Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Distributed Gradient Temporal Difference Off-policy Learning With Eligibility Traces: Weak Convergence

Miloš S. Stanković * **Marko Beko** ** **Srdjan S. Stanković** ***

* *Innovation Center, School of Electrical Engineering, University of Belgrade; Vlatacom Institute, Belgrade; and Singidunum University, Belgrade, Serbia (e-mail: milstank@gmail.com).*
** *COPELABS, Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal; and UNINOVA, Caparica, Portugal.*
*** *School of Electrical Engineering, University of Belgrade, Serbia; and Vlatacom Institute, Belgrade, Serbia.*

Abstract: In this paper we propose two novel distributed algorithms for multi-agent off-policy learning of linear approximation of the value function in Markov decision processes. The algorithms differ in the way of how distributed consensus iterations are incorporated in a basic, recently proposed, single agent scheme. The proposed completely decentralized off-policy learning schemes subsume local eligibility traces, and allow applications in which all the agents may have different behavior policies while evaluating a single target policy. Under nonrestrictive assumptions on the time-varying network topology and the individual state-visiting distributions of the agents, we prove that the parameter estimates of the algorithms weakly converge to a consensus. The variance reduction properties of the proposed algorithms are demonstrated. We also formulate specific guidelines on how to design the network weights and topology. The results are illustrated using simulations.

Keywords: Reinforcement learning; Distributed consensus; Value function approximation; Convergence; Eligibility traces; Off-policy learning; Weak convergence; Multi-agent systems.

1. INTRODUCTION

Recently, interest in decentralized multi-agent algorithms has grown dramatically mainly due to their fundamental role in cutting edge technologies such as Cyber-Physical Systems (CPS) and Internet of Things (IoT). Distributed estimation and optimization methods play an essential role in development of these algorithms; a large class of them are based on consensus-based collaborations (e.g. Stanković et al. (2011); Stanković et al. (2015); Kushner and Yin (1987); Stanković et al. (2018b,a); Nedić and Olshevsky (2015); Stanković et al. (2016, 2020) and references therein).

Reinforcement learning (RL) is a powerful methodology for decision making in uncertain environments which typically uses Markov Decision Process (MDP) modeling, as well as approximate and adaptive dynamic programming (Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996). A fundamentally important issue is the approximation of the value function, under large state spaces and off-policy learning (e.g. Sutton et al. (2009); Geist and Scherrer (2014); Yu (2017)).

Distributed and multi-agent RL methods have received a lot of attention (see, e.g. Busoniu et al. (2008); Zhang et al. (2019) and references therein). A typical distributed setup assumes that each agent can access the same MDP with different agents' behaviors, and without mutual interactions through the MDP. Similar setups have been adopted in several recent works (Mathkar and Borkar, 2017; Kar et al., 2013; Macua et al., 2015; Lee et al., 2018; Zhang and Zavlanos, 2019; Stanković and Stanković, 2016; Suttle et al., 2019), but introducing several restricitve and/or simplifying assumptions: the same behavior of all the agents (Mathkar and Borkar, 2017), presence of a global controller (Kar et al., 2013), independent sampling from the underlying stationary distributions, and without eligibility traces (Stanković and Stanković, 2016; Macua et al., 2015; Lee et al., 2018), mean-square convergence under fixed communication graph (Macua et al., 2015). The authors of (Zhang and Zavlanos, 2019; Suttle et al., 2019) developed actor-critic schemes based on similar reasoning with consensus-based collaboration.

In this paper we propose new distributed algorithms for *iterative multi-agent off-policy learning* of linear approximation of the value function in MDPs. The algorithms represent a generalization of the recently proposed single agent off-policy algorithms (Sutton et al., 2009; Geist and Scherrer, 2014; Yu, 2017), including the algorithms with *eligibility traces*, to decentralized multi-agent framework. The main idea is to incorporate linear distributed dynamic consensus iterations over the underlying network of agents which can communicate only with their corresponding

⋆ This work was partially supported by Fundação para a Ciência e a Tecnologia under Project IF/00325/2015, Project foRESTER PCIF/SSI/0102/2017, and Project UIDB/04111/2020.

neighbors, avoiding in such a way dependence on any type of fusion center. In the adopted distributed framework, it is of fundamental importance that the proposed algorithms are off-policy since this allows applications to scenarios which are typical in practice, in which all the agents may have different behavior policies while evaluating a single target policy. Another important property of the proposed algorithms is that the local recursions of each agent can be based on eligibility traces (Geist and Scherrer, 2014; Yu, 2017), where each agent may choose different λ parameters, which can be *history dependent*. The linear value function parameterization is performed a priori in such a way that all the agents use the same feature vectors. Under nonrestrictive assumptions on the time-varying network topology and the individual behavior policies, we prove that the parameter estimates of the algorithms *weakly converge* to a consensus point. In the convergence proofs, the stochastic dynamics of the underlying MDP are rigoroulsy taken into account. The denoising effect of the scheme is verified by theoretical analysis of the algorithms' rate of convergence. We also formulate specific guidelines on how to design the network weights and topology such that a desired convergence point is achieved.

