2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Decentralized Multi-Agent Multi-Task Q-Learning with Function Approximation for POMDPs

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

***Abstract*— In this paper we propose a novel distributed gradient-based two-time-scale algorithm for decentralized multi-agent multi-task learning (MTL) using a linear approximation of the optimal action value function (Q-function) in POMDPs. The algorithm is based on the idea of using in a concurrent way recursive Bayesian state belief filters for estimation of the system model parameters, prediction of the hidden state and definition of the optimal approximation parameters of the local Q-functions. The main MTL algorithm is composed of: 1) local parameter updates based on an off-policy gradient-based learning algorithm with target policy belonging to the greedy or Gibbs classes, and 2) a linear stochastic time-varying consensus scheme for parameters shared between the agents in order to achieve the MTL goal. It is proved, under general assumptions, that the parameter estimates generated by the proposed algorithm weakly converge to a bounded invariant set of the corresponding ordinary differential equations (ODE). Simulation results illustrate the effectiveness of the algorithm.**

I. Introduction

Reinforcement learning (RL) for Markov Decision Processes (MDPs) has become a widely accepted problem solving sample-based tool applicable to unknown and stochastic environments. Numerous successful RL methods for large state and action spaces are based on *(action) value function* and/or *policy function approximation*, using a limited number of *parameters* and reducing the problem to finding *optimal parameter values* (see e.g. [1]). Also, *decentralized and distributed multi-agent* RL algorithms are currently in the focus due to great theoretical and practical challenges [2]–[4]. *Partially observable Markov decision processes* (POMDPs) are a natural generalization of MDPs which assumes *partial state observation* [5]–[7]. When the POMDP parameters are given, the optimal policy is determined by using dynamic programming in *the belief state* [8]. There are numerous contributions to the problem of finding approximate solutions, e.g. [6], [9]. As the belief state is continuous, the optimization problem is computationally very challenging. An intuitively appealing idea is to develop a recursive system model parameter estimator for a POMDP, to predict hidden states, and to obtain the optimal parameters in parallel [8].

M. S. Stanković is with Singidunum University, Belgrade, Serbia; and Universidade Lusófona, Lisboa, Portugal; e-mail: milstank@gmail.com

M. Beko is with Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal; and Universidade Lusófona, 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 No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms - ECOSwarm, and by the Fundação para a Ciência e a Tecnologia under Grant 2022.07530.CEECIND and Project UIDB/04111/2020.

Q-learning has been recognized as a basic, but always promising RL tool [1]. Q-learning iteratively estimates the optimal *Q-function* (*action value function*) [1]. However, applications have been limited to the problems with relatively small state and action spaces. To overcome this, Q-learning with *function approximation* has been treated in many papers [10]–[12], but convergence is guaranteed usually under fairly strong assumptions.

In this paper we present a comprehensive and intuitively logical method for solving *multi-agent multi-task learning* problems for POMDPs in a *recursive, distributed and decentralized* way, using *Q-learning* [10], [11] in conjunction with a dynamic consensus scheme [12], [13]. The method is in the form of a multi-agent network connecting POMDPs with different characteristics performing different tasks, aimed at setting up the inductive bias across tasks by designing a *parameterized common policy* optimizing a *proxy objective* [14]. The proposed methodology incorporates *system parameter identification* at the local level using *belief based prediction* [15] and distributed recursive estimation of parameters of a *linear approximation* of the local Q-functions [11], [12].

More specifically, the paper contains the following main contributions:

a) formulation of a *proxy criterion* for the MTL problem involving *linear Q-function approximation*;

b) proposal of an original consensus-based decentralized two-time-scale algorithm for estimation of the parameters of local Q-function approximations based on belief-based functions derived from the Baum-Welch method [16];

c) formulation of a weak convergence theorem of the parameter estimates to a set of limit ODEs using the Kushner-Yin methodology [17], [18];

d) proposal to apply the Borkar-Meyn fluid model methodology to the stability analysis of the derived ODEs [19].

To the authors knowledge, the formulated distributed multi-agent problem has not yet been treated in the literature using Q-learning with function approximation in the context of POMDPs. Utilization of the *joint conditional probability* of successive states in the construction of the proposed estimation schemes is different from similar schemes found in the literature and possesses superior quality (compare to e.g. [8]). Also, the given formulation of MTL involving *two types of parameters* (shared and local) in the context of consensus-based estimation of the parameters of Q-functions approximation is novel, leading to new methodological elements [17], [18]. The proposed algorithm can also be an efficient general tool for parallelization of Q-function approximation in POMDPs [12], [20].

