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

Anas Barakat, Pascal Bianchi, Julien Lehmann

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

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

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 stepsizes. 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 ϵ-approximate stationary point of the non-concave performance function using at most $\mathcal{O}(\epsilon^{-3}\ln^3(\frac{1}{\epsilon}))$ samples compared with $\mathcal{O}(\epsilon^{-2}\ln(\frac{1}{\epsilon}))$ 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, Algorithm 2] 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,

Dalal et al., 2018, Kaledin et al., 2020]. These works opened the way to the recent development of a flurry of nonasymptotic results for actor-critic algorithms [Yang et al., 2018, Qiu et al., 2019, Kumar et al., 2019, Hong et al., 2020, Xu et al., 2020b, Xu et al., 2020a, Wang et al., 2020, Wu et al., 2020, Shen et al., 2020]. Regarding online one-step actor-critic algorithms, Wu et al. [Wu et al., 2020] provided a finite-time analysis 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 (A3C) algorithm in the discounted setting. 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 [Lee and He, 2019, Zhang et al., 2021].

The summary table 1 compiles some key features of our work to situate it in the literature and highlights our contributions with respect to (w.r.t.) the closest related works. We also mention that alternative update rules are also possible for actor-critic algorithms. Other common variants in practice use different policy gradients estimates based directly on the critic estimate instead of using it for bootstrapping (see for e.g. a recent discussion in [Wen et al., 2021]). Such a modification of the actor would not impact our critic analysis but would induce a different bias for the policy gradient estimate (impacting namely Th. 5.4 and Th. 6.2 below). Our analysis can also be adapted to this setting with a suitable analysis of the induced bias.

3 PRELIMINARIES

Notation. For every finite set $\mathcal{X}$, we use the notation $\mathcal{P}(\mathcal{X})$ for the set of probability measures on $\mathcal{X}$. The cardinality of a finite set $\mathcal{Y}$ is denoted by $|\mathcal{Y}|$. For two sequences of nonnegative reals (x_n) and (y_n), the notation $x_n = \mathcal{O}(y_n)$ means that there exists a constant C independent of n such that $x_n \leq Cy_n$ 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 $\langle x, Bx \rangle$. 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, Szepesvári, 2010] 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 $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ be the state transition probability kernel and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the immediate reward function. A randomized stationary policy, which we will simply call a policy in the rest of the paper, is a mapping $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ specifying for each $s \in \mathcal{S}, a \in \mathcal{A}$ the 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 $\mathbb{P}_{\rho,\pi}$ the probability distribution of the Markov chain (S_t, A_t) issued from the MDP controlled by the policy π with initial state distribution ρ. The notation $\mathbb{E}_{\rho,\pi}$ refers to the associated expectation. We will use $\mathbb{E}_\pi$ whenever there is no dependence on ρ. 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 π, the long-term expected cumulative discounted reward is quantified by the state-value function $V_\pi : \mathcal{S} \to \mathbb{R}$ and the action-value function $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 $Q_\pi(s, a) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t R_{t+1}|S_0 = s, A_0 = a]$.We also define the advantage function $\Delta_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by $\Delta_\pi(s, a) := Q_\pi(s, a) - V_\pi(s)$. Given an initial probability distribution ρ over states for the initial state S_0, the goal of the agent is to find a policy π 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 π 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 θ. 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: $\max_{\theta \in \mathbb{R}^d} J(\theta)$.Whenever it exists, define for every $\theta \in \mathbb{R}^d$ the function $\psi_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

Table 1: Comparison to closest related works.

	Discounted reward	Actor critic	Markovian sampling[1]	Target variable	Asymptotic results	Finite-time analysis	Timescales
[Lillicrap et al., 2016]	✓	✓	✘	✓	✘	✘	1
[Lee and He, 2019]	✓	✘	✘	✓	✓	✓[2]	1
[Wu et al., 2020]	✘	✓	✓	✘	✘	✓	3
[Shen et al., 2020]	✓	✓	✓	✘	✘	✓	2
[Zhang et al., 2021]	✓	✘	✓	✓	✓	✘	2
This paper	✓	✓	✓	✓	✓	✓	3

[1] refers to the use of samples generated from the MDP and the acting policy, this excludes experience replay as in [Lillicrap et al., 2016] and identically independently distributed (i.i.d.) samples used in theoretical analysis.
[2] [Lee and He, 2019] provide a finite-time analysis for a target-based TD-learning algorithm (for policy evaluation) based on the periodic update style of the target variable used in [Mnih et al., 2015] involving two loops. They highlight that a finite-time analysis of the Polyak-averaging style update rule [Lillicrap et al., 2016] is an open question. Here, we address this question in the control setting.

by:

$$\psi_\theta(s,a) := \nabla \ln \pi_\theta(a|s)\,,$$

where ∇ denotes the gradient w.r.t. θ. 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 ψ_θ is well defined.

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

(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_π-Lipschitz continuous.
(c) The function $\theta \mapsto \psi_\theta(s,a)$ is bounded and L_ψ-Lipschitz.

Assumption 3.1 is satisfied for instance by the Gibbs (or softmax) policy and the Gaussian policy (see [Zhang et al., 2020a, Sec. 3] 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 θ given by:

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

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

$$\mu_{\rho,\theta}(s,a) := d_{\rho,\theta}(s)\, \pi_\theta(a|s) \tag{2}$$

$$\text{where} \quad d_{\rho,\theta}(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\rho,\pi_\theta}(S_t = s) \tag{3}$$

is a probability measure over the state space $\mathcal{S}$ known as the discounted state-occupancy measure. Note that under Assumption 3.1, the policy gradient ∇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 $(\tilde{S}_t,\tilde{A}_t)_{t\in\mathbb{N}}$ issued from the artificial MDP whose transition kernel $\tilde{p} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is defined for every $(s,a) \in \mathcal{S} \times \mathcal{A}$ by

$$\tilde{p}(\cdot|s,a) := \gamma\, p(\cdot|s,a) + (1-\gamma)\, \rho(\cdot)\,, \tag{4}$$

and which is controlled by the policy π_θ 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 $(\tilde{S}_t,\tilde{A}_t)_{t\in\mathbb{N}}$ 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 $p(\cdot|\tilde{S}_t,\tilde{A}_t)$ with probability γ and from ρ otherwise. For this purpose, at each time step t, we draw a Bernoulli random variable $B_t \in \{0,1\}$ with parameter γ which is independent of all the past random variables generated until time t.

Then, using the definition of the advantage function, Eq. (1) becomes:

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}[(R(\tilde{S},\tilde{A}) + \gamma V_{\pi_\theta}(S) - V_{\pi_\theta}(\tilde{S}))\, \psi_\theta(\tilde{S},\tilde{A})]\,, \tag{5}$$

where $(\tilde{S},\tilde{A}) \sim \mu_{\rho,\theta}$ 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

$$\delta_{t+1}^V = R_{t+1} + \gamma\, V(S_{t+1}) - V(\tilde{S}_t)\,, \tag{6}$$

where S_{t+1} is drawn from the distribution $p(\cdot|\tilde{S}_t,\tilde{A}_t)$ and $(\tilde{S}_t,\tilde{A}_t)_{t\in\mathbb{N}}$ is the Markov chain induced by the artificial MDP described in Eq. (4) and controlled by the policy π_θ. 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_{π_θ} is known, it stems from Eq. (5) that a natural estimator of the gradient $\nabla J(\theta)$ is $\delta_{t+1}^{V_{\pi_\theta}} \psi_\theta(\tilde{S}_t,\tilde{A}_t)/(1-\gamma)$. This estimator is only biased because the distribution of our sampled Markov chain $(\tilde{S}_t,\tilde{A}_t)_t$ is not exactly $\mu_{\rho,\theta}$ but converges geometrically to this one. However, the state-value function V_{π_θ} is unknown. Given an estimate $V_{\omega_t} \in \mathbb{R}^n$ of $V_{\pi_{\theta_t}}$ and a positive stepsize α_t, the actor updates its parameter as follows:

$$\theta_{t+1} = \theta_t + \alpha_t \frac{1}{1-\gamma} \delta_{t+1}^{V_{\omega_t}} \psi_{\theta_t}(\tilde{S}_t,\tilde{A}_t)\,. \tag{7}$$

4.2 Critic update

The state-value function V_{π_θ} is approximated for every state $s \in \mathcal{S}$ 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 $\omega = (\omega_1,\cdots,\omega_m)^T \in \mathbb{R}^m$ for some integer $m \ll n = |\mathcal{S}|$ and $\phi(s) = (\phi^1(s),\cdots,\phi^m(s))^T$ is the feature vector of the state $s \in \mathcal{S}$. We compactly represent the feature vectors as a matrix of features Φ of size $n \times m$ whose ith row corresponds to the row vector $\phi(s)^T$ for some $s \in \mathcal{S}$.

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_θ 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 $s, s' \in \mathcal{S}$ by $P_\theta(s'|s) := \sum_{a\in\mathcal{A}} p(s'|s,a)\pi_\theta(a|s)$ .Consider the vector $R_\theta = (R_\theta(s_1),\cdots R_\theta(s_n))$ whose ith coordinate is provided by $R_\theta(s_i) = \sum_{a\in\mathcal{A}} \pi_\theta(a|s_i) R(s_i,a)$. Let $D_{\rho,\theta}$ be the diagonal matrix with elements $d_{\rho,\theta}(s_i)$, $i = 1,\cdots n$ along its diagonal. Define also the Bellman operator $T_\theta : \mathbb{R}^n \mapsto \mathbb{R}^n$ for every $V \in \mathbb{R}^n$ by $T_\theta V := R_\theta + \gamma P_\theta V$. The true value function V_{π_θ} 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 $\mathcal{E}_\theta(\omega) := \frac{1}{2}\|T_\theta V_\omega - V_\omega\|^2_{D_{\rho,\theta}}$ where $V_\omega = \Phi\omega$. The gradient of the MSBE w.r.t. ω can be written as $\nabla_\omega \mathcal{E}_\theta(\omega) = \mathbb{E}_{\tilde{S}\sim d_{\rho,\theta}}[(T_\theta V_\omega(\tilde{S}) - V_\omega(\tilde{S}))(\mathbb{E}_{S\sim P_\theta(\cdot|\tilde{S})}[\gamma \nabla_\omega V_\omega(S)] - \nabla_\omega V_\omega(\tilde{S}))]$. As explained in [Bertsekas and Tsitsiklis, 1996, p. 369], omitting the gradient term $\nabla_\omega T_\theta V_\omega(\tilde{S}) = \mathbb{E}_{S\sim P_\theta(\cdot|\tilde{S})}[\gamma \nabla_\omega V_\omega(S)]$ in $\nabla_\omega \mathcal{E}_\theta(\omega)$ 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 [Barnard, 1993, Appendix 1] for a proof). The idea of target-based TD learning is to consider a modified version of the MSBE $\tilde{\mathcal{E}}_\theta(\omega,\bar{\omega}) := \frac{1}{2}\|T_\theta V_{\bar{\omega}} - V_\omega\|^2_{D_{\rho,\theta}}$. Observe that the term $T_\theta V_\omega$ depending on ω in the MSBE is now freezed in $\tilde{\mathcal{E}}_\theta(\omega,\bar{\omega})$ thanks to the target variable $\bar{\omega}$. We now need to introduce a new sequence $\bar{\omega}_t$ to define a sample-based version of $T_\theta V_{\bar{\omega}} - V_\omega$ which will be a modified version of the standard TD-error

$$\bar{\delta}_{t+1} = R_{t+1} + \gamma \phi(S_{t+1})^T \bar{\omega}_t - \phi(\tilde{S}_t)^T \omega_t\,. \tag{8}$$

Then, a stochastic gradient descent on $\tilde{\mathcal{E}}$ w.r.t. ω yields the critic update

$$\omega_{t+1} = \omega_t + \beta_t \bar{\delta}_{t+1} \phi(\tilde{S}_t)\,. \tag{9}$$

The target variable sequence $\bar{\omega}_t$ needs to be a slowed down version of the critic parameter ω_t. For this purpose, instead of using a periodical synchronization of the target variable $\bar{\omega}_t$ with ω_t through a copy as in DQN, we use the Polyak-averaging update rule proposed by [Lillicrap et al., 2016]

$$\bar{\omega}_{t+1} = \bar{\omega}_t + \xi_t(\omega_{t+1} - \bar{\omega}_t)\,, \tag{10}$$

where ξ_t is a positive stepsize chosen s.t. the sequence $(\bar{\omega}_t)$ evolves on a slower timescale than the sequence (ω_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 $\delta_{t+1} := \delta_{t+1}^{V_{\omega_t}}$ from now on.

Remark 1. We can simplify Algorithm 1 by using only the target-based TD error $\bar{\delta}_{t+1}$ instead of maintaining

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**
$\tilde{A}_t \sim \pi_{\theta_t}(\cdot|\tilde{S}_t)$; $S_{t+1} \sim p(\cdot|\tilde{S}_t, \tilde{A}_t)$
$\delta_{t+1} = R_{t+1} + \gamma\, \phi(S_{t+1})^T\omega_t - \phi(\tilde{S}_t)^T\omega_t$ ▷ classical TD error
$\bar{\delta}_{t+1} = R_{t+1} + \gamma\, \phi(S_{t+1})^T\bar{\omega}_t - \phi(\tilde{S}_t)^T\omega_t$ ▷ target-based TD error
$\theta_{t+1} = \theta_t + \alpha_t \frac{1}{1-\gamma}\delta_{t+1}\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)$ ▷ actor
$\omega_{t+1} = \omega_t + \beta_t\bar{\delta}_{t+1}\phi(\tilde{S}_t)$ ▷ critic
$\bar{\omega}_{t+1} = \bar{\omega}_t + \xi_t(\omega_{t+1} - \bar{\omega}_t)$ ▷ target variable
$S^\rho_{t+1} \sim \rho$; $B_{t+1} \sim \mathcal{B}(\gamma)$
$\tilde{S}_{t+1} = B_{t+1}S_{t+1} + (1 - B_{t+1})S^\rho_{t+1}$
end for
Output: Policy and value function parameters θ_T and ω_T.

both TD errors $\bar{\delta}_{t+1}$ and δ_{t+1}. The proofs can be easily adapted, note for this that $(\bar{\omega}_t)$ and (ω_t) track the same target $\bar{\omega}_*(\theta_t)$ (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 $\tilde{K}_\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be the transition matrix over the state-action pairs defined for every $(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$ by $\tilde{K}_\theta(s',a'|s,a) = \tilde{p}(s'|s,a)\pi_\theta(a'|s')$. Let $\mathcal{K} := \{\tilde{K}_\theta : \theta \in \mathbb{R}^d\}$ and let $\bar{\mathcal{K}}$ be its closure. Every element of $\bar{\mathcal{K}}$ 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 $K \in \bar{\mathcal{K}}$, the Markov chain induced by K is ergodic.

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

Algorithm 1 involves three different timescales. The actor parameter θ_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 ω_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.

Assumption 5.2 (stepsizes)**.** The sequences of positive stepsizes $(\alpha_t), (\beta_t)$ and (ξ_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 setting (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 (ω_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 (θ_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 $\epsilon\,\|\theta\|^2/2$ (ϵ 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, Shen 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 Φ 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 (ω_t) evolving on the fastest timescale, i.e., with the slowly vanishing stepsizes β_t (see Assumption 5.2). The main idea behind the proofs is that $\theta_t, \bar{\omega}_t$ 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 (ω_t) is associated to the ODE

$$\begin{cases} \dot{\omega}(s) &= \bar{h}(\theta(s), \bar{\omega}(s)) - \bar{G}(\theta(s))\,\omega(s)\,, \\ \dot{\theta}(s) &= 0\,, \\ \dot{\bar{\omega}}(s) &= 0\,, \end{cases} \qquad \text{(ODE-}\omega\text{)}$$

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

$$\bar{h}(\theta, \bar{\omega}) := \Phi^T D_{\rho,\theta}(R_\theta + \gamma P_\theta \Phi\, \bar{\omega}) \,\text{and}\, \bar{G}(\theta) := \Phi^T D_{\rho,\theta} \Phi\,. \tag{11}$$

Recall that the matrices $D_{\rho,\theta}, P_\theta$ and the vector R_θ 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-ω) that ω_t tracks a slowly moving target $\omega_*(\theta_t, \bar{\omega}_t)$ governed by the slower iterates θ_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 $\omega_*(\theta, \bar{\omega})$ for every $\theta \in \mathbb{R}^d, \bar{\omega} \in \mathbb{R}^m$ and $\lim_t \|\omega_t - \omega_*(\theta_t, \bar{\omega}_t)\| = 0$ *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 (θ_t) and slower than the sequence (ω_t). At the timescale ξ_t, everything happens as if the quantity ω_t in Eq. (10) could be replaced by $\omega_*(\theta_t, \bar{\omega}_t)$ 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

$$\begin{cases} \dot{\bar{\omega}}(s) &= \bar{G}(\theta(s))^{-1}(h(\theta(s)) - G(\theta(s))\bar{\omega}(s))\,, \\ \dot{\theta}(s) &= 0\,, \end{cases} \qquad \text{(ODE-}\bar{\omega}\text{)}$$

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

$$h(\theta) := \Phi^T D_{\rho,\theta} R_\theta \quad \text{and} \quad G(\theta) := \Phi^T D_{\rho,\theta}(I_n - \gamma P_\theta)\Phi\,. \tag{12}$$

We show in the appendix that the matrix $-G(\theta)$ 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 $\tilde{p}$ and the policy π_θ in lieu of the original transition kernel p. Then, we prove that $-\bar{G}(\theta)^{-1}G(\theta)$ is also stable, which suggests from (ODE-$\bar{\omega}$) that $\bar{\omega}_t$ tracks an other slowly moving target $\bar{\omega}_*(\theta_t)$. 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 $G(\theta)\bar{\omega} = h(\theta)$ has a unique solution $\bar{\omega}_*(\theta)$ and $\lim_t \|\bar{\omega}_t - \bar{\omega}_*(\theta_t)\| = 0$ *w.p.*1. Moreover, for every $\theta \in \mathbb{R}^d$, $\Phi\,\bar{\omega}_*(\theta)$ is a fixed point of the projected Bellman operator, i.e., $\Pi_\theta T_\theta(\Phi\,\bar{\omega}_*(\theta)) = \Phi\,\bar{\omega}_*(\theta)$, where $\Pi_\theta = \Phi(\Phi^T D_{\rho,\theta}\Phi)^{-1}\Phi^T D_{\rho,\theta}$ 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 ω_t tracks the same target $\bar{\omega}_*(\theta_t)$.

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

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

Moreover, this limit implies the following: $\lim_t \|\Pi_{\theta_t} T_{\theta_t}(\Phi\,\omega_t) - \Phi\,\omega_t\| = 0 \quad w.p.1\,.$

Remark 3. When the actor parameter θ_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

$$\liminf_t \left(\|\nabla J(\theta_t)\| - \|b(\theta_t)\|\right) \leq 0\,,$$

where for every $\theta \in \mathbb{R}^d, (s,a) \in \mathcal{S} \times \mathcal{A}$, $b(\theta) := \frac{1}{1-\gamma}\mathbb{E}_{\mu_{\rho,\theta}}[\psi_\theta(\tilde{S},\tilde{A})(\hat{Q}_\theta(\tilde{S},\tilde{A}) - Q_{\pi_\theta}(\tilde{S},\tilde{A}))]$ and $\hat{Q}_\theta(s,a) := R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)\phi(s')^T \bar{\omega}_*(\theta)$.

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 (θ_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 ($m = |\mathcal{S}|$ and the features spanning $\mathbb{R}^{|\mathcal{S}|}$) 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 $\mathcal{S}$ (and still finite action space) upon supposing that the feature map ϕ 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 π_θ 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 $t \in \mathbb{N}, \theta \in \mathbb{R}^d$,

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

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 $-\bar{G}(\theta)$ and $-\bar{G}(\theta)^{-1}G(\theta)$. While we only need these matrices to be stable for our asymptotic results, we actually show in the appendix that $-\bar{G}(\theta)$ is even negative definite uniformly in θ. We suppose that the second matrix $-\bar{G}(\theta)^{-1}G(\theta)$ is also negative definite uniformly in θ.

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

We are now ready to state our critic convergence rate.