The paper is organized as follows. In Section 2 we formulate the problem and define the proposed algorithms. In Section 3 rigorous weak convergence analysis, including convergence rate, is presented, while in Section 4 the results of simulations demonstrating the effectiveness of the proposed algorithms are shown.

2. DISTRIBUTED GRADIENT BASED TEMPORAL DIFFERENCE ALGORITHMS

Consider N *autonomous agents*, each acting on a separate Markov Decision Process (MDP), denoted as $\text{MDP}^{(i)}$, $i = 1, \dots, N$, characterized by the quadruplet $\{\mathcal{S}, \mathcal{A}, p(s'|s,a), R(s,a,s')\}$, where $\mathcal{S} = \{1, \dots, M\}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $p(s'|s,a)$ is a function defining probabilities of moving from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ by applying action $a \in \mathcal{A}$, and $R(s,a,s')$ are appropriate random rewards. Let $\text{MDP}^{(0)}$, characterized by the same quadruplet, represent a reference MDP. Each $\text{MDP}^{(i)}$, $i = 0, 1, \dots, N$, has an associated fixed stationary *policy* $\pi^{(i)}(a|s)$ (indicating the probability of taking action a at state s), so that the resulting state processes $\{S_i(n)\}$, where $n \geq 1$ denotes integer transition times, are time homogenous Markov chains. The goal of the agents is to learn the *state value function* for a given *target policy* $\pi^{(0)}$ in $\text{MDP}^{(0)}$, where each agent i can observe only state transitions and rewards in $\text{MDP}^{(i)}$ with *behavior policy* $\pi^{(i)}$, $i = 1, \dots, n$. Let $P^{(0)}$ and $P^{(i)}$ denote the resulting transition matrices of the Markov chains $\{S_0(n)\}$ and $\{S_i(n)\}$, respectively. Therefore, we are dealing with a *cooperative off-policy reinforcement learning* problem (Sutton and Barto, 1998; Stanković and Stanković, 2016).

The desired *state value function* is defined in accordance with the classical literature, using *discount factors* $\gamma(s) \in [0,1]$, $s \in \mathcal{S}$; we describe the value function using a vector $v_{\pi^{(0)}} \in \mathcal{R}^M$ (Yu et al., 2019). Denote by Γ the $M \times M$ diagonal matrix with $\gamma(s)$ as diagonal entries. By the MDP theory (Sutton and Barto, 1998; Yu, 2017; Yu et al., 2019) $v_{\pi^{(0)}}$ uniquely satisfies the *Bellman equation*

$$v_{\pi^{(0)}} = r_{\pi^{(0)}} + P^{(0)} \Gamma v_{\pi^{(0)}}, \tag{1}$$

where $r_{\pi^{(0)}}$ is the vector of one-stage expected rewards at each state $s \in \mathcal{S}$ under policy $\pi^{(0)}$. Besides (1), $v_{\pi^{(0)}}$ also satisfies a family of *generalized Bellman equations*, $v_{\pi^{(0)}} = T^{(0,\lambda)} v_{\pi^{(0)}}$, where $T^{(0,\lambda)}$ is the *generalized Bellman operator*, $T^{(0,\lambda)} v = r^{(\lambda)}_{\pi^{(0)}} + P^{(0,\lambda)} v$, $\forall v \in \mathcal{R}^M$, for a given vector $r^{(\lambda)}_{\pi^{(0)}}$ and a substochastic matrix $P^{(0,\lambda)}$, where $\lambda \in [0,1]$ are the so-called λ-parameters (Yu, 2017; Yu et al., 2019).

Introduce the local *importance sampling ratios* $\rho_i(s,s') = P^{(0)}_{ss'}/P^{(i)}_{ss'}$ for $s, s' \in \mathcal{S}$ (with $0/0 = 0$). The following assumption ensures well defined value function and the importance ratios:

(A1) (*Assumptions on target and behavior policies*)
a) $P^{(0)}$ is such that $I - P^{(0)}\Gamma$ is nonsingular;
b) $P^{(i)}$ are irreducible and such that for all $s, s' \in \mathcal{S}$ $P^{(i)}_{ss'} = 0 \Rightarrow P^{(0)}_{ss'} = 0$, $i = 1, \dots, N$.

