Convergent Distributed Actor-Critic Algorithm Based on Gradient Temporal Difference

Miloš S. Stanković, Marko Beko and Srdjan S. Stanković

***Abstract*— In this paper a new distributed off-policy Actor-Critic algorithm for reinforcement learning is proposed. It is composed of the Gradient Temporal Difference GTD(1) algorithm at the Critic stage, and a complementary consensus-based exact policy gradient algorithm at the Actor stage, derived from the global objective in the form of a sum of weighted local state-value functions. Weak convergence of the algorithm to the invariant set of a corresponding attached ODE is demonstrated under mild conditions. An experimental verification of the algorithm properties is presented, showing that the algorithm can represent an efficient tool for practice, enabling parallel execution and fusion of local exploration spaces.**

I. Introduction

Reinforcement learning (RL) has become a widely accepted tool for *decision making* in unknown and stochastic environments (see e.g. [1], [2]). The existing RL algorithms are often based on *function approximation*, reducing the learning problem to finding optimal parameter values of lower dimensions than the state space, using either *on-policy* or *off-policy* scenario [3]–[5]. The so-called *Actor-Critic* (AC) methods have appeared as a consistent response to the requirements for approximate dynamic programming optimization [6]–[9]. They are composed of a *value function estimator* (*Critic*) under the given policy, and a *policy function estimator* (*Actor*) aimed at improving the policy function parameters (see e.g. [6]–[8]).

In this paper, we focus on a *decentralized and distributed AC* methods in the form of *network of agents* communicating among themselves in real-time in order to achieve an agreement on the *optimal policy*. In general, distributed multi-agent algorithms are of great importance nowadays due to their fundamental role in diverse *signal processing and control problems*, especially within the scope of Cyber-Physical Systems (CPS), Internet of Things (IoT) and Wireless Sensor Networks (WSN) (see e.g. [5], [10]–[12] and references therein). In particular, distributed AC algorithms have been treated in [13]–[16], but within different settings. Compared to our approach, [13], [15] did not treat the case of strict information structure constraints, while in [14], [16] restrictive on-policy settings have been assumed. We start from a finite set of Markov Decision Processes (MDPs), assigned to the corresponding agents, characterized by finite state and action spaces, transitional probability functions and locally generated random rewards. We assume the *off-policy* scenario with strict information structure constraints, and propose a new AC algorithm in which each agent applies a local *behavior policy* and generates a *linear approximation* for a predefined *target policy* (Critic), and implements iteratively the *exact policy gradient algorithm*, explicitly derived from the algorithm implemented at the Critic stage, providing improved policy parameters (Actor). A *dynamic consensus scheme* is applied at the Actor stage, asymptotically providing agreement on the policy parameters [14], [16]. Using the gradient temporal-difference algorithm GTD(1) [2], [4] at the Critic stage, a new corresponding *distributed policy gradient* algorithm is obtained formulating the global objective as a *sum of weighted local state-value functions*. The whole proposed multi-agent algorithm is new, extending the approach in [9] related to the single-agent case. A rigorous convergence analysis is provided, proving, under a set of nonrestrictive conditions, that the proposed algorithm weakly converges to the invariant set of an asymptotic *ordinary differential equation (ODE)*. As a prerequisite, the Feller-Markov properties are proved for the new algorithm at the Actor stage. Illustrative experimental results demonstrate that the proposed algorithm represents an efficient tool for real problems in distributed intelligent signal processing and control.

The paper is organized as follows. Section II contains the problem formulation and definition of the main criteria. Section III is devoted to the Critic stage, while Section IV contains the formulation of the entire AC algorithm. Section V is devoted to the weak convergence proof of the algorithm, paying special attention to the trace variables. In Section VI illustrative simulation results are provided.

II. Problem Formulation

Consider a set of Markov Decision Processes $\text{MDP}^{(i)}$, $i = 1, \ldots, N$, characterized by common finite sets of states $\mathcal{S}$ and actions $\mathcal{A}$, probability $P(s'|s,a)$ (to move to state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ by applying action $a \in \mathcal{A}$) and the random rewards $R^i(s,a,s')$, characterized by the distribution $p(\cdot|s',a,s)$, with the expectation $r(s',a,s)$ [2].

Communications between the agents are represented by a *strongly connected digraph* $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ is the set of *nodes* (agents) and $\mathcal{E}$ the set of directed arcs representing inter-node communications. Let A_G be the

M. S. Stanković is with University Singidunum, Belgrade, Serbia; and Vlatacom Institute, Belgrade, Serbia; e-mail: milstank@gmail.com

M. Beko is with Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal; and COPELABS, Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal; e-mail: beko.marko@gmail.com

S. S. Stanković is with School of Electrical Engineering, University of Belgrade, Serbia; e-mail: stankovic@etf.rs

This research was supported by the Science Fund of the Republic of Serbia, Grant #6524745, AI-DECIDE, and by the Fundação para a Ciência e a Tecnologia under Project UIDB/04111/2020.

constant *adjacency matrix*, and $\mathcal{N}_i \subset \mathcal{N}$ the *in-neighborhood* of node i [11], [17], [18].

