Consensus on the Auxiliary Variables in Distributed Gradient-Based Temporal Difference Algorithms

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

***Abstract*—In this paper we discuss important properties of two novel distributed algorithms for iterative multi-agent off-policy learning of linear value function approximation in Markov Decision Processes (MDP). The algorithms are derived using the off-policy Gradient Reinforcement Learning (GRL) methodology, together with linear dynamic consensus iterations over an underlying inter-agent communication network represented by directed graphs. The proposed algorithms are entirely decentralized, offering new possibilities for choosing different behavior policies while evaluating one single target policy. The presented algorithms formally differ only in the way of applying consensus iterations to the so-called auxiliary variables. The presented proof of weak convergence of both algorithms represents a firm basis for deriving relevant conclusions concerning the role of the consensus iterations. It is shown that the algorithm utilizing consensus on the auxiliary variables shows slightly inferior asymptotic properties, but can provide a higher convergence rate. The figure of merit of each of the algorithms is presented and discussed using the theoretical results obtained under generally nonrestrictive assumptions.**

I. Introduction

Decentralized multi-agent decision making algorithms have recently gained much popularity due to their high effectiveness in dealing with uncertain and dynamic environments typical for the emerging areas of Cyber-Physical Systems (CPS) and Internet of Things (IoT). Numerous distributed estimation, optimization and adaptation methods have been successfully developed using recursive collaborations aimed at achieving a consensus on variables of interest (e.g. [1]–[15] and references therein).

Reinforcement learning (RL) is a general methodology for decision making in uncertain environments based on models in the form of Markov Decision Process (MDP) based on utilization of approximate dynamic programming [16], [17]. A very important issue in this domain is the problem of approximation of the value function under very large state space and the presence of a discrepancy between the behavior policy of an agent and a policy that is currently targeted for evaluation (off-policy learning, e.g. [18]). Recently, in [19]–[22] several fast gradient-based algorithms for temporal-difference (TD) learning have been proposed. Distributed and multi-agent RL methods have become very popular very recently (see, e.g. [1], [23]–[25] and references therein). Different setups have been adopted in a number of recent works [26]–[32].

In this paper we shall present and discuss two distributed algorithms for *iterative multi-agent off-policy learning* of linear approximation of the value function in MDPs [1]. The algorithms represent generalizations of the recently proposed single agent off-policy gradient algorithm GTD2(λ) [1], [19]–[21], incorporating a distributed consensus scheme operating over a network of typically sparsely connected agents. Another important property of the algorithms is that the local recursions of each agent can be based on eligibility traces [20], [21], where each agent may choose different λ parameters. We provide a firm theoretical background in the form of a proof that the parameter estimates *weakly converge* to consensus points [1], [19], [28], [29], [31], [33], under nonrestrictive connectivity assumption on the topology of the underlying digraph and on the state-visiting distributions of the agents (their behavior policies). The main focus of this paper is placed on the dilemma whether or not to apply consensus to the *auxiliary variables* in the DGTD2(λ)-type algorithms with one-time-scale (see [19], [28], [31]). Notice that the paper [33] deals with the basically two-time-scale algorithms of DGTDC(λ)-type. The given analysis will be exclusively concerned with the limit points of the mean asymptotic ODEs: in this sense the behavior of the estimates for large, but finite t, including the derivation of the corresponding ODEs, can be found in [1], [33]. The limit sets are analyzed by formulating appropriate Lyapunov functions, following the line of thought of [21]. A discussion on role of convexification of the auxiliary variables is provided apart, showing that the two algorithms converge to the same limit points only in special cases. Application of consensus to the auxiliary variables causes, in principle, inferior asymptotic performance, having in mind that the implicitly imposed constraint increases the achievable estimation error. On the other hand, introduction of consensus can contribute to the overall convergence rate at the global level; however, the global convergence rate depends largely on the network connectivity.

The paper is organized as follows. In Section II we formulate the problem and define the algorithms. In Section III a rigorous weak convergence analysis is presented focused on the limit points, while Section IV is devoted to a general discussion on the application of consensus to the auxiliary variables in the DGTD2(λ) algorithms.

M. S. Stanković is with Singidunum University, 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 Faculty of Information Technology and Engineering, University Union Nikola Tesla, Belgrade, Serbia; e-mail: beko.marko@gmail.com.

N. Ilić is with College of Applied Technical Sciences, Kruševac, Serbia; and Vlatacom Institute, Belgrade, Serbia; e-mail: nemiliexp@yahoo.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.

 ISBN 978-86-7466-930-3

II. Distributed Gradient Temporal Difference Algorithms

A. Problem Formulation. Definition of the Algorithm

Consider *N autonomous agents* learning linear approximation to the *state value function* for a given policy in an MDP (denoted as $\mathrm{MDP}^{(0)}$), using observations of sample transitions in additional N *independent* MDPs, denoted as $\mathrm{MDP}^{(i)}$, $i = 1, \ldots, N$. Assume a finite state space $\mathcal{S} = \{1, \ldots, M\}$, and that $\mathrm{MDP}^{(0)}$ has the transition matrix P, and $\mathrm{MDP}^{(i)}$ the transition matrices $P^{(i)}$, $i = 1, \ldots, N$; these chains are induced by π and $\pi^{(i)}$, and referred to as the *target policy* and *behavior policies*, respectively. We are, therefore, dealing with a *cooperative off-policy learning* problem [1], [16], [18], [31].