Theorem 6.1. Let Assumptions 3.1, 5.1 and 5.3 to 6.2 hold. Let $c_1, c_2, c_3, \alpha, \xi, \beta$ be positive constants s.t. $0 < \beta < \xi < \alpha < 1$. Set $\alpha_t = \frac{c_1}{(1+t)^\alpha}$, $\xi_t = \frac{c_2}{(1+t)^\xi}$ and $\beta_t = \frac{c_3}{(1+t)^\beta}$. Then, the sequences (ω_t) and (θ_t) from Algorithm 1 satisfy for every integer $T \geq 1$,

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

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 $\mathcal{O}(T^{\beta-1})$ 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. 5.1) is crucial for the proof (see Sec. B.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 $\epsilon_{\text{FA}} \geq 0$ s.t. for every $\theta \in \mathbb{R}^d$, $\|V_{\pi_\theta} - \Phi\,\bar{\omega}_*(\theta)\|_{D_{\rho,\theta}} \leq \epsilon_{\text{FA}}$.

Observe that $\epsilon_{\text{FA}} = 0$ if the true value function V_{π_θ} belongs to the linear function space spanned by the feature functions for every $\theta \in \mathbb{R}^d$.

Theorem 6.2. Let Assumptions 3.1 , 5.1 , 5.3 to 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 $\alpha_t = \frac{c_1}{(1+t)^\alpha}$, $\xi_t = \frac{c_2}{(1+t)^\xi}$ and $\beta_t = \frac{c_3}{(1+t)^\beta}$. Then, for every integer $T \geq 1$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-\alpha}}\right) + \mathcal{O}\left(\frac{\ln^2 T}{T^\alpha}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2]\right) + \mathcal{O}\left(\epsilon_{\text{FA}}\right) .$$

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 $T \geq 1$,

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

Moreover, if we set $\alpha = \frac{2}{3}$, $\xi = \frac{1}{2}$ and $\beta = \frac{1}{3}$ to define the stepsizes (α_t), (ξ_t) and (β_t), the actor parameter sequence (θ_t) generated by Algorithm 1 within $T = \mathcal{O}(\epsilon^{-3}\ln^3(\frac{1}{\epsilon}))$ steps, satisfies

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

As a consequence, since Algorithm 1 uses a single sample from the MDP per iteration, its sample complexity is $\mathcal{O}(\epsilon^{-3}\ln^3(\frac{1}{\epsilon}))$. This is to compare with the best $\mathcal{O}(\epsilon^{-2}\ln(\frac{1}{\epsilon}))$ 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 [Qiu et al., 2019] (up to logarithmic factors) and better than the $\mathcal{O}(\epsilon^{-4})$ 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 $\mathcal{O}(T^{2(\alpha-\xi)})$ and $\mathcal{O}(T^{2(\xi-\beta)})$ 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 4 also applies to the function approximation error ϵ_{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

[Barnard, 1993] Barnard, E. (1993). Temporal-difference methods and markov models. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(2):357–365.

[Barto et al., 1983] Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive ele-

ments that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846.

[Benaim, 1996] Benaim, M. (1996). A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2):437–472.

[Benveniste et al., 1990] Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the French by Stephen S. Wilson.

[Bertsekas and Tsitsiklis, 1996] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, 1st edition.

[Bhandari et al., 2018] Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1691–1692. PMLR.

[Bhatnagar et al., 2009] Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.

[Borkar, 2008] Borkar, V. S. (2008). *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi. A dynamical systems viewpoint.

[Borkar and Meyn, 2000] Borkar, V. S. and Meyn, S. P. (2000). The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

[Dalal et al., 2018] Dalal, G., Thoppe, G., Szörényi, B., and Mannor, S. (2018). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1199–1233. PMLR.

[Fujimoto et al., 2018] Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

[Gupta et al., 2019] Gupta, H., Srikant, R., and Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Haarnoja et al., 2018] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR.

[Heess et al., 2015] Heess, N., Hunt, J. J., Lillicrap, T. P., and Silver, D. (2015). Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455*.

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

[Horn and Johnson, 1994] Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

[Kaledin et al., 2020] Kaledin, M., Moulines, E., Naumov, A., Tadic, V., and Wai, H. (2020). Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Abernethy, J. D. and Agarwal, S., editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR.

[Karmakar and Bhatnagar, 2018] Karmakar, P. and Bhatnagar, S. (2018). Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Math. Oper. Res.*, 43(1):130–151.

[Konda, 2002] Konda, V. R. (2002). *Actor-Critic Algorithms*. PhD thesis, USA. AAI0804543.

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

[Konda and Tsitsiklis, 2003a] Konda, V. R. and Tsitsiklis, J. N. (2003a). Linear stochastic approximation

driven by slowly varying markov chains. *Systems & Control Letters*, 50(2):95–102.

[Konda and Tsitsiklis, 2003b] Konda, V. R. and Tsitsiklis, J. N. (2003b). On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166.

[Kumar et al., 2019] Kumar, H., Koppel, A., and Ribeiro, A. (2019). On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*.

[Kushner and Yin, 2003] Kushner, H. J. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition. Stochastic Modelling and Applied Probability.

[Lakshminarayanan and Bhatnagar, 2017] Lakshminarayanan, C. and Bhatnagar, S. (2017). A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 79:108–114.

[Lee and He, 2019] Lee, D. and He, N. (2019). Target-based temporal-difference learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3713–3722. PMLR.

[Lillicrap et al., 2016] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *ICLR 2016*.

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

[Mnih et al., 2016] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA. PMLR.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[Peters and Schaal, 2008] Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190.

[Puterman, 2014] Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[Qiu et al., 2019] Qiu, S., Yang, Z., Ye, J., and Wang, Z. (2019). On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*.

[Shen et al., 2020] Shen, H., Zhang, K., Hong, M., and Chen, T. (2020). Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. *arXiv preprint arXiv:2012.15511*.

[Srikant and Ying, 2019] Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and td learning. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2803–2830, Phoenix, USA. PMLR.

[Sutton, 1988] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

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

[Sutton et al., 2009] Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 993–1000, New York, NY, USA. Association for Computing Machinery.

[Sutton et al., 2000] Sutton, R. S., Mcallester, D., Singh, S., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, volume 99, pages 1057–1063. MIT Press.

[Szepesvári, 2010] Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103.

[Tsitsiklis and Van Roy, 1997] Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation.

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

[Wang et al., 2020] Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2020). Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*.

[Wen et al., 2021] Wen, J., Kumar, S., Gummadi, R., and Schuurmans, D. (2021). Characterizing the gap between actor-critic and policy gradient. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11101–11111. PMLR.

[Wu et al., 2020] Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17617–17628. Curran Associates, Inc.

[Xu et al., 2020a] Xu, T., Wang, Z., and Liang, Y. (2020a). Improving sample complexity bounds for (natural) actor-critic algorithms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4358–4369. Curran Associates, Inc.

[Xu et al., 2020b] Xu, T., Wang, Z., and Liang, Y. (2020b). Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*.

[Xu et al., 2019] Xu, T., Zou, S., and Liang, Y. (2019). Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems*, pages 10634–10644.

[Yang et al., 2019] Yang, Z., Fu, Z., Zhang, K., and Wang, Z. (2019). Convergent reinforcement learning with function approximation: A bilevel optimization perspective.

[Yang et al., 2018] Yang, Z., Zhang, K., Hong, M., and Başar, T. (2018). A finite sample analysis of the actor-critic algorithm. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2759–2764.

[Zhang et al., 2020a] Zhang, K., Koppel, A., Zhu, H., and Başar, T. (2020a). Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control Optim.*, 58(6):3586–3612.

[Zhang et al., 2020b] Zhang, S., Liu, B., Yao, H., and Whiteson, S. (2020b). Provably convergent two-timescale off-policy actor-critic with function approximation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11204–11213, Virtual. PMLR.

[Zhang et al., 2021] Zhang, S., Yao, H., and Whiteson, S. (2021). Breaking the deadly triad with a target network. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12621–12631. PMLR.

[Zou et al., 2019] Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for sarsa with linear function approximation. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 (ω_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 (ω_t) evolving on the fastest timescale, i.e., with the stepsizes β_t which are converging the slowest to zero (see Assumption 5.2). We rewrite the slower sequences $(\theta_t), (\bar{\omega}_t)$ with the stepsizes β_t. In this timescale, $(\theta_t), (\bar{\omega}_t)$ are quasi-static from the point of view of the evolution of the sequence (ω_t). We deduce from this first step that ω_t tracks a slowly moving target $\omega_*(\theta_t, \bar{\omega}_t)$ governed by the slower iterates θ_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 $(\bar{\omega}_t)$ which is evolving in a faster timescale than the sequence (θ_t) and slower than the sequence (ω_t). Similarly, we show that $\bar{\omega}_t$ tracks an other slowly moving target $\bar{\omega}_*(\theta_t)$. 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 ω_t tracks the same target $\bar{\omega}_*(\theta_t)$.

A.1.1 Proof of Prop. 5.1

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

$$\begin{aligned}\omega_{t+1} &= \omega_t + \beta_t \bar{\delta}_{t+1} \phi(\tilde{S}_t) \\ &= \omega_t + \beta_t (R_{t+1} + \gamma \phi(S_{t+1})^T \bar{\omega}_t - \phi(\tilde{S}_t)^T \omega_t) \phi(\tilde{S}_t) \\ &= \omega_t + \beta_t w(\bar{\omega}_t, \omega_t, Z_t) + \beta_t \eta_{t+1}^{(1)}\,, \end{aligned} \tag{13}$$

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

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

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

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

As can be seen in Eq. (13), the sequence (ω_t) can be written as a linear stochastic approximation scheme controlled by the slowly varying Markov chains (θ_t) and $(\bar{\omega}_t)$. 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 $Z = (\tilde{S}, \tilde{A})$ 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 $\bar{G} : \mathbb{R}^d \to \mathbb{R}^{m \times m}$ from Eq. (11), for every $\theta \in \mathbb{R}^d, \bar{\omega} \in \mathbb{R}^m$

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

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

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

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

$$\begin{aligned}
\mathbb{E}_{Z \sim \mu_{\rho,\theta}}[w(\bar{\omega}, \omega, Z)] &= \mathbb{E}_{Z \sim \mu_{\rho,\theta}} \left[\left(R(\tilde{S}, \tilde{A}) + \gamma \sum_{s' \in \mathcal{S}} p(s'|\tilde{S}, \tilde{A}) \phi(s')^T \bar{\omega} \right) \phi(\tilde{S}) - \phi(\tilde{S}) \phi(\tilde{S})^T \omega \right] \\
&= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu_{\rho,\theta}(s,a) \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \phi(s')^T \bar{\omega} \right) \phi(s) - \phi(s)\phi(s)^T \omega \\
&= \sum_{s \in \mathcal{S}} d_{\rho,\theta}(s) \left(R_\theta(s)\phi(s) + \gamma \sum_{s' \in \mathcal{S}} p_\theta(s'|s) \phi(s')^T \bar{\omega} \phi(s) - \phi(s)\phi(s)^T \omega \right) \\
&= \bar{h}(\theta, \bar{\omega}) - \bar{G}(\theta)\omega\,,
\end{aligned}$$

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 (θ_t) and $(\bar{\omega}_t)$ that

$$\chi_{t+1} = \chi_t + \beta_t \varepsilon_t\,, \tag{16}$$

where $\varepsilon_t = \left(\frac{\alpha_t}{\beta_t} \frac{1}{1-\gamma} \delta_{t+1} \psi_{\theta_t}(Z_t), \frac{\xi_t}{\beta_t}(\omega_{t+1} - \bar{\omega}_t) \right)$. 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 , (ω_t) and (hence) $(\bar{\omega}_t)$ are a.s. bounded by Assumption 5.3 , (R_t) is bounded by U_R, $\theta \mapsto \psi_\theta(s,a)$ is bounded by Assumption 3.1 and $\mathcal{S}, \mathcal{A}$ are finite.

Let $\zeta_t = (\chi_t, \omega_t)$, $\zeta = (\theta, \bar{\omega}, \omega) \in \mathbb{R}^{d+2m}$, $W(\zeta, z) = (0, w(\bar{\omega}, \omega, z))$, $\varepsilon'_t = (\varepsilon_t, 0)$ and $\tilde{\eta}^{(1)}_{t+1} = (0, \eta^{(1)}_{t+1})$. 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:

$$\zeta_{t+1} = \zeta_t + \beta_t [W(\zeta_t, Z_t) + \varepsilon'_t + \tilde{\eta}^{(1)}_{t+1}]\,, \tag{17}$$

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 (ζ_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

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

i.e.,

$$\begin{cases} \frac{d}{ds}\chi(s) &= 0\,, \\ \frac{d}{ds}\omega(s) &= \bar{h}(\chi(s)) - \bar{G}(\theta(s))\omega(s)\,. \end{cases} \tag{18}$$

As we will show that the second ODE governing ω has a unique asymptotically stable equilibrium $\omega_*(\theta, \bar{\omega})$ for every constant function $\chi(t) = \chi = (\theta, \bar{\omega})$, it follows that (χ_t, ω_t) converges a.s. towards the set $\{(\chi, \omega_*(\chi)) : \chi \in \mathbb{R}^{d+m}\}$. In other words, $\lim_t \|\omega_t - \omega_*(\theta_t, \bar{\omega}_t)\| = 0$, 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 (θ_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 $\mathcal{S}$ is finite.

(iii) (A3): $(\tilde{\eta}_{t+1}^{(1)})$ 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. $\mathbb{E}[\|\tilde{\eta}_{n+1}^{(1)}\|^2 | \mathcal{F}_t] \leq K(1 + \|\omega_t\|^2 + \|\bar{\omega}_t\|^2)$.

(iv) (A4): The stepsizes (β_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 $\bar{G}(\theta)$ exists thanks to Assumptions 5.1 and 5.4. For all $\chi = (\theta, \bar{\omega}) \in \mathbb{R}^{d+m}$, we now show that the ODE $\frac{d}{ds}\omega(s) = \bar{h}(\chi) - \bar{G}(\theta)\omega(s)$ has a unique globally asymptotically stable equilibrium $\omega_*(\chi) = \bar{G}(\theta)^{-1}\bar{h}(\chi)$. The aforementioned ODE is stable if and only if the matrix $\bar{G}(\theta)$ is Hurwitz. We actually show that we have a stronger result in Lem. A.2 under Assumptions 5.1 and 5.4. We briefly explicit why the assumption as formulated in the rest of (A6') holds.

Define the function $L(\chi, \omega) = \frac{1}{2}\|\bar{G}(\theta)\omega - \bar{h}(\chi)\|^2$. For every $\chi = (\theta, \bar{\omega}) \in \mathbb{R}^{d+m}$, the function $L(\chi, \cdot)$ is a Lyapunov function for ODE (18). Indeed, using Lem. A.2 below, we can write

$$\frac{d}{ds}L(\chi, \omega(s)) = -\langle \bar{h}(\chi) - \bar{G}(\theta)\omega(s), \bar{G}(\theta)(\bar{h}(\chi) - \bar{G}(\theta)\omega(s))\rangle \leq -\varepsilon\|\bar{G}(\theta)\omega(s) - \bar{h}(\chi)\|^2 .$$

(vii) (A7): The stability Assumption 5.3 ensures that $\sup_t(\|\omega_t\| + \|\theta_t\|) < +\infty$ w.p.1. As a consequence, it also follows from the update rule of $(\bar{\omega}_t)$ that $\sup_t \|\bar{\omega}_t\| < +\infty$.

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 $\mathcal{K} := \{\tilde{K}_\theta : \theta \in \mathbb{R}^d\}$ where for every $\theta \in \mathbb{R}^d$, $\tilde{K}_\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is the transition matrix over the state-action pairs defined for every $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$ by $\tilde{K}_\theta(s', a'|s, a) = \tilde{p}(s'|s, a)\pi_\theta(a'|s')$. We also denoted by $\bar{\mathcal{K}}$ 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 $K \in \bar{\mathcal{K}}$.

We first show that the map $K \mapsto \mu_K$ is continuous over the set $\bar{\mathcal{K}}$. 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 μ_K satisfies:

$$M(K)\mu_K = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{where} \quad M(K) := \begin{bmatrix} K^T - I \\ \mathbb{1} \end{bmatrix} .$$

As a consequence, since $M(K)$ has full column rank thanks to Assumption 5.1, the matrix $M(K)^T M(K)$ is invertible and we obtain a closed form expression for μ_K given by:

$$\mu_K = (M(K)^T M(K))^{-1} M(K)^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{\mathrm{com}(M(K)^T M(K))^T}{\det(M(K)^T M(K))} M(K)^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

where $\mathrm{com}(A)$ 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 $M(K)^T M(K)$, and the determinant operator is continuous.

It follows from Assumption 5.1 that for every $K \in \bar{\mathcal{K}}$ 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 $\bar{\mathcal{K}}$ that $\inf_{K \in \bar{\mathcal{K}}} \mu_K(s, a) > 0$. Since $\tilde{K}_\theta \in \bar{K}$ for

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 $\tilde{K}_\theta$. As a consequence, since $d_{\rho,\theta}(s) = \sum_{a\in\mathcal{A}} \mu_{\rho,\theta}(s,a)$, it also holds that

$$\inf_\theta d_{\rho,\theta}(s) > 0\,.$$

Therefore, for every $\theta \in \mathbb{R}^d$, $\omega \in \mathbb{R}^m$:

$$\omega^T \bar{G}(\theta)\omega = (\Phi\,\omega)^T D_{\rho,\theta}(\Phi\,\omega) \geq \min_{s\in\mathcal{S}} \inf_\theta d_{\rho,\theta}(s)\|\Phi\omega\|^2 \geq \min_{s\in\mathcal{S}} \inf_\theta d_{\rho,\theta}(s)\lambda_{\min}(\Phi^T\Phi)\|\omega\|^2\,,$$

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 $\varepsilon := \lambda_{\min}(\Phi^T\Phi)\cdot \min_{s\in\mathcal{S}} \inf_\theta d_{\rho,\theta}(s) > 0$ which is independent of θ. □

A.1.2 Proof of Prop. 5.2

Recall the definitions of the vector $h(\theta)$ and the matrix $G(\theta)$ from Eq. (12):

$$h(\theta) := \Phi^T D_{\rho,\theta} R_\theta \quad \text{and} \quad G(\theta) := \Phi^T D_{\rho,\theta}(I_n - \gamma P_\theta)\Phi\,. \tag{19}$$

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 $\theta \in \mathbb{R}^d$, $\omega \in \mathbb{R}^m$,

$$\omega^T G(\theta)\omega = (\Phi\omega)^T D_{\rho,\theta}(I_n - \gamma P_\theta)\Phi\omega = (\Phi\omega)^T D_{\rho,\theta}(\Phi\omega) - \gamma(\Phi\omega)^T D_{\rho,\theta} P_\theta(\Phi\omega)\,. \tag{20}$$

Then, the Cauchy-Schwarz inequality yields

$$(\Phi\omega)^T D_{\rho,\theta} P_\theta(\Phi\omega) = (\Phi\omega)^T D_{\rho,\theta}^{\frac{1}{2}} D_{\rho,\theta}^{\frac{1}{2}} P_\theta(\Phi\omega) \leq \|\Phi\omega\|_{D_{\rho,\theta}} \|P_\theta\Phi\omega\|_{D_{\rho,\theta}}\,. \tag{21}$$

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_θ but it is instead associated to the artificial kernel $\tilde{P}_\theta$. 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

$$\|P_\theta V\|^2_{D_{\rho,\theta}} \leq \frac{1}{\gamma}\|V\|^2_{D_{\rho,\theta}} - \frac{1-\gamma}{\gamma}\|V\|^2_\rho \leq \frac{1}{\gamma}\|V\|^2_{D_{\rho,\theta}}\,.$$

Proof. It follows from Jensen's inequality that

$$\|P_\theta V\|^2_{D_{\rho,\theta}} = \sum_{i=1}^n d_{\rho,\theta}(s_i)\Big(\sum_{j=1}^n P_\theta(s_j|s_i)V_j\Big)^2 \leq \sum_{i=1}^n d_{\rho,\theta}(s_i)\sum_{j=1}^n P_\theta(s_j|s_i)V_j^2.$$

Then, observe that $\tilde{P}_\theta = \gamma P_\theta + (1-\gamma)\mathbb{1}\rho^T$ as a consequence of Eq. (4). By plugging this formula and then using the fact that $d^T_{\rho,\theta}\tilde{P}_\theta = d^T_{\rho,\theta}$, we obtain

$$\begin{aligned}
\sum_{i=1}^n d_{\rho,\theta}(s_i)\sum_{j=1}^n P_\theta(s_j|s_i)V_j^2 &= \frac{1}{\gamma}\Bigg[\Bigg(\sum_{j=1}^n\sum_{i=1}^n d_{\rho,\theta}(s_i)\tilde{P}_\theta(s_j|s_i)V_j^2\Bigg) - (1-\gamma)\sum_{j=1}^n \rho(s_j)V_j^2\Bigg] \\
&= \frac{1}{\gamma}\Bigg[\sum_{j=1}^n d_{\rho,\theta}(s_j)V_j^2 - (1-\gamma)\sum_{j=1}^n \rho(s_j)V_j^2\Bigg] \\
&= \frac{1}{\gamma}\|V\|^2_{D_{\rho,\theta}} - \frac{1-\gamma}{\gamma}\|V\|^2_\rho\,,
\end{aligned}$$