The agents *learn from data* acquired by *interacting with their local environments* and *communicating* among themselves. In the *on-policy* scenario, agent i at time t executes the action $a_t^i \sim \pi^i(\cdot|s_t^i)$, where $\pi^i : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the *target policy function*. As a consequence, the environment of agent i changes the state to s_{t+1}^i and produces a random reward R_{t+1}^i, $i = 1, \ldots, N$ [2]. The local *state value functions* under policy π^i are defined by

$$V^{\pi^i}(s) = E_{\pi^i}\left\{ R_{t+1}^i + \sum_{j=1}^{\infty} \prod_{k=1}^{j} \gamma_{t+k}^i R_{t+j+1}^i | s_t^i = s \right\}, \tag{1}$$

with the discount factors $\gamma_{t+k}^i = \gamma(s_{t+k}^i) \in [0,1]$, where $E_{\pi^i}\{\cdot\}$ is the expectation over data generated by the Markov chains induced by π^i in MDP$^{(i)}$, $i = 1, \ldots, N$. If $V^i = [V^i(s_1) \cdots V^i(s_{|\mathcal{S}|})]^T$, the Bellman operator is defined as $T^{(\pi^i)}V^i = r^{\pi^i} + P^{\pi^i}\Gamma V^i$, where $P^{\pi^i}(s'|s) = \sum_{a \in \mathcal{A}} \pi^i(a|s) P(s'|s,a)$ is the $|\mathcal{S}| \times |\mathcal{S}|$ *local state transition matrix*, Γ an $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrix with $\gamma(s)$ at the diagonal and r^{π^i} an $|\mathcal{S}|$-vector composed of $r^{\pi^i}(s)$, $s = s_1, \ldots, s_{|\mathcal{S}|}$, the *one stage expected rewards* [2], [19]. Under standard MDP assumptions [19], the solution of the equation $V^i = T^{(\pi^i)}V^i$ is unique and given by $V^i = V^{\pi^i}$.

According to the AC methodology [9], the first goal of every agent i is to generate, using the local MDP observations, an approximation of $V^{\pi^i}(s)$ by $V_{\theta^i}(s) = \theta^{iT}\varphi(s)$, which is linear in the *parameter vector* θ^i, where $\varphi(s) \in \mathcal{R}^d$ is a *feature vector*, satisfying $d << |\mathcal{S}|$. Policies π^i are also parameterized as $\pi^i = \pi_{w^i}$, where $w^i \in \mathcal{R}^n$ are *policy parameter vectors*, $n << |\mathcal{S}|$, so that the main goal is to is to learn w^i using local observations and inter-agent communications in order to achieve an overall *policy improvement*.

In this paper we assume the *off-policy* scenario, in which each agent interacts with its environment using its local *behavior policy* π_b^i. Introducing the *importance ratios* $\rho_t^i = \pi_{w^i}(a_t^i|s_t^i)/\pi_b^i(a_t^i|s_t^i)$, the following modification of (1) is obtained: $V^{\pi_{w^i}}(s) = E_i\{\rho_t^i[R_{t+1}^i + \gamma_{t+1}^i V^{\pi_{w^i}}(s_{t+1}^i)]|s_t^i = s\}$, where $E_i\{\cdot\}$ is the mathematical expectation w.r.t. $d_b^i(s)$, the stationary distribution of the local Markov chain induced in MDP$^{(i)}$ by π_b^i [19].

We adopt the following standard assumptions:

(A1) a) P^{π^i} is such that $I - P^{\pi^i}\Gamma$ is nonsingular; b) P^{π^i} is irreducible and such that $\forall s, s' \in \mathcal{S}$, $P_{ss'}^{\pi_b^i} = 0 \Rightarrow P_{ss'}^{\pi^i} = 0$; for all $w^i \in \mathcal{R}^n$, $i = 1, \ldots, N$.

Formally, we write $V^{\theta^i} = \Phi\theta^i$, where $\Phi \in R^{|\mathcal{S}| \times d}$ is a feature matrix whose s-th row is the feature vector $\varphi^T(s)$.

(A2) a) the column vectors of Φ are *linearly independent*, b) all the feature vectors $\varphi(s)$ are bounded and have a *unit feature value* of 1 as their d-th element [9].

