Analysis of a Target-Based Actor-Critic Algorithm
with Linear Function Approximation

Anas Barakat  Pascal Bianchi  Julien Lehmann
LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Abstract

Actor-critic methods integrating target networks have exhibited a stupendous empirical success in deep reinforcement learning. However, a theoretical understanding of the use of target networks in actor-critic methods is largely missing in the literature. In this paper, we reduce this gap between theory and practice by proposing the first theoretical analysis of an online target-based actor-critic algorithm with linear function approximation in the discounted reward setting. Our algorithm uses three different timescales: one for the actor and two for the critic. Instead of using the standard single timescale temporal difference (TD) learning algorithm as a critic, we use a two timescales target-based version of TD learning closely inspired from practical actor-critic algorithms implementing target networks. First, we establish asymptotic convergence results for both the critic and the actor under Markovian sampling. Then, we provide a finite-time analysis showing the impact of incorporating a target network into actor-critic methods.

1 INTRODUCTION

Actor-critic algorithms [Barto et al., 1983, Konda and Borkar, 1999, Konda and Tsitsiklis, 2003b, Peters and Schaal, 2008, Bhatnagar et al., 2009] are a class of reinforcement learning (RL) [Sutton and Barto, 2018, Bertsekas and Tsitsiklis, 1996] methods to find an optimal policy maximizing the total expected reward in a stochastic environment modelled by a Markov Decision Process (MDP) [Puterman, 2014]. In this type of algorithms, two main processes interplay: the actor and the critic. The actor updates a parameterized policy in a direction of performance improvement whereas the critic estimates the current policy of the actor by estimating the unknown state-value function. In turn, the critic estimation is used to produce the update rule of the actor. Combined with deep neural networks as function approximators of the value function, actor-critic algorithms witnessed a tremendous success in a range of challenging tasks [Heess et al., 2015, Lillicrap et al., 2016, Mnih et al., 2016, Fujimoto et al., 2018, Haarnoja et al., 2018]. Apart from using neural networks for function approximation (FA), one of the main features underlying their remarkable empirical achievements is the use of target networks for the critic estimation of the value function. Introduced by the seminal work of Mnih et al. [Mnih et al., 2015] to stabilize the training process, this target innovation consists in using two neural networks maintaining two copies of the estimated value function: A so-called target network tracking a main network with some delay computes the target values for the value function update.

Despite their resounding empirical success in deep RL, a theoretical understanding of the use of target networks in actor-critic methods is largely missing in the literature. Theoretical contributions investigating the use of a target network are very recent and limited to temporal difference (TD) learning for policy evaluation [Lee and He, 2019] and critic-only methods such as Q-learning for control [Zhang et al., 2021]. In particular, these works are not concerned with actor-critic algorithms and leave the question of the finite-time analysis open.

In the present work, we reduce this gap between theory and practice by proposing the first theoretical analysis of an online target-based actor-critic algorithm in the discounted reward setting. We consider the linear FA setting where a linear combination of pre-selected feature (or basis) functions estimates the value function...
in the critic. An analysis of this setting is an insightful first step before tackling the more challenging nonlinear FA setting aligned with the use of neural networks. We conduct our study in the multiple timescales framework. In the standard two timescales actor-critic algorithms [Konda and Tsitsiklis, 2003b, Bhatnagar et al., 2009], at each iteration, the actor and the critic are updated simultaneously but the critic evolves faster than the actor which uses smaller step sizes. We face two main challenges due to the integration of the target variable mechanism. First, in contrast to standard two timescales actor-critic algorithms, our algorithm uses three different timescales: one for the actor and two for the critic. Instead of using the single timescale TD learning algorithm as a critic, we use a two timescales target-based version of TD learning closely inspired from practical actor-critic algorithms implementing target networks. Second, incorporating a target variable into the critic results in the intricate interplay between three processes evolving on three different timescales. In particular, the use of a target variable significantly modifies the dynamics of the actor-critic algorithm and deserves a careful analysis accordingly.

Our main contributions are summarized as follows. First, we prove asymptotic convergence results for both the critic and the actor. More precisely, as the actor parameter changes slowly compared to the critic one, we show that the critic using a target variable tracks a slowly moving target corresponding to a TD-like solution [Tsitsiklis and Van Roy, 1997]. Our development is based on the ordinary differential equation (ODE) method of stochastic approximation (see, for e.g., [Benveniste et al., 1990, Borkar, 2008]). Then, we show that the actor parameter visits infinitely often a region of the parameter space where the norm of the policy gradient is dominated by a bias due to linear FA. Second, we conduct a finite-time analysis of our actor-critic algorithm which shows the impact of using a target variable on the convergence rates and the sample complexity. Loosely speaking, up to a FA error, we show that our target-based algorithm converges in expectation to an $\epsilon$-approximate stationary point of the non-concave performance function using at most $O(\epsilon^{-3} \ln^3(1/\epsilon))$ samples compared with $O(\epsilon^{-2} \ln(1/\epsilon))$ for the best known complexity for two timescales actor-critic algorithms without a target network. All the proofs are deferred to the appendix.

2 RELATED WORK

In this section, we briefly discuss the most relevant related works to ours. Existing theoretical results in the literature can be divided into two classes.

Asymptotic results. Almost sure convergence results are referred to as asymptotic. Konda & Tsitsiklis [Konda and Tsitsiklis, 2003b] [Konda, 2002] provided almost sure (with probability one) convergence results for a two timescales actor-critic algorithm in which the critic estimates the action-value function via linear FA. Our algorithm is closer to an actor-critic algorithm introduced by Bhatnagar et al. [Bhatnagar et al., 2009] in the average reward setting. However, unlike [Bhatnagar et al., 2009], we integrate a target variable mechanism into our critic and consider the discounted reward setting. Moreover, as previously mentioned, the target variable for the critic adds an additional timescale in comparison to [Konda and Tsitsiklis, 2003b, Bhatnagar et al., 2009] which only involve two different timescales. Regarding theoretical results considering target networks, Lee & He [Lee and He, 2019] proposed a family of single timescale target-based TD learning algorithms for policy evaluation. Our critic corresponds to a two timescales version of the single timescale target-based TD learning algorithm of Lee & He [Lee and He, 2019] called Averaging TD. In [Lee and He, 2019 Th. 1], this single timescale algorithm is shown to converge with probability one (w.p.1) towards the standard TD solution solving the projected Bellman equation (see [Tsitsiklis and Van Roy, 1997] for a precise statement). Besides the timescales difference with [Lee and He, 2019], in this article, we are concerned with a control setting in which the policy changes at each timestep via the actor update. Yang et al. [Yang et al., 2019] proposed a bilevel optimization perspective to analyze Q-learning with a target network and an actor-critic algorithm without any target network. More recently, Zhang et al. [Zhang et al., 2021] investigated the use of target networks in Q-learning with linear FA and a target variable with Ridge regularization. Their analysis covers the average and discounted reward settings and establishes asymptotic convergence results for policy evaluation and control. This recent work [Zhang et al., 2021] focuses on the critic-only Q-learning method with a target network update rule, showing the role of the target network in the off-policy setting. In particular, this work is not concerned with actor-critic algorithms.

Finite-time analysis. The second type of results consists in establishing time-dependent bounds on some error or performance quantities such as the average expected norm of the gradient of the performance function. These are referred to as finite-time analysis. In the last few years, several works proposed finite-time analysis for TD learning [Bhandari et al., 2018, Srikant and Ying, 2019] for two timescales TD methods [Xu et al., 2019] and even more generally for two timescales linear stochastic approximation algorithms [Gupta et al., 2019].
These works opened the way to the recent development of a flurry of nonasymptotic results for actor-critic algorithms. None of the mentioned works uses a target network. In this work, we conduct a finite-time analysis of our target-based actor-critic algorithm. Such new results are missing in all theoretical results investigating the use of a target network.

The summary table compiles some key features of the standard two timescales actor-critic algorithm [Bhatnagar et al., 2009] Algorithm 1] in the average reward setting with linear FA. Shen et al. [Shen et al., 2020] conducted a similar study for a revisited version of the asynchronous advantage actor-critic algorithm [Yang et al., 2018, Qiu et al., 2019,

For every finite set $\mathcal{S}$, we use the notation $\mathcal{P}(\mathcal{S})$ for the set of probability measures on $\mathcal{S}$. The cardinality of a finite set is denoted by $|\cdot|$. For two sequences of nonnegative reals $(x_n)$ and $(y_n)$, the notation $x_n = O(y_n)$ means that there exists a constant $C$ independent of $n$ such that $x_n \leq C y_n$ for all $n \in \mathbb{N}$. For any integer $p$, the euclidean space $\mathbb{R}^p$ is equipped with its usual inner product $\langle \cdot, \cdot \rangle$ and its corresponding 2-norm $\| \cdot \|$. For any integer $d$ and any matrix $A \in \mathbb{R}^{d \times p}$, we use the notation $\|A\|$ for the operator norm induced by the euclidean vector norm. For a symmetric positive semidefinite matrix $B \in \mathbb{R}^{p \times p}$ and a vector $x \in \mathbb{R}^p$, the notation $\|x\|_B^2$ refers to the quantity $\langle x, B x \rangle$. The transpose of the vector $x$ is denoted by $x^T$ and $I_p$ is the identity matrix.

### 3.1 Markov decision process and problem formulation

Consider the RL setting [Sutton and Barto, 2018] Bertsekas and Tsitsiklis, 1996 where a learning agent interacts with an environment modeled as an infinite horizon discrete-time discounted MDP. We denote by $\mathcal{S} = \{s_1, \cdots, s_n\}$ the finite set of states and $\mathcal{A}$ the finite set of actions. Let $p: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$ be the state transition probability kernel and $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ the immediate reward function. A randomized stationary policy, which we will simply call a policy in the rest of the paper, is a mapping $\pi: \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ specifying for each $s \in \mathcal{S}$ a probability $\pi(a|s)$ of selecting action $a$ in state $s$. At each time step $t \in \mathbb{N}$, the RL agent in a state $S_t \in \mathcal{S}$ executes an action $A_t \in \mathcal{A}$ with probability $\pi(A_t|S_t)$, transitions into a state $S_{t+1} \in \mathcal{S}$ with probability $p(S_{t+1}|S_t, A_t)$ and observes a random reward $R_{t+1} \in [-U_R, U_R]$ where $U_R$ is a positive real. We denote by $\mathbb{E}_{p,\pi}$ the probability distribution of the Markov chain $(S_t, A_t)$ issued from the MDP controlled by the policy $\pi$ with initial state distribution $\rho$. The notation $\mathbb{E}_{p,\pi}$ refers to the associated expectation. We will use $\mathbb{E}_\pi$ whenever there is no dependence on $\rho$. The sequence $(R_t)$ is such that (s.t.) $\mathbb{E}_\pi[R_{t+1}|S_t, A_t] = R(S_t, A_t)$. Let $\gamma \in (0, 1)$ be a discount factor. Given a policy $\pi$, the long-term expected cumulative discounted reward is quantified by the state-value function $V_\pi: \mathcal{S} \rightarrow \mathbb{R}$ and the action-value function $Q_\pi: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defined for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ by $V_\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s]$ and $Q_\pi(s, a) := \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s, A_0 = a]$. We also define the advantage function $\Delta_\pi: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ by $\Delta_\pi(s, a) := Q_\pi(s, a) - V_\pi(s)$. Given an initial probability distribution $\rho$ over states for the initial state $S_0$, the goal of the agent is to find a policy $\pi$ maximizing the expected long-term return $J(\pi) := \sum_{s \in \mathcal{S}} \rho(s) V_\pi(s)$. For this purpose, the agent has only access to realizations of the random variables $S_t, A_t$ and $R_t$ whereas the state transition kernel $p$ and the reward function $R$ are unknown.

### 3.2 Policy Gradient framework

From now on, we restrict the policy search to the set of policies $\pi$ parameterized by a vector $\theta \in \mathbb{R}^d$ for some integer $d > 0$ and optimize the performance criterion $J$ over this family of parameterized policies $\{\pi_\theta: \theta \in \mathbb{R}^d\}$. The policy dependent function $J$ can also be seen as a function of the parameter $\theta$. We use the notation $J(\theta)$ for $J(\pi_\theta)$ by abuse of notation. The problem that we are concerned with can be written as: $\max_{\theta \in \mathbb{R}^d} J(\theta)$.
by:
\[
\psi_\theta(s,a) := \nabla \ln \pi_\theta(a|s),
\]
where \(\nabla\) denotes the gradient w.r.t. \(\theta\). We introduce an assumption on the regularity of the parameterized family of policies which is a standard requirement in policy gradients (see, for eg., Zhang et al., 2020a, Assumption 3.1 [Konda and Tsitsiklis, 2003b, Assumption 2.1]). In particular, it ensures that \(\psi_\theta\) is well defined.

**Assumption 3.1.** The following conditions hold true for every \((s,a) \in S \times A\).

(a) For every \(\theta \in \mathbb{R}^d\), \(\pi_\theta(a|s) > 0\).
(b) The function \(\theta \mapsto \pi_\theta(a|s)\) is continuously differentiable and \(L_\psi\)-Lipschitz continuous.
(c) The function \(\theta \mapsto \psi_\theta(s,a)\) is bounded and \(L_\psi\)-Lipschitz.

Assumption 3.1 is satisfied for instance by the Gibbs (or softmax) policy and the Gaussian policy (see Zhang et al., 2020a, Assumption 3.1) and the references therein for details). Under Assumption 3.1, the policy gradient theorem [Sutton et al., 2000, Konda, 2002, Th. 2.13] with the state-value function as a baseline provides an expression for the gradient of the performance metric \(J\) w.r.t. the policy parameter \(\theta\) given by:

\[
\nabla J(\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{(S,A) \sim \mu_{\rho,\theta}} \left[ \Delta \pi_\theta(\tilde{S}, \tilde{A}) \psi_\theta(\tilde{S}, \tilde{A}) \right].
\]

Here, the couple of random variables \((\tilde{S}, \tilde{A})\) follows the discounted state-action occupancy measure \(\mu_{\rho,\theta} \in \mathcal{P}(S,A)\) defined for all \((s,a) \in S \times A\) by:

\[
\mu_{\rho,\theta}(s,a) := d_{\rho,\theta}(s) \pi_\theta(a|s)
\]

where \(d_{\rho,\theta}(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\rho,\pi_\theta}(S_t = s)\) is a probability measure over the state space \(S\) known as the discounted state-occupancy measure. Note that under Assumption 3.1, the policy gradient \(\nabla J\) is Lipschitz continuous (see Zhang et al., 2020a, Lem. 4.2).

## 4 TARGET-BASED ACTOR-CRITIC ALGORITHM

In this section, we gradually present our actor-critic algorithm.

### 4.1 Actor update

First, we need an estimate of the policy gradient \(\nabla J(\theta)\) of Eq. (1) in view of using stochastic gradient ascent to solve the maximization problem. Given Eq. (1) and following previous works, we recall how to sample according to the distribution \(\mu_{\rho,\theta}\). As described in Konda, 2002 Sec. 2.4], the distribution \(\mu_{\rho,\theta}\) is the stationary distribution of a Markov chain \((\tilde{S}_t, \tilde{A}_t)\) issued from the artificial MDP whose transition kernel \(\tilde{p} : S \times A \rightarrow \mathcal{P}(S)\) is defined for every \((s,a) \in S \times A\) by:

\[
\tilde{p}(|s,a) := \gamma \ p(|s,a) + (1 - \gamma) \rho(^\cdot), \quad (4)
\]

and which is controlled by the policy \(\pi_\theta\) generating the action sequence \((\tilde{A}_t)\). We will later state conditions to ensure its existence and uniqueness. Therefore, under suitable conditions, the distribution of the Markov chain \((\tilde{S}_t, \tilde{A}_t)\) will converge geometrically towards its stationary distribution \(\mu_{\rho,\theta}\). This justifies the following sampling procedure. Given a state \(\tilde{S}_t\) and an action \(\tilde{A}_t\), we sample a state \(\tilde{S}_{t+1}\) according to this artificial MDP by sampling from \(p(|\tilde{S}_t, \tilde{A}_t)\) with probability \(\gamma\) and from \(\rho\) otherwise. For this purpose, at each time step \(t\), we draw a Bernoulli random variable \(B_t \in \{0,1\}\) with parameter \(\gamma\) which is independent of all the past random variables generated until time \(t\).
Then, using the definition of the advantage function, Eq. (1) becomes:

\[
\nabla J(\theta) = \frac{1}{1-\gamma} \mathbb{E}[(R(\tilde{S}, \tilde{A}) + \gamma V_{\pi_\theta}(S) - V_{\pi_\theta}(\tilde{S})) \psi_\theta(\tilde{S}, \tilde{A})],
\]

where \((\tilde{S}, \tilde{A}) \sim \mu_{p, \theta}\) and \(S \sim p(\cdot|\tilde{S}, \tilde{A})\). From this equation, it is natural to define for every \(V \in \mathbb{R}^n\) the temporal difference (TD) error

\[
\delta V_{t+1} = R_{t+1} + \gamma V(S_{t+1}) - V(\tilde{S}_t),
\]

where \(S_{t+1}\) is drawn from the distribution \(p(\cdot|\tilde{S}_t, \tilde{A}_t)\) and \((\tilde{S}_t, \tilde{A}_t) \in \mathbb{N}\) is the Markov chain induced by the artificial MDP described in Eq. (4) and controlled by the policy \(\pi_\theta\). Notice here from Eq. (5) that we need two different sequences \((S_t)\) and \((\tilde{S}_t)\) respectively sampled from the kernels \(p\) and \(\tilde{p}\). In our discounted reward setting, using only the sequence \((\tilde{S}_t)\) issued from the artificial kernel \(\tilde{p}\) would result in a bias with a sampling error of the order \(1 - \gamma\) (see Shen et al., 2020 Eq. (14) and Lem. 7)).

Supposing for now that the value function \(V_{\pi_\theta}\) is known, it stems from Eq. (5) that a natural estimator of the gradient \(\nabla J(\theta)\) is \(\delta V_{t+1} \psi_\theta(\tilde{S}_t, \tilde{A}_t)/(1 - \gamma)\). This estimator is only biased because the distribution of our sampled Markov chain \((S_t, A_t)\) is not exactly \(\mu_{p, \theta}\) but converges geometrically to this one. However, the state-value function \(V_{\pi_\theta}\) is unknown. Given an estimate \(\hat{V}_{\pi_\theta} \in \mathbb{R}^n\) of \(V_{\pi_\theta}\) and a positive stepsize \(\alpha_t\), the actor updates its parameter as follows:

\[
\theta_{t+1} = \theta_t + \alpha_t \frac{1}{1 - \gamma} \delta V_{t+1} \psi_\theta(\tilde{S}_t, \tilde{A}_t).
\]

### 4.2 Critic update

The state-value function \(V_{\pi_\theta}\) is approximated for every state \(s \in S\) by a linear function of carefully chosen feature vectors as follows: \(V_{\pi_\theta}(s) \approx V_\omega(s) = \omega^T \phi(s) = \sum_{i=1}^m \omega_i \phi_i(s)\), where \(\omega = (\omega_1, \ldots, \omega_m)^T \in \mathbb{R}^m\) for some integer \(m \ll n = |S|\) and \(\phi(s) = (\phi_1(s), \ldots, \phi_m(s))^T\) is the feature vector of the state \(s \in S\). We compactly represent the feature vectors as a matrix of features \(\Phi\) of size \(n \times m\) whose \(i\)th row corresponds to the row vector \(\phi(s)_i^T\) for some \(s \in S\).

Now, before completing the presentation of our algorithm, we motivate the use of a target variable for the critic. As previously mentioned, instead of a standard TD learning algorithm [Sutton, 1988] for the critic, we use a target-based TD learning algorithm. We follow a similar exposition to Lee and He, 2019 Secs. 2.3, 2.4 and 3] to introduce the target variable for the critic. Let us introduce some additional notations for this purpose. Fix \(\theta \in \mathbb{R}^d\). Let \(P_\theta\) be the transition matrix over the finite state space associated to the Markov chain \((S_t)\), i.e., the matrix of size \(n \times n\) defined for every \(s, s' \in S\) by \(P_\theta(s'|s) := \sum_{a \in A} p(a|s, s') \pi_\theta(a|s)\). Consider the vector \(\bar{R}_\theta = (R_{\theta}(s_1), \ldots, R_{\theta}(s_n))^T\) whose \(i\)th coordinate is provided by \(R_{\theta}(s_i) = \sum_{a \in A} p_i(a|s_i) R_{\theta}(a, s_i)\) and \(R_{\theta}(s_i)\) is the diagonal matrix with elements \(d_{p, \theta}(s_i), i = 1, \ldots, n\) along its diagonal. Define also the Bellman operator \(T_\theta : \mathbb{R}^n \to \mathbb{R}^n\) for every \(V \in \mathbb{R}^n\) by \(T_\theta V := R_\theta + \gamma p_\theta V\).

The true value function \(V_{\pi_\theta}\) satisfies the celebrated Bellman equation \(V_{\pi_\theta} = T_{\theta} V_{\pi_\theta}\). This naturally leads to minimize the mean-square Bellman error (MSBE) [Sutton et al., 2009, Sec. 3] defined for every \(\omega \in \mathbb{R}^m\) by \(\mathcal{E}_\theta(\omega) := \frac{1}{2} ||T_{\theta} V_\omega - V_\omega||^2_{p_\theta, \omega}\) where \(V_\omega = \Phi \omega\). The gradient of the MSBE w.r.t. \(\omega\) can be written as \(\nabla_\omega \mathcal{E}_\theta(\omega) = \mathbb{E}_{S \sim p_\theta, \omega} [T_{\theta} V_\omega(\tilde{S}) - V_\omega(\tilde{S})](\mathbb{E}_{S \sim p_\theta, \omega} [\gamma V_\omega(\tilde{S})] - \nabla_\omega V_\omega(\tilde{S}))\).

As explained in Bertsekas and Tsitsiklis, 1996 p. 369, omitting the gradient term \(\nabla_\omega T_{\theta} V_\omega(\tilde{S}) = \mathbb{E}_{S \sim p_\theta, \omega} [\gamma V_\omega(\tilde{S})]\) in \(\mathbb{E}_{\tilde{S} \sim T_{\theta} \omega} [\gamma V_\omega(\tilde{S})]\) yields the standard TD learning update rule \(\omega_{t+1} = \omega_t + \delta_{t+1} \phi(\tilde{S}_t)\). The TD learning update does not coincide with a stochastic gradient descent on the MSBE or even any other objective function (see Barmard, 1993 Appendix 1 for a proof). The idea of target-based TD learning is to consider a modified version of the MSBE \(\bar{\mathcal{E}}_\theta(\omega, \tilde{\omega}) := \frac{1}{2} ||T_{\theta} V_\omega - V_\omega||^2_{p_\theta, \omega}\). Observe that the term \(T_{\theta} V_\omega\) depending on \(\omega\) in the MSBE is now frozen in \(\bar{\mathcal{E}}_\theta(\omega, \tilde{\omega})\) thanks to the target variable \(\tilde{\omega}\). We now need to introduce a new sequence \(\tilde{\omega}_t\) to define a sample-based version of \(T_{\theta} V_\omega - V_\omega\) which will be a modified version of the standard TD-error

\[
\bar{\delta}_{t+1} = R_{t+1} + \gamma \phi(S_{t+1})^T \tilde{\omega}_t - \phi(\tilde{S}_t)^T \omega_t.
\]

Then, a stochastic gradient descent on \(\bar{\mathcal{E}}\) w.r.t. \(\omega\) yields the critic update

\[
\omega_{t+1} = \omega_t + \beta_t \bar{\delta}_{t+1} \phi(\tilde{S}_t).
\]

The target variable sequence \(\tilde{\omega}_t\) needs to be a slowed down version of the critic parameter \(\omega_t\). For this purpose, instead of using a periodical synchronization of the target variable \(\tilde{\omega}_t\) with \(\omega_t\) through a copy as in DQN, we use the Polyak-averaging update rule proposed by Lillicrap et al., 2015

\[
\tilde{\omega}_{t+1} = \tilde{\omega}_t + \xi_t (\omega_{t+1} - \tilde{\omega}_t),
\]

where \(\xi_t\) is a positive stepsize chosen s.t. the sequence \(\tilde{\omega}_t\) evolves on a slower timescale than the sequence \(\omega_t\) to track it. The update rules of the actor and the critic collected together from Eqs. (6) to (9) give rise to Algorithm 1. We will use the shorthand notation \(\delta_{t+1} := \delta_{V_{t+1}}\) from now on.

**Remark 1.** We can simplify Algorithm 1 by using only the target-based TD error \(\delta_{t+1}\) instead of maintaining
both TD errors $\tilde{\delta}_{t+1}$ and $\delta_{t+1}$. The proofs can be easily adapted, note for this that $(\bar{\omega}_t)$ and $(\omega_t)$ track the same target $\bar{\omega}_t(\theta_t)$ (see Prop. 5.2 Th. 5.3). For clarity of exposition, we present the algorithm with both TD errors, since the classical TD error stems directly from the policy gradient whereas the target-based TD error comes from the use of the target network.

5 CONVERGENCE ANALYSIS

In this section, we provide asymptotic convergence guarantees for the critic and the actor of Algorithm 1 successively. For every $\theta \in \mathbb{R}^d$, let $K_\theta \in \mathbb{R}^{[S] \times [A] \times [S] \times [A]}$ be the transition matrix over the state-action pairs defined for every $(s, a), (s', a') \in S \times A$ by $K_\theta(s', a'|s, a) = \bar{p}(s'|s, a)\pi_\theta(a'|s)$. Let $\mathcal{K} := \{K_\theta : \theta \in \mathbb{R}^d\}$ and let $\mathcal{K}$ be its closure. Every element of $\mathcal{K}$ defines a Markov chain on the state-action space. We make the following assumption (see also Zhang et al., 2021 Marbach and Tsitsiklis, 2001).

Assumption 5.1. For every $K \in \mathcal{K}$, the Markov chain induced by $K$ is ergodic.

In particular, it ensures the existence of a unique invariant distribution $\mu_{\mu, \theta}$ for the kernel $K_\theta$ for every $\theta \in \mathbb{R}^d$. Note that we can replace $\bar{p}$ by $p$ in Assumption 5.1.

Algorithm 1 involves three different timescales. The actor parameter $\theta_t$ is updated on a slower timescale (i.e., with smaller stepsizes) than the target variable $\bar{\omega}_t$ which itself uses smaller stepsizes than the main critic parameter $\omega_t$. This is guaranteed by a specific choice of the three stepsize schedules. The following assumption is a three timescales version of the standard assumption used for two timescales stochastic approximation [Borkar, 2008 Chap. 6] and plays a pivotal role in our analysis.

**Algorithm 1** Target-based actor-critic.

**Initialization:** $\theta_0 \in \mathbb{R}^d, \omega_0 \in \mathbb{R}^m$.

for $t = 0, 1, 2, \cdots, T - 1$ do

$A_t \sim \pi_{\theta_t}(\cdot|\bar{S}_t); S_{t+1} \sim p(\cdot|\bar{S}_t, A_t)$

$\theta_{t+1} = R_{t+1} + \gamma \phi(S_{t+1})^T \omega_t - \phi(S_t)^T \omega_t$  \> classical TD error

$\tilde{\delta}_{t+1} = R_{t+1} + \gamma \phi(S_{t+1})^T \bar{\omega}_t - \phi(S_t)^T \bar{\omega}_t$  \> target-based TD error

$\tilde{\theta}_{t+1} = \theta_t + \alpha_t \frac{1}{\gamma} \tilde{\delta}_{t+1} \psi_\theta(S_t, A_t)$  \> actor

$\omega_{t+1} = \omega_t + \beta_t \tilde{\delta}_{t+1} \phi(\bar{S}_t)$  \> critic

$\tilde{\omega}_{t+1} = \tilde{\omega}_t + \xi_t(\omega_{t+1} - \tilde{\omega}_t)$  \> target variable

$S_{t+1}^\theta \sim \rho; B_{t+1} \sim \mathcal{B}(\gamma)$

$\bar{S}_{t+1} = B_{t+1} S_{t+1} + (1 - B_{t+1}) S_{t+1}^\theta$

end for

**Output:** Policy and value function parameters $\theta_T$ and $\omega_T$.

Assumption 5.2 (stepsizes). The sequences of positive stepsizes $(\alpha_t), (\beta_t)$ and $(\xi_t)$ satisfy:

(a) $\sum_t \alpha_t = \sum_t \beta_t = \sum_t \xi_t = +\infty$ ,

(b) $\sum_t (\alpha_t^2 + \beta_t^2 + \xi_t^2) < +\infty$,

(c) $\lim_{t \to \infty} \alpha_t / \xi_t = \lim_{t \to \infty} \xi_t / \beta_t = 0$.

We also need the following stability assumption.

**Assumption 5.3.** $\sup_t (||\omega_t|| + ||\theta_t||) < +\infty$ w.p.1.

The almost sure boundedness assumption is classical Konda and Borkar, 1999 Borkar, 2008 Bhatnagar et al., 2009 Karmakar and Bhatnagar, 2018. The stability question could be addressed in a look up table representation (for e.g., $m = n$). Nevertheless, this question seems out of reach in the FA setting without any modification of the algorithm. Indeed, as discussed in [Bhatnagar et al., 2009] p. 2478-2479, FA makes it hard to find a Lyapunov function to apply the stochastic Lyapunov function method [Kushner and Yin, 2003] whereas the function $J$ can be readily used in the tabular case. Under a modification of the actor update of the algorithm and slightly stronger assumptions inspired from [Konda and Tsitsiklis, 2003a Konda and Tsitsiklis, 2003b], the almost sure boundedness of the sequence $(\bar{\omega}_t)$ can be relaxed using a generalization to three timescales of the rescaling technique of [Borkar and Meyn, 2000] which was extended by [Lakshminarayanan and Bhatnagar, 2017] to two timescales stochastic approximation in the case of i.i.d. samples. For simplicity of exposition, we defer the technical details regarding this question to the appendix (see Appendix C). Concerning the sequence $(\theta_t)$, as previously mentioned, it seems out of reach without modifying the algorithm, [Lakshminarayanan and Bhatnagar, 2017] (see their Section 6) propose for example to regularize the objective function $J$ by adding a quadratic penalty $\epsilon ||\theta||^2 / 2$ (positive leading to an additional $\epsilon \theta_t$ term in the actor update of the standard actor-critic algorithm 1 of [Bhatnagar et al., 2009]. We do not make use of this trick which modifies the critical points of the performance function. It is also worth mentioning that several works enforce the boundedness via a projection of the iterates on some compact set [Bhandari et al., 2018 Wu et al., 2020 Shien et al., 2020 Zhang et al., 2021]. The drawback of this procedure is that it modifies the dynamics of the iterates and could possibly introduce spurious equilibria.

