mathbb{X},$$

where $P(\theta)$ denotes the transition probability matrix for the Markov chain $\{X_k\}$ under the RSP θ. (We use here the notation $[P]_{xx^*}$ to denote the (x, x^*) entry of a matrix P.)

The first three parts of the above assumption are natural and easy to verify. The fourth part assumes that the probability of reaching x^*, in a number of transitions that is independent of θ, is uniformly bounded away from zero. This assumption is satisfied if part (c) of the assumption holds, and the policy probabilities $\mu_\theta(u|x)$ are all bounded away from zero uniformly in θ (see [11]).

Consider the average cost function $\bar{\alpha} : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\bar{\alpha}(\theta) = \sum_{x \in \mathbb{X}, u \in \mathbb{U}} c(x, u) \eta_\theta(x, u).$$

A natural approach to minimize $\bar{\alpha}(\theta)$ over RSPs θ, is to start with a policy θ_0 and improve it using gradient descent. To do this, we will rely on a formula for $\nabla \bar{\alpha}(\theta)$ to be presented shortly.

For each $\theta \in \mathbb{R}^n$, let $V_\theta : \mathbb{X} \to \mathbb{R}$ be a “differential cost function”, i.e., a solution of the Poisson equation:

$$\bar{\alpha}(\theta) + V_\theta(x) = \sum_u \mu_\theta(u|x) \left[c(x, u) + \sum_y p(y|x, u) V_\theta(y) \right].$$

Intuitively, $V_\theta(x)$ can be viewed as the “disadvantage” of state x: it is the expected future excess cost – on top of the average cost – incurred if we start at state x. It plays a role similar to that played by the more familiar value function that arises in total or discounted cost Markov decision problems. Finally, for every $\theta \in \mathbb{R}^n$, we define the Q-value function $Q_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$, by

$$Q_\theta(x, u) = c(x, u) - \bar{\alpha}(\theta) + \sum_y p(y|x, u) V_\theta(y).$$

We recall the following result, as stated in [16]. (Such a result has been established in various forms in [7, 8, 10] and elsewhere.)

Theorem 2.2. *We have*

$$\nabla \bar{\alpha}(\theta) = \sum_{x,u} \eta_\theta(x, u) Q_\theta(x, u) \psi_\theta(x, u), \tag{2.1}$$

where

$$\psi_\theta(x, u) = \nabla \ln \mu_\theta(u|x).$$

In [16], the quantity $Q_\theta(x, u)$ in the above formula is interpreted as the expected excess cost incurred over a certain renewal period of the Markov chain $\{X_n, U_n\}$, under the RSP μ_θ, and is then estimated by means of simulation, leading to actor-only algorithms. Here, we provide an alternative interpretation of the formula in Theorem 2.2, as an inner product, and arrive at a different set of algorithms.

For any $\theta \in \mathbb{R}^n$, we define the inner product $\langle \cdot, \cdot \rangle_\theta$ of two real-valued functions Q_1, Q_2 on $\mathbb{X} \times \mathbb{U}$, viewed as vectors in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, by

$$\langle Q_1, Q_2 \rangle_\theta = \sum_{x,u} \eta_\theta(x,u) Q_1(x,u) Q_2(x,u).$$

(We will be using the above notation for vector or matrix-valued functions as well.) With this notation, we can rewrite the formula (2.1) as

$$\frac{\partial}{\partial \theta_i} \bar{\alpha}(\theta) = \langle Q_\theta, \psi_\theta^i \rangle_\theta, \qquad i = 1, \ldots, n,$$

where ψ_θ^i stands for the ith component of ψ_θ. Let $\|\cdot\|_\theta$ denote the norm induced by this inner product on $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$. For each $\theta \in \mathbb{R}^n$, let Ψ_θ denote the span of the vectors $\{\psi_\theta^i;\ 1 \leq i \leq n\}$ in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$.

An important observation is that although the gradient of $\bar{\alpha}$ depends on the function Q_θ, which is a vector in a possibly very high dimensional space $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, the dependence is only through its inner products with vectors in Ψ_θ. Thus, instead of "learning" the function Q_θ, it suffices to learn its projection on the low dimensional subspace Ψ_θ.

Indeed, let $\Pi_\theta : \mathbb{R}^{|\mathbb{X}||\mathbb{U}|} \mapsto \Psi_\theta$ be the projection operator defined by

$$\Pi_\theta Q = \arg \min_{\hat{Q} \in \Psi_\theta} \|Q - \hat{Q}\|_\theta.$$

Since

$$\langle Q_\theta, \psi_\theta^i \rangle_\theta = \langle \Pi_\theta Q_\theta, \psi_\theta^i \rangle_\theta, \qquad i = 1, \ldots, n, \tag{2.2}$$

it is enough to know the projection of Q_θ onto Ψ_θ to compute $\nabla \bar{\alpha}$.

3. Actor-critic algorithms. We view actor-critic algorithms as stochastic gradient algorithms on the parameter space of the actor. When the actor parameter vector is θ, the job of the critic is to compute an approximation of the projection $\Pi_\theta Q_\theta$, which is then used by the actor to update its policy in an approximate gradient direction. The analysis in [21, 22] shows that this is precisely what Temporal Difference (TD) learning algorithms try to do, i.e., to compute the projection of an exact value function onto a subspace spanned by feature vectors. This allows us to implement the critic by using a TD algorithm. (Note, however, that other types of critics are possible, e.g., based on batch solution of least squares problems, as long as they aim at computing the same projection.)

We note some minor differences with the common usage of TD. In our context, we need the projection of q-functions, rather than value functions. But this is easily achieved by replacing the Markov chain $\{x_t\}$ in [21, 22] by the Markov chain $\{X_k, U_k\}$. A further difference is that [21, 22] assume that the decision policy and the feature vectors are fixed. In our algorithms, the decision policy as well as the features need to change as the actor updates its parameters. As suggested by the results of [12, 6, 14], this need not pose any problems, as long as the actor parameters are updated on a slower time scale.

We are now ready to describe two actor-critic algorithms, which differ only as far as the critic updates are concerned. In both variants, the critic is a TD algorithm with a linearly parameterized approximation architecture for the Q-value function, of the form

$$Q_\theta^r(x,u) = \sum_{j=1}^m r^j \phi_\theta^j(x,u),$$

where $r = (r^1, \dots, r^m) \in \mathbb{R}^m$ denotes the parameter vector of the critic. The features ϕ_θ^j, $j = 1, \dots, m$, used by the critic are dependent on the actor parameter vector θ, and are chosen so that the following assumptions are satisfied.

Assumption 3.1. (Critic Features)

(a) *For every $(x,u) \in \mathbb{X} \times \mathbb{U}$ the map $\theta \to \phi_\theta(x,u)$ is bounded and differentiable, with a bounded derivative.*

(b) *The span of the vectors ϕ_θ^j, $j = 1, \dots, m$, in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, denoted by Φ_θ, contains Ψ_θ.*

Note that the formula (2.2) still holds if Π_θ is redefined as the projection onto Φ_θ, as long as Φ_θ contains Ψ_θ. The most straightforward choice would be to let the number m of critic parameters be equal to the number n of actor parameters, and $\phi_\theta^i = \psi_\theta^i$ for each i. Nevertheless, we allow the possibility that $m > n$ and that Φ_θ properly contains Ψ_θ, so that the critic can use more features than are actually necessary. This added flexibility may turn out to be useful in a number of ways:

(a) It is possible that for certain values of θ, the feature vectors ψ_θ^i are either close to zero or are almost linearly dependent. For these values of θ, the operator Π_θ becomes ill-conditioned which can have a negative effect on the performance of the algorithms. This might be avoided by using a richer set of features ϕ_θ^i.

(b) For the second algorithm that we propose, which involves a TD(λ) critic with $\lambda < 1$, the critic can only compute an approximate – rather than exact – projection. The use of additional features can result in a reduction of the approximation error.

To avoid the above first possibility, we choose features for the critic so that our next assumption is satisfied. To understand that assumption, note that if the functions $\underline{1}$ and ϕ_θ^j, $j = 1, \dots, m$, are linearly independent for each θ, then there exists a positive function $a(\theta)$ such that

$$||r'\hat{\phi}_\theta||_\theta^2 \geq a(\theta)|r|^2,$$

where $|r|$ is the Euclidean norm of r, and $\hat{\phi}_\theta$ is the projection of ϕ_θ on the subspace orthogonal to the function $\underline{1}$. (Here and throughout the rest of the paper, $\underline{1}$ stands for a function which is identically equal to 1.) Our assumption below involves the stronger requirement that the function $a(\cdot)$ be uniformly bounded away from zero.

Assumption 3.2. *There exists $a > 0$, such that for every $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$,*

$$||r'\hat{\phi}_\theta||_\theta^2 \geq a|r|^2,$$

where

$$\hat{\phi}_\theta(x,u) = \phi_\theta(x,u) - \sum_{\bar{x},\bar{u}} \eta_\theta(\bar{x},\bar{u})\phi_\theta(\bar{x},\bar{u}).$$

Along with the parameter vector r, the critic stores some auxiliary parameters: a scalar estimate α of the average cost, and an m-vector $\hat{Z}$ which represents Sutton's eligibility trace [5, 19]. The actor and critic updates take place in the course of a simulation of a single sample path of the Markov decision process. Let $r_k, \hat{Z}_k, \alpha_k$ be the parameters of the critic, and let θ_k be the parameter vector of the actor, at time k. Let $(\hat{X}_k, \hat{U}_k)$ be the state-action pair at that time. Let $\hat{X}_{k+1}$ be the new state, obtained after action $\hat{U}_k$ is applied. A new action $\hat{U}_{k+1}$ is generated according to the RSP corresponding to the actor parameter vector θ_k. The critic carries out an update similar to the average cost temporal-difference method of [22]:

$$\begin{aligned} \alpha_{k+1} &= \alpha_k + \gamma_k(c(\hat{X}_{k+1}, \hat{U}_{k+1}) - \alpha_k), \\ r_{k+1} &= r_k + \gamma_k d_k \hat{Z}_k, \end{aligned} \tag{3.1}$$

where the temporal difference d_k is defined by

$$d_k = c(\hat{X}_k, \hat{U}_k) - \alpha_k + r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - r_k' \phi_{\theta_k}(\hat{X}_k, \hat{U}_k),$$

and where γ_k is a positive step-size parameter. The two variants of the critic differ in their update of $\hat{Z}_k$, which is as follows.