We define the following *local criteria for policy evaluation*

$$J^i(w^i) = \sum_{s \in \mathcal{S}} d_b^i(s) V_{\theta^i}^i(s) = \theta^{iT} E_i\{\varphi_t^i\}, \tag{2}$$

where $\theta^i = \theta^i(w^i)$ and $\varphi_t^i = \varphi(s_t^i)$. Let

$$J(w^1, \ldots, w^N) = \sum_i c^i J^i(w_i), \tag{3}$$

where $c^i \in \mathcal{R}^+$ are a priori chosen weights, so that the overall goal is to find

$$\bar{w} = \mathrm{Argmax}_w \{ \sum_i \nabla_{w^i} J(w^1, \ldots, w^N) = 0 |_{w^1 = \cdots = w^N = w} \}.$$

III. Critic

In the Critic part the agents generate θ^i iterates independently, using locally available data, $i = 1, \ldots, N$. Although any type of temporal-difference algorithms can be applied, we follow an idea from [9] and reduce the choice to the GTD(1) algorithm, due to the possibility to construct its complementary algorithm at the Actor stage based on exact policy gradients derived from $J(w^1, \ldots, w^N)$.

A. Algorithm GTD(1)

Introducing the *bootstrapping parameters* λ^i, we come to the generalized Bellman operators $T^{(\pi^i, \lambda^i)}$ [20]. The local gradient TD-algorithms (GTD(λ)) are generated using the objective function $J_{GTD}^i(\theta^i) = \frac{1}{2}\|\Pi^i(T^{(\pi^i,\lambda^i)}V_{\theta^i}^i - V_{\theta^i}^i)\|_{d_b^i}^2$, where Π^i is the projection operator onto the approximation space $\mathcal{L}_\Phi$ w.r.t. the weighted Euclidian norm $\|\cdot\|_{d_b^i}$ (recall (A2)) [19]. For $\lambda^i = 1$, we have $T^{(\pi^i,1)}V^i = (I - P^{\pi^i}\Gamma)^{-1}r^{\pi^i}$ for any $V^i \in \mathcal{R}^{|\mathcal{S}|}$. The locally optimal parameter vectors $\bar{\theta}^i$ are solutions of the linear equations

$$\nabla_{\theta^i} J_{GTD}^i(\theta^i)|_{\lambda^i=1} = -C_{GTD}^i\theta^i + b_{GTD}^i = 0, \tag{4}$$

where $D_b^i = \mathrm{diag}\{d_b^i\}$, $C_{GTD}^i = \Phi^T D_b^i \Phi$ and $b_{GTD}^i = \Phi^T D_b^i (I - P^{\pi^i}\Gamma)^{-1} r^{\pi^i}$, $i = 1, \ldots, N$.

The statistical form of (4) is $E_i\{\rho_t^i \delta_t^i e_t^i\} = 0$, where $\delta_t^i = R_{t+1}^i + \gamma_{t+1}^i \theta_t^{iT}\varphi_{t+1}^i - \theta_t^{iT}\varphi_t^i$ represents the *temporal difference* and $e_t^i = \varphi_t^i + \gamma_t^i \rho_{t-1}^i e_{t-1}^i$ the *trace vector* ($e_{-1}^i = 0$), while θ_t^i denotes the estimate of θ^i at time t. The corresponding local parameter iterates, denoted as AlgC, are

$$\theta_{t+1}^i = \theta_t^i + \alpha_t^i \rho_t^i \delta_t^i e_t^i, \tag{5}$$

where $\alpha_t^i > 0$ is the step size (to be specified later).

IV. Actor

The algorithm for the Actor part will be derived below from N GTD(1) algorithms implemented in the Critic part. The resulting Actor part, consists of N corresponding *policy gradient algorithms* interconnected through a *dynamic consensus scheme* ensuring asymptotic agreement on the optimal policy parameters.

A. Policy Gradient

We start from $\nabla_{w^i} E_i\{\rho_t^i \delta_t^i e_t^i\} = 0$ and obtain an expression for $\nabla_{w^i}\theta^{iT}$ according to [9]. The local policy gradient obtained from (2) is given by

$$\begin{aligned} \nabla_{w^i} J^i(w^i) &= \nabla_{w^i}\theta^{iT} E_i\{\varphi_t^i\} \\ &= E_i\{\rho_t^i \delta_t^i [f_t^i \nabla_{w^i} \log \pi_{w^i}(a_t^i|s_t^i) + \nabla_{w^i} e_t^{iT} g^i]\}, \end{aligned} \tag{6}$$

where $h^i = (H^{iT})^{-1} E_i\{\varphi_t^i\}$, $H^i = E_i\{\rho_t^i(\varphi_t^i - \gamma^i \varphi_{t+1}^i) e_t^{iT}\}$ and $f_t^i = e_t^{iT} h^i$. We also have

$$E_i\{\rho_t^i(\varphi_t^i - \gamma_{t+1}^i \varphi_{t+1}^i)\hat{f}^i(s_t^i)\} = E_i\{\varphi_t^i\}, \tag{7}$$

where $\hat{f}^i(s) = E_i\{f_t^i | s_t^i = s\} = E_i\{e_t^i | s_t^i = s\}^T h^i$.

In general, implementation of the policy gradient (6) can be faced with problems of generating f_t^i and $(\nabla_{w^i} e_t^i) h^i$. However, when the Critic part uses GTD(1), the solution becomes elegant.