First, we will analyze the critic before investigating the convergence properties of the actor.
5.1 Critic analysis

The following assumption regarding the family of basis functions is a standard requirement [Bhatnagar et al., 2009; Konda and Tsitsiklis, 2003b; Tsitsiklis and Van Roy, 1997].

Assumption 5.4 (critic features). The matrix $\Phi$ has full column rank.

We follow the strategy of [Borkar, 2008, Chap. 6, Lem. 1] for the analysis of multi-timescale stochastic approximation schemes based on the ODE method. We start by analyzing the sequence $(\omega_t)$ evolving on the fastest timescale, i.e., with the slowly vanishing step-sizes $\beta_t$ (see Assumption 5.2). The main idea behind the proofs is that $\theta_t, \omega_t$ can be considered as quasi-static in this timescale. Then, loosely speaking (see Appendix for a rigorous statement and proof), we can show from its update rule Eq. (9) that $(\omega_t)$ is associated to the ODE

$$
\begin{align*}
\dot{\omega}(s) &= \bar{h}(\theta(s), \omega(s)) - \bar{G}(\theta(s)) \omega(s), \\
\dot{\theta}(s) &= 0, \\
\hat{\omega}(s) &= 0,
\end{align*}
$$

(ODE-$\omega$)

where $\bar{h} : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m$ and $\bar{G} : \mathbb{R}^d \to \mathbb{R}^{m \times m}$ are defined for every $\theta \in \mathbb{R}^d$, $\bar{\omega} \in \mathbb{R}^m$ by

$$
\bar{h}(\theta, \bar{\omega}) := \Phi^T \bar{D}_{\rho, \bar{\theta}} (R_\theta + \gamma P_\rho \bar{\Phi}) \bar{\omega} \quad \text{and} \quad \bar{G}(\theta) := \Phi^T \bar{D}_{\rho, \bar{\theta}} \bar{\Phi}.
$$

(11)

Recall that the matrices $D_{\rho, \bar{\theta}}, P_\rho$ and the vector $R_\theta$ are defined in Sec. 4.2.

Remark 2. Under Assumptions 5.1 and 5.4, the matrix $-\bar{G}(\theta)$ is Hurwitz for every $\theta \in \mathbb{R}^d$, i.e., all its eigenvalues have negative real parts. In particular, it is invertible.

The matrix $-\bar{G}(\theta)$ being Hurwitz, it follows from (ODE-$\omega$) that $\omega_t$ tracks a slowly moving target $\omega_\ast(\theta_t, \bar{\omega}_t)$ governed by the slower iterates $\theta_t$ and $\bar{\omega}_t$. The detailed proof in the appendix makes use of a result from [Karmakar and Bhatnagar, 2018] to handle the Markovian noise.

Proposition 5.1. Under Assumptions 3.1 and 5.1 to 5.4, the linear equation $\bar{G}(\theta) \omega = \bar{h}(\theta, \bar{\omega})$ has a unique solution $\omega_\ast(\theta, \bar{\omega})$ for every $\theta \in \mathbb{R}^d$, $\bar{\omega} \in \mathbb{R}^m$ and $\lim_t \|\omega_t - \omega_\ast(\theta_t, \bar{\omega}_t)\| = 0$ w.p.1.

In a second step, we analyze the target variable sequence $(\bar{\omega}_t)$ which is evolving on a faster timescale than the sequence $(\theta_t)$ and slower than the sequence $(\omega_t)$. At the timescale $\xi_t$, everything happens as if the quantity $\omega_t$ in Eq. (10) could be replaced by $\omega_\ast(\theta_t, \bar{\omega}_t)$ thanks to Prop. 5.1. Thus, in a sense that is made precise in the appendix, we can show from Eq. (10) that $(\bar{\omega}_t)$ is related to the ODE

$$
\begin{align*}
\dot{\bar{\omega}}(s) &= \bar{G}(\theta(s))^{-1}(\bar{h}(\theta(s)) - \bar{G}(\theta(s))\bar{\omega}(s)), \\
\dot{\theta}(s) &= 0,
\end{align*}
$$

(ODE-$\bar{\omega}$)

where $h : \mathbb{R}^d \to \mathbb{R}^n$ and $G : \mathbb{R}^d \to \mathbb{R}^{m \times m}$ are defined for every $\theta \in \mathbb{R}^d$ by

$$
\bar{h}(\theta) := \Phi^T D_{\rho, \theta} R_\theta \quad \text{and} \quad \bar{G}(\theta) := \Phi^T D_{\rho, \theta} (I_n - \gamma P_\theta) \Phi.
$$

(12)

We show in the appendix that the matrix $-\bar{G}(\theta)$ is Hurwitz. This result differs from [Bertsekas and Tsitsiklis, 1996, Lem. 6.6. p.300] or [Tsitsiklis and Van Roy, 1997, Lem. 9] because the matrix $D_{\rho, \theta}$ corresponds to the stationary distribution associated to the artificial kernel $\bar{p}$ and the policy $\pi_\theta$ in lieu of the original transition kernel $p$. Then, we prove that $-\bar{G}(\theta)^{-1}\bar{G}(\theta)$ is also stable, which suggests from (ODE-$\bar{\omega}$) that $\bar{\omega}_t$ tracks an other slowly moving target $\bar{\omega}_\ast(\bar{\theta}_t)$. This is established in the next proposition.

Proposition 5.2. Under Assumptions 3.1 and 5.1 to 5.4, for every $\theta \in \mathbb{R}^d$, the linear equation $\bar{G}(\theta) \omega = \bar{h}(\theta)$ has a unique solution $\bar{\omega}_\ast(\theta)$ and $\lim_t \|\bar{\omega}_t - \bar{\omega}_\ast(\theta_t)\| = 0$ w.p.1. Moreover, for every $\theta \in \mathbb{R}^d$, $\bar{\Phi}$ $\bar{\omega}_\ast(\theta)$ is a fixed point of the projected Bellman operator, i.e., $\Pi_\theta T_\theta(\bar{\Phi} \bar{\omega}_\ast(\theta)) = \bar{\Phi} \bar{\omega}_\ast(\theta)$, where $\Pi_\theta = \Phi^T D_{\rho, \theta} (\Phi^T D_{\rho, \theta} \Phi)^{-1} \Phi^T D_{\rho, \theta}$ is the projection matrix on the space $\{\Phi \omega : \omega \in \mathbb{R}^m\}$ of all vectors of the form $\Phi \omega$ for $\omega \in \mathbb{R}^m$. w.r.t. the norm $\|\cdot\|_{D_{\rho, \theta}}$.

Combining the results from Props. 5.1 and 5.2, we prove that $\omega_t$ tracks the same target $\omega_\ast(\theta_t)$.

Theorem 5.3. Let Assumptions 3.1 and 5.1 to 5.4 hold true. Then, we have

$$
\lim_t \|\omega_t - \omega_\ast(\theta_t)\| = 0 \quad \text{w.p.1}.
$$

Moreover, this limit implies the following:

$$
\lim_t \|\Pi_\theta T_\theta(\bar{\Phi} \omega_t) - \bar{\Phi} \omega_t\| = 0 \quad \text{w.p.1}.
$$

Remark 3. When the actor parameter $\theta_t$ is fixed (i.e., we are back to a policy evaluation problem), the second part of the above convergence result coincides with the widely known interpretation of the limit of the TD learning algorithm provided in [Tsitsiklis and Van Roy, 1997] (see also [Bertsekas and Tsitsiklis, 1996, p. 303-304]).

5.2 Actor analysis

Theorem 5.4. Let Assumptions 3.1 and 5.1 to 5.4 hold true. Then, w.p.1

$$
\liminf_t (\|\nabla J(\theta_t)\| - \|b(\theta_t)\|) \leq 0,
$$
where for every $\theta \in \mathbb{R}^d$, $(s,a) \in S \times A$, $b(\theta) := \frac{1}{Z_\theta} \mathbb{E}_{\mu_\theta}(\psi_\theta(S,A)|Q_\theta(S,A) - Q_{\pi_\theta}(S,A))$ and $Q_\theta(s,a) := R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\phi(s')^T \bar{\omega}_\varphi(\theta)$.

Th. 5.4 is analog to [Konda, 2002, Th. 5.5] which is established for the standard on-policy actor-critic in the average reward setting and [Zhang et al., 2020b, Th. 3] for an off-policy actor-critic without any target network. The result states that the sequence $(\theta_t)$ generated by our actor-critic algorithm visits any neighborhood of the set $\{\theta \in \mathbb{R}^d : \|\nabla J(\theta)\| \leq \|b(\theta)\|\}$ infinitely often. The bias $b(\theta)$ corresponds to the difference between the gradient $\nabla J(\theta)$ and the steady state expectation of the actor’s update direction. The estimate used to update the actor in Eq. (7) is only a biased estimate of $\nabla J(\theta)$ because of linear FA.

Remark 4. The bias $b(\theta)$ disappears in the tabular setting ($m = |S|$) and the features spanning $\mathbb{R}^{|S|}$ when we do not use FA and in the linear FA setting when the value function belongs to the class of linear functions spanned by the pre-selected feature (or basis) functions. Beyond these particular settings, considering compatible features as introduced in [Sutton et al., 2000, Konda and Tsitsiklis, 2003b] can be a solution to cancel the bias $b(\theta)$ incurred by Algorithm 1. We do not investigate this direction in this work.

6 FINITE-TIME ANALYSIS

Our analysis in this section should be valid for a continuous state space $S$ (and still finite action space) upon supposing that the feature map $\phi$ defined in Section 4.2 has bounded norm (i.e., $\|\phi(\cdot)\| \leq 1$) and slightly adapting our notations and definitions to this more general setting (see also for e.g., [Wu et al., 2020]). To stay concise and consistent with the first part of our analysis in Section 5 we restrict ourselves to the finite state space setting.

6.1 Critic analysis

For every $\theta \in \mathbb{R}^d$, we suppose that the Markov chain $(\tilde{S}_t)$ induced by the policy $\pi_\theta$ and the transition kernel $\tilde{p}$ mixes at a geometric rate.

Assumption 6.1. There exist constants $c > 0$ and $\sigma \in (0,1)$ s.t. for every $t \in \mathbb{N}$, $\theta \in \mathbb{R}^d$,

$$\sup_{s \in S} d_{TV}(P(\tilde{S}_t = s|\tilde{S}_0, \pi_\theta), d_{\rho,\theta}) \leq ca^t,$$

where $d_{TV}(\cdot, \cdot)$ denotes the total-variation distance between two probability measures.

This assumption is used to control the Markovian noise induced by sampling transitions from the MDP under a dynamically changing policy. It was considered first in [Bhandari et al., 2018] in a policy evaluation setting for the finite-time analysis of TD learning. It was later used for instance in [Zou et al., 2019, Wu et al., 2020, Shen et al., 2020].

We have seen in Sec. 5.1 that the dynamics of the critic is driven by two key matrices $-\tilde{G}(\theta)$ and $-\tilde{G}(\theta)^{-1} \tilde{G}(\theta)$. While we only need these matrices to be stable for our asymptotic results, we actually show in the appendix that $-\tilde{G}(\theta)$ is even negative definite uniformly in $\theta$. We suppose that the second matrix $-\tilde{G}(\theta)^{-1} \tilde{G}(\theta)$ is also negative definite uniformly in $\theta$.

Assumption 6.2. There exists $\zeta > 0$ s.t. for every $\theta \in \mathbb{R}^d$, $\omega \in \mathbb{R}^m$, $\omega^T \tilde{G}(\theta)^{-1} \tilde{G}(\theta) \omega \geq \zeta \|\omega\|_2^2$.

We are now ready to state our critic convergence rate.

Theorem 6.1. Let Assumptions 5.1, 6.1 and 6.3 hold. Let $c_1, c_2, c_3, \alpha, \xi, \beta$ be positive constants s.t. $0 < \beta < \xi < \alpha < 1$. Set $\alpha_t = \frac{c_1}{(1 + \xi)^t}$, $\xi_t = \frac{c_2}{(1 + \alpha)^t}$ and $\beta_t = \frac{c_3}{(1 + \xi)^t}$. Then, the sequences $(\omega_t)$ and $(\theta_t)$ from Algorithm 1 satisfy for every integer $T \geq 1$,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\omega_t - \omega^*(\theta_t)\|_2^2] = \mathcal{O}\left(\frac{1}{T^{1-\xi}}\right) + \mathcal{O}\left(\frac{\ln T}{T^{\beta}}\right) + \mathcal{O}\left(\frac{1}{T^{2(\alpha - \xi)}}\right) + \mathcal{O}\left(\frac{1}{T^{2(\xi - \beta)}}\right).$$

The bound of Th. 6.1 shows the impact of using a target variable. First, the last two terms impose the conditions $\alpha > \xi$ and $\xi > \beta$. At least with linear FA, this may provide a theoretical justification to the common practice of updating the target network at a slower rate compared to the main network for the critic. Second, compared to [Wu et al., 2020, Th. 4.7] which is concerned with the standard actor-critic in the average reward setting, we have the slower $\mathcal{O}(T^{\xi - 1})$ instead of $\mathcal{O}(T^{\xi - 1})$ and our bound comprises four error terms. These are also consequences of the use of a target variable.

Remark 5. Although we use similar proof techniques to [Wu et al., 2020] for our finite-time analysis, notice that our novel asymptotic analysis of the critic (Sec. 6.1) is crucial for the proof (see Sec. 6.1 for details).

6.2 Actor analysis

We suppose that the critic approximation error induced by linear FA is uniformly bounded (see also [Qiu et al., 2019, Wu et al., 2020, Xu et al., 2020a]).

Assumption 6.3. There exists $c_{FA} \geq 0$ s.t. for every $\theta \in \mathbb{R}^d$, $\|V_{\pi_\theta} - \Phi(\theta)\|_{D_{\rho,\theta}} \leq c_{FA}$. 
Theorem 6.2. Let Assumptions 5.1, 5.3 and 6.3 hold. Let \( c_1, c_2, c_3, \alpha, \xi, \beta \) be positive constants s.t. \( 0 < \beta < \xi < \alpha < 1 \). Set \( \alpha_t = \frac{c_1}{(1 + \beta)^t} \), \( \xi_t = \frac{c_2}{(1 + \beta)^t} \) and \( \beta_t = \frac{c_3}{(1 + \beta)^t} \). Then, for every integer \( T \geq 1 \),

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = O\left(\frac{1}{T^{1-\alpha}}\right) + O\left(\frac{\ln^2 T}{T^\alpha}\right)
+ O\left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|w_t - \bar{w}_t(\theta_t)\|^2]\right) + O\left(\epsilon_{FA}\right).
\]

Combining Th. 6.1 and Th. 6.2 we obtain the following result.

Corollary 6.3. Under the setting and the assumptions of Ths. 6.1 and 6.2, we have for every \( T \geq 1 \),

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = O\left(\frac{1}{T^{1-\alpha}}\right) + O\left(\frac{\ln T}{T^{\beta}}\right)
+ O\left(\frac{1}{T^{2(\alpha - \xi)}}\right) + O\left(\frac{1}{T^{2(\xi - \beta)}}\right) + O\left(\epsilon_{FA}\right).
\]

Moreover, if we set \( \alpha = \frac{2}{3}, \xi = \frac{1}{2} \) and \( \beta = \frac{1}{3} \) to define the stepsizes sequence \( (\alpha_t), (\xi_t) \) and \( (\beta_t) \), the actor parameter sequence \( (\theta_t) \) generated by Algorithm 1 within \( T = O(\epsilon^{-3} \ln^3(\frac{1}{\epsilon})) \) steps, satisfies

\[
\min_{0 \leq t \leq T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \leq O(\epsilon_{FA}) + \epsilon.
\]

As a consequence, since Algorithm 1 uses a single sample from the MDP per iteration, its sample complexity is \( O(\epsilon^{-3} \ln^3(\frac{1}{\epsilon})) \). This is to compare with the best \( O(\epsilon^{-2} \ln^2(\frac{1}{\epsilon})) \) sample complexity known in the literature (to the best of our knowledge) for actor-critic algorithms up to the linear FA error [Xu et al., 2020a Th. 2]. Although the use of a target variable seems to deteriorate the sample complexity w.r.t. the best known result for target-free actor-critic methods, note that it is still aligned with the complexity reported in [Qu et al., 2019] (up to logarithmic factors) and better than the \( O(\epsilon^{-4}) \) sample complexity obtained in [Kumar et al., 2019] with i.i.d. sampling. Notice that we do not make use of mini-batching of samples (even from a single sample path) or nested loops as in [Xu et al., 2020a]. We refer to [Wu et al., 2020 Section 4.4] and [Xu et al., 2020a Table 1] for further discussion. We briefly comment on the origin of this deteriorated sample complexity stemming from our finite-time bounds. Due to the use of a target variable, instead of the \( O(T^{2(\alpha - \beta)}) \) error term of the standard actor-critic (see [Wu et al., 2020] Cor. 4.9) or [Shen et al., 2020 Ths.3-4]), we have two error terms \( O(T^{2(\alpha - \xi)}) \) and \( O(T^{2(\xi - \beta)}) \) slowing down the convergence because of the condition \( \beta < \xi < \alpha \). Interestingly, at least in the linear FA setting, this corroborates the practical intuition that the use of a target network may slow down learning as formulated for instance in [Lillicrap et al., 2016 Section 3] (even if constant stepsizes are used in practice).

Remark 6. Remark 3 also applies to the function approximation error \( \epsilon_{FA} \).

7 CONCLUSION AND FUTURE WORK

This paper provides the first convergence analysis of an actor-critic algorithm incorporating a target network, establishing both asymptotic and finite-time results under Markovian sampling. Motivated by the success of actor-critic methods using target networks in deep RL, our analysis shows that this target network mechanism is theoretically sound in the linear FA setting. Although our analysis does not demonstrate a particular advantage of target-based actor-critic methods over non-target based counterpart in the linear FA setting, our results pave the road for the nonlinear FA setting. There are several interesting directions for future research. A theoretical justification of the use of a target network in the nonlinear FA setting beyond linear FA is a challenging problem that merit further investigation. In particular, as practical algorithms in deep RL seem to indicate, it would be interesting to see if such a trick can be a theoretically grounded alternative to the failure of temporal difference learning with nonlinear FA. Another possible avenue for future work to close the gap between theory and practice is to address the case of off-policy target-based actor-critic algorithms which have enjoyed great empirical success [Fujimoto et al., 2018 Haarnoja et al., 2018].

Acknowledgements

The authors would like to thank the anonymous referees for their useful feedback. Anas Barakat was supported by the "Futur & Ruptures" research program which is jointly funded by the IMT, the Mines-Télécom Foundation and the Carnot TSN Institute.

References


Analysis of a Target-Based Actor-Critic Algorithm with Linear Function Approximation

IEEE transactions on automatic control, 42(5):674–690.


Supplementary Material:
Analysis of a Target-Based Actor-Critic Algorithm
with Linear Function Approximation

A Proofs for Sec. 5: asymptotic convergence results

A.1 Critic analysis

The objective of this section is to prove Th. 5.3. First, we recall the outline of the proof. Our actor-critic algorithm features three different timescales associated to three different stepsizes converging to zero with different rates, each one associated to one of the sequences \((\theta_t), (\bar{\omega}_t)\) and \((\hat{\omega}_t)\). In spirit, we follow the strategy of [Borkar, 2008 Chap. 6, Lem. 1] for the analysis of two timescales stochastic approximation schemes. We make use of the results of [Karmakar and Bhatnagar, 2018] which handles controlled Markov noise. The proof is divided into three main steps:

(i) We start by analyzing the sequence \((\bar{\omega}_t)\) evolving on the fastest timescale, i.e., with the stepsizes \(\beta_t\) which are converging the slowest to zero (see Assumption 5.2). We rewrite the slower sequences \((\theta_t), (\hat{\omega}_t)\) with the stepsizes \(\beta_t\). In this timescale, \((\theta_t), (\hat{\omega}_t)\) are quasi-static from the point of view of the evolution of the sequence \((\bar{\omega}_t)\). We deduce from this first step that \(\omega_t\) tracks a slowly moving target \(\omega_*(\theta_t, \bar{\omega}_t)\) governed by the slower iterates \(\theta_t\) and \(\bar{\omega}_t\). This is the purpose of Prop. 5.1 which is proved in Sec. A.1.1 below.

(ii) In a second step, we analyze the sequence \((\hat{\omega}_t)\) which is evolving in a faster timescale than the sequence \((\theta_t)\) and slower than the sequence \((\omega_t)\). Similarly, we show that \(\hat{\omega}_t\) tracks an other slowly moving target \(\omega_*(\hat{\omega}_t)\). This is established in the proof of Prop. 5.2 in Sec. A.1.2

(iii) We conclude in Sec A.1.3 by combining the results from the first two steps, proving that the sequence \(\omega_t\) tracks the same target \(\omega_*(\theta_t)\).

A.1.1 Proof of Prop. 5.1

Let \(\mathcal{F}_t\) be the \(\sigma\)-field generated by the random variables \(S_t, \hat{S}_t, \bar{A}_t, \theta_t, \bar{\omega}_t, \omega_t\) for \(t \leq T\). For each time step \(t\), let \(Z_t = (\bar{S}_t, \bar{A}_t)\). Our objective here is to show that the critic sequence \((\bar{\omega}_t)\) tracks the slowly moving target \(\omega_*(\theta_t, \bar{\omega}_t)\) defined in Prop. 5.1. From the update rule of the sequence \((\bar{\omega}_t)\), we have

\[
\bar{\omega}_{t+1} = \bar{\omega}_t + \beta_t \bar{\theta}_{t+1} \phi(\bar{S}_t) = \omega_t + \beta_t (R_{t+1} + \gamma \phi(S_{t+1})^T \bar{\omega}_t - \phi(\bar{S}_t) \phi(\bar{\omega}_t))
\]

\[=
\omega_t + \beta_t w(\bar{\omega}_t, \omega_t, Z_t) + \beta_t \eta_{t+1}^{(1)}, \tag{13}
\]

where for every \(\bar{\omega}, \omega \in \mathbb{R}^m, z = (s, a) \in S \times A,

\[
w(\bar{\omega}, \omega, z) := \left( R(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) \phi(s')^T \bar{\omega} \right) \phi(s) - \phi(s) \phi(s)^T \omega \tag{14}
\]

and \(\eta_{t+1}^{(1)}\) is a martingale difference sequence defined as

\[
\eta_{t+1}^{(1)} = (R_{t+1} - R(\bar{S}_t, \bar{A}_t)) \phi(\bar{S}_t) + \gamma \bar{\omega}_t^T (\phi(S_{t+1}) - \mathbb{E}[\phi(S_{t+1}) | \mathcal{F}_t]) \phi(\bar{S}_t). \tag{15}
\]

As can be seen in Eq. (13), the sequence \((\omega_t)\) can be written as a linear stochastic approximation scheme controlled by the slowly varying Markov chains \((\theta_t)\) and \((\hat{\omega}_t)\). In view of characterizing its asymptotic behavior, we compute
for fixed $\bar{\omega}, \omega \in \mathbb{R}^m$ the expectation of the quantity $w(\bar{\omega}, \omega, Z)$ (see Eq. (14)) where $Z = (\bar{S}, \bar{A})$ is a random variable (on $\mathcal{S} \times \mathcal{A}$) following the stationary distribution $\mu_{\rho, \theta}$ (see Eq. (2)) of the Markov chain $(Z_t)$. Recall the definitions of $\bar{h} : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m$ and $\tilde{G} : \mathbb{R}^d \to \mathbb{R}^{m \times m}$ from Eq. (11), for every $\theta \in \mathbb{R}^d, \bar{\omega} \in \mathbb{R}^m$

$$\bar{h}(\theta, \bar{\omega}) := \Phi^T D_{\rho, \theta} (R_{\theta} + \gamma P_{\theta} \bar{\omega}) \quad \text{and} \quad \tilde{G}(\theta) := \Phi^T D_{\rho, \theta} \Phi.$$

**Lemma A.1.** Under Assumption 5.1, for every $\bar{\omega}, \omega \in \mathbb{R}^m$, we have

$$\mathbb{E}_{Z \sim \mu_{\rho, \theta}}[w(\bar{\omega}, \omega, Z)] = \bar{h}(\theta, \bar{\omega}) - \tilde{G}(\theta) \omega.$$

**Proof.** We obtain from the definitions of $w$ in Eq. (14) and $\mu_{\rho, \theta}$ in Eq. (2) that

$$\mathbb{E}_{Z \sim \mu_{\rho, \theta}}[w(\bar{\omega}, \omega, Z)] = \mathbb{E}_{Z \sim \mu_{\rho, \theta}} \left[ \left( R(\bar{S}, \bar{A}) + \gamma \sum_{s' \in \mathcal{S}} p(s'|\bar{S}, \bar{A}) \phi(s')^T \bar{\omega} \right) \phi(\bar{S}) - \phi(\bar{S}) \phi(\bar{S})^T \omega \right]$$

$$= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu_{\rho, \theta}(s, a) \left( R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \phi(s')^T \bar{\omega} \right) \phi(s) - \phi(s) \phi(s)^T \omega$$

$$= \sum_{s \in \mathcal{S}} d_{\rho, \theta}(s) \left( R_{\theta}(s) \phi(s) + \gamma \sum_{s' \in \mathcal{S}} p_{\theta}(s'|s) \phi(s')^T \bar{\omega} \phi(s) - \phi(s) \phi(s)^T \omega \right)$$

$$= \bar{h}(\theta, \bar{\omega}) - \tilde{G}(\theta) \omega,$$

where the penultimate equation stems from recalling that $R_{\theta}(s) = \sum_{a \in \mathcal{A}} R(s, a) \pi_{\theta}(a|s)$ and $p_{\theta}(s'|s) = \sum_{a \in \mathcal{A}} p(s'|s, a) \pi_{\theta}(a|s)$ for every $s \in \mathcal{S}$.

Defining $\chi_t = (\theta_t, \bar{\omega}_t)$, we obtain from the update rules of $(\theta_t)$ and $(\bar{\omega}_t)$ that

$$\chi_{t+1} = \chi_t + \beta_t \epsilon_t,$$

where $\epsilon_t = \left( \frac{\alpha_t}{\beta_t} \frac{1}{\gamma - \eta_t} \psi_t(Z_t), \frac{\xi_t}{\beta_t} (\bar{\omega}_{t+1} - \bar{\omega}_t) \right).$ Notice that $\epsilon_t \to 0$ as $t \to \infty$. This is because $\frac{\alpha_t}{\beta_t} \to 0$, $\frac{\xi_t}{\beta_t} \to 0$ by Assumption 5.2. $(\omega_t)$ and (hence) $(\bar{\omega}_t)$ are a.s. bounded by Assumption 5.3. $(R_t)$ is bounded by $d_{R, \theta} \to \psi_R(s, a)$ is bounded by Assumption 5.1 and $\mathcal{S}, \mathcal{A}$ are finite.

Let $\zeta_t = (\chi_t, \bar{\omega}_t)$, $\zeta = (\theta, \bar{\omega}) \in \mathbb{R}^{d+2m}$, $W(\zeta, z) = (0, w(\bar{\omega}, \omega, z))$, $\epsilon'_t = (\epsilon_t, 0)$ and $\eta^{(1)}_{t+1} = (0, \eta^{(1)}_{t+1})$. Then, we can write Eqs. (16) and (13) in the framework of [Karmakar and Bhatnagar, 2018 Sec. 3, Eq.(14), Lem. 9], i.e., as a single timescale controlled Markov noise stochastic approximation scheme:

$$\zeta_{t+1} = \zeta_t + \beta_t [W(\zeta_t, Z_t) + \epsilon'_t + \eta^{(1)}_{t+1}],$$

(17) with $\epsilon'_t \to 0$. Under the assumptions of [Karmakar and Bhatnagar, 2018] that we will verify at the end of the proof, we obtain that the sequence $(\zeta_t)$ converges to an internally chain transitive set (i.e., a compact invariant set which has no proper attractor, see definition in [Karmakar and Bhatnagar, 2018 Sec. 2.1] or [Benaim, 1996 Sec. 1 p. 439]) of the ODE

$$\frac{d}{ds} \zeta(s) = W(\zeta(s)) \quad \text{where} \quad W(\zeta) = (0, \bar{h}(\chi) - \tilde{G}(\theta) \omega),$$

i.e.,

$$\begin{cases}
\frac{d}{ds} \chi(s) = 0, \\
\frac{d}{ds} \omega(s) = \bar{h}(\chi(s)) - \tilde{G}(\theta(s)) \omega(s).
\end{cases}$$

(18)

As we will show that the second ODE governing $\omega$ has a unique asymptotically stable equilibrium $\omega_*(\theta, \bar{\omega})$ for every constant function $\chi(t) = \chi = (\theta, \bar{\omega})$, it follows that $(\chi_t, \omega_t)$ converges a.s. towards the set $\{ (\chi, \omega_*(\chi)) : \chi \in \mathbb{R}^{d+m} \}$. In other words, $\lim_{t \to \infty} \| \omega_t - \omega_*(\theta_t, \bar{\omega}_t) \| = 0$, which is the desired result.