TD(1) Critic:.

$$\begin{aligned} \hat{Z}_{k+1} &= \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), \qquad \text{if} \quad \hat{X}_{k+1} \neq x^*, \\ &= \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), \qquad \text{otherwise}, \end{aligned}$$

where x^* is the special state introduced in Assumption 2.1.

TD(λ) Critic, $0 < \lambda < 1$**:.**

$$\hat{Z}_{k+1} = \lambda \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}).$$

Actor:. Finally, the actor updates its parameter vector according to

$$\theta_{k+1} = \theta_k - \beta_k \Gamma(r_k) r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) \psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), \tag{3.2}$$

where $\Gamma(\cdot)$ is a scalar that controls the step-size β_k of the actor, taking into account the current estimate r_k of the critic.

Note that we have used $\hat{X}_k$, $\hat{U}_k$ and $\hat{Z}_k$ to denote the simulated processes in the above algorithm. Throughout the paper we will use hats to denote the simulated processes that are used to update the parameters in the algorithm, and use X_k, U_k, and Z_k to denote processes in which a fixed RSP θ is used.

To understand the actor update, recall the formulas (2.1) and (2.2). According to these formulas, if the projection $\hat{Q}_\theta$ of Q_θ onto the subspace Φ_θ (which contains Ψ_θ) was known for the current value of $\theta \in \mathbb{R}^n$, then $\hat{Q}_{\theta_k}(\hat{X}_k, \hat{U}_k)\psi_{\theta_k}(\hat{X}_k, \hat{U}_k)$ would be a reasonable estimate of $\nabla \bar{\alpha}(\theta_k)$, because the steady-state expected value of the former is equal to the latter. However, $\hat{Q}_{\theta_k}(\hat{X}_k, \hat{U}_k)$ is not known, and it is natural to use in its place the critic's current estimate, which is $Q_{\theta_k}^{r_k}(\hat{X}_k, \hat{U}_k) = r_k' \phi_\theta(\hat{X}_k, \hat{U}_k)$. For the above scheme to converge, it is then important that the critic's estimate be accurate (at least asymptotically). This will be indeed established in Section 5, under the following assumption on the step-sizes.

Assumption 3.3.

(a) The step-sizes β_k and γ_k are deterministic, non-increasing, and satisfy

$$\sum_k \beta_k = \sum_k \gamma_k = \infty,$$

$$\sum_k \beta_k^2 < \infty, \quad \sum_k \gamma_k^2 < \infty \quad \textit{and} \quad \sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty,$$

for some $d > 0$.

(b) *The function $\Gamma(\cdot)$ is assumed to satisfy the following inequalities for some positive constants $C_1 < C_2$:*

$$|r|\Gamma(r) \in [C_1, C_2], \quad \forall r \in \mathbb{R}^m,$$
$$|\Gamma(r) - \Gamma(\hat{r})| \leq \frac{C_2|r - \hat{r}|}{1 + |r| + |\hat{r}|} \quad \forall r, \hat{r} \in \mathbb{R}^n. \tag{3.3}$$

The following result on the convergence properties of the actor is established in Section 6, in much greater generality.

THEOREM 3.4. *Under Assumptions 2.1 and 3.1-3.3, the following hold.*

(a) *In the actor-critic algorithm with a TD(1) critic, $\liminf_k |\nabla\bar{\alpha}(\theta_k)| = 0$, w.p.1.*

(b) *For each $\epsilon > 0$, there exists λ sufficiently close to 1 such that, in the actor-critic algorithm with a TD(λ) critic, $\liminf_k |\nabla\bar{\alpha}(\theta_k)| < \epsilon$, w.p.1.*

The algorithms introduced in this section are only two out of many possible variations. For instance, one can also consider "episodic" problems, in which one starts from a given initial state x^* and runs the process until a random termination time (at which time the process is reinitialized at x^*), with the objective of minimizing the expected total cost until termination. In this setting, the average cost estimate α_k is unnecessary and is removed from the critic update formula. If the critic parameter r_k were to be reinitialized each time that x^* is entered, one would obtain a method closely related to Williams' REINFORCE algorithm [23]. Such a method does not involve any value function learning, because the observations during one episode do not affect the critic parameter r during another episode. In contrast, in our approach, the observations from all past episodes affect the current critic parameter r, and in this sense, the critic is "learning". This can be advantageous because, as long as θ is changing slowly, the observations from recent episodes carry useful information on the Q-value function under the current policy.

The analysis of actor-critic methods for total and/or discounted cost problems is similar (in fact, a little simpler) than that for the average cost case; see [20, 11].

4. Algorithms for Polish state and action spaces.

In this section, we consider actor-critic algorithms for Markov decision processes with Polish (complete, separable, metric) state and action spaces. The algorithms are the same as for the case of finite state and action spaces and therefore will not be repeated in this section. However, we will restate our assumptions in the general setting, as the notation and the theory is quite technical. Throughout, we will use the abbreviation *w.p.1.* for the phrase *with probability 1.* We will denote norms on real Euclidean spaces with $|\cdot|$, and norms on Hilbert spaces by $||\cdot||$. For a probability measure ν and a ν-integrable function f, $\nu(f)$ will denote the expectation of f with respect to ν. Finally, for any Polish space $\mathbb{X}$, $\mathcal{B}(\mathbb{X})$ denotes its countably generated Borel σ-field.

4.1. Preliminaries.

Consider an MDP in which the state space $\mathbb{X}$ and the action space $\mathbb{U}$ are Polish spaces, and with a transition kernel $p(dy|x, u)$ which for every (x, u) defines a probability measure on $\mathbb{X}$. In the finite case, we had considered a parameterized family of randomized stationary policies (RSPs) described by a parameterized family of probability mass functions. Similarly, we now consider a family of parameterized RSPs specified by a parameterized family of probability density functions. More specifically, let ν be a fixed measure on the action space $\mathbb{U}$. Let $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ be a family of positive measurable functions on $\mathbb{X} \times \mathbb{U}$ such that for each $x \in \mathbb{X}$, $\mu_\theta(\cdot|x)$ is a probability density function with respect to $\nu(du)$, i.e.,

$$\int \mu_\theta(u|x)\nu(dy) = 1, \quad \forall x, \theta.$$

This parameterized family of density functions can be viewed as a parameterized family of RSPs where, for each $\theta \in \mathbb{R}^n$, the probability distribution of an action at state x under RSP θ is given by $\mu_\theta(u|x)\nu(du)$.

Note that the state-action process $\{X_k, U_k\}$ of an MDP controlled by any fixed RSP is a Markov chain. For each θ, let $\mathbf{P}_{\theta,x}$ denote the probability law of the state-action process $\{X_k, U_k\}$ in which the starting state X_0 is x. Let $\mathbf{E}_{\theta,x}$ denote expectation w.r.t. $\mathbf{P}_{\theta,x}$.

Assumption 4.1. (Irreducibility and aperiodicity) *For each $\theta \in \mathbb{R}^n$, the process $\{X_k\}$ controlled by RSP θ is irreducible and aperiodic.*

For the details on the notion of irreducibility for general state space Markov chains, see [17]. Under Assumption 4.1, it follows from Theorem 5.2.2 of [17] that for each $\theta \in \mathbb{R}^n$, there exists a set of states $\mathbb{X}_0(\theta) \in \mathcal{B}(\mathbb{X})$, a positive integer $N(\theta)$, a constant $\delta_\theta > 0$, and a probability measure ϑ_θ on $\mathbb{X}$, such that $\vartheta_\theta(\mathbb{X}_0(\theta)) = 1$ and

$$\mathbf{P}_{\theta,x}(X_{N(\theta)} \in B) \geq \delta_\theta \vartheta_\theta(B), \qquad \forall \theta \in \mathbb{R}^n, \quad x \in \mathbb{X}_0(\theta), \quad B \in \mathcal{B}(\mathbb{X}).$$

We will now assume that such a condition holds uniformly in θ. This is one of the most restrictive of our assumptions. It corresponds to a "stochastic stability" condition, which holds uniformly over all policies.

Assumption 4.2. (Uniform Geometric Ergodicity)

(a) There exists a positive integer N, a set $\mathbb{X}_0 \in \mathcal{B}(\mathbb{X})$, a constant $\delta > 0$, and a probability measure ϑ on $\mathbb{X}$, such that

$$\mathbf{P}_{\theta,x}(X_N \in B) \geq \delta \vartheta(B), \qquad \forall \theta \in \mathbb{R}^n, \quad x \in \mathbb{X}_0, \quad B \in \mathcal{B}(\mathbb{X}). \tag{4.1}$$

(b) There exists a function $L : \mathbb{X} \to [1, \infty)$ and constants $0 \leq \rho < 1$, $b > 0$, such that for each $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_{\theta,x}[L(X_1)] \leq \rho L(x) + b I_{\mathbb{X}_0}(x), \qquad \forall x \in \mathbb{X}, \tag{4.2}$$

where $I_{\mathbb{X}_0}(\cdot)$ is the indicator function of the set $\mathbb{X}_0$. We call a function L satisfying the above condition a stochastic Lyapunov function.