Lemma 1 ([9]): Under (A1) and (A2), for AlgC defined by (5), the solution of (7) is defined as $\hat{f}^i(s) = E\{f_t^i | s_t^i = s\}$, where $f_t^i = 1 + \gamma_t^i \rho_{t-1}^i f_{t-1}^i, \quad t \geq 0, \quad f_{-1}^i = 0$; also, the elements of g^i are zero except the d-th element which is equal to 1.

Then, we have the following statistical form for the policy gradient algorithm $\nabla_{w^i} J^i(w^i) = \lim_{t\to\infty} E_i\{\rho_t^i \delta_t^i \tilde{e}_t^i\} = 0$, where $\tilde{e}_t^i = f_t^i \nabla_{w^i} \log \pi_{w^i}(a_t^i | s_t^i) + \gamma_t^i \rho_{t-1}^i \tilde{e}_{t-1}^i$, $t \geq 0$, $\tilde{e}_{-1}^i = 0$.

B. Consensus Based Actor-Critic Algorithm

Using the above derivations, we propose the following novel distributed consensus-based AC algorithm consisting of: 1) *Critic part (AlgC)*, in the form of N independent recursions (5); and 2) *Actor part (AlgA)*, defined by the following iterates

$$\tilde{w}_t^i = w_t^i + \beta_t^i \rho_t^i \delta_t^i \tilde{e}_t^i, \quad w_{t+1}^i = \sum_{j\in\mathcal{N}_i} a_t^{ij} \tilde{w}_t^j, \tag{8}$$

where $A_t = [a_t^{ij}]$, $i, j = 1, \ldots, N$, $a_t^{ij} \geq 0$, is a row-stochastic matrix ($\forall t \geq 0$), with $a_t^{ij} = 0$ for all $(j, i) \notin \mathcal{N}$. AlgA from (8) is composed of two parts: a) update of w_t^i using the currently observed local trajectory tuple, and b) convexification of the estimates from the node neighborhood, aimed at achieving convergence to consensus, when $\bar{w}^1 = ... = \bar{w}^N = \bar{w}$. The step size sequence $\{\beta_t^i\}$ satisfies $\beta_t^i << \alpha_t^i$, $\forall t \geq 0$, implying two different time scales (see e.g. [21]).

V. Convergence Analysis

A. AlgC

We first recapitulate some results on the convergence of GTD(λ)-based algorithms [4], [19].

According to [19], [22], $\{Z_t^i\} = \{(s_t^i, a_t^i, e_t^i)\}$, $t \geq 0$, forms a weak Feller Markov chain on $S \times A \times R^d$, bounded in probability, and having a unique probability measure ζ^i. For each Z_0^i the sequence of the averages $\frac{1}{t}\sum_{k=0}^{t-1} f(Z_k^i)$ converges in mean and a.s. to $E_{\zeta^i}\{f(Z_0^i)\}$ for any continuous function f. Let $g^i(\theta^i, \xi^i) = \rho^i(s,a)\bar{\delta}(\theta^i, s, a, s') e^i$, where $\xi^i = (s, a, s', e^i)$, $\bar{\delta}(\theta^i, s, a, s') = r(s, a, s') + \gamma(s')\varphi^T(s')\theta^i - \varphi^T(s)\theta^i$ and $r(s, a, s')$ denotes the one-step expected reward. Notice that $\delta_t^i = \bar{\delta}(\theta_t^i, s_t^i, a_t^i, s_{t+1}^i) + \omega_{t+1}^i$, where ω_{t+1}^i is a zero mean sequence modeling randomness in R_{t+1}^i. Following [19], we introduce $\bar{g}^i(\theta^i) = E_{\zeta^i}\{g^i(\theta^i, \xi_0^i)\}$, and obtain $\bar{g}^i(\theta^i) = -C_{GTD}^i \theta^i + b_{GTD}^i$, in accordance with (4).

(A3) Matrix C_{GTD}^i is nonsingular, $i = 1, \ldots, N$.

Theorem 1 ([23]): Let (A1)–(A3) hold. For each constant $\alpha^i > 0$ let $\{\theta_t^i\}$ be generated by AlgC from any e_{-1}^i. Let $\{k_{\alpha^i}\}$ be a sequence of nonnegative integers such that $\alpha^i k_{\alpha^i} \to \infty$ as $\alpha^i \to 0$. Then there exists a sequence T_{α^i} with $T_{\alpha^i} \to \infty$ as $\alpha^i \to 0$, such that for any $\delta^i > 0$,

$$\lim_{\alpha^i\to 0} \sup P(\theta_t^i \notin \mathcal{N}_{\delta^i}(\bar{\theta}^i), \text{some } t \in [k_{\alpha^i}, k_{\alpha^i} + T_{\alpha^i}/\alpha^i]) = 0, \tag{9}$$

where $\mathcal{N}_{\delta^i}(\cdot)$ denotes the δ^i-neighborhood of an indicated set and $\bar{\theta}^i = (C_{GTD}^i)^{-1} b_{GTD}^i$, according to (4).