We now conclude the proof by verifying among (A1) to (A7) of [Karmakar and Bhatnagar, 2018] the assumptions under which [Karmakar and Bhatnagar, 2018] Lemmas 9 and 10 hold.
(i) (A1): \((Z_t)\) takes values in a compact metric space. Note that it is a finite state-action Markov chain controlled by the sequence \((\theta_t)\).

(ii) (A2): It is easy to see from Eq. (14) that the drift function \(w\) is Lipschitz continuous w.r.t. the variables \(\bar{\omega}, \omega\) uniformly w.r.t. the last variable \(z\) because \(p\) is a probability kernel and the set of states \(S\) is finite.

(iii) (A3): \((\eta^{(1)}_t)\) is a martingale difference sequence w.r.t. the filtration \((\mathcal{F}_t)\). Moreover, since \((R_t)\) is bounded, there exists \(K > 0\) s.t. \(\mathbb{E}[\|\eta^{(1)}_{t+1}\|^2|\mathcal{F}_t] \leq K (1 + \|\omega_t\|^2 + \|\bar{\omega}_t\|^2)\).

(iv) (A4): The stepsizes \((\beta_t)\) satisfy \(\sum_t \beta_t = +\infty\) and \(\sum_t \beta_t^2 < \infty\) as formulated in Assumption 5.2.

(v) (A5): The transition kernel associated to the controlled Markov process \((Z_t)\) is continuous w.r.t. the variables \(z \in \mathcal{S} \times \mathcal{A}, \chi \in \mathbb{R}^{d+m}, \omega \in \mathbb{R}^m\). Continuity (w.r.t. to the metric of the weak convergence of probability measures) is a consequence of the fact that we have a finite-state MDP.

(vi) (A6'): We first note that the inverse of the matrix \(K\) follows from the update rule of (14) that the drift function \(\bar{\omega}\) \(\bar{\omega}\) uniformly w.r.t. the last variable \(z\) because \(p\) is a probability kernel and the set of states \(S\) is finite.

Define the function \(L(\chi, \omega) = \frac{1}{2}\|\bar{G}(\bar{\omega}) - \bar{h}(\bar{\omega})\|^2\). For every \(\chi = (\theta, \bar{\omega}) \in \mathbb{R}^{d+m}\), the function \(L(\chi, \cdot)\) is a Lyapunov function for ODE (15). Indeed, using Lem. A.2 below, we can write
\[
\frac{d}{ds} L(\chi, \omega(s)) = -\langle \bar{h}(\bar{\omega}) - \bar{G}(\bar{\omega})\omega(s), \bar{G}(\bar{\omega}) - \bar{G}(\bar{\omega})\omega(s) \rangle 
\leq -\varepsilon \|\bar{G}(\bar{\omega})\omega(s) - \bar{h}(\bar{\omega})\|^2.
\]

(vii) (A7): The stability Assumption 5.3 ensures that \(\sup_{\theta \in \mathbb{R}^d} \|\bar{G}(\theta)^{-1}\| < \infty\). As a consequence, it also follows from the update rule of \((\bar{\omega}_t)\) that \(\sup_{\theta \in \mathbb{R}^d} \|\bar{\omega}_t\| < +\infty\).

Lemma A.2. Under Assumptions 5.1 and 5.4, there exists \(\varepsilon > 0\) s.t. for all \(\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^m\),
\[
\omega^T \bar{G}(\theta) \omega \geq \varepsilon \|\omega\|^2.
\]

In particular, it holds that \(\sup_{\theta \in \mathbb{R}^d} \|\bar{G}(\theta)^{-1}\| < \infty\).

Proof. Recall that \(\mathcal{K} := \{ \bar{K}_\theta : \theta \in \mathbb{R}^d \}\) where for every \(\theta \in \mathbb{R}^d, \bar{K}_\theta \in \mathbb{R}^{\mathcal{S}|\mathcal{A}| \times \mathcal{S}|\mathcal{A}|}\) is the transition matrix over the state-action pairs defined for every \((s, a), (s', a') \in \mathcal{S} \times \mathcal{A}\) by \(\bar{K}_\theta(s', a'|s, a) = \bar{p}(s'|s, a)\pi_\theta(a'|s)\). We also denote by \(\bar{K}\) the closure of \(\mathcal{K}\). Under Assumption 5.1, there exists a unique stationary distribution \(\mu_K \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}\) for every \(K \in \bar{K}\).

We first show that the map \(K \mapsto \mu_K\) is continuous over the set \(\bar{K}\). The proof of this fact is similar to the proofs of [Zhang et al., 2021 Lem. 9] and [Marbach and Tsitsiklis, 2001 Lem. 1]. We reproduce a similar argument here for completeness. Observe first that \(\mu_K\) satisfies:
\[
M(K)\mu_K = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{where} \quad M(K) := \begin{bmatrix} K^T - I \\ 1 \end{bmatrix}.
\]

As a consequence, since \(M(K)\) has full column rank thanks to Assumption 5.1, the matrix \(M(K)^TM(K)\) is invertible and we obtain a closed form expression for \(\mu_K\) given by:
\[
\mu_K = (M(K)^TM(K))^{-1}M(K)^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{\text{com}(M(K)^TM(K))}{\det(M(K)^TM(K))}M(K)^T \begin{bmatrix} 0 \\ 1 \end{bmatrix},
\]

where \(\text{com}(A)\) stands for the comatrix of the matrix \(A\). Then, it can be seen from this expression that the map \(K \mapsto \mu_K\) is continuous. Note for this that the entries of the comatrix are polynomial functions of the entries of \(M(K)^TM(K)\), and the determinant operator is continuous.

It follows from Assumption 5.1 that for every \(K \in \bar{K}\) and every \((s, a) \in \mathcal{S} \times \mathcal{A}\), \(\mu_K(s, a) > 0\). We deduce from the continuity of the map \(K \mapsto \mu_K\) over the compact set \(\bar{K}\) that \(\inf_{K \in \bar{K}} \mu_K(s, a) > 0\).
every \( \theta \in \mathbb{R}^d \), we obtain that \( \inf_{\theta} \mu_{\rho,\theta}(s,a) > 0 \) where we recall that \( \mu_{\rho,\theta} \) is the unique stationary distribution of the Markov chain induced by \( K_{\theta} \). As a consequence, since \( d_{\rho,\theta}(s) = \sum_{a \in A} \mu_{\rho,\theta}(s,a) \), it also holds that
\[
\inf_{\theta} d_{\rho,\theta}(s) > 0.
\]
Therefore, for every \( \theta \in \mathbb{R}^d \), \( \omega \in \mathbb{R}^m \):
\[
\omega^T \bar{G}(\theta) \omega = (\Phi \omega)^T D_{\rho,\theta} (\Phi \omega) \geq \min_{\theta \in S} \inf_{\theta} d_{\rho,\theta}(s) \|\Phi \omega\|^2 \geq \min_{\theta \in S} \inf_{\theta} d_{\rho,\theta}(s) \lambda_{\min}(\Phi^T \Phi) \|\omega\|^2,
\]
where \( \lambda_{\min}(\Phi^T \Phi) > 0 \) corresponds to the smallest eigenvalue of the symmetric positive definite matrix \( \Phi^T \Phi \) which is invertible thanks to Assumption \( 5.4 \). The proof is concluded by setting \( \epsilon := \lambda_{\min}(\Phi^T \Phi) \cdot \min_{s \in S} \inf_{\theta} d_{\rho,\theta}(s) > 0 \) which is independent of \( \theta \). 

\[ \square \]

### A.1.2 Proof of Prop. 5.2

Recall the definitions of the vector \( h(\theta) \) and the matrix \( G(\theta) \) from Eq. (19):
\[
h(\theta) := \Phi^T D_{\rho,\theta} R_\theta \quad \text{and} \quad G(\theta) := \Phi^T D_{\rho,\theta} (I_n - \gamma P_\theta) \Phi.
\]

We begin the proof by showing the existence of a unique solution \( \bar{\omega}_*(\theta) \) to the linear system \( G(\theta) \bar{\omega} = h(\theta) \). The following lemma establishes the uniform positive definiteness of the matrix \( G(\theta) \) Note that we do not include symmetry in our definition of positive definiteness as in [Bertsekas and Tsitsiklis, 1996]. As a matter of fact, the matrix \( G(\theta) \) is not symmetric in general.

**Lemma A.3.** If Assumptions 5.1 and 5.4 hold, there exists \( \kappa > 0 \) s.t. for all \( \theta \in \mathbb{R}^d \) and \( \omega \in \mathbb{R}^m \),
\[
\omega^T G(\theta) \omega \geq \kappa \|\omega\|^2.
\]

In particular, the matrix \( G(\theta) \) is invertible.

**Proof.** First, we have for every \( \theta \in \mathbb{R}^d \), \( \omega \in \mathbb{R}^m \),
\[
\omega^T G(\theta) \omega = (\Phi \omega)^T D_{\rho,\theta} (I_n - \gamma P_\theta) \Phi \omega = (\Phi \omega)^T D_{\rho,\theta} (\Phi \omega) - \gamma (\Phi \omega)^T D_{\rho,\theta} P_\theta (\Phi \omega).
\]

Then, the Cauchy–Schwarz inequality yields
\[
(\Phi \omega)^T D_{\rho,\theta} P_\theta (\Phi \omega) = (\Phi \omega)^T D_{\rho,\theta}^2 D_{\rho,\theta} P_\theta (\Phi \omega) \leq \|\Phi \omega\|_{D_{\rho,\theta}} \|P_\theta \Phi \omega\|_{D_{\rho,\theta}}.
\]

Notice now that we cannot use the classical result [Tsitsiklis and Van Roy, 1997, Lem. 1] to obtain that \( \|P_\theta V\|_{D_{\rho,\theta}} \leq \|V\|_{D_{\rho,\theta}} \) for any \( V \in \mathbb{R}^n \) because \( D_{\rho,\theta} \) is not the stationary distribution of the kernel \( P_\theta \) but it is instead associated to the artificial kernel \( P_\theta \). Nevertheless, the following lemma provides an analogous result with a similar proof.

**Lemma A.4.** For every \( \theta \in \mathbb{R}^d \), \( V \in \mathbb{R}^n \), we have
\[
\|P_\theta V\|^2_{D_{\rho,\theta}} \leq \frac{1}{\gamma} \|V\|^2_{D_{\rho,\theta}} - \frac{1 - \gamma}{\gamma} \|V\|_\rho^2 \leq \frac{1}{\gamma} \|V\|^2_{D_{\rho,\theta}}.
\]

**Proof.** It follows from Jensen’s inequality that
\[
\|P_\theta V\|^2_{D_{\rho,\theta}} = \sum_{i=1}^n d_{\rho,\theta}(s_i) \left( \sum_{j=1}^n \rho(s_j | s_i) V_j \right)^2 \leq \sum_{i=1}^n d_{\rho,\theta}(s_i) \sum_{j=1}^n \rho(s_j | s_i) V_j^2.
\]

Then, observe that \( \hat{P}_\theta = \gamma P_\theta + (1 - \gamma) \mathbb{1} \rho^T \) as a consequence of Eq. (4). By plugging this formula and then using the fact that \( d_{\rho,\theta}^T \hat{P}_\theta = d_{\rho,\theta}^T \rho^T \), we obtain
\[
\sum_{i=1}^n d_{\rho,\theta}(s_i) \sum_{j=1}^n \rho(s_j | s_i) V_j^2 = \frac{1}{\gamma} \left[ \left( \sum_{i=1}^n d_{\rho,\theta}(s_i) \hat{P}_\theta(s_j | s_i) V_j^2 \right) - (1 - \gamma) \sum_{j=1}^n \rho(s_j) V_j^2 \right] = \frac{1}{\gamma} \left[ \left( \sum_{i=1}^n d_{\rho,\theta}(s_i) \rho(s_j) V_j^2 \right) - (1 - \gamma) \sum_{j=1}^n \rho(s_j) V_j^2 \right] = \frac{1}{\gamma} \|V\|^2_{D_{\rho,\theta}} - \frac{1 - \gamma}{\gamma} \|V\|_\rho^2.
\]
which concludes the proof of Lem. A.4.

We now complete the proof of Lem. A.3. From Eq. (21), Lem. A.4 with $V = \Phi \omega$ yields
\[
(\Phi \omega)^T D_{p,0} \Phi \omega \leq \frac{1}{\sqrt{\gamma}} \|\Phi \omega\|^2_{D_{p,0}} = \frac{1}{\sqrt{\gamma}} (\Phi \omega)^T D_{p,0} (\Phi \omega).
\]
Whence, we obtain from Eq. (20) that
\[
\omega^T G(\theta) \omega \geq (1 - \sqrt{\gamma}) (\Phi \omega)^T D_{p,0} (\Phi \omega) \geq \varepsilon (1 - \sqrt{\gamma}) \|\omega\|^2,
\]
where the last inequality stems from Lem. A.2.

We now prove the remaining convergence results. We start with the first result showing that the sequence $(\bar{\omega}_t)$ tracks $\bar{\omega}_\ast(\theta_t)$. From the update rules of the sequences $(\bar{\omega}_t)$ and $(\omega_t)$ (Eqs. (9)-(10)), we can introduce the quantity $\omega_\ast(\theta_t, \bar{\omega}_t)$ as defined in Prop. 5.1 to obtain
\[
\bar{\omega}_{t+1} = \bar{\omega}_t + \xi_t (\bar{\omega}_{t+1} - \bar{\omega}_t) = \bar{\omega}_t + \xi_t (\omega_\ast(\theta_t, \bar{\omega}_t) - \bar{\omega}_t)
\]
Then, using the expressions of $\bar{h}, \bar{G}$ in Eq. (11) and $h, G$ in Eq. (12), we can write
\[
\omega_\ast(\theta_t, \bar{\omega}_t) - \bar{\omega}_t = \bar{G}(\theta)^{-1} (\bar{h}(\theta_t, \bar{\omega}_t) - \bar{G}(\theta_t) \bar{\omega}_t) = \bar{G}(\theta)^{-1} (h(\theta_t) - G(\theta) \bar{\omega}_t).
\]
As a consequence,
\[
\bar{\omega}_{t+1} = \bar{\omega}_t + \xi_t \bar{G}(\theta)^{-1} (h(\theta_t) - G(\theta_t) \bar{\omega}_t) + \xi_t (\omega_t - \omega_\ast(\theta_t, \bar{\omega}_t) + \beta \omega_t \omega_t, \omega_t, Z_t) + \eta_t \beta \eta_t^{(1)}.
\]
Therefore, the sequence $(\bar{\omega}_t)$ satisfies a linear stochastic approximation scheme driven by the slowly varying Markov chain $(\theta_t)$ evolving on a slower timescale than the iterates $(\bar{\omega}_t)$. We proceed similarly to the proof of Prop. 5.1.

Recall the notation $\chi_t = (\theta_t, \bar{\omega}_t)$. Let $\chi = (\theta, \bar{\omega}) \in \mathbb{R}^{d+m}$, $U(\chi) = (0, \bar{G}(\theta)^{-1} (h(\theta) - G(\theta) \bar{\omega})))$. Then,
\[
\chi_{t+1} = \chi_t + \xi_t U(\chi_t) + \tilde{\varepsilon}_t
\]
where $\tilde{\varepsilon}_t = (\frac{\alpha_t}{\xi_t} + \frac{1}{\sqrt{\gamma}} \delta t + \psi_t (\bar{S}_t, \bar{A}_t), \omega_t - \omega_\ast(\theta_t, \bar{\omega}_t) + \beta \omega_t \omega_t, \omega_t, Z_t) + \eta_t \beta \eta_t^{(1)}$.

It can be shown that $\tilde{\varepsilon}_t \to 0$ as $t \to +\infty$. Note for this that $\alpha_t / \xi_t \to 0$ and $\beta_t \to 0$ by Assumption 3.2, the boundedness of the reward function $R$ and the fact that the sets $S, A$ are finite. Moreover, Assumption 5.2 ensures that $\sum \xi_t = +\infty$ and $\sum \xi_t^2 < +\infty$.

Furthermore, one can show that the function $U$ is Lipschitz continuous. For this, remark that:

(a) The function $U$ is affine in $\bar{\omega}$.

(b) The functions $\theta \mapsto R_0$ and $\theta \mapsto P_0$ are Lipschitz continuous as $P_0(s'|s) = p(s'|s,a)\pi_0(a|s)$, $R_0(s) = \sum_{a \in A} R(s,a)\pi_0(a|s)$ and Assumption 3.1(b) guarantees that $\theta \mapsto \pi_0(a|s)$ is Lipschitz continuous for every $(s,a) \in S \times A$.

(c) The function $\theta \mapsto D_{p,0}$ is Lipschitz continuous. We refer to [Zhang et al., 2021] Lem. 9] for a proof.

(d) The function $\theta \mapsto \bar{G}(\theta)^{-1}$ is Lipschitz continuous. Observe for this that for every $\theta, \theta' \in \mathbb{R}^d$, $\bar{G}(\theta)^{-1} - \bar{G}(\theta')^{-1} = \bar{G}(\theta)^{-1} (\bar{G}(\theta') - \bar{G}(\theta)) \bar{G}(\theta')^{-1}$ and that $\sup_\theta \|\bar{G}(\theta)^{-1}\| < \infty$ using Lem. A.2.

(e) The reward function $R$ is bounded and the entries of the matrices $D_{p,0}$ and $P_0$ are bounded by one.
Using classical stochastic approximation results (see, for e.g., [Benaim, 1996 Th.1.2]), we obtain that the sequence \((\chi_t)\) converges a.s. towards an internally chain transitive set of the ODE \(\frac{d}{ds} \chi(s) = U(\chi(s)),\) i.e.,
\[
\begin{align*}
\frac{d}{ds} \theta(s) &= 0, \\
\frac{d}{ds} \omega(s) &= \bar{G}(\theta(s))^{-1}(h(\theta(s)) - G(\theta(s))\omega(s)).
\end{align*}
\]
(25)

We conclude by showing that for every \(\theta \in \mathbb{R}^d\), the ODE \(\frac{d}{ds} \omega(s) = \bar{G}(\theta)^{-1}(h(\theta) - G(\theta)\omega(s))\) has a globally asymptotically stable equilibrium \(\bar{\omega}_*\). This result holds if the matrix \(-\bar{G}(\theta)^{-1}G(\theta)\) is Hurwitz, i.e., all its eigenvalues have negative real parts. We show this result in Lem. A.5 below.

Then, it follows that \(\chi_t = (\theta_t, \bar{\omega}_t)\) converges a.s. towards the set \(\{(\theta, \bar{\omega}_*) : \theta \in \mathbb{R}^d\}\). This yields the desired result \(\lim_t \|\bar{\omega}_t - \bar{\omega}_*(\theta_t)\| = 0\).

**Lemma A.5.** For every \(\theta \in \mathbb{R}^d\), the matrix \(-\bar{G}(\theta)^{-1}G(\theta)\) is Hurwitz.

**Proof.** We first recall Lyapunov’s theorem which characterizes Hurwitz matrices (see, for e.g., [Horn and Johnson, 1994, Th.2.2.1 p. 96]). A complex matrix \(A\) is Hurwitz if and only if there exists a positive definite matrix \(M = M^*\) s.t. \(A^*M + MA\) is negative definite, where \(M^*\) and \(A^*\) are the complex conjugate transposes of \(M\) and \(A\). We use this theorem with \(A = -\bar{G}(\theta)^{-1}G(\theta)\) and \(M = G(\theta)\) which is symmetric by definition and positive definite thanks to Lem. A.2. Then, we obtain that
\[
A^*M + MA = -G(\theta)^T \bar{G}(\theta)^{-1}G(\theta) - G(\theta) \bar{G}(\theta)^{-1}G(\theta) = -(G(\theta)^T + G(\theta)).
\]
We conclude the proof by showing that \(G(\theta)^T + G(\theta)\) is a (symmetric) positive definite matrix. For that, observe that for every nonzero vector \(\omega \in \mathbb{R}^m\), it holds that \(\omega^T (G(\theta)^T + G(\theta))\omega = 2\omega^TG(\theta)\omega > 0\) where the positivity stems from Lem. A.3.\]

The last result states that for every \(\theta \in \mathbb{R}^d\), \(\Phi\bar{\omega}_*(\theta)\) is a fixed point of the projected Bellman operator \(\Pi_\theta T_\theta\). This is a consequence of the following derivations:
\[
\Pi_\theta T_\theta(\Phi\bar{\omega}_*(\theta)) = \Phi\bar{G}(\theta)^{-1}\Phi^T D_{\theta^0} T_\theta(\Phi\bar{\omega}_*(\theta)) = \Phi\bar{G}(\theta)^{-1}\Phi^T D_{\theta^0}(R_\theta + \gamma P_\theta\Phi\bar{\omega}_*(\theta)) = \Phi\bar{G}(\theta)^{-1}h(\theta) + \Phi\bar{G}(\theta)^{-1}(\bar{G}(\theta) - G(\theta))\bar{G}(\theta)^{-1}h(\theta) = \Phi\bar{G}(\theta)^{-1}h(\theta) + \Phi G(\theta)^{-1}h(\theta) - \Phi\bar{G}(\theta)^{-1}h(\theta) = \Phi\bar{G}(\theta)^{-1}h(\theta) = \Phi\bar{\omega}_*(\theta),
\]
(26)
where the first equality uses the expression of the projection \(\Pi_\theta\), the second one uses the definition of the Bellman operator \(T_\theta\) and the third one stems from the definitions of the matrices \(\bar{G}(\theta)\) and \(G(\theta)\) (see Eqs. \(11\) and \(12\)).

**A.1.3 Proof of Th. 5.3**

The proof of Th. 5.3 uses both Prop. 5.1 and Prop. 5.2.

In order to show that \(\lim_t \|\bar{\omega}_t - \bar{\omega}_*(\theta_t)\| = 0\) w.p.1, we prove the following two results:

(a) \(\lim_t \|\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_*(\theta_t))\| = 0\) w.p.1.

(b) \(\bar{\omega}_*(\theta, \bar{\omega}_*(\theta)) = \bar{\omega}_*(\theta)\) for all \(\theta \in \mathbb{R}^d\).

(a) We have the decomposition
\[
\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_*(\theta_t)) = [\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_t)] + [\bar{\omega}_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_*(\theta_t, \bar{\omega}_*(\theta_t))],
\]
\[
= [\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}(\bar{h}(\theta_t, \bar{\omega}_t) - h(\theta_t, \bar{\omega}_*(\theta_t)))
\]
\[
= [\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}\Phi^T D_{\theta^0} P_{\theta^0} \Phi(\bar{\omega}_t - \bar{\omega}_*(\theta_t))
\]
\[
= [\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}(\bar{G}(\theta_t) - G(\theta_t))(\bar{\omega}_t - \bar{\omega}_*(\theta_t))
\]
\[
= [\bar{\omega}_t - \bar{\omega}_*(\theta_t, \bar{\omega}_t)] + (I_{\bar{G}} - \bar{G}(\theta_t)^{-1}G(\theta_t))(\bar{\omega}_t - \bar{\omega}_*(\theta_t)).
\]
(27)
It follows from Prop. 5.1 that the first term in the above decomposition goes to zero. Then, observe that sup_θ ||G(θ)^{-1}|| < ∞ given Lem. 3.2 and sup_θ ||G(θ)|| < ∞ thanks to the boundedness of the matrices F_θ and D_{ρ,θ} uniformly in θ. As a consequence, the second term also converges to zero using Prop. 5.2.

(b) Using the definitions of the functions ω_s and ¯ω_s, we can write for every θ ∈ R^d,

$$\omega_s(θ, ¯ω_s(θ)) = \bar{G}(θ)^{-1}h(θ, ¯ω_s(θ))$$

$$= \bar{G}(θ)^{-1}\Phi^T D_{ρ,θ}(R_θ + γP_θ \Phi G(θ)^{-1}h(θ))$$

$$= \bar{G}(θ)^{-1}(h(θ) + \gamma \Phi^T D_{ρ,θ}P_θ \Phi G(θ)^{-1}h(θ))$$

$$= \bar{G}(θ)^{-1}I_n + \gamma \Phi^T D_{ρ,θ}P_θ \Phi G(θ)^{-1}h(θ)$$

$$= \bar{G}(θ)^{-1}(G(θ) + \gamma \Phi^T D_{ρ,θ}P_θ \Phi G(θ)^{-1}h(θ))$$

$$= \bar{G}(θ)^{-1}G(θ)G(θ)^{-1}h(θ)$$

$$= \omega_s(θ).$$

For the last result, we write

$$\|Π_{θ_t}T_{θ_t}(Φω_t) - Φω_t\| = \|Φ (G(θ_t)^{-1}\Phi^T D_{ρ,θ_t}T_{θ_t}(Φω_t) - ω_t)\|$$

$$= \|Φ (G(θ_t)^{-1}\Phi^T D_{ρ,θ_t}(T_{θ_t}(Φω_t) - Φω_t))\|$$

$$= \|Φ (G(θ_t)^{-1}(h(θ_t) - G(θ_t)ω_t))\|$$

$$= \|Φ\|\|G(θ_t)^{-1}\|\|G(θ_t)\|\|ω_t - ¯ω_s(θ_t)\|.$$  

(28)

Then, as previously mentioned in the proof, observe that sup_θ ||G(θ)^{-1}|| < ∞ and sup_θ ||G(θ)|| < ∞. Since ω_t - ¯ω_s(θ_t) → 0 as t → ∞, the result follows.

### A.2 Proof of Th. 5.4 actor analysis

In this subsection, we present a proof of Th. 5.4 which is similar in spirit to the proof in [Konda and Tsitsiklis, 2003](#) Sec. 6. Recall the notation Z_t = (S_t, A_t). Note that (Z_t) is a Markov chain. The actor parameter θ_t iterates as follows:

$$θ_{t+1} = θ_t + α_t \frac{1}{1 - γ} δ_{t+1}ψ_{θ_t}(S_t, A_t)$$

$$= θ_t + α_t \frac{1}{1 - γ} (R_{t+1} + (γφ(S_{t+1}) - φ(S_t))^T ω_t)ψ_{θ_t}(S_t, A_t)$$

$$= θ_t + α_t \frac{1}{1 - γ} (R(\tilde{S}_t, \tilde{A}_t)ψ_{θ_t}(\tilde{S}_t, \tilde{A}_t) + H_θ(Z_t)ω_t) + α_t \frac{1}{1 - γ} \tilde{η}_{t+1},$$

where for every θ ∈ R^d, z = (s, a) ∈ S × A,

$$H_θ(z) = ψ_θ(s, a) \left( γ \sum_{s' ∈ S} p(s'|s, a)φ(s') - φ(s) \right)^T,$$

and (\tilde{η}_{t+1}) is an R^d-valued F_t-martingale difference sequence defined by

$$\tilde{η}_{t+1} = (R_{t+1} - \mathbb{E}[R_{t+1}|F_t])ψ_{θ_t}(\tilde{S}_t, \tilde{A}_t) + γψ_{θ_t}(\tilde{S}_t, \tilde{A}_t)(φ(S_{t+1}) - \mathbb{E}[φ(S_{t+1})|F_t])^T ω_t.$$  

(29)

We now introduce the steady-state expectation of the main term H_θ(Z_t)ω_t + R(\tilde{S}_t, \tilde{A}_t)ψ_θ(\tilde{S}_t, \tilde{A}_t). Recall that μ_{ρ,θ} is the stationary distribution of the Markov chain (Z_t). Define the functions H : R^d → R^{d×m} and u : R^d → R^d for every θ ∈ R^d by

$$H(θ) = \mathbb{E}_{Z \sim μ_{ρ,θ}}[H_θ(Z)],$$

(30)

$$u(θ) = \mathbb{E}_{Z \sim μ_{ρ,θ}}[R(\tilde{S}, \tilde{A})ψ_θ(\tilde{S}, \tilde{A})],$$

(31)
where $Z = (\tilde{S}, \tilde{A})$ is a random variable following the distribution $\mu_{p, \theta}$.

Then, we introduce the quantity $\bar{\omega}_s(\theta_i)$ which approximates well $\omega_t$ for large $t$ (in the sense of Th. 5.3) and only depends on the actor parameter $\theta_i$. We obtain the following decomposition

$$
\theta_{t+1} = \theta_t + \alpha_t f(\theta_t) + \alpha_t \frac{1}{1 - \gamma} (\tilde{\eta}_{t+1} + e^{(1)}_t + e^{(2)}_t),
$$

(32)

where the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and the error terms $e^{(1)}_t$ and $e^{(2)}_t$ are defined as follows

$$
f(\theta) = \frac{1}{1 - \gamma} (\tilde{H}(\theta) \bar{\omega}_s(\theta) + u(\theta)),
$$

(33)

$$
e^{(1)}_t = \langle R(\tilde{S}_t, \tilde{A}_t) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) + H_{\theta_t}(Z_t) \bar{\omega}_s(\theta_t) - (\tilde{H}(\theta_t) \bar{\omega}_s(\theta_t) + u(\theta_t)),
$$

(34)

$$
e^{(2)}_t = H_{\theta_t}(Z_t) (\omega_t - \bar{\omega}_s(\theta_t)).
$$

(35)

The bias induced by the approximation of $\nabla J(\theta)$ by our actor-critic algorithm is defined for every $\theta \in \mathbb{R}^d$ by

$$
b(\theta) := f(\theta) - \nabla J(\theta).
$$

(36)

This bias is due to the linear FA of the true state-value function. It is defined as the difference between the steady-state expectation of the actor update given by the function $f$ defined in Eq. (33) and the gradient $\nabla J(\theta)$ we are interested in. The following lemma provides a more explicit and interpretable expression for the bias $b(\theta)$. The state-value function $V_{\pi_\theta}$ will be seen as a vector of $\mathbb{R}^{|S|}$.