We introduce the one-stage *reward function* $r_\pi : \mathcal{S} \to \mathcal{R}$, specifying the expected reward at each state $s \in \mathcal{S}$, where $\mathcal{R}$ is the set of real numbers [16], [21]. The associated discounted *total reward criterion* (*value function*), with the state dependent discount factors $\gamma(s) \in [0, 1]$, $s \in \mathcal{S}$, is given by

$$v_\pi(s) = E_s^\pi\{r_\pi(S_0) + \sum_{n=1}^{\infty} \gamma(S_1)\gamma(S_2)\cdots\gamma(S_n) \cdot r_\pi(S_n)\}, \tag{1}$$

where $E_s^\pi\{\cdot\}$ indicates the expectation w.r.t. to the Markov chain $\{S_n\}_{n>0}$ induced by π, with the initial state $S_0 = s$. Denote by Γ the $M \times M$ diagonal matrix with $\gamma(s)$ as diagonal entries and $v_\pi = [v_\pi(s_1) \cdots v_\pi(s_M)]^T$.

We assume the following [21]:

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

By the MDP theory [1], [16], [21], [34], v_π uniquely satisfies the *Bellman equation* $v_\pi = r_\pi + P\Gamma v_\pi$ (see e.g. [21], [34]). Within the framework of the *temporal-difference* (TD) algorithms, it is usual to consider the Bellman equation depending on the so-called *λ-parameters*, procedurally introduced by the so-called *eligibility traces*. In this sense, $v_\pi = T^{(\lambda)}v_\pi$ is considered as a *generalized Bellman equation*, where $T^{(\lambda)}v = r_\pi^\lambda + P^{(\lambda)}v$, $\forall v \in R^{|S|}$, is the *generalized Bellman operator* for a vector $r_\pi^{(\lambda)}$ and a substochastic matrix $P^{(\lambda)}$ [16], [21].

Let $\phi : S \to R^p$ be a function that maps each state to a p-dimensional feature vector $\phi = [\phi_1 \cdots \phi_p]^T$; let the subspace spanned by feature vectors ϕ be $\mathcal{L}_\phi$. In general, TD algorithms look for some function $v \in \mathcal{L}_\phi$ that satisfies $v \approx T^{(\lambda)}v$. We assume that the approximation functions are parameterized as $v(s) = \phi(s)^T\theta$, $s \in \mathcal{S}$ using parameters $\theta \in \mathcal{R}^p$, so that the algorithms learn the vector θ. If we define the $M \times p$ matrix Φ as a matrix composed of p-vectors $\phi(s)$ as row vectors, we have $v_\theta = \Phi\theta$.

In order to construct a distributed algorithm for finding an approximation function $v_\theta \in \mathcal{L}_\phi$ by using observations from $\mathrm{MDP}^{(i)}$, $i = 1, \ldots, N$, we define the following *global objective function*

$$J(\theta) = \sum_{i=1}^{N} q_i J_i(\theta) = \frac{1}{2}\sum_{i=1}^{N} q_i \|\Pi_{\xi_i}(T^{(\lambda_i)}v_\theta - v_\theta)\|_{\xi_i}^2, \tag{2}$$

where $J_i(\theta)$ are the *local objective functions*, $q_i > 0$ the *a priori* defined weighting coefficients, λ_i is the local λ-parameter and Π_{ξ_i} denotes the projection onto the subspace $\mathcal{L}_\phi$ w.r.t. the weighted Euclidean norm $\|v\|_{\xi_i}^2 = \sum_{s\in\mathcal{S}} \xi_{i;s} v(s)^2$ for a positive M-dimensional vector ξ_i with components $\xi_{i;s}$ (see [21], [31]). In accordance with [21], [34], we take ξ_i to be the invariant probability distribution for the local Markov chain $\mathrm{MDP}^{(i)}$, with the transition matrix $P^{(i)}$ induced by $\pi^{(i)}$ ($\xi_i^T P^{(i)} = \xi_i^T$). It follows that

$$\nabla J(\theta) = \sum_{i=1}^{N} q_i(\Phi^T\Xi_i(P^{(\lambda_i)} - I)\Phi)^T w_i(\theta), \tag{3}$$

where Ξ_i is an $M \times M$ diagonal matrix with the components of ξ_i on the diagonal, and $w_i(\theta)$ the unique solution (in w_i) of the equation

$$\Phi w_i = \Pi_{\xi_i}(T^{(\lambda_i)}v_\theta - v_\theta), \tag{4}$$

assuming that $w_i \in \mathrm{span}\{\phi(S)\}$.

In the off-policy scenario, we introduce the local *importance sampling ratios* $\rho_i(s, s') = P_{ss'}^i / P_{ss'}$ for $s, s' \in \mathcal{S}$, $i = 1, \ldots, N$; denote $\rho_i(n) = \rho_i(S_n, S_{n+1})$, as well as $\gamma(n) = \gamma(S_n)$ [21], [34]. The local *temporal-difference term* is defined by

$$\delta_i(v_\theta; n) = \rho_i(n)(R(n+1) + \gamma(n+1)v_\theta(S_{n+1}) - v_\theta(S_n)). \tag{5}$$

The local *eligibility trace vectors* $\{e_i(n)\}$ are generated by

$$e_i(n) = \lambda_i(n)\gamma(n)\rho_i(n-1)e_i(n-1) + \phi(S_n), \tag{6}$$

where $\lambda_i(n) \in [0, 1]$ are the local *λ-parameters*, $i = 1, \ldots, N$ [21], [34].