We note that in the finite case, Assumption 2.1(d) implies that Assumption 4.2 holds. Indeed, the first part of Assumption 4.2 is immediate, with $\mathbb{X}_0 = \{x^*\}$, $\delta_\theta = \epsilon_0$, and with ϑ equal to a point mass at state x^*. To verify the second part, consider the first hitting time τ of the state x^*. For a sequence $\{\theta_k\}$ of values of the actor parameter, consider the time-varying Markov chain obtained by using policy θ_k at time k. For $s > 1$, consider the function

$$L(x) = \sup_{\{\theta_k\}} E\left[s^\tau \,|\, X_0 = x \right].$$

Assumption 2.1(d) guarantees that $L(\cdot)$ is finite when s is sufficiently close to 1. Then, it is a matter of simple algebraic calculations to see that $L(\cdot)$ satisfies (4.2).

Using geometric ergodicity results (Theorem 15.0.1) in [17], it can be shown that if Assumption 4.2 is satisfied then, for each $\theta \in \mathbb{R}^n$, the Markov chains $\{X_k\}$ and $\{X_k, U_k\}$ have steady-state distributions $\pi_\theta(dx)$ and

$$\eta_\theta(dx, du) = \pi_\theta(dx)\mu_\theta(u|x)\nu(du),$$

respectively. Moreover, the steady state is reached at a geometric rate (see Lemma 4.3 below). For any $\theta \in \mathbb{R}^n$, we will use $\langle \cdot, \cdot \rangle_\theta$ and $||\cdot||_\theta$ to denote the inner product and the norm, respectively, on $\mathcal{L}^2(\eta_\theta)$. Finally, for any $\theta \in \mathbb{R}^n$, we define the operator P_θ on $\mathcal{L}^2(\eta_\theta)$ by

$$\begin{aligned}(P_\theta Q)(x, u) &= \mathbf{E}_\theta[Q(X_1, U_1) \mid X_0 = x, U_0 = u] \\ &= \int Q(y, \bar{u})\mu_\theta(\bar{u}|y)p(dy|x, u)\nu(d\bar{u}), \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U},\ Q \in \mathcal{L}^2(\eta_\theta).\end{aligned}$$

For the finite case, we introduced certain boundedness assumptions on the maps $\theta \mapsto \psi_\theta(x, u)$ and $\theta \mapsto \phi_\theta(x, u)$, and their derivatives. For the more general case considered here, these bounds may depend on the state-action pair (x, u). We wish to bound the rate of growth of such functions, as (x, u) changes, in terms of the stochastic Lyapunov function L. Towards this purpose, we introduce a class $\mathcal{D}$ of functions that satisfy the desired growth conditions.

We will say that a parameterized family of functions $f_\theta : \mathbb{X} \times \mathbb{U} \mapsto \mathbb{R}$ belongs to $\mathcal{D}$ if there exists a function $q : \mathbb{X} \times \mathbb{U} \mapsto \mathbb{R}$ and constants C, K_d $(d \geq 1)$, such that

$$f_\theta(x, u) \leq Cq(x, u), \qquad \forall\ x \in \mathbb{X},\ u \in \mathbb{U},\ \theta \in \mathbb{R}^n,$$

and

$$\mathbf{E}_{\theta,x}\left[|q(x, U_0)|^d\right] \leq K_d L(x), \qquad \forall\ \theta, x, d \geq 1.$$

For easy reference, we collect here various useful properties of the class $\mathcal{D}$. The proof is elementary and is omitted.

LEMMA 4.3. *Consider a process $\{\hat{X}_k, \hat{U}_k\}$ driven by RSPs θ_k which change with time, but in a non-anticipative manner (i.e., θ_k is completely determined by $(\hat{X}_l, \hat{U}_l)$, $l \leq k$). Assume that $\mathbf{E}[L(\hat{X}_0)] < \infty$.*

(a) The sequence $\mathbf{E}[L(\hat{X}_k)]$, $k = 1, 2, \ldots$, is bounded.

(b) If the parametric class of functions f_θ belongs to $\mathcal{D}$, then for any $d \geq 1$ and any (possibly random) sequence $\{\tilde{\theta}_k\}$

$$\sup_k \mathbf{E}\left[\left|f_{\tilde{\theta}_k}(\hat{X}_k, \hat{U}_k)\right|^d\right] < \infty.$$

(c) In particular, the above boundedness property holds when θ_k and $\tilde{\theta}_k$ are held fixed at some θ, for all k, so that the process $\{\hat{X}_k, \hat{U}_k\}$ is time-homogeneous.

(d) If $f_\theta \in \mathcal{D}$ then the maps $(x, u) \to \mathbf{E}_{\theta,x}[f_\theta(x, U_0)]$ and $(x, u) \to (P_\theta f_\theta)(x, u)$ also belong to $\mathcal{D}$, and

$$f_\theta \in \mathcal{L}^d(\eta_\theta), \qquad \forall \theta \in \mathbb{R}^n,\ d \geq 1.$$

(e) For any function $f \in \mathcal{D}$, the steady-state expectation $\pi_\theta(f)$ is well-defined, it is a bounded function of θ, and there exists a constant $C > 0$ such that

$$|\mathbf{E}_{\theta,x}[f(X_k, U_k)] - \pi_\theta(f)| \leq C\rho^k L(x), \qquad \forall x \in \mathbb{X},\ \theta \in \mathbb{R}^n. \tag{4.3}$$

(f) *If the parametric classes of functions f_θ and g_θ belong to $\mathcal{D}$, then*

$$f_\theta + g_\theta \in \mathcal{D}, \qquad f_\theta g_\theta \in \mathcal{D}.$$

The next two assumptions will be used to show that the average cost is a smooth function of the policy parameter θ. In the finite case, their validity is an automatic consequence of Assumption 2.1.

ASSUMPTION 4.4. (Differentiability)

(a) *For every $x \in \mathbb{X}$, $u \in \mathbb{U}$, and $\theta \in \mathbb{R}^n$, we have $\mu_\theta(u|x) > 0$.*

(b) *The mapping $\theta \mapsto \mu_\theta(u|x)$ is twice differentiable. Furthermore, $\psi_\theta(x,u) = \nabla \ln \mu_\theta(u|x)$ and its derivative belong to $\mathcal{D}$.*

(c) *For every θ_0, there exists $\epsilon > 0$ such that the class of functions*

$$\{\nabla\mu_\theta(u|x)/\mu_{\bar\theta}(u|x),\ |\theta-\theta_0| \le \epsilon,\ |\bar\theta - \theta_0| \le \epsilon\}$$

(parameterized by θ and $\bar\theta$) belongs to $\mathcal{D}$.

ASSUMPTION 4.5. *The cost function $c(\cdot,\cdot)$ belongs to $\mathcal{D}$.*

Under the above assumptions we wish to prove that a gradient formula similar to (2.1) is again valid. By Assumption 4.5 and Lemma 4.3, $c \in \mathcal{L}^2(\eta_\theta)$ and therefore the average cost function can be written as

$$\bar\alpha(\theta) = \int c(x,u)\pi_\theta(dx)\mu_\theta(u|x)\nu(du) = \langle c, \underline{1}\rangle_\theta.$$

We say that $Q \in \mathcal{L}^2(\eta_\theta)$ is a solution of the Poisson equation with parameter θ if Q satisfies

$$Q = c - \bar\alpha(\theta)\underline{1} + P_\theta Q. \tag{4.4}$$

Using Proposition 17.4.1 from [17], one can easily show that a solution to Poisson equation with parameter θ exists and is unique up to a constant. That is, if Q_1, Q_2 are two solutions, then $Q_1 - Q_2$ and $\underline{1}$ are collinear in $\mathcal{L}^2(\eta_\theta)$. One obvious family of solutions to the Poisson equation is

$$Q_\theta(x,u) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x}\left[(c(X_k,U_k) - \bar\alpha(\theta)) \middle| U_0 = u\right].$$

(The convergence of the above series is a consequence of (4.3).)

There are other (e.g., regenerative) representations of solutions to the Poisson equation which are useful both for analysis and for derivation of algorithms. For example, Glynn and L'Ecuyer [9] use regenerative representations to show that the steady state expectation of a function is differentiable under certain assumptions. We use similar arguments to prove that the average cost function $\bar\alpha(\cdot)$ is twice differentiable with bounded derivatives. Furthermore, it can be shown that there exist solutions $\hat{Q}_\theta(x,u)$ to the Poisson equation that are differentiable in θ. From a technical point of view, our assumptions are similar to those provided by Glynn and L'Ecuyer [9]. The major difference is that [9] concerns Markov chains $\{X_k\}$ that have the recursive representation

$$X_{k+1} = f(X_k, W_k),$$

where W_k are i.i.d., whereas we allow the distribution of W_k (which is U_k in our case) to depend on X_k. Furthermore, the formula for the gradient of steady state expectations that we derive here is quite different from that of [9], and makes explicit the role of the Poisson equation in gradient estimation. The following theorem holds for any solution $Q_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ of the Poisson equation with parameter θ. We only provide an outline of the proof and refer the reader to [15] for the details.