Lemma A.6. For every $\theta \in \mathbb{R}^d$,

$$
b(\theta) = \gamma \sum_{s \in S, a \in A} \mu_{p, \theta}(s, a) \psi_\theta(s, a) \sum_{s' \in S} p(s'|s, a) (\phi(s')^T \bar{\omega}_s(\theta) - V_{\pi_\theta}(s')).
$$

Proof. The expression follows from using the definition of $b(\theta)$ and computing both the function $\tilde{H}$ defined in Eq. (30) and the gradient of the function $J$.

First, we explicit the function $\tilde{H}$, writing

$$
\tilde{H}(\theta) = \mathbb{E}_{Z \sim \mu_{p, \theta}} [H_\theta(Z)] = \mathbb{E}_{Z \sim \mu_{p, \theta}} \left[ \psi_\theta(\tilde{S}, \tilde{A}) \left( \gamma \sum_{s' \in S} p(s'|\tilde{S}, \tilde{A}) \phi(s') - \phi(\tilde{S}) \right)^T \right]
$$

$$
= \sum_{s \in S, a \in A} \mu_{p, \theta}(s, a) \psi_\theta(s, a) \left( \gamma \sum_{s' \in S} p(s'|s, a) \phi(s')^T - \phi(s)^T \right)
$$

$$
= \gamma \sum_{s \in S, a \in A} \mu_{p, \theta}(s, a) \psi_\theta(s, a) \sum_{s' \in S} p(s'|s, a) \phi(s')^T,
$$

(37)

where the last equality stems from remarking that $\sum_{a \in A} \mu_{p, \theta}(s, a) \psi_\theta(s, a) = 0$.

Then, the policy gradient theorem as formulated in Eq. (1) and the definition of the advantage function provide

$$
(1 - \gamma) \nabla J(\theta) = \mathbb{E}_{Z \sim \mu_{p, \theta}} [\Delta_{\pi_\theta}(\tilde{S}, \tilde{A}) \psi_\theta(\tilde{S}, \tilde{A})]
$$

$$
= \mathbb{E}_{Z \sim \mu_{p, \theta}} [(R(\tilde{S}, \tilde{A}) + \gamma \sum_{s' \in S} p(s'|\tilde{S}, \tilde{A}) V_{\pi_\theta}(s') - V_{\pi_\theta}(\tilde{S})) \psi_\theta(\tilde{S}, \tilde{A})]
$$

$$
= \sum_{s, a} \mu_{p, \theta}(s, a) (R(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_{\pi_\theta}(s') - V_{\pi_\theta}(s)) \psi_\theta(s, a)
$$

$$
= u(\theta) + \gamma \sum_{s \in S, a \in A} \mu_{p, \theta}(s, a) \psi_\theta(s, a) \sum_{s' \in S} p(s'|s, a) V_{\pi_\theta}(s').
$$

(38)

The result stems from using the definition of $b(\theta)$ together with Eqs. (37) and (38).
Using a second-order Taylor expansion of the $\hat{L}$-Lipschitz function $\nabla J$ (again see [Zhang et al., 2020a, Lem. 4.2]) together with Eq. (32), we can derive the following inequalities
\[
J(\theta_{t+1}) \geq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle - L\|\theta_{t+1} - \theta_t\|^2, \\
\geq J(\theta_t) + \alpha_t \langle \nabla J(\theta_t), f(\theta_t) \rangle + \frac{\alpha_t}{1-\gamma} \langle \nabla J(\theta_t), \tilde{\eta}_{t+1} \rangle + e_t^{(1)} + e_t^{(2)} + \tilde{L} \frac{\alpha_t^2}{(1-\gamma)^2} \|\delta_{t+1} \psi_\theta(\tilde{S}_t, \tilde{A}_t)\|^2.
\]  
(39)

The above inequality consists of a main term involving the function $f$ and noise terms. The following lemma controls these noise terms which are shown to be negligible.

**Lemma A.7.** (a) $\sum_{t=0}^{\infty} \alpha_t \langle \nabla J(\theta_t), e_t^{(1)} \rangle < \infty$ w.p.1, 
(b) $\sum_{t=0}^{\infty} \alpha_t \langle \nabla J(\theta_t), \tilde{\eta}_{t+1} \rangle < \infty$ w.p.1, 
(c) $\lim_{t \to \infty} e_t^{(2)} = 0$ w.p.1, 
(d) $\sum_{t=0}^{\infty} \alpha_t^2 \|\delta_{t+1} \psi_\theta(\tilde{S}_t, \tilde{A}_t)\|^2 < \infty$ w.p.1.

**Proof.** (a) The proof is based on the classical decomposition of the Markov noise term $e_t^{(1)}$ using the Poisson equation [Benveniste et al., 1990, p. 222-229]. We refer to [Zhang et al., 2020b, Lem. 7 and Sec. A.8.3] for a detailed proof using this technique. The proof of our result here follows the same line. For conciseness, we only describe the necessary tools, pointing out the differences with [Zhang et al., 2020b, Lem. 7 and Sec. A.8.3] which is concerned with a different algorithm.

Let $Z := \mathcal{S} \times \mathcal{A}$. First, define the functions $g^*_\theta : Z \to \mathbb{R}^d$ and $\bar{g} : \mathbb{R}^d \to \mathbb{R}^d$ by:
\[
g^*_\theta(z) := R(z) \psi_\theta(z) + H_\theta(z) \bar{\omega}_\theta(z), \\
\bar{g}(\theta) := u(\theta) + \tilde{H}(\theta) \bar{\omega}_\theta(z),
\]
(40)
(41)
for every $z = (s, a) \in Z, \theta \in \mathbb{R}^d$. Observe in particular that $e_t^{(1)} = g^*_\theta(\tilde{S}_t, \tilde{A}_t) - \bar{g}(\theta_t)$. Recall that for every $\theta \in \mathbb{R}^d$, the kernel transition $K_\theta$ is defined for every $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$ by $K_\theta(s', a') = \bar{p}(s', s) \pi_{\theta}(a'|s)$ (see Assumption 5.1). The idea of the proof is to introduce for each integer $i = 1, \ldots, d$ a Markov Reward Process (MRP) [Puterman, 2014, Sec. 8.2] on the space $Z$ induced by the transition kernel $K_\theta$ and the reward function $g^*_\theta, i$ (ith coordinate of the function $g^*_\theta$). As a consequence, the corresponding average reward is given by $g_i(\theta)$ (ith coordinate of $\bar{g}(\theta)$). Then, the differential value function of the MRP is provided by $v_{g, i} := (I - K_\theta + \mu_{\theta}^T K_\theta)^{-1} (1 - I - \mu_{\theta}^T g^*_\theta) g^*_\theta$. As shown for instance in [Puterman, 2014, Sec. 8.2]. The functions $v_{g, i}$ for $i = 1, \ldots, d$ define together a vector valued function $v_\theta : Z \to \mathbb{R}^d$. Under Assumption 5.1, using similar arguments to the proof of Lemma A.2 (see also [Zhang et al., 2021, Sec. 4.2]), we can show that the function $K \in \hat{C} \mapsto (I - K + \mu_{\theta}^T K)^{-1} (I - 1 - \mu_{\theta}^T g^*_\theta)$ is continuous on the compact set $\hat{C}$. It follows that $\sup_{\theta, z} \|v_\theta(z)\| < \infty$ because $K_\theta$ is bounded for every $\theta \in \mathbb{R}^d$ and $g^*_\theta$ is uniformly bounded w.r.t. $\theta$ under our assumptions. Moreover, the differential value function satisfies the crucial Bellman equation:
\[
v_\theta(z) = g^*_\theta(z) - \bar{g}(\theta) + \sum_{z' \in Z} K_\theta(z'|z) v_\theta(z'),
\]
for every $z \in Z$. We use the above Poisson equation to express $e_t^{(1)} = g^*_\theta(\tilde{S}_t, \tilde{A}_t) - \bar{g}(\theta_t)$ using $v_\theta$. The rest of the proof follows the same line as [Zhang et al., 2020b, Lem. 7 and Sec. A.8.3].

(b) First, recall that $(\tilde{\eta}_t)$ is a martingale difference sequence adapted to $\mathcal{F}_t$ and so is $(\langle \nabla J(\theta_t), \tilde{\eta}_{t+1} \rangle)$. Using the boundedness of the function $\theta \mapsto \psi_\theta(s, a)$ guaranteed by Assumption 3.1(c) with the boundedness of the rewards sequence $(R_t)$, the sequence $(\omega_t)$ (Assumption 5.3) and the gradient $\nabla J$, one can show by Cauchy-Schwarz inequality that there exists a constant $C > 0$ s.t. $\mathbb{E}\|\langle \nabla J(\theta_t), \tilde{\eta}_{t+1} \rangle\|^2 |\mathcal{F}_t| \leq C$ a.s. Then, using that $\sum_{t} \alpha_t^2 < \infty$ (Assumption 5.2), it follows that $\sum_{t} \mathbb{E}\|\alpha_t \langle \nabla J(\theta_t), \tilde{\eta}_{t+1} \rangle\|^2 |\mathcal{F}_t| < \infty$ a.s. We deduce from Doob's convergence theorem that item (b) holds.

(c) As for item (c), we first observe that $\tilde{H}(\theta_t)$ is bounded since $\theta \mapsto \psi_\theta(s, a)$ is bounded for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ thanks again to Assumption 3.1(c). Then, item (c) stems from the fact that $\omega_t - \bar{\omega}_\theta(\theta_t) \to 0$ as shown in Th. 5.3.
(d) Similarly to $\tilde{H}(\theta_t)$, upon noticing that the reward sequence $(R_t)$ is bounded by $U_R$ and the sequence $(\omega_t)$ is a.s. bounded by Assumption 5.3, the quantity $\delta_{t+1} \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)$ is also a.s. bounded. Then, item (d) is a consequence of the square summability of the stepsizes $\alpha_t \left( \sum \alpha_t^2 < \infty \right)$ as guaranteed by Assumption 5.2.

The end of the proof follows the same line as [Konda and Tsitsiklis, 2003b, p. 1163] (see also [Konda, 2002, p. 86]). We reproduce the argument here for completeness. Let $T > 0$. Define a sequence $k_t$ by

$$k_0 = 0, \quad k_{t+1} = \min \left\{ k \geq k_t : \sum_{i=k_t}^{k} \alpha_i \geq T \right\} \quad \text{for } t > 0.$$  

Using Eq. (39) together with the Cauchy-Schwarz inequality and Eq. (36), we can write

$$J(\theta_{k_{t+1}}) \geq J(\theta_{k_t}) + \sum_{k=k_t}^{k_{t+1}-1} \alpha_k \left( \| \nabla J(\theta_k) \|^2 - \| b(\theta_k) \| \cdot \| \nabla J(\theta_k) \| \right) + v_t,$$

where $v_t$ is defined by

$$v_t = \sum_{k=k_t}^{k_{t+1}-1} \left( \frac{\alpha_k}{1 - \gamma} \left( \nabla J(\theta_k), \tilde{\eta}_{k+1} + \epsilon_k^{(1)} + \epsilon_k^{(2)} \right) - \frac{\alpha_k^2}{(1 - \gamma)^2} \| \delta_{k+1} \psi_{\theta_k}(\tilde{S}_k, \tilde{A}_k) \|^2 \right).$$

It stems from Lem. A.7 that $v_t \to 0$ as $t \to +\infty$. By contradiction, if the result does not hold, the sequence $J(\theta_k)$ would increase indefinitely. This contradicts the boundedness of the function $J$ (note that $\theta \mapsto V_{\pi_{\theta}}$ is bounded since the rewards are bounded).

**B  Proofs for Sec. 6: finite-time analysis**

Throughout our finite-time analysis, we will not track all the constants although these can be precisely determined. We will in particular explicit the dependence on the effective horizon $1/(1 - \gamma)$ and the cardinal $|A|$ of the action space. The universal constant $C$ may change from line to line and from inequality to inequality. It may depend on constants of the problem s.t. the Lipschitz constants of the functions $J, \theta \mapsto \psi_{\theta}, \theta \mapsto \pi_{\theta}$, upperbounds of the rewards and the score function $\psi_{\theta}$.

**B.1  Proof of Th. 6.1** finite-time analysis of the critic

The proof is inspired from the recent works [Wu et al., 2020] [Shen et al., 2020]. However, it significantly deviates from these works because of the use of a target variable $\tilde{\omega}$ in Algorithm 1. In particular, as previously mentioned, Algorithm 1 involves three different timescales whereas the actor-critic algorithms considered in [Wu et al., 2020] [Shen et al., 2020] only use two different timescales respectively associated to the critic and the actor.

We follow a similar strategy to our asymptotic analysis of the critic. Indeed, our non-asymptotic analysis consists of two main steps based on the following decomposition:

$$\omega_t - \tilde{\omega}_s(\theta_t) = \omega_t - \omega_s(\theta_t, \tilde{\omega}_t) + \omega_s(\theta_t, \tilde{\omega}_t) - \tilde{\omega}_s(\theta_t)$$

$$= \omega_t - \omega_s(\theta_t, \tilde{\omega}_t) + \omega_s(\theta_t, \tilde{\omega}_t) - \omega_s(\theta_t, \omega_s(\theta_t))$$

$$= \omega_t - \omega_s(\theta_t, \tilde{\omega}_t) + \tilde{G}(\theta_t)^{-1} (\tilde{h}(\theta_t, \tilde{\omega}_t) - \tilde{h}(\theta_t, \omega_s(\theta_t))).$$

Hence, it is sufficient to obtain a control of the convergence rates of the quantities $\omega_t - \omega_s(\theta_t, \tilde{\omega}_t)$ and $\tilde{\omega}_t - \omega_s(\theta_t)$. We already know that these quantities converge a.s. to zero thanks to Props. 5.1 and 5.2. We conduct a finite-time analysis of each of the terms separately in the subsections below and combine the obtained results to conclude the proof.
We start by introducing a few useful shorthand notations. Let $\tilde{x}_t := (\tilde{S}_t, \tilde{A}_t, S_{t+1})$. Define for every $\tilde{x} = (\tilde{s}, \tilde{a}, s) \in S \times A \times S$ and every $\tilde{\omega}, \omega \in \mathbb{R}^m$:

$$
\bar{d}(\tilde{x}, \tilde{\omega}, \omega) = R(\tilde{s}, \tilde{a}) + \gamma \phi(s)^T \tilde{\omega} - \phi(\tilde{s})^T \omega, \\
g(\tilde{x}, \tilde{\omega}, \omega) = \bar{d}(\tilde{x}, \tilde{\omega}, \omega) \phi(\tilde{s}).
$$

(43)

(44)

Finally, define for every $\theta \in \mathbb{R}^d$ the steady-state expectation:

$$
\bar{g}(\theta, \tilde{\omega}, \omega) = \mathbb{E}_{\tilde{s} \sim \tilde{a}, \tilde{a} \sim \pi_\theta(s \sim p(|\tilde{s}, \tilde{a}|)}[g(\tilde{x}, \tilde{\omega}, \omega)] = \bar{h}(\theta, \omega) - G(\theta)\omega.
$$

(45)

**B.1.1 Control of the first error term $\omega_t - \omega_*(\theta_t, \bar{\omega}_t)$**

We introduce an additional shorthand notation for brevity:

$$
\nu_t := \omega_t - \omega_*(\theta_t, \bar{\omega}_t).
$$

**Decomposition of the error.** Using the update rule of the critic gives

$$
\|\nu_{t+1}\|^2 = \|\omega_t + \beta_t g(\tilde{x}_t, \tilde{\omega}_t, \omega_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\|^2 \\
= \|\nu_t + \beta_t g(\tilde{x}_t, \tilde{\omega}_t, \omega_t) + \omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\|^2.
$$

Then, we develop the squared norm and use the classical inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to obtain

$$
\|\nu_{t+1}\|^2 \leq \|\nu_t\|^2 + 2\beta_t \langle \nu_t, g(\tilde{x}_t, \tilde{\omega}_t, \omega_t) \rangle + 2\|\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\|^2 + 2\|\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\|^2 + 2C\beta_t^2.
$$

(46)

Now, we decompose the first inner product into a main term generating a repelling effect and a second Markov noise term as follows

$$
\langle \nu_t, g(\tilde{x}_t, \tilde{\omega}_t, \omega_t) \rangle = \langle \nu_t, \bar{g}(\theta_t, \tilde{\omega}_t, \omega_t) \rangle + \Lambda(\theta_t, \bar{\omega}_t, \omega_t, \tilde{x}_t),
$$

(47)

where we used the shorthand notation

$$
\Lambda(\theta, \tilde{\omega}, \omega, \tilde{x}) := \langle \omega - \omega_*(\theta, \bar{\omega}), g(\tilde{x}, \tilde{\omega}, \omega) - \bar{g}(\theta, \omega, \bar{\omega}) \rangle.
$$

(48)

We control the first term in Eq. (47) as follows

$$
\langle \nu_t, \bar{g}(\theta_t, \tilde{\omega}_t, \omega_t) \rangle = \langle \nu_t, \bar{g}(\theta_t, \tilde{\omega}_t, \omega_t) \rangle - \bar{g}(\theta_t, \tilde{\omega}_t, \omega_*(\theta_t, \bar{\omega}_t)) = -\langle \nu_t, \bar{G}(\theta_t) \nu_t \rangle \leq -\varepsilon \|\nu_t\|^2.
$$

(49)

We used the fact that $\bar{g}(\theta_t, \tilde{\omega}_t, \omega_*(\theta_t, \bar{\omega}_t)) = 0$ for the first equality and Lem. A.2 for the inequality. Then, it can be shown that

$$
\|\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\| \leq C(\|\bar{\omega}_t - \bar{\omega}_{t+1}\| + \|\tilde{\omega}_t - \bar{\omega}_{t+1}\|) \leq C \left( \frac{\alpha_t}{1 - \gamma} + \xi_t \right).
$$

(50)

Combining Eqs. (46) to (50) leads to

$$
\|\nu_{t+1}\|^2 \leq (1 - 2\varepsilon\beta_t)\|\nu_t\|^2 + 2\beta_t \|\Lambda(\theta_t, \bar{\omega}_t, \omega_t, \tilde{x}_t)\| + C \left( \frac{\alpha_t}{1 - \gamma} + \xi_t \right) \|\nu_t\| + C \left( \frac{\alpha_t^2}{(1 - \gamma)^2} + \xi_t^2 + \beta_t^2 \right).
$$

(51)

**Control of the Markov noise term $\Lambda(\theta_t, \bar{\omega}_t, \omega_t, \tilde{x}_t)$**. We decompose the noise term using a similar technique to Zou et al., 2019 which was then used in Wu et al., 2020 Shen et al., 2020. Let $T > 0$. Define the mixing time

$$
\tau_T := \min\{t \in \mathbb{N}, t \geq 1 : c \tau_T^{t-1} \leq \min\{\alpha_T, \xi_T, \beta_T\} \}.
$$

(52)

In the remainder of the proof, we will use the notation $\tau$ for $\tau_T$ (interchangeably). In order to control the difference between the update rule of the critic and its steady-state expectation, we introduce an auxiliary chain
Analysis of a Target-Based Actor-Critic Algorithm with Linear Function Approximation

which coincides with $\tilde{x}_t$ except for the $\tau$ last steps where the policy is fixed to $\pi_{\theta_{t-\tau}}$. The auxiliary chain will be denoted by $\tilde{x}_t := (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1})$ where $\tilde{S}_{t+1} \sim p(\cdot|\tilde{S}_t, \tilde{A}_t)$ and $(\tilde{S}_t, \tilde{A}_t)$ is generated as follows:

$$\tilde{S}_{t-\tau} \xrightarrow{\theta_{t-\tau}} \tilde{A}_{t-\tau} \xrightarrow{\tilde{p}} \tilde{S}_{t-\tau+1} \xrightarrow{\theta_{t-\tau+1}} \tilde{A}_{t-\tau+1} \xrightarrow{\tilde{p}} \tilde{S}_{t-\tau+2} \xrightarrow{\theta_{t-\tau+2}} \tilde{A}_{t-\tau+2} \xrightarrow{\tilde{p}} \cdots \xrightarrow{\tilde{p}} \tilde{S}_t \xrightarrow{\theta_{t-\tau}} \tilde{A}_t \xrightarrow{\tilde{p}} \tilde{S}_{t+1}.$$ 

Compared to this chain, the original chain has a drifting policy, i.e., at each time step, the actor parameter $\theta_t$ is updated and so is the policy $\pi_{\theta_t}$ and we recall that it is given by:

$$\tilde{S}_{t-\tau} \xrightarrow{\theta_{t-\tau}} \tilde{A}_{t-\tau} \xrightarrow{\tilde{p}} \tilde{S}_{t-\tau+1} \xrightarrow{\theta_{t-\tau+1}} \tilde{A}_{t-\tau+1} \xrightarrow{\tilde{p}} \tilde{S}_{t-\tau+2} \xrightarrow{\theta_{t-\tau+2}} \tilde{A}_{t-\tau+2} \xrightarrow{\tilde{p}} \cdots \xrightarrow{\tilde{p}} \tilde{S}_t \xrightarrow{\theta_{t-\tau}} \tilde{A}_t \xrightarrow{\tilde{p}} \tilde{S}_{t+1}.$$ 

Using the shorthand notation $z_t := (\tilde{\omega}_t, \omega_t)$, the Markov noise term can be decomposed as follows:

$$\Lambda(\theta_t, \tilde{\omega}_t, \omega_t, \tilde{x}_t) = (\Lambda(\theta_t, z_t, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)) + (\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)) + \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t). \quad (53)$$ 

We control each one of the terms successively.

(a) **Control of $\Lambda(\theta_t, z_t, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)$:** Using that $\omega_t$ and $\tilde{g}$ are Lipschitz in all their arguments, $g$ is Lipschitz in its two last arguments and $\omega_t, \omega_t, g$ and $\tilde{g}$ are all bounded, one can show after tedious decompositions that

$$|\Lambda(\theta_t, z_t, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)| \leq C (||\theta_t - \theta_{t-\tau}|| + ||\tilde{\omega}_t - \tilde{\omega}_{t-\tau}|| + ||\omega_t - \omega_{t-\tau}||). \quad (54)$$

Then, recalling that the sequence $(\alpha_t)$ is nonincreasing, remark that

$$||\theta_t - \theta_{t-\tau}|| \leq \sum_{t=1}^{t-1}||\theta_{j+1} - \theta_j|| \leq C \frac{t-1}{1-\gamma} \sum_{t=1}^{t-1} \alpha_j \leq C \frac{t-1}{1-\gamma} \alpha_{t-\tau}.$$ 

Similarly, we have $||\tilde{\omega}_t - \tilde{\omega}_{t-\tau}|| \leq C t \xi_{t-\tau}$, $||\omega_t - \omega_{t-\tau}|| \leq C t \beta_{t-\tau}$ and we can therefore deduce from Eq. (54) that

$$|\Lambda(\theta_t, z_t, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)| \leq C t^2 \left( \frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau} \right). \quad (55)$$

(b) **Control of $\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)$:** following similar arguments to [Wu et al., 2020, Shen et al., 2020], we upperbound the conditional expectation of this error term w.r.t. $\tilde{S}_{t+1}, \tilde{\omega}_{t-\tau}, \omega_{t-\tau}$ and $\theta_{t-\tau}$. Note that our definition of $\tilde{x}_t$ is slightly different from the ones used in the two aforementioned references of the third component of $\tilde{x}_t$ (and also $\tilde{x}_t$) which is generated according to the original kernel $p$ instead of the artificial kernel $\tilde{p}$. We have

$$\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)|\tilde{S}_{t+1}, \theta_{t-\tau}] = \mathbb{E}[\nu_{t-\tau} g(\tilde{x}_t, z_{t-\tau}) - g(\tilde{x}_t, z_{t-\tau})|\tilde{S}_{t+1}, \theta_{t-\tau}]$$

$$\leq C d_{TV} (\mathbb{P}(\tilde{x}_t \in |\tilde{S}_{t+1}, \theta_{t-\tau}), \mathbb{P}(\tilde{x}_t \in |\tilde{S}_{t+1}, \theta_{t-\tau}))$$

$$\leq \frac{C}{\mathbb{E}[|\theta_t - \theta_{t-\tau}||\tilde{S}_{t+1}, \theta_{t-\tau}]}, \quad (56)$$

where the first equality stems from the definition of $\Lambda$, the first inequality uses the definition of the total variation distance $d_{TV}$ between two probability measures and the last inequality is a consequence of [Wu et al., 2020, Lem. B.2, p.17] (see also [Shen et al., 2020, Lem. 2 p.12]).

Then, we have

$$\sum_{i=t-\tau}^{t} \mathbb{E}[||\theta_i - \theta_{i-\tau}|||\tilde{S}_{i+1}, \theta_{i-\tau}] \leq \sum_{i=t-\tau}^{t} \sum_{j=t-\tau}^{i-1} \mathbb{E}[||\theta_{j+1} - \theta_j|||\tilde{S}_{i+1}, \theta_{i-\tau}]$$

$$\leq \frac{C}{1-\gamma} \sum_{i=t-\tau}^{t} \sum_{j=t-\tau}^{i-1} \alpha_j \leq \frac{C}{1-\gamma} \alpha_{t-\tau} \sum_{i=0}^{t} i \leq \frac{C}{1-\gamma} \alpha_{t-\tau} (\tau + 1)^2.$$
As a consequence of these derivations, Eq. (56) yields
\[ \mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)] - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \leq \frac{C}{1-\gamma}|A|\alpha_{t-\tau}(\tau + 1)^2, \tag{57} \]

(c) Control of $\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)$: Define $\tilde{x}_t := (\tilde{S}_t, \tilde{A}_t, S_{t+1})$ where $\tilde{S}_t \sim d_{\rho, \theta_{t-\tau}}, \tilde{A}_t \sim \pi_{\theta_{t-\tau}}$ and $S_{t+1} \sim \rho(\cdot|\tilde{S}_t, \tilde{A}_t)$. Observing that $\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)|\tilde{S}_{t+1}, \theta_{t-\tau}] = 0$, we obtain
\[ \mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)|\tilde{S}_{t+1}, \theta_{t-\tau}] = \mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t)|\tilde{S}_{t+1}, \theta_{t-\tau}] \]
\[ = \mathbb{E}[(\nu_{t-\tau}, g(\tilde{x}_t, z_{t-\tau}) - g(\tilde{x}_t, z_{t-\tau})|\tilde{S}_{t+1}, \theta_{t-\tau}] \]
\[ \leq C d_{TV}(\mathbb{P}(\tilde{x}_t \in |\tilde{S}_{t+1}, \theta_{t-\tau}), \mathbb{P}(\tilde{x}_t \in |\tilde{S}_{t+1}, \theta_{t-\tau})) \]
\[ \leq C d_{TV}(\mathbb{P}(\tilde{S}_t \in |\tilde{S}_{t+1}, \theta_{t-\tau}), d_{\rho, \theta_{t-\tau}}) \leq C \sigma^{\tau-1} \leq C \alpha_T, \tag{58} \]

where the first inequality stems again from the definition of the total variation norm and the last two ones follow from Assumption 6.1 and the definition of the mixing time $\tau = \tau_T$ (see Eq. (52)).

Given the decomposition of Eq. (53), collecting Eqs. (55), (57), (58) and taking total expectation leads to the conclusion of this subsection
\[ \mathbb{E}[\Lambda(\theta_{t}, z_{t}, \tilde{x}_t)] \leq C \left( \tau \left( \frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau} \right) + |A|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau + 1)^2 + \alpha_T \right). \tag{59} \]

Derivation of the convergence rate of the mean error term $\frac{1}{T} \sum_{t=1}^{T} \|\nu_t\|^2$. We obtain from taking the total expectation in Eq. (51) together with Eq. (59) that
\[ \mathbb{E}[\|\nu_{t+1}\|^2] \leq (1 - 2\varepsilon \beta_t)\mathbb{E}[\|\nu_t\|^2] + 2C\beta_t \left( \tau \left( \frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau} \right) + |A|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau + 1)^2 + \alpha_T \right) \]
\[ + C \left( \frac{\alpha_t}{1-\gamma} + \xi_t \right) \mathbb{E}[\|\nu_t\|] + C \left( \frac{\alpha_t^2}{(1-\gamma)^2} + \xi_t^2 + \beta_t^2 \right). \tag{60} \]

Rearranging the inequality and summing for $t$ between $\tau_T$ and $T$, we get
\[ 2\varepsilon \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nu_t\|^2] \leq I_1(T) + I_2(T) + I_3(T) + I_4(T), \tag{61} \]

where
\[ I_1(T) := \sum_{t=\tau_T}^{T} \frac{1}{\beta_t} \left( \mathbb{E}[\|\nu_t\|^2] - \mathbb{E}[\|\nu_{t+1}\|^2] \right), \tag{62} \]
\[ I_2(T) := \sum_{t=\tau_T}^{T} 2C \left( \tau \left( \frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau} \right) + |A|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau + 1)^2 + \alpha_T \right) \]
\[ I_3(T) := C \sum_{t=\tau_T}^{T} \left( \frac{\alpha_t}{1-\gamma} + \frac{\xi_t}{\beta_t} \right) \mathbb{E}[\|\nu_t\|] \tag{64} \]
\[ I_4(T) := C \sum_{t=\tau_T}^{T} \frac{\alpha_t^2}{(1-\gamma)^2} + \frac{\xi_t^2}{\beta_t} + \beta_t. \tag{65} \]

We derive estimates of each one of the terms $I_i(T)$ for $i = 1, 2, 3, 4$. 