The distributed algorithms for learning linear approximation to the state value function for a given policy π we are going to analyze consist of two main parts: 1) *local parameter updates* based on the *gradient descent* methodology using local state transition observations from $\mathrm{MDP}^{(i)}$, and 2) interchange of the current parameter estimates aimed at achieving *consensus* between the agents. The local parameter updates are defined by

$$\begin{aligned}\theta_i'(n) =& \theta_i(n) + \alpha_i(n)q_i\rho_i(n)(\phi(S_n) \\ &- \gamma(n+1)\phi(S_{n+1}))e_i(n)^T w_i(n)\end{aligned} \tag{7}$$

$$\begin{aligned}w_i'(n) =& w_i(n) + \beta_i(n)(e_i(n)\delta_i(v_{\theta_i(n)}; n) \\ &- \phi(S_n)\phi(S_n)^T w_i(n))\end{aligned} \tag{8}$$

where $v_{\theta_i(n)} = \Phi\theta_i(n)$; $\theta_i(0)$ is chosen arbitrarily, while for $w_i(0)$ and $e_i(0)$ we have $w_i(0)$, $e_i(0) \in \mathrm{span}\{\phi(S)\}$ [21]. Notice that the algorithm incorporates the *auxiliary variables* $w_i(n)$ and $w_i'(n)$; their role is essential for this paper [19], [21].

The second, communication part of the algorithm performs the following convexification w.r.t. the approximation parameter θ, leaving local auxiliary parameters unchanged, *i.e.*,

$$\theta_i(n+1) = \sum_{j=1}^{N} a_{ij}(n)\theta_j'(n), \qquad w_i(n+1) = w_i'(n), \tag{9}$$

where $a_{ij}(n)$ are random variables, elements of a random matrix $A(n) = [a_{ij}(n)]$ [11], [31], [35]. If one adopts that the

 ISBN 978-86-7466-930-3

available N MDP's are connected by communication links in accordance with a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ the set of arcs, then matrix $A(n)$ has zeros at the same places as the graph adjacency matrix A_G and is *row-stochastic, i.e.*, $\sum_{j=1}^{N} a_{ij}(n) = 1$, $i = 1, \ldots, N$, $\forall n \geq 0$. The algorithm (7), (8) incorporating consensus only w.r.t. θ according to (9) will be denoted as AlgA.

We also consider a modification of AlgA, denoted as AlgB, obtained by applying convexification to both θ_i and w_i, $i = 1, \ldots, N$, in such a way that the second relation in (9) becomes

$$w_i(n+1) = \sum_{j=1}^{N} a_{ij}(n) w'_j(n). \tag{10}$$

A comparative analysis of AlgA and AlgB is in the main focus of this paper.

III. Convergence Analysis

A. Prerequisites

1) Choice of λ-parameters: The results given below are applicable to both *state-dependent* and *history-dependent* λ_i. In the first case, we have simply $\lambda_i(n) = \lambda_i(S_n)$ for a given function $\lambda_i : \mathcal{S} \to [0, 1]$, while in the second case $\lambda_i(n) = \lambda(y_i(n), e_i(n-1))$, $y_i(n) = f(y_i(n-1), S_n)$, where $y(n)$, $n \geq 0$, is a memory state summarizing the history of the past states up to time n (see [21], [34]).

Different choices of λ_i lead to different generalized Bellman operators. For example, in the case of state-dependent λ_i, we have:

$$T^{(\lambda_i)} v = (I - P\Gamma\Lambda_i)^{-1} r_\pi + (I - P\Gamma\Lambda_i)^{-1} P\Gamma(I - \Lambda_i)v, \tag{11}$$

where Λ_i is an $M \times M$ diagonal matrix with entries $\lambda_i(s)$; therefore, we have $P^{(\lambda_i)} = (I - P\Gamma\Lambda_i)^{-1} P\Gamma(I - \Lambda_i)$. For the history-dependent λ_i-parameters, the formulation is more complex [21], [34] (the details are out of the scope of this paper).

2) Properties of the State-Trace Processes: Under the behavior policies $\pi^{(i)}$, the *state-trace processes* are defined as $\{S_n, e_i(n)\}$. These state-trace processes are Markov chains with the weak Feller property [21], [34]. Let $Z_i(n) = (S_n, e_i(n), S_{n+1})$. We have the following important result [21], [34]: a) the state-trace weak Feller-Markov chain process $Z_i(n)$ has a unique invariant probability measure ζ_i; for each initial condition the *occupation probability measure* converges weakly to ζ_i [21, Theorem 2.1(i)]; b) if E_{ζ_i} denotes the expectation of the stationary state-trace process with initial distribution ζ_i, then $E_{\zeta_i}\{\|f(Z_0)\|\} < \infty$ and $\frac{1}{n}\sum_{j=0}^{n-1} f(Z_i(j))$ converges to $E_{\zeta_i}\{f(Z_i(0))\}$ in mean and a.s., where $f(z)$ is a Lipschitz continuous function in the trace variable e [21, Theorem 2.1(ii)].

The results a) and b) are used to prove the following:
1) $E_{\zeta_i}\{\phi(S_0)\phi(S_0)^T\} = \Phi^T \Xi_i \Phi$;
2) $E_{\zeta_i}\{e_i(0)\ \delta_i(v; 0)\} = \Phi^T \Xi_i (T^{(\lambda_i)} v - v)$, $\forall v \in \mathcal{R}^M$
3) $E_{\zeta_i}\{e_i(0)\rho_i(0)(\phi^j(S_0) - \gamma(1)\phi^j(S_1))\} = \Phi^T \Xi_i (I - P^{(\lambda_i)})\Phi_j$, $1 \leq j \leq p$;
4) $E_{\zeta_i}\{e_i(0)\rho_i(0)(1 - \lambda_i(1))\gamma(1)\phi^j(S_1)\} = \Phi^T \Xi_i P^{(\lambda_i)} \Phi_j$, $1 \leq j \leq p$;

where $\phi^j(\cdot)$ is the j-th component of $\phi(\cdot)$ and Φ_j the j-th column vector of Φ [21, Proposition 2.1].