THEOREM 4.6. *Under Assumptions 4.1, 4.2, 4.4, and 4.5,*

$$\nabla \bar{\alpha}(\theta) = \langle \psi_\theta, Q_\theta \rangle_\theta.$$

Furthermore, $\nabla \bar{\alpha}(\theta)$ *has bounded derivatives.*

Proof. (Outline) Using regenerative representations and likelihood ratio methods, we can show that $\bar{\alpha}(\theta)$ is differentiable and that there exists a parameterized family $\{\hat{Q}_\theta(x,u)\}$ of solutions to the Poisson equation, belonging to $\mathcal{D}$, such that the map $\theta \to \hat{Q}_\theta(x,u)$ is differentiable for each (x,u), and such that the family of functions $\nabla \hat{Q}_\theta(x,u)$ belongs to $\mathcal{D}$ (see [15]). Then, one can differentiate both sides of equation (4.4) with respect to θ, to obtain

$$\nabla \bar{\alpha}(\theta)\underline{1} + \nabla \hat{Q}_\theta = P_\theta(\psi_\theta \hat{Q}_\theta) + P_\theta(\nabla \hat{Q}_\theta).$$

(This step involves an interchange of differentiation and integration is justified by uniform integrability). Taking inner product with $\underline{1}$ on both sides of the above equation, and using the facts that $\nabla \hat{Q}_\theta \in \mathcal{L}^2(\eta_\theta)$ and

$$\langle \underline{1}, P_\theta f \rangle_\theta = \langle \underline{1}, f \rangle_\theta, \qquad \forall f \in \mathcal{L}^2(\eta_\theta),$$

we obtain $\nabla \bar{\alpha}(\theta) = \langle \hat{Q}_\theta, \psi_\theta \rangle_\theta = \langle Q_\theta, \psi_\theta \rangle_\theta$, where the second equality follows from the fact that $Q_\theta - \hat{Q}_\theta$ and $\underline{1}$ are necessarily collinear, and the easily verified fact $\langle \underline{1}, \psi_\theta \rangle_\theta = 0$.

Since ψ_θ and $\hat{Q}_\theta$ are both differentiable w.r.t. θ, with the derivatives belonging to $\mathcal{D}$, the formula

$$\nabla \bar{\alpha}(\theta) = \langle \psi_\theta, \hat{Q}_\theta \rangle_\theta = \langle \underline{1}, \psi_\theta \hat{Q}_\theta \rangle$$

implies that $\nabla \bar{\alpha}(\theta)$ is also differentiable with bounded derivative. ☐

Before we move on to present the algorithms for Polish state and action spaces, we illustrate how the above assumptions can be verified in the context of a simple inventory control problem.

Example 4.7. Consider a facility with $X_k \in \mathbb{R}$ amount of stock at the beginning of the k-th period, with negative stock representing the unsatisfied (or backlogged) demand. Let $D_k \geq 0$ denote the random demand during the k-th period. The problem is to determine the amount of stock to be ordered at the beginning of the k-th period, based on the current stock and the previous demands. If $U_k \geq 0$ represents the amount of stock ordered at the beginning of the k-th period, then the cost incurred is assumed to be

$$c(X_k, U_k) = h \max(0, X_k) + b \max(0, -X_k) + pU_k,$$

where p is the price of the material per unit, b is the cost incurred per unit of backlogged demand, and h is the holding cost per unit of stock in the inventory. Moreover, the evolution of the stock X_k is given by

$$X_{k+1} = X_k + U_k - D_k, \quad k = 0, 1, \ldots .$$

If we assume that the demands D_k, $\quad k = 0, 1\dots$ are nonnegative and i.i.d. with finite mean, then it is well known (e.g. see [4]) that there is an optimal policy μ^* of the form

$$\mu^*(x) = \max(S - x, 0)$$

for some $S > 0$ depending on the distribution of D_k. A good approximation for policies having the above form is the family of randomized policies in which S is chosen at random from the density

$$p_\theta(s) = \frac{1}{2T}\mathrm{sech}^2\left(\frac{s - \bar{s}(\theta)}{T}\right)$$

where $\bar{s}(\theta) = e^\theta C/(1 + e^\theta)$. The constant C is picked based on our prior knowledge of an upper bound on the parameter S in an optimal policy. To define the family of density functions $\{\mu_\theta\}$ for the above family of policies, let $\nu(du)$ be the sum of the Dirac measure at 0 and the Lebesgue measure on $[0, \infty)$. Then, the density functions are given by

$$\begin{aligned}
\mu_\theta(0|x) &= \frac{1}{2}\left(1 + \tanh\left(\frac{x - \bar{s}(\theta)}{T}\right)\right), \\
\mu_\theta(u|x) &= \frac{1}{2T}\mathrm{sech}^2\left(\frac{x + u - \bar{s}(\theta)}{T}\right), \qquad u > 0.
\end{aligned}$$

The dynamics of the stock in the inventory, when controlled by policy μ_θ, are described by

$$X_{k+1} = \max(X_k, S_k) - D_k, \qquad k = 0, 1\dots,$$

where the $\{S_k\}$ are i.i.d. with density p_θ and independent of the demands D_k and the stock X_k. It is easy to see that the Markov chain $\{X_k\}$ is irreducible. To prove that the Markov chain is aperiodic it suffices to show that (4.1) holds with $N = 1$. Indeed, for $\mathbb{X}_0 = [-a, a]$, $x \in \mathbb{X}_0$, and a Borel set B consider

$$\begin{aligned}
\mathbf{P}_{\theta,x}(X_1 \in B) &= \mathbf{P}_{\theta,x}(\max(x, S_0) - D_0 \in B), \\
&\geq \mathbf{P}_{\theta,x}(S_0 - D_0 \in B, S_0 \geq a), \\
&\geq \int_B \int_{a-t}^{\infty} \left(\inf_\theta p_\theta(t + y)\right) D(dy)dt,
\end{aligned}$$

where $D(dy)$ is the probability distribution of D_0 and $\vartheta(dy)$ is the r.h.s appropriately normalized. This normalization is possible because the above integral is positive when $B = \mathbb{X}_0$.

To prove the Lyapunov condition (4.2), assume that D_k has exponentially decreasing tails. In other words, assume that there exists $\gamma > 0$ such that

$$\mathbf{E}[\exp(\gamma D_0)] < \infty.$$

We first argue intuitively that the function $L(x) = \exp(\bar{\gamma}|x|)$, for some positive $\bar{\gamma}$ less than $\min(\gamma, \frac{1}{T})$, is a good candidate Lyapunov function. To see this, note that the desired inequality (4.2) requires the Lyapunov function to decrease by a common factor outside some set $\mathbb{X}_0$. Let us try the set $\mathbb{X}_0 = [-a, a]$ for a sufficiently larger than C. If the inventory starts with a stock larger than a, then no stock is ordered with very high probability (since S_0 is most likely less than C) and therefore the stock decreases by D_0, decreasing the Lyapunov function by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$. If the inventory starts with a large backlogged demand then most likely new stock will be ordered to satisfy all the backlogged demand decreasing the Lyapunov function to almost 1. This can be made precise as follows:

$$\begin{aligned}
\mathbf{E}_{\theta,x}[L(X_1)] &= \mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|\max(x,S_0)-D_0|)] \\
&= \exp(\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \le x)\mathbf{E}_{\theta,x}[\exp(-\bar{\gamma}D_0); D_0 \le x] \\
&\qquad + \exp(-\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \le x)\mathbf{E}_{\theta,x}[\exp(\bar{\gamma}D_0); D_0 > x] \\
&\qquad\quad + \mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|S_0-D_0|); S_0 > x].
\end{aligned}$$

Note that the third term is bounded uniformly in θ, x since $\bar{\gamma} < \min(\frac{1}{T}, \gamma)$. The first term is bounded when x is negative and the second term is bounded when x is positive. Therefore the Lyapunov function decreases by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$ when $x > a$ and decreases by a factor of $\mathbf{P}(S_0 \le -a)\mathbf{E}[\exp(\bar{\gamma}D_0)] < 1$ for a sufficiently large. The remaining assumptions are easy to verify.

4.2. Critic. In the finite case, the feature vectors were assumed to be bounded. This assumption is seldom satisfied for infinite state spaces. However, it is reasonable to impose some bounds on the growth of the feature vectors, as in the next assumption.

Assumption 4.8. (Critic features)

(a) *The family of functions $\phi_\theta(x,u)$ belongs to $\mathcal{D}$.*

(b) *For each (x,u), the map $\theta \mapsto \phi_\theta(x,u)$is differentiable, and the family of functions $\nabla\phi_\theta(x,u)$ belongs to $\mathcal{D}$.*

(c) *There exists some $a > 0$, such that*

$$||r'\hat{\phi}_\theta||_\theta^2 \ge a|r|^2, \qquad \forall\theta \in \mathbb{R}^n, r \in \mathbb{R}^m \tag{4.5}$$

where $\hat{\phi}_\theta = \phi_\theta - \langle\phi_\theta, \underline{1}\rangle_\theta\underline{1}$.

(d) *For each $\theta \in \mathbb{R}^n$, the subspace Φ_θ in $\mathcal{L}^2(\eta_\theta)$ spanned by the features ϕ_θ^i, $i = 1, \dots, m$, of the critic contains the subspace Ψ_θ spanned by the functions ψ_θ^j, $j = 1, \dots, n$, i.e.,*

$$\Phi_\theta \supset \Psi_\theta, \quad \forall\theta \in \mathbb{R}^n.$$

4.2.1. TD(1) critic. For the TD(1) critic, we will strengthen Assumption 4.2, by adding the following condition.

Assumption 4.9. *The set $\mathbb{X}_0$ consists of a single state x^*, and $\mathbf{E}_{\theta,x^*}[\phi_\theta(x^*, U_0)] = 0$ for all $\theta \in \mathbb{R}^n$.*

The requirement that there is a single state that is hit with positive probability is quite strong, but is satisfied in many practical situations involving queueing systems, as well as for systems that have been made regenerative using the splitting techniques of [1] and [18]. The assumption that the expected value of the features at x^* is zero is automatically satisfied in the special case where $\phi_\theta = \psi$. Furthermore, for features of the form $\phi_\theta(x)$ that do not depend on u, the assumption is easily satisfied by enforcing the condition $\phi_\theta(x^*) = 0$. It is argued in [11] that besides ψ_θ, there is little benefit in using additional features that depend on u. Therefore, the assumption imposed here is not a major restriction.