Anas Barakat, Pascal Bianchi, Julien Lehmann
(1) Since \( \nu_t \) is a bounded sequence,

\[
I_1(T) = \sum_{t=\tau_T}^{T} \left( \frac{1}{\beta_t} - \frac{1}{\beta_{t-1}} \right) \mathbb{E}[\|\nu_t\|^2] + \frac{1}{\beta_{\tau_T}} \mathbb{E}[\|\nu_{\tau_T}\|^2] - \frac{1}{\beta_{\tau_T}} \mathbb{E}[\|\nu_{T+1}\|^2]
\]

\[
\leq C \left[ \sum_{t=\tau_T}^{T} \left( \frac{1}{\beta_t} - \frac{1}{\beta_{t-1}} \right) + \frac{1}{\beta_{\tau_T}} \right] = \frac{C}{\beta_T} = \mathcal{O}(T^\gamma). \quad (66)
\]

Then, since \( \tau_T = \mathcal{O}(\ln T) \), it follows that

\[
\frac{1}{1 + T - \tau_T} I_1(T) \leq \frac{1}{1 + T - \tau_T} \frac{C}{\beta_T} = \frac{1}{T(\frac{1}{\beta_T} + 1 - \frac{\tau_T}{T})} \frac{C}{\beta_T} = \mathcal{O}(T^{\beta-1}).
\]

(2) Using the inequality \( \sum_{k=1}^{p} k^{-\beta} \leq \frac{p^{1-\beta}}{1-\beta} \) for \( 1 \leq l < p \) and the fact that \( \tau_T = \mathcal{O}(\ln T) \), we have

\[
I_2(T) \leq C \left( \tau_T \sum_{t=0}^{T-\tau} \left( \frac{\alpha_t}{1 - \gamma} + \beta_t + \xi_t \right) + |A| \alpha (T + 1)^{\frac{2}{1 - \gamma}} \sum_{t=0}^{T-\tau} \alpha_t + (T - \tau) \alpha_T \right)
\]

\[
\leq \frac{C}{1 - \gamma} \left( (T + 1)^{1-\beta} + (T + 1)^{2} |A| (1 + T)^{1-\alpha} \right)
\]

\[
= \mathcal{O} \left( \frac{\ln T}{1 - \gamma} T^{1-\beta} \right) + \mathcal{O} \left( \frac{|A|}{1 - \gamma} \ln^2(T) T^{1-\alpha} \right) = \mathcal{O} \left( \frac{|A|}{1 - \gamma} \ln(T) T^{1-\beta} \right), \quad (67)
\]

where we recall for the second inequality that \( 0 < \beta < \xi < \alpha < 1 \) and for the last equality, we recall that \( |A| \) is finite. As a consequence,

\[
\frac{1}{1 + T - \tau_T} I_2(T) = \mathcal{O} \left( \frac{|A|}{1 - \gamma} \ln(T) T^{-\beta} \right).
\]

(3) Using the Cauchy-Schwarz inequality, we can write:

\[
I_3(T) = \sum_{t=\tau_T}^{T} C \left( \frac{\alpha_t}{1 - \gamma} + \frac{\xi_t}{\beta_t} \right) \mathbb{E}[\|\nu_t\|]
\]

\[
\leq C \sqrt{\sum_{t=\tau_T}^{T} \left( \frac{\alpha_t}{1 - \gamma} + \frac{\xi_t}{\beta_t} \right)^2 \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nu_t\|^2]}. \quad (68)
\]

Then, observing that the sequences \( \left( \frac{\alpha_t}{\beta_t} \right) \) and \( \left( \frac{\xi_t}{\beta_t} \right) \) are nonincreasing, we have:

\[
\frac{1}{1 + T - \tau_T} \sum_{t=\tau_T}^{T} \left( \frac{\alpha_t}{1 - \gamma} + \frac{\xi_t}{\beta_t} \right)^2 \leq \frac{2}{1 + T - \tau_T} \sum_{t=\tau_T}^{T} \left( \frac{\alpha_t}{1 - \gamma} \right)^2 + \left( \frac{\xi_t}{\beta_t} \right)^2
\]

\[
= \frac{2}{1 + T - \tau_T} \sum_{t=0}^{T-\tau_T} \left( \frac{\alpha_t+\tau_T}{1 - \gamma} \right)^2 \left( \frac{\xi_t+\tau_T}{\beta_t+\tau_T} \right)^2
\]

\[
\leq \frac{2}{T - \tau_T + 1} \sum_{t=0}^{T-\tau_T} \left( \frac{\alpha_t}{1 - \gamma} \right)^2 + \left( \frac{\xi_t}{\beta_t} \right)^2
\]

\[
\leq \frac{(T - \tau_T + 1)^{-2(\alpha-\beta)}}{(1 - \gamma)^2(1 - 2(\alpha - \beta))} + \frac{(T - \tau_T + 1)^{-2(\xi-\beta)}}{1 - 2(\xi - \beta)}
\]

\[
= \mathcal{O} \left( \frac{T^{-2(\alpha-\beta)}}{(1 - \gamma)^2} + T^{-2(\xi-\beta)} \right) \quad (69)
\]
(4) Similarly to item (3), to control the fourth term, we write:

\[
\frac{1}{1 + T - \tau_T} \sum_{t=\tau_T}^T \left( \frac{\alpha_t^2}{(1 - \gamma)^2 \beta_t} + \frac{\xi_t^2}{\beta_t} + \beta_t \right) \leq \frac{1}{1 + T - \tau_T} \sum_{t=0}^{T-\tau_T} \left( \frac{\alpha_t^2}{(1 - \gamma)^2 \beta_t} + \frac{\xi_t^2}{\beta_t} + \beta_t \right)
\]

\[
\leq \frac{(1 + T - \tau_T)^{(2\alpha - \beta)}}{(1 - \gamma)^2(1 - (2\alpha - \beta))} + \frac{(1 + T - \tau_T)^{-2(\xi - \beta)}}{1 - (2\xi - \beta)} + \frac{(1 + T - \tau_T)^{-\beta}}{1 - \beta}
\]

\[
= \mathcal{O}\left(\frac{T^{-2(\alpha - \beta)}}{(1 - \gamma)^2} + T^{-2(\xi - \beta)} + T^{-\beta}\right).
\]

Hence,

\[
\frac{1}{1 + T - \tau_T} I_4(T) = \mathcal{O}\left(\frac{T^{-2(\alpha - \beta)}}{(1 - \gamma)^2} + T^{-2(\xi - \beta)} + T^{-\beta}\right).
\]

Define:

\[
N(T) := \frac{1}{1 + T - \tau_T} \sum_{t=\tau_T}^T \mathbb{E}[\|\nu_t\|^2],
\]

(72)

\[
F(T) := \frac{1}{1 + T - \tau_T} \sum_{t=\tau_T}^T \left( \frac{\alpha_t}{(1 - \gamma)\beta_t} \right)^2 + \left( \frac{\xi_t}{\beta_t} \right)^2,
\]

(73)

\[
G(T) := \frac{1}{1 + T - \tau_T} (I_1(T) + I_2(T) + I_4(T)).
\]

(74)

Using items (1) to (4), we have:

\[
F(T) = \mathcal{O}\left(\frac{T^{-2(\alpha - \beta)}}{(1 - \gamma)^2} + T^{-2(\xi - \beta)}\right),
\]

(75)

\[
G(T) = \mathcal{O}(T^{\beta - 1}) + \mathcal{O}\left(\frac{|A|}{1 - \gamma} \ln(T)T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha - \beta)}}{(1 - \gamma)^2} + T^{-2(\xi - \beta)} + T^{-\beta}\right),
\]

(76)

From Eq. (61) and items (1) to (4) above, we have:

\[
2\varepsilon N(T) \leq C \sqrt{F(T)} \sqrt{N(T)} + G(T).
\]

Solving this inequality yields:

\[
N(T) = \mathcal{O}(F(T) + G(T)).
\]

Remark that \(0 < 2(\alpha - \beta) < 2\alpha - \beta\) and \(0 < 2(\xi - \beta) < 2\xi - \beta\), we obtain:

\[
N(T) = \mathcal{O}(T^{\beta - 1}) + \mathcal{O}\left(\frac{|A|}{1 - \gamma} \ln(T)T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha - \beta)}}{(1 - \gamma)^2} + T^{-2(\xi - \beta)} + T^{-\beta}\right).
\]

Then, we conclude that:

\[
\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nu_t\|^2] = \mathcal{O}(\ln(T)T^{-1}) + \mathcal{O}(N(T)) = \mathcal{O}(N(T)).
\]

B.1.2 Control of the second error term \(\tilde{\omega}_t - \bar{\omega}_s(\theta_t)\)

Consider the shorthand notation \(\tilde{\nu}_t := \tilde{\omega}_t - \bar{\omega}_s(\theta_t)\).
Using the update rules of $\bar{\omega}_t$, $\omega_t$ and developing the squared norm gives:

$$\|\bar{\nu}_{t+1}\|^2 = \|\bar{\nu}_t + \xi_t(\omega_{t+1} - \bar{\omega}_t) - \bar{\omega}_*(\theta_{t+1})\|^2$$

$$= \|\bar{\nu}_t + \xi_t(\omega_t + \beta g(\bar{x}_t, \bar{\omega}_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\|^2$$

$$= \|\bar{\nu}_t\|^2 + 2\langle \bar{\nu}_t, \xi_t(\omega_t + \beta g(\bar{x}_t, \bar{\omega}_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\rangle + \|\bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\|^2.$$

(77)

Since the sequences $(\nu_t)$, $(\bar{\omega}_t)$ and the functions $g, \omega_t$ are bounded and the function $\bar{\omega}_*$ is Lipschitz continuous, the last squared norm term can be bounded by:

$$C\xi_t^2 + \xi_t^2 + \frac{\alpha^2_t}{(1 - \gamma)^2}.$$

We now control the scalar product in Eq. (77). We decompose this term into four different terms:

(a) Using Assumption 6.2, it holds that:

$$2\xi_t \langle \bar{\nu}_t, \bar{\omega}_*(\theta_t) - \bar{\omega}_t \rangle = -2\xi_t \langle \bar{\nu}_t, G(\theta_t)^{-1}G(\theta_t)\bar{\nu}_t \rangle \leq -2\xi_t \|\bar{\nu}_t\|^2.$$

(b) The boundedness of the function $g$ implies that:

$$2\xi_t \beta_t \langle \bar{\nu}_t, g(\bar{x}_t, \bar{\omega}_t, \omega_t) \rangle \leq C\xi_t \beta_t \|\bar{\nu}_t\|.$$

(c) Applying the Cauchy-Schwarz inequality gives:

$$2\xi_t \|\bar{\nu}_t\| \leq 2\xi_t \|\bar{\nu}_t\| \cdot \|\nu_t\|.$$

(d) Since $\bar{\omega}_*$ is Lipschitz continuous, we can write:

$$2\langle \bar{\nu}_t, \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\rangle \leq C\frac{\alpha_t}{1 - \gamma} \|\bar{\nu}_t\|.$$

Collecting the bounds from items (a) to (d) and incorporating them into Eq. (77), we obtain:

$$\|\bar{\nu}_{t+1}\|^2 \leq (1 - 2\xi_t)\|\bar{\nu}_t\|^2 + C \left(\xi_t \beta_t + \frac{\alpha_t}{1 - \gamma}\right) \|\bar{\nu}_t\| + 2\xi_t \|\bar{\nu}_t\| \cdot \|\nu_t\| + C \left(\xi_t^2 \beta_t^2 + \xi_t^2 + \frac{\alpha^2_t}{(1 - \gamma)^2}\right).$$

(78)

Rearranging Ineq. (78) leads to:

$$2\xi_t \|\bar{\nu}_t\|^2 \leq \frac{1}{\xi_t} (\|\bar{\nu}_t\|^2 - \|\bar{\nu}_{t+1}\|^2) + C \left(\beta_t + \frac{\alpha_t}{(1 - \gamma)\xi_t}\right) \|\bar{\nu}_t\| + 2\|\bar{\nu}_t\| \cdot \|\nu_t\| + C \left(\xi_t \beta_t^2 + \xi_t + \frac{\alpha^2_t}{(1 - \gamma)^2}\right).$$

(79)

Summing this inequality for $t$ between 1 and $T$ and taking total expectation yield:

$$\frac{2\xi_t}{T} \sum_{t=1}^T \mathbb{E}[\|\bar{\nu}_t\|^2] \leq \Sigma_1(T) + \Sigma_2(T) + \Sigma_3(T) + \Sigma_4(T),$$

(80)

where

$$\Sigma_1(T) := \frac{1}{T} \sum_{t=1}^T \frac{1}{\xi_t} \left(\mathbb{E}[\|\bar{\nu}_t\|^2] - \mathbb{E}[\|\bar{\nu}_{t+1}\|^2]\right),$$

$$\Sigma_2(T) := C \frac{T}{T} \sum_{t=1}^T \left(\beta_t + \frac{\alpha_t}{(1 - \gamma)\xi_t}\right) \mathbb{E}[\|\bar{\nu}_t\|],$$

$$\Sigma_3(T) := \frac{2}{T} \sum_{t=1}^T \mathbb{E}[\|\bar{\nu}_t\| \cdot \|\nu_t\|],$$

$$\Sigma_4(T) := C \frac{T}{T} \sum_{t=1}^T \left(\xi_t \beta_t^2 + \xi_t + \frac{\alpha^2_t}{(1 - \gamma)^2}\right).$$

(81-84)

Similarly to Sec. B.1.1, we control each one of the terms $\Sigma_i, i = 1, 2, 3, 4$ successively.
(i) First, using the boundedness of \((\bar{\nu}_t)\), we estimate \(\Sigma_1\) as follows:

\[
\Sigma_1(T) = \frac{1}{T} \left[ \sum_{t=1}^{T} \left( \frac{1}{\xi_t} - \frac{1}{\xi_{t-1}} \right) \mathbb{E}[\|\bar{\nu}_t\|^2] + \mathbb{E}[\|\bar{\nu}_1\|^2] - \frac{1}{\xi_T} \mathbb{E}[\|\bar{\nu}_{T+1}\|^2] \right] \leq \frac{C}{T\xi_T} = \mathcal{O}(T^{\xi-1}).
\]

(ii) Cauchy-Schwarz inequality implies:

\[
\Sigma_2(T) \leq \frac{C}{T} \sqrt{\sum_{t=1}^{T} \left( \beta_t + \frac{\alpha_t}{(1-\gamma)\xi_t} \right)^2} \mathbb{E}[\|\bar{\nu}_t\|^2] \leq C \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left( \beta_t^2 + \left( \frac{\alpha_t}{(1-\gamma)\xi_t} \right)^2 \right)} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\bar{\nu}_t\|^2]}.
\]

Moreover,

\[
\frac{1}{T} \sum_{t=1}^{T} \left( \beta_t^2 + \left( \frac{\alpha_t}{(1-\gamma)\xi_t} \right)^2 \right) \leq \frac{1}{T} \left( (T+1)^{1-2\beta} \frac{T+1}{1-2\beta} + \frac{(T+1)^{-2(\alpha-\xi)}}{(1-\gamma)^2(1-2(\alpha-\xi))} \right) = \mathcal{O}(T^{-2\beta}) + \mathcal{O} \left( \frac{T^{\alpha-\xi}}{(1-\gamma)^2} \right).
\]

(iii) Invoking the Cauchy-Schwarz inequality again yields:

\[
\Sigma_3(T) \leq 2 \sqrt{\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\bar{\nu}_t\|^2]} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nu_t\|^2]}.
\]

(iv) Similarly to item (ii), we obtain

\[
\Sigma_4(T) = \mathcal{O}(T^{-\xi-2\beta}) + \mathcal{O}(T^{-\xi}) + \mathcal{O} \left( \frac{T^{\xi-2\alpha}}{(1-\gamma)^2} \right).
\]

Define for every \(T > 0\) the following quantities:

\[
W(T) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nu_t\|^2], \quad (85)
\]

\[
X(T) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\bar{\nu}_t\|^2], \quad (86)
\]

\[
Y(T) := \frac{1}{T} \sum_{t=1}^{T} \left( \beta_t^2 + \left( \frac{\alpha_t}{(1-\gamma)\xi_t} \right)^2 \right), \quad (87)
\]

\[
Z(T) := \Sigma_1(T) + \Sigma_4(T). \quad (88)
\]

It follows from items (i) to (iv) and Sec. B.1.1 (for the last estimate) that

\[
Y(T) = \mathcal{O}(T^{-2\beta}) + \mathcal{O} \left( \frac{T^{\alpha-\xi}}{(1-\gamma)^2} \right), \quad (89)
\]

\[
Z(T) = \mathcal{O}(T^{\xi-1}) + \mathcal{O}(T^{-\xi-2\beta}) + \mathcal{O}(T^{-\xi}) + \mathcal{O} \left( \frac{T^{\xi-2\alpha}}{(1-\gamma)^2} \right), \quad (90)
\]

\[
W(T) = \mathcal{O}(T^{\beta-1}) + \mathcal{O} \left( \frac{|A|}{1-\gamma} \ln(T) T^{-\beta} \right) + \mathcal{O} \left( \frac{T^{2(\beta-\alpha)}}{(1-\gamma)^2} \right) + \mathcal{O}(T^{2(\beta-\xi)}). \quad (91)
\]
Eq. \((80)\) can be written:

\[
2\zeta X(T) \leq C \left( \sqrt{Y(T)} + \sqrt{W(T)} \right) \sqrt{X(T)} + Z(T).
\]

Solving this inequality implies:

\[
X(T) = O(Y(T) + W(T) + Z(T)).
\]  \(\text{(92)}\)

Since \(0 < \beta < \xi < \alpha < 1\), we obtain:

\[
X(T) = O(T^{\xi-1}) + O \left( \frac{|A|}{1-\gamma} \ln(T) T^{-\beta} \right) + O \left( \frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2} \right) + O(T^{-2(\xi-\beta)}).
\]  \(\text{(93)}\)

### B.1.3 End of Proof of Th. 6.1

We conclude our finite-time analysis of the critic by combining both previous sections  \(\text{(B.1.1) and (B.1.2)}\):

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\omega_t - \bar{\omega}_t(\theta_t)\|^2] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nu_t + \omega_t(\theta_t, \bar{\omega}_t) - \bar{\omega}_t(\theta_t)\|^2]
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nu_t + \omega_t(\theta_t, \bar{\omega}_t) - \bar{\omega}_t(\theta_t, \bar{\omega}_t)\|^2]
\]

\[
\leq 2W(T) + CX(T)
\]

\[
= O(T^{\xi-1}) + O \left( \frac{|A|}{1-\gamma} \ln(T) T^{-\beta} \right) + O \left( \frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2} \right) + O(T^{-2(\xi-\beta)}),
\]  \(\text{(94)}\)

where the second equality follows from using the identity \(w_t(\theta_t, \bar{\omega}_t) = \bar{\omega}_t(\theta_t)\) for every \(\theta_t \in \mathbb{R}^d\), the inequality stems from using the classical inequality \(\|a + b\|^2 \leq 2\|a\|^2 + \|b\|^2\) together with the fact that \(\omega_t\) is Lipschitz continuous, the penultimate equality is a consequence of Eq. \(\text{(92)}\) and the last equality is the result of the previous section (see Eq. \(\text{(93)}\)).

### B.2 Proof of Th. 6.2 finite-time analysis of the actor

Recall the notation \(\tilde{x}_t := (\tilde{S}_t, \tilde{A}_t, S_{t+1})\). In this section, we overload this notation with the reward sequence \((R_t)\), i.e., \(\tilde{x}_t := (\tilde{S}_t, \tilde{A}_t, S_{t+1}, R_{t+1})\). Let us fix some additional convenient notations. Define for every \(\tilde{x} = (\tilde{s}, \tilde{a}, s, r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [-U_R, U_R]\), and every \(\omega \in \mathbb{R}^m, \theta \in \mathbb{R}^d\):

\[
\hat{\delta}(\tilde{x}, \omega) := r + \gamma \phi(s)^T \omega - \phi(\tilde{s})^T \omega
\]

\[
\hat{\delta}(\tilde{x}, \theta) := r + \gamma V_\omega(\tilde{s}) - V_\omega(s).
\]  \(\text{(95)}\)

\(\text{(96)}\)

Note that the TD error \(\delta_{t+1}\) used in Algorithm 1 coincides with \(\hat{\delta}(\tilde{x}_t, \omega_t)\).

Recall that \(\theta \mapsto \nabla J(\theta)\) and \(\theta \mapsto V_\omega(s)\) (for every \(s \in \mathcal{S}\)) are Lipschitz continuous. Throughout the proof, \(L_{\nabla J}\) (resp. \(L_V\)) stands for the Lipschitz constant of \(\theta \mapsto \nabla J(\theta)\) (resp. \(\theta \mapsto V_\omega(s)\) for every \(s \in \mathcal{S}\)) and \(C_{\nabla J}\) (resp. \(C_V\)) denotes the upperbound of \(\theta \mapsto \|\nabla J(\theta)\|\) (resp. \(\theta \mapsto V_\omega(s)\) for every \(s \in \mathcal{S}\)). Since the function \(\nabla J\) is \(L_{\nabla J}\)-Lipschitz continuous, a classical Taylor inequality combined with the update rule of \(\theta_t\) yields:

\[
J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{1-\gamma} (\nabla J(\theta_t), \hat{\delta}(\tilde{x}_t, \omega_t))_{\psi_t}(\tilde{S}_t, \tilde{A}_t) - \frac{L_{\nabla J}}{2} \frac{\alpha_t^2}{(1-\gamma)^2} \|\hat{\delta}(\tilde{x}_t, \omega_t)_{\psi_t}(\tilde{S}_t, \tilde{A}_t)\|^2.
\]  \(\text{(97)}\)

Recalling that \(\theta \mapsto \psi_\theta(s, a)\) is bounded by Assumption 3.1\((c)\) \((R_t)\) and \((\omega_t)\) are bounded (see Assumption 5.3) and \(\mathcal{S}, \mathcal{A}\) are finite, we obtain from Eq. \(\text{(97)}\) that there exists a constant \(C\) s.t.:

\[
J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{1-\gamma} (\nabla J(\theta_t), \hat{\delta}(\tilde{x}_t, \omega_t))_{\psi_t}(\tilde{S}_t, \tilde{A}_t) - CL_{\nabla J} \frac{\alpha_t^2}{(1-\gamma)^2}.
\]  \(\text{(98)}\)
Now, we decompose the TD error by introducing both the moving target $\bar{\omega}_t(\theta_t)$ and the TD error $\delta(\tilde{x}_t, \theta_t)$ associated to the true value function $V_{\pi_{\theta}}$:

$$\hat{\delta}(\tilde{x}_t, \omega_t) = [\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_t(\theta_t))] + [\hat{\delta}(\tilde{x}_t, \bar{\omega}_t(\theta_t)) - \delta(\tilde{x}_t, \theta_t)] + \delta(\tilde{x}_t, \theta_t).$$  \hspace{1cm} (99)

Incorporating this decomposition into Eq. \hspace{1cm} (99) gives:

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{1 - \gamma} \langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_t(\theta_t))) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \rangle$$

$$+ \frac{\alpha_t}{1 - \gamma} \langle \nabla J(\theta_t), \delta(\tilde{x}_t, \theta_t) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) - \nabla J(\theta_t) \rangle + \frac{\alpha_t}{1 - \gamma} \| \nabla J(\theta_t) \|^2 - C L \nabla J \frac{\alpha_t^2}{(1 - \gamma)^2}.$$ \hspace{1cm} (100)

In Eq. (100), the first inner product corresponds to the bias introduced by the critic. The second one represents the linear FA error and the third translates the Markovian noise. Our task now is to control each one of these error terms in Eq. (100).

For the first term, observing that $\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_t(\theta_t)) = (\gamma \phi(S_{t+1}) - \phi(\tilde{S}_t)(\omega_t - \bar{\omega}_t(\theta_t)))$, the Cauchy-Schwarz inequality leads to:

$$\mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_t(\theta_t))) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \rangle] \geq -C \sqrt{\mathbb{E}[\| \nabla J(\theta_t) \|^2]} \sqrt{\mathbb{E}[\| \omega_t - \bar{\omega}_t(\theta_t) \|^2]}.$$ \hspace{1cm} (101)

Then, we control each one of the second and third terms in Eq. (100) in the following sections successively.

**B.2.1 Control of the Markovian bias term**

We introduce a specific convenient notation for the second term, for every $\tilde{x} = (\tilde{s}, \tilde{a}, s, r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [-U_R, U_R]$, and every $\theta \in \mathbb{R}^d$:

$$\Gamma(\tilde{x}, \theta) := \langle \nabla J(\theta), \delta(\tilde{x}, \theta) \psi_{\theta}(\tilde{s}, \tilde{a}) - \nabla J(\theta) \rangle.$$ \hspace{1cm}

Recall from Sec. B.1.1 the auxiliary Markov chain ($\tilde{x}_t$), the Markov chain ($\tilde{x}_t$) induced by the stationary distribution and the mixing time $\tau$ defined in Eq. (52).

Similarly to Sec. B.1.1 we introduce the following decomposition:

$$\mathbb{E}[\Gamma(\tilde{x}_t, \theta_t)] = \mathbb{E}[\Gamma(\tilde{x}_t, \theta_t) - \Gamma(\tilde{x}_t, \theta_{t-\tau})] + \mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau}) - \Gamma(\tilde{x}_t, \theta_{t-\tau})]$$

$$+ \mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau}) - \Gamma(\tilde{x}_t, \theta_{t-\tau})] + \mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau})].$$ \hspace{1cm} (102)

We address each term of this decomposition successively.

(a) For this first term, we write:

$$\Gamma(\tilde{x}_t, \theta_t) - \Gamma(\tilde{x}_t, \theta_{t-\tau}) = \langle \nabla J(\theta_t) - \nabla J(\theta_{t-\tau}), \delta(\tilde{x}_t, \theta_t) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) - \nabla J(\theta_t) \rangle$$

$$+ \langle \nabla J(\theta_{t-\tau}), (\delta(\tilde{x}_t, \theta_t) - \delta(\tilde{x}_t, \theta_{t-\tau})) \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \rangle$$

$$+ \langle \nabla J(\theta_{t-\tau}), (\delta(\tilde{x}_t, \theta_{t-\tau}) - \psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)) \rangle$$

Moreover, note that:

$$\delta(\tilde{x}_t, \theta_t) - \delta(\tilde{x}_t, \theta_{t-\tau}) = \gamma (V_{\pi_{\theta}}(S_{t+1}) - V_{\pi_{\theta_{t-\tau}}}(S_{t+1})) + V_{\pi_{\theta_{t-\tau}}}(\tilde{S}_t) - V_{\pi_{\theta}}(\tilde{S}_t).$$

Remark that $\nabla J, \theta \mapsto \psi_{\theta}$ and $\theta \mapsto V_{\pi_{\theta}}$ are bounded functions under Assumption 3.1. Since $\nabla J, V_{\pi_{\theta}}, \psi_{\theta}$ are in addition Lipschitz continuous as functions of $\theta$ (see, for e.g., [Shen et al., 2020], Lem. 3) for a proof for $V_{\pi_{\theta}}$ under Assumption 3.1, one can show after tedious inequalities that:

$$|\Gamma(\tilde{x}_t, \theta_t) - \Gamma(\tilde{x}_t, \theta_{t-\tau})| \leq (L_{\nabla J}(C(1 + C_V) + C_{\nabla J}) + CC_{\nabla J}L_V + C(1 + C_V)C_{\nabla J} + C_{\nabla J}L_{\nabla J}) \| \theta_t - \theta_{t-\tau} \|$$

$$\leq CC_{1 - \gamma} \| \theta_t - \theta_{t-\tau} \|,$$ \hspace{1cm} (103)
where $C_{1-\gamma} := \max(L_{\mathcal{C}V}, L_{\mathcal{C}V}C_{\mathcal{C}V}, L_{\mathcal{C}V}C_{\mathcal{C}V}, C_{\mathcal{C}V}C_{\mathcal{C}V})$. Note here that the last notation highlights that the constant depends on $1 - \gamma$ due to the dependence on $1 - \gamma$ of the constants defining $C_{1-\gamma}$. We will explicit this dependence later on in the proof.

(b) For the second term, we have:
\[
|E[\Gamma(\bar{x}_t, \theta_{t-\tau}) - \Gamma(\bar{x}_t, \theta_{t-\tau})]| = |E[(\nabla J(\theta_{t-\tau}), \delta(\bar{x}_t, \theta_{t-\tau}))\psi_{\theta_{t-\tau}}(\bar{S}_t, \bar{A}_t) - \delta(\bar{x}_t, \theta_{t-\tau})\psi_{\theta_{t-\tau}}(\bar{S}_t, \bar{A}_t)]| \\
= |E[(\nabla J(\theta_{t-\tau}), \delta(\bar{x}_t, \theta_{t-\tau}))\psi_{\theta_{t-\tau}}(\bar{S}_t, \bar{A}_t) - \delta(\bar{x}_t, \theta_{t-\tau})\psi_{\theta_{t-\tau}}(\bar{S}_t, \bar{A}_t)]|_{\bar{S}_{t-\tau+1}, \theta_{t-\tau}}| \\
\leq C_{TV}C_{\mathcal{C}V}\mathbb{E}[d_{TV}(\mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}))] \\
\leq C_{TV}C_{\mathcal{C}V}|A| \sum_{i=t-\tau}^{t} \mathbb{E}[|\theta_i - \theta_{t-\tau}|].
\] (104)

Here, the first inequality is a consequence of the definition of the total variation distance whereas the second inequality follows from applying [Wu et al., 2020, Lem. B.2]. Indeed, using this last lemma, to show the last inequality, it is sufficient to write:
\[
d_{TV}(\mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau})) \\
= d_{TV}(\mathbb{P}((\bar{S}_t, \bar{A}_t) \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}((\bar{S}_t, \bar{A}_t) \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau})) \\
\leq d_{TV}(\mathbb{P}(\bar{S}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{S}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau})) + \frac{1}{2}|A|\mathbb{E}[|\theta_t - \theta_{t-\tau}|].
\]

Iterating this inequality gives the desired result of Eq. (104). We conclude from this item that:
\[
E[\Gamma(\bar{x}_t, \theta_{t-\tau}) - \Gamma(\bar{x}_t, \theta_{t-\tau})] \geq -C_{TV}C_{\mathcal{C}V}|A| \sum_{i=t-\tau}^{t} \mathbb{E}[|\theta_i - \theta_{t-\tau}|].
\]

(c) Regarding the third term, similarly to item (b), we can write:
\[
E[\Gamma(\bar{x}_t, \theta_{t-\tau}) - \Gamma(\bar{x}_t, \theta_{t-\tau})] \geq -C_{TV}C_{\mathcal{C}V}\mathbb{E}[d_{TV}(\mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}))] \\
= -C_{TV}C_{\mathcal{C}V}\mathbb{E}[d_{TV}(\mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), d_p\theta_{t-\tau} \otimes \pi_{\theta_{t-\tau}} \otimes p)] \\
= -C_{TV}C_{\mathcal{C}V}\mathbb{E}[d_{TV}(\mathbb{P}(\bar{x}_t \in |\bar{S}_{t-\tau+1}, \theta_{t-\tau}), d_p\theta_{t-\tau} \otimes \pi_{\theta_{t-\tau}} \otimes p)] \\
\geq -C_{TV}C_{\mathcal{C}V}|A|\sigma^{-1},
\] (105)

where the equalities follow from the definitions of $\bar{x}_t$, $\bar{x}_t$ and the last inequality stems from Assumption 6.1.