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

$$(\Phi\omega)^T D_{\rho,\theta} P_\theta (\Phi\omega) \leq \frac{1}{\sqrt{\gamma}} \|\Phi\omega\|^2_{D_{\rho,\theta}} = \frac{1}{\sqrt{\gamma}} (\Phi\omega)^T D_{\rho,\theta} (\Phi\omega) \,.$$

Whence, we obtain from Eq. (20) that

$$\omega^T G(\theta)\omega \geq (1-\sqrt{\gamma})(\Phi\omega)^T D_{\rho,\theta}(\Phi\omega) \geq \varepsilon(1-\sqrt{\gamma})\|\omega\|^2 \,,$$

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 $\bar{\omega}_*(\theta_t)$. From the update rules of the sequences $(\bar{\omega}_t)$ and (ω_t) (Eqs. (9)-(10)), we can introduce the quantity $\omega_*(\theta_t, \bar{\omega}_t)$ as defined in Prop. 5.1 to obtain

$$\begin{aligned}
\bar{\omega}_{t+1} &= \bar{\omega}_t + \xi_t(\omega_{t+1} - \bar{\omega}_t) \\
&= \bar{\omega}_t + \xi_t(\omega_t + \beta_t w(\bar{\omega}_t, \omega_t, Z_t) + \beta_t \eta^{(1)}_{t+1} - \bar{\omega}_t) \\
&= \bar{\omega}_t + \xi_t(\omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_t) + \xi_t(\omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \beta_t w(\bar{\omega}_t, \omega_t, Z_t)) + \xi_t\beta_t\eta^{(1)}_{t+1} \,.
\end{aligned} \tag{22}$$

Then, using the expressions of $\bar{h}, \bar{G}$ in Eq. (11) and h, G in Eq. (12), we can write

$$\omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_t = \bar{G}(\theta_t)^{-1}(\bar{h}(\theta_t, \bar{\omega}_t) - \bar{G}(\theta_t)\bar{\omega}_t) = \bar{G}(\theta_t)^{-1}(h(\theta_t) - G(\theta_t)\bar{\omega}_t) \,.$$

As a consequence,

$$\bar{\omega}_{t+1} = \bar{\omega}_t + \xi_t \bar{G}(\theta_t)^{-1}(h(\theta_t) - G(\theta_t)\bar{\omega}_t) + \xi_t(\omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \beta_t w(\bar{\omega}_t, \omega_t, Z_t)) + \xi_t\beta_t\eta^{(1)}_{t+1} \,. \tag{23}$$

Therefore, the sequence $(\bar{\omega}_t)$ satisfies a linear stochastic approximation scheme driven by the slowly varying Markov chain (θ_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,

$$\chi_{t+1} = \chi_t + \xi_t[U(\chi_t) + \tilde{\varepsilon}_t] \,, \tag{24}$$

where $\tilde{\varepsilon}_t = (\frac{\alpha_t}{\xi_t}\frac{1}{1-\gamma}\delta_{t+1}\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t), \omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \beta_t w(\bar{\omega}_t, \omega_t, Z_t) + \beta_t\eta^{(1)}_{t+1})$.

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 5.2 , $\omega_t - \omega_*(\theta_t, \bar{\omega}_t) \to 0$ as proved in Prop. 5.1 and $\delta_{t+1}\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t), w(\bar{\omega}_t, \omega_t, Z_t)$ are bounded by Assumptions 3.1-*(c)* , 5.3, the boundedness of the reward function R and the fact that the sets $\mathcal{S}, \mathcal{A}$ are finite. Moreover, Assumption 5.2 ensures that $\sum_t \xi_t = +\infty$ and $\sum_t \xi_t^2 < +\infty$.

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 $\theta \mapsto R_\theta$ and $\theta \mapsto P_\theta$ are Lipschitz continuous as $P_\theta(s'|s) = p(s'|s,a)\pi_\theta(a|s)$, $R_\theta(s) = \sum_{a\in\mathcal{A}} R(s,a)\pi_\theta(a|s)$ and Assumption 3.1-*(b)* guarantees that $\theta \mapsto \pi_\theta(a|s)$ is Lipschitz continuous for every $(s,a) \in \mathcal{S}\times\mathcal{A}$.

(c) The function $\theta \mapsto D_{\rho,\theta}$ 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 $D_{\rho,\theta}$ and P_θ are bounded by one.

Using classical stochastic approximation results (see, for e.g., [Benaim, 1996, Th.1.2]), we obtain that the sequence (χ_t) converges a.s. towards an internally chain transitive set of the ODE $\frac{d}{ds}\chi(s) = U(\chi(s))$, i.e.,

$$\begin{cases} \frac{d}{ds}\theta(s) &= 0\,, \\ \frac{d}{ds}\bar{\omega}(s) &= \bar{G}(\theta(s))^{-1}(h(\theta(s)) - G(\theta(s))\bar{\omega}(s))\,. \end{cases} \tag{25}$$

We conclude by showing that for every $\theta \in \mathbb{R}^d$, the ODE $\frac{d}{ds}\bar{\omega}(s) = \bar{G}(\theta)^{-1}(h(\theta) - G(\theta)\bar{\omega}(s))$ has a globally asymptotically stable equilibrium $\bar{\omega}_*(\theta)$. 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 $\{(\theta, \bar{\omega}_*(\theta)) : \theta \in \mathbb{R}^d\}$. 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 $M = \bar{G}(\theta)$ which is symmetric by definition and positive definite thanks to Lem. A.2. Then, we obtain that

$$A^*M + MA = -G(\theta)^T\bar{G}(\theta)^{-1}\bar{G}(\theta) - \bar{G}(\theta)\bar{G}(\theta)^{-1}G(\theta) = -(G(\theta)^T + G(\theta))\,.$$

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 $\omega^T(G(\theta)^T + G(\theta))\omega = 2\omega^T G(\theta)\omega > 0$ 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:

$$\begin{aligned} \Pi_\theta T_\theta(\Phi\bar{\omega}_*(\theta)) &= \Phi\bar{G}(\theta)^{-1}\Phi^T D_{\rho,\theta}T_\theta(\Phi\bar{\omega}_*(\theta)) \\ &= \Phi\bar{G}(\theta)^{-1}\Phi^T D_{\rho,\theta}(R_\theta + \gamma P_\theta\Phi\bar{\omega}_*(\theta)) \\ &= \Phi\bar{G}(\theta)^{-1}h(\theta) + \Phi\bar{G}(\theta)^{-1}(\bar{G}(\theta) - G(\theta))G(\theta)^{-1}h(\theta) \\ &= \Phi\bar{G}(\theta)^{-1}h(\theta) + \Phi G(\theta)^{-1}h(\theta) - \Phi\bar{G}(\theta)^{-1}h(\theta) \\ &= \Phi G(\theta)^{-1}h(\theta) \\ &= \Phi\bar{\omega}_*(\theta)\,, \end{aligned} \tag{26}$$

where the first equality uses the expression of the projection Π_θ, the second one uses the definition of the Bellman operator T_θ 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 $\lim_t \|\omega_t - \bar{\omega}_*(\theta_t)\| = 0$ *w.p.*1 ,we prove the two following results:

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

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

(a) We have the decomposition

$$\begin{aligned} \omega_t - \omega_*(\theta_t, \bar{\omega}_*(\theta_t)) &= [\omega_t - \omega_*(\theta_t, \bar{\omega}_t)] + [\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_t, \bar{\omega}_*(\theta_t))], \\ &= [\omega_t - \omega_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}(\bar{h}(\theta_t, \bar{\omega}_t) - \bar{h}(\theta_t, \bar{\omega}_*(\theta_t))) \\ &= [\omega_t - \omega_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}\Phi^T D_{\rho,\theta_t}P_{\theta_t}\Phi(\bar{\omega}_t - \bar{\omega}_*(\theta_t)) \\ &= [\omega_t - \omega_*(\theta_t, \bar{\omega}_t)] + \bar{G}(\theta_t)^{-1}(\bar{G}(\theta_t) - G(\theta_t))(\bar{\omega}_t - \bar{\omega}_*(\theta_t)) \\ &= [\omega_t - \omega_*(\theta_t, \bar{\omega}_t)] + (I_m - \bar{G}(\theta_t)^{-1}G(\theta_t))(\bar{\omega}_t - \bar{\omega}_*(\theta_t))\,. \end{aligned} \tag{27}$$

It follows from Prop. 5.1 that the first term in the above decomposition goes to zero. Then, observe that $\sup_\theta \|\bar{G}(\theta)^{-1}\| < \infty$ given Lem. A.2 and $\sup_\theta \|G(\theta)\| < \infty$ thanks to the boundedness of the matrices P_θ and $D_{\rho,\theta}$ 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_* and $\bar{\omega}_*$, we can write for every $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\omega_*(\theta, \bar{\omega}_*(\theta)) &= \bar{G}(\theta)^{-1}\bar{h}(\theta, \bar{\omega}_*(\theta)) \\
&= \bar{G}(\theta)^{-1}\Phi^T D_{\rho,\theta}(R_\theta + \gamma P_\theta \Phi G(\theta)^{-1}h(\theta)) \\
&= \bar{G}(\theta)^{-1}(h(\theta) + \gamma \Phi^T D_{\rho,\theta} P_\theta \Phi G(\theta)^{-1}h(\theta)) \\
&= \bar{G}(\theta)^{-1}(I_n + \gamma \Phi^T D_{\rho,\theta} P_\theta \Phi G(\theta)^{-1})h(\theta) \\
&= \bar{G}(\theta)^{-1}(G(\theta) + \gamma \Phi^T D_{\rho,\theta} P_\theta \Phi)G(\theta)^{-1}h(\theta) \\
&= \bar{G}(\theta)^{-1}\bar{G}(\theta)G(\theta)^{-1}h(\theta) \\
&= \bar{\omega}_*(\theta)\,.
\end{aligned}$$

For the last result, we write

$$\begin{aligned}
\|\Pi_{\theta_t} T_{\theta_t}(\Phi\omega_t) - \Phi\omega_t\| &= \|\Phi\left(\bar{G}(\theta_t)^{-1}\Phi^T D_{\rho,\theta_t} T_{\theta_t}(\Phi\omega_t) - \omega_t\right)\| \\
&= \|\Phi\left(\bar{G}(\theta_t)^{-1}\Phi^T D_{\rho,\theta_t}(T_{\theta_t}(\Phi\omega_t) - \Phi\omega_t)\right)\| \\
&= \|\Phi\left(\bar{G}(\theta_t)^{-1}(h(\theta_t) - G(\theta_t)\omega_t)\right)\| \\
&= \|\Phi\bar{G}(\theta_t)^{-1}G(\theta_t)(\omega_t - \bar{\omega}_*(\theta_t))\| \\
&\leq \|\Phi\|\|\bar{G}(\theta_t)^{-1}\|\|G(\theta_t)\|\|\omega_t - \bar{\omega}_*(\theta_t)\|\,. \qquad (28)
\end{aligned}$$

Then, as previously mentioned in the proof, observe that $\sup_\theta \|\bar{G}(\theta)^{-1}\| < \infty$ and $\sup_\theta \|G(\theta)\| < \infty$. Since $\bar{\omega}_t - \bar{\omega}_*(\theta_t) \to 0$ as $t \to \infty$, 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, 2003b, Sec. 6]. Recall the notation $Z_t = (\tilde{S}_t, \tilde{A}_t)$. Note that (Z_t) is a Markov chain. The actor parameter θ_t iterates as follows:

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha_t \frac{1}{1-\gamma}\delta_{t+1}\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \\
&= \theta_t + \alpha_t \frac{1}{1-\gamma}(R_{t+1} + (\gamma\phi(S_{t+1}) - \phi(\tilde{S}_t))^T\omega_t)\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \\
&= \theta_t + \alpha_t \frac{1}{1-\gamma}(R(\tilde{S}_t, \tilde{A}_t)\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) + H_{\theta_t}(Z_t)\omega_t) + \alpha_t \frac{1}{1-\gamma}\tilde{\eta}_{t+1}\,,
\end{aligned}$$

where for every $\theta \in \mathbb{R}^d$, $z = (s, a) \in \mathcal{S} \times \mathcal{A}$,

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

and $(\tilde{\eta}_{t+1})$ is an $\mathbb{R}^d$-valued $\mathcal{F}_t$-martingale difference sequence defined by

$$\tilde{\eta}_{t+1} = (R_{t+1} - \mathbb{E}[R_{t+1}|\mathcal{F}_t])\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) \; + \gamma\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)(\phi(S_{t+1}) - \mathbb{E}[\phi(S_{t+1})|\mathcal{F}_t])^T\omega_t\,. \qquad (29)$$

We now introduce the steady-state expectation of the main term $H_\theta(Z_t)\omega_t + R(\tilde{S}_t, \tilde{A}_t)\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)$. Recall that $\mu_{\rho,\theta}$ is the stationary distribution of the Markov chain (Z_t). Define the functions $\bar{H} : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ and $u : \mathbb{R}^d \to \mathbb{R}^d$ for every $\theta \in \mathbb{R}^d$ by

$$\bar{H}(\theta) = \mathbb{E}_{Z\sim\mu_{\rho,\theta}}[H_\theta(Z)]\,, \qquad (30)$$

$$u(\theta) = \mathbb{E}_{Z\sim\mu_{\rho,\theta}}[R(\tilde{S}, \tilde{A})\psi_\theta(\tilde{S}, \tilde{A})]\,, \qquad (31)$$

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

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

$$\theta_{t+1} = \theta_t + \alpha_t f(\theta_t) + \alpha_t \frac{1}{1-\gamma}(\tilde{\eta}_{t+1} + e_t^{(1)} + e_t^{(2)}), \tag{32}$$

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

$$f(\theta) = \frac{1}{1-\gamma}(\bar{H}(\theta)\,\bar{\omega}_*(\theta) + u(\theta)), \tag{33}$$

$$e_t^{(1)} = (R(\tilde{S}_t, \tilde{A}_t)\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) + H_{\theta_t}(Z_t)\bar{\omega}_*(\theta_t)) - (\bar{H}(\theta_t)\,\bar{\omega}_*(\theta_t) + u(\theta_t)), \tag{34}$$

$$e_t^{(2)} = H_{\theta_t}(Z_t)(\omega_t - \bar{\omega}_*(\theta_t)). \tag{35}$$

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

$$b(\theta) := f(\theta) - \nabla J(\theta). \tag{36}$$

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_{π_θ} will be seen as a vector of $\mathbb{R}^{|\mathcal{S}|}$.

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

$$b(\theta) = \frac{\gamma}{1-\gamma} \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu_{\rho,\theta}(s,a)\psi_\theta(s,a) \sum_{s'\in\mathcal{S}} p(s'|s,a)(\phi(s')^T \bar{\omega}_*(\theta) - V_{\pi_\theta}(s')).$$

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

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

$$\begin{aligned}
\bar{H}(\theta) = \mathbb{E}_{Z\sim\mu_{\rho,\theta}}[H_\theta(Z)] &= \mathbb{E}_{Z\sim\mu_{\rho,\theta}}\left[\psi_\theta(\tilde{S},\tilde{A})\left(\gamma \sum_{s'\in\mathcal{S}} p(s'|\tilde{S},\tilde{A})\phi(s') - \phi(\tilde{S})\right)^T\right] \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu_{\rho,\theta}(s,a)\psi_\theta(s,a)\left(\gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\phi(s')^T - \phi(s)^T\right) \\
&= \gamma \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu_{\rho,\theta}(s,a)\psi_\theta(s,a) \sum_{s'\in\mathcal{S}} p(s'|s,a)\phi(s')^T,
\end{aligned} \tag{37}$$

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

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

$$\begin{aligned}
(1-\gamma)\nabla J(\theta) &= \mathbb{E}_{Z\sim\mu_{\rho,\theta}}[\Delta_{\pi_\theta}(\tilde{S},\tilde{A})\psi_\theta(\tilde{S},\tilde{A})] \\
&= \mathbb{E}_{Z\sim\mu_{\rho,\theta}}[(R(\tilde{S},\tilde{A}) + \gamma \sum_{s'\in\mathcal{S}} p(s'|\tilde{S}_t,\tilde{A}_t)V_{\pi_\theta}(s') - V_{\pi_\theta}(\tilde{S}))\psi_\theta(\tilde{S},\tilde{A})] \\
&= \sum_{s,a} \mu_{\rho,\theta}(s,a)(R(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V_{\pi_\theta}(s') - V_{\pi_\theta}(s))\psi_\theta(s,a) \\
&= u(\theta) + \gamma \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu_{\rho,\theta}(s,a)\psi_\theta(s,a) \sum_{s'\in\mathcal{S}} p(s'|s,a)V_{\pi_\theta}(s').
\end{aligned} \tag{38}$$

The result stems from using the definition of $b(\theta)$ together with Eqs. (37) and (38). □

Using a second-order Taylor expansion of the $\tilde{L}$-Lipschitz function ∇J (again see [Zhang et al., 2020a, Lem. 4.2]) together with Eq. (32), we can derive the following inequalities

$$\begin{aligned} J(\theta_{t+1}) &\geq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle - L\|\theta_{t+1} - \theta_t\|^2 \,, \\ &\geq J(\theta_t) + \alpha_t \langle \nabla J(\theta_t), f(\theta_t) \rangle \\ &\quad + \frac{\alpha_t}{1-\gamma} \langle \nabla J(\theta_t), \tilde{\eta}_{t+1} + e_t^{(1)} + e_t^{(2)} \rangle - \tilde{L} \frac{\alpha_t^2}{(1-\gamma)^2} \|\delta_{t+1} \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\|^2 \,. \end{aligned} \tag{39}$$

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) $\sum_{t=0}^{\infty} \alpha_t^2 \|\delta_{t+1} \psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\|^2 < \infty$ 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 $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$. First, define the functions $g_\theta^* : \mathcal{Z} \to \mathbb{R}^d$ and $\bar{g} : \mathbb{R}^d \to \mathbb{R}^d$ by:

$$g_\theta^*(z) := R(z)\psi_\theta(z) + H_\theta(z)\bar{\omega}_*(\theta) \,, \tag{40}$$

$$\bar{g}(\theta) := u(\theta) + \bar{H}(\theta)\bar{\omega}_*(\theta) \,, \tag{41}$$

for every $z = (s,a) \in \mathcal{Z}, \theta \in \mathbb{R}^d$. Observe in particular that $e_t^{(1)} = g_{\theta_t}^*(\tilde{S}_t, \tilde{A}_t) - \bar{g}(\theta_t)$. Recall that for every $\theta \in \mathbb{R}^d$, the kernel transition $\tilde{K}_\theta$ is defined for every $(s,a),(s',a') \in \mathcal{S} \times \mathcal{A}$ by $\tilde{K}_\theta(s',a') = \tilde{p}(s'|s,a)\pi_\theta(a'|s')$ (see Assumption 5.1). The idea of the proof is to introduce for each integer $i = 1, \cdots, d$ a Markov Reward Process (MRP) [Puterman, 2014, Sec. 8.2] on the space $\mathcal{Z}$ induced by the transition kernel $\tilde{K}_\theta$ and the reward function $g_{\theta,i}^*$ (ith coordinate of the function g_θ^*). As a consequence, the corresponding average reward is given by $\bar{g}_i(\theta)$ (ith coordinate of $\bar{g}(\theta)$). Then, the differential value function of the MRP is provided by $v_{\theta,i} := (I - \tilde{K}_\theta + \mathbb{1}\mu_{\rho,\theta}^T)^{-1}(I - \mathbb{1}\mu_{\rho,\theta}^T) g_{\theta,i}^*$ as shown for instance in [Puterman, 2014, Sec. 8.2]. The functions $v_{\theta,i}$ for $i = 1, \cdots, d$ define together a vector valued function $v_\theta : \mathcal{Z} \to \mathbb{R}^d$. Under Assumption 5.1, using similar arguments to the proof of Lem. A.2 (see also [Zhang et al., 2021, Proof of Lem. 4, p. 26]), we can show that the function $K \in \bar{\mathcal{K}} \mapsto (I - K + \mathbb{1}\mu_K^T)^{-1}(I - \mathbb{1}\mu_K^T)$ is continuous on the compact set $\bar{\mathcal{K}}$. It follows that $\sup_{\theta,z} \|v_\theta(z)\| < \infty$ because $\tilde{K}_\theta \in \bar{\mathcal{K}}$ for every $\theta \in \mathbb{R}^d$ and $g_{\theta,i}^*$ is uniformly bounded w.r.t. θ under our assumptions . Moreover, the differential value function satisfies the crucial Bellman equation:

$$v_\theta(z) = g_\theta^*(z) - \bar{g}(\theta) + \sum_{z' \in \mathcal{Z}} \tilde{K}_\theta(z'|z) v_\theta(z) \,,$$

for every $z \in \mathcal{Z}$. We use the above Poisson equation to express $e_t^{(1)} = g_{\theta_t}^*(\tilde{S}_t, \tilde{A}_t) - \bar{g}(\theta_t)$ using v_θ. 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 $\theta \to \psi_\theta(s,a)$ guaranteed by Assumption 3.1-*(c)* with the boundedness of the rewards sequence (R_t), the sequence (ω_t) (Assumption 5.3) and the gradient ∇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 $\bar{H}(\theta_t)$ 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 $\omega_t - \bar{\omega}_*(\theta_t) \to 0$ as shown in Th. 5.3 .