Under (A1), the results from [21, Proposition 2.2] also show that the sequences of traces $\{e_i(n)\}$ satisfy the condition $E\{\|e_i(n) - \hat{e}_i(n)\|\} \leq c(n)$, where $c(n) \to 0$ when $n \to \infty$, while $\{e_i(n)\}$ and $\{\hat{e}_i(n)\}$ are obtained using the same trajectory of states, but with different initial conditions $e_i(0)$ and $\hat{e}_i(0)$. Also, $\{e_i(n)\}$ is uniformly integrable, and, consequently, the random variables $\{Z_i(n)\}$, $n \geq 0$, are tight [36].

Let

$$g_i(\theta_i, w_i, Z_i) = \rho_i(s, s')(\phi(s) - \gamma(s')\phi(s'))e_i^T w_i \tag{12}$$

and

$$k_i(\theta_i, w_i, Z_i) = e_i \bar{\delta}_i(s, s', v_{\theta_i}) - \phi(s)\phi(s)^T w_i, \tag{13}$$

where $\bar{\delta}_i(s, s', v_{\theta_i}) = \rho_i(s, s')(r(s, s') + \gamma(s')v_{\theta_i}(s') - v_{\theta_i}(s))$. We also have:

$$\begin{aligned} \bar{g}_i(\theta_i, w_i) &= E_{\zeta_i}\{g_i(\theta_i, w_i, Z_i(0))\} \\ &= (\Phi^T \Xi_i (I - P^{(\lambda_i)})\Phi)^T w_i, \end{aligned} \tag{14}$$

$$\begin{aligned} \bar{k}_i(\theta_i, w_i) &= E_{\zeta_i}\{k_i(\theta_i, w_i, Z_i(0))\} \\ &= \Phi^T \Xi_i (T^{(\lambda_i)} v_{\theta_i} - v_{\theta_i}) - \Phi^T \Xi_i \Phi w_i, \end{aligned} \tag{15}$$

and

$$\bar{g}_i(\theta_i, w_i(\theta_i)) = (\Phi^T \Xi_i (I - P^{(\lambda_i)})\Phi)^T w_i(\theta_i). \tag{16}$$

Comparison with (3) shows that $\bar{g}_i(\theta_i, w_i(\theta_i) = -\nabla J_i(\theta_i)$.

Based on the above definitions and the results from [21], we have the following important ergodic properties:

Lemma 1 ([21]): Under (A1), the following holds for each θ_i and w_i and each compact set $D_i \subset \mathcal{Z}_i$:
a) $\lim_{m,n\to\infty} \frac{1}{m}\sum_{s=n}^{n+m-1} E_n\{k_i(\theta_i, w_i, Z_i(s+1)) - \bar{k}_i(\theta_i, w_i)\} I(Z_i(n) \in D_i) = 0$ in mean,
b) $\lim_{m,n\to\infty} \frac{1}{m}\sum_{s=n}^{n+m-1} E_n\{g_i(\theta_i, w_i, Z_i(s+1)) - \bar{g}_i(\theta_i, w_i)\} I(Z_i(n) \in D_i) = 0$ in mean,
c) $\lim_{m,n\to\infty} \frac{1}{m}\sum_{s=n}^{n+m-1} E_n\{g_i(\theta_i, w_i(\theta_i), Z_i(s+1)) - \bar{g}_i(\theta_i, w_i(\theta_i))\} I(Z_i(n) \in D_i) = 0$ in mean,
where $E_n\{\cdot\}$ denotes the conditional expectation given the history $(Z_i(0), \cdots, Z_i(n))$ and $I(\cdot)$ denotes the indicator function.

B. Global Model

Let $X(n) = [\Theta(n)^T \vdots W(n)^T]^T$, $\Theta(n) = [\theta_1(n)^T \cdots \theta_N(n)^T]^T$, $W(n) = [w_1(n)^T \cdots w_N(n)^T]^T$; similarly, $X'(n) = [\Theta'(n)^T \vdots W'(n)^T]^T$, together with the corresponding vector components. Then, we have for AlgA the following global model at the network level

$$\begin{aligned} X'(n) &= X(n) + \Gamma(n) F(X(n), n), \\ X(n+1) &= \mathrm{diag}\{(A(n) \otimes I_p), I_{Np}\} X'(n), \end{aligned} \tag{17}$$

where $\otimes$ denotes the Kronecker's product, while $\Gamma(n) = \mathrm{diag}\{\alpha_1(n), \ldots, \alpha_N(n), \beta_1(n), \ldots, \beta_N(n)\} \otimes I_p$,

$$F(X(n), n) = [F^\theta(X(n), n)^T \vdots F^w(X(n), n)^T]^T,$$

$$\begin{aligned} F^\theta(X(n), n) = [&q_1 g_1(\theta_1(n), w_1(n), Z_1(n))^T \cdots \\ &q_N g_N(\theta_N(n), w_N(n), Z_N(n))^T]^T, \end{aligned}$$

 ISBN 978-86-7466-930-3

$$F^w(X(n),n) = [k_1(\theta_1(n), w_1(n), Z_1(n))^T + e_1(n)^T\omega_1(n+1)\cdots k_N(\theta_N(n), w_N(n), Z_N(n)) + e_N(n)^T\omega_N(n+1)]^T$$

with $g_i(\cdot)$ defined by (12).