II. Problem Definition

A. General Setting

Consider N *autonomous agents*, attached one-to-one to *partially observable Markov Decision Processes* (POMDPs) denoted as POMDP$^{(i)}$, $i = 1, \ldots, N$. All these POMDPs are characterized by the septuplets $\Sigma^i = \{\mathcal{S}, \mathcal{A}, P^i(s'|s,a), R^i(s,a,s'), \mathcal{Y}, Y^i(y,s), \gamma^i\}$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ a finite action space, $P^i(s'|s,a) = P\{s^i_{t+1} = s'|s^i_t = s, a^i_t = a)\}$ a state transition probability for POMDP$^{(i)}$ of moving from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ by applying action $a \in \mathcal{A}$, $R^i(s,a,s')$ a reward model with distribution $q^i(\cdot|s',a,s)$, $\mathcal{Y}$ a finite observation space, $Y^i(y,s) = P\{y^i_t = y|s^i_t = s)\}$ the observation probability for POMDP$^{(i)}$ and $\gamma^i \in [0.1)$ a discount factor. At each time $t \in \mathcal{I}^+$ agent i observes $y^i_t \in \mathcal{Y}$ at the state $s^i_t \in \mathcal{S}$, performs action $a^i_t \in \mathcal{A}$ and gets a reward $R^i_t \in \mathcal{R}$ according to $R^i(s,a,s')$. Each POMDP$^{(i)}$, $i = 1, \ldots, N$, applies a local *stationary behavior policy* $\pi^i(a|y)$ (probability of taking action $a^i_t = a$ at observation $y^i_t = y$), implying that the state processes $\{s^i_t\}$ and the state-action processes $\{s^i_t, a^i_t\}$, $i = 1, \ldots, N$, represent time-homogenous Markov chains.

We assume that the agents communicate between them in order to achieve a common goal. The inter-agent communications are formally represented by a *strongly connected digraph* $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ is the set of *nodes* attached to the agents (POMDPs) and $\mathcal{E}$ the set of directed arcs. Let $\mathcal{N}_i \subset \mathcal{N}$ be the *in-neighborhood* of node i [2], [21]. We shall assume *strict Information Structure Constraints* (sISCs) restricting agent i to *observation, action and reward information* only from the local POMDP$^{(i)}$, $i = 1, \ldots, N$.

Recall that the *target policy* for POMDP$^{(i)}$ taken apart is characterized by a stationary distribution $\pi^i(a|s)$. The corresponding *action value function* (*Q-function*) is defined by

$$Q^{\pi^i}(s,a) = E_{\pi^i}\left\{\sum_{t=0}^{\infty} \gamma^{it} r^i(s^i_t, a^i_t)|s_0 = s, a_0 = a\right\}. \quad (1)$$

The *optimal Q-function* $Q^{i*}(s,a)$ satisfies the Bellman equation

$$Q^{i*}(s,a) = r^i(s,a) + \gamma^i \sum_{s'} P^i(s'|s,a) \max_{a'} Q^{i*}(s',a'),$$

where $r^i(s,a)$ denotes the one-step expected reward. The corresponding optimal policy is defined by $\pi^{i*} = \mathrm{Arg}\max_\pi Q^{i*}(s,a)$. In general, the optimal value Q^{i*} can be found using dynamic programming. If the MDP model is unknown, it can be computed by *stochastic approximation*. The so-called *Q-learning* algorithm iteratively provides optimal Q-values in a tabular form [1].

Let $\phi^i : \mathcal{S} \times \mathcal{A} \to \mathcal{R}^{p^i}$ be a function that maps each state-action pair (s,a) to a *feature vector* $\phi^i(s,a)$. We shall consider the *linear approximation* of the Q-function in the form $Q^i_\theta(s,a) = \theta^{iT}\phi^i(s,a)$, $\|\phi^i(s,a)\| < \infty$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\theta^i \in \mathcal{R}^{p^i}$ is a parameter vector ($p^i << |\mathcal{S} \times \mathcal{A}|$). Following [11], we introduce a class of stationary stochastic target policies $\pi^i_\theta(\cdot|s)$ to denote *the action selection probability distribution* at state s. In this paper, we shall focus on: 1) the greedy class, where $\pi^i_\theta(\cdot|s)$ is such that $\mathrm{argmax}_{a' \in \mathcal{A}} Q^i_\theta(s,a')$ is chosen w.p.1, and 2) the Gibbs class, when $\pi^i_\theta(a|s) \sim e^{\kappa(Q^i_\theta(s,a))}$ for some $\kappa(\cdot)$.

B. Multi-Task Learning: Performance Criteria

The goal of multi-task RL is to learn a *common policy* for *multiple tasks* so that it *generalizes well* across all of them [4], [22]. A common way to introduce some *shared parameters* between tasks is to optimize a *proxy objective* [14]. There are many successful approaches to multi-task learning [4], [23], [24]. However, most of the proposed methods assume access to all data to all tasks.

We shall propose in this paper a new completely *decentralized and distributed off-policy method for multi-task RL learning*. Formally, we introduce N *local learning tasks*, aimed at minimizing the following *local criteria* in the form of projected Bellman errors

$$J_i(\theta^i) = \|\Pi_i T^{\pi_{\theta^i}} Q^i_{\theta^i} - Q^i_{\theta^i}\|^2_{\mu_i}, \quad (2)$$