(d) Similarly to $\bar{H}(\theta_t)$, upon noticing that the reward sequence (R_t) is bounded by U_R and the sequence (ω_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 α_t ($\sum_t \alpha_t^2 < \infty$) 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

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

where υ_t is defined by

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

It stems from Lem. A.7 that $\upsilon_t \to 0$ 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 $|\mathcal{A}|$ 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 ψ_θ.

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 $\bar{\omega}$ 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:

$$\begin{aligned}
\omega_t - \bar{\omega}_*(\theta_t) &= \omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_*(\theta_t) \\
&= \omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_t, \bar{\omega}_*(\theta_t)) \\
&= \omega_t - \omega_*(\theta_t, \bar{\omega}_t) + \bar{G}(\theta_t)^{-1}(\bar{h}(\theta_t, \bar{\omega}_t) - \bar{h}(\theta_t, \bar{\omega}_*(\theta_t)))\,. \qquad (42)
\end{aligned}$$

Hence, it is sufficient to obtain a control of the convergence rates of the quantities $\omega_t - \omega_*(\theta_t, \bar{\omega}_t)$ and $\bar{\omega}_t - \bar{\omega}_*(\theta_t)$. 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 $\tilde{x} = (\tilde{s}, \tilde{a}, s) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and every $\bar{\omega}, \omega \in \mathbb{R}^m$:

$$\bar{\delta}(\tilde{x}, \bar{\omega}, \omega) = R(\tilde{s}, \tilde{a}) + \gamma \phi(s)^T \bar{\omega} - \phi(\tilde{s})^T \omega \,, \tag{43}$$

$$g(\tilde{x}, \bar{\omega}, \omega) = \bar{\delta}(\tilde{x}, \bar{\omega}, \omega) \phi(\tilde{s}) \,. \tag{44}$$

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

$$\bar{g}(\theta, \bar{\omega}, \omega) = \mathbb{E}_{\tilde{s} \sim d_{\rho,\theta}, \tilde{a} \sim \pi_\theta, s \sim p(\cdot|\tilde{s},\tilde{a})}[g(\tilde{x}, \bar{\omega}, \omega)] = \bar{h}(\theta, \bar{\omega}) - \bar{G}(\theta)\omega \,. \tag{45}$$

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

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

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

$$\begin{aligned} \|\nu_{t+1}\|^2 \leq \|\nu_t\|^2 + 2\beta_t \langle \nu_t, g(\tilde{x}_t, \bar{\omega}_t, \omega_t)\rangle + 2\langle \nu_t, \omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\rangle \\ + 2\|\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\|^2 + 2C\beta_t^2 \,. \end{aligned} \tag{46}$$

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

$$\langle \nu_t, g(\tilde{x}_t, \bar{\omega}_t, \omega_t)\rangle = \langle \nu_t, \bar{g}(\theta_t, \bar{\omega}_t, \omega_t)\rangle + \Lambda(\theta_t, \bar{\omega}_t, \omega_t, \tilde{x}_t) \,, \tag{47}$$

where we used the shorthand notation

$$\Lambda(\theta, \bar{\omega}, \omega, \tilde{x}) := \langle \omega - \omega_*(\theta, \bar{\omega}), g(\tilde{x}, \bar{\omega}, \omega) - \bar{g}(\theta, \bar{\omega}, \omega)\rangle \,. \tag{48}$$

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

$$\langle \nu_t, \bar{g}(\theta_t, \bar{\omega}_t, \omega_t)\rangle = \langle \nu_t, \bar{g}(\theta_t, \bar{\omega}_t, \omega_t) - \bar{g}(\theta_t, \bar{\omega}_t, \omega_*(\theta_t, \bar{\omega}_t))\rangle = -\langle \nu_t, \bar{G}(\theta_t)\nu_t\rangle \leq -\varepsilon \|\nu_t\|^2 \,. \tag{49}$$

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

$$\|\omega_*(\theta_t, \bar{\omega}_t) - \omega_*(\theta_{t+1}, \bar{\omega}_{t+1})\| \leq C(\|\theta_t - \theta_{t+1}\| + \|\bar{\omega}_t - \bar{\omega}_{t+1}\|) \leq C\left(\frac{\alpha_t}{1-\gamma} + \xi_t\right) \,. \tag{50}$$

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

$$\|\nu_{t+1}\|^2 \leq (1 - 2\varepsilon\beta_t)\|\nu_t\|^2 + 2\beta_t \Lambda(\theta_t, \bar{\omega}_t, \omega_t, \tilde{x}_t) + C\left(\frac{\alpha_t}{1-\gamma} + \xi_t\right)\|\nu_t\| + C\left(\frac{\alpha_t^2}{(1-\gamma)^2} + \xi_t^2 + \beta_t^2\right) \,. \tag{51}$$

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

$$\tau_T := \min\{t \in \mathbb{N}, t \geq 1 : c\sigma^{t-1} \leq \min\{\alpha_T, \xi_T, \beta_T\}\} \,. \tag{52}$$

In the remainder of the proof, we will use the notation τ for τ_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

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

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

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

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

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

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

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 ω_* and $\bar{g}$ are Lipschitz in all their arguments, g is Lipschitz in its two last arguments and ω_t, ω_*, g and $\bar{g}$ are all bounded, one can show after tedious decompositions that

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

Then, recalling that the sequence (α_t) is nonincreasing, remark that

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

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

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

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

$$\begin{aligned}\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \check{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] &= \mathbb{E}[\langle \nu_{t-\tau}, g(\tilde{x}_t, z_{t-\tau}) - g(\check{x}_t, z_{t-\tau})\rangle|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \\ &\leq C d_{TV}(\mathbb{P}(\tilde{x}_t \in \cdot|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\check{x}_t \in \cdot|\tilde{S}_{t-\tau+1}, \theta_{t-\tau})) \\ &\leq \frac{C}{2}|\mathcal{A}|L_\pi \sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\||\tilde{S}_{t-\tau+1}, \theta_{t-\tau}]\,, \quad (56)\end{aligned}$$

where the first equality stems from the definition of Λ, 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

$$\begin{aligned}\sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\||\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] &\leq \sum_{i=t-\tau}^{t} \sum_{j=t-\tau}^{i-1} \mathbb{E}[\|\theta_{j+1} - \theta_j\||\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \\ &\leq \frac{C}{1-\gamma} \sum_{i=t-\tau}^{t} \sum_{j=t-\tau}^{i-1} \alpha_j \leq \frac{C}{1-\gamma} \alpha_{t-\tau} \sum_{i=0}^{\tau} i \leq \frac{C}{1-\gamma} \alpha_{t-\tau} (\tau+1)^2\,.\end{aligned}$$

As a consequence of these derivations, Eq. (56) yields

$$\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \tilde{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \check{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \leq \frac{C}{1-\gamma}|\mathcal{A}|\alpha_{t-\tau}(\tau+1)^2\,, \tag{57}$$

(c) **Control of $\Lambda(\theta_{t-\tau}, z_{t-\tau}, \check{x}_t)$:** Define $\bar{x}_t := (\bar{S}_t, \bar{A}_t, S_{t+1})$ where $\bar{S}_t \sim d_{\rho,\theta_{t-\tau}}$, $\bar{A}_t \sim \pi_{\theta_{t-\tau}}$ and $S_{t+1} \sim p(\cdot|\bar{S}_t, \bar{A}_t)$. Observing that $\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \bar{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] = 0$, we obtain

$$\begin{aligned}
\mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \check{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] &= \mathbb{E}[\Lambda(\theta_{t-\tau}, z_{t-\tau}, \check{x}_t) - \Lambda(\theta_{t-\tau}, z_{t-\tau}, \bar{x}_t)|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \\
&= \mathbb{E}[\langle \nu_{t-\tau}, g(\check{x}_t, z_{t-\tau}) - g(\bar{x}_t, z_{t-\tau})\rangle|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}] \\
&\leq C d_{TV}(\mathbb{P}(\check{x}_t \in \cdot|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{x}_t \in \cdot|\tilde{S}_{t-\tau+1}, \theta_{t-\tau})) \\
&= C d_{TV}(\mathbb{P}(\tilde{S}_t \in \cdot|\tilde{S}_{t-\tau+1}, \theta_{t-\tau}), d_{\rho,\theta_{t-\tau}}) \\
&\leq C\sigma^{\tau-1} \\
&\leq C\alpha_T\,,
\end{aligned} \tag{58}$$

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

$$\mathbb{E}[\Lambda(\theta_t, z_t, \tilde{x}_t)] \leq C\left(\tau\left(\frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau}\right) + |\mathcal{A}|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau+1)^2 + \alpha_T\right)\,. \tag{59}$$

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

$$\begin{aligned}
\mathbb{E}[\|\nu_{t+1}\|^2] \leq (1-2\varepsilon\beta_t)\mathbb{E}[\|\nu_t\|^2] &+ 2C\beta_t\left(\tau\left(\frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau}\right) + |\mathcal{A}|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau+1)^2 + \alpha_T\right) \\
&+ C\left(\frac{\alpha_t}{1-\gamma} + \xi_t\right)\mathbb{E}[\|\nu_t\|] + C\left(\frac{\alpha_t^2}{(1-\gamma)^2} + \xi_t^2 + \beta_t^2\right)\,.
\end{aligned} \tag{60}$$

Rearranging the inequality and summing for t between τ_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

$$I_1(T) := \sum_{t=\tau_T}^{T}\frac{1}{\beta_t}(\mathbb{E}[\|\nu_t\|^2] - \mathbb{E}[\|\nu_{t+1}\|^2])\,, \tag{62}$$

$$I_2(T) := \sum_{t=\tau_T}^{T} 2C\left(\tau\left(\frac{\alpha_{t-\tau}}{1-\gamma} + \beta_{t-\tau} + \xi_{t-\tau}\right) + |\mathcal{A}|\frac{\alpha_{t-\tau}}{1-\gamma}(\tau+1)^2 + \alpha_T\right) \tag{63}$$

$$I_3(T) := C\sum_{t=\tau_T}^{T}\left(\frac{\alpha_t}{(1-\gamma)\beta_t} + \frac{\xi_t}{\beta_t}\right)\mathbb{E}[\|\nu_t\|] \tag{64}$$

$$I_4(T) := C\sum_{t=\tau_T}^{T}\frac{\alpha_t^2}{(1-\gamma)^2\beta_t} + \frac{\xi_t^2}{\beta_t} + \beta_t\,. \tag{65}$$

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

(1) Since (ν_t) is a bounded sequence,

$$I_1(T) = \sum_{t=\tau_T}^{T} \left(\frac{1}{\beta_t} - \frac{1}{\beta_{t-1}}\right) \mathbb{E}[\|\nu_t\|^2] + \frac{1}{\beta_{\tau_T-1}} \mathbb{E}[\|\nu_{\tau_T}\|^2] - \frac{1}{\beta_{\tau_T}} \mathbb{E}[\|\nu_{T+1}\|^2]$$
$$\leq C \left[\sum_{t=\tau_T}^{T} \left(\frac{1}{\beta_t} - \frac{1}{\beta_{t-1}}\right) + \frac{1}{\beta_{\tau_T-1}}\right] = \frac{C}{\beta_T} = \mathcal{O}(T^{\beta}) \,. \quad (66)$$

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

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

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

$$I_2(T) \leq C \left(\tau_T \sum_{t=0}^{T-\tau} \left(\frac{\alpha_t}{1-\gamma} + \beta_t + \xi_t\right) + |\mathcal{A}| \frac{(\tau+1)^2}{1-\gamma} \sum_{t=0}^{T-\tau} \alpha_t + (1+T-\tau)\alpha_T\right)$$
$$\leq \frac{C}{1-\gamma} (\tau(1+T)^{1-\beta} + (\tau+1)^2 |\mathcal{A}|(1+T)^{1-\alpha})$$
$$= \mathcal{O}\left(\frac{\ln T}{1-\gamma} T^{1-\beta}\right) + \mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma} \ln^2(T) T^{1-\alpha}\right) = \mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma} \ln(T) T^{1-\beta}\right) , \quad (67)$$

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

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

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

$$I_3(T) = \sum_{t=\tau_T}^{T} C \left(\frac{\alpha_t}{(1-\gamma)\beta_t} + \frac{\xi_t}{\beta_t}\right) \mathbb{E}[\|\nu_t\|]$$
$$\leq C \sqrt{\sum_{t=\tau_T}^{T} \left(\frac{\alpha_t}{(1-\gamma)\beta_t} + \frac{\xi_t}{\beta_t}\right)^2} \sqrt{\sum_{t=\tau_T}^{T} \mathbb{E}[\|\nu_t\|^2]} \,. \quad (68)$$

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

$$\frac{1}{1+T-\tau_T} \sum_{t=\tau_T}^{T} \left(\frac{\alpha_t}{(1-\gamma)\beta_t} + \frac{\xi_t}{\beta_t}\right)^2 \leq \frac{2}{1+T-\tau_T} \sum_{t=\tau_T}^{T} \left(\left(\frac{\alpha_t}{(1-\gamma)\beta_t}\right)^2 + \left(\frac{\xi_t}{\beta_t}\right)^2\right)$$
$$= \frac{2}{1+T-\tau_T} \sum_{t=0}^{T-\tau_T} \left(\left(\frac{\alpha_{t+\tau_T}}{(1-\gamma)\beta_{t+\tau_T}}\right)^2 + \left(\frac{\xi_{t+\tau_T}}{\beta_{t+\tau_T}}\right)^2\right)$$
$$\leq \frac{2}{T-\tau_T+1} \sum_{t=0}^{T-\tau_T} \left(\left(\frac{\alpha_t}{(1-\gamma)\beta_t}\right)^2 + \left(\frac{\xi_t}{\beta_t}\right)^2\right)$$
$$\leq \frac{(T-\tau_T+1)^{-2(\alpha-\beta)}}{(1-\gamma)^2(1-2(\alpha-\beta))} + \frac{(T-\tau_T+1)^{-2(\xi-\beta)}}{1-2(\xi-\beta)}$$
$$= \mathcal{O}\left(\frac{T^{-2(\alpha-\beta)}}{(1-\gamma)^2} + T^{-2(\xi-\beta)}\right) . \quad (69)$$

(4) Similarly to item (3), to control the fourth term, we write:

$$\begin{aligned}\frac{1}{1+T-\tau_T}\sum_{t=\tau_T}^{T}\left(\frac{\alpha_t^2}{(1-\gamma)^2\beta_t}+\frac{\xi_t^2}{\beta_t}+\beta_t\right) &\leq \frac{1}{1+T-\tau_T}\sum_{t=0}^{T-\tau_T}\left(\frac{\alpha_t^2}{(1-\gamma)^2\beta_t}+\frac{\xi_t^2}{\beta_t}+\beta_t\right)\\ &\leq \frac{(1+T-\tau_T)^{-(2\alpha-\beta)}}{(1-\gamma)^2(1-(2\alpha-\beta))}+\frac{(1+T-\tau_T)^{-(2\xi-\beta)}}{1-(2\xi-\beta)}\\ &\quad+\frac{(1+T-\tau_T)^{-\beta}}{1-\beta}\\ &= \mathcal{O}\left(\frac{T^{-(2\alpha-\beta)}}{(1-\gamma)^2}+T^{-(2\xi-\beta)}+T^{-\beta}\right). \end{aligned} \tag{70}$$

Hence,

$$\frac{1}{1+T-\tau_T}I_4(T)=\mathcal{O}\left(\frac{T^{-(2\alpha-\beta)}}{(1-\gamma)^2}+T^{-(2\xi-\beta)}+T^{-\beta}\right). \tag{71}$$

Define:

$$N(T):=\frac{1}{1+T-\tau_T}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\nu_t\|^2], \tag{72}$$

$$F(T):=\frac{1}{1+T-\tau_T}\sum_{t=\tau_T}^{T}\left(\left(\frac{\alpha_t}{(1-\gamma)\beta_t}\right)^2+\left(\frac{\xi_t}{\beta_t}\right)^2\right), \tag{73}$$

$$G(T):=\frac{1}{1+T-\tau_T}(I_1(T)+I_2(T)+I_4(T)). \tag{74}$$

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

$$F(T)=\mathcal{O}\left(\frac{T^{-2(\alpha-\beta)}}{(1-\gamma)^2}+T^{-2(\xi-\beta)}\right), \tag{75}$$

$$G(T)=\mathcal{O}(T^{\beta-1})+\mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma}\ln(T)T^{-\beta}\right)+\mathcal{O}\left(\frac{T^{-(2\alpha-\beta)}}{(1-\gamma)^2}+T^{-(2\xi-\beta)}+T^{-\beta}\right). \tag{76}$$

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:

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

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

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

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 $\bar{\omega}_t-\bar{\omega}_*(\theta_t)$

Consider the shorthand notation $\bar{\nu}_t:=\bar{\omega}_t-\bar{\omega}_*(\theta_t)$.

Using the update rules of $(\bar{\omega}_t)$, (ω_t) and developing the squared norm gives:

$$\begin{aligned}
\|\bar{\nu}_{t+1}\|^2 &= \|\bar{\omega}_t + \xi_t(\omega_{t+1} - \bar{\omega}_t) - \bar{\omega}_*(\theta_{t+1})\|^2 \\
&= \|\bar{\nu}_t + \xi_t(\omega_t + \beta_t g(\tilde{x}_t, \bar{\omega}_t, \omega_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\|^2 \\
&= \|\bar{\nu}_t + (\xi_t(\nu_t + \beta_t g(\tilde{x}_t, \bar{\omega}_t, \omega_t) + \omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1}))\|^2 \\
&= \|\bar{\nu}_t\|^2 + 2\langle \bar{\nu}_t, \xi_t(\nu_t + \beta_t g(\tilde{x}_t, \bar{\omega}_t, \omega_t) + \omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\rangle \\
&+ \|\xi_t(\nu_t + \beta_t g(\tilde{x}_t, \bar{\omega}_t, \omega_t) + \omega_*(\theta_t, \bar{\omega}_t) - \bar{\omega}_t) + \bar{\omega}_*(\theta_t) - \bar{\omega}_*(\theta_{t+1})\|^2 . \qquad (77)
\end{aligned}$$

Since the sequences $(\nu_t), (\bar{\omega}_t)$ and the functions g, ω_* are bounded and the function $\bar{\omega}_*$ is Lipschitz continuous, the last squared norm term can be bounded by: $C(\xi_t^2\beta_t^2 + \xi_t^2 + \frac{\alpha_t^2}{(1-\gamma)^2})$.