Let $\phi : \mathcal{S} \to \mathcal{R}^p$ be a function that maps each state to a p-dimensional feature vector ϕ; let the subspace spanned by these vectors be $\mathcal{L}_\phi$. Our goal is to find $v = [v_1 \dots v_M]^T \in \mathcal{L}_\phi$ that satisfies $v \approx T^{(0,\lambda)} v$, assuming that $v_s = \phi(s)^T \theta$, where $s \in \mathcal{S}$ and $\theta \in \mathcal{R}^p$ is a parameter vector. Let Φ be an $M \times p$ matrix composed of p-vectors $\phi(s)$ as row vectors, so that we introduce $v_\theta = \Phi\theta$.

We define the following *global objective function*

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

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, $T^{(\lambda_i)}$ stands for $T^{(0,\lambda_i)}$ and Π_{ξ_i} denotes the projection onto the subspace $\mathcal{L}_\phi$ w.r.t. the weighted Euclidean norm $\|v\|^2_{\xi_i} = \sum_{s \in \mathcal{S}} \xi_{i;s} v_s^2$ for a positive M-dimensional vector ξ_i with components $\xi_{i;s}$ (see Stanković and Stanković (2016); Yu (2017)). Let ξ_i be the invariant probability distribution for the local Markov chain $\{S^{(i)}(n)\}$ ($\xi_i^T P^{(i)} = \xi_i^T$). 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 $P^{(\lambda_i)}$ stands for $P^{(0,\lambda_i)}$, 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 $w_i \in \text{span}\{\phi(s)\}$.

Let $\rho_i(n) = \rho_i(S_i(n), S_i(n+1))$ and $\gamma_i(n) = \gamma(S_i(n))$ (Yu, 2017; Yu et al., 2019). The local *temporal-difference terms* are given by $\delta_i(v_\theta; n) = \rho_i(n)(R_i(n+1) + \gamma_i(n+1) v_\theta(S_i(n+1)) - v_\theta(S_i(n)))$, where $R_i(n+1)$ is the local random reward defined by the function $R(s,a,s')$ and the corresponding transition, and $v_\theta(S_i(n))$ is the approximation of the value function obtained by the i-th agent using θ.

The sequences of the local *eligibility trace vectors* $\{e_i(n)\}$ are supposed to be locally available. They are defined as

$$e_i(n) = \lambda_i(n)\gamma_i(n)\rho_i(n-1) e_i(n-1) + \phi(S_i(n)). \tag{5}$$

We propose algorithms composed of two main parts: 1) *local parameter updates*, based on *gradient descent* recursion starting from (3), and local observations of state transitions and rewards associated to MDP$^{(i)}$, and 2) real-time incorporation of estimates communicated from neighboring agents, aimed at achieving *consensus*.

The first proposed algorithm is derived from (3) and denoted as D1-GTD2(λ), according to the algorithm GTD2 proposed in (Sutton et al., 2009). The local updates are defined by

$$\begin{aligned}\theta_i'(n) =&\theta_i(n) + \alpha(n)q_i\rho_i(n)(\phi(S_i(n)) \\ &- \gamma_i(n+1)\phi(S_i(n+1)))e_i(n)^T w_i(n) \quad (6)\\ w_i'(n) =&w_i(n) + \beta(n)(e_i(n)\delta_i(v_{\theta_i(n)}; n) \\ &- \phi(S_i(n))\phi(S_i(n))^T w_i(n)) \quad (7)\end{aligned}$$

where $v_{\theta_i(n)} = \Phi\theta_i(n)$. The initial values $\theta_i(0)$ are chosen arbitrarily; however, $w_i(0)$, as well as $e_i(0)$, have to satisfy $w_i(0)$, $e_i(0) \in \text{span}\{\phi(s)\}$ in order to achieve the desired convergence properties (discussed in the next section). Sequences $\{\alpha(n)\}$ and $\{\beta(n)\}$ are positive step sizes, which can be either equal (single time-scale) or satisfying $\alpha(n) << \beta(n)$ (two time-scales), see (Yu, 2017).

The second step of the algorithm performs the following convexification:

$$\theta_i(n+1) = \sum_{j=1}^{N} a_{ij}(n)\theta_j'(n), \quad w_i(n+1) = w_i'(n), \quad (8)$$

where $a_{ij}(n)$ are random variables, elements of a random matrix $A(n) = [a_{ij}(n)]$(Stanković et al., 2011, 2016; Stanković and Stanković, 2016). If one adopts that the agents 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.*, $\sum_{j=1}^{N} a_{ij}(n) = 1$, $i = 1, \ldots, N$, $\forall n \geq 0$.

We also consider a modification of D1-GTD2(λ), denoted as D2-GTD2(λ), obtained by applying convexification to both θ_i and w_i, $i = 1, \ldots, N$, from (6), in such a way that in (8) $w_i(n+1) = \sum_{j=1}^{N} a_{ij}(n)w_j'(n)$.