Introduce dummy variables $X = [\Theta^T \vdots W^T]^T$, together with $\bar{F}(X) = [\bar{F}^\theta(\Theta, W)^T \vdots \bar{F}^w(\Theta, W)^T]^T$, $F^\theta(\Theta,\ W) = [q_1\bar{g}_1(\theta_1, w_1)^T \cdots q_N\bar{g}_N(\theta_N, w_N)^T]^T$, with $\bar{g}_i(\cdot,\cdot)$ defined by (14), $F^w(\Theta, W) = [\bar{k}_1(\theta_1, w_1)^T \cdots \bar{k}_N(\theta_N, w_N)^T]^T$, with $\bar{k}_i(\cdot,\cdot)$ defined by (15).

In the case of AlgB, we have a slightly modified model (17): instead of $\mathrm{diag}\{(A(n)\otimes I_p), I_{Np}\}$ in the second relation in (17), we have $\mathrm{diag}\{(A(n)\otimes I_p), (A(n)\otimes I_p)\}$, as a consequence of consensus w.r.t. w_i.

C. Communication Part of the Algorithm

The result of this subsection is a slight generalization of the results in [5], based on [11].

Define $\Psi(n|k) = A(n)\cdots A(k)$ for $n \geq k$, $\Psi(n|n+1) = I_N$. Let $\mathcal{F}_n$ be an increasing sequence of σ-algebras such that $\mathcal{F}_n$ measures $\{X(k), k \leq n, A(k), k < n\}$.

(A2) There is a scalar $\alpha_0 > 0$ such that $a_{ii}(n) \geq \alpha_0$, and, for $i \neq j$, either $a_{ij}(n) = 0$ or $a_{ij}(n) \geq \alpha_0$.

(A3) There are a scalar $p_0 > 0$ and an integer n_0 such that $P_{\mathcal{F}_n}\{$agent j communicates to agent i on the interval $[n, n+n_0]\} \geq p_0$, for all n and $i = 1,\ldots N$, $j \in \mathcal{N}_i$.

(A4) The digraph $\mathcal{G}$ is strongly connected.

According to [5], [11], it is possible to show that (A2)-(A4) imply that $\Psi(k) = \lim_n \Psi(n|k)$ exists w.p.1; moreover, its rows are equal and $E\{|\Psi(n|k) - \Psi(k)|\}$, $E_{\mathcal{F}_n}\{|\Psi(n|k) - \Psi(k)|\} \to 0$ geometrically as $n - k \to \infty$, uniformly in k and ω (w.p.1). In addition, $E_{\mathcal{F}_n}\{\Psi(n|k)\}$ converges to $\Psi(k)$ geometrically, uniformly in ω and k, as $n \to \infty$.

D. Convergence Proofs

(A5) Sequence $\{A(n)\}$ is independent of the processes in MDP^i, $i = 1,\ldots,N$.

(A6) There is a $N\times N$ matrix $\bar{\Psi}$ such that $E\{|E_k\{\Psi(n)\} - \bar{\Psi}|\} \to 0$ as $n - k \to \infty$, which, under the conditions of Lemma 1, has the form

$$\bar{\Psi} = \begin{bmatrix} \bar{\psi_1} & \cdots & \bar{\psi_N} \\ & \cdots & \\ \bar{\psi_1} & \cdots & \bar{\psi_N} \end{bmatrix} = \begin{bmatrix} \hat{\Psi} \\ \vdots \\ \hat{\Psi} \end{bmatrix},$$

where $\sum_i \bar{\psi}_i = 1$ ($|\cdot|$ denotes the infinity norm).

(A7) Sequence $\{X(n)\}$ is tight.

1) AlgA):

Theorem 1: Let (A1)–(A7) hold. Let $X^\alpha(n)$ be generated by AlgA, (7), (8), (9), with $\beta_i(n) = \alpha_i(n) = \alpha$ and define for $t \geq 0$, $t \in \mathcal{R}$, $X^\alpha(t) = X(n)$ for $t \in [(n-n_\alpha)\alpha, (n-n_\alpha+1)\alpha)$. Let $w_i^\alpha(0) = w_{i,0}^\alpha$, $e_i(0) = e_{i,0} \in \mathrm{span}\{\phi(S)\}$. Then, for any integers n'_α such that $\alpha n'_\alpha \to \infty$ as $\alpha \to 0$, there exist positive numbers $\{T_\alpha\}$ with $T_\alpha \to \infty$ as $\alpha \to 0$ such that for any $\epsilon > 0$

$$\limsup_{\alpha\to 0} P\{(X^\alpha(n'_\alpha + k)) \notin N_\epsilon(\bar{\Sigma}) \text{ for some } k \in [0, T_\alpha/\alpha]\} = 0, \quad (18)$$

$i = 1,\ldots,N$, where $N_\epsilon(\cdot)$ denotes the ϵ-neighborhood, while $\bar{\Sigma} = \bar{\Sigma}_\theta \times \cdots \bar{\Sigma}_\theta \times \bar{\Sigma}_{w_1} \times \cdots \bar{\Sigma}_{w_N}$ is the set of points $\bar{\theta},\ldots,\bar{\theta},\bar{w}_1,\ldots,\bar{w}_N$ satisfying

$$\begin{aligned} \sum_{i=1}^N \bar{\psi}_i q_i G_i^T \bar{w}_i &= 0, \\ G_1\bar{\theta} + b_1 - H_1\bar{w}_1 &= 0, \\ &\vdots \\ G_N\bar{\theta} + b_N - H_N\bar{w}_N &= 0, \end{aligned} \quad (19)$$

where $G_i = \Phi^T\Xi_i(P^{(\lambda_i)} - I)\Phi$, $b_i = \Phi^T\Xi_i r_\pi^{(\lambda_i)}$, $r_\pi^{(\lambda_i)}$ is a constant M-vector in the affine function $T^{(\lambda_i)}(\cdot)$, while $H_i = \Phi^T\Xi_i\Phi$, $i = 1,\ldots,N$.