B. AlgA

At the network level, we introduce $W_t = [w_t^{1T} \cdots w_t^{NT}]^T$ and $W_t' = [w_t'^{1T} \cdots w_t'^{NT}]^T$ and obtain the following global model:

$$W_t' = W_t + \beta_t \tilde{F}_t(W_t), \qquad W_{t+1} = (A_t \otimes I_p) W_t', \tag{10}$$

where $\beta_t^i = \beta_t$, $\tilde{F}_t(W_t) = [\tilde{g}_t^{1T} \cdots \tilde{g}_t^{NT}]^T$ and $\otimes$ denotes the Kronecker's product.

1) Consensus Part: Following [17], [24], we define $\Psi(t|k) = A_t \cdots A_k$ for $t \geq k$, $\Psi(t|t+1) = I_N$. Let $\tilde{\mathcal{F}}_t$ be an increasing sequence of σ-algebras such that $\tilde{\mathcal{F}}_t$ measures both $\{W_k, k \leq t, A_k, k < t\}$.

(A4) There is a scalar $\varsigma_0 > 0$ such that $a_t^{ii} \geq \varsigma_0$, and, for $i \neq j$, either $a_t^{ij} = 0$ or $a_t^{ij} \geq \varsigma_0$.

(A5) Graph $\mathcal{G}$ is strongly connected.

(A6) There are a scalar $p_0 > 0$ and an integer t_0 such that $P_{\tilde{\mathcal{F}}_t}\{\text{agent } j \text{ communicates to agent } i \text{ on the interval } [t, t+t_0]\} \geq p_0$, for all t and i, j for which $A_{\mathcal{G}}^{ij} \neq 0$.

(A7) There is a $N \times N$ matrix $\bar{\Psi}$ such that $E\{|E_{\tilde{\mathcal{F}}_k}\{\Psi_t\} - \bar{\Psi}|\} \to 0$ as $|t-k| \to \infty$, which, according to Lemma 2, has the form $\bar{\Psi} = [\hat{\Psi}^T \cdots \hat{\Psi}^T]^T$, where $\hat{\Psi} = [\bar{\psi}_1 \cdots \bar{\psi}_N]^T$.

(A8) Sequence $\{A_t\}$ is independent of the processes in MDP$^{(i)}$, $i = 1, \ldots, N$.

2) Properties of the Trace Variables: Introduce:

(A9) For every $(s, a) \in \mathcal{S} \times \mathcal{A}$ the mappings $w^i \longmapsto \pi_{w^i}$ are twice differentiable, $i = 1, \ldots, N$.

(A10) $\sup_{w^i, s^i, a^i} \|\nabla \log \pi_{w^i}(a^i | s^i)\| < \infty$ and $\nabla \log \pi_{w^i}(a^i | s^i)$ has a bounded derivative $\forall (s^i, a^i) \in \mathcal{S} \times \mathcal{A}$.

Lemma 2: Let $\tilde{x}_t^i = [\tilde{e}_t^{iT}, f_t^i]^T$. Then, for any given $\tilde{x}_{-1}^i$: a) $\sup_{t\geq 0} E\{\|\tilde{x}_t^i\|\} < \infty$, and b) the zero input response of $\tilde{x}_t^i$ tends to zero a.s.

Proof: The parameter iterates give rise to the following model

$$\tilde{x}_t^i = \gamma_t^i \rho_{t-1}^i \tilde{S}_t^i \tilde{x}_{t-1}^i + \tilde{G}_t^i, \tag{11}$$

where $\tilde{S}_t^i = \begin{bmatrix} I & \vdots & \phi_t^i \\ \cdots & \cdots & \cdots \\ 0 & \vdots & 1 \end{bmatrix}$, $\tilde{G}_t^i = \begin{bmatrix} \phi_t^i \\ \cdots \\ 1 \end{bmatrix}$ and $\phi_t^i = \nabla_{w^i} \log \pi_{w^i}(a_t^i | s_t^i)$. It is straightforward to verify using (11) that $E\{\prod_{j=k}^{t-1} \|\tilde{S}_j^i\|\} \leq L_1^i (t-k)$, $0 < L_1^i < \infty$, as $\sup_t \|\phi_t^i\| < \infty$; also, using [19, Lemmas A.1 and A.2, Proposition A.1], we find out that for $|t-k|$ large enough $E\{E\{\prod_{j=k}^{t-1} \gamma_{j+1}^i \rho_j^i | \mathcal{F}_k\}\} \leq L_2^i (\sigma^i)^{t-k}$, where $\mathcal{F}_k$ is the σ-algebra generated by the states up to time t and actions up to time $t-1$, and $0 < L_2^i < \infty$ and $|\sigma^i| < 1$. Therefore, $E\{\|\tilde{x}_t^i\|\} \leq L_3 t(\sigma^i)^t + L_4 \sum_{k=1}^t (t-k)(\sigma^i)^{t-k} < \infty$ ($0 < L_3$, $L_4 < \infty$).