3. CONVERGENCE ANALYSIS

3.1 *Preliminaries*

1. Properties of the State-Trace Processes

We shall consider state-dependent λ_i, when $\lambda_i(n) = \lambda_i(S_i(n))$ for a given function $\lambda_i : \mathcal{S} \to [0, 1]$. Let $Z_i(n) = (S_i(n), e_i(n), S_i(n+1))$. It has been shown that the state-trace processes are Markov chains with the weak Feller property (see Yu (2017); Yu et al. (2019) for details).

According to (6), denoting $z = (s, e, s')$, introduce functions

$$g_i(\theta, w, z) = \rho_i(s, s')(\phi(s) - \gamma(s')\phi(s'))e^T w \quad (9)$$

and

$$k_i(\theta, w, z) = e\bar{\delta}_i(s, s', v_\theta) - \phi(s)\phi(s)^T w, \quad (10)$$

where $\bar{\delta}_i(s, s', v_\theta) = \rho_i(s, s')(r_i(s, s') + \gamma(s')v_\theta(s') - v_\theta(s))$ is the temporal difference term with averaged reward $r_i(s, s')$ (under policy $\pi^{(i)}$, with possible presence of additive zero-mean white noise term in the actual rewards). Using the results from (Yu, 2017), we have:

$$\bar{g}_i(\theta, w) = (\Phi^T\Xi_i(I - P^{(\lambda_i)})\Phi)^T w \quad (11)$$

$$\bar{k}_i(\theta, w) = \Phi^T\Xi_i(T^{(\lambda_i)}v_\theta - v_\theta) - \Phi^T\Xi_i\Phi w. \quad (12)$$

For any given θ_i there is a unique solution $w_i(\theta_i)$ to the linear equation $\bar{k}_i(\theta_i, w_i) = 0$, $w_i \in \text{span}\{\phi(s)\}$, so that we obtain $\bar{g}_i(\theta_i, w_i(\theta_i)) = -\nabla J_i(\theta_i)$.

The results from (Yu, 2017) may be applied *in extenso* to our analysis. We shall mention only the following basic result:
Under (A1), for each θ_i and w_i on each compact set $D_i \in \text{domain}(Z_i)$, $\bar{k}_i(\theta_i, w_i) = \lim_{m,n\to\infty} \frac{1}{m} \sum_{s=n}^{n+m-1} E_n\{k_i(\theta_i, w_i, Z_i(s+1))\}I(Z_i(n) \in D_i) = 0$ (in mean) and $\bar{g}_i(\theta_i, w_i) = \lim_{m,n\to\infty} \frac{1}{m} \sum_{s=n}^{n+m-1} E_n\{g_i(\theta_i, w_i, Z_i(s+1))\}I(Z_i(n) \in D_i) = 0$ (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 (for details, see Yu (2017)).

The last result is basic for the convergence analysis given below, allowing direct verification of the underlying general assumptions given in (Kushner and Yin, 1987, 2003).

2. 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 the following global model at the network level:

$$\begin{aligned}X'(n) &= X(n) + \Gamma(n)F(X(n), n), \quad (13)\\ X(n+1) &= \text{diag}\{(A(n) \otimes I_p), I_{Np}\}X'(n),\end{aligned}$$

where $\otimes$ denotes the Kronecker's product,
- $\Gamma(n) = \text{diag}\{\alpha(n), \beta(n)\} \otimes I_{Np}$,
- $F(X(n), n) = [F^\theta(X(n), n)^T \vdots F^w(X(n), n)^T]^T$,
- $F^\theta(X(n), n) = [q_1 g_1(\theta_1(n), w_1(n), Z_1(n))^T \vdots q_N g_N(\theta_N(n), w_N(n), Z_N(n))^T]^T$,
- $F^w(X(n), n) = [k_1(\theta_1(n), w_1(n), Z_1(n))^T + e_1(n)^T\omega_1(n+1) \cdots k_N(\theta_N(n), w_N(n), Z_N(n)) + e_N(n)^T \omega_N(n+1)]^T$
- $\omega_i(n+1) = \rho_i(n)(R_i(n+1) - r(S_i(n), S_i(n+1)))$.

3. Communication Part of the Algorithm

The results given here follow from (Kushner and Yin, 1987) and (Stanković et al., 2016).

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

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

Lemma 1. (Kushner and Yin (1987)). Let (A2)–(A4) hold. Then $\Psi(k) = \lim_n \Psi(n|k)$ exists w.p.1 and its rows are all equal; moreover, $E\{|\Psi(n|k) - \Psi(k)|\}$ and $E_{\mathcal{F}_n}\{|\Psi(n|k) - \Psi(k)|\} \to 0$ geometrically as $n - k \to \infty$, uniformly in k and ω (w.p.1); also, $E_{\mathcal{F}_n}\{\Psi(n|k)\}$ converges to $\Psi(k)$ geometrically, uniformly in ω and k, as $t \to \infty$.