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:

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

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

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

(c) Applying the Cauchy-Schwarz inequality gives:

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

(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:

$$\|\bar{\nu}_{t+1}\|^2 \leq (1 - 2\zeta\xi_t)\|\bar{\nu}_t\|^2 + C\left(\xi_t\beta_t + \frac{\alpha_t}{1-\gamma}\right)\|\bar{\nu}_t\| + 2\xi_t\|\bar{\nu}_t\| \cdot \|\nu_t\| + C\left(\xi_t^2\beta_t^2 + \xi_t^2 + \frac{\alpha_t^2}{(1-\gamma)^2}\right) . \qquad (78)$$

Rearranging Ineq. (78) leads to:

$$2\zeta\|\bar{\nu}_t\|^2 \leq \frac{1}{\xi_t}(\|\bar{\nu}_t\|^2 - \|\bar{\nu}_{t+1}\|^2) + C\left(\beta_t + \frac{\alpha_t}{(1-\gamma)\xi_t}\right)\|\bar{\nu}_t\| + 2\|\bar{\nu}_t\| \cdot \|\nu_t\| + C\left(\xi_t\beta_t^2 + \xi_t + \frac{\alpha_t^2}{(1-\gamma)^2\xi_t}\right) . \qquad (79)$$

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

$$\frac{2\zeta}{T}\sum_{t=1}^{T}\mathbb{E}[\|\bar{\nu}_t\|^2] \leq \Sigma_1(T) + \Sigma_2(T) + \Sigma_3(T) + \Sigma_4(T) , \qquad (80)$$

where

$$\Sigma_1(T) := \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\xi_t}(\mathbb{E}[\|\bar{\nu}_t\|^2] - \mathbb{E}[\|\bar{\nu}_{t+1}\|^2]) , \qquad (81)$$

$$\Sigma_2(T) := \frac{C}{T}\sum_{t=1}^{T}\left(\beta_t + \frac{\alpha_t}{(1-\gamma)\xi_t}\right)\mathbb{E}[\|\bar{\nu}_t\|] , \qquad (82)$$

$$\Sigma_3(T) := \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\|\bar{\nu}_t\| \cdot \|\nu_t\|] , \qquad (83)$$

$$\Sigma_4(T) := \frac{C}{T}\sum_{t=1}^{T}\left(\xi_t\beta_t^2 + \xi_t + \frac{\alpha_t^2}{(1-\gamma)^2\xi_t}\right) . \qquad (84)$$

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 Σ_1 as follows:

$$\Sigma_1(T) = \frac{1}{T}\left[\sum_{t=1}^{T}\left(\frac{1}{\xi_t} - \frac{1}{\xi_{t-1}}\right)\mathbb{E}[\|\bar{\nu}_t\|^2] + \frac{1}{\xi_0}\mathbb{E}[\|\bar{\nu}_1\|^2] - \frac{1}{\xi_T}\mathbb{E}[\|\bar{\nu}_{T+1}\|^2]\right] \leq \frac{C}{T\xi_T} = \mathcal{O}(T^{\xi-1})\,.$$

(ii) Cauchy-Schwarz inequality implies:

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

Moreover,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\beta_t^2 + \left(\frac{\alpha_t}{(1-\gamma)\xi_t}\right)^2\right) &\leq \frac{1}{T}\left(\frac{(T+1)^{1-2\beta}}{1-2\beta} + \frac{(T+1)^{1-2(\alpha-\xi)}}{(1-\gamma)^2(1-2(\alpha-\xi))}\right) \\ &= \mathcal{O}(T^{-2\beta}) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right)\,.\end{aligned}$$

(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:

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

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

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

$$Z(T) := \Sigma_1(T) + \Sigma_4(T)\,. \tag{88}$$

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

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

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

$$W(T) = \mathcal{O}(T^{\beta-1}) + \mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma}\ln(T)\,T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{2(\beta-\alpha)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{2(\beta-\xi)})\,. \tag{91}$$

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:

$$X(T) = \mathcal{O}(Y(T) + W(T) + Z(T))\,. \tag{92}$$

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

$$X(T) = \mathcal{O}(T^{\xi-1}) + \mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma}\ln(T)\,T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{-2(\xi-\beta)})\,. \tag{93}$$

B.1.3 End of Proof of Th. 6.1

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

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nu_t + \omega_*(\theta_t,\bar{\omega}_t) - \bar{\omega}_*(\theta_t)\|^2] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nu_t + \omega_*(\theta_t,\bar{\omega}_t) - \omega_*(\theta_t,\bar{\omega}_*(\theta_t))\|^2] \\
&\leq 2W(T) + CX(T) \\
&= \mathcal{O}(X(T)) \\
&= \mathcal{O}(T^{\xi-1}) + \mathcal{O}\left(\frac{|\mathcal{A}|}{1-\gamma}\ln(T)\,T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{-2(\xi-\beta)})\,,
\end{aligned} \tag{94}$$

where the second equality follows from using the identity $w_*(\theta,\bar{\omega}_*(\theta)) = \bar{\omega}_*(\theta)$ for every $\theta \in \mathbb{R}^d$, the inequality stems from using the classical inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ together with the fact that ω_* is Lipschitz continuous, the penultimate equality is a consequence of Eq. (92) and the last equality is the result of the previous section (see Eq. (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 $\omega \in \mathbb{R}^m, \theta \in \mathbb{R}^d$:

$$\hat{\delta}(\tilde{x},\omega) := r + \gamma\phi(s)^T\omega - \phi(\tilde{s})^T\omega \tag{95}$$

$$\delta(\tilde{x},\theta) = r + \gamma V_{\pi_\theta}(s) - V_{\pi_\theta}(\tilde{s})\,. \tag{96}$$

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

Recall that $\theta \mapsto \nabla J(\theta)$ and $\theta \mapsto V_{\pi_\theta}(s)$ (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. $\theta \mapsto V_{\pi_\theta}(s)$ for every $s \in \mathcal{S}$) and $C_{\nabla J}$ (resp. C_V) denotes the upperbound of $\theta \mapsto \|\nabla J(\theta)\|$ (resp. $\theta \mapsto V_{\pi_\theta}(s)$ for every $s \in \mathcal{S}$). Since the function ∇J is $L_{\nabla J}$-Lipschitz continuous, a classical Taylor inequality combined with the update rule of (θ_t) yields:

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{1-\gamma}\langle\nabla J(\theta_t), \hat{\delta}(\tilde{x}_t,\omega_t)\psi_{\theta_t}(\tilde{S}_t,\tilde{A}_t)\rangle - \frac{L_{\nabla J}}{2}\frac{\alpha_t^2}{(1-\gamma)^2}\|\hat{\delta}(\tilde{x}_t,\omega_t)\psi_{\theta_t}(\tilde{S}_t,\tilde{A}_t)\|^2\,. \tag{97}$$

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

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{1-\gamma}\langle\nabla J(\theta_t), \hat{\delta}(\tilde{x}_t,\omega_t)\psi_{\theta_t}(\tilde{S}_t,\tilde{A}_t)\rangle - CL_{\nabla J}\frac{\alpha_t^2}{(1-\gamma)^2}\,. \tag{98}$$

Now, we decompose the TD error by introducing both the moving target $\bar{\omega}_*(\theta_t)$ and the TD error $\delta(\tilde{x}_t, \theta_t)$ associated to the true value function $V_{\pi_{\theta_t}}$:

$$\hat{\delta}(\tilde{x}_t, \omega_t) = [\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t))] + [\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t)] + \delta(\tilde{x}_t, \theta_t)\,. \tag{99}$$

Incorporating this decomposition (99) into Eq. (98) gives:

$$\begin{aligned} J(\theta_{t+1}) \geq J(\theta_t) &+ \frac{\alpha_t}{1-\gamma}\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\rangle \\ &+ \frac{\alpha_t}{1-\gamma}\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\rangle \\ &+ \frac{\alpha_t}{1-\gamma}\langle \nabla J(\theta_t), \delta(\tilde{x}_t, \theta_t)\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) - \nabla J(\theta_t)\rangle + \frac{\alpha_t}{1-\gamma}\|\nabla J(\theta_t)\|^2 - CL_{\nabla J}\frac{\alpha_t^2}{(1-\gamma)^2}\,. \end{aligned} \tag{100}$$

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 $\hat{\delta}(\tilde{x}_t, \omega_t) - \hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) = (\gamma\phi(S_{t+1}) - \phi(\tilde{S}_t))^T(\omega_t - \bar{\omega}_*(\theta_t))$, the Cauchy-Schwarz inequality leads to:

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

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 $\theta \in \mathbb{R}^d$:

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

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

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

$$\begin{aligned} \mathbb{E}[\Gamma(\tilde{x}_t, \theta_t)] = \mathbb{E}[\Gamma(\tilde{x}_t, \theta_t) - \Gamma(\tilde{x}_t, \theta_{t-\tau})] + \mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau}) - \Gamma(\check{x}_t, \theta_{t-\tau})] \\ + \mathbb{E}[\Gamma(\check{x}_t, \theta_{t-\tau}) - \Gamma(\bar{x}_t, \theta_{t-\tau})] + \mathbb{E}[\Gamma(\bar{x}_t, \theta_{t-\tau})]\,. \end{aligned} \tag{102}$$

We address each term of this decomposition successively.

(a) For this first term, we write:

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

Moreover, note that:

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

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 θ (see, for e.g., [Shen et al., 2020, Lem. 3] for a proof for V_{π_θ}) under Assumption 3.1 , one can show after tedious inequalities that:

$$\begin{aligned} |\Gamma(\tilde{x}_t, \theta_t) - \Gamma(\tilde{x}_t, \theta_{t-\tau})| &\leq (L_{\nabla J}(C(1+C_V) + C_{\nabla J}) + CC_{\nabla J}L_V + C(1+C_V)C_{\nabla J} + C_{\nabla J}L_{\nabla J})\|\theta_t - \theta_{t-\tau}\| \\ &\leq CC_{1-\gamma}\|\theta_t - \theta_{t-\tau}\|\,, \end{aligned} \tag{103}$$

where $C_{1-\gamma} := \max(L_{\nabla J} C_V, L_{\nabla J} C_{\nabla J}, L_V C_{\nabla J}, C_V C_{\nabla J})$. 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:

$$\begin{aligned}
&|\mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau}) - \Gamma(\check{x}_t, \theta_{t-\tau})]| \\
&= |\mathbb{E}[\langle \nabla J(\theta_{t-\tau}), \delta(\tilde{x}_t, \theta_{t-\tau}) \psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t) - \delta(\check{x}_t, \theta_{t-\tau}) \psi_{\theta_{t-\tau}}(\check{S}_t, \check{A}_t) \rangle]| \\
&= |\mathbb{E}[\langle \nabla J(\theta_{t-\tau}), \delta(\tilde{x}_t, \theta_{t-\tau}) \psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t) - \delta(\check{x}_t, \theta_{t-\tau}) \psi_{\theta_{t-\tau}}(\check{S}_t, \check{A}_t) \rangle | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}]| \\
&\leq C C_V C_{\nabla J} \mathbb{E}[d_{TV}(\mathbb{P}(\tilde{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\check{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}))] \\
&\leq C C_V C_{\nabla J} |\mathcal{A}| \sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\|] \,. \qquad (104)
\end{aligned}$$

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:

$$\begin{aligned}
&d_{TV}(\mathbb{P}(\tilde{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\check{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau})) \\
&= d_{TV}(\mathbb{P}((\tilde{S}_t, \tilde{A}_t) \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}((\check{S}_t, \check{A}_t) \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau})) \\
&\leq d_{TV}(\mathbb{P}(\tilde{S}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\check{S}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau})) + \frac{1}{2} |\mathcal{A}| L_\pi \mathbb{E}[\|\theta_t - \theta_{t-\tau}\|] \,.
\end{aligned}$$

Iterating this inequality gives the desired result of Eq. (104). We conclude from this item that:

$$\mathbb{E}[\Gamma(\tilde{x}_t, \theta_{t-\tau}) - \Gamma(\check{x}_t, \theta_{t-\tau})] \geq -C C_V C_{\nabla J} |\mathcal{A}| \sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\|] \,.$$

(c) Regarding the third term, similarly to item (b), we can write:

$$\begin{aligned}
\mathbb{E}[\Gamma(\check{x}_t, \theta_{t-\tau}) - \Gamma(\bar{x}_t, \theta_{t-\tau})] &\geq -C C_V C_{\nabla J} \mathbb{E}[d_{TV}(\mathbb{P}(\check{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), \mathbb{P}(\bar{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}))] \\
&= -C C_V C_{\nabla J} \mathbb{E}[d_{TV}(\mathbb{P}(\check{x}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), d_{\rho, \theta_{t-\tau}} \otimes \pi_{\theta_{t-\tau}} \otimes p)] \\
&= -C C_V C_{\nabla J} \mathbb{E}[d_{TV}(\mathbb{P}(\check{S}_t \in \cdot | \tilde{S}_{t-\tau+1}, \theta_{t-\tau}), d_{\rho, \theta_{t-\tau}})] \\
&\geq -C C_V C_{\nabla J} \sigma^{\tau-1} \,, \qquad (105)
\end{aligned}$$

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

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

We conlude this section from Eq. (102) by collecting Eqs. (103) to (105) (items (a) to (d)) to obtain:

$$\begin{aligned}
\mathbb{E}[\Gamma(\tilde{x}_t, \theta_t)] &\geq -C C_{1-\gamma} \mathbb{E}[\|\theta_t - \theta_{t-\tau}\|] - C C_V C_{\nabla J} \sum_{i=t-\tau+1}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\|] - C C_V C_{\nabla J} \sigma^{\tau-1} \\
&\geq -C C_{1-\gamma} \sum_{i=t-\tau+1}^{t} \mathbb{E}[\|\theta_i - \theta_{i-1}\|] - C C_V C_{\nabla J} \sum_{i=t-\tau+1}^{t} \sum_{j=t-\tau+1}^{i} \mathbb{E}[\|\theta_j - \theta_{j-1}\|] - C C_V C_{\nabla J} \sigma^{\tau-1} \\
&\geq -C C_{1-\gamma} \sum_{i=t-\tau+1}^{t} \mathbb{E}[\|\theta_i - \theta_{i-1}\|] - C C_V C_{\nabla J} \sum_{i=t-\tau+1}^{t} \sum_{j=t-\tau+1}^{t} \mathbb{E}[\|\theta_j - \theta_{j-1}\|] - C C_V C_{\nabla J} \sigma^{\tau-1} \\
&\geq -C (C_{1-\gamma} + C_V C_{\nabla J} \tau) \sum_{i=t-\tau+1}^{t} \mathbb{E}[\|\theta_i - \theta_{i-1}\|] - C C_V C_{\nabla J} \sigma^{\tau-1} \\
&\geq -C \left((C_{1-\gamma} \tau + C_V C_{\nabla J} \tau^2) \frac{\alpha_{t-\tau}}{1-\gamma} + C_V C_{\nabla J} \alpha_T \right) , \qquad (106)
\end{aligned}$$

where the last inequality uses the definition of the mixing time τ and the fact that the sequence (α_t) is nonincreasing.

B.2.2 Control of the linear FA error term

Recall that $\theta \mapsto \psi_\theta$ is Lipschitz continuous, ∇J is bounded and remark that the quantity $\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t)$ is bounded. Therefore, using the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
&\mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\rangle] \\
&= \mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))(\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t) - \psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t))\rangle] \\
&+ \mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)\rangle] \\
&\geq -C(1 + C_V)C_{\nabla J}\mathbb{E}[\|\theta_t - \theta_{t-\tau}\|] + \mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)\rangle] \,. \qquad (107)
\end{aligned}$$

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:

$$\Delta(\tilde{x}, \theta) := \langle \nabla J(\theta), (\hat{\delta}(\tilde{x}, \bar{\omega}_*(\theta)) - \delta(\tilde{x}, \theta))\psi_{\theta_{t-\tau}}(\tilde{S}_t, \tilde{A}_t)\rangle \,.$$

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:

$$\begin{aligned}
\Delta(\tilde{x}_t, \theta_t) = (\Delta(\tilde{x}_t, \theta_t) - \Delta(\tilde{x}_t, \theta_{t-\tau})) + (\Delta(\tilde{x}_t, \theta_{t-\tau}) - \Delta(\check{x}_t, \theta_{t-\tau})) \\
+ (\Delta(\check{x}_t, \theta_{t-\tau}) - \Delta(\bar{x}_t, \theta_{t-\tau})) + \Delta(\bar{x}_t, \theta_{t-\tau}) \,. \qquad (108)
\end{aligned}$$

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

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

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

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

Using similar manipulations to the previous section, we get:

(ii)

$$\mathbb{E}[\Delta(\tilde{x}_t, \theta_{t-\tau}) - \Delta(\check{x}_t, \theta_{t-\tau})] \geq -CC_{\nabla J}(1 + C_V)|\mathcal{A}| \sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\|] \,. \qquad (109)$$

(iii)

$$\mathbb{E}[\Delta(\check{x}_t, \theta_{t-\tau}) - \Delta(\bar{x}_t, \theta_{t-\tau})] \geq -CC_{\nabla J}(1 + C_V)\sigma^{\tau-1} \,. \qquad (110)$$

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

$$\mathbb{E}[\Delta(\bar{x}_t, \theta_{t-\tau})|\theta_{t-\tau}] \geq -C\|\nabla J(\theta_{t-\tau})\| \cdot \mathbb{E}[|\hat{\delta}(\bar{x}_t, \bar{\omega}_*(\theta_{t-\tau})) - \delta(\bar{x}_t, \theta_{t-\tau})||\theta_{t-\tau}] \,. \qquad (111)$$

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

$$\begin{aligned}
\hat{\delta}(\bar{x}_t, \bar{\omega}_*(\theta_{t-\tau})) - \delta(\bar{x}_t, \theta_{t-\tau}) = \gamma(\phi(S_{t+1})^T\bar{\omega}_*(\theta_{t-\tau}) - V_{\pi_{\theta_{t-\tau}}}(S_{t+1})) \\
+ (V_{\pi_{\theta_{t-\tau}}}(\bar{S}_t) - \phi(\bar{S}_t)^T\bar{\omega}_*(\theta_{t-\tau})) \,. \qquad (112)
\end{aligned}$$

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

$$\mathbb{E}[|\hat{\delta}(\bar{x}_t, \bar{\omega}_*(\theta_{t-\tau})) - \delta(\bar{x}_t, \theta_{t-\tau})||\theta_{t-\tau}] \leq C\epsilon_{\text{FA}} \,.$$

As a consequence, noticing that ∇J is bounded, we obtain from Eq. (111):

$$\mathbb{E}[\Delta(\bar{x}_t, \theta_{t-\tau})] \geq -CC_{\nabla J}\epsilon_{\text{FA}} \,.$$

Combining items (i) to (iv) with the boundedness of the function ∇J, we conclude from this section that:

$$\begin{aligned}
&\mathbb{E}[\langle \nabla J(\theta_t), (\hat{\delta}(\tilde{x}_t, \bar{\omega}_*(\theta_t)) - \delta(\tilde{x}_t, \theta_t))\psi_{\theta_t}(\tilde{S}_t, \tilde{A}_t)\rangle] \\
&\geq -C((1+C_V)C_{\nabla J} + \tilde{C}_{1-\gamma})\mathbb{E}[\|\theta_t - \theta_{t-\tau}\|] - CC_{\nabla J}(1+C_V)|\mathcal{A}| \sum_{i=t-\tau}^{t} \mathbb{E}[\|\theta_i - \theta_{t-\tau}\|] \\
&\quad - CC_{\nabla J}(1+C_V)\sigma^{\tau-1} - CC_{\nabla J}\epsilon_{\mathrm{FA}} \\
&\geq -C\left(\left(((1+C_V)C_{\nabla J} + \tilde{C}_{1-\gamma})\tau + C_{\nabla J}(1+C_V)|\mathcal{A}|\tau^2\right)\frac{\alpha_{t-\tau}}{1-\gamma} + C_{\nabla J}(1+C_V)\alpha_T + C_{\nabla J}\epsilon_{\mathrm{FA}}\right) \qquad (113)
\end{aligned}$$

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:

$$\begin{aligned}
\mathbb{E}[J(\theta_{t+1})] \geq \mathbb{E}[J(\theta_t)] + \frac{\alpha_t}{1-\gamma}\mathbb{E}[\|\nabla J(\theta_t)\|^2] - C\frac{\alpha_t}{1-\gamma}\sqrt{\mathbb{E}[\|\nabla J(\theta_t)\|^2]}\sqrt{\mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2]} \\
- C\frac{\alpha_t}{1-\gamma}((C^1_{1-\gamma}\tau + C^2_{1-\gamma}\tau^2)\frac{\alpha_{t-\tau}}{1-\gamma} + C^3_{1-\gamma}\alpha_T + C_{\nabla J}\epsilon_{\mathrm{FA}}) - CL_{\nabla J}\frac{\alpha_t^2}{(1-\gamma)^2}\,, \qquad (114)
\end{aligned}$$

where $C^1_{1-\gamma} := (1+C_V)C_{\nabla J} + \tilde{C}_{1-\gamma} + C_{1-\gamma}$, $C^2_{1-\gamma} := C_V C_{\nabla J} + C_{\nabla J}(1+C_V)|\mathcal{A}|$ and $C^3_{1-\gamma} := C_V C_{\nabla J} + C_{\nabla J}(1+C_V)$.