Define

$$\Delta_t^i = t\prod_{j=1}^{t}\gamma_j^i\rho_{j-1}^i = \frac{t}{t-\mu}\prod_{j=t-\mu+1}^{t}(\gamma_j^i\rho_{j-1}^i)\Delta_{t-\mu}^i, \quad (12)$$

where μ is a positive integer since $\prod_{j=1}^{t}\|\tilde{S}_j^i\| \sim t$. Assume that μ is chosen in such a way that $E\{\prod_{j=t-\mu}^{t}(\gamma_j^i\rho_{j-1}^i)|F_{t-\mu}\} = ((P^{\pi^i}\Gamma)^\mu\mathbf{1})(s_{t-\mu}^i) < 1$ for all $s_k^i \in \mathcal{S}$. Then, $\exists t_0 > 0$ such that for all $t > t_0$ $\frac{t}{t-\mu}E\{\prod_{j=t-\mu}^{t}(\gamma_j^i\rho_{j-1}^i)|F_{t-\mu}\} \leq 1$, so that the convergence theorem for nonnegative supermartingales can be applied, i.e. $\Delta_t^i \to \Delta_\infty^i$ a.s. As $E\{\Delta_t^i\} \to_{t\to\infty} 0$, we conclude that $\Delta_\infty^i = 0$ a.s. Hence, the result follows. ■

From Lemma 2 and the results from [22, Subsection 3.1] it follows that $\{\tilde{Z}_t^i\}$ is a Feller Markov chain bounded in probability and having a unique probability measure $\tilde{\zeta}^i$, and that for each $\tilde{Z}_0^i$ the sequence $\frac{1}{t}\sum_{k=0}^{t-1} f(\tilde{Z}_k^i)$ converges in mean and a.s. to $E_{\tilde{\zeta}^i}\{f(\tilde{Z}_0^i)\}$ for any continuous f.

3) Convergence Proof: Let $\tilde{g}^i(w^i, \bar{\theta}^i(w^i), \tilde{\xi}^i) = \rho^i(s,a)\tilde{\delta}(\bar{\theta}(w^i), s, a, s')\tilde{e}^i(w^i)$, where $\tilde{\xi}^i = (s, a, s', \tilde{x}^i)$. We also have $\tilde{g}_t^i = \tilde{g}^i(w_t^i, \bar{\theta}^i(w_t^i), \tilde{\xi}_t^i)$, $\bar{\tilde{g}}^i(w^i, \bar{\theta}^i(w^i)) = E_{\tilde{\zeta}^i}\{\tilde{g}^i(w^i, \bar{\theta}^i(w^i), \tilde{\xi}_0^i)\}$, and let $\bar{\tilde{F}}_t(W) = [\bar{\tilde{g}}^1(w^1, \theta^1(w^1))^T \cdots \bar{\tilde{g}}^N(w^N, \theta^N(w^N))^T]^T$.

Theorem 2: Let (A1)–(A10) hold. Let W_t be generated by (8) with $\beta_t = \beta > 0$. Then, $W^\beta(\tau) = W_t$ for $\tau \in [t\beta, (t+1)\beta)$, $\tau \in R^+$, is tight and converges weakly when $\beta \to 0$ to $W(\tau) = [w(\tau)^T \cdots w(\tau)^T]^T$, $\tau \in \mathcal{R}^+$, where $w(\tau)$ is generated by

$$\dot{w} = \sum_{i=1}^{N}\bar{\psi}^i c^i \bar{\tilde{g}}^i(w, \bar{\theta}^i(w)), \quad w(0) = w_0, \quad (13)$$

and $\bar{\theta}^i(w)$ is a unique solution of (4). There exists a sequence T_β with $T_\beta \to \infty$ as $\beta \to 0$, such that for any $\delta > 0$,

$$\lim_{\beta\to 0}\sup P(w_t^i \notin \mathcal{N}_\delta(\bar{\mathcal{W}}), \text{some } t \in [k_\beta, k_\beta + T_\beta/\beta]) = 0, \quad (14)$$

where $\bar{\mathcal{W}}$ is the closure of the set of points $\bar{w}$ defined by $\sum_{i=1}^{N}\bar{\psi}^i c^i \bar{\tilde{g}}^i(\bar{w}, \bar{\theta}^i(\bar{w})) = 0$.

Proof Sketch: The proof follows the general line of reasoning from [24], [25], including additional technical aspects from [11], [22], [23] and Lemma 2. The essential part is the verification that the general assumptions C(3.2) and C(3.3') from [24] and A.8.1.11 from [25] hold. According to [24], the first part of the proof shows that $W^\beta(\cdot)$ is tight. In the second part, the asymptotic mean ODE (13) is derived using the arguments from [11], [24], [25]. The third part is directly related to the convergence points, in accordance with [11] ■

VI. Simulation Results

The simulated MDP environment is assumed to be a version of the Boyan's chain which can be used to model a travel decision making problem (see e.g. [4], [11]), whose diagram is shown in Fig. 1.