(d) Since the Markov chain $\bar{x}_t$ is built s.t. $\bar{S}_t \sim d_p\theta_{t-\tau}, \bar{A}_t \sim \pi_{\theta_{t-\tau}}, \bar{S}_{t+1} \sim p(\cdot |\bar{S}_t, \bar{A}_t)$, one can see that
\[
E[\Gamma(\bar{x}_t, \theta_{t-\tau})] = 0.
\]

We conclude this section from Eq. (102) by collecting Eqs. (103) to (105) (items (a) to (d)) to obtain:
\[
E[\Gamma(\bar{x}_t, \theta_t)] \geq -C_{1-\gamma}E[|\theta_t - \theta_{t-\tau}|] - C_{TV}C_{\mathcal{C}V}J \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{t-\tau}|] - C_{TV}C_{\mathcal{C}V}|A|\sigma^{-1} \\
\geq -C_{1-\gamma} \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{i-1}|] - C_{TV}C_{\mathcal{C}V}J \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{i-1}|] - C_{TV}C_{\mathcal{C}V}|A|\sigma^{-1} \\
\geq -C_{1-\gamma} \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{i-1}|] - C_{TV}C_{\mathcal{C}V}J \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{i-1}|] - C_{TV}C_{\mathcal{C}V}|A|\sigma^{-1} \\
\geq -C(C_{1-\gamma} + C_{TV}C_{\mathcal{C}V}|A|\sigma^2) \sum_{i=t-\tau+1}^{t} \mathbb{E}[|\theta_i - \theta_{i-1}|] - C_{TV}C_{\mathcal{C}V}|A|\sigma^{-1} \\
\geq -C \left( (C_{1-\gamma} + C_{TV}C_{\mathcal{C}V}|A|\sigma^2) \frac{\alpha_{t-\tau}}{1-\gamma} + C_{TV}C_{\mathcal{C}V}|A|\alpha_T \right),
\] (106)

where the last inequality uses the definition of the mixing time $\tau$ and the fact that the sequence $\alpha_t$ is nonincreasing.
B.2.2 Control of the linear FA error term

Recall that $\theta \mapsto \psi_\theta$ is Lipschitz continuous, $\nabla J$ is bounded and remark that the quantity $\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t)$ is bounded. Therefore, using the Cauchy-Schwarz inequality, we have:

$$\mathbb{E}[(\nabla J(\theta_t), (\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t))]$$

$$= \mathbb{E}[(\nabla J(\theta_t), (\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t))(\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) - \psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)))]$$

$$+ \mathbb{E}[(\nabla J(\theta_t), (\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t))]$$

$$\geq -C(1 + C_V)C_{\nabla J}\mathbb{E}[||\theta_t - \theta_{t-\tau}||] + \mathbb{E}[(\nabla J(\theta_t), (\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t))]. \quad (107)$$

Let us introduce for every $\tilde{x} = (\tilde{s}, \tilde{a}, s, r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [-U_R, U_R]$, and every $\theta \in \mathbb{R}^d$ the shorthand notation:

$$\Delta(\tilde{x}, \theta) := (\nabla J(\theta), (\delta(x_t, \omega_\theta(\theta_t)) - \delta(x_t, \theta_t))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)).$$

Note here that the term $\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)$ in the notation above is fixed in adequacy with Eq. [107]. The following decomposition holds:

$$\Delta(x_t, \theta_t) = (\Delta(x_t, \theta_t) - \Delta(x_t, \theta_{t-\tau})) + (\Delta(x_t, \theta_{t-\tau}) - \Delta(x_t, \theta_{t-\tau})) + (\Delta(x_t, \theta_{t-\tau}) - \Delta(x_t, \theta_{t-\tau}) + \Delta(x_t, \theta_{t-\tau}). \quad (108)$$

Similar derivations to the previous section allow us to control each one of the error terms.

(i) Using that the mappings $\nabla J, \theta \mapsto V_{\omega}(s)$ (for every $s \in \mathcal{S}$) and $\theta \mapsto \omega_\theta(\theta)$ are $L_{\nabla J}$ (resp. $L_V, L_{\omega_\theta}$)-Lipschitz continuous, we obtain:

$$\Delta(x_t, \theta_t) - \Delta(x_t, \theta_{t-\tau}) \geq -C\tilde{C}_{1-\gamma}||\theta_t - \theta_{t-\tau}||,$$

where $\tilde{C}_{1-\gamma} := L_{\nabla J}(1 + C_V) + C_{\nabla J}(L_V + L_{\omega}).$

Using similar manipulations to the previous section, we get:

(ii) $\mathbb{E}[\Delta(x_t, \theta_{t-\tau}) - \Delta(x_t, \theta_{t-\tau})] \geq -CC_{\nabla J}(1 + C_V)|A| \sum_{i=t-\tau}^t \mathbb{E}[||\theta_i - \theta_{t-\tau}||]. \quad (109)$

(iii) $\mathbb{E}[\Delta(x_t, \theta_{t-\tau}) - \Delta(x_t, \theta_{t-\tau})] \geq -CC_{\nabla J}(1 + C_V)\sigma^{\tau-1}. \quad (110)$

(iv) For the last term, we can write:

$$\mathbb{E}[\Delta(x_t, \theta_{t-\tau})|\theta_{t-\tau}] \geq -C||\nabla J(\theta_{t-\tau})|| \cdot \mathbb{E}[||\delta(x_t, \omega_\theta(\theta_{t-\tau})) - \delta(x_t, \theta_{t-\tau})|||\theta_{t-\tau}]. \quad (111)$$

Then, recall that $\tilde{x}_t = (\tilde{S}_t, \tilde{A}_t, S_{t+1})$ where $S_{t+1} \sim p(.|\tilde{S}_t, \tilde{A}_t)$ and observe that:

$$\delta(x_t, \omega_\theta(\theta_{t-\tau})) - \delta(x_t, \theta_{t-\tau}) = \gamma(V_{\sigma_{t-\tau}}(S_{t+1}))T\tilde{\omega}_{\theta_{t-\tau}} - V_{\sigma_{t-\tau}}(\tilde{S}_t + \phi(\tilde{S}_t)^T\tilde{\omega}_{\theta_{t-\tau}}). \quad (112)$$

Recalling that $\tilde{p} = \gamma p + (1 - \gamma)\rho$ and using Assumption 6.3, one can then easily show that:

$$\mathbb{E}[||\delta(x_t, \omega_\theta(\theta_{t-\tau})) - \delta(x_t, \theta_{t-\tau})|||\theta_{t-\tau}] \leq C_{\epsilon_{\text{FA}}}.$$

As a consequence, noticing that $\nabla J$ is bounded, we obtain from Eq. [111]:

$$\mathbb{E}[\Delta(x_t, \theta_{t-\tau})] \geq -CC_{\nabla J}\epsilon_{\text{FA}}.$$
Combining items (i) to (iv) with the boundedness of the function \( \nabla J \), we conclude from this section that:
\[
\mathbb{E}[(\nabla J(\theta_t), (\delta(\hat{x}_t, \hat{\omega}_t(\theta_t)) - \delta(\hat{x}_t, \tilde{\theta}_t))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t))]
\geq -C((1 + C_V)C_{\nabla J} + \tilde{C}_{1-\gamma})|\mathbb{E}[\theta_t - \theta_{t-\tau}]| - CC_{\nabla J}(1 + C_V)|A| \sum_{i=t-\tau}^{t} \mathbb{E}[|\theta_i - \theta_{t-\tau}|]
- CC_{\nabla J}(1 + C_V)\sigma^{-1} - CC_{\nabla J}\epsilon_{FA}
\geq -C \left( ((1 + C_V)C_{\nabla J} + \tilde{C}_{1-\gamma})\tau + C_{\nabla J}(1 + C_V)|A|\tau^2 \frac{\alpha_{t-\tau}}{1 - \gamma} + C_{\nabla J}(1 + C_V)\alpha_T + C_{\nabla J}\epsilon_{FA} \right) \quad (113)
\]
where the last inequality has already been established in Sec. B.1 with the choice of the mixing time \( \tau = \tau_T \).

### B.2.3 End of the proof of Th. 6.2

Combining Eq. (100) with Eqs. (101), (106) and (113) yields:
\[
\mathbb{E}[J(\theta_{t+1})] \geq \mathbb{E}[J(\theta_t)] + \frac{\alpha_t}{1 - \gamma} \mathbb{E}[|\nabla J(\theta_t)|^2] - C \frac{\alpha_t}{1 - \gamma} \sqrt{\mathbb{E}[|\nabla J(\theta_t)|^2]} \sqrt{\mathbb{E}[|\omega_t - \tilde{\omega}_t(\theta_t)|^2]}
- C \frac{\alpha_t}{1 - \gamma} ((C_{1-\gamma}^1 + C_{1-\gamma}^2 \tau^2) \frac{\alpha_{t-\tau}}{1 - \gamma} + C_{1-\gamma}^3 \alpha_T + C_{\nabla J}\epsilon_{FA}) - CL_{\nabla J} \frac{\alpha_t^2}{(1 - \gamma)^2}, \quad (114)
\]
where \( C_{1-\gamma}^1 := (1 + C_V)C_{\nabla J} + \tilde{C}_{1-\gamma} + C_{1-\gamma}^2 := C_{\nabla J} + C_{\nabla J}(1 + C_V)|A| \) and \( C_{1-\gamma}^3 := C_{\nabla J} + C_{\nabla J}(1 + C_V) \).

Rearranging and summing this inequality for \( t = \tau_T \) to \( T \) lead to:
\[
\frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[|\nabla J(\theta_t)|^2] \leq U_1(T) + U_2(T) + U_3(T) + CC_{\nabla J}\epsilon_{FA}, \quad (115)
\]
where
\[
U_1(T) := \frac{1 - \gamma}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \frac{1 - \gamma}{\alpha_t} (\mathbb{E}[J(\theta_{t+1})] - \mathbb{E}[J(\theta_t)]),
\]
\[
U_2(T) := \frac{C}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \left( (C_{1-\gamma}^1 \tau_T + C_{1-\gamma}^2 \tau_T^2) \frac{\alpha_{t-\tau_T}}{1 - \gamma} + C_{1-\gamma}^3 \alpha_T + L_{\nabla J} \frac{\alpha_t}{1 - \gamma} \right),
\]
\[
U_3(T) := \frac{C}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \sqrt{\mathbb{E}[|\nabla J(\theta_t)|^2]} \sqrt{\mathbb{E}[|\omega_t - \tilde{\omega}_t(\theta_t)|^2]}.
\]

Let us now provide estimates of each one of the quantities \( U_i(T) \) for \( i = 1, 2, 3 \).

1. Since the function \( J \) is bounded by \( \frac{U_R}{1 - \gamma} \) and the sequence \( (\alpha_t) \) is nonincreasing, the first term can be controlled as follows:
\[
U_1(T) = \frac{1 - \gamma}{T - \tau_T + 1} \left( \frac{1}{\alpha_T} \mathbb{E}[J(\theta_{T+1})] - \frac{1}{\alpha_{\tau_T-1}} \mathbb{E}[J(\theta_{\tau_T})] + \sum_{t=\tau_T}^{T} \left( \frac{1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \right) \mathbb{E}[J(\theta_t)] \right)
\leq \frac{U_R}{T - \tau_T + 1} \left( \frac{1}{\alpha_T} + \frac{1}{\alpha_{\tau_T-1}} + \frac{1}{\alpha_T} - \frac{1}{\alpha_{\tau_T-1}} \right)
\leq \frac{U_R}{T - \tau_T + 1} \frac{2}{\alpha_T}
= O \left( \frac{T^{\alpha^{-1}}}{\alpha_T} \right).
\]

2. We can observe from the policy gradient that \( C_{\nabla J} = O((1 - \gamma)^{-2}), L_V = O((1 - \gamma)^{-2}) \) and from the definition of the value function that \( C_V = O((1 - \gamma)^{-1}) \). Moreover, it follows from [Zhang et al., 2020a] Lem. 4.2] that \( L_{\nabla J} = O((1 - \gamma)^{-3}) \). As a consequence, we have that:
\[
C_{1-\gamma} = O((1 - \gamma)^{-5}), \quad \tilde{C}_{1-\gamma} = O((1 - \gamma)^{-4}), \quad C_{1-\gamma}^1 = O((1 - \gamma)^{-5}), \quad C_{1-\gamma}^2 = O((1 - \gamma)^{-3}), \quad C_{1-\gamma}^3 = O((1 - \gamma)^{-3}).
\]

We conclude the proof by remarking that items (1) to (3) above imply:

\[ U_2(T) = \frac{C}{T - \tau_T + 1} \left( (C^1_{1-\gamma} \tau_T + C^2_{1-\gamma} \tau_T^2) \sum_{t=\tau_T}^{T} \alpha_t \tau_T + C^3_{1-\gamma} (T - \tau_T + 1) \alpha_T + L_{\nabla J} \sum_{t=\tau_T}^{T} \alpha_t \right) \]

\[ \leq \frac{C}{T - \tau_T + 1} \left( (C^1_{1-\gamma} \tau_T + C^2_{1-\gamma} \tau_T^2) \sum_{t=0}^{T-\tau_T} \alpha_t + C^3_{1-\gamma} (T - \tau_T + 1) \alpha_T + L_{\nabla J} \sum_{t=0}^{T-\tau_T} \alpha_t \right) \]

\[ \leq \frac{C}{T - \tau_T + 1} \left( (C^1_{1-\gamma} \tau_T + C^2_{1-\gamma} \tau_T^2) + L_{\nabla J} \cdot (T - \tau_T + 1)^{1-\alpha} \frac{1}{1 - \alpha} + C^3_{1-\gamma} (T - \tau_T + 1) \alpha_T \right) \]

\[ = O \left( \frac{\ln^2 T}{(1 - \gamma)^6 T^{-\alpha}} \right) \quad (120) \]

3. Using the Cauchy-Schwarz inequality, we have:

\[ U_3(T) \leq \frac{C}{T - \tau_T + 1} \left( \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \right)^{\frac{1}{2}} \left( \sum_{t=\tau_T}^{T} \mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2] \right)^{\frac{1}{2}} \quad (121) \]

Define the quantities:

\[ F(T) := \frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \quad (122) \]

\[ E(T) := \frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2] \quad (123) \]

\[ K(T) := U_1(T) + U_2(T) + CC_{\nabla J} \epsilon_{FA} \quad (124) \]

Using these definitions, we can rewrite Eq. (115) as follows:

\[ F(T) \leq C \sqrt{F(T)} \sqrt{E(T)} + K(T) \]

Solving this inequality yields:

\[ F(T) = O(E(T)) + O(K(T)) \quad (125) \]

We conclude the proof by remarking that items (1) to (3) above imply:

\[ K(T) = O(T^{\alpha-1}) + O \left( \frac{\ln^2 T}{(1 - \gamma)^6 T^{-\alpha}} \right) + O \left( \frac{\epsilon_{FA}}{(1 - \gamma)^2} \right) \quad (126) \]

Eqs. (125) and (126) combined can be explicitly written as follows:

\[ \frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = O(T^{\alpha-1}) + O \left( \frac{\ln^2 T}{(1 - \gamma)^6 T^{-\alpha}} \right) + O \left( \frac{\epsilon_{FA}}{(1 - \gamma)^2} \right) \]

\[ + O \left( \frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2] \right) \]

Thus, by combining with the result of Theorem [6.1] we have:

\[ \frac{1}{T - \tau_T + 1} \sum_{t=\tau_T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = O(T^{\xi-1}) + O \left( \frac{\ln T}{1 - \gamma} T^{-\beta} \right) + O \left( \frac{T^{-2(\alpha-\xi)}}{(1 - \gamma)^2} \right) + O(T^{-2(\xi-\beta)}) \quad (127) \]
Then, we can write
\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = \frac{1}{T} \left( \sum_{t=1}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2] + \sum_{t=T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \right)
\leq C \ln T + \mathcal{O}\left(\frac{1}{T^{-\tau_T} + 1} \sum_{t=T}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \right)
= \mathcal{O}(T^{-\alpha-1}) + \mathcal{O}\left(\frac{\ln T}{(1-\gamma)^6 T^{-\beta}}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{-2(\xi-\beta)}) + \mathcal{O}\left(\frac{\epsilon_{FA}}{(1-\gamma)^2}\right).
\]
This completes the proof.

**B.2.4 Proof of Cor. 6.3**

The result is a consequence of combining Ths. \[6.1\] and \[6.2\] and simplifying the obtained rate using the fact that \(0 < \beta < \xi < \alpha < 1\).

**C Proof of the stability result**

The proof is inspired from the techniques used in \[Konda and Tsitsiklis, 2003a\] and \[Lakshminarayanan and Bhatnagar, 2017\]. Note though that our proof deviates from a simple application of these results. On the one hand, the approach of Konda and Tsitsiklis \[Konda and Tsitsiklis, 2003a\] is not sufficient to tackle the case of our three timescales algorithms which is more involved than the standard two timescales actor-critic algorithm. On the other hand, the result of \[Lakshminarayanan and Bhatnagar, 2017\] extending the rescaling technique of \[Borkar and Meyn, 2000\] to two timescales stochastic approximation algorithms does not handle the Markovian noise and only addresses the case of additive martingale noise.

Before proceeding with the proof, we state the stability result with all the required assumptions.

**C.1 Assumptions and stability theorem**

We first introduce a useful assumption regarding the increments of the actor iterates.

**Assumption C.1.** There exists a constant \(C > 0\) s.t. for every \(t \in \mathbb{N}\), \(\|\theta_{t+1} - \theta_{t}\| \leq \alpha_t C\).

In order to satisfy this assumption, one can slightly change the update rule of the actor sequence \((\theta_t)\) of our algorithm to bound its increments. This trick was previously used in \[Konda, 2002\] p. 80 for instance and considered later in \[Zhang et al., 2020\]. Let \(\Gamma : \mathbb{R}^m \rightarrow \mathbb{R}\) be a function assumed to satisfy the following inequalities for some positive constants \(C_1 < C_2\): for every \(\omega \in \mathbb{R}^m\), \(\|\omega\| \Gamma(\omega) \leq [C_1, C_2]\), and for every \(\omega_1, \omega_2 \in \mathbb{R}^m\), \(|\Gamma(\omega_1) - \Gamma(\omega_2)| \leq \frac{C_3 \|\omega_1 - \omega_2\|}{\mathbb{E}[\|\omega_1 - \omega_2\|]}\). An example of such function as provided in \[Konda, 2002\] is for instance the function defined for every \(\omega \in \mathbb{R}^m\) by:
\[
\Gamma(\omega) := 1_{\|\omega\| \leq C_0} + \frac{1 + C_0}{1 + \|\omega\|} 1_{\|\omega\| \geq C_0},
\]
where \(C_0\) is some given positive constant. Given such a projection-like function \(\Gamma\), we replace the update rule of the actor of our actor-critic algorithm (see Algorithm \[\|\]) by a modified update rule guaranteeing Assumption C.1 above as follows:
\[
\theta_{t+1} = \theta_t + \alpha_t \frac{1}{1 - \gamma} \Gamma(\omega_t) \delta_{t+1} \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t).
\]

We introduce an additional assumption on the stepsizes complementing Assumption \[5.2\]

**Assumption C.2.** The sequences of positive stepsizes satisfy the following:

(i) The sequences \((\beta_t)\), \((\alpha_t)\) and \((\xi_t)\) are nonincreasing.

(ii) For every \(t \in \mathbb{N}\), \(0 < \xi_t \leq 1\).
Theorem C.1. Let Assumptions 3.1, 5.1, 5.2, 5.4, 6.2, C.1 and C.2 hold true. Then, sup_k ([|\omega_k|] + |\omega_k|) < \infty, a.s., i.e., Assumption 5.3 holds true.

The proof of this result proceeds as for our convergence result: we address the faster timescale first before analyzing the slower one.

C.2 Faster timescale analysis

In this section, our goal is to bound the norm of the sequence (\omega_t) evolving on the fast timescale driven by the step-sizes (\beta_t) using the norm of the sequence (\tilde{\omega}_t) updated in a slower timescale defined by the stepsizes (\xi_t). In order to use a rescaling technique inspired from [Borkar and Meyn, 2000, Lakshminarayanan and Bhatnagar, 2017], we introduce a few useful notations. Define for every \theta \in \mathbb{R}^d the functions \h_\theta : \mathbb{R} \times \mathcal{S}^2 \rightarrow \mathbb{R}^{2m} and \G_\theta : \mathbb{R} \times \mathcal{S}^2 \rightarrow \mathbb{R}^{2m \times 2m}

for every \( y = (r, s, s') \in \mathbb{R} \times \mathcal{S}^2 \) by:

\[
\h_\theta(y) := \begin{bmatrix} r \phi(\tilde{s}) \\ 0 \end{bmatrix}, \quad \G_\theta(y) := \begin{bmatrix} \phi(\tilde{s}) \phi(\tilde{s})^T & -\gamma \phi(\tilde{s}) \phi(s')^T \end{bmatrix}.
\]

Consider the sequences \( r_k := (\omega_k^T, \tilde{\omega}_k^T)^T \) and \( Y_{k+1} := (\tilde{S}_k, S_{k+1}, R_{k+1}) \). Given the update rules of the sequences (\omega_k) and (\tilde{\omega}_k) from our algorithm, we have the following decomposition:

\[
r_{k+1} = r_k + \beta_k \left( h_{\theta_k}(Y_{k+1}) - G_{\theta_k}(Y_{k+1}) r_k \right) + \beta_k M_{k+1} r_k + \beta_k \eta_{k+1},
\]

where \( (M_{k+1}) \) is a \( 2m \times 2m \)-matrix valued martingale difference sequence w.r.t. the filtration \( (\mathcal{F}_k) \) (where the \( \sigma \)-field is generated by all the r.v.s up to time \( k \)) defined for every \( k \in \mathbb{N} \) by:

\[
M_{k+1} := \begin{bmatrix} 0 & \gamma \phi(S_k) \phi(S_{k+1}) - \mathbb{E} [\phi(S_{k+1}) | \mathcal{F}_k]^T \end{bmatrix},
\]

and \( (\eta_{k+1}) \) is a \( 2m \)-vector valued sequence defined for every \( k \in \mathbb{N} \) by:

\[
\eta_{k+1} := \frac{\xi_k}{\beta_k} \begin{bmatrix} 0 \\ \omega_{k+1} - \tilde{\omega}_k \end{bmatrix}.
\]

Consider now the functions \( \hat{h} : \mathbb{R}^d \rightarrow \mathbb{R}^{2m} \) and \( \hat{G} : \mathbb{R}^d \rightarrow \mathbb{R}^{2m \times 2m} \) defined for every \( \theta \in \mathbb{R}^d \) by:

\[
\hat{h}(\theta) := \frac{h(\theta)}{\theta}, \quad \hat{G}(\theta) := \left[ \begin{array}{c} \hat{G}(\theta) - \gamma \Phi^T D_{p,\theta} P_\theta \Phi \\ 0 \end{array} \right],
\]

where we recall that \( h(\theta) = \Phi^T D_{p,\theta} R_\theta \) and \( \hat{G}(\theta) = \Phi^T D_{p,\theta} \Phi \).

Let the sequence of nonnegative integers \( (k_j^\beta) \) be defined by:

\[
k_0^\beta = 0, \quad k_{j+1}^\beta = \min \left\{ k > k_j^\beta : \sum_{l=k_j}^{k-1} \beta_l > T \right\}, \quad (128)
\]

where \( T \) is a positive constant that will be chosen appropriately later on. For notational convenience, in the rest of Section C.2 we will simply use the notation \( (k_j) \) for the sequence \( (k_j^\beta) \). The superscript \( \beta \) will be useful when considering a different timescale in the upcoming section.

Then, for any \( j \in \mathbb{N} \), we can introduce the rescaled iterates \( \hat{r}_k^j = \frac{r_k}{\max(1, |r_k|)} \) defined for every \( k \geq k_j \) and which satisfy the following recurrence relation:

\[
\hat{r}_{k+1}^j = \hat{r}_k^j + \beta_k \left( \frac{\hat{h}(\theta_k)}{\max(1, |r_k|)} - \hat{G}(\theta_k) \hat{r}_k^j \right) + \beta_k \hat{r}_{k+1}^j + \beta_k \frac{\eta_{k+1}}{\max(1, |r_k|)},
\]

where for \( k \geq k_j \), the term \( \hat{r}_{k+1}^j \) is defined by:

\[
\hat{r}_{k+1}^j := \left( \frac{h_{\theta_k}(Y_{k+1}) - \hat{h}(\theta_k)}{\max(1, |r_k|)} - (G_{\theta_k}(Y_{k+1}) - \hat{G}(\theta_k)) \hat{r}_k^j \right) + \beta_k M_{k+1} \hat{r}_k^j.
\]
We also introduce the iterates \((r_k^j)\) defined as follows: 
\[ r_{k+1}^j = r_k^j + \beta_k \left( \frac{\tilde{h}(\theta_k)}{\max(1, \|r_k^j\|)} - \tilde{G}(\theta_k)r_k^j \right) + \beta_k \frac{\eta_{k+1}}{\max(1, \|r_k^j\|)}. \]

Observing that the sequence \(r_k^j\) can be written as \((\omega_k^j, \bar{\omega}_k^j)\) and given the update rule of \((r_k^j)\), we have the following for every \(j \in \mathbb{N}, k \geq k_j:\)

\[
\begin{align*}
\omega_{k+1}^j & = \omega_k^j + \beta_k \left( \frac{h(\theta_k)}{\max(1, \|r_k^j\|)} + \gamma \Phi^T D_\rho \theta_k P_\theta \Phi \omega_k^j - \tilde{G}(\theta_k) \omega_k^j \right), \\
\bar{\omega}_{k+1}^j & = \bar{\omega}_k^j + \xi_k (\omega_{k+1}^j - \omega_k^j),
\end{align*}
\]

Before proceeding, we recall two useful lemmas which we will repeatedly use in the proofs.

**Lemma C.2.** Let \(\lambda \in [0, 1)\). Suppose that \((u_k)\) and \((\varepsilon_k)\) are nonnegative sequences satisfying \(u_{k+1} \leq \lambda u_k + \varepsilon_k\).

If \(\sup_k \varepsilon_k < \infty\), then \(\sup_k u_k < \infty\).

**Lemma C.3.** Let \(G \in \mathbb{R}^{m \times m}\) be a matrix verifying for every \(\omega \in \mathbb{R}^m\), \(\omega^T G \omega \geq \varepsilon \|\omega\|^2\) where \(\varepsilon > 0\) is a constant. Then, for sufficiently small \(\gamma > 0\), \(||(I - \gamma G)\omega|| \leq (1 - \frac{1}{2}\gamma \varepsilon)\|\omega\| \leq e^{-\frac{1}{2}\gamma \varepsilon}\|\omega\||\).  

**Lemma C.4.** We have the following:

(i) There exists a constant \(C > 0\) s.t. \(\sup_j \max_{k_j \leq k \leq k_{j+1}} \|r_k^j\| \leq C\).

(ii) \(\lim_j \max_{k_j \leq k \leq k_{j+1}} \|r_k^j - r_k^j\| = 0\), a.s.

(iii) There exists a constant \(C' > 0\) s.t. \(\sup_j \max_{k_j \leq k \leq k_{j+1}} \|r_k^j\| \leq C', \) a.s.

**Proof.** (i) Let us show that there exists a positive constant \(\bar{C} > 0\) s.t. \(\sup_j \max_{k_j \leq k \leq k_{j+1}} \|\omega_k^j\| \leq \bar{C}\). For \(j\) sufficiently large s.t. Lem. C.3 holds and for \(k\) between \(k_j\) and \(k_{j+1}\), we have

\[
\|\omega_{k+1}^j\| \leq \|(I - \beta_k \bar{G}(\theta_k)) \omega_k^j\| + \beta_k \frac{\|h(\theta_k)\|}{\max(1, \|r_k^j\|)} + \beta_k \gamma \bar{G}(\theta_k)\omega_k^j + \beta_k C_2 \|\bar{\omega}_k^j\| \leq (1 - \frac{1}{2}\beta_k \varepsilon)\|\omega_k^j\| + \beta_k C_1 \|\bar{\omega}_k^j\| \]

\[
\leq e^{-\frac{1}{2}\gamma \varepsilon} \left( \sum_{i=k}^{k_{j+1}} \beta_i \|\omega_i^j\| + \left( \sum_{i=k}^{k_{j+1}} \beta_i \right) \frac{C_1}{\max(1, \|r_k^j\|)} + C_2 \left( \sum_{i=k}^{k_{j+1}} \beta_i \|\omega_i^j\| \right) \right) \]

\[
\leq 1 + T' C_1 + C_2 \left( \sum_{i=k}^{k_{j+1}} \beta_i \|\omega_i^j\| \right) \]

where \(C_1, C_2\) are two positive constants, \(T'\) is a positive constant (which we do not explicit) s.t. \(T' > T\), the second inequality follows from the fact that the matrix \(\bar{G}(\theta)\) is \(\varepsilon\)-uniformly positive definite (i.e., for every \(\omega \in \mathbb{R}^m\), \(\omega^T \bar{G}(\theta)\omega \geq \varepsilon \|\omega\|^2\)) together with Lem. C.3 and the last inequality stems from the fact that \(\|\bar{\omega}_k^j\| \leq 1\) by definition.