Rearranging and summing this inequality for $t = \tau_T$ to T lead to:

$$\frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] \leq U_1(T) + U_2(T) + U_3(T) + CC_{\nabla J}\epsilon_{\mathrm{FA}}\,, \qquad (115)$$

where

$$U_1(T) := \frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\frac{1-\gamma}{\alpha_t}(\mathbb{E}[J(\theta_{t+1})] - \mathbb{E}[J(\theta_t)])\,, \qquad (116)$$

$$U_2(T) := \frac{C}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\left((C^1_{1-\gamma}\tau_T + C^2_{1-\gamma}\tau_T^2)\frac{\alpha_{t-\tau_T}}{1-\gamma} + C^3_{1-\gamma}\alpha_T + L_{\nabla J}\frac{\alpha_t}{1-\gamma}\right)\,, \qquad (117)$$

$$U_3(T) := \frac{C}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\sqrt{\mathbb{E}[\|\nabla J(\theta_t)\|^2]}\sqrt{\mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2]}\,. \qquad (118)$$

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 (α_t) is nonincreasing, the first term can be controlled as follows:

$$\begin{aligned}
U_1(T) &= \frac{1-\gamma}{T-\tau_T+1}\left(\frac{1}{\alpha_T}\mathbb{E}[J(\theta_{T+1})] - \frac{1}{\alpha_{\tau_T-1}}\mathbb{E}[J(\theta_{\tau_T})] + \sum_{t=\tau_T}^{T}\left(\frac{1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\right)\mathbb{E}[J(\theta_t)]\right) \\
&\leq \frac{U_R}{T-\tau_T+1}\left(\frac{1}{\alpha_T} + \frac{1}{\alpha_{\tau_T-1}} + \frac{1}{\alpha_T} - \frac{1}{\alpha_{\tau_T-1}}\right) \\
&\leq \frac{U_R}{T-\tau_T+1}\frac{2}{\alpha_T} \\
&= \mathcal{O}\left(T^{\alpha-1}\right)\,. \qquad (119)
\end{aligned}$$

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

$C_{1-\gamma} = \mathcal{O}((1-\gamma)^{-5})$, $\tilde{C}_{1-\gamma} = \mathcal{O}((1-\gamma)^{-4})$; $C^1_{1-\gamma} = \mathcal{O}((1-\gamma)^{-5})$; $C^2_{1-\gamma} = \mathcal{O}((1-\gamma)^{-3})$; $C^3_{1-\gamma} = \mathcal{O}((1-\gamma)^{-3})$.

Recalling that the sequence of stepsizes (α_t) is nonincreasing and that $\tau_T = \mathcal{O}(\ln T)$, the second term can be estimated by the following derivations:

$$\begin{aligned}
U_2(T) &= \frac{C}{T-\tau_T+1}\left((C^1_{1-\gamma}\tau_T + C^2_{1-\gamma}\tau_T^2)\sum_{t=\tau_T}^{T}\frac{\alpha_{t-\tau_T}}{1-\gamma} + C^3_{1-\gamma}(T-\tau_T+1)\alpha_T + L_{\nabla J}\sum_{t=\tau_T}^{T}\frac{\alpha_t}{1-\gamma}\right)\\
&\leq \frac{C}{T-\tau_T+1}\left((C^1_{1-\gamma}\tau_T + C^2_{1-\gamma}\tau_T^2)\sum_{t=0}^{T-\tau_T}\frac{\alpha_t}{1-\gamma} + C^3_{1-\gamma}(T-\tau_T+1)\alpha_T + L_{\nabla J}\sum_{t=0}^{T-\tau_T}\frac{\alpha_t}{1-\gamma}\right)\\
&\leq \frac{C}{T-\tau_T+1}\left(\frac{(C^1_{1-\gamma}\tau_T + C^2_{1-\gamma}\tau_T^2)+L_{\nabla J}}{1-\gamma}\cdot\frac{(T-\tau_T+1)^{1-\alpha}}{1-\alpha} + C^3_{1-\gamma}(T-\tau_T+1)\alpha_T\right)\\
&= \mathcal{O}\left(\frac{\ln^2 T}{(1-\gamma)^6}T^{-\alpha}\right) \qquad (120)
\end{aligned}$$

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

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

Define the quantities:

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

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

$$K(T) := U_1(T) + U_2(T) + CC_{\nabla J}\epsilon_{\mathrm{FA}}\,. \qquad (124)$$

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:

$$F(T) = \mathcal{O}(E(T)) + \mathcal{O}(K(T))\,. \qquad (125)$$

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

$$K(T) = \mathcal{O}\left(T^{\alpha-1}\right) + \mathcal{O}\left(\frac{\ln^2 T}{(1-\gamma)^6}T^{-\alpha}\right) + \mathcal{O}\left(\frac{\epsilon_{\mathrm{FA}}}{(1-\gamma)^2}\right)\,. \qquad (126)$$

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

$$\begin{aligned}
\frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] &= \mathcal{O}\left(T^{\alpha-1}\right) + \mathcal{O}\left(\frac{\ln^2 T}{(1-\gamma)^6}T^{-\alpha}\right) + \mathcal{O}\left(\frac{\epsilon_{\mathrm{FA}}}{(1-\gamma)^2}\right)\\
&\quad + \mathcal{O}\left(\frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\omega_t - \bar{\omega}_*(\theta_t)\|^2]\right)\,.
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] &= \mathcal{O}\left(T^{\alpha-1}\right) + \mathcal{O}\left(\frac{\ln^2 T}{(1-\gamma)^6}T^{-\alpha}\right) + \mathcal{O}\left(\frac{\epsilon_{\mathrm{FA}}}{(1-\gamma)^2}\right)\\
&\quad + \mathcal{O}(T^{\xi-1}) + \mathcal{O}\left(\frac{\ln T}{1-\gamma}T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{-2(\xi-\beta)})\,. \qquad (127)
\end{aligned}$$

Then, we can write

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] &= \frac{1}{T}\left(\sum_{t=1}^{\tau_T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2] + \sum_{t=\tau_T}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]\right)\\ &\leq \frac{C\ln T}{T} + \mathcal{O}\left(\frac{1}{T-\tau_T+1}\sum_{t=\tau_T}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]\right)\\ &= \mathcal{O}\left(T^{\alpha-1}\right) + \mathcal{O}\left(\frac{\ln T}{(1-\gamma)^6}T^{-\beta}\right) + \mathcal{O}\left(\frac{T^{-2(\alpha-\xi)}}{(1-\gamma)^2}\right) + \mathcal{O}(T^{-2(\xi-\beta)}) + \mathcal{O}\left(\frac{\epsilon_{\text{FA}}}{(1-\gamma)^2}\right).\end{aligned}$$

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, 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 (θ_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., 2020b]. Let $\Gamma : \mathbb{R}^m \to \mathbb{R}$ be a function assumed to satisfy the following inequalities for some positive constants $C_1 < C_2$: for every $\omega \in \mathbb{R}^m$, $\|\omega\|\cdot\Gamma(\omega) \in [C_1, C_2]$, and for every $\omega, \omega' \in \mathbb{R}^m$, $|\Gamma(\omega) - \Gamma(\omega')| \leq \frac{C_2\|\omega-\omega'\|}{1+\|\omega\|+\|\omega'\|}$. An example of such function as provided in [Konda, 2002] is for instance the function defined for every $\omega \in \mathbb{R}^m$ by:

$$\Gamma(\omega) := \mathbb{1}_{\|\omega\|\leq C_0} + \frac{1+C_0}{1+\|\omega\|}\mathbb{1}_{\|\omega\|\geq C_0},$$

where C_0 is some given positive constant. Given such a projection-like function Γ, we replace the update rule of the actor of our actor-critic algorithm (see Algorithm 1) 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 (ξ_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, $\sup_k(\|\bar{\omega}_k\| + \|\omega_k\|) < \infty, a.s.$, 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 (ω_t) evolving on the fast timescale driven by the stepsizes (β_t) using the norm of the sequence $(\bar{\omega}_t)$ updated in a slower timescale defined by the stepsizes (ξ_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 $h_\theta : \mathbb{R}\times\mathcal{S}^2 \to \mathbb{R}^{2m}$ and $G_\theta : \mathbb{R}\times\mathcal{S}^2 \to \mathbb{R}^{2m\times 2m}$ for every $y = (r, \tilde{s}, s') \in \mathbb{R} \times \mathcal{S}^2$ by:

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

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

$$r_{k+1} = r_k + \beta_k\Big(h_{\theta_k}(Y_{k+1}) - G_{\theta_k}(Y_{k+1})r_k\Big) + \beta_k M_{k+1} r_k + \beta_k \eta_{k+1},$$

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

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

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

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

Consider now the functions $\tilde{h} : \mathbb{R}^d \to \mathbb{R}^{2m}$ and $\tilde{G} : \mathbb{R}^d \to \mathbb{R}^{2m\times 2m}$ defined for every $\theta \in \mathbb{R}^d$ by:

$$\tilde{h}(\theta) := \begin{bmatrix} h(\theta) \\ 0 \end{bmatrix}, \quad \tilde{G}(\theta) := \begin{bmatrix} \bar{G}(\theta) & -\gamma\Phi^T D_{\rho,\theta} P_\theta \Phi \\ 0 & 0 \end{bmatrix},$$

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

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

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

where T is a positive constant that will be chosen appropiately 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^β). The superscript β 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 $\hat{r}_k^j = \frac{r_k}{\max(1, \|r_{k_j}\|)}$ defined for every $k \geq k_j$ and which satisfy the following recurrence relation:

$$\hat{r}_{k+1}^j = \hat{r}_k^j + \beta_k\left(\frac{\tilde{h}(\theta_k)}{\max(1, \|r_{k_j}\|)} - \tilde{G}(\theta_k)\hat{r}_k^j\right) + \beta_k\hat{\epsilon}_{k+1}^j + \beta_k\frac{\eta_{k+1}}{\max(1, \|r_{k_j}\|)},$$

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

$$\hat{\epsilon}_{k+1}^j := \left(\frac{h_{\theta_k}(Y_{k+1}) - \tilde{h}(\theta_k)}{\max(1, \|r_{k_j}\|)} - (G_{\theta_k}(Y_{k+1}) - \tilde{G}(\theta_k))\hat{r}_k^j\right) + \beta_k M_{k+1}\hat{r}_k^j.$$

We also introduce the iterates (r_k^j) defined as follows: $r_{k_j}^j = \hat{r}_{k_j}$ and

$$r_{k+1}^j = r_k^j + \beta_k \left(\frac{\tilde{h}(\theta_k)}{\max(1, \|r_{k_j}\|)} - \tilde{G}(\theta_k) r_k^j \right) + \beta_k \frac{\eta_{k+1}}{\max(1, \|r_{k_j}\|)}.$$

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 $j \in \mathbb{N}, k \geq k_j$:

$$\begin{cases} \omega_{k+1}^j &= \omega_k^j + \beta_k \left(\frac{h(\theta_k)}{\max(1, \|r_{k_j}\|)} + \gamma \Phi^T D_{\rho,\theta_k} P_{\theta_k} \Phi \bar{\omega}_k^j - \bar{G}(\theta_k) \omega_k^j \right), \\ \bar{\omega}_{k+1}^j &= \bar{\omega}_k^j + \xi_k (\omega_{k+1}^j - \bar{\omega}_k^j), \end{cases} \tag{129}$$

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 (ε_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 $\omega \in \mathbb{R}^m$, $\omega^T G \omega \geq \epsilon \|\omega\|^2$ where $\epsilon > 0$ is a constant. Then, for sufficiently small $\gamma > 0$, $\|(I - \gamma G)\omega\| \leq (1 - \frac{1}{2}\gamma\epsilon)\|\omega\| \leq e^{-\frac{1}{2}\gamma\epsilon}\|\omega\|$.

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) $\lim_j \max_{k_j \leq k \leq k_{j+1}} \|\hat{r}_k^j - r_k^j\| = 0$, *a.s.*
(iii) There exists a constant $C' > 0$ s.t. $\sup_j \max_{k_j \leq k \leq k_{j+1}} \|\hat{r}_k^j\| \leq C'$, *a.s.*.

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

$$\begin{aligned} \|\omega_{k+1}^j\| &\leq \|(I - \beta_k \bar{G}(\theta_k))\omega_k^j\| + \beta_k \frac{\|h(\theta_k)\|}{\max(1, \|r_{k_j}\|)} + \beta_k \|\gamma \Phi^T D_{\rho,\theta_k} P_{\theta_k} \Phi \bar{\omega}_k^j\| \\ &\leq (1 - \frac{1}{2}\beta_k \epsilon)\|\omega_k^j\| + \beta_k \frac{C_1}{\max(1, \|r_{k_j}\|)} + \beta_k C_2 \|\bar{\omega}_k^j\| \\ &\leq e^{-\frac{1}{2}\epsilon \sum_{i=k_j}^k \beta_i} \|\omega_{k_j}^j\| + \left(\sum_{i=k_j}^k \beta_i \right) \frac{C_1}{\max(1, \|r_{k_j}\|)} + C_2 \left(\sum_{i=k_j}^k \beta_i \|\bar{\omega}_i^j\| \right) \\ &\leq 1 + T' C_1 + C_2 \left(\sum_{i=k_j}^k \beta_i \|\bar{\omega}_i^j\| \right), \end{aligned} \tag{130}$$

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 ϵ-uniformly positive definite (i.e., for every $\omega \in \mathbb{R}^m$, $\omega^T \bar{G}(\theta)\omega \geq \epsilon\|\omega\|^2$) together with Lem. C.3 and the last inequality stems from the fact that $\|\omega_{k_j}^j\| \leq 1$ by definition.

We now relate the term $\|\bar{\omega}_i^j\|$ to the quantity $\max_{k_j \leq l \leq i} \|\omega_l^j\|$. For every $i \in \{k_j \cdots, k_{j+1} - 1\}$,

$$\|\bar{\omega}_{i+1}^j\| \leq \|\bar{\omega}_i^j\| + \xi_i \|\omega_{i+1}^j\| \leq \|\bar{\omega}_{k_j}^j\| + \sum_{l=k_j}^i \xi_l \|\omega_{l+1}^j\| \leq 1 + T' \left(\max_{k_j \leq l \leq i} \frac{\xi_l}{\beta_l} \right) \left(\max_{k_j \leq l \leq i+1} \|\omega_l^j\| \right).$$

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

$$\begin{aligned} \max_{k_j \leq k \leq k_{j+1}} \|\omega_k^j\| &\leq (1 + T' C_1 + C_2 T') + C_2 T' \max_{k_j \leq k \leq k_{j+1}} \left(\sum_{i=k_j}^k \beta_i \left(\max_{k_j \leq l \leq i} \frac{\xi_l}{\beta_l} \right) \left(\max_{k_j \leq l \leq i} \|\omega_l^j\| \right) \right) \\ &\leq (1 + T' C_1 + C_2 T') + C_2 T'^2 \left(\max_{k_j \leq k \leq k_{j+1}} \frac{\xi_k}{\beta_k} \right) \left(\max_{k_j \leq k \leq k_{j+1}} \|\omega_k^j\| \right). \end{aligned}$$

Since the sequence $(\frac{\xi_k}{\beta_k})$ converges to 0 by Assumption 5.2 , there exists $\upsilon > 0$ s.t. for j sufficiently large, $C_2 T'^2 (\max_{k_j \leq k \leq k_{j+1}} \frac{\xi_k}{\beta_k}) \leq 1 - \upsilon$. Thus,

$$\max_{k_j \leq k \leq k_{j+1}} \|\omega_k^j\| \leq \frac{1 + T' C_1 + C_2 T'}{\upsilon},$$

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 $\tilde{G}(\theta)$ 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 $(\max_{k_j \leq k \leq k_{j+1}} \|\hat{r}_k^j - r_k^j\|)_j$ is a.s. bounded.

□

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

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

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 $\tilde{\omega}_j^*(\bar{\omega}, \theta)$ of $\omega_*(\bar{\omega}, \theta)$ as follows:

$$\tilde{\omega}_j^*(\bar{\omega}, \theta) := \bar{G}(\theta)^{-1} \left(\frac{h(\theta)}{\max(1, \|r_{k_j}\|)} + \gamma \Phi^T D_{\rho,\theta} P_\theta \Phi \bar{\omega} \right) . \tag{131}$$

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

$$\max \left(\|\omega_*(\bar{\omega}, \theta)\|, \|\tilde{\omega}_j^*(\bar{\omega}, \theta)\| \right) \leq C^* (1 + \|\bar{\omega}\|) . \tag{132}$$

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,

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

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:

$$\begin{aligned}
\|r_{k_j}\| = \|(\omega_{k_j}, \bar{\omega}_{k_j})\| &= \sqrt{\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j}) + \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + \|\bar{\omega}_{k_j}\|^2} \\
&\leq \sqrt{2\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + 2\|\omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\|^2 + \|\bar{\omega}_{k_j}\|^2} \\
&\leq \sqrt{2}\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\| + \sqrt{2C^{*2}(1 + \|\bar{\omega}_{k_j}\|)^2 + \|\bar{\omega}_{k_j}\|^2} \\
&\leq \sqrt{2}\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\| + \sqrt{2}C^* + (\sqrt{2}C^* + 1)\|\bar{\omega}_{k_j}\|) .
\end{aligned}$$

As a consequence, we have:

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

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

$$\frac{\|\omega_{k_{j+1}} - \omega_*(\bar{\omega}_{k_{j+1}}, \theta_{k_{j+1}})\|}{\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\|} = \frac{\|\hat{\omega}_{k_{j+1}}^j - \tilde{\omega}_j^*(\hat{\bar{\omega}}_{k_{j+1}}^j, \theta_{k_{j+1}})\|}{\|\hat{\omega}_{k_j}^j - \tilde{\omega}_j^*(\hat{\bar{\omega}}_{k_j}^j, \theta_{k_j})\|} \leq C_2 (\|\hat{\omega}_{k_{j+1}}^j - \omega_{k_{j+1}}^j\| + \|\omega_{k_{j+1}}^j - \tilde{\omega}_j^*(\hat{\bar{\omega}}_{k_{j+1}}, \theta_{k_{j+1}})\|) \tag{133}$$

Since the first term of the right-hand side converges a.s. to zero as j goes to infinity by Lem. C.4 , there exists $j_0 \in \mathbb{N}$ s.t. for every $j \geq j_0$,

$$\|\hat{\omega}^j_{k_{j+1}} - \omega^j_{k_{j+1}}\| \leq \frac{1}{4C_2}\,, a.s. \tag{134}$$

We now establish a bound for the second term in Eq. (133). For every $k_j \leq k < k_{j+1}$, we have that:

$$\begin{aligned}
\|\omega^j_{k+1} - \tilde{\omega}^*_j(\bar{\omega}^j_{k+1}, \theta_{k+1})\| &= \|\omega^j_k - \tilde{\omega}^*_j(\bar{\omega}^j_k, \theta_k) + \beta_k(\frac{h(\theta_k)}{\max(1, \|r_{k_j}\|)} + \gamma \Phi^T D_{\rho,\theta_k} P_{\theta_k} \Phi \bar{\omega}^j_k - \bar{G}(\theta_k)\omega^j_k) \\
&\quad + \tilde{\omega}^*_j(\bar{\omega}^j_k, \theta_k) - \tilde{\omega}^*_j(\bar{\omega}^j_{k+1}, \theta_{k+1})\| \\
&\leq \|\omega^j_k - \tilde{\omega}^*_j(\bar{\omega}^j_k, \theta_k) - \beta_k \bar{G}(\theta_k)(\omega^j_k - \tilde{\omega}^*_j(\bar{\omega}^j_k, \theta_k))\| \\
&\quad + \|\tilde{\omega}^*_j(\bar{\omega}^j_k, \theta_k) - \tilde{\omega}^*_j(\bar{\omega}^j_{k+1}, \theta_{k+1})\| \\
&\leq \|I - \beta_k \bar{G}(\theta_k)\| \|\omega^j_k - \tilde{\omega}^*_j(\bar{\omega}^j_{k_j}, \theta_k)\| + C(\xi_k + \alpha_k)
\end{aligned}$$

where $C > 0$ is a constant coming from Lem. C.4 and the last inequality stems from the fact that the function $(\bar{\omega}, \theta) \mapsto \tilde{\omega}^*_j(\bar{\omega}, \theta)$ 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:

$$\begin{aligned}
\|\omega^j_{k_{j+1}} - \tilde{\omega}^*_j(\bar{\omega}^j_{k_{j+1}}, \theta_{k_{j+1}})\| &\leq e^{-\frac{1}{2}\epsilon T} \|\omega^j_{k_j} - \tilde{\omega}^*_j(\bar{\omega}^j_{k_j}, \theta_{k_j})\| + C \sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k) \\
&\leq e^{-\frac{1}{2}\epsilon T} \left(\|\omega^j_{k_j}\| + \|\tilde{\omega}^*_j(\bar{\omega}^j_{k_j}, \theta_{k_j})\| \right) + C \sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k)\,.
\end{aligned} \tag{135}$$

By definition, $\|\omega^j_{k_j}\| \leq 1$, $\|\bar{\omega}^j_{k_j}\| \leq 1$, and it stems from Eq. (132) that $\|\omega^j_{k_j}\| + \|\tilde{\omega}^*_j(\bar{\omega}^j_{k_j}, \theta_{k_j})\| \leq C'$ for some $C' > 0$. Choosing $T \geq \frac{2\ln(4C'/C_2)}{\epsilon}$, we obtain: $e^{-\frac{1}{2}\epsilon T} \left(\|\omega^j_{k_j}\| + \|\tilde{\omega}^*_j(\bar{\omega}^j_{k_j}, \theta_{k_j})\| \right) \leq \frac{1}{4C_2}$. We also have that for every $j \in \mathbb{N}$, $\sum_{k=k_j}^{k_{j+1}} (\xi_k + \alpha_k) \leq \max_{k_j \leq k \leq k_{j+1}} \frac{\xi_k + \alpha_k}{\beta_k} T'$. 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 $j \geq \max(j_0, j_1)$,

$$\|\omega^j_{k_{j+1}} - \tilde{\omega}^*_j(\bar{\omega}^j_{k_{j+1}}, \theta_{k_{j+1}})\| \leq \frac{1}{2C_2}\,. \tag{136}$$

Combining Eq. (133) with Eqs. (134) and (136) yields for every $j \geq \max(j_0, j_1)$,

$$\frac{\|\omega_{k_{j+1}} - \omega_*(\bar{\omega}_{k_{j+1}}, \theta_{k_{j+1}})\|}{\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\|} \leq C_2 \left(\frac{1}{4C_2} + \frac{1}{2C_2} \right) = \frac{3}{4}\,,$$

which is the desired inequality.