Proof: The proof is based on [1], [33] and the general results from [5], [11], [36]. In order to apply the proof of Theorem 3.1 in [5], it is essential to verify whether the basic assumptions from [5] concerning $F(X(n), n)$ hold in our case. We can easily conclude that Lemma 1 implies that the assumptions (C3.2) and C(3.3') from Section 3 in [5] hold. Following further [5], it follows that the Skorokhod embedding implies that we have the limit process $X^\alpha(\cdot) \to X(\cdot)$, where $\dot{X} = \mathrm{diag}\{\bar{\Psi}\otimes I_p, I_{Np}\}\bar{F}(X)$ [5]. By Lemma 1 and (A6), all the rows of $\bar{\Psi}$ are equal. Consequently, $\Theta(\cdot) = [\theta(\cdot)^T \cdots \theta(\cdot)^T]$, $\forall\theta(\cdot) \in \mathcal{R}^p$, implying that $\dot{\theta} = \bar{\psi}_1 q_1\bar{g}_1(\theta, w_1) + \cdots + \bar{\psi}_N q_N\bar{g}_N(\theta, w_N)$; we also have $\dot{w}_1 = \bar{k}_1(\theta, w_1), \ldots \dot{w}_N = \bar{k}_N(\theta, w_N)$, having in mind that consensus is not applied to the auxiliary variables.

In order to prove (18), we study the limit set

$$E = \cap_{\tau\geq 0}\mathrm{cl}\{\theta(t), w_1(t), \ldots, w_N(t) | \theta(0), w_1(0), \ldots, w_N(0) \in \mathcal{R}^{(N+1)p}, t \geq \tau\}. \quad (20)$$

where $\mathrm{cl}\{\cdot\}$ denotes the closure of a given set. Following [21] (Proposition 4.1), we introduce the Lyapunov function

$$V(\theta, w_1, \ldots w_N) = \frac{1}{2}\|\theta - \bar{\theta}\|^2 + \frac{1}{2}\sum_{i=1}^N q_i\bar{\psi}_i\|w_i - \bar{w}_i\|^2, \quad (21)$$

where $\bar{\theta}$ and $\bar{w}_i$ are given by (19). We have directly

$$\begin{aligned} \dot{V}(\theta, w_1, \ldots w_N) =& \langle\theta - \bar{\theta}, -\sum_{i=1}^N q_i\bar{\psi}_i G_i^T w_i\rangle \\ &+ \sum_{i=1}^N q_i\bar{\psi}_i\langle w_i - \bar{w}_i, G_i\theta + \bar{g}_i - \bar{H}_i w_i\rangle \\ =& -\sum_{i=1}^N q_i\bar{\psi}_i\langle w_i - \bar{w}_i, H_i(w_i - \bar{w}_i)\rangle. \end{aligned} \quad (22)$$

Therefore, $\dot{V}(\theta, w_1, \ldots w_N) < 0$ for $w_i \in \mathrm{span}\{\phi(S)\}$ and $w_i \neq \bar{w}_i$, implying that $\hat{w}_i = \bar{w}_i$ if $[\hat{\theta}^T\hat{w}_1^T \ldots \hat{w}_N^T]^T \in$

 ISBN 978-86-7466-930-3

E and $\hat{w}_i \in \text{span}\{\phi(S)\}$, $i = 1, \dots, N$. Similarly, if $[\hat{\theta}^T \bar{w}_1^T \dots \bar{w}_N^T]^T \in E$, then $\hat{\theta} = \bar{\theta}$. In such a way we conclude that for initial conditions $w_i(0) \in \text{span}\{\phi(S)\}$ the limit set E is indeed the set $\bar{\Sigma}$ of points satisfying (19).

The steps remaining to prove (18) are standard for the stochastic approximation theory (see [1], [21], [36]). ■

2) AlgB):

Theorem 2: Let (A1)–(A7) hold. Let $X^\alpha(n)$ be generated by AlgB (7), (8), (9) and (10), with $\beta_i(n) = \alpha_i(n) = \alpha$, and let both $w_i^\alpha(0) = w_{i,0}^\alpha$ and $e_i(0) = e_{i,0} \in \text{span}\{\phi(S)\}$. Then, for any integers n'_α such that $\alpha n'_\alpha \to \infty$ as $\alpha \to 0$, there exist positive numbers $\{T_\alpha\}$ with $T_\alpha \to \infty$ as $\alpha \to 0$ such that for any $\epsilon > 0$

$$\limsup_{\alpha \to 0} P\{ \begin{bmatrix} \theta_i^\alpha(n'_\alpha + k) \\ w_i^\alpha(n'_\alpha + k) \end{bmatrix} \notin N_\epsilon(\bar{\Sigma}) \quad \text{for some} \quad k \in [0, T_\alpha/\alpha]\} = 0, \tag{23}$$