The discount factor is set to $\gamma = 0.9$. The (stationary) policy to be optimized is the probability of the exit action a^{exit} at state s: $\pi(a^{\text{exit}}|s)$. If a^{exit} is chosen, the probability

Fig. 1. Diagram of the simulated MDP.

of being stuck in a traffic jam is fixed to 0.2, with the reward $r(s, a^{\text{exit}}, s') = -2.5$ for all s and s', while if a^h is chosen, the reward is $r(s, a^h, s') = -1$ for all s and s', and the probability of being stuck grows as $1 - \frac{1}{s}$.

10 agents are simulated with a sparse neighbors-based communication graph. We used the linear function approximation for the critic based on 7-features Gaussian radial basis representation $\varphi_i(s) = e^{-\frac{(s-z_i)^2}{2\sigma^2}}$, $i = 1, ..., 7$, $z_i \in \{1, 3, 5, 7, 9, 11, 13\}$, $\sigma^2 = 2$. For the actor, we parameterize the policy $\pi_w(a|s)$ using the Gibbs parameterization $\pi_w(a|s) = \frac{e^{w^T\varphi_p(s,a)/\tau}}{\sum_{a'\in\mathcal{A}} e^{w^T\varphi_p(s,a')/\tau}}$, where $\varphi_p(s,a)$ is a d_p-dimensional feature vector and τ is the "temperature" parameter to be specified. Since the chain has an absorbing state we run the algorithms in multiple episodes.

In the first experiment, the agents are individually not able to find the optimal policy since we restrict their behavior such that they can visit only (complementary) subsets of the states (e.g. agent 1 starts in state 7 and stops in state 13, etc.). Their behavior policies are different with $\pi^i(a^{\text{exit}}|s)$ set to $[0.15, 0.24, 0.13, 0.38, 0.55, 0.89, 0.64, 0.97, 0.75, 0.69]$, respectively. In this experiment we choose tabular policy features $\varphi_p(s,a)$ with dimensionality $d_p = 15 \times 2$ (i.e. we don't loose any "information", so that the agents should be able to converge to the optimal policy). In Fig. 2 the evolution of the exact value function (exactly calculated in each time step) corresponding to the agents optimal policy estimates and averaged over all agents and states, for step sizes $\alpha_i = 0.02$ and $\beta_i = 0.0002$, and for $\tau = 1/16$, is shown. The red horizontal line represents the optimal value function. It can be concluded that the agents collectively successfully converge to the optimal policy despite the individual state-visiting restrictions and the critic approximation. This can also be concluded from Fig. 3 where we show the final value function approximations obtained by the critics, the exact value functions corresponding to the final policy estimates of each agent (which have converged to the same values due to the actor consensus), as well as the true optimal value function.

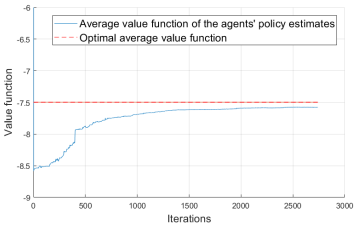

Fig. 2. Experiment 1. Evolution of the exact value function of the agents' optimal policy estimates averaged over all the agents and states. The red line is the optimal value.

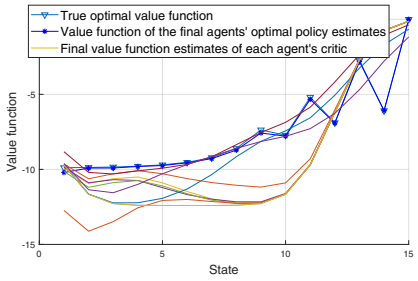

Fig. 3. Experiment 1. The exact value functions corresponding to the final optimal policy estimates obtained by the agents, together with the true optimal value function and the final value function approximations obtained by the critics.

In the second experiment, we assume that there are no restrictions on the agents' state-visiting possibilities, i.e. they are all capable to travel from the first to the last state with positive probability. The agents' behavior policies are assumed to be the same as above. However, here we choose policy features $\varphi_p(s,a)$ as binary codes (a unique code for each state) which lowers dimensionality to $d_p = 4 \times 2$ [7]; hence, we do not obtain convergence to the exact optimal policy as can be seen in Fig. 4, for $\alpha_i = 0.007$ and $\beta_i = 0.0001$, and for $\tau = 1/4$. In general, it has been confirmed in simulations that the rate of convergence and the variance of the estimates is improved using the proposed scheme, compared to the local agents working independently [11].