□

Theorem C.6. There exists a constant $C > 0$ s.t. for every $j \in \mathbb{N}$,

(i) $\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\| \leq C(1 + \|\bar{\omega}_{k_j}\|), a.s.$
(ii) $\|\omega_{k_j}\| \leq C(1 + \|\bar{\omega}_{k_j}\|), a.s.$
(iii) $\max_{k_j \leq k \leq k_{j+1}} \|\omega_k\| \leq C(1 + \|\bar{\omega}_{k_j}\|), a.s.$

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_l) for which $C_{j_l} \uparrow \infty$ as $l \to \infty$ and $\|\omega_{k_{j_l}}\| \geq C_{j_l}(1 + \|\bar{\omega}_{k_{j_l}}\|)$. Now, from Lem. C.5 , we know that if $\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\| > C_1(1 + \|\bar{\omega}_{k_j}\|)$, then $\|\omega_{k_i} - \omega_*(\bar{\omega}_{k_i}, \theta_{k_i})\|$ for $i \geq j$ falls at an exponential rate until it is within the ball of radius $C_1(1 + \|\bar{\omega}_{k_j}\|)$. Thus, corresponding to the sequence (j_l), there must exist another sequence (j'_l) s.t. $j_{l-1} \leq j'_l \leq j_l$ and $\|\omega_{k_{j'_{l-1}}} - \omega_*(\bar{\omega}_{k_{j'_{l-1}}}, \theta_{k_{j'_{l-1}}})\|$ is within the ball of radius $C_1(1 + \|\bar{\omega}_{k_{j'_{l-1}}}\|)$ and $\|\omega_{k'_l} - \omega_*(\bar{\omega}_{k'_l}, \theta_{k'_l})\|$

is greater than $C_{j_l}(1+\|\bar{\omega}_{k_{j'_l}}\|)$. However, we know from Lem. C.4 that the iterates can only grow by a factor of C' between the time $k_{j'_l-1}$ and $k_{j'_l}$. This leads to a contradiction. We conclude that $\|\omega_{k_j} - \omega_*(\bar{\omega}_{k_j}, \theta_{k_j})\| \leq \bar{C}(1+\|\bar{\omega}_{k_j}\|)$ 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 $(\hat{\omega}_k^j)$ 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:

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

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

□

C.3 Slower timescale analysis

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

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

Given a constant $T > 0$, let (k_j^β) be defined as in Eq. (128) and define the sequence (k_n^ξ) (which we will sometimes simply denote (k_n) in the rest of this section when unambiguous) as follows:

$$k_0^\xi = 0, \qquad k_{n+1}^\xi = \min\left\{ k_j^\beta > k_n \,:\, j \in \mathbb{N}\,, \sum_{l=k_n}^{k_j^\beta - 1} \xi_l > T \right\}.$$

Since ξ_k/β_k converges to 0, there exists $C_\xi > 0$ such that $T \leq \sum_{l=k_n}^{k_{n+1}} \xi_l \leq C_\xi T$.

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

$$\begin{cases} \hat{\omega}_{k_n}^n &= \frac{\omega_{k_n}}{\max(1,\|r_{k_n}\|)} \\ \hat{\omega}_{k+1}^n &= \hat{\omega}_k^n + \beta_k\phi(\tilde{S}_k)\left(\frac{R_{k+1}}{\max(1,\|r_{k_n}\|)} + \gamma\phi(S_{k+1})^T\hat{\omega}_k^n - \phi(\tilde{S}_k)^T\hat{\omega}_k^n\right) \end{cases}; \quad \begin{cases} \hat{\bar{\omega}}_{k_n}^n &= \frac{\bar{\omega}_{k_n}}{\max(1,\|r_{k_n}\|)} \\ \hat{\bar{\omega}}_{k+1}^n &= \hat{\bar{\omega}}_k^n + \xi_k(\hat{\omega}_{k+1}^n - \hat{\bar{\omega}}_k^n)\,, \end{cases} \tag{137}$$

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

$$\begin{cases} \omega_{k_n}^n &= \hat{\omega}_{k_n}^n \\ \omega_{k+1}^n &= \omega_k^n + \beta_k(\frac{h(\theta_k)}{\max(1,\|r_{k_n}\|)} + \gamma\Phi^T D_{\rho,\theta_k} P_{\theta_k}\Phi\bar{\omega}_k^n - \bar{G}(\theta_k)\omega_k^n) \end{cases}; \quad \begin{cases} \bar{\omega}_{k_n}^n &= \hat{\bar{\omega}}_{k_n}^n \\ \bar{\omega}_{k+1}^n &= \bar{\omega}_k^n + \xi_k(\omega_{k+1}^n - \bar{\omega}_k^n)\,. \end{cases} \tag{138}$$

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

Lemma C.7. The following assertions hold true:
(i) $\sup_n \max_{k_n \leq k \leq k_{n+1}}(\|\hat{\bar{\omega}}_k^n\| + \|\hat{\omega}_k^n\|) < \infty\,, a.s.$
(ii) $\sup_n \max_{k_n \leq k \leq k_{n+1}}(\|\bar{\omega}_k^n\| + \|\omega_k^n\|) < \infty\,, a.s.$

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

$$\|\hat{\bar{\omega}}_{k+1}^n\| \leq (1-\xi_k)\|\hat{\bar{\omega}}_k^n\| + \xi_k\|\hat{\omega}_{k+1}^n\| \leq \|\hat{\bar{\omega}}_k^n\| + \xi_k C(1+\|\hat{\bar{\omega}}_{k_j^\beta}^n\|)\,,$$

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.,

$$\|\hat{\bar{\omega}}_{k+1}^n\| \leq \left(1 + C\sum_{l=k_j^\beta}^{k} \xi_l\right) \|\hat{\bar{\omega}}_{k_j^\beta}^n\| + C\sum_{l=k_j^\beta}^{k} \xi_l \leq e^{C\sum_{l=k_j^\beta}^{k} \xi_l} \|\hat{\bar{\omega}}_{k_j^\beta}^n\| + C\sum_{l=k_j^\beta}^{k} \xi_l \,.$$

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.,

$$\|\hat{\bar{\omega}}_{k_{j+1}^\beta}^n\| \leq e^{Cu_j} \|\hat{\bar{\omega}}_{k_j^\beta}^n\| + Cu_j \,. \tag{139}$$

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 $j \in U(k_n, k_{n+1})$, a.s,

$$\begin{aligned}
\|\hat{\bar{\omega}}_{k_{j+1}^\beta}^n\| &\leq \left(\prod_{i \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} e^{Cu_i}\right) \|\hat{\bar{\omega}}_{k_n}^n\| + C \sum_{p \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} \left(\prod_{i \in \mathcal{U}(k_{p+1}^\beta, k_{j+1}^\beta - 1)} e^{Cu_i}\right) u_p \\
&= e^{C\sum_{l=k_n}^{k_{j+1}^\beta - 1} \xi_l} \|\hat{\bar{\omega}}_{k_n}^n\| + C \sum_{p \in \mathcal{U}(k_n, k_{j+1}^\beta - 1)} e^{C\sum_{l=k_{p+1}^\beta}^{k_{j+1}^\beta - 1} \xi_l} u_p \\
&\leq e^{CC_\xi T} + Ce^{CC_\xi T} C_\xi T \,,
\end{aligned}$$

where the last inequality comes from the facts that $\|\hat{\bar{\omega}}_{k_n}^n\|$ is bounded by 1 and that $\sum_{l=k_n}^{k_{n+1}} \xi_l \leq C_\xi T$. 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 $\|\hat{\omega}_k^n\|$.

(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.

□

Lemma C.8. $\lim_n \max_{k_n \leq k \leq k_{n+1}} \left\|(\hat{\omega}_k^n, \hat{\bar{\omega}}_k^n) - (\omega_k^n, \bar{\omega}_k^n)\right\| = 0 \,.$

Proof. Let $n \in \mathbb{N}$. Consider the shorthand notations $x_k^n := \hat{\omega}_k^n - \omega_k^n$ and $y_k^n := \hat{\bar{\omega}}_k^n - \bar{\omega}_k^n$ for $k \geq k_n^\xi$. Note that for every $k \geq k_n^\xi$, the sequences $(x_k^n)_k$ and $(y_k^n)_k$ satisfy the recurrence relations:

$$\begin{cases} x_{k+1}^n &= x_k^n + \beta_k(\gamma \Phi^T D_{\rho,\theta_k} P_{\theta_k} \Phi y_k^n - \bar{G}(\theta_k) x_k^n) + \beta_k \hat{\epsilon}_k^n \,, \\ y_{k+1}^n &= y_k^n + \xi_k(x_{k+1}^n - y_k^n) \,, \end{cases} \tag{140}$$

where the Markovian noise sequence $(\hat{\epsilon}_k^n)_k$ is defined for every $k \geq k_n^\xi$ by:

$$\hat{\epsilon}_k^n := \frac{1}{\max(1, \|r_{k_n^\xi}\|)} \left[\phi(\tilde{S}_k) R_{k+1} - h(\theta_k)\right] + \gamma \left[\phi(\tilde{S}_k)\phi(S_{k+1})^T - \Phi^T D_{\rho,\theta_k} P_{\theta_k} \Phi\right] \hat{\bar{\omega}}_k^n + \left[\bar{G}(\theta_k) - \phi(\tilde{S}_k)\phi(\tilde{S}_k)^T\right] \hat{\omega}_k^n \,. \tag{141}$$

It is clear that the sequence $(\hat{\epsilon}_k^n)$ 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:

$$x^*(y, \theta) := \gamma \bar{G}(\theta)^{-1} \Phi^T D_{\rho,\theta} P_\theta \Phi y \,. \tag{142}$$

Then, we have the following decomposition for every $k \geq k_n^\xi$:

$$\begin{aligned}
y_{k+1}^n &= y_k^n + \xi_k(x^*(y_k^n, \theta_k) - y_k^n) + \xi_k(x_{k+1}^n - x_k^n) + \xi_k(x_k^n - x^*(y_k^n, \theta_k)) \\
&= (I_m - \xi_k \bar{G}(\theta_k)^{-1} G(\theta_k)) y_k^n + \xi_k(x_{k+1}^n - x_k^n) + \xi_k(x_k^n - x^*(y_k^n, \theta_k)) \,.
\end{aligned}$$

Since $\bar{G}(\theta)^{-1}G(\theta)$ is uniformly (in θ) κ-positive definite (see Assumption 6.2), Lem. C.3 implies that there exists $\kappa > 0$ s.t. for sufficiently large n and $k \in \{k_n^\xi, \cdots, k_{n+1}^\xi - 1\}$,

$$\begin{aligned}\|y_{k+1}^n\| &\leq e^{-\frac{1}{2}\kappa T}\|y_{k_n^\xi}^n\| + \sum_{l=k_n^\xi}^{k} \xi_l \|x_{l+1}^n - x_l^n\| + \xi_l \|x_l^n - x^*(y_l^n, \theta_l)\| \\ &= \sum_{l=k_n^\xi}^{k} \xi_l \|x_{l+1}^n - x_l^n\| + \xi_l \|x_l^n - x^*(y_l^n, \theta_l)\| \\ &\leq \sum_{l=k_n^\xi}^{k} \xi_l \beta_l C + C_\xi T \max_{l \in \{k_n^\xi, \cdots, k\}} \|x_l^n - x^*(y_l^n, \theta_l)\|\,,\end{aligned}$$

where the equality comes from the fact that $y_{k_n^\xi}^n = 0$ by definition and the last inequality comes from the fact that $x_{l+1}^n - x_l^n = \beta_l(\gamma \Phi^T D_{\rho,\theta_l} P_{\theta_l} \Phi y_l^n - \bar{G}(\theta_l) x_l^n + \hat{\epsilon}_l^n)$ and the a.s. boundedness of the sequences $(x_k^n)_k$,$(y_k^n)_k$ and $(\hat{\epsilon}_k^n)_k$ resulting from Lem. C.7. Observe then that:

$$\sum_{l=k_n^\xi}^{k_{n+1}^\xi - 1} \xi_l \beta_l = \sum_{l=k_n^\xi}^{k_{n+1}^\xi - 1} \frac{\xi_l}{\beta_l} \beta_l^2 = \sum_{l=k_{j_n}^\beta}^{k_{j_{n+1}}^\beta - 1} \frac{\xi_l}{\beta_l} \beta_l^2 \leq \max_{k_{j_n}^\beta \leq l \leq k_{j_{n+1}}^\beta} \left(\frac{\xi_l}{\beta_l}\right) \sum_{l=k_{j_n}^\beta}^{+\infty} \beta_l^2\,. \tag{143}$$

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

$$\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \|y_k^n\| = 0, \quad a.s. \tag{144}$$

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:

$$x_{k+1}^n = \sum_{l=k_n^\xi}^{k} \left[\prod_{p=l+1}^{k} (I_m - \beta_p \bar{G}(\theta_p))\right] \beta_l \left(\gamma \Phi^T D_{\rho,\theta_l} P_{\theta_l} \Phi y_l^n + \hat{\epsilon}_l^n\right)\,.$$

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

$$\begin{aligned}\|x_{k+1}^n\| &\leq C \sum_{l=k_n^\xi}^{k} \left[\prod_{p=l+1}^{k} (1 - \frac{1}{2}\epsilon\beta_p)\right] \beta_l \|y_l^n\| + \left\| \sum_{l=k_n^\xi}^{k} \left[\prod_{p=l+1}^{k} (I_m - \beta_p \bar{G}(\theta_p))\right] \beta_l \hat{\epsilon}_l^n \right\| \\ &\leq C \frac{2}{\epsilon} \max_{k_n^\xi \leq l \leq k_{n+1}^\xi} \|y_l^n\| + \left\| \sum_{l=k_n^\xi}^{k} \left[\prod_{p=l+1}^{k} (I_m - \beta_p \bar{G}(\theta_p))\right] \beta_l \hat{\epsilon}_l^n \right\|\,.\end{aligned}$$

where the first inequality stems from the fact that the matrix $\bar{G}(\theta)$ 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:

$$\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \|x_k^n\| = 0, \quad a.s.\,,$$

which concludes the proof. □

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

$$\|\bar{\omega}_{k_{n+1}}^n\| \leq e^{-\frac{1}{2}\kappa T}\|\bar{\omega}_{k_n}^n\| + \delta_n + \frac{C}{\max(1, \|r_{k_n}\|)}\,.$$

Proof. Recall from Eq. (131) that $\tilde{\omega}_n^*(\bar{\omega},\theta) := \bar{G}(\theta)^{-1}\left(\frac{h(\theta)}{\max(1,\|r_{k_n}\|)} + \gamma\Phi^T D_{\rho,\theta}P_\theta\Phi\bar{\omega}\right)$ for every $n\in\mathbb{N},\bar{\omega}\in\mathbb{R}^m,\theta\in\mathbb{R}^d$. It is clear that for every $k\geq k_n$:

$$\bar{\omega}_{k+1}^n = \bar{\omega}_k^n + \xi_k(\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)-\bar{\omega}_k^n) + \xi_k(\omega_{k+1}^n-\omega_k^n) + \xi_k(\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k))\,.$$

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

$$\bar{\omega}_{k+1}^n = (I-\xi_k\bar{G}(\theta_k)^{-1}G(\theta_k))\bar{\omega}_k^n + \xi_k\frac{\bar{G}(\theta_k)^{-1}h(\theta_k)}{\max(1,\|r_{k_n}\|)} + \xi_k(\omega_{k+1}^n-\omega_k^n) + \xi_k(\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k))\,.$$

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

$$\|\bar{\omega}_{k+1}^n\| \leq (1-\frac{1}{2}\kappa\xi_k)\|\bar{\omega}_k^n\| + \xi_k\frac{\|\bar{G}(\theta_k)^{-1}h(\theta_k)\|}{\max(1,\|r_{k_n}\|)} + \xi_k\|\omega_{k+1}^n-\omega_k^n\| + \xi_k\|\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\|\,. \tag{145}$$

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

$$\|\omega_{k+1}^n-\omega_k^n\| = \beta_k\left\|\frac{h(\theta_k)}{\max(1,\|r_{k_n}\|)} + \gamma\Phi^T D_{\rho,\theta_k}P_{\theta_k}\Phi\bar{\omega}_k^n - \bar{G}(\theta_k)\omega_k^n\right\| \leq \beta_k C\,.$$

Therefore, for sufficiently large n,

$$\|\bar{\omega}_{k_{n+1}^\xi}^n\| \leq e^{-\frac{1}{2}\kappa T}\|\bar{\omega}_{k_n^\xi}^n\| + \frac{CC_\xi T}{\max(1,\|r_{k_n}\|)} + CC_\xi T\beta_{k_n} + \sum_{k=k_n^\xi}^{k_{n+1}^\xi}\xi_k\|\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\|\,.$$

It remains to show that $\sum_{k=k_n^\xi}^{k_{n+1}^\xi}\xi_k\|\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\|$ converges to 0 as $n\to\infty$. For this purpose, we adopt the same strategy used for studying the sequence $(\bar{\omega}_k^n)$. First, we write for every $k\geq k_n$,

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

Then, applying Lem. C.3, for sufficiently large n, there exists $\epsilon>0$ s.t. for every $k\in\{k_n,\cdots,k_{n+1}\}$,

$$\|\omega_{k+1}^n-\tilde{\omega}_n^*(\bar{\omega}_{k+1}^n,\theta_{k+1})\| \leq (1-\beta_k\frac{1}{2}\epsilon)\|\omega_k^n-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\| + \|\tilde{\omega}_n^*(\bar{\omega}_{k+1}^n,\theta_{k+1})-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\|\,.$$

We can show that, for every n, the function $(\bar{\omega},\theta)\mapsto\tilde{\omega}_n^*(\bar{\omega},\theta)$ 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' s.t. for every $k\in\{k_n^\xi,\cdots,k_{n+1}^\xi\}$,

$$\|\tilde{\omega}_n^*(\bar{\omega}_{k+1}^n,\theta_{k+1})-\tilde{\omega}_n^*(\bar{\omega}_k^n,\theta_k)\| \leq C\xi_k\|\omega_{k+1}^n-\bar{\omega}_k^n\| + C\alpha_k \leq C'\xi_k\,,$$

where the last inequality comes from the boundedness of the sequences ω_k^n and $\bar{\omega}_k^n$ for $k\in\{k_n^\xi,\cdots,k_{n+1}^\xi\}$, 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. $\|\omega_{k_n^\xi}^n-\tilde{\omega}_n^*(\bar{\omega}_{k_n^\xi}^n,\theta_{k_n^\xi})\|\leq C$, it is easy to check that

$$\|\omega_{k+1}^n-\tilde{\omega}_n^*(\bar{\omega}_{k+1}^n,\theta_{k+1})\| \leq e^{-\frac{1}{2}\epsilon\sum_{l=k_n^\xi}^k\beta_l}C + C'\sum_{l=k_n^\xi}^k e^{-\frac{1}{2}\epsilon\sum_{p=l+1}^k\beta_p}\xi_l\,.$$