$i = 1, \dots, N$, where $N_\epsilon(\cdot)$ denotes the ϵ-neighborhood, while $\bar{\Sigma} = \bar{\Sigma}_\theta \times \bar{\Sigma}_w$ is the set of points $\bar{x} = [\bar{\theta}^T \bar{w}^T]^T \in \mathcal{R}^{2p}$ satisfying

$$\bar{G}\bar{\theta} + \bar{g} - \bar{H}\bar{w} = 0, \qquad \bar{G}^T \bar{w} = 0, \tag{24}$$

where $\bar{G} = \sum_{i=1}^N \bar{\psi}_i q_i \Phi^T \Xi_i (P^{(\lambda_i)} - I)\Phi$, $\bar{b} = \Phi^T \sum_{i=1}^N \bar{\psi}_i q_i \Xi_i r_\pi^{(\lambda_i)}$, $r_\pi^{(\lambda_i)}$ is a constant M-vector in the affine function $T^{(\lambda_i)}(\cdot)$, while $\bar{H} = \sum_{i=1}^N \bar{\psi}_i q_i \Phi^T \Xi_i \Phi$.

Proof: AlgA differs from AlgB only in the communication part of the algorithm. Formally, the procedure of the proof remains the same as in Theorem 1, after replacing $\text{diag}\{(A(n) \otimes I_p), I_{Np}\}$ by $\text{diag}\{(A(n) \otimes I_p), (A(n) \otimes I_p)\}$. This implies that, asymptotically, instead of $\text{diag}\{(\hat{\Psi} \otimes I_p), I_{Np}\}$ we have now $\text{diag}\{(\hat{\Psi} \otimes I_p), (\hat{\Psi} \otimes I_p)\}$. In this sense, we obtain $X(\cdot) = [\theta(\cdot)^T \cdots \theta(\cdot)^T w(\cdot)^T \cdots w(\cdot)^T]^T$, where $\theta(\cdot)$ and $w(\cdot)$ satisfy the following ODE:

$$\begin{bmatrix} \dot{\theta} \\ \dot{w} \end{bmatrix} = \bar{\psi}_1 q_1 \begin{bmatrix} \bar{g}_1(\theta, w) \\ \bar{k}_1(\theta, w) \end{bmatrix} + \cdots + \bar{\psi}_N q_N \begin{bmatrix} \bar{g}_N(\theta, w) \\ \bar{k}_N(\theta, w) \end{bmatrix} \tag{25}$$

The limit points (24) follow from (25), according to Theorem 1. Namely, we define the Lyapunov function

$$V(\theta, w_1, \dots w_N) = \frac{1}{2}\|\theta - \bar{\theta}\|^2 + \frac{1}{2}\|w - \bar{w}\|^2, \tag{26}$$

where $\bar{\theta}$ and $\bar{w}$ are given by (24) and obtain for the derivative that

$$\begin{aligned} \dot{V}(\theta, w) &= \langle \theta - \bar{\theta}, \bar{G}^T w \rangle + \langle w - \bar{w}, \bar{G}\theta + \bar{b} - \bar{H}w \rangle \\ &= -\langle w - \bar{w}, \bar{H}(w_- \bar{w}) \rangle. \end{aligned} \tag{27}$$

Therefore, $\dot{V}(\theta, w) < 0$ for $w_i \in \text{span}\{\phi(S)\}$ and $w \neq \bar{w}$, implying that $\hat{w} = \bar{w}$ if $[\hat{\theta}^T \hat{w}^T]^T \in E$ and $\hat{w} \in \text{span}\{\phi(S)\}$. In the same way, if $[\hat{\theta}^T \bar{w}^T]^T \in E$, then $\hat{\theta} = \bar{\theta}$. Consequently, for the initial conditions $w_i(0) \in \text{span}\{\phi(S)\}$, the limit set of ODE (25) is the set $\bar{\Sigma}$ satisfying (24). ■

IV. Discussion

The preceding section has been devoted to the weak convergence of the proposed distributed temporal difference learning algorithms. The role of convexification of w_i remains to be clarified. It is clear, from the definition of the criterion function (2) and the algorithm construction, that AlgA follows from the basic local relations (4), providing for all i unique solutions $w_i(\theta)$ for all θ. However, AlgB is based on the introduction of an additional constraint that $w_1(\theta) = \cdots = w_N(\theta) = w(\theta)$, where $w(\theta)$ is the unique solution of

$$\Phi^T (\sum_{i=1}^N \bar{\psi}_i q_i \Xi_i) \Phi w(\theta) = \sum_{i=1}^N \bar{\psi}_i q_i \Pi_{\xi_i} (T^{(\lambda_i)} v_\theta - v_\theta). \tag{28}$$

It is straightforward to observe from (28) that we have for any given θ

$$\Phi^T (\sum_{i=1}^N \bar{\psi}_i q_i \Xi_i) \Phi w(\theta) = \sum_{i=1}^N \bar{\psi}_i q_i \Phi \Xi_i \Phi w_i(\theta). \tag{29}$$

This property is verified by the above theorems.

Consequently, convergence points of θ are different for these two cases. In order to get a clearer insight, assume that $\det H_i \neq 0$. Then, for AlgA we have

$$\sum_{i=1}^N \bar{\psi}_i q_i G_i^T w_i(\theta) = \sum_{i=1}^N \bar{\psi}_i q_i G_i^T [H_i^{-1}(G_i\theta + b_i)] = 0, \tag{30}$$

resulting in

$$\sum_{i=1}^N \bar{\psi}_i q_i G_i^T H_i^{-1} G_i \theta = \sum_{i=1}^N \bar{\psi}_i q_i G_i^T H_i^{-1} b_i = 0, \tag{31}$$

while for AlgB we obtain

$$\bar{G^T} \bar{H}^{-1} \bar{G}\theta = \bar{G^T} \bar{H}^{-1} \bar{b}. \tag{32}$$

Notice that in the case of equal λ-parameters and equal behavior policies for all the agents, both algorithms provide the same solution.