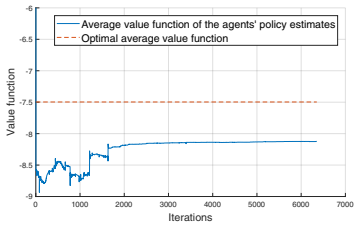

Fig. 4. Experiment 2. Evolution of the exact value function of the agents' optimal policy estimates averaged over all the agents and states. The red line is the optimal value.

VII. Conclusion

In this paper a new distributed off-policy AC algorithm has been proposed using the GTD(1) algorithm at the Critic stage, and a complementary exact consensus-based policy gradient algorithm derived from GTD(1) at the Actor stage, starting from a global objective in the form of a weighted sum of local state-value functions. After proving that the Feller-Markov properties hold for the derived algorithm at the Actor stage, a proof of the weak convergence of the entire algorithm to the invariant set of the attached ODE has been derived. Simulation results demonstrate that the algorithm can be an efficient tool for practice, enabling parallelization and complementarity of agent actions (off-policy setting).

Further efforts can be directed towards extensions to multi-task reinforcement learning problems.

References

[1] V. Mnih., K. Karavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmilelr, A. K. Fiedjeland, and O. O. *et al*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 1307, 2015.

[2] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 2017.

[3] M. Geist and B. Scherrer, "Off-policy learning with eligibility traces: A survey," *Journal of Machine Learning Research*, vol. 15, pp. 289–333, 2014.

[4] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proc. 26th Int. Conf. on Machine Learning*, 2009, pp. 993–1000.

[5] M. S. Stanković, M. Beko, and S. S. Stanković, "Distributed consensus-based multi-agent off-policy temporal-difference learning," in *Proc. 60th IEEE Conf. Decision and Control*, 2021, pp. 5976–5981.

[6] V. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, pp. 1143–14 166, 2003.

[7] S. Bhatnagar, R. S. Sutton, R. Ghavamzadeh, and M. Lee, "Natural actor-critic algorithms," *Automatica*, vol. 45, pp. 2471–2482, 2009.

[8] T. Degris, M. White, and R. S. Sutton, "Off policy actor critic," in *Proc. Int. Conf. Machine Learning*, 2012, pp. 179–186.

[9] H. R. Maei, "Convergent actor-critic algorithms under off-policy training and function approximation," *arXiv:1802.07842*, 2018.

[10] D. Lee, H. Yoon, and N. Hovakimyan, "Primal-dual algorithm for distributed reinforcement learning: Distributed GTD," in *IEEE Conf. Decision and Control*, 2018, pp. 1967–1972.

[11] M. S. Stanković, M. Beko, and S. S. Stanković, "Distributed value function approximation for collaborative multiagent reinforcement learning," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 3, pp. 1270–1280, 2021.

[12] S. S. Stanković, N. Ilić, and M. S. Stanković, "Adaptive consensus-based distributed system for multisensor multitarget tracking," *IEEE Transactions on Aerospace and Electronic Systems*, 2021.

[13] W. Suttle, Z. Yang, K. Zhang, Z. Wang, T. Basar, and J. Liu, "A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning," *arXiv:1908.03963*, 2019.

[14] P. Pennesi and I. Paschalidis, "A distributed actor-critic algorithm and applications to mobile sensor network coordination problems," *IEEE Trans. Autom. Control*, vol. 55, pp. 492–497, 2010.

[15] Y. Zhang and M. M. Zavlanos, "Distributed off-policy actor-critic reinforcement learning with policy consensus," *arXiv:1903.09255*, 2019.

[16] S. Valcarcel Macua, A. Tukiainen, D. Garcia-Ocana Hernandez, D. Baldazo, E. Munoz de Cote, and S. Zazo, "Diff-DAC: Distributed actor-critic for average multitask deep reinforcement learning," *arXiv 1710.10363*, 2019.

[17] M. S. Stanković, N. Ilić, and S. S. Stanković, "Distributed stochastic approximation: Weak convergence and network design," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4069–4074, 2016.

[18] M. S. Stanković, S. S. Stanković, and K. H. Johansson, "Asynchronous distributed blind calibration of sensor networks under noisy measurements," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 571–582, 2018.

[19] H. Yu, "On convergence of some gradient-based temporal-differences algorithms for off-policy learning," *arXiv:1712.09652*, 2017.

[20] H. Yu, A. Mahmood, and R. Sutton, "On generalized Bellman equations and temporal-difference learning," *Journal of Machine Learning Research*, vol. 19, pp. 1–49, 2019.

[21] V. S. Borkar, "Stochastic approximation with two time scales," *Systems & Control Letters*, vol. 29, no. 5, pp. 291–294, 1997.

[22] H. Yu, "On convergence of emphatic temporal-difference learning," ser. Proceedings of Machine Learning Research, vol. 40, 2015, pp. 1724–1751.

[23] ——, "Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize," *Journal of Machine Learning Research*, vol. 17, pp. 1–58, 2016.

[24] H. J. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM J. Control Optim.*, vol. 25, pp. 1266–1290, 1987.

[25] ——, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.