5. Convergence of the critic. In this section, we analyze the convergence of the critic in the algorithms described above, under the assumptions introduced in Section 4, together with Assumption 3.3 on the step-sizes. If θ_k was held constant at some value θ, it would follow (similar to [22], which dealt with the finite case) that the critic parameters converge to some $\bar{r}(\theta)$. In our case, θ_k changes with k, but slowly, and this will allow us to show that $r_k - \bar{r}(\theta_k)$ converges to zero. To establish this, we will cast the update of the critic as a linear stochastic approximation driven by Markov noise, specifically in the form of Equation (A.1) in Appendix A. We will show that the critic update satisfies all the hypotheses of Theorem A.7 of Appendix A, and the desired result (Theorem 5.7) will follow. The assumptions of the result in Appendix A are similar to the assumptions of a result (Theorem 2) used in [22]. Therefore, the proof we present here is similar to that in [22], modulo the technical difficulties due to more general state and action spaces. We start with some notation.

For each time k, let

$$\hat{Y}_{k+1} = (\hat{X}_k, \hat{U}_k, \hat{Z}_k),$$
$$R_k = \begin{pmatrix} L\alpha_k \\ r_k \end{pmatrix},$$

for some deterministic constant $L > 0$, whose purpose will be clear later. Let $\mathcal{F}_k$ be the σ-field generated by $\{Y_l, R_l, \theta_l,\ l \leq k\}$. For $y = (x, u, z, \bar{x}, \bar{u})$, define

$$h_\theta(y) = \begin{pmatrix} Lc(x,u) \\ zc(x,u) \end{pmatrix},$$

$$G_\theta(y) = \begin{pmatrix} 1 & 0 \\ z/L & \tilde{G}_\theta(y) \end{pmatrix},$$

where

$$\tilde{G}_\theta(y) = z\left(\phi_\theta'(x,u) - (P_\theta\phi_\theta)'(x,u)\right).$$

It will be shown later that the steady-state expectation of $\tilde{G}_\theta(y)$ is positive definite. The constant L is introduced because when it is chosen small enough, we will be able to show that the steady-state expectation of $G_\theta(y)$ is also positive definite.

The update (3.1) for the critic can be written as

$$R_{k+1} = R_k + \gamma_k(h_{\theta_k}(\hat{Y}_{k+1}) - G_{\theta_k}(\hat{Y}_{k+1})R_k + \xi_k R_k),$$

which is a linear iteration with Markov-modulated coefficients and ξ_k is a martingale difference given by

$$\xi_k = \begin{bmatrix} 0 \\ \hat{Z}_k\left(\phi_{\theta_k}'(\hat{X}_{k+1}, \hat{U}_{k+1}) - (P_{\theta_k}\phi_{\theta_k}')(\hat{X}_k, \hat{U}_k)\right) \end{bmatrix}.$$

To apply Theorem A.7 to this update equation, we need to prove that it satisfies Assumptions A.1-A.6. We will verify these assumptions for the two cases $\lambda = 1$ and $\lambda < 1$ separately.

Assumption A.1 follows from our Assumption 3.3. Assumption A.2 is trivially satisfied. To verify Assumption A.4, we use the actor iteration (3.2), to identify H_{k+1} with $\Gamma(r_k)r_k'\phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})\psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})$. Because of Assumption 3.3(b), the term $\Gamma(r_k)r_k$ is bounded. Furthermore, since ψ_θ and ϕ_θ belong to $\mathcal{D}$ (Assumptions 4.4 and 4.8), Lemma 4.3(b) implies that $\mathbf{E}[|H_k|^d]$ is bounded. This, together with Assumption 3.3(a), shows that Assumption A.4 is satisfied. In the next two subsections, we will concentrate on showing that Assumptions A.3, A.5 and A.6 are satisfied.

5.1. TD(1) Critic. Define a process Z_k in terms of the process $\{X_k, U_k\}$ of Section 4.1 (in which the policy is fixed) as follows:

$$Z_0 = \phi_\theta(X_0, U_0), \quad Z_{k+1} = I\{X_{k+1} \neq x^*\}Z_k + \phi_\theta(X_{k+1}, U_{k+1}),$$

where I is the indicator function. Note that the process $\{Z_k\}$ depends on the parameter θ. Whenever we use this process inside an expectation or a probability measure, we will assume that the parameter of this process is the same as the parameter of the probability or expectation. It is easy to see that $Y_{k+1} = (X_k, U_k, Z_k)$ is a Markov chain. Furthermore, the transition kernel of this process, when the policy parameter is θ, is the same as that of $\{\hat{Y}_k\}$ when the actor parameter is fixed at θ.

Let τ be the stopping time defined by

$$\tau = \min\{k > 0 \mid X_k = x^*\}.$$

For any $\theta \in \mathbb{R}^n$, define T_θ and Q_θ by

$$T_\theta(x, u) = \mathbf{E}_{\theta,x}[\tau \mid U_0 = u],$$
$$Q_\theta(x, u) = \mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\tau-1}(c(X_k, U_k) - \bar{\alpha}(\theta)) \;\Big|\; U_0 = u\right].$$

LEMMA 5.1. *The families of functions T_θ and Q_θ both belong to $\mathcal{D}$.*

Proof. The fact that $T_\theta \in \mathcal{D}$ follows easily from the assumption that $\mathbb{X}_0 = x^*$ (Assumption 4.9), and the uniform ergodicity Assumption 4.2. Using Theorem 15.2.5 of [17], we obtain that $\mathbf{E}_{\theta,x}[Q_\theta(x, U_0)]^d \leq K'_d L(x)$ for some $K'_d > 0$, so that $\mathbf{E}_{\theta,x}[Q_\theta(x, U_0)]^d$ also belongs to $\mathcal{D}$. Since

$$Q_\theta(x, u) = c(x, u) - \bar{\alpha}(\theta) + \mathbf{E}_{\theta,x}[Q_\theta(X_1, U_1)|U_0 = u]$$

is a sum of elements of $\mathcal{D}$, it follows that Q_θ also belongs to $\mathcal{D}$. $\square$

Using simple algebraic manipulations and Assumption 4.9, we obtain, for every $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}\Big(\big(c(X_k, U_k) - \bar{\alpha}(\theta)\big)Z_k - \langle Q_\theta, \phi_\theta\rangle_\theta\Big)\right] = 0,$$

$$\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}\Big(Z_k\big(\phi'_\theta(X_k, U_k) - \phi'_\theta(X_{k+1}, U_{k+1})\big) - \langle \phi_\theta, \phi'_\theta\rangle_\theta\Big)\right] = 0.$$

This implies that the steady-state expectations of $h_\theta(y)$ and $G_\theta(y)$ are given by

$$\bar{h}(\theta) = \begin{pmatrix} L\bar{\alpha}(\theta) \\ \bar{h}_1(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix},$$
$$\bar{G}(\theta) = \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/L & \bar{G}_1(\theta) \end{pmatrix},$$

where

$$\bar{h}_1(\theta) = \langle Q_\theta, \phi_\theta \rangle_\theta, \qquad \bar{Z}(\theta) = \langle T_\theta, \phi_\theta \rangle_\theta, \qquad \bar{G}_1(\theta) = \langle \phi_\theta, \phi'_\theta \rangle_\theta.$$

For $y = (x, u, z)$, we define

$$\hat{h}_\theta(y) = \mathbf{E}_{\theta,\bar{x}} \left[\sum_{k=0}^{\tau-1} (h_\theta(Y_k) - \bar{h}(\theta)) \;\Big|\; Y_0 = y \right],$$
$$\hat{G}_\theta(y) = \mathbf{E}_{\theta,\bar{x}} \left[\sum_{k=0}^{\tau-1} (G_\theta(Y_k) - \bar{G}(\theta)) \;\Big|\; Y_0 = y \right],$$

and it can be easily verified that part (a) of Assumption A.3 is satisfied. Note that we have been working with families of functions that belong to $\mathcal{D}$, and which therefore have steady-state expectations that are bounded functions of θ (Lemma 4.3(e)). In particular, $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ are bounded, and part (b) of Assumption A.3 is satisfied.

To verify the other parts of Assumption A.3, we will need the following result.

LEMMA 5.2. *For every $d > 1$, $\sup_k \mathbf{E}[|\hat{Z}_k|^d] < \infty$.*

Proof. Let $\hat{W}_k$ denote the vector $(\hat{X}_k, \hat{U}_k, \hat{Z}_k, r_k, \alpha_k, \theta_k)$. Since the step-size sequences $\{\gamma_k\}$ and $\{\beta_k\}$ are deterministic, $\{\hat{W}_k\}$ forms a time-varying Markov chain. For each k, let $\mathbf{P}_{k,\hat{w}}$ denote the conditional law of the process $\{\hat{W}_n\}$ given that $\hat{W}_k = \hat{w}$. Define a sequence of stopping times for the process $\{\hat{W}_n\}$ by letting

$$\hat{\tau}_k = \min\{n > k : \hat{X}_n = x^*\}.$$

For $1 < t < 1/\rho$, define

$$V_k^{(d)}(\hat{w}) = \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) \right],$$

which can be verified to be finite, due to uniform geometric ergodicity and the assumption that ϕ_θ belongs to $\mathcal{D}$. It is easy to see that $V_k^{(d)}(\hat{W}_k) \geq |\hat{Z}_k|^d$. Therefore, it is sufficient to prove that $\mathbf{E}[V_k^{(d)}(\hat{W}_k)]$ is bounded.