3.2 ***Weak Convergence Proof***

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

(A6) There is a $N \times N$ matrix $\bar{\Psi}$ such that $E\{|E_{\mathcal{F}_k}\{\Psi(n)\} - \bar{\Psi}|\} \to 0$ as $n - k \to \infty$, uniformly in k. Under the conditions of Lemma 1,

$$\bar{\Psi} = \begin{bmatrix} \bar{\psi}_1 & \cdots & \bar{\psi}_N \\ & \cdots & \\ \bar{\psi}_1 & \cdots & \bar{\psi}_N \end{bmatrix} = \begin{bmatrix} \hat{\Psi} \\ \vdots \\ \hat{\Psi} \end{bmatrix},$$

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

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

Because of the lack of space, we shall pay attention only to D1-GTD2(λ) with single time-scale. Define $X^\alpha(\cdot)$ as $X^\alpha(t) = X(n)$ for $t \in [(n - n_\alpha)\alpha, (n - n_\alpha + 1)\alpha)$, where n_α is a sequence tending to ∞ and satisfying $\alpha^{\frac{1}{2}} n_\alpha \to 0$ (for details, see Kushner and Yin (1987)).

Theorem 1. Let (A1)–(A7) hold. Let $X^\alpha(n)$ be generated by (6), (7) and (8), with $\beta_i(n) = \alpha_i(n) = \alpha$. Let $w_i^\alpha(0) = w_{i,0}^\alpha$, $e_i(0) = e_{i,0} \in \text{span}\{\phi(s)\}$. Define $X^\alpha(0)$ by $\lim_{\alpha \to 0} X_0^\alpha = [\theta_0^T \cdots \theta_0^T w_{1,0}^T \cdots w_{N,0}^T]^T$. Then $X^\alpha(\cdot)$ is tight and converges weakly to a process $X^\alpha(\cdot) = [\theta(\cdot)^T \cdots \theta(\cdot)^T w_1(\cdot)^T \cdots w_N(\cdot)^T]^T$, where $\theta(\cdot), w_1(\cdot), \dots, w_N(\cdot)$ satisfy the following system of ODE's

$$\dot{\theta} = \bar{\psi}_1 q_1 \bar{g}_1(\theta, w_1) + \cdots + \bar{\psi}_N q_N \bar{g}_N(\theta, w_N), \quad (14)$$
$$\dot{w}_1 = \bar{k}_1(\theta, w_1), \ \cdots, \ \dot{w}_N = \bar{k}_N(\theta, w_N),$$

with initial conditions $\theta_0, w_{1,0}, \dots, w_{N,0}$.

Moreover, 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 $\epsilon > 0$ $\limsup_{\alpha \to 0} P\{(X^\alpha(n'_\alpha + k)) \notin N_\epsilon(\bar{\Sigma}) \text{ for some } k \in [0, T_\alpha/\alpha]\} = 0$, $i = 1, \dots, N$, 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 $\bar{\theta}, \dots, \bar{\theta}, \bar{w}_1, \dots, \bar{w}_N$ satisfying

$$\sum_{i=1}^N \bar{\psi}_i q_i G_i^T \bar{w}_i = 0, \quad (15)$$
$$G_1\bar{\theta} + b_1 - H_1\bar{w}_1 = 0, \ \cdots, \ G_N\bar{\theta} + b_N - H_N\bar{w}_N = 0,$$

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 $H_i = \Phi^T \Xi_i \Phi$, $i = 1, \dots, N$.

Proof. *Part 1.* Iterating (13) back, one obtains

$$X(n+1) = X_0^\alpha + \alpha \sum_{k=n_\alpha}^n \text{diag}\{\Psi(k) \otimes I_p, I_{Np}\} F(X(k), k)$$
$$+ \alpha \varrho^\alpha(n) + \text{diag}\{[\Psi(n|0) - \Psi(n_\alpha|0)] \otimes I_p, I_{Np}\} X_0, \quad (16)$$

where $\varrho^\alpha(n) = \sum_{k=0}^n \text{diag}\{[\Psi(n|k) - \Psi(k)] \otimes I_p, I_{Np}\} F(X(k), k)$. A direct comparison of (16) with the analogous expression in the proof of Theorem 3.1 in (Kushner and Yin, 1987), shows that the main formal difference lies in the specific form of the model (13) and the replacement of $\Psi(k)$ by $\Psi(k+1)$. Having in mind general properties of the matrix $\Psi(k)$ (see also Stanković et al. (2016)), it becomes evident that the results of Theorem 3.1 in (Kushner and Yin, 1987) can be almost directly applied to (13). What essentially has to be verified is whether or not the basic assumptions from (Kushner and Yin, 1987) concerning $F(X(n), n)$ hold in our case. Coming back to the preliminary part of this section, we can easily conclude that the exposed properties of the local transitions formulated in Paragraph 2 of Subsection 3.1 imply that the assumptions (C3.2) and C(3.3') from Section 3 in (Kushner and Yin, 1987) are satisfied in our case.