We now relate the term \(\|\bar{\omega}_k^j\|\) to the quantity \(\max_{k \leq k \leq k_{j+1}} \|\omega_k^j\|\). For every \(i \in \{k_j, \cdots, k_{j+1} - 1\},\)

\[
\|\bar{\omega}_{k+1}^j\| \leq \|\omega_0^j\| + \xi_i \|\omega_{k+1}^j\| \leq \|\omega_0^j\| + \sum_{i=k_j}^{k_{j+1}} \xi_i \|\omega_{k+1}^j\| \leq 1 + T' \left( \max_{k \leq k \leq k_{j+1}} \|\omega_k^j\| \right) \]

Notice then that \(\|\bar{\omega}_k^j\|\) is bounded whenever \(\|\omega_k^j\|\) is bounded. It remains to show that the sequence \((\omega_k^j)\) is bounded. For this purpose, combining the above inequality with Eq. (130) yields

\[
\max_{k \leq k \leq k_{j+1}} \|\omega_k^j\| \leq (1 + T' C_1 + C_2 T') + C_2 T' \left( \sum_{i=k_j}^{k_{j+1}} \beta_i \left( \max_{k \leq k \leq k_{j+1}} \|\omega_k^j\| \right) \right) \]

\[
\leq (1 + T' C_1 + C_2 T') + C_2 T' \left( \max_{k \leq k \leq k_{j+1}} \|\omega_k^j\| \right). \]
As a consequence, we have:

\[
\max_{k_j \leq k \leq k_{j+1}} \|\omega_k\| \leq \frac{1 + T'C_1 + C_2 T'}{v},
\]

which concludes the proof.

(ii) This result is a consequence of applying [Konda and Tsitsiklis, 2003a, Lem. 9] to the sequence \((r_t)\). Note that Assumption 6 in [Konda and Tsitsiklis, 2003a] is not needed for this result to hold since we proved item one. This means that the matrix \(G(\theta)\) is not required to be uniformly positive definite (see [Konda and Tsitsiklis, 2003a, Assumption 6]). We leave the verification of the remaining technical assumptions to the reader.

(iii) This item follows from combining the two first items with the triangular inequality. Remark that the second item implies that the sequence \((\max_{k_j \leq k \leq k_{j+1}} \|r_k^j - r_{k-1}^j\|)\) is a.s. bounded.

Recall that for every \(\bar{\omega} \in \mathbb{R}^m, \theta \in \mathbb{R}^d\),

\[
\omega_* (\bar{\omega}, \theta) = G(\theta)^{-1} \left( h(\theta) + \gamma \Phi^T D_{\rho, \theta} P \Phi \bar{\omega} \right).
\]

Now, we define for every \(j \in \mathbb{N}\) and for every \(\bar{\omega} \in \mathbb{R}^m, \theta \in \mathbb{R}^d\) a rescaled version \(\hat{\omega}^* (\bar{\omega}, \theta)\) of \(\omega_* (\bar{\omega}, \theta)\) as follows:

\[
\hat{\omega}^* (\bar{\omega}, \theta) := G(\theta)^{-1} \left( h(\theta) \max(1, \|r_k\|) + \gamma \Phi^T D_{\rho, \theta} P \Phi \bar{\omega} \right). \tag{131}
\]

Notice that there exists a constant \(C^* > 0\) s.t. for every \(j \in \mathbb{N}\), for every \(\bar{\omega} \in \mathbb{R}^m, \theta \in \mathbb{R}^d\),

\[
\max (\|\omega_* (\bar{\omega}, \theta)\|, \|\hat{\omega}^* (\bar{\omega}, \theta)\|) \leq C^* (1 + \|\bar{\omega}\|). \tag{132}
\]

**Lemma C.5.** There exists \(j_* \in \mathbb{N}, T_* > 0\) s.t for every integer \(j \geq j_*\) and \(T \geq T_*\) (\(T\) as in the definition of \(k_j\)), if \(\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\| > C_1 (1 + \|\bar{\omega}_{k_j}\|)\) for some constant \(C_1 > 0\), then,

\[
\|\omega_{k_{j+1}} - \omega_* (\bar{\omega}_{k_{j+1}}, \theta_{k_{j+1}})\| \leq \frac{3}{4} \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|, \text{a.s.}
\]

**Proof.** Notice that if \(\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\| > C_1 (1 + \|\bar{\omega}_{k_j}\|)\), using Eq. (132), we obtain that:

\[
\|r_{k_j}\| = \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\| = \sqrt{\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + \|\omega_{k_{j+1}} - \omega_* (\bar{\omega}_{k_{j+1}}, \theta_{k_{j+1}})\|^2} \\
\leq \sqrt{2 \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + 2 \|\omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|^2} \\
\leq \sqrt{2} \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\| + \sqrt{2C^* (1 + \|\bar{\omega}_{k_j}\|)^2 + \|\omega_{k_j}\|^2} \\
\leq \sqrt{2} \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\| + \sqrt{2C^* + (1 + \|\bar{\omega}_{k_j}\|)} |\omega_{k_j}|.
\]

As a consequence, we have:

\[
\frac{\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|}{\max(1, \|r_{k_j}\|)} \geq \frac{\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|}{\sqrt{2 \|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + (\sqrt{2C^* + 1})(1 + \|\bar{\omega}_{k_j}\|)}} \geq \frac{1}{\sqrt{2 + \frac{2C^*+1}{C_1}}}. \tag{133}
\]

Then, setting \(C_2 := \sqrt{2 + \frac{2C^*+1}{C_1}}\), it follows that:

\[
\frac{\|\omega_{k_{j+1}} - \omega_* (\bar{\omega}_{k_{j+1}}, \theta_{k_{j+1}})\|}{\|\omega_{k_j} - \omega_* (\bar{\omega}_{k_j}, \theta_{k_j})\|} \leq C_2 (\|\hat{\omega}^*_{k_{j+1}} - \hat{\omega}^*_{k_j}\| + \|\omega^*_{k_{j+1}} - \hat{\omega}^*_{k_{j+1}}\|)
\]

\[
\leq C_2 (\|\hat{\omega}^*_{k_{j+1}} - \hat{\omega}^*_{k_j}\| + \|\omega^*_{k_{j+1}} - \hat{\omega}^*_{k_{j+1}}\|)
\]

\[
\leq C_2 (\|\hat{\omega}^*_{k_{j+1}} - \hat{\omega}^*_{k_j}\| + \|\omega^*_{k_{j+1}} - \hat{\omega}^*_{k_{j+1}}\|)
\]
Since the first term of the right-hand side converges a.s. to zero as \( j \) goes to infinity by Lem. C.3, there exists \( j_0 \in \mathbb{N} \) s.t. for every \( j \geq j_0 \),
\[
\|\tilde{\omega}_{k+1}^j - \omega_{k+1}^j\| \leq \frac{1}{4C_2}, \text{ a.s.}
\] (134)

We now establish a bound for the second term in Eq. (133). For every \( k \leq k < k_{j+1} \), we have that:
\[
\|\omega_{k+1}^j - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\| = \|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_{k}) + \beta_k(\frac{h(\theta_k)}{\max(1, \|r_k\|)} + \frac{\gamma \Phi^T D \rho \beta \Phi \tilde{\omega}_{k}^j - G(\theta_k)\omega_{k}^j}{C}) + \frac{\gamma \Phi^T D \rho \beta \Phi \tilde{\omega}_{k+1}^j - G(\theta_k)\omega_{k+1}^j}{C})\|
\]
\[
+ \|\omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k) - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\| 
\]
\[
\leq \|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k) - \beta_k(\tilde{\omega}_{k}^j - \tilde{\omega}_{k+1}^j(\omega_{k}^j, \theta_k))\| 
\]
\[
+ \|\omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k) - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\| 
\]
\[
\leq \|I - \beta_k \tilde{G}(\theta_k)\| \|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| + C(\xi_k + \alpha_k)
\]

where \( C > 0 \) is a constant coming from Lem. C.4 and the last inequality stems from the fact that the function \((\omega, \theta) \mapsto \omega(\tilde{\omega}, \theta)\) is Lipschitz continuous for every \( j \) (by the same arguments as for the proof showing that the function \( U \) is Lipschitz before Lemma A.5). Similarly to the proof of the first item of Lem. C.4, we have:
\[
\|\omega_{k+1}^j - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\| \leq e^{-\frac{1}{2}T} \|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| + C \sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k)
\]
\[
\leq e^{-\frac{1}{2}T} \left( \|\omega_{k}^j\| + \|\omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| \right) + C \sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k)
\]
(135)

By definition, \( \|\omega_{k}^j\| \leq 1 \), \( \|\tilde{\omega}_{k}^j\| \leq 1 \), and it stems from Eq. (132) that \( \|\omega_{k}^j\| + \|\omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| \leq C' \) for some \( C' > 0 \). Choosing \( T \geq \frac{2\ln(4C'/C_2)}{e} \), we obtain: \( e^{-\frac{1}{2}T} \left( \|\omega_{k}^j\| + \|\omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| \right) \leq \frac{1}{4C_2} \). We also have that for every \( j \in \mathbb{N} \),
\[
\sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k) \leq \max_{k \leq k \leq k_{j+1}} \frac{\xi_k + \alpha_k}{\beta} k^{\epsilon T'}
\]
Since \( (\xi_k + \alpha_k)/\beta_k \to 0 \), there exists \( j_1 \in \mathbb{N} \) s.t., for every \( j \geq j_1 \),
\[
C \sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k) \leq \frac{1}{4C_2}
\]
As a consequence, Eq. (135) implies that for every \( j \geq \max(j_0, j_1) \),
\[
\|\omega_{k+1}^j - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\| \leq \frac{1}{2C_2}
\]
(136)

Combining Eq. (133) with Eqs. (134) and (136) yields for every \( j \geq \max(j_0, j_1) \),
\[
\frac{\|\omega_{k+1}^j - \omega_{k}^j(\tilde{\omega}_{k+1}^j, \theta_{k+1})\|}{\|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\|} \leq C_2 \left( \frac{1}{4C_2} + \frac{1}{2C_2} \right) = \frac{3}{4},
\]
which is the desired inequality.

**Theorem C.6.** There exists a constant \( C > 0 \) s.t. for every \( j \in \mathbb{N} \),
\[
(i) \ \|\omega_{k}^j - \omega_{k}^j(\tilde{\omega}_{k}^j, \theta_k)\| \leq C(1 + \|\tilde{\omega}_{k}^j\|), \text{ a.s.}
\]
\[
(ii) \ \|\omega_{k}^j\| \leq C(1 + \|\tilde{\omega}_{k}^j\|), \text{ a.s.}
\]
\[
(iii) \ \max_{k \leq k \leq k_{j+1}} \|\omega_{k}^j\| \leq C(1 + \|\tilde{\omega}_{k_{j+1}}\|), \text{ a.s.}
\]

**Proof.** (i) The proof follows exactly the same path than the proof of [Lakshminarayanan and Bhatnagar, 2017 Th. 7-(ii)]. We reproduce it here for completeness. On a set of positive probability, let us assume on the contrary that there exists a monotonically increasing sequence \((j_i)\) for which \(C_{j_i} \uparrow \infty\) as \( i \to \infty \) and \( \|\omega_{k_{j_i}}\| \geq C_{j_i}(1 + \|\tilde{\omega}_{k_{j_i}}\|) \). Now, from Lem. C.5, we know that if \( \|\omega_{k_{j_i}} - \omega_{k_{j_i}}(\tilde{\omega}_{k_{j_i}}, \theta_{k_{j_i}})\| > C_{1}(1 + \|\tilde{\omega}_{k_{j_i}}\|) \), then \( \|\omega_{k_{j_i}} - \omega_{k_{j_i}}(\tilde{\omega}_{k_{j_i}}, \theta_{k_{j_i}})\| \) for \( i \geq j \) falls at an exponential rate until it is within the ball of radius \( C_{1}(1 + \|\tilde{\omega}_{k_{j_i}}\|) \). Thus, corresponding to the sequence \((j_i)\), there must exist another sequence \((j'_i)\) s.t. \( j_{i-1} \leq j'_i \leq j_i \) and \( \|\omega_{k_{j'_i}} - \omega_{k_{j'_i}}(\tilde{\omega}_{k_{j'_i}}, \theta_{k_{j'_i}})\| \) is within the ball of radius \( C_{1}(1 + \|\tilde{\omega}_{k_{j'_i}}\|) \) and \( \|\omega_{k_{j_i}} - \omega_{k_{j_i}}(\tilde{\omega}_{k_{j_i}}, \theta_{k_{j_i}})\| \)
is greater than $C_3(1 + \|\hat{\omega}_{k_j}\|)$. However, we know from Lem. C.4 that the iterates can only grow by a factor of $C'$ between the time $k_{j-1}$ and $k_j$. This leads to a contradiction. We conclude that $\|\omega_{k_j} - \omega_n(\hat{\omega}_{k_j}, \theta_k)\| \leq \bar{C}(1 + \|\hat{\omega}_{k_j}\|)$ for some $\bar{C} > 0$.

(ii) The inequality is a consequence of the first item combined with Eq. (132).

(iii) Using the definition of the sequence $(\hat{\omega}^n_j)$ and the third item of Lem. C.4 (providing the constant $C'$) combined with the second item of the present theorem, we obtain the desired result as follows:

$$\|\omega_k\| = \max(1, \|(\omega_{k_j}, \hat{\omega}_{k_j})\|)\|\xi\| \leq (1 + \|\omega_{k_j}\|) + \|\hat{\omega}_{k_j}\| C' \leq C'(1 + \|\hat{\omega}_{k_j}\|),$$

where $C := C'(1 + \bar{C})$ and $\bar{C}$ comes from the proof of the first item.

\[\square\]

C.3 Slower timescale analysis

We now turn to the analysis of the sequence $(\hat{\omega}_i)$ evolving in a slower timescale than that of the sequence $(\omega_i)$. Recall the update rule of the sequence $(\hat{\omega}_i)$:

$$\hat{\omega}_{k+1} = \hat{\omega}_k + \xi_k(\omega_{k+1} - \hat{\omega}_k).$$

Given a constant $T > 0$, let $(k^\beta_j)$ be defined as in Eq. (125) and define the sequence $(k^\xi_n)$ (which we will sometimes simply denote $(\hat{k}_n)$) in the rest of this section when unambiguous) as follows:

$$k^\xi_0 = 0, \quad k^\xi_{n+1} = \min \left\{ k^\beta_j : j \in \mathbb{N}, \sum_{l=k_n}^{k^\beta_j} \xi_l > T \right\}.$$

Since $\xi_k/\beta_k$ converges to 0, there exists $C_\xi > 0$ such that $T \leq \sum_{l=k_n}^{k^\xi_n} \xi_l \leq C_\xi T$.

Similarly to the previous section, for every $n \in \mathbb{N}$, we define the rescaled iterates $(\hat{\omega}^n_k)$ and $(\hat{\omega}^n_k)$ for every $k \geq k_n$ as follows:

$$\begin{cases} 
\hat{\omega}^n_{k_{n+1}} = \frac{\omega_{k_n}}{\max(1, \|r_{k_n}\|)} + \frac{R_{k_{n+1}}}{\max(1, \|r_{k_{n}}\|)} + \gamma \phi(S_{k^\xi_{n+1}} T \hat{\omega}^n_k - \phi(S_{k^\xi_{n}} T \hat{\omega}^n_k)) \hat{\omega}^{\gamma}_{k_{n+1}} = \frac{\omega_{k_n}}{\max(1, \|r_{k_n}\|)} \hat{\omega}^n_{k_{n+1}} = \omega_{k_n} + \theta_k(\hat{\omega}^n_k - \hat{\omega}^n_{k^\xi_{n+1}}),
\end{cases}$$

and their noiseless counterparts $(\omega^n_k)$ and $(\hat{\omega}^n_k)$ are defined for every $n \in \mathbb{N}, k \geq k_n$ by:

$$\begin{cases} 
\omega^n_{k_{n+1}} = \frac{\omega_{k_n}}{\max(1, \|r_{k_n}\|)} + \frac{h(\theta_{k_n})}{\max(1, \|r_{k_n}\|)} + \gamma \Phi T D_p \theta_k \Phi \omega^n_k - G(\theta_k) \omega^n_k \omega_{k_{n+1}} = \omega_{k_n} + \xi_k(\hat{\omega}^n_k - \hat{\omega}^n_{k_{n+1}}),
\end{cases}$$

The following lemma states the almost sure boundedness of the above rescaled and noiseless iterates.

\textbf{Lemma C.7.} The following assertions hold true:

(i) $\sup_n \max_{k_n \leq k \leq k^\xi_n(\|\hat{\omega}^n_k\| + \|\omega^n_k\|) < \infty, a.s.$

(ii) $\sup_n \max_{k_n \leq k \leq k^\xi_n(\|\hat{\omega}^n_k\| + \|\omega^n_k\|) < \infty, a.s.$

\textbf{Proof.} (i) Let $n \in \mathbb{N}$. By definition of the sequence $(k_n)$, there exists $j \in \mathbb{N}$ s.t. $k_n = k^\beta_j$. There exists $C > 0$ (independent of $n$) s.t. for every $k \in \{k^\beta_j, \cdots, k^\beta_{j+1} - 1\}$, a.s.,

$$\|\hat{\omega}^n_{k_{n+1}}\| \leq (1 - \xi_k)\|\hat{\omega}^n_k\| + \xi_k\|\omega^n_{k_{n+1}}\| \leq \|\hat{\omega}^n_k\| + \xi_k(1 + \|\hat{\omega}^n_k\|),$$

where we used Th. C.6 (iii) for the last inequality. It follows that for every \( k \in \{ k_j^\beta, \cdots, k_{j+1}^\beta - 1 \} \), a.s.,

\[
\| \hat{w}^n_{k+1} \| \leq \left( 1 + C \sum_{l=k_j^\beta}^k \xi_l \right) \| \hat{w}^n_{k_j^\beta} \| + C \sum_{l=k_j^\beta}^k \xi_l \leq e^{C \sum_{l=k_j^\beta}^k \xi_l} \| \hat{w}^n_{k_j^\beta} \| + C \sum_{l=k_j^\beta}^k \xi_l.
\]

As a consequence, using the notation \( u_j := \sum_{l=k_j^\beta}^{k_{j+1}^\beta - 1} \xi_l \) for every \( j \in \mathbb{N} \), we obtain that a.s.,

\[
\| \hat{w}^n_{k+1} \| \leq e^{Cu_j} \| \hat{w}^n_{k_j^\beta} \| + Cu_j, \tag{139}
\]

For every \( l, p \in \mathbb{N} \), let \( \mathcal{U}(l, p) \) be the set of integers \( j \) s.t. \( l \leq k_j^\beta \leq p \). Recall that for every \( n \in \mathbb{N} \), there exist integers \( j_{n+1} > j_n \) s.t. \( k_n = k_{j_n}^\beta \) and \( k_{n+1} = k_{j_{n+1}}^\beta \) by definition of the sequence \( (k_n) \). Then, using Eq. (139), we have for every \( j \in \{ j_n, k_{n+1} \} \), a.s.,

\[
\| \hat{w}^n_{k+1} \| \leq \left( \prod_{i \in \mathcal{U}(k_n, k_{j_n}^\beta - 1)} e^{Cu_i} \| \hat{w}^n_{k_n} \| + C \sum_{p \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} \left( \prod_{i \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} e^{Cu_i} \right) u_p \right) = e^{C \sum_{l=k_n}^{k_{j+1}^\beta - 1} \xi_l} \| \hat{w}^n_{k_n} \| + C \sum_{p \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} e^{C \sum_{l=k_{j+1}^\beta}^{k_{j+1}^\beta - 1} \xi_l} u_p \leq e^{CC \xi T} + Ce^{CC \xi T} C \xi T,
\]

where the last inequality comes from the facts that \( \| \hat{w}^n_{k_n} \| \) is bounded by 1 and that \( \sum_{l=k_n}^{k_{j+1}^\beta} \xi_l \leq C \xi T \). To conclude, notice that this bound also holds for any \( k \in \{ k_n, \cdots, k_{n+1} \} \) and use Th. C.6 (iii) to bound \( \| \hat{w}^n_k \| \).

(ii) The proof of this item follows a similar path to the first one. Notice that the iterates considered in this item are noiseless versions of their counterparts which were shown to be bounded in the first item.

\[\square\]

**Lemma C.8.** \( \lim_n \max_{k_n \leq k \leq k_{n+1}} \left| \left( \hat{w}^n_k, \hat{w}^n_{k+1} \right) - \left( w^n_k, w^n_{k+1} \right) \right| = 0 \).

**Proof.** Let \( n \in \mathbb{N} \). Consider the shorthand notations \( x^n_k := \hat{x}^n_k - w^n_k \) and \( y^n_k := \hat{y}^n_k - w^n_k \) for \( k \geq k_n^\beta \). Note that for every \( k \geq k_n^\beta \), the sequences \( (x^n_k)_k \) and \( (y^n_k)_k \) satisfy the recurrence relations:

\[
\begin{align*}
x_{k+1}^n &= x_k^n + \beta_k (\gamma \Phi^T D_{\rho, \theta_k} P_{\theta_k} \Phi y_k^n - \tilde{G}(\theta_k) x_k^n) + \beta_k e_k^n, \\
y_{k+1}^n &= y_k^n + \xi_k (x_{k+1}^n - y_k^n),
\end{align*}
\tag{140}
\]

where the Markovian noise sequence \( (e_k^n)_k \) is defined for every \( k \geq k_n^\beta \) by:

\[
e_k^n := \frac{1}{\max(1, \| r_k^n \|)} \left[ \phi(\tilde{S}_k) R_{k+1} - h(\theta_k) \right] + \gamma \left[ \phi(\tilde{S}_k) \phi(S_{k+1})^T - \Phi^T D_{\rho, \theta_k} P_{\theta_k} \Phi \right] \hat{w}^n_k + \left[ \tilde{G}(\theta_k) - \phi(\tilde{S}_k) \phi(\tilde{S}_k)^T \right] \tilde{w}^n_k.
\tag{141}
\]

It is clear that the sequence \( (e_k^n)_k \) is a.s. bounded using Lem. C.7 Define the mapping \( x^*: \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^m \) for every \( \theta \in \mathbb{R}^d, y \in \mathbb{R}^m \) by:

\[x^*(y, \theta) := \gamma \tilde{G}(\theta)^{-1} \Phi^T D_{\rho, \theta} P_{\theta} \Phi y.\]

Then, we have the following decomposition for every \( k \geq k_n^\beta \):

\[
y_{k+1}^n = y_k^n + \xi_k (x^*(y_k^n, \theta_k) - y_k^n) + \xi_k (x_{k+1}^n - x_k^n) + \xi_k (x_k^n - x^*(y_k^n, \theta_k)) = (I_m - \xi_k \tilde{G}(\theta_k)^{-1} G(\theta_k)) y_k^n + \xi_k (x_{k+1}^n - x_k^n) + \xi_k (x_k^n - x^*(y_k^n, \theta_k)).
\tag{142}
\]
Since $\tilde{G}(\theta)^{-1}G(\theta)$ is uniformly (in $\theta$) $\kappa$-positive definite (see Assumption 6.2), Lem. C.3 implies that there exists $\kappa > 0$ s.t. for sufficiently large $n$ and $k \in \{k_n^\epsilon, \ldots, k_{n+1}^\epsilon - 1\}$,

$$
\|y_{k+1}^n\| \leq e^{-\frac{1}{2} \kappa T} \|y_{k^\epsilon}^n\| + \sum_{l=k^\epsilon}^k \xi_l \|x_{l+1}^n - x_l^n\| + \xi_l \|x_l^n - x^*(y_l^n, \theta_l)\|
$$

$$
= \sum_{l=k^\epsilon}^k \xi_l \|x_{l+1}^n - x_l^n\| + \xi_l \|x_l^n - x^*(y_l^n, \theta_l)\|
$$

$$
\leq \sum_{l=k^\epsilon}^k \xi_l \beta_l C + C_\epsilon T \max_{l \in \{k^\epsilon, \ldots, k\}} \|x_l^n - x^*(y_l^n, \theta_l)\|
$$

where the equality comes from the fact that $y_{k^\epsilon}^n = 0$ by definition and the last inequality comes from the fact that $x_{l+1}^n - x_l^n = \beta_l (\gamma T D_{p, \theta} P_0 \Phi y_l^n - \tilde{G}(\theta_l) x_l^n + \epsilon_l^n)$ and the a.s. boundedness of the sequences $(x_k^n)_k, (y_k^n)_k$ and $(\epsilon_k^n)_k$ resulting from Lem. C.7. Observe then that:

$$
\sum_{l=k^\epsilon}^{k_{n+1}^\epsilon - 1} \xi_l \beta_l = \sum_{l=k^\epsilon}^{k_{n+1}^\epsilon - 1} \frac{\xi_l}{\beta_l} \beta_l^2 \leq \max_{k^\epsilon \leq l \leq k_{n+1}^\epsilon} \left( \frac{\xi_l}{\beta_l} \right) \sum_{l=k^\epsilon}^{k_{n+1}^\epsilon - 1} \beta_l^2.
$$

(143)

Since $\sum_{n} \beta_n^2 < \infty$ and $\xi_n/\beta_n$ $\rightarrow$ 0, it follows that $\sum_{k=k^\epsilon}^{k_{n+1}^\epsilon - 1} \beta_k \epsilon_k \rightarrow 0$. Combining this result with Lem. C.11 below yields:

$$
\lim_{n \rightarrow \infty} \max_{k^\epsilon \leq k \leq k_{n+1}^\epsilon} \|y_k^n\| = 0, \ a.s.
$$

(144)

We now show the same result for the sequence $(x_k^n)_k$. First, observing that $x_{k_n}^n = 0$, we obtain by iterating Eq. (140) that:

$$
x_{k+1}^n = \sum_{l=k^\epsilon}^k \left[ \prod_{p=l+1}^k (I_m - \beta_p \tilde{G}(\theta_p)) \right] \beta_l \left( \gamma T D_{p, \theta} P_0 \Phi y_l^n + \epsilon_l^n \right).
$$

Then, similarly to the first part of the proof, there exist $C > 0$ and $\epsilon > 0$ s.t. for sufficiently large $n$ and $k_n \leq k \leq k_{n+1}^\epsilon$,

$$
\|x_{k+1}^n\| \leq C \sum_{l=k^\epsilon}^k \left[ \prod_{p=l+1}^k (1 - \frac{1}{2} \epsilon \beta_p) \right] \beta_l \|y_l^n\| + \left[ \prod_{l=k^\epsilon}^k (I_m - \beta_p \tilde{G}(\theta_p)) \right] \beta_l \epsilon_l^n
$$

$$
\leq C \frac{2}{\epsilon} \max_{k^\epsilon \leq l \leq k_{n+1}^\epsilon} \|y_l^n\| + \left[ \prod_{l=k^\epsilon}^k (I_m - \beta_p \tilde{G}(\theta_p)) \right] \beta_l \epsilon_l^n
$$

where the first inequality stems from the fact that the matrix $\tilde{G}(\theta)$ is uniformly positive definite and Lem. C.3 and the last inequality is a consequence of [Kaledin et al., 2020, Lem. 12]. Eq. (144) and Lem. C.12 below entail together that:

$$
\lim_{n \rightarrow \infty} \max_{k^\epsilon \leq k \leq k_{n+1}^\epsilon} \|x_k^n\| = 0, \ a.s.,
$$

which concludes the proof.