We will now show that $V_k^{(d)}(\hat{w})$ acts as a Lyapunov function for the algorithm. Indeed,

$$\begin{aligned}
V_k^{(d)}(\hat{w}) &\geq \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k+1}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) \right] \\
&= \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k+1}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) I\{\hat{X}_{k+1} \neq x^*\} \right] \\
&= t\mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) I\{\hat{X}_{k+1} \neq x^*\} \right] \\
&= t\mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) \right] - t\mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) I\{\hat{X}_{k+1} = x^*\} \right].
\end{aligned}$$

Using the geometric ergodicity condition (4.2), some algebraic manipulations, and the fact that ϕ_θ belongs to $\mathcal{D}$, we can verify that $\mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_1) I\{\hat{X}_1 = x^*\} \right]$ is bounded by some constant C. We take expectations of both sides of the preceding inequality, with $\hat{w}$ distributed as the random variable $\hat{W}_k$, and use the property

$$\mathbf{E}\Big[\mathbf{E}_{k,\hat{W}_k}\big[V_{k+1}^{(d)}(\hat{W}_{k+1})\big]\Big] = \mathbf{E}[V_{k+1}^{(d)}(\hat{W}_{k+1})],$$

we obtain

$$\mathbf{E}[V_k^{(d)}(\hat{W}_k)] \geq t\mathbf{E}[V_{k+1}^{(d)}(\hat{W}_{k+1})] - C.$$

Since $t > 1$, $\mathbf{E}[V_k^{(d)}(\hat{W}_k)]$ is bounded, and the result follows. $\square$

To verify part (c) of Assumption A.3, note that $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot), h_\theta(\cdot)$ and $G_\theta(\cdot)$ are affine in z, of the form

$$f_\theta^{(1)}(\cdot) + z f_\theta^{(2)}(\cdot),$$

for some functions $f_\theta^{(i)}$ that belong to $\mathcal{D}$. Therefore, Holder's inequality and Lemma 5.2 can be used to verify part (c) of Assumption A.3. As in the proof of Theorem 4.6, likelihood ratio methods can be used to verify Assumptions parts (d) and (e) of Assumption A.3; see [15] for details. Assumption A.5 follows from Holder's inequality, Lemma 5.2, and part (b) of Lemma 4.3.

Finally, the following lemma verifies Assumption A.6.

LEMMA 5.3. *There exist L and $\epsilon > 0$ such that for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$,*

$$R'\bar{G}(\theta)R \geq \epsilon |R|^2.$$

Proof. Let $R = (\alpha, r)$, where $\alpha \in \mathbb{R}$ and $r \in \mathbb{R}^m$. Using the definition of $\bar{G}(\theta)$, and Assumption 4.8(c) for the first inequality, we have

$$\begin{aligned} R'\bar{G}(\theta)R &= ||r'\phi_\theta||_\theta^2 + |\alpha|^2 + r'\bar{Z}(\theta)\alpha/L \\ &\geq a|r|^2 + |\alpha|^2 - r'\bar{Z}(\theta)\alpha/L \\ &\geq \min(a,1)|R|^2 - |\bar{Z}(\theta)|(|r|^2 + |\alpha|^2)/2L \\ &= \left(\min(a,1) - \frac{\bar{Z}(\theta)}{2L}\right)|R|^2. \end{aligned}$$

We can now choose $L > \sup_\theta |\bar{Z}(\theta)|/\min(a,1)$, which is possible because $\bar{Z}(\theta)$ is bounded (it is the steady-state expectation of a function in $\mathcal{D}$). $\square$

5.2. TD(λ) Critic. To analyze the TD(λ) critic, with $0 < \lambda < 1$, we redefine the process Z_k as

$$Z_{k+1} = \lambda Z_k + \phi_\theta(X_{k+1}, U_{k+1}).$$

As in the case of TD(1), we consider the steady-state expectations

$$\bar{h}(\theta) = \begin{pmatrix} L\bar{\alpha}(\theta) \\ \bar{h}_1(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix}, \qquad \bar{G}(\theta) = \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/L & \bar{G}_1(\theta) \end{pmatrix},$$

of $h_\theta(Y_k)$ and $G_\theta(Y_k)$. For the present case, the entries of $\bar{h}$ and $\bar{G}$ are given by

$$\bar{h}_1(\theta) = \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k c - \bar{\alpha}(\theta)\underline{1}, \phi_\theta \rangle_\theta,$$

$$\bar{G}_1(\theta) = \langle \phi_\theta, \phi'_\theta \rangle_\theta - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^{k+1} \phi_\theta, \phi'_\theta \rangle_\theta,$$

and $\bar{Z}(\theta) = (1-\lambda)^{-1} \langle \underline{1}, \phi_\theta \rangle_\theta$. As in Assumption 4.8(c), let $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta \underline{1}$. Then, $P_\theta \phi_\theta - \phi_\theta = P_\theta \hat{\phi}_\theta - \hat{\phi}_\theta$, and $\bar{G}_1(\theta)$ can also be written as

$$\bar{G}_1(\theta) = \langle \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^{k+1} \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta.$$

By an argument similar to the one used for the case of TD(1), we can see that $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ are bounded and, therefore, part (b) of Assumption A.3 is satisfied.

LEMMA 5.4. *There exists a positive constant C, such that for all $k \geq 0$, θ, x, λ, we have*

(a) $\left| \mathbf{E}_{\theta,x}\left[(c(X_k, U_k) - \bar{\alpha}(\theta)) Z_k \right] - \bar{h}_1(\theta) \right| \leq Ck \max(\lambda, \rho)^k L(x)$,

(b) $\left| \mathbf{E}_{\theta,x}\left[Z_k (\phi'_\theta(X_k, U_k) - \phi'_\theta(X_{k+1}, U_{k+1})) \right] - \bar{G}(\theta) \right| \leq Ck \max(\lambda, \rho)^k L(x)$.

Proof. We have

$$\begin{aligned}
&\left| \mathbf{E}_{\theta,x}\left[(c(X_k, U_k) - \bar{\alpha}(\theta)) Z_k \right] - \bar{h}_1(\theta) \right| \\
&\quad \leq \sum_{l=0}^{k} \lambda^l \left| \mathbf{E}_{\theta,x}\left[(c(X_k, U_k) - \bar{\alpha}(\theta)) \phi_\theta(X_{k-l}, U_{k-l}) \right] - \langle P_\theta^l c - \bar{\alpha}(\theta)\underline{1}, \phi_\theta \rangle_\theta \right| \\
&\qquad + C' \lambda^k \\
&\quad \leq \sum_{l=0}^{k} C' \lambda^l \rho^{k-l} L(x) + C' \lambda^k L(x) \\
&\quad \leq \sum_{l=0}^{k} 2C' \max(\lambda, \rho)^k L(x),
\end{aligned}$$

where the second inequality makes use of Lemma 4.3(e) and the assumption $L(x) \geq 1$. This proves part (a). The proof of part (b) is similar. □

From the previous lemma, it is clear that for $\theta \in \mathbb{R}^n$ and $y = (x, u, z)$,

$$\hat{h}_\theta(y) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x} \left[(h_\theta(Y_k) - \bar{h}(\theta)) \; \middle| Y_0 = y \right],$$
$$\hat{G}_\theta(y) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x} \left[(G_\theta(Y_k) - \bar{G}(\theta)) \; \middle| Y_0 = y \right],$$

are well-defined, and it is easy to check that part 1 of Assumption A.3 is satisfied.

To verify part (c) of Assumption A.3, we have the following counterpart of Lemma 5.2.

LEMMA 5.5. *For every $d > 1$, we have* $\sup_k \mathbf{E}[|\hat{Z}_k|^d] < \infty$.

Proof. We have, using Jensen's inequality,

$$\begin{aligned}
|\hat{Z}_k|^d &= \frac{1}{(1-\lambda)^d} \left| (1-\lambda) \sum_{l=0}^{k} \lambda^{k-l} \phi_{\theta_k}(\hat{X}_k, \hat{U}_k) \right|^d \\
&\leq \frac{1}{(1-\lambda)^d} (1-\lambda) \sum_{l=0}^{k} \lambda^{k-l} \left| \phi_{\theta_k}(\hat{X}_k, \hat{U}) \right|^d.
\end{aligned}$$

We note that $\mathbf{E}[|\phi_{\theta_k}(\hat{X}_k, \hat{U}_k)|^d]$ is bounded (Lemma 4.3(b)), from which it follows that $\mathbf{E}[|\hat{Z}_k|^d]$ is bounded. $\square$

The verification of parts (d) and (e) of Assumption A.3 is tedious and we only provide an outline (see [15] for the details). The idea is to write the components of $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot)$ that are linear in z in the form

$$\sum_{k=0}^{\infty} \lambda^k \mathbf{E}_{\theta,x}[f_\theta(Y_k) \mid U_0 = u, Z_0 = z],$$

for suitably defined functions f_θ, and show that the map $\theta \mapsto \mathbf{E}_\theta[f_\theta(Y_k) \mid U_0 = u, Z_0 = z]$ is Lipschitz continuous, with Lipschitz constant at most polynomial in k. The "forgetting" factor λ^k dominates the polynomial in k, and thus the sum will be Lipschitz continuous in θ. Assumption A.5 follows from Holder's inequality, the previous lemma and part (b) of Lemma 4.3. For the components that are not linear in z, likelihood ratio methods are used.