Furthermore, we introduce

$$M_f(t) = f(X(t)) - f(X(0)) + \quad (17)$$
$$\int_0^t f'_X(X(s)) \text{diag}\{\bar{\Psi} \otimes I_p, I_{Np}\} \bar{F}(X(s)) ds$$

for each real valued function $f(\cdot)$ with compact support and continuous second derivatives. It is possible to show that $M_f(t)$ is a continuous martingale, and that, in addition, $M_f(t) = 0$, implying that $\dot{X} = \text{diag}\{\bar{\Psi} \otimes I_p, I_{Np}\} \bar{F}(X)$. By Lemma 1 and (A6), all the rows of $\bar{\Psi}$ are equal, so that the p-dimensional vector components of Θ must be equal, *i.e.*, we obtain $\Theta(\cdot) = [\theta(\cdot)^T \cdots \theta(\cdot)^T]^T$, $\forall \theta(\cdot) \in \mathcal{R}^p$. This implies that $\theta(\cdot)$ satisfies the first ODE from (14). The remaining ODE's follow from the general results in (Kushner and Yin, 2003, Theorem 8.2.2).

Part 2. We study the limit set of the ODE (14) using the methodology from (Yu, 2017, Proposition 4.1). We introduce

$$V(\theta, w_1, \dots w_N) = \frac{1}{2}\|\theta - \bar{\theta}\|^2 + \frac{1}{2}\sum_{i=1}^N q_i \bar{\psi}_i \|w_i - \bar{w}_i\|^2, \quad (18)$$

where $\bar{\theta}$ and $\bar{w}_i$ are given by (15), giving

$$\dot{V}(\theta, w_1, \dots w_N) = -\sum_{i=1}^N q_i \bar{\psi}_i \langle w_i - \bar{w}_i, H_i(w_i - \bar{w}_i)\rangle. \quad (19)$$

Following (Yu, 2017), we infer that for initial conditions $w_i(0) \in \text{span}\{\phi(s)\}$ the limit set of ODE (14) is the set $\bar{\Sigma}$ satisfying (15).

The remaining steps are standard for the stochastic approximation theory (Yu (2017), (Kushner and Yin, 2003, Theorem 8.2.2)).

Remark 1. A theorem, analogous to Theorem 1, can be formulated for D2-GTD2(λ), in which consensus is applied to both θ_i and w_i. A detailed presentation will be omitted. We shall note here only that consensus on w_i introduces the additional asymptotic constraint $\bar{w}_1(\theta) = \cdots = \bar{w}_N(\theta) = \bar{w}(\theta)$ for any given θ. It is possible to show that the algorithm converges weakly to the set of points $\bar{x} = [\bar{\theta}^T \bar{w}^T]^T \in \mathcal{R}^{2p}$ defined by

$$\bar{G}\bar{\theta} + \bar{g} - \bar{H}\bar{w} = 0, \quad \bar{G}^T\bar{w} = 0, \quad (20)$$

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

$T^{(\lambda_i)}(\cdot)$ and $\bar{H} = \sum_{i=1}^{N} \bar{\psi}_i q_i \Phi^T \Xi_i \Phi$. This implies that the convergence points of θ are, accordingly, different. When all the agents have equal λ-parameters and equal behavior policies, both algorithms provide the same solution for θ.

Remark 2. The proposed multi-agent algorithm can be considered from two viewpoints: 1) as a tool for organizing complementary actions of multiple agents contributing to a common goal, and 2) as a parallelization tool, allowing faster convergence and reduction of the noise influence (by the averaging of the consensus scheme), especially in the problems of large dimensions. In the first case, the agents can have different behavior policies (in the sense of different probabilities of visiting the MDP states), as well as different ways of defining local λ-parameters. The weighting factors $\bar{\psi}_i q_i$ can help to place more emphasis on those agents that can provide more significant contribution to the overall goal. Adequate weights, covering possibly non-overlapping subsets of MDP states, can contribute significantly to the overall rate of convergence. However, one should have in mind that diffusion of information over the network is a dynamic process: the higher the connectedness of the network, the higher the overall convergence rate (Stanković et al., 2016, 2011)

Remark 3. The coefficients $\bar{\psi}_i$ can be defined by appropriate network design, including the network topology and the numerical values of the convexification coefficients. If one adopts a time invariant network with $A(n) = A$, the problem reduces to the definition of the elements of an $N \times N$ matrix A which provides $\bar{\psi}_i = \frac{1}{N}$ (having in mind the freedom in selecting the weights q_i). Then, one has the the standard equation $\mathbf{1}^T A = \mathbf{1}^T$, where $\mathbf{1}^T = [1 \cdots 1]^T$ (see (Stanković et al., 2016) for details). The adopted algorithm formulation allows stochastic sequences $A(n)$, and, therefore, adequate treatment of communication dropouts and different forms of asynchronous gossip communications.