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

$$\lim_{n\to\infty}\sum_{k=k_n^\xi}^{k_{n+1}^\xi}\xi_k\left(e^{-\frac{1}{2}\epsilon\sum_{l=k_n^\xi}^{k-1}\beta_l} + \sum_{l=k_n^\xi}^{k-1}e^{-\frac{1}{2}\epsilon\sum_{p=l+1}^{k-1}\beta_p}\xi_l\right) = 0\,.$$

The proof of this technical result is deferred to Lem. C.15 below. □

Theorem C.10. We have the following:

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

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,

$$\|\hat{\bar{\omega}}_{k_{n+1}}^n\| \leq e^{-\frac{1}{2}\kappa T}\|\hat{\bar{\omega}}_{k_n}^n\| + \hat{\delta}_n + \frac{C}{\max(1, \|r_{k_n}\|)}\,.$$

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

$$\max(1, \|r_{k_n}\|) \leq 1 + \|\omega_{k_n}\| + \|\bar{\omega}_{k_n}\| \leq (1 + C')(1 + \|\bar{\omega}_{k_n}\|)\,,$$

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

$$\|\bar{\omega}_{k_{n+1}}\| \leq (e^{-\frac{1}{2}\kappa T} + (1 + C')\hat{\delta}_n)\|\bar{\omega}_{k_n}\| + (1 + C')\hat{\delta}_n + C\,.$$

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 $(\hat{\bar{\omega}}_k^n)$, observe that:

$$\begin{aligned}\sup_k \|\bar{\omega}_k\| &= \sup_n \max_{k_n \leq k \leq k_{n+1}} \|\bar{\omega}_k\| \\ &= \sup_n \max_{k_n \leq k \leq k_{n+1}} \{\max(1, \|(\omega_{k_n}, \bar{\omega}_{k_n})\|) \cdot \|\hat{\bar{\omega}}_k^n\|\}\,. \end{aligned} \tag{146}$$

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:

$$\sup_k \|\bar{\omega}_k\| \leq \sup_n (1 + \|\omega_{k_n}\| + \|\bar{\omega}_{k_n}\|)(\max_{k_n \leq k \leq k_{n+1}} \|\bar{\omega}_k^n\| + \max_{k_n \leq k \leq k_{n+1}} \|\hat{\bar{\omega}}_k^n - \bar{\omega}_k^n\|)\,. \tag{147}$$

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

$$\sup_k \|\bar{\omega}_k\| \leq \sup_n \tilde{C}(1 + \|\bar{\omega}_{k_n}\|)(\max_{k_n \leq k \leq k_{n+1}} \|\bar{\omega}_k^n\| + \max_{k_n \leq k \leq k_{n+1}} \|\hat{\bar{\omega}}_k^n - \bar{\omega}_k^n\|)\,. \tag{148}$$

The result follows from the boundedness of the sequences $(\bar{\omega}_{k_n})$ (see the first item) and $(\bar{\omega}_k^n)$ (see the second item) and Lem. C.8.

□

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:

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

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

Proof. Recall that $x_{k_n^\xi}^n = y_{k_n^\xi}^n = 0$. 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 $k \geq k_n^\xi$ the recurrence relation:

$$v_{k+1}^n = (I_m - \beta_k \bar{G}(\theta_k))v_k^n + (x^*(y_k^n, \theta_k) - x^*(y_{k+1}^n, \theta_{k+1})) + \beta_k \epsilon_k^n\,.$$

Iterating this equality for $k \geq k_n^\xi$ and observing that $v_{k_n^\xi}^n = 0$ leads to the identity:

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

It can be shown that the function $(\bar{\omega}, \theta) \mapsto x^*(\bar{\omega}, \theta)$ 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 $\bar{G}(\theta)$ is uniformly positive definite, applying Lem. C.3 yields the existence of $\epsilon > 0$ s.t. for sufficiently large n and for $k \geq k_n^\xi$,

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

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

$$\|v_{k+1}^n\| \leq LC' \sum_{p=k_n^\xi}^{k} e^{-\frac{1}{2}\epsilon \sum_{l=p+1}^{k} \beta_l} \xi_p + \left\| \sum_{p=k_n^\xi}^{k} \left[\prod_{l=p+1}^{k} (I_m - \beta_l \bar{G}(\theta_l)) \right] \beta_p \hat{\epsilon}_p^n \right\|. \tag{149}$$

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^ξ) that there exist $j_n, j_{n+1} \in \mathbb{N}$ s.t. $k_n^\xi = k_{j_n}^\beta$ and $k_{n+1}^\xi = k_{j_{n+1}}^\beta$. Observe also that for every $k \in \{k_n^\xi, \cdots, k_{n+1}^\xi - 1\}$, there exists $i_k \in \{j_n, \cdots, j_{n+1} - 1\}$ s.t. $k \in \{k_{i_k}^\beta, \cdots, k_{i_k+1}^\beta - 1\}$. Then, we can rewrite the first term in the above inequality as follows:

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

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

$$\sum_{i=j_n}^{i_k - 1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} e^{-\frac{1}{2}\epsilon \sum_{l=p+1}^{k} \beta_l} \xi_p \leq \sum_{i=j_n}^{i_k - 1} e^{-\frac{1}{2}\epsilon T(i_k - 1 - i)} \sum_{p=k_i^\beta}^{k_{i+1}^\beta - 1} \xi_p \leq \frac{C_\xi T}{1 - e^{-\frac{1}{2}\epsilon T}} \max_{k_n^\xi \leq p \leq k_{n+1}^\xi} \frac{\xi_p}{\beta_p} .$$

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

$$\sum_{p=k_n^\xi}^{k} e^{-\frac{1}{2}\epsilon \sum_{l=p+1}^{k} \beta_l} \xi_p \leq C \max_{k_n^\xi \leq p \leq k_{n+1}^\xi} \frac{\xi_p}{\beta_p} .$$

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.,

$$\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \sum_{p=k_n^\xi}^{k} e^{-\frac{1}{2}\epsilon \sum_{l=p+1}^{k} \beta_l} \xi_p = 0 .$$

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. $\lim_{n \to \infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \left\| \sum_{p=k_n^\xi}^{k} \left[\prod_{l=p+1}^{k} (I_m - \beta_l \bar{G}(\theta_l)) \right] \beta_p \hat{\epsilon}_p^n \right\| = 0, a.s.$

Proof. Let $\bar{G}_{p+1:k} := \prod_{l=p+1}^{k}(I_m - \beta_l \bar{G}(\theta_l))$ for every $p \in \{k_n^\xi, \cdots, k_{n+1}^\xi\}$ and $p \leq k-1$. As in the proof of Lem. C.11, we begin by the observation that there exist j_n and j_{n+1} s.t. $k_{j_n}^\beta = k_n^\xi, k_{j_{n+1}}^\beta = k_{n+1}^\xi$ and that for every $k \in \{k_n^\xi, \cdots, k_{n+1}^\xi - 1\}$, there exists $i_k \in \{j_n, \cdots, j_{n+1}-1\}$ s.t. $k \in \{k_{i_k}^\beta, \cdots, k_{i_k+1}^\beta - 1\}$. Then, we can write for every $k \in \{k_n^\xi, \cdots, k_{n+1}^\xi - 1\}$,

$$\left\| \sum_{p=k_n^\xi}^{k} \bar{G}_{p+1:k}\beta_p \hat{\epsilon}_p^n \right\| \leq \left\| \sum_{i=j_n}^{i_k-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta} \bar{G}_{p+1:k}\beta_p \hat{\epsilon}_p^n \right\| + \left\| \sum_{p=k_{i_k}^\beta}^{k} \bar{G}_{p+1:k}\beta_p \hat{\epsilon}_p^n \right\| .$$

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 $k \in \{k_n^\xi, \cdots, k_{n+1}^\xi - 1\}$,

$$\left\| \sum_{i=j_n}^{i_k-1} \sum_{p=k_i^\beta}^{k_{i+1}^\beta} \bar{G}_{p+1:k}\beta_p \hat{\epsilon}_p^n \right\| \leq \sum_{i=j_n}^{i_k-1} \left\| \bar{G}_{k_{i+1}^\beta:k} \right\| \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} \bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p \hat{\epsilon}_p^n \right\| . \tag{150}$$

Lem C.3 implies that for sufficiently large n, for $k \in \{k_{i_k}^\beta, \cdots, k_{i_k+1}^\beta\} \subset \{k_n^\xi, \cdots, k_{n+1}^\xi\}$ and $i \in \{j_n, \cdots, i_k - 1\}$

$$\left\| \bar{G}_{k_{i+1}^\beta:k} \right\| \leq e^{-\frac{1}{2}\epsilon \sum_{l=k_{i+1}^\beta}^{k} \beta_l} \leq e^{-\frac{1}{2}\epsilon T(i_k - 1 - i)} . \tag{151}$$

Recall now from Eq. (141) the definition of $\hat{\epsilon}_p^n$ for $p \geq k_n^\xi$,

$$\hat{\epsilon}_p^n := \frac{1}{\max(1, \|r_{k_n^\xi}\|)} \left[\phi(\tilde{S}_p) R_{p+1} - h(\theta_p) \right] + \gamma \left[\phi(\tilde{S}_p)\phi(S_{p+1})^T - \Phi^T D_{\rho,\theta_p} P_{\theta_p} \Phi \right] \hat{\omega}_p^n + \left[\bar{G}(\theta_p) - \phi(\tilde{S}_p)\phi(\tilde{S}_p)^T \right] \hat{\omega}_p^n .$$

In the following, we control this Markovian noise using the decomposition technique of [Benveniste et al., 1990] which was also used in [Konda and Tsitsiklis, 2003a]. We use similar notations to those of the proof of [Konda and Tsitsiklis, 2003a, Lem. 8]. Define the Markov chain $Y_{p+1} := (\tilde{S}_p, \tilde{A}_p)$. The perturbation $\hat{\epsilon}_p^n$ is of the form

$$F_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1}) - \bar{F}_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n) + M_{p+1}^{(1)} \hat{\bar{\omega}}_p^n + M_{p+1}^{(2)} ,$$

where $\bar{F}_\theta(\omega, \bar{\omega})$ is the steady state expectation of $F_\theta(\omega, \bar{\omega}, (\bar{S}_p, \bar{A}_p))$, where $\bar{S}_p$ is a Markov chain with transition kernel P_θ, and where $M_{p+1}^{(i)}$ for $i = 1, 2$ are martingale difference sequences. For every $\theta \in \mathbb{R}^d, \omega, \bar{\omega} \in \mathbb{R}^m$, there exists a solution $\hat{F}_\theta(\omega, \bar{\omega})$ to the so-called Poisson equation:

$$F_\theta(\omega, \bar{\omega}, y) - \bar{F}_\theta(\omega, \bar{\omega}) = \hat{F}_\theta(\omega, \bar{\omega}, y) - (P_\theta \hat{F}_\theta)(\omega, \bar{\omega}, y) .$$

Using this equation, the perturbation can be decomposed as follows for any fixed $n \in \mathbb{N}$ and $p \geq k_n$,

$$\begin{aligned}
\hat{\epsilon}_p^n &= M_{p+1}^{(1)} \hat{\bar{\omega}}_p^n + M_{p+1}^{(2)} + F_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1}) - \bar{F}_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n) \\
&= M_{p+1}^{(1)} \hat{\bar{\omega}}_p^n + M_{p+1}^{(2)} + \hat{F}_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1}) - (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1}) \\
&= (M_{p+1}^{(1)} \hat{\bar{\omega}}_p^n + M_{p+1}^{(2)} + (\hat{F}_{\theta_p}(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1})) - (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_p)) && (152) \\
&\quad + ((P_{\theta_{p-1}}\hat{F}_{\theta_{p-1}})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_{p-1}^n, Y_p) - (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_{p+1})) && (153) \\
&\quad + (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_p^n, \hat{\bar{\omega}}_p^n, Y_p) - (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_p^n, Y_p) && (154) \\
&\quad + (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_p^n, Y_p) - (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_{p-1}^n, Y_p) && (155) \\
&\quad + (P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_{p-1}^n, Y_p) - (P_{\theta_{p-1}}\hat{F}_{\theta_{p-1}})(\hat{\omega}_{p-1}^n, \hat{\bar{\omega}}_{p-1}^n, Y_p) . && (156)
\end{aligned}$$

Eqs. (150),(151) and (156) imply that the proof is complete if we show that:

$$\lim_{n\to\infty} \max_{k_n^\xi \leq k \leq k_{n+1}^\xi} \sum_{i=j_n}^{i_k-1} e^{-\frac{1}{2}\epsilon T(i_k - 1 - i)} \left\| \sum_{p=k_i^\beta}^{k_{i+1}^\beta-1} \bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p \hat{\epsilon}_p^n \right\| = 0 , \quad a.s.$$

For this, it is sufficient to prove the following inequality:

$$\mathbb{E}\left[\max_{j_n\le i\le j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p\hat{\epsilon}_p^n\right\|^2\right]\le C\sum_{p=k_n^\xi}^{k_{n+1}^\xi-1}\beta_p^2\,. \tag{157}$$

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

$$\mathbb{P}\left(\max_{j_n\le i\le j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p\hat{\epsilon}_p^n\right\|\ge\delta\right)\le\frac{C}{\delta^2}\sum_{p=k_n^\xi}^{k_{n+1}^\xi-1}\beta_p^2\,,$$

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

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

To prove that Ineq. 157 holds, it is sufficient to show that the desired inequality holds when $\hat{\epsilon}_p^n$ 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 $n\in\mathbb{N}$,

$$\mathbb{E}\left[\max_{j_n\le i\le j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_pZ_{p+1}^n\right\|^2\right]\le C\sum_{p=k_n^\xi}^{k_{n+1}^\xi-1}\beta_p^2\,,$$

where for every $p\ge k_n$, $Z_{p+1}^n:=M_{p+1}^{(1)}\hat{\omega}_p^n+M_{p+1}^{(2)}+(\hat{F}_{\theta_p}(\hat{\omega}_p^n,\hat{\omega}_p^n,Y_{p+1})-(P_{\theta_p}\hat{F}_{\theta_p})(\hat{\omega}_p^n,\hat{\omega}_p^n,Y_p))\,.$

Proof. In this proof, we suppress the superscript n of Z_{p+1}^n to simplify notation. Note that n is fixed throughout the proof. Define $M_{k_i^\beta}^k:=\sum_{l=k_i^\beta}^k\beta_lZ_{l+1}$ for every $i\in\mathbb{N},k>k_i^\beta$. This is a zero mean, square integrable martingale for $k\in\{k_n^\xi+1,\cdots,k_{n+1}^\xi\}$. By summation by part, we have for every $j_n\le i\le j_{n+1}-1$,

$$\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_pZ_p=M_{k_i^\beta}^{k_{i+1}^\beta-1}-\sum_{p=k_i^\beta}^{k_{i+1}^\beta-2}(\bar{G}_{p+1:k_{i+1}^\beta-1}-\bar{G}_{p:k_{i+1}^\beta-1})M_{k_i^\beta}^p\,.$$

Notice that $\bar{G}_{p+1:k_{i+1}^\beta-1}-\bar{G}_{p:k_{i+1}^\beta-1}=\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}\bar{G}(\theta_p)$. Hence, bounding the max by the sum, we obtain the following inequality:

$$\max_{j_n\le i\le j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_pZ_{p+1}\right\|^2\le2\sum_{i=j_n}^{j_{n+1}-1}\left\|M_{k_i^\beta}^{k_{i+1}^\beta-1}\right\|^2+2\sum_{i=j_n}^{j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}\bar{G}(\theta_p)M_{k_i^\beta}^p\right\|^2 \tag{158}$$

We have that $\sup_{\theta\in\mathbb{R}^d}\|\bar{G}(\theta)\|<\infty$. Moreover, using Lem. C.3 , one can show that there exists $C>0$ s.t. for every integers $q>p$, $\|\bar{G}_{p:q}\|\le C$. Thus, we obtain the following upper bound using the triangle inequality:

$$\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}\bar{G}(\theta_p)M_{k_i^\beta}^p\right\|^2\le C^2\left(\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\beta_p\left\|M_{k_i^\beta}^p\right\|\right)^2\le C^2T'^2\left(\max_{k_i^\beta\le p\le k_{i+1}^\beta-1}\|M_{k_i^\beta}^p\|\right)^2\,.$$

Taking the expectation in Eq. (158) and using Doob's inequality yields:

$$\mathbb{E}\left[\max_{j_n\le i\le j_{n+1}-1}\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p Z_{p+1}\right\|^2\right] \le \left(2+8C^2T'^2\right)\sum_{i=j_n}^{j_{n+1}-1}\mathbb{E}\left[\left\|M_{k_i^\beta}^{k_{i+1}^\beta-1}\right\|^2\right]$$
$$\le \left(2+8C^2T'^2\right)C_Z\sum_{i=j_n}^{j_{n+1}-1}\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\beta_p^2$$
$$= \left(2+8C^2T'^2\right)C_Z\sum_{p=k_n^\xi}^{k_{n+1}^\xi-1}\beta_p^2\,,$$

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.:

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

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

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

Notice that $\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}-\beta_{p-1}\bar{G}_{p:k_{i+1}^\beta-1} = (\beta_p-\beta_{p-1})\bar{G}_{p+1:k_{i+1}^\beta-1}+\beta_{p-1}\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}\bar{G}(\theta_p)$. Then, similarly to the proof of the previous lemma, recall that $\sup_{\theta\in\mathbb{R}^d}\|\bar{G}(\theta)\|<\infty$ and that Lem. C.3 entails the existence of a constant $C_G>0$ s.t. for sufficiently large p and for every integers $q>p$, $\max(\|\bar{G}(\theta_p)\bar{G}_{p+1:q}\|,\|\bar{G}_{p:q}\|)\le C_G$. Using the previous remarks with Eq. (159) yields for $j_n\le i\le j_{n+1}-1$,

$$\left\|\sum_{p=k_i^\beta}^{k_{i+1}^\beta-1}\bar{G}_{p+1:k_{i+1}^\beta-1}\beta_p(X_{p+1}-X_p)\right\|^2 \le 4\beta_{k_{i+1}^\beta-1}^2\left\|X_{k_{i+1}^\beta}\right\|^2+4C_G^2\beta_{k_i^\beta}^2\left\|X_{k_i^\beta}\right\|^2$$
$$+4\left\|\sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1}(\beta_p-\beta_{p-1})\bar{G}_{p+1:k_{i+1}^\beta-1}X_p\right\|^2+4\left\|\sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1}\beta_{p-1}\beta_p\bar{G}_{p+1:k_{i+1}^\beta-1}\bar{G}(\theta_p)X_p\right\|^2\,. \quad (160)$$

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:

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

where the constant $C_X>0$ bounds the second moment of X_k (i.e., $\sup_k\mathbb{E}\|X_k\|^2\le C_X$) 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 $\|\bar{G}_{p:q}\|$ for $q > p$ yields for n sufficiently large and $j_n \leq i \leq j_{n+1} - 1$,

$$\left\| \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} (\beta_p - \beta_{p-1}) \bar{G}_{p+1:k_{i+1}^\beta-1} X_p \right\|^2 \leq C_G^2 \left(\sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} |\beta_{p-1} - \beta_p| \cdot \|X_p\| \right)^2 .$$

Then, it follows that:

$$\max_{j_n \leq i \leq j_{n+1}-1} \left\| \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} (\beta_p - \beta_{p-1}) \bar{G}_{p+1:k_{i+1}^\beta-1} X_p \right\|^2 \leq C_G^2 \sum_{i=j_n}^{j_{n+1}-1} \left(\sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} |\beta_{p-1} - \beta_p| \cdot \|X_p\| \right)^2 .$$

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

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

where $C' := C_G^2 C_X$. 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:

$$\max_{j_n \leq i \leq j_{n+1}} \left\| \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} \beta_{p-1}\beta_p \bar{G}_{p+1:k_{i+1}^\beta-1} \bar{G}(\theta_p) X_p \right\|^2 \leq C_G^2 \sum_{i=j_n}^{j_{n+1}-1} \sum_{p=k_i^\beta+1}^{k_{i+1}^\beta-1} \sum_{q=k_i^\beta+1}^{k_{i+1}^\beta-1} \beta_p \beta_{p-1} \beta_q \beta_{q-1} \|X_p\| \|X_q\| ,$$

and taking the expectation implies:

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

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

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

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