Finally, we will verify Assumption A.6 in the following lemma.

LEMMA 5.6. *There exist L and $\epsilon > 0$ such that for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$,*

$$R'\bar{G}(\theta)R \geq \epsilon |R|^2.$$

Proof. Recall the definition $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta \underline{1}$ of $\hat{\phi}_\theta$. Using Lemma 4.3(e) and the fact $\pi_\theta(\hat{\phi}_\theta) = 0$, we obtain, for some constant C,

$$||P_\theta^k \hat{\phi}_\theta^j||_\theta \leq C\rho^k, \qquad \forall \theta,\ k.$$

Therefore, for any $r \in \mathbb{R}^m$, we have

$$\begin{aligned}
\left\| P_\theta^k \left(r'\hat{\phi}_\theta\right) \right\|_\theta &= \left\| \sum_j r_j P_\theta^k \hat{\phi}_\theta^j \right\|_\theta \\
&\leq \sum_j |r_j| \cdot ||P_\theta^k \hat{\phi}_\theta^j||_\theta \\
&\leq C_1 \rho^k |r|.
\end{aligned}$$

We note that the transition operator P_θ is nonexpanding, i.e., $\|P_\theta f\|_\theta \leq \|f\|_\theta$, for every $f \in \mathcal{L}^2(\eta_\theta)$; see, e.g., [21]. Using this property and some algebraic manipulations, we obtain

$$\begin{aligned}
r'\bar{G}_1(\theta)r &= r'\langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta r - (1-\lambda)\sum_{k=0}^{\infty} \lambda^k r' \langle P_\theta^k \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta r \\
&= ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda)\sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k(r'\hat{\phi}_\theta), r'\hat{\phi}_\theta \rangle_\theta \\
&\geq ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda)\left\{ \sum_{k=0}^{k_0-1} \lambda^k ||r'\hat{\phi}_\theta||_\theta^2 + \sum_{k \geq k_0} C_1 \lambda^k \rho^k ||r'\hat{\phi}_\theta||_\theta |r| \right\} \\
&\geq ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda^{k_0})||r'\phi_\theta||_\theta^2 - C_1(\lambda\rho)^{k_0} \frac{(1-\lambda)}{(1-\rho\lambda)} ||(r'\phi_\theta)||_\theta |r| \\
&\geq |r|^2 \lambda^{k_0} \left(a - \frac{C_2 \rho^{k_0}(1-\lambda)}{(1-\rho\lambda)} \right),
\end{aligned}$$

where the last step made use of the uniform positive definiteness property (Assumption 4.8(c)). We choose k_0 so that

$$\rho^{k_0} < \frac{a(1-\rho\lambda)}{C_2(1-\lambda)},$$

and conclude that $\bar{G}_1(\theta)$ is uniformly positive definite. From this point on, the proof is identical to the proof of Lemma 5.3. □

Having verified all the hypotheses of Theorem A.7, we have proved the following result.

THEOREM 5.7. *Under Assumptions 3.3 and 4.1-4.9, and for any TD critic, the sequence R_k is bounded, and* $\lim_k |\bar{G}(\theta_k)R_k - \bar{h}(\theta_k)| = 0$.

6. Convergence of the Actor. For every $\theta \in \mathbb{R}^n$ and $(x,u) \in \mathbb{X} \times \mathbb{U}$, let

$$H_\theta(x,u) = \psi_\theta(x,u)\phi'_\theta(x,u), \qquad \bar{H}(\theta) = \langle \psi_\theta, \phi'_\theta \rangle_\theta.$$

Note that H_θ belongs to $\mathcal{D}$, and consequently $\bar{H}(\theta)$ is bounded. Let $\bar{r}(\theta)$ be such that $\bar{h}_1(\theta) = \bar{G}_1(\theta)\bar{r}(\theta)$, so that $\bar{r}(\theta)$ is the limit of the critic parameter r if the policy parameter θ was held fixed. Then, the recursion for the actor parameter θ can be written as The recursion for the actor parameter θ can be written as

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \beta_k H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k\Gamma(r_k)) \\
&= \theta_k - \beta_k \bar{H}(\theta_k)\left(\bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))\right) \\
&\quad -\beta_k(H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k))(r_k\Gamma(r_k)) \\
&\quad -\beta_k\bar{H}(\theta_k)(r_k\Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))).
\end{aligned}$$

Let

$$\begin{aligned}
f(\theta) &= \bar{H}(\theta)\bar{r}(\theta), \\
e_k^{(1)} &= (H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k))r_k\Gamma(r_k), \\
e_k^{(2)} &= \bar{H}(\theta_k)(r_k\Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))).
\end{aligned}$$

Using Taylor's series expansion, one can see that

$$\begin{aligned}
\bar{\alpha}(\theta_{k+1}) \le{}& \bar{\alpha}(\theta_k) - \beta_k\Gamma(\bar{r}(\theta))\nabla\bar{\alpha}(\theta_k)\cdot f(\theta_k) - \beta_k\nabla\bar{\alpha}(\theta_k)\cdot e_k^{(1)} \\
&-\beta_k\nabla\bar{\alpha}(\theta_k)\cdot e_k^{(2)} + C\beta_k^2\left|H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k\Gamma(r_k))\right|^2,
\end{aligned} \tag{6.1}$$

where C reflects a bound on the Hessian of $\bar{\alpha}(\theta)$.

Note that $\bar{r}(\theta)$ and $f(\theta)$ depend on the parameter λ of the critic. The following lemma characterizes this dependence.

LEMMA 6.1. *If a TD(λ) critic is used, with* $0 < \lambda \le 1$*, then* $f(\theta) = \nabla\bar{\alpha}(\theta) + \varepsilon(\lambda,\theta)$*, where* $\sup_\theta |\varepsilon(\lambda,\theta)| \le C(1-\lambda)$*, and where the constant* $C > 0$ *is independent of* λ.

Proof. Consider first the case of a TD(1) critic. By definition, $\bar{r}(\theta)$ is the solution to the linear equation $\bar{G}_1(\theta)\bar{r}(\theta) = \bar{h}_1(\theta)$, or

$$\langle \phi_\theta, \phi'_\theta\bar{r}(\theta)\rangle_\theta = \langle \phi_\theta, Q_\theta\rangle_\theta.$$

Thus, $\phi'_\theta\bar{r}(\theta) - Q_\theta$ is orthogonal to ϕ_θ, in $\mathcal{L}^2(\eta_\theta)$. By Assumption 4.8(d), the components of ψ_θ are contained in the subspace spanned by the components of ϕ_θ. It follows that $\phi'_\theta\bar{r}(\theta) - Q_\theta$ is also orthogonal to ψ_θ. Therefore,

$$\bar{H}(\theta)\bar{r}(\theta) = \langle \psi_\theta, \phi'_\theta\rangle_\theta\bar{r}(\theta) = \langle \psi_\theta, Q_\theta\rangle_\theta = \nabla\bar{\alpha}(\theta),$$

where the last equality is the gradient formula in Theorem 4.6.

For $\lambda < 1$, let us write $\bar{G}_1^\lambda(\theta)$ and $\bar{h}_1^\lambda(\theta)$ for $\bar{G}_1(\theta)$ and $\bar{h}_1(\theta)$, defined in Section 5.2, to show explicitly the dependence on λ. Let $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta \underline{1}$. Then, it is easy to see that

$$|\bar{G}_1^\lambda(\theta) - \langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta| = (1-\lambda) \left| \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k \hat{\phi}_\theta, \hat{\phi}_\theta \rangle_\theta \right| \leq C \left(\frac{1-\lambda}{1-\rho\lambda} \right),$$

where the inequality follows from the geometric ergodicity condition (4.3). Similarly, one can also see that $|\bar{h}_1^\lambda(\theta) - \langle Q_\theta, \hat{\phi}_\theta \rangle_\theta| \leq C(1-\lambda)$. Let $\bar{r}(\theta)$ and $\bar{r}^\lambda(\theta)$ be solutions of the linear equations $\langle \hat{\phi}_\theta, \hat{\phi}_\theta' r \rangle_\theta = \langle Q_\theta, \phi_\theta \rangle_\theta$ and $\bar{G}_1^\lambda(\theta) r = \bar{h}_1^\lambda(\theta)$, respectively. Then,

$$\langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta (\bar{r}(\theta) - \bar{r}^\lambda(\theta)) = (\bar{h}_1(\theta) - \bar{h}_1^\lambda(\theta)) + (\bar{G}_1^\lambda(\theta) - \langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta) \bar{r}^\lambda(\theta),$$

which implies that $|\bar{r}(\theta) - \bar{r}^\lambda(\theta)| \leq C(1-\lambda)$. The rest follows from the observation that $\bar{H}(\theta)\bar{r}(\theta) = \nabla \bar{\alpha}(\theta)$. □

Lemma 6.2. *(Convergence of the noise terms)*

(a) $\sum_{k=0}^{\infty} \beta_k \nabla \bar{\alpha}(\theta_k) \cdot e_k^{(1)}$ *converges w.p.1.*

(b) $\lim_k e_k^{(2)} = 0$ *w.p.1.*

(c) $\sum_k \beta_k^2 |H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)|^2 < \infty$ *w.p.1.*