Remark 4. Estimation algorithms based on consensus have, in general, nice "denoising" properties, consisting of the reduction of additive noise influence obtained by averaging over the network. In order to clarify this effect in the case of the proposed algorithm, the *asymptotic convergence rate* of the proposed algorithm can be studied using the results from (Kushner and Yin, 1987, Section 6.1). It is possible to show that, under additional assumptions, we have that asymptotically (when $\alpha \to 0$ and $n \to \infty$) the normalized errors $U^\alpha(n) = \frac{Y^\alpha(n) - \bar{Y}}{\sqrt{\alpha}}$, where $Y^\alpha(n) = [x_1(n)^T \cdots x_N(n)^T]^T$, $x_i(n) = [\theta_i(n)^T w_i(n)^T]^T$ and $\bar{Y} = [\bar{x}^T \cdots \bar{x}^T]^T$, $\bar{x} = [\bar{\theta}^T \bar{w}^T]^T$ converge (in the sense of Theorem 1) to $U(\cdot) = [u(\cdot)^T \cdots u(\cdot)^T]^T$, where $u(\cdot)$ satisfies the following Ito stochastic differential equation

$$du = M u \, dt + dv \tag{21}$$

where $v(\cdot)$ is a Wiener process and M a Jacobian matrix. By analyzing the stationary covariance of u one concludes that the stationary error covariance of the proposed algorithm is lower than the local stationary error covariance in the single agent case, by the factor $\Sigma_{i=1}^{N} E\{\psi_i(n)^2\} < 1$, where $\psi_i(n)$ are elements of each row of the matrix $\Psi(n)$ for n large enough. This is an important property due to the large variance problem of the TD(λ) based algorithms.

Fig. 1. Diagram of the simulated MDP

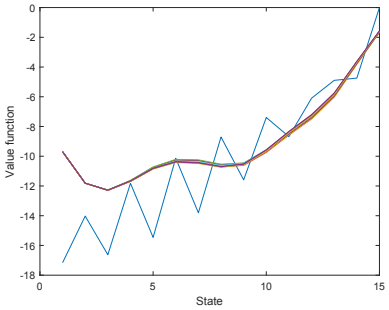

Fig. 2. Value function approximation obtained using D1-GTD2(λ); the agents' behavior policies are such that they can individually visit only a subset of the states. True value function is shown using blue line.

4. SIMULATION RESULTS

In this section we illustrate the main properties of the proposed algorithms by applying them to a version of the Boyan's chain, e.g. (Sutton et al., 2009; Stanković and Stanković, 2016). The diagram of the underlaying Markov chain is shown in Fig. 1 (Stanković and Stanković, 2016).

The discount factor is set to $\gamma = 0.85$. The policy that can be chosen in each state is the probability of choosing the exit action a^{exit} at state s: $\pi(s, a^{\text{exit}})$. The reward for exiting is $r(s, a^{\text{exit}}, s') = -4$ for all s and s', but the probability of staying in the same state is fixed to 0.2. If we choose action a^h the reward is $r(s, a^h, s') = -1$ for all s and s', but the probability of staying in the same state grows with the state number as $1 - \frac{1}{s}$, where s it the state number. The *target policy* is the stationary policy $\pi(s, a^{\text{exit}}) = 0.8$. We assume that there are 10 agents communicating only with a few randomly chosen neighbors with equal weights. They use 7-features Gaussian radial basis representations of the state vector as functions of distances to the states 1, 3, 5, 7, 9, 11 and 13. Note that the chain has an absorbing state; hence we run the algorithms in multiple episodes.

In the first experiment, we demonstrate the case in which the agents, individually, are not able to estimate the value function due to their behavior policies; however, they are able to obtain convergent estimates using the proposed algorithm. The agents can individually visit only a subset of the states, with the following agents' starting and stopping states [(1,3),(2,4),(4,7),(5,15),(5,15),(3,14), (8,15), (1,6),(5,10),(6,11)], with the following stationary behavior policies $[\pi_1(s, a^{\text{exit}}), \pi_2(s, a^{\text{exit}}), ..., \pi_{10}(s, a^{\text{exit}})] = [0.64, 0.75, 0.5, 0.81, 0.85, 0.8, 0.3, 0.55, 0.45, 0.6]$. In Fig. 2 the value function approximation obtained for this case, using D1-GTD2(λ) algorithm, with different λ parameters for each agent: $[0.6, 0.1, 0.25, 0.5, 0.05, 0.01, 0.3, 0.5, 0.4, 0.7]$ and for step sizes $\alpha = \beta = 0.5$, is shown.