where $\theta^i \in \mathcal{R}^{p^i}$, $i = 1, \ldots, N$, is the local parameter vector attached to POMDP$^{(i)}$, $T^{\pi_{\theta^i}}$ the Bellman operator and $\|Q^i_{\theta^i}\|^2_{\mu^i} = \sum_{s,a} Q^{i2}_{\theta^i}(s,a)\mu^i(s,a)$, where $\mu^i(s,a)$ is the steady-state distribution of the underlying Markov chain, and Π_i is an operator that projects Q-functions into the linear space $\mathcal{F}_i = \{Q^i_{\theta^i} : \theta^i \in R^{p_i}\}$ w.r.t. $\|\cdot\|_{\mu^i}$, i.e. $\Pi_i \hat{Q}^i = \mathrm{Arg}\min_{f_i \in \mathcal{F}_i} \|\hat{Q}^i - f_i\|_{\mu^i}$. In addition, we adopt the principle of *hard parameter sharing* [14], [25] and assume that $\theta^i = [\bar{\theta}^{iT} \vdots \tilde{\theta}^{iT}]^T$, $\dim(\bar{\theta}^i) = \bar{p}$ and $\dim(\tilde{\theta}^i) = \tilde{p}$ (this implies w.l.o.g. that $p^i = p = \bar{p} + \tilde{p}$ in order to simplify notation), where $\bar{\theta}^i$ is the approximation vector *shared with all the remaining agents* and $\tilde{\theta}^i$ are *task-specific* vectors ($i = 1, \ldots, N$). Formally, we have now $J^i(\theta^i) = J^i(\bar{\theta}^i, \tilde{\theta}^i)$ and $Q^i_{\theta^i} = \theta^{iT}\phi^i(s,a) = \bar{\theta}^{iT}\bar{\phi}^i(s,a) + \tilde{\theta}^{iT}\tilde{\phi}^i(s,a)$, where $\bar{\phi}^i(s,a)$ and $\tilde{\phi}^i(s,a)$ are preselected *specific feature vectors*.

At the network level, we introduce the global parameter vector $\Theta = [\theta^{1T} \cdots \theta^{NT}]^T$ and define the following *optimization problem* including consensus w.r.t. $\bar{\theta}^i$

$$\begin{gathered}\min_\Theta J(\Theta) = \min_{\theta^1, \cdots, \theta^N} \sum_{i=1}^N q^i J_i(\theta^i) \quad (3)\\ \text{Subject to } \bar{\theta}^1 = \cdots = \bar{\theta}^N,\end{gathered}$$

where $q^i > 0$ are *a priori* defined weights. The optimal parameter vector $\Theta^* = \arg\min_\Theta J(\Theta)$ provides $\bar{\Theta}^* = [\bar{\theta}^{*T} \cdots \bar{\theta}^{*T}]^T$, $\tilde{\Theta}^* = [\tilde{\theta}^{1*T} \cdots \tilde{\theta}^{N*T}]^T$.

III. HMM Estimation

In this section, we shall provide a short insight into some basic Hidden Markov Model (HMM) concepts applicable to the introduced MTL optimization problem (3). We shall also introduce probabilistic functions derived from the Baum-Welch algorithm relevant for the paper.

Notice that the HMM property is induced into the POMDPs by the behavior policies π^i_b. The corresponding Markov chains are characterized by the state transition probability $P^i_{s's} = P\{s^i_{t+1} = s'|s^i_t = s; \Phi^i\}$, $\forall s, s'$, where Φ^i are the parameters describing Σ^i, and the extended observation

probability $X^i_{xs} = P\{x^i_t = x | s^i_t = s; \Phi^i\}$ matrices, where $x^i = (y^i, a^i, R^i)$ denotes the *extended observation* [8].

(A1) The transition probability matrices $P^i_{s's}$ are aperiodic and irreducible [1].

A. HMM Parameter Identification

When the system model parameters Φ^i are available, the *belief vectors* $u^i_t = [u^i_{t,1} \dots u^i_{t,|\mathcal{S}|}]^T$ are calculated, where $u_{t,j} = P\{s^i_t = j | \mathcal{X}^i_{t-1}; \Phi^i\}$ and $\mathcal{X}^i_{t-1} = (x^i_0, \dots, x^i_{t-1})$, using the recursive Baum-Welch state predictor [8], [15], [26]

$$u^i_{t+1} = \frac{P^{iT} B^i(x^i_t) u^i_t}{b^{iT}(x^i_t) u^i_t}, \tag{4}$$

where $b^i(x^i_t) = [b^i_1(x^i_t) \cdots b^i_{|\mathcal{S}|}(x^i_t)]^T, b^i_j(x^i_t) = P\{x^i_t | s^i_t = j\} = P\{y^i_t | s^i_t = j\} P\{a^i_t | x^i_t\} P\{R^i_t | s^i_t, a^i_t\}$ and $B^i(x^i_t) = \mathrm{diag}\{b^i(x^i_t)\}$.

Identification of the HMM parameter vectors Φ^i can be based on the conditional log-likelihood of a sequence of extended observations $L^i_t(\Phi^i) = \frac{1}{t+1} \log p(l^i_t, \dots, l^i_1; \Phi^i)$, where $p(l^i_t, \dots, l^i_1; \Phi^i) = P\{x^i_t = l^i_t, \ ..., x^i_1 = l^1_t | \ s^i_t, \dots, s^i_1; \Phi^i\}$. It is proved in [15] that $L^i_t(\Phi^i) = \frac{1}{t+1} \sum_{k=1}^t \log(b^{iT}(x^i_k) u^i_k)$, according to (4), having in mind that $p(l^i_t, \dots, l^i_1; \Phi^i) = \prod_{k=1}^t P\{x^i_k = l^i_k | \mathcal{X}_{k-1}; \Phi^i\}$. Following this line of thought, one can formulate a gradient scheme for HMM identification based on stochastic gradient ascent maximizing $L^i_t(\Phi^i)$ [8], [15], [26].