Proof. Since r_k is bounded and $\Gamma(\cdot)$ satisfies the condition (3.3), it is easy to see that $r\Gamma(r)$ is bounded and $|r\Gamma(r) - \hat{r}\Gamma(\hat{r})| < C|r - \hat{r}|$ for some constant C. The proof of part (a) is now similar to the proof of Lemma 2, in p. 224 of [3]. Part (b) follows from Theorem 5.7 and the fact that $\bar{H}(\cdot)$ is bounded. Part (c) follows from the inequality

$$|H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)| \leq C |H_{\theta_k}(\hat{X}_k, \hat{U}_k)|$$

for some $C > 0$ and the boundedness of $\mathbf{E}\left[|H_{\theta_k}(\hat{X}_k, \hat{U}_k)|^2\right]$ (part (b) of Lemma 4.3). □

Theorem 6.3. (Convergence of Actor-Critic algorithms) *Let Assumptions 3.3 and 4.1-4.9 hold.*

(a) If a TD(1) critic is used, then $\liminf_k |\nabla \bar{\alpha}(\theta_k)| = 0$, *w.p.1.*

(b) For any $\epsilon > 0$, *there exists some* λ *sufficiently close to 1, so that the algorithm that uses a TD(*λ*) critic (with* $0 < \lambda < 1$*) satisfies* $\liminf_k |\nabla \bar{\alpha}(\theta_k)| < \epsilon$, *w.p.1.*

Proof. The proof is standard [24], and we only provide an outline. Fix some $T > 0$, and define a sequence k_j by

$$k_0 = 0, \qquad k_{j+1} = \min\left\{ k \geq k_j \;\middle|\; \sum_{l=k_j}^{k} \beta_k \geq T \right\} \quad \text{for} \quad j > 0.$$

Using Eq. (6.1), we have

$$\bar{\alpha}(\theta_{k_{j+1}}) \leq \bar{\alpha}(\theta_{k_j}) - \sum_{k=k_j}^{k_{j+1}-1} \beta_k \left(|\nabla \bar{\alpha}(\theta_k)|^2 - C(1-\lambda)|\nabla \bar{\alpha}(\theta_k)| \right) + \delta_j,$$

where δ_j is defined by

$$\delta_j = \sum_{k=k_j}^{k_{j+1}-1} \left(\beta_k \nabla \bar{\alpha}(\theta_k) \cdot (e_k^{(1)} + e_k^{(2)}) + C\beta_k^2 |H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)|^2 \right).$$

Lemma 6.2 implies that δ_j goes to zero. If the result fails to hold, it can be shown that the sequence $\bar{\alpha}(\theta_k)$ would decrease indefinitely, contradicting the boundedness of $\bar{\alpha}(\theta)$. The result follows easily. $\square$

Appendix A. A result on linear stochastic approximation.

We recall the following result from [14]. Consider a stochastic process $\{\hat{Y}_k\}$ taking values in a Polish space $\mathbb{Y}$ with Borel σ-field denoted by $\mathcal{B}(\mathbb{Y})$. Let $\{P_\theta(y, d\bar{y}); \theta \in \mathbb{R}^n\}$ be a parameterized family of transition kernels on $\mathbb{Y}$. Consider the following iterations to update a vector $R \in \mathbb{R}^m$ and the parameter $\theta \in \mathbb{R}^n$:

$$\begin{aligned} R_{k+1} &= R_k + \gamma_k(h_{\theta_k}(\hat{Y}_{k+1}) - G_{\theta_k}(\hat{Y}_{k+1})R_k + \xi_{k+1}R_k), \\ \theta_{k+1} &= \theta_k + \beta_k H_{k+1}. \end{aligned} \tag{A.1}$$

In the above iteration, $\{h_\theta(\cdot), G_\theta(\cdot) : \theta \in \mathbb{R}^n\}$ is a parameterized family of m-vector valued and $m \times m$-matrix valued measurable functions on $\mathbb{Y}$. We introduce the following assumptions.

ASSUMPTION A.1. *The step-size sequence $\{\gamma_k\}$ is deterministic, non-increasing, and satisfies*

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

Let $\mathcal{F}_k$ be the σ-field generated by $\{\hat{Y}_l, H_l, r_l, \theta_l, l \le k\}$.

ASSUMPTION A.2. *For a measurable set $A \subset \mathbb{Y}$,*

$$\mathbf{P}(\hat{Y}_{k+1} \in A \mid \mathcal{F}_k) = \mathbf{P}(\hat{Y}_{k+1} \in A \mid \hat{Y}_k, \theta_k) = P_{\theta_k}(\hat{Y}_k, A).$$

For any measurable function f on $\mathbb{Y}$, let $P_\theta f$ denote the measurable function $y \mapsto \int P_\theta(y, d\bar{y}) f(\bar{y})$, and for any vector r, let $|r|$ denote its Euclidean norm.

ASSUMPTION A.3. (Existence and properties of solutions to the Poisson Equation) *For each θ, there exist functions $\bar{h}(\theta) \in \mathbb{R}^m$, $\bar{G}(\theta) \in \mathbb{R}^{m \times m}$, $\hat{h}_\theta : \mathbb{Y} \to \mathbb{R}^m$, and $\hat{G}_\theta : \mathbb{Y} \to \mathbb{R}^{m \times m}$ that satisfy the following:*

(a) For each $y \in \mathbb{Y}$,

$$\begin{aligned} \hat{h}_\theta(y) &= h_\theta(y) - \bar{h}(\theta) + (P_\theta \hat{h}_\theta)(y), \\ \hat{G}_\theta(y) &= G_\theta(y) - \bar{G}(\theta) + (P_\theta \hat{G}_\theta)(y). \end{aligned}$$

(b) For some constant C and for all θ, we have

$$\max(|\bar{h}(\theta)|, |\bar{G}(\theta)|) \le C.$$

(c) For any $d > 0$, there exists $C_d > 0$ such that

$$\sup_k \mathbf{E}[|f_{\theta_k}(\hat{Y}_k)|^d] \le C_d,$$

where $f_\theta(\cdot)$ represents any of the functions $\hat{h}_\theta(\cdot), h_\theta(\cdot), \hat{G}_\theta(\cdot), G_\theta(\cdot)$.

(d) For some constant $C > 0$, and for all $\theta, \bar{\theta} \in \mathbb{R}^n$,

$$\max(|\bar{h}(\theta) - \bar{h}(\bar{\theta})|, |\bar{G}(\theta) - \bar{G}(\bar{\theta})|) \le C|\theta - \bar{\theta}|.$$

(e) *There exists a positive measurable function $C(\cdot)$ on $\mathbb{Y}$ such that for each $d > 0$,*

$$\sup_k \mathbf{E}[C(\hat{Y}_k)^d] < \infty,$$

and

$$|P_\theta f_\theta(y) - P_{\bar{\theta}} f_{\bar{\theta}}(y)| \leq C(y)|\theta - \bar{\theta}|,$$

where $f_\theta(\cdot)$ is any of the functions $\hat{h}_\theta(\cdot)$ and $\hat{G}_\theta(\cdot)$.

ASSUMPTION A.4. (Slowly changing environment) *The (random) process $\{H_k\}$ satisfies*

$$\sup_k \mathbf{E}\left[|H_k|^d\right] < \infty,$$

for all $d > 0$. Furthermore, the sequence $\{\beta_k\}$ is deterministic and satisfies

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty,$$

for some $d > 0$.

ASSUMPTION A.5. *The sequence $\{\xi_k\}$ is an $m \times m$-matrix valued $\mathcal{F}_k$-martingale difference, with bounded moments, i.e.,*

$$\mathbf{E}\left[\xi_{k+1}|\mathcal{F}_k\right] = 0, \qquad \sup_k \mathbf{E}\left[|\xi_{k+1}|^d\right] < \infty$$

for each $d > 0$.

ASSUMPTION A.6. (Uniform positive definiteness) *There exists $a > 0$ such that for all $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$,*

$$r'\bar{G}(\theta)r \geq a|r|^2.$$

THEOREM A.7. *If Assumptions A.1-A.6 are satisfied, then the sequence R_k is bounded and*

$$\lim_k \left|R_k - \bar{G}(\theta_k)^{-1}\bar{h}(\theta_k)\right| = 0.$$

REFERENCES

[1] K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, 245:493–501, 1978.

[2] A. Barto, R. Sutton, and C. Anderson. Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, 13:835–846, 1983.

[3] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin-Heidelberg, 1990.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 1995.

[5] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[6] V. S. Borkar. Stochastic approximation with two time scales. *Systems and Control Letters*, 29:291–294, 1996.

[7] X. R. Cao and H. F. Chen. Perturbation realization, potentials, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, 42:1382–1393, 1997.

[8] P. W. Glynn. Stochastic approximation for Monte Carlo optimization. In *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

[9] P. W. Glynn and P. L'Ecuyer. Likelihood ratio gradient estimation for stochastic recursions. *Advances in applied probability*, 27:1019–1053, 1995.

[10] T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithms for partially observable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 7, pages 345–352, San Francisco, CA, 1995. Morgan Kaufman.

[11] V. R. Konda. *Actor-Critic Algorithms*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.

[12] V. R. Konda and V. S. Borkar. Actor-critic like learning algorithms for Markov decision processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

[13] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12, pages 1008-1014, 2000.

[14] V. R. Konda and J. N. Tsitsiklis. Linear stochastic approximation driven by slowly varying Markov chains. Submitted, June 2002.

[15] V. R. Konda and J. N. Tsitsiklis. Appendix to 'Actor-critic methods'. `http://web.mit.edu/jnt/www/Papers.html/actor-app.pdf`, July 2002.

[16] P. Marbach and J. N. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, February 2001.

[17] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[18] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrscheinlichkeitstheorie and Verw. Geb.*, 43:119–143, 1978.

[19] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[20] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063, 2000.

[21] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[22] J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

[23] R. Williams. Simple statistical gradient following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[24] B. T. Polyak. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 34(3):377-397, 1973.