Observe that better approximation is obtained for the latter states because the agents visit these states more frequently (with higher probability).

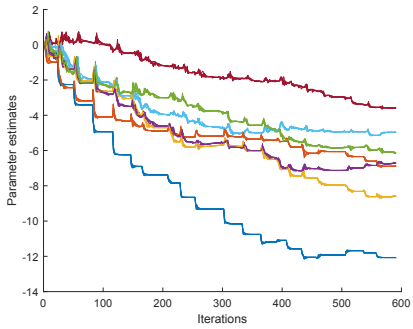

Fig. 3. Parameter estimates for all the agents using D2-GTD2(λ) in the second experiment.

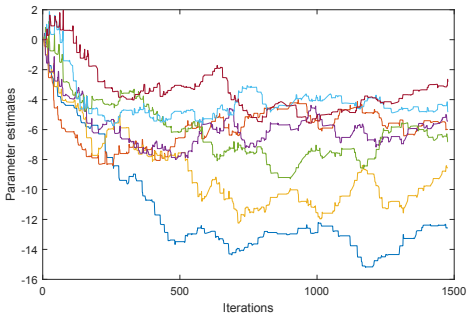

Fig. 4. Parameter estimates obtained in single-agent case using GTD2(λ) algorithm.

In the second experiment we demonstrate the denoising effect of the proposed algorithms. We now assume that the all agents start in state 1 and are able to advance to the final state 15, using the above behavior policies. Fig. 3 shows the parameter estimates $\theta_i(n)$ as functions of the number of iterations n. Observe that 20 episodes were needed for the obtained approximation, which is much less compared to the single agent case (Fig. 4), which also has much larger variance.

REFERENCES

Bertsekas, D.P. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Busoniu, L., Babuska, R., and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(2), 156–172.

Geist, M. and Scherrer, B. (2014). Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Research*, 15, 289–333.

Kar, S., Moura, J.M., and Poor, H.V. (2013). QD-Learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Trans. Signal Proc.*, 61(7), 1848–1862.

Kushner, H.J. and Yin, G. (1987). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM J. Control Optim.*, 25, 1266–1290.

Kushner, H.J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer.

Lee, D., Yoon, H., and Hovakimyan, N. (2018). Primal-dual algorithm for distributed reinforcement learning: Distributed GTD. In *IEEE Conf. Decision and Control*, 1967–1972.

Macua, S.V., Chen, J., Zazo, S., and Sayed, A.H. (2015). Distributed policy evaluation under multiple behavior strategies. *IEEE Trans. Autom. Control*, 60(5), 1260–1274.

Mathkar, A. and Borkar, V.S. (2017). Distributed reinforcement learning via gossip. *IEEE Trans. Autom. Control*, 62(3), 1465–1470.

Nedić, A. and Olshevsky, A. (2015). Distributed optimization over time-varying directed graphs. *IEEE Trans. Autom. Control*, 60, 601 – 615.

Stanković, M.S., Ilić, N., and Stanković, S.S. (2016). Distributed stochastic approximation: Weak convergence and network design. *IEEE Trans. Autom. Control*, 61(12), 4069–4074.

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

Stanković, M.S., Stanković, S.S., and Johansson, K.H. (2018a). Distributed time synchronization for networks with random delays and measurement noise. *Automatica*, 93, 126 – 137.

Stanković, M.S., Stanković, S.S., Johansson, K.H., Beko, M., and Camarinha-Matos, L.M. (2018b). On consensus-based distributed blind calibration of sensor networks. *Sensors*, 18(11).

Stanković, M.S., Stanković, S.S., and Stipanović, D.M. (2015). Consensus-based decentralized real-time identification of large-scale systems. *Automatica*, 60, 219–226.

Stanković, S.S., Beko, M., and Stanković, M.S. (2020). Nonlinear robustified stochastic consensus seeking. *Systems & Control Letters*, 139.

Stanković, S.S., Stanković, M.S., and Stipanović, D.M. (2011). Decentralized parameter estimation by consensus based stochastic approximation. *IEEE Trans. Autom. Control*, 47, 531–543.

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

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

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 *Proc. 26th Int. Conf. on Machine Learning*, 993–1000.

Yu, H., Mahmood, A., and Sutton, R. (2019). On generalized Bellman equations and temporal-difference learning. *Journal of Machine Learning Research*, 19, 1–49.

Yu, H. (2017). On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv:1712.09652*.

Zhang, K., Yang, Z., and Basar, T. (2019). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv:1911.10635*.

Zhang, Y. and Zavlanos, M.M. (2019). Distributed off-policy actor-critic reinforcement learning with policy consensus. *arXiv:1903.09255*.