**Lemma C.9.** There exists a sequence $(\delta_n)$ that converges to 0 when $n \rightarrow \infty$ and a constant $C > 0$ s.t. for every $n \in \mathbb{N}$,

$$
\|\tilde{u}_{k+1}^n\| \leq e^{-\frac{1}{2} \kappa T} \|\tilde{u}_k^n\| + \delta_n + \frac{C}{\max(1, \|r_k^n\|)}.
$$
We can show that, for every $n \in \mathbb{N}$, $\omega_n := G(\theta)^{-1} \left( \frac{h(\theta)}{\max(1, \|r_n\|)} + \gamma \Phi^T D_{p, \theta} P_{\theta} \Phi \omega \right)$ for every $n \in \mathbb{N}$, $\omega \in \mathbb{R}^m$, $\theta \in \mathbb{R}^d$. It is clear that for every $k \geq k_n$:

$$\bar{\omega}^n_{k+1} = \bar{\omega}^n_k + \xi_k (\bar{\omega}^n_k, \theta_k) - \omega^*_n(\bar{\omega}^n_k, \theta_k) + \xi_k (\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_k)) + \xi_k (\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_k)).$$

Rewriting this equation using the definition of $\bar{\omega}^*_n$ gives us:

$$\bar{\omega}^n_{k+1} = (I - \xi_k G(\theta))^{-1} G(\theta) \bar{\omega}^n_k + \xi_k G(\theta)^{-1} h(\theta) \frac{\max(1, \|r_n\|)}{\xi_k (\omega^*_{n+1} - \omega^*_k) + \xi_k (\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_k))).$$

Remember that $G(\theta)^{-1} h(\theta)$ is (uniformly) positive definite. Thus, for sufficiently large $n$, Lem. [C.3] ensures the existence of $\kappa > 0$ s.t.:

$$\|\bar{\omega}^n_{k+1}\| \leq (1 - \frac{1}{2}\kappa\xi_k) \|\bar{\omega}^n_k\| + \xi_k \|G(\theta)^{-1} h(\theta)\| \max(1, \|r_n\|) + \xi_k \|\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_k).$$

(145)

Since the sequences $(\omega^*_n)$, $(\bar{\omega}^*_n)$ and $h(\theta)$ are bounded and $\sup_{\theta \in \mathbb{R}^d} \|G(\theta)^{-1}\| < \infty$, there exists $C > 0$ s.t. for every $k \in \{k_n, \cdots, k^\xi_n\}$, $\|G(\theta)^{-1} h(\theta)\| \leq C$ and:

$$\|\omega^*_{n+1} - \omega_n^*\| = \beta_k \max(1, \|r_n\|) + \gamma \Phi^T D_{p, \theta} P_{\theta} \Phi \omega^*_n - G(\theta) \omega^*_n \leq \beta_k C.$$

Therefore, for sufficiently large $n$,

$$\|\bar{\omega}^n_{k+1}\| \leq C \|\omega^*_n\| + \frac{CC \xi}{\max(1, \|r_n\|)} + C \xi T \beta_k + \sum_{k = k_n}^{k^\xi_n} \xi_k \|\omega^*_n - \omega^*_n(\bar{\omega}^n_k, \theta_k).$$

It remains to show that $\sum_{k = k_n}^{k^\xi_n} \xi_k \|\omega^*_n - \omega^*_n(\bar{\omega}^n_k, \theta_k)\|$ converges to 0 as $n \to \infty$. For this purpose, we adopt the same strategy used for studying the sequence $(\omega^*_n)$. First, we write for every $k \geq k_n$,

$$\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_{k+1}) = (I - \beta_k G(\theta_k)) (\omega^*_{n+1} - \omega^*_n(\bar{\omega}^n_k, \theta_k)) - (\omega^*_n(\bar{\omega}^n_{k+1}, \theta_{k+1}) - \omega^*_n(\bar{\omega}^n_k, \theta_k)).$$

Then, applying Lem. [C.3] for sufficiently large $n$, there exists $\epsilon > 0$ s.t. for every $k \in \{k_n, \cdots, k^\xi_n\}$:

$$\|\omega^*_{n+1} - \omega^*_{n}(\bar{\omega}^n_{k+1}, \theta_{k+1})\| \leq (1 - \beta_k \frac{1}{2}) \|\omega^*_{n} - \omega^*_{n}(\bar{\omega}^n_k, \theta_k)\| + \|\omega^*_{n}(\bar{\omega}^n_{k+1}, \theta_{k+1}) - \omega^*_{n}(\bar{\omega}^n_k, \theta_k).$$

We can show that, for every $n$, the function $(\omega, \theta) \mapsto \omega^*_n(\bar{\omega}, \theta)$ is Lipschitz continuous (same arguments as the proof showing that the function $U$ is Lipschitz before Lem. [A.5]). It follows that there exists positive constants $C$ and $C^\prime$ s.t. for every $k \in \{k_n^\xi, \cdots, k_n^\xi_n\}$,

$$\|\omega^*_n(\bar{\omega}^n_{k+1}, \theta_{k+1}) - \omega^*_n(\bar{\omega}^n_k, \theta_k)\| \leq C \xi_k \|\omega^*_{n+1} - \omega^*_{n}\| + C \alpha_k \leq C^\prime \xi_k,$$

where the last inequality comes from the boundedness of the sequences $\omega^*_n$ and $\omega^*_n$ for $k \in \{k_n^\xi, \cdots, k_n^\xi_n\}$, and the fact that there exists $C > 0$ s.t. for every $k$, $\alpha_k \leq C \xi_k$. Therefore, noticing that there exists $C > 0$ s.t. $\|\omega^*_n - \omega^*_n(\bar{\omega}^n_k, \theta_k)\| \leq C$, it is easy to check that

$$\|\omega^*_{n+1} - \omega^*_{n}(\bar{\omega}^n_{k+1}, \theta_{k+1})\| \leq e^{-\frac{1}{4} \sum_{l = k_n^\xi}^{k_n^\xi_n} \beta_l} C + C^\prime \sum_{l = k_n^\xi}^{k_n^\xi_n} (e^{-\frac{1}{4} \sum_{p = l+1}^{l+1} \beta_p} \xi_l).$$

To conclude the proof, it is sufficient to show that:

$$\lim_{n \to \infty} \sum_{k = k_n^\xi}^{k_n^\xi_n} \xi_k \left( e^{-\frac{1}{4} \sum_{l = k_n^\xi}^{k_n^\xi_n} \beta_l} + \sum_{l = k_n^\xi}^{k_n^\xi_n} e^{-\frac{1}{4} \sum_{p = l+1}^{l+1} \beta_p} \xi_l \right) = 0.$$
Theorem C.10. We have the following:

(i) \( \sup_n \| \hat{\omega}_{kn} \| < \infty, \text{a.s.} \)
(ii) \( \sup_n \max_{k_n \leq k \leq k_{n+1}} \| \omega^n_k \| < \infty, \text{a.s.} \)
(iii) \( \sup_k \| \hat{\omega}_k \| < \infty, \text{a.s.} \)

Proof. (i) Combining Lem. C.9 with Lem. C.8 implies the existence of a sequence \((\hat{\delta}_n)\) converging to zero a.s. s.t. for sufficiently large \(n\),

\[
\| \hat{\omega}_{kn+1}^n \| \leq e^{-\frac{1}{2}T} \| \hat{\omega}_{kn}^n \| + \hat{\delta}_n + \frac{C}{\max(1, \| r_{kn} \|)}.
\]

Multiplying both sides by \(\max(1, \| r_{kn} \|)\) and using the fact that a.s.:

\[
\max(1, \| r_{kn} \|) \leq 1 + \| \omega_{kn} \| + \| \hat{\omega}_{kn} \| \leq (1 + C')(1 + \| \hat{\omega}_{kn} \|),
\]

where \(C' > 0\) in the last inequality is a constant stemming from Th. C.6 (iii) we obtain a.s.:

\[
\| \hat{\omega}_{kn+1}^n \| \leq (e^{-\frac{1}{2}T} + (1 + C')\hat{\delta}_n)\| \hat{\omega}_{kn}^n \| + (1 + C')\hat{\delta}_n + C.
\]

The result follows from Lem. C.2.

(ii) This result can be proven following similar arguments to the first item by exploiting Eq. (145) in the proof of Lem. C.9 and the results therein.

(iii) First, using the definition of \((\hat{\omega}_k^n)\), observe that:

\[
\sup_k \| \hat{\omega}_k \| = \sup_n \max_{k_n \leq k \leq k_{n+1}} \| \hat{\omega}_k \|
\]

\[
= \sup_n \max_{k_n \leq k \leq k_{n+1}} \{ \max(1, \| \omega_{kn} \|, \| \hat{\omega}_{kn} \|) \cdot \| \hat{\omega}_k^n \| \}.
\]

Then, using that \(\max(a,b) \leq a + b\) for any nonnegative reals \(a, b\), together with the triangular inequality, it follows from Eq. (146) that:

\[
\sup_k \| \hat{\omega}_k \| \leq \sup_n (1 + \| \omega_{kn} \| + \| \hat{\omega}_{kn} \|) \left( \max_{k_n \leq k \leq k_{n+1}} \| \hat{\omega}_k^n \| + \max_{k_n \leq k \leq k_{n+1}} \| \hat{\omega}_k^n - \omega_k^n \| \right).
\]

Given Th. C.6 (ii) there exists a constant \(\tilde{C} > 0\) s.t. a.s.:

\[
\sup_k \| \hat{\omega}_k \| \leq \sup_n \tilde{C}(1 + \| \omega_{kn} \|) \left( \max_{k_n \leq k \leq k_{n+1}} \| \hat{\omega}_k^n \| + \max_{k_n \leq k \leq k_{n+1}} \| \hat{\omega}_k^n - \omega_k^n \| \right).
\]

The result follows from the boundedness of the sequences \((\omega_{kn})\) (see the first item) and \((\hat{\omega}_k^n)\) (see the second item) and Lem. C.8.

\[\square\]

C.4 Technical lemmas

Lemma C.11. With \((x_k^n)\) and \((y_k^n)\) defined as in the proof of Lem. C.8 it holds that:

\[
\lim_{n \to \infty} \max_{k_n \leq k \leq k_{n+1}} \| x_k^n - x^*(y_k^n, \theta_k) \| = 0, \text{ a.s.},
\]

where we recall that for every \(y \in \mathbb{R}^m, \theta \in \mathbb{R}^d, x^*(y, \theta) = \gamma \tilde{G}(\theta)^{-1} \Phi^T D_{\rho, \theta} P_\theta \Phi y\) as previously defined in Eq. (142).

Proof. Recall that \(x_k^n = y_k^n = 0\). Throughout this proof, we will use the shorthand notation \(v_k^n := x_k^n - x^*(y_k^n, \theta_k)\).

Recall that \((x_k^n)\) and \((y_k^n)\) are bounded sequences in the sense of Lem. C.7 and so is the sequence \((v_k^n)\). Using Eq. (140), it is easy to check that the sequence \((v_k^n)\) satisfies for every \(k \geq k_n\) the recurrence relation:

\[
v_{k+1}^n = (I_m - \beta_k \tilde{G}(\theta_k)) v_k^n + (x^*(y_k^n, \theta_k) - x^*(y_{k+1}^n, \theta_{k+1})) + \beta_k \epsilon_k^n.
\]
Iterating this equality for $k \geq k_n^\xi$ and observing that $v_{k_n}^n = 0$ leads to the identity:

$$v_{k+1}^n = \sum_{p=k_n^\xi}^k \left[ \prod_{l=p+1}^k (I_m - \beta_l \hat{G}(\theta_l)) \right] ((x^*(y_p^n, \theta_p) - x^*(y_{p+1}^n, \theta_{p+1})) + \beta_p \epsilon_p^n).$$

It can be shown that the function $(\hat{\omega}, \theta) \mapsto x^*(\hat{\omega}, \theta)$ is $L$-Lipschitz continuous for some $L > 0$ (using the same arguments as for the proof showing that the function $U$ is Lipschitz before Lem. A.5). Furthermore, since $\hat{G}(\theta)$ is uniformly positive definite, applying Lem. C.3 yields the existence of $\epsilon > 0$ s.t. for sufficiently large $n$ and for $k \geq k_n^\xi$,

$$\|v_{k+1}^n\| \leq L \sum_{p=k_n^\xi}^k e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \| \epsilon_{i+1}^p \| (y_{p+1}^n, \theta_{p+1}) \| + \left\| \sum_{p=k_n^\xi}^k \left[ \prod_{l=p+1}^k (I_m - \beta_l \hat{G}(\theta_l)) \right] \beta_p \epsilon_p^n \right\| .$$

It can be easily checked that there exist $C > 0$ and $C' > 0$ s.t. for every $k \in \{k_n^\xi, \ldots, k_{n+1}^\xi - 1\}$, $\| (y_{k+1}^n, \theta_{k+1}) \| \leq C (\xi_k + \alpha_k) \leq C' \xi_k$. As a consequence, we obtain for every $k \in \{k_n^\xi, \ldots, k_{n+1}^\xi - 1\}$,

$$\|v_{k+1}^n\| \leq L C' \sum_{p=k_n^\xi}^k e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p + \sum_{p=k_n^\xi}^k \left\| \prod_{l=p+1}^k (I_m - \beta_l \hat{G}(\theta_l)) \right\| \beta_p \epsilon_p^n.$$ (149)

To prove the lemma, it is sufficient to show that both terms on the r.h.s. of the above inequality converge a.s. to 0. For this, recall first from the definition of the sequence $(k_n^\xi)$ that there exist $j_n, j_{n+1} \in \mathbb{N}$ s.t. $k_n^\xi = k_j^\xi$ and $k_{n+1}^\xi = k_{j_{n+1}}^\xi$. Observe also that for every $k \in \{k_n^\xi, \ldots, k_{n+1}^\xi - 1\}$, there exists $i_k \in \{j_n, \ldots, j_{n+1} - 1\}$ s.t. $k \in \{k_{i_k}^\beta, \ldots, k_{i_k+1}^\beta - 1\}$. Then, we can rewrite the first term in the above inequality as follows:

$$\sum_{p=k_n^\xi}^k e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p = \sum_{i=j_n}^{i_k-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p + \sum_{i_k}^{k_{i_k+1}^\beta-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p .$$

The second term on the r.h.s. of the above equation can be easily upperbounded by $\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} \epsilon_{i+1}^p \leq (T + \beta_{k_{i+1}}^\beta) \max_{p=k_i^\beta}^{k_{i+1}^\beta} \epsilon_{p}$. Now, for the first term of the above equation, notice that for $i \in \{j_n, \ldots, i_k-1\}$, $p \in \{k_i^\beta, \ldots, k_{i+1}^\beta-1\}$ and $k \in \{k_i^\beta, \ldots, k_{i+1}^\beta\}$, $\sum_{i=p+1}^k \beta_i \geq \sum_{i=k_i^\beta}^{k_{i+1}^\beta-1} \beta_i \geq T (i_k - i - 1)$ and this implies:

$$\sum_{i=j_n}^{i_k-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p \leq \sum_{i=j_n}^{i_k-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p \leq \frac{C T}{1 - e^{-\frac{1}{2} T k_i^\beta}} \max_{k_i^\beta \leq \epsilon_p} \epsilon_{p}. $$

We conclude from the above derivations that there exists $C > 0$ (independent of $n$) s.t. for every $k \in \{k_n^\xi, \ldots, k_{n+1}^\xi - 1\}$,

$$\sum_{p=k_n^\xi}^k e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p \leq C \max_{k_i^\beta \leq \epsilon_p} \epsilon_{p}. $$

Given Assumption 5.2, we deduce from this inequality that the first term on the r.h.s. of Eq. (149) converges to 0, i.e.,

$$\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \sum_{p=k_n^\xi}^k e^{-\frac{1}{2} \sum_{i=p+1}^k \beta_i} \epsilon_{i+1}^p = 0.$$

As for the second term on the r.h.s. of Eq. (149), we control it in the following lemma (Lem. C.12). □

Lemma C.12. $\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \left\| \sum_{p=k_n^\xi}^k \left[ \prod_{l=p+1}^k (I_m - \beta_l \hat{G}(\theta_l)) \right] \beta_p \epsilon_p^n \right\| = 0, a.s.$
Proof. Let \( \tilde{G}_{p+1:k} := \prod_{i=p}^{k} (I_m - \beta_{i} \tilde{G}(\theta_{i})) \) for every \( p \in \{k_n^\beta, \ldots, k_{n+1}^\beta\} \) and \( p \leq k - 1 \). As in the proof of Lem. C.11, we begin by observing that for every \( k \in \{k_n^\beta, \ldots, k_{n+1}^\beta - 1\} \), there exists \( k^\beta_{j_{n+1}} = k_{j_{n+1}}^\beta = k_{j_{n+1}+1}^\beta \) and that for every \( k \in \{k_{n}^\beta, \ldots, k_{n+1}^\beta - 1\} \), there exists \( k_{i_{n+1}} \in \{j_{n}, \ldots, j_{n+1} - 1\} \) s.t. \( k_{j_{n+1}}^\beta = k_{i_{n+1}}^\beta = k_{j_{n+1}+1}^\beta \). Then, we can write for every \( k \in \{k_{n}^\beta, \ldots, k_{n+1}^\beta - 1\} \),

\[
\left| \sum_{p=k_{i}^\beta}^{k} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| \leq \left| \sum_{j_{n}=p=k_{i}^\beta}^{k_{j_{n}}^\beta-1} \sum_{p=k_{i}^\beta}^{k_{j_{n}}^\beta} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| + \left| \sum_{p=k_{i}^\beta}^{k_{j_{n}}^\beta} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| .
\]

We will show that the first term on the r.h.s. of the above inequality converges to 0 a.s. when \( n \to \infty \). A slight change in the following proof will establish the convergence to zero of the second term. Notice that for \( k \in \{k_{n}^\beta, \ldots, k_{n+1}^\beta - 1\} \),

\[
\left| \sum_{j_{n}=p=k_{i}^\beta}^{k_{j_{n}}^\beta-1} \sum_{p=k_{i}^\beta}^{k_{j_{n}}^\beta} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| \leq \left| \sum_{j_{n}=p=k_{i}^\beta}^{k_{j_{n}}^\beta-1} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| \leq \left| \sum_{p=k_{i}^\beta}^{k_{j_{n}}^\beta} \tilde{G}_{p+1:k} \beta_{p} \hat{\epsilon}_p^n \right| .
\]

Eq. (150) and the proof is complete if we show that:

\[
\lim_{n \to \infty} \max_{k_{n}^\beta \leq k \leq k_{n+1}^\beta} \sum_{l=1}^{i_{n+1}-i_{n}} e^{-\frac{1}{2} \epsilon T(i_{n+1}-i)} \left| \sum_{p=k_{i}^\beta}^{k_{j_{n+1}}^\beta-1} \tilde{G}_{p+1:k_{i}^\beta} \beta_{p} \hat{\epsilon}_p^n \right| = 0, \ a.s.
\]
For this, it is sufficient to prove the following inequality:

\[
E \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \right\|^2 \right] \leq C \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p^2.
\]  

(157)

Indeed, the Chebyshev inequality implies that for every \( \delta > 0 \),

\[
P \left( \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \right\| \geq \delta \right) \leq \frac{C}{\delta^2} \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p^2,
\]

and applying the Borel-Cantelli lemma with the summability of the series \( \sum_k \beta_k^2 \) yields:

\[
\lim_{n \to \infty} \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \right\| = 0, \quad \text{a.s.}
\]

To prove that Ineq. (157) holds, it is sufficient to show that the desired inequality holds when \( \hat{c}_p^\beta \) is replaced by each one of the terms of its decomposition. For the first term which is a martingale difference with bounded second moment, we establish the sought-after inequality in Lem. C.13. The last three terms are of the order \( O(\beta_p) \), \( O(\xi_p) \) and \( O(\alpha_p) \), respectively. The remaining term is the summand of a telescopic series with bounded moment and we address its particular case in Lem. C.14 below.

Lemma C.13. There exists \( C > 0 \) s.t. for every \( n \in \mathbb{N} \),

\[
E \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \right\|^2 \right] \leq C \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p^2,
\]

where for every \( p \geq k_n \), \( Z_p^\beta := M_p^{(1)} \hat{\omega}_p + M_p^{(2)} \hat{\vartheta}_p + \hat{\vartheta}_p \hat{\omega}_p, Y_p ) - ( P_{\theta_p} \hat{\vartheta}_p ) ( \hat{\omega}_p, \hat{\omega}_p, Y_p ) \).

Proof. In this proof, we suppress the superscript \( n \) of \( Z_p^\beta \) to simplify notation. Note that \( n \) is fixed throughout the proof. Define \( M_p^{k_i^\beta} := \sum_{i=k_i^\beta}^{k_{i+1}^\beta} \beta_i Z_{i+1} \) for every \( i \in \mathbb{N}, k > k_i^\beta \). This is a zero mean, square integrable martingale for \( k \in \{ k_n^\beta + 1, \ldots, k_{n+1}^\beta \} \). By summation by part, we have for every \( j_n \leq i \leq j_{n+1} - 1 \),

\[
\sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta = M_p^{k_{i+1}^\beta} - \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 2} ( \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta ) M_p^{k_{i+1}^\beta}.
\]

Notice that \( \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \leq \beta_p \hat{G}_{p+1:k_{i+1}^\beta} \hat{\vartheta}_p \). Hence, bounding the max by the sum, we obtain the following inequality:

\[
\max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \hat{G}_{p+1:k_{i+1}^\beta} - \beta_p \hat{M}_p^\beta \right\|^2 \leq 2 \sum_{i=j_n}^{j_{n+1} - 1} \left\| M_{k_i^\beta}^{k_{i+1}^\beta} \right\|^2 + 2 \sum_{i=j_n}^{j_{n+1} - 1} \left\| \beta_p \hat{G}_{p+1:k_{i+1}^\beta} \right\|^2
\]

(158)

We have that \( \sup_{\theta \in \mathcal{B}_d} \| \hat{G}(\theta) \| < \infty \). Moreover, using Lem. C.3, one can show that there exists \( C > 0 \) s.t. for every integers \( q > p \), \( \| \hat{G}_{p,\theta} \| \leq C \). Thus, we obtain the following upper bound using the triangle inequality:

\[
\left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p \hat{G}_{p+1:k_{i+1}^\beta} \hat{\vartheta}_p \hat{\omega}_p \right\|^2 \leq 2 \left( \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p \| \hat{M}_p^\beta \| \right)^2 \leq C^2 T^2 \left( \max_{k_i^\beta \leq p \leq k_{i+1}^\beta - 1} \| \hat{M}_p^\beta \| \right)^2.
\]
Taking the expectation in Eq. (158) and using Doob’s inequality yields:

\[ \mathbb{E} \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \bar{G}_{p+1} k_{i+1}^\beta - 1 \beta_p Z_{p+1} \right\|^2 \right] \leq (2 + 8C^2 T^2) \sum_{i=j_n}^{j_{n+1} - 1} \mathbb{E} \left[ \left\| M_{k_i^\beta}^{k_{i+1}^\beta - 1} \right\|^2 \right] \]

\[ \leq (2 + 8C^2 T^2) C \sum_{i=j_n}^{j_{n+1} - 1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \beta_p^2 = (2 + 8C^2 T^2) C \sum_{p=k_n^\beta}^{k_{n+1}^\beta - 1} \beta_p^2 , \]

where the last inequality comes from the bounded second moment of $Z_{p+1}$.

**Lemma C.14.** Let $(X_k)$ be an $\mathbb{R}^m$-valued random sequence with bounded second moment. Then, there exists $C > 0$ s.t.:

\[ \mathbb{E} \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \bar{G}_{p+1} k_{i+1}^\beta - 1 \beta_p (X_{p+1} - X_p) \right\|^2 \right] \leq C \sum_{p=k_n^\beta}^{k_{n+1}^\beta - 1} \beta_p^2 . \]

**Proof.** Summation by parts yields for $j_n \leq i \leq j_{n+1} - 1$,

\[ \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \bar{G}_{p+1} k_{i+1}^\beta - 1 \beta_p (X_{p+1} - X_p) = \beta_{k_{i+1}^\beta - 1} X_{k_{i+1}^\beta - 1} - \beta_{k_i^\beta} \bar{G}_{k_i^\beta+1} k_{i+1}^\beta - 1 X_{k_i^\beta} + \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta - 1} (\beta_p \bar{G}_{p+1} k_{i+1}^\beta - 1 - \beta_{p-1} \bar{G}_{p} k_{i+1}^\beta - 1) X_p . \] (159)

Notice that $\beta_p \bar{G}_{p+1} k_{i+1}^\beta - 1 - \beta_{p-1} \bar{G}_{p} k_{i+1}^\beta - 1 = (\beta_p - \beta_{p-1}) \bar{G}_{p+1} k_{i+1}^\beta - 1 + \beta_{p-1} \bar{G}_{p} k_{i+1}^\beta - 1 - \bar{G}(\theta_p)$. Then, similarly to the proof of the previous lemma, recall that $\sup_{\theta \in \mathbb{R}^q} \| \bar{G}(\theta) \| < \infty$ and that Lem. C.3 entails the existence of a constant $C_G > 0$ s.t. for sufficiently large $p$ and for every integers $q > p$, $\max(||\bar{G}(\theta_p)\bar{G}_{p+1,q}||, ||\bar{G}_{p,q}||) \leq C_G$.

Using the previous remarks with Eq. (159) yields for $j_n \leq i \leq j_{n+1} - 1$,

\[ \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \bar{G}_{p+1} k_{i+1}^\beta - 1 \beta_p (X_{p+1} - X_p) \right\|^2 \leq 4 \beta_{k_{i+1}^\beta - 1}^2 \left\| X_{k_{i+1}^\beta - 1} \right\|^2 + 4 C^2 G^2 \beta_{k_i^\beta}^2 \left\| X_{k_i^\beta} \right\|^2 + 4 \left\| \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta - 1} (\beta_p - \beta_{p-1}) \bar{G}_{p+1} k_{i+1}^\beta - 1 X_p \right\|^2 + 4 \left\| \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta - 1} \beta_{p-1} \bar{G}_{p} k_{i+1}^\beta - 1 \bar{G}(\theta_p) X_p \right\|^2 . \] (160)

To prove the lemma, it is sufficient to show that the desired inequality holds when the l.h.s. is replaced by each of the terms on the r.h.s. of the above equation. Consider the first term:

\[ \mathbb{E} \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \beta_{k_{i+1}^\beta - 1}^2 \left\| X_{k_{i+1}^\beta} \right\|^2 \right] \leq \sum_{i=j_n}^{j_{n+1} - 1} \beta_{k_{i+1}^\beta - 1}^2 \mathbb{E} \left[ \left\| X_{k_i^\beta} \right\|^2 \right] \leq C \sum_{i=j_n}^{j_{n+1} - 1} \beta_{k_{i+1}^\beta - 1}^2 \leq C \sum_{p=k_n^\beta}^{k_{n+1}^\beta - 1} \beta_p^2 , \] (161)

where the constant $C_X > 0$ bounds the second moment of $X_k$ (i.e., $\sup_k \mathbb{E} \left\| X_k \right\|^2 \leq C_X$) and $C$ is also a positive constant independent of $p$ and $n$. The second term is treated analogously.
Let us consider now the third term. Using the triangle inequality combined with the boundedness of $\|\tilde{G}_{p,q}\|$ for $q > p$ yields for $n$ sufficiently large and $j_n \leq i \leq j_{n+1} - 1$,

$$\left\| \sum_{p=k_n^q+1}^{k_{n+1}^q} (\beta_p - \beta_p - 1) \tilde{G}_{p+1;k_{n+1}^q-1} X_p \right\|^2 \leq C_2^2 \left( \sum_{p=k_n^q+1}^{k_{n+1}^q} |\beta_{p-1} - \beta_p| \|X_p\| \right)^2.$$

Then, it follows that:

$$\max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_n^q+1}^{k_{n+1}^q} (\beta_p - \beta_p - 1) \tilde{G}_{p+1;k_{n+1}^q-1} X_p \right\|^2 \leq C_2^2 \sum_{i=j_n}^{j_{n+1} - 1} \left( \sum_{p=k_n^q+1}^{k_{n+1}^q} |\beta_{p-1} - \beta_p| \|X_p\| \right)^2.$$

We obtain the desired inequality by taking the expectation and using the boundedness of the second moment of the r.v. $X_k$:

$$\mathbb{E} \left[ \max_{j_n \leq i \leq j_{n+1} - 1} \left\| \sum_{p=k_n^q+1}^{k_{n+1}^q} (\beta_p - \beta_p - 1) \tilde{G}_{p+1;k_{n+1}^q-1} X_p \right\|^2 \right] \leq C' \sum_{i=j_n}^{j_{n+1} - 1} \left( \sum_{p=k_n^q+1}^{k_{n+1}^q} |\beta_{p-1} - \beta_p| \|X_p\| \right)^2 \leq C' \sum_{p=k_n^q+1}^{k_{n+1}^q} \beta_p^2,$$

where $C' := C_2^2 C_X$. It only remains to show that the desired inequality also holds for the fourth term in Ineq. 160. Using similar manipulations as above, we have:

$$\max_{j_n \leq i \leq j_{n+1}} \left\| \sum_{p=k_n^q+1}^{k_{n+1}^q} (\beta_p - \beta_p - 1) \tilde{G}_{p+1;k_{n+1}^q-1} (\theta_p) X_p \right\|^2 \leq C_2^2 \sum_{i=j_n}^{j_{n+1} - 1} \sum_{p=k_n^q+1}^{k_{n+1}^q} \sum_{q=k_n^q+1}^{k_{n+1}^q} \beta_{p-1} \beta_{p-1} \beta_q \beta_{q-1} \|X_p\| \|X_q\|,$$

and taking the expectation implies:

$$\mathbb{E} \left[ \max_{j_n \leq i \leq j_{n+1}} \left\| \sum_{p=k_n^q+1}^{k_{n+1}^q} (\beta_p - \beta_p - 1) \tilde{G}_{p+1;k_{n+1}^q-1} (\theta_p) X_p \right\|^2 \right] \leq C' \left( \sum_{p=k_n^q+1}^{k_{n+1}^q} \beta_p^2 \right)^2 \leq \tilde{C} \sum_{p=k_n^q+1}^{k_{n+1}^q} \beta_p^2,$$

where $\tilde{C} := C' \sum_{k=0}^{\infty} \beta_k^2$. Thus, the lemma holds for $C \geq 4C_X + 4C_2^2 + 4C' + 4\tilde{C}$. 

**Lemma C.15.** \( \lim_{n \to \infty} \sum_{k=k_n}^{k_{n+1}} \xi_k \left( e^{-\frac{1}{2} \epsilon \sum_{l=k_n}^{k} \beta_l} + e^{-\frac{1}{2} \epsilon \sum_{l=k_n}^{k} \beta_l \beta_{l-1}} \right) = 0 \).

**Proof.** We have already proved that \( \lim_{n \to \infty} \sum_{k=k_n}^{k_{n+1}} \xi_k e^{-\frac{1}{2} \epsilon \sum_{l=k_n}^{k} \beta_l \beta_{l-1}} = 0 \) (see the proof of Lem. C.11). The convergence of the second term in the lemma is proven in the same manner. \( \square \)