We shall, therefore, take into account two possibilities in practice: 1) identification of the system parameters is done in real time, simultaneously with the learning algorithm, 2) system parameters are considered to be known and fixed.

B. Belief-based Functions

Besides the *a priori* belief function given by (4) (used for prediction), it is possible to define the *a posteriori belief function* $v^i_t = [v^i_{t,1} \cdots v^i_{t,|\mathcal{S}|}]^T$, where $v^i_{t,j} = P\{s^i_t = j | \mathcal{X}^i_t; \Phi^i\}$, $j = 1, \dots, |\mathcal{S}|$, defined by

$$v^i_{t+1} = \frac{B^i(x^i_{t+1}) P^{iT} v^i_t}{b^{iT}(x^i_{t+1}) P^{iT} v^i_t} = \frac{B^i(x^i_{t+1}) u^i_{t+1}}{b^{iT}(x^i_{t+1}) u^i_{t+1}}. \tag{5}$$

A similar forward recursion for the $|\mathcal{S}|$-vector $\alpha^i_{t;j} = P\{x^i_t, \dots, x^i_1, s^i_t = j\}$ is a standard part of the Baum-Welch algorithms [16]:

$$\alpha^i_{t+1} = B^i(x^i_{t+1}) P^{iT} \alpha^i_t, \tag{6}$$

$t \geq 1$, with $\alpha^i_1 = [P\{s^i_1 = 1\} b_1(x_1), \dots, P\{s^i_1 = |\mathcal{S}|\} b_{|\mathcal{S}|}(x_1)]^T$.

The following matrix connected to the belief function will play a fundamental role in the construction of the learning algorithm proposed in the next section

$$\Xi^i_t = [\xi^i_t(j,k)] = \frac{\mathrm{diag}\{\alpha^i_t\} P^i \, \mathrm{diag}\{b^i(x^i_{t+1})\}}{\alpha^{iT}_t P^i b^i(x_{t+1})}, \tag{7}$$

where $\xi^i_t(j,k) = P\{s^i_t = j, s^i_{t+1} = k | \mathcal{X}^i_{t+1}, \Phi^i\}$ [16]. Notice that (7) enables applications to recursions involving one-step look ahead processing such as value iteration or temporal difference methods. Applying (7), we shall be able to get a methodologically consistent link between the original state spaces and the belief spaces for the proposed algorithm.

IV. Algorithm

The main focus of the paper is on the proposal of a new algorithm for decentralized multi-agent multi-task learning based on a linear approximation of Q-functions for POMDPs. To the authors knowledge, explicit application of the joint conditional probability (7) of successive states has not been treated in the literature in connection with temporal difference recursions (the approach to the iterative Q-learning from [8] is similar, but devoted to the single-agent case and utilizes a different belief-based function). The application of (7) has its theoretical, as well as practical advantages.

A. Algorithm Derivation

Starting from (2) we obtain the Fréchet sub-gradients of J_i using the arguments from [11]. Let $\delta^i_{t+1}(\theta^i) = R^i_{t+1} + \gamma^i \hat{\phi}^{iT}_{t+1} \theta^i - \phi^{iT}_t \theta^i$ be the *temporal difference error*, with $\phi^i_t = \phi(s^i_t, a^i_t)$ and $\hat{\phi}^i_{t+1} = \phi(s^i_{t+1}, \hat{a}^i_{t+1})$, where $\hat{a}^i_{t+1}$ is obtained using either the greedy policy or the Gibbs policy [11]. Following further [11], we define $d^i_{t+1} = \gamma \hat{\phi}^i_{t+1} - \phi^i_t$, and obtain

$$\begin{aligned} \partial J_i(\theta^i) &= E\{d^i_{t+1} \phi^{iT}_t\} [E\{\phi^i_t \phi^{iT}_t\}]^{-1} E\{\delta^i_{t+1}(\theta^i) \phi^i_t\} \\ &= E\{\delta^i_{t+1}(\theta^i) \phi^i_t\} + \gamma^i E\{\hat{\phi}^i_{t+1} \phi^{iT}_t\} w^{i*}(\theta^i), \end{aligned} \tag{8}$$

where $w^{i*}(\theta^i) = E\{\phi^i_t \phi^{iT}_t\}^{-1} E\{\delta^i_{t+1}(\theta^i) \phi^i_t\}$. Notice that, in such a way, we come to the condition $\sum_{i=1}^N q^i \partial J_i(\theta_i) = 0$ subject to $\bar{\theta}^1 = \cdots = \bar{\theta}^N$. The new *distributed algorithm* is composed of *two main parts*: 1) *local parameter updates*, based on the *gradient descent* methodology using realizations of (8) and 2) *convexification* of current parameter estimates based on inter-agent communications. The algorithm represents a decentralized multi-agent generalization of the Greedy-GQ algorithm proposed in [11], [12], supposing, in addition, *partial state observation*.

The update part of the algorithm is defined by

$$\theta^i_{t'} = \theta^i_t + \mu^i_t g^i_t(\theta^i_t, w^i_t); \quad w^{i}_{t'} = w^i_t + \nu^i_t h^i_t(\theta^i_t, w^i_t) \tag{9}$$

where

$$g^i_t(\theta^i, w^i) = \sum_j \sum_k \xi^i_t(j,k) \hat{g}^i_t(j,k; \theta^i, w^i), \tag{10}$$

with $\hat{g}^i_t(j,k;\theta^i,w^i) = \delta^i_{t+1}(j,k;\theta^i)\phi^i_t(j) - \gamma^i \hat{\phi}^i_{t+1}(k;\theta^i)\phi^{iT}_t(j) w^i$ with $\delta^i_{t+1}(j,k;\theta^i) = \delta^i_{t+1}(s^i_t = j, s^i_{t+1} = k; \theta^i) = R^i_{t+1} + \gamma^i \hat{\phi}^i_{t+1}(k,\theta^i)^T \theta^i - \phi^{iT}_t(j)\theta^i$, $\phi^i_t(j) = \phi^i(s^i_t = j, a^i_t)$ (a^i_t is generated by the behavior policy π^i_b), $\hat{\phi}^i_{t+1}(k;\theta^i) = \phi^i(s^i_{t+1} = k, \tilde{a}^i_{t+1})$, where $\tilde{a}^i_{t+1}$ is generated by one of the two adopted target policies (e.g.

$$\tilde{a}^i_{t+1} = \mathrm{argmax}_{a'} \sum_j \sum_k \xi^i_t(j,k) \phi^{iT}(s^i_{t+1} = k, a') \theta^i_t \tag{11}$$

for the greedy policy), and

$$h^i_t(\theta^i, w^i) = \sum_j \sum_k \xi^i_t(j,k) \hat{h}^i_t(j,k;\theta^i, w^i) \tag{12}$$

with $\hat{h}^i_t(j,k;\theta^i,w^i) = \delta^i_{t+1}(j,k;\theta^i) - \phi^{iT}_t(j) w^i] \phi^i_t(j)$. The initial values $\theta_i(0)$ and $w_i(0)$ are chosen arbitrarily. The step size sequences $\{\mu^i_t\}$ and $\{\nu^i_t\}$ are composed of positive

numbers which satisfy $\mu_t^i << \nu_t^i$, introducing two time-scales in the algorithm.

The second (convexification) part of the algorithm is given by

$$\bar{\theta}_{t+1}^i = \sum_{j=1}^N a_t^{ij} \bar{\theta}_t'^j; \quad \tilde{\theta}_{t+1}^i = \tilde{\theta}_t'^i, \quad w_{t+1}^i = w_t'^i, \tag{13}$$

having in mind that $\theta^i = [\bar{\theta}^{iT} \tilde{\theta}^{iT}]^T$. In (13) convexification is applied only to the $\bar{\theta}$-iterates in order to achieve consensus w.r.t. $\bar{\theta}$. We shall assume that $a_t^{ij} \geq 0$ are random variables, elements of an $N \times N$ time-varying random matrix $A_t = [a_t^{ij}]$, with general properties specified later (see e.g. [21]).

According to the definition of the joint probability $\xi_t^i(j,k)$ in (7), we conclude that

$$\sum_j \sum_k \xi_t^i(j,k) f_1(j) f_2(k) = E\{f_1(s_t^i) f_2(s_{t+1}^i) | \mathcal{X}_{t+1}^i\} \tag{14}$$

for arbitrary integrable functions $f_1(x_t^i = j)$ and $f_2(x_{t+1}^i = k)$. Therefore, we can write $g_t^i(\theta^i, w^i) = q^i E\{\delta_{t+1}^i(\theta^i)\phi_t^i - \gamma^i \hat{\phi}_{t+1}^i(\theta^i)\phi_t^{iT} w^i | \mathcal{X}_{t+1}^i\}$ and $h_t^i(\theta^i, w^i) = E\{[\delta_{t+1}^i(\theta^i) - \phi_t^{iT} w^i]\phi_t^i | \mathcal{X}_{t+1}^i\}$.

B. Global Model

Let $X_t = [\Theta_t^T \vdots W_t^T]^T$, $\bar{\Theta}_t = [\bar{\theta}_t^{1T} \cdots \bar{\theta}_t^{NT}]^T$, $\tilde{\Theta}_t = [\tilde{\theta}_t^{1T} \cdots \tilde{\theta}_t^{NT}]^T$, $W_t = [w_t^{1T} \cdots w_t^{NT}]^T$ and $X_t' = [\Theta_t'^T \vdots W_t'^T]^T$. Then, we have at the network level the following global model

$$X_t' = X_t + \Gamma_t F_t(X_t), \qquad X_{t+1} = \tilde{C}(A_t) X_t', \tag{15}$$

$X(0) = X_0$, $\Gamma_t = \mathrm{diag}\{\mu_t^1, \ldots, \mu_t^N, \nu_t^1, \ldots, \nu_t^N\} \otimes I_p$, ($\otimes$ denotes the Kronecker's product), $F_t(X_t) = [F_t^{\theta T}(X_t) \vdots F_t^{wT}(X_t)]^T$, $F_t^\theta(X_t) = [F_t^{\theta 1 T}(X_t) \cdots F_t^{\theta N T}(X_t)]^T$, $F_t^w(X_t) = [F_t^{w1T}(X_t) \cdots F_t^{wNT}(X_t)]^T$, with $F_t^{\theta,i}(X_t) = g_t^i(\theta^i, w^i)$ and $F_t^{w,i}(X_t) = h_t^i(\theta^i, w^i)$. In the second relation, $\tilde{C}(A_t) = \left[\begin{array}{c:c} T^{pT} & 0 \\ \hdashline 0 & I_{Np} \end{array}\right] \mathrm{block\,diag}\{A_t \otimes I_{\bar{p}}, I_{N(\bar{p}+p)}\} \left[\begin{array}{c:c} T^p & 0 \\ \hdashline 0 & I_{Np} \end{array}\right]$, where T^p is an $Np \times Np$ permutation matrix satisfying $T^p \Theta = \left[\begin{array}{c} \bar{\Theta} \\ \tilde{\Theta} \end{array}\right]$.

Applying the expectation operator $E_i\{\cdot\}$, where sub-index i denotes the expectation under the probability law induced in POMDP$^{(i)}$, we obtain $\bar{g}^i(\theta^i, w^i) = E_i\{g_t^i(\theta^i, w^i)\}$, $\bar{h}^i(\theta^i, w^i) = E_i\{h_t^i(\theta^i, w^i)\}$; more specifically, we have

$$\bar{g}^i(\theta^i, w^i) = b^i - D^i(\theta^i)\theta^i - \gamma^i B^i(\theta^i) w^i \tag{16}$$

$$\bar{h}^i(\theta^i, w^i) = b^i - D^i(\theta^i)\theta^i - C^i w^i \tag{17}$$

where
$b^i = E_i\{R^i(s_t^i, a_t^i, s_{t+1}^i)\}$, $C^i = E_i\{E\{\phi_t^i \phi_t^{iT} | \mathcal{X}_{t+1}^i\}\}$, $D^i(\theta^i) = C^i - \gamma^i B^i(\theta^i)$, $B^i(\theta^i) = E_i\{E\{\hat{\phi}_{t+1}^i \phi_t^{iT} | \mathcal{X}_{t+1}^i\}\}$.

Accordingly, we introduce $\bar{F}(X) = [\bar{F}^\theta(X)^T \vdots \bar{F}^w(X)^T]^T$, where $\bar{F}^{\theta i}(X) = q^i \bar{g}^i(\theta^i, w^i)$ and $\bar{F}^{wi}(X) = \bar{h}^i(\theta^i, w^i)$, $i = 1, \ldots, N$.

V. Convergence Analysis

Convergence analysis of the proposed distributed algorithm is based on the *weak convergence methodology*. We start from the general results of Kushner and Yin [17], [18] and focus only to the specific properties of the algorithm, including the parameter structure implied by the MTL problem posed. Analysis of the fixed points of the algorithm is shortly considered using the methodology of Borkar and Meyn [19].

A. Consensus Part; Assumptions

(A2) Graph $\mathcal{G}$ is strongly connected.

Define $\Psi_{t|k} = A_t \cdots A_k$ for $t \geq k$, $\Psi_{t|t+1} = I_N$. Let $\mathcal{F}_t$ be an increasing sequence of σ-algebras such that $\mathcal{F}_t$ measures $\{X_k, k \leq t, A_k, k < t\}$.

(A3) There is a scalar $a_0 > 0$ such that $a^{jj}(n) \geq a_0$, and, for $j \neq k$, either $a_t^{jk} = 0$ or $a_t^{ij} \geq a_0$, as well as a scalar $p_0 > 0$ and an integer t_0 such that $P_{\mathcal{F}_t}$(agent k communicates to agent j on the interval $[t, t+t_0]) \geq p_0$, for all t and for $j, k = 1, \ldots N$.

By [18, Lemma 2.1] and [21], $\Psi_k = \lim_t \Psi_{t|k}$ exists with probability 1 (w.p.1) and its rows are all equal; moreover, $E\{|\Psi_{t|k} - \Psi_k|\}$ and $E_{\mathcal{F}_k}\{|\Psi_{t|k} - \Psi_k|\} \to 0$ geometrically as $t - k \to \infty$, uniformly in k (w.p.1); also, $E_{\mathcal{F}_k}\{\Psi_{t|k}\}$ converges to Ψ_k geometrically, uniformly in k, as $t \to \infty$ ($|\cdot|$ denotes the infinity norm).

(A4) There is a $N \times N$ matrix $\bar{\Psi}$ such that $E\{|E_{\mathcal{F}_k}\{\Psi_t\} - \bar{\Psi}|\} \to 0$ as $t - k \to \infty$ (it has the form $\bar{\Psi} = [\hat{\Psi}^T \cdots \hat{\Psi}^T]^T$, where $\hat{\Psi} = [\bar{\psi}_1 \cdots \bar{\psi}_N]^T$).

In the following, we shall adopt that $\bar{\psi}_i = 1/N$, $i = 1, \ldots, N$, in order to avoid ambiguities w.r.t. to the weights q^i in the products $\bar{\psi}_i q^i$.

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

(A6) Sequence $\{X_t\}$ is tight.

(A7) Matrices C^i are nonsingular, $i = 1, \ldots, N$.

(A8) Matrices $\sum_{i=1}^N q_i D^i$ and D^i, $i = 1, \ldots, N$, are nonsingular and bounded.

B. Convergence Proof

Theorem 1: Let (A1)–(A8) hold.

a) *Fast time-scale.* Let W_t^ν be generated by (9) and (13) for arbitrary $\theta_t^i = \theta_0^i$, $i = 1, \ldots, N$, using $\nu_t = \nu > 0$. Then, $W^\nu(\tau) = W_t$ for $\tau \in [t\nu, (t+1)\nu)$, $\tau \in \mathcal{R}^+$, is tight and converges weakly to $W(\cdot) = [w^1(\cdot)^T \cdots w^N(\cdot)^T]^T$ generated by

$$\dot{w}^i = \bar{h}^i(\theta^i, w^i), \tag{18}$$

for any given $\theta_t^i = \theta_0 \ \forall t \in \mathcal{I}^+$ and w_0^i, $i = 1, \ldots, N$.

b) *Slow time-scale.* Let Θ_t^μ be generated by (9) and (13) for $w_t^i = w_t^{i*}$, $i = 1, \ldots, N$, using step size $\mu_t = \mu > 0$, $\mu << \nu$. Then $\Theta^\mu(\tau) = \Theta_t$ for $\tau \in [(t - t_\mu)\mu, (t - t_\mu + 1)\mu)$, where $\mu^{\frac{1}{2}} t_\mu \to 0$ as $\mu \to 0$, is tight and converges weakly to $\Theta(\cdot) = [\theta^1(\cdot)^T \cdots \theta^N(\cdot)^T]^T$, $\theta^i(\cdot) = [\bar{\theta}(\cdot)^T \ \tilde{\theta}^i(\cdot)^T]^T$, where

$$\dot{\bar{\theta}} = \tfrac{1}{N} \textstyle\sum_{i=1}^N q^i \bar{g}_1^i(\bar{\theta}, \tilde{\theta}^i, w^{i*}(\bar{\theta}, \tilde{\theta}^i)), \tag{19}$$

$$\dot{\tilde{\theta}}^i = \bar{g}_2^i(\bar{\theta}, \tilde{\theta}^i, w^{i*}(\bar{\theta}, \tilde{\theta}^i)), \tag{20}$$

$i = 1, \ldots, N$, with arbitrary initial conditions $\bar{\theta}_0$ and $\tilde{\theta}_0^i$, where $\bar{g}^i = [\bar{g}_1^{iT} \bar{g}_2^{iT}]^T$, $\dim(\bar{g}_1^i) = \bar{p}$, $\dim(\bar{g}_2^i) = \tilde{p}$; $w^{i*}(\theta^i)$ is the unique solution of the equation $\bar{h}_i(\theta^i, w^i) = b^i - D^i(\theta^i)\theta^i - C^i w^i = 0$.

Proof: At the start, we conclude that the assumptions C(3.2), C(3.3) and C(3.4) from [18, Theorem 3.1, Part 1] are satisfied. Consequently, it is possible to show using [18] that $\sup_{\mu, t \geq t_\mu} \frac{1}{\mu^2} E\{|X_{t+1} - X_t|^2\} < \infty$ and $\{\frac{1}{\mu}|X_{t+1} - X_t|, t \geq t_\mu\}$ is uniformly integrable, $\{X^\mu(\cdot)\}$ is tight and the limit paths are Lipschitz continuous.

a) ODE in (18) follows directly from the assumption about two time-scales in the algorithm. Existence and uniqueness of the solution to $\bar{h}^i(\theta^i, w^i) = 0$ w.r.t. w^i follows from (A7) implying that the ODE $\dot{w}^i = \bar{h}^i(\theta^i, w^i)$ admits a unique, globally asymptotically stable equilibrium w^{i*}, given any fixed value of θ^i.

b) Let $\bar{F}^{[1]}(\cdot)$ be obtained from $\bar{F}(\cdot)$ as $\bar{\Theta}$ is obtained from Θ. Following [18], the asymptotic mean ODE (19) can be obtained by demonstrating that the $M(\tau)$, $\tau \in \mathcal{R}^+$, defined by

$$M(\tau) = f(\bar{\Theta}(\tau)) - f(\bar{\Theta}(0)) + \int_0^\tau f'(\bar{\Theta}(s))\{\bar{\Psi} \otimes I_{\bar{p}}\}\bar{F}^{[1]}(\bar{\Theta}(s))ds, \quad (21)$$

is a Lipschitz-continuous martingale where $f(\cdot)$ a real valued function [18]. The technical part of the derivation is based on the Skorokhod embedding. As $X(\tau)$ is Lipschitz continuous and $M(0) = 0$, it follows that $M(\tau) = 0$. This implies that $\dot{\bar{\Theta}} = (\bar{\Psi} \otimes I_{\bar{p}})\bar{F}^{[1]}(\bar{\Theta})$. By (A2)–(A4), all the rows of $\bar{\Psi}$ are equal. It follows that $\bar{\Theta}(\cdot) = [\bar{\theta}(\cdot)^T \cdots \bar{\theta}(\cdot)^T]$ and that $\bar{\theta}(\cdot)$ satisfies the ODE from (19). ODEs (20) generating $\tilde{\theta}_t^i$ are obtained directly from the corresponding recursions. ■

In the following, we shall shortly indicate the main line of thought for the analysis of the fixed points of the ODEs (19) using fluid models [19]. Namely, if the following limits exist

$$\lim_{c\to\infty} \frac{1}{cN} \sum_{i=1}^N q^i \bar{g}_1^i(c\theta^i, w_i^*(c\theta^i)) = \bar{g}_1^{i,\infty}(\theta^i),$$
$$\lim_{c\to\infty} \frac{1}{c} \bar{g}_2^i(c\theta^i, w_i^*(c\theta^i)) = \bar{g}_2^{i,\infty}(\theta^i), \quad (22)$$

$i = 1, \ldots, N$, and the convergence is uniform, we want to show that zero is the unique globally exponentially stable equilibrium to $\dot{\bar{\Theta}} = \bar{G}^\infty(\bar{\Theta})$, where $\bar{G}^\infty = [\bar{g}^{1,\infty T} \cdots \bar{g}^{N,\infty T}]^T$.

Let $\tilde{D}^{i,\infty}(\theta^i) = (C^i)^{-\frac{1}{2}} D^{i,\infty}(\theta^i) = \left[\begin{array}{c:c} \tilde{D}^{i,\infty}_{[11]} & \tilde{D}^{i,\infty}_{[12]} \\ \hdashline \tilde{D}^{i,\infty}_{[21]} & \tilde{D}^{i,\infty}_{[22]} \end{array}\right]$, with $\dim(\tilde{D}_{[11]}) = \bar{p} \times \bar{p}$, $\dim(\tilde{D}_{[12]}) = \bar{p} \times \tilde{p}$, $\dim(\tilde{D}_{[21]}) = \tilde{p} \times \bar{p}$, $\dim(\tilde{D}_{[22]}) = \tilde{p} \times \tilde{p}$.

(A9) $D^{i,\infty}(\theta^i) = \lim_{c\to\infty} D^i(c\theta^i)$ and the convergence is uniform.

Define $\tilde{N} = [\tilde{N}^{jk}]$ with $(N+1) \times (N+1)$ block-matrices $\tilde{N}^{jk}$ with compatible dimensions defined by: $\tilde{N}^{11} = \frac{1}{N} \sum_{i=1}^N q^i \tilde{D}^{i,\infty T}_{[11]} \tilde{D}^{i,\infty}_{[11]}$, $\tilde{N}^{1k} = \tilde{D}^{k-1,\infty T}_{[11]} \tilde{D}^{k-1,\infty}_{[12]}$, $\tilde{N}^{j,1} = \tilde{D}^{j-1,\infty T}_{[22]} \tilde{D}^{j-1,\infty}_{[21]}$, $\tilde{N}^{jj} = \tilde{D}^{j-1,\infty T}_{[22]} \tilde{D}^{j-1,\infty}_{[22]}$, $j, k = 2, \ldots, N+1$.

Fig. 1. Diagram of the simulated POMDPs.

Following [11], we have that

$$\bar{g}_1^{i,\infty}(\theta^i) = \frac{1}{N} \sum_{i=1}^N q^i \nabla_{\bar{\theta}} \| \frac{1}{c}(b^i - B^i(c\theta^i)w^{i*}) - D^i(c\theta^i)\theta^i \|^2_{(C^i)^{-1}},$$
$$\bar{g}_2^{i,\infty}(\theta^i) = \nabla_{\tilde{\theta}^i} \| \frac{1}{c}(b^i - B^i(c\theta^i)w^{i*}) - D^i(c\theta^i)\theta^i \|^2_{(C^i)^{-1}}.$$

Letting $c \to \infty$, our task is, therefore, to analyze asymptotic stability of

$$\dot{\Theta} = -\tilde{N}\Theta. \quad (23)$$

Properties of the composite matrix $\tilde{N}$ depend on the correlation between the *feature vectors* $\bar{\phi}^i(s,a)$ and $\tilde{\phi}^i(s,a)$. Notice that in the case when there are no task specific parameters ($\theta^i = \bar{\theta}^i$, $i = 1, \ldots, N$), the matrix $\tilde{N}$ reduces to the block $\tilde{N}^{11}$, so that the results from [11], [19] can be directly applied. The general stability problem remains to be elaborated elsewhere, due to the lack of space.

Remark 1: It should be pointed out that the proposed multi-agent algorithm can be considered as an efficient