It is difficult to make any general conclusion about the relative advantage of one of the two presented algorithms. It is to be noticed that this issue has not been directly treated in the literature; all the examples of distributed TD algorithms subsume that the consensus operator is applied to both θ and w, without mentioning any alternative (*e.g.*, [28] with the references therein). As far as the limit points of the corresponding ODEs are concerned, it should be noticed that a better approximation could be, in general, expected when consensus is not applied to w. Our experience confirms this statement; however it does not show any significant difference from this point of view. In some cases it could be expected that the application of consensus to w may improve the convergence rate of the algorithm. It is hard to draw any definite conclusion, in general, having in mind that connectedness of the underlying network can play an essential role from this point of view. A comprehensive Monte Carlo analysis could practically resolve the remaining dilemmas. It would be also interesting to analyze the discussed problem in the two-time-scale cases (see [1]).

 ISBN 978-86-7466-930-3

V. Conclusion

In this paper we have presented and discussed two novel algorithms for distributed off-policy gradient-based value function approximation within a collaborative multi-agent reinforcement learning setting. The algorithms are based on an integration of linear dynamic consensus schemes into local gradient-based recursions, involving the so called *auxiliary variables*. We presented rigorous proofs that, under nonrestrictive assumptions, the parameter estimates weakly converge to consensus. Based on these proofs, a discussion is provided of the incorporation of consensus w.r.t. auxiliary variables, defining clearly the figure of merit of the alternative approaches.

References

[1] M. S. Stanković, M. Beko, and S. S. Stanković, "Distributed value function approximation for collaborative multi-agent reinforcement learning," *IEEE Trans. Control Networked Systems*, vol. (Early Access), 2021.

[2] M. S. Stanković, S. S. Stanković, and D. M. Stipanović, "Consensus-based decentralized real-time identification of large-scale systems," *Automatica*, vol. 60, pp. 219–226, 2015.

[3] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Dep. Electrical Eng. Comput. Sci., M.I.T., Cambridge, MA, 1984.

[4] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, pp. 803–812, 1986.

[5] 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.

[6] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, pp. 48–61, 2009.

[7] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. Inf. Theory*, vol. 59, pp. 7405–7418, 2013.

[8] M. S. Stanković, S. S. Stanković, and K. H. Johansson, "Distributed time synchronization for networks with random delays and measurement noise," *Automatica*, vol. 93, pp. 126 – 137, 2018.

[9] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 60, pp. 601 – 615, 2015.

[10] S. Tu and A. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, pp. 6217–6233, 2012.

[11] 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.

[12] N. Ilić, S. S. Stanković, M. S. Stanković, and K. H. Johansson, "Consensus based distributed change detection using generalized likelihood ratio methodology," *Signal Processing*, vol. 92, no. 7, pp. 1715 – 1728, 2012.

[13] M. S. Stanković, S. S. Stanković, K. H. Johansson, M. Beko, and L. M. Camarinha-Matos, "On consensus-based distributed blind calibration of sensor networks," *Sensors*, vol. 18, no. 11, 2018.

[14] M. S. Stanković, K. H. Johansson, and D. M. Stipanović, "Distributed seeking of Nash equilibria with applications to mobile sensor networks," *IEEE Trans. Autom. Control*, vol. 57, no. 4, pp. 904–919, 2012.

[15] S. S. Stanković, M. Beko, and M. S. Stanković, "Nonlinear robustified stochastic consensus seeking," *Systems & Control Letters*, vol. 139, p. 104667, 2020.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[17] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[18] D. Precup, R. S. Sutton, and S. Dasgupta, "Off-policy temporal-difference learning with function approximation," in *Proc. 18th Int. Conf. on Machine Learning*, 2001, pp. 417–424.

[19] 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.

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

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

[22] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, J. Chen, and L. Song, "SBEED: Convergent reinforcement learning with nonlinear function approximation," *arXiv:1712.10285*, 2017.

[23] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156–172, 2008.

[24] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Autonomous Agents and Multiagent Systems*, G. Sukthankar and J. A. Rodriguez-Aguilar, Eds. Cham: Springer International Publishing, 2017, pp. 66–83.

[25] A. OroojlooyJadid and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *arXiv:1908.03963*, 2019.

[26] A. Mathkar and V. S. Borkar, "Distributed reinforcement learning via gossip," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1465–1470, 2017.

[27] S. Kar, J. M. Moura, and H. V. Poor, "QD-Learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations," *IEEE Trans. Signal Proc.*, vol. 61, no. 7, pp. 1848–1862, 2013.

[28] S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed, "Distributed policy evaluation under multiple behavior strategies," *IEEE Trans. Autom. Control*, vol. 60, no. 5, pp. 1260–1274, 2015.

[29] 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.

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

[31] M. S. Stanković and S. S. Stanković, "Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies," in *2016 American Control Conference (ACC)*, 2016, pp. 167–172.

[32] T. Doan, S. Maguluri, and J. Romberg, "Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2019, pp. 1626–1635.

[33] M. S. Stanković, M. Beko, N. Ilić, and S. S. Stanković, "Distributed multi-agent reinforcement learning algorithm based on gradient correction," in *Proc. IcETRAN Conference*, 2020.

[34] 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.

[35] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 47, pp. 531–543, 2011.

[36] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

 ISBN 978-86-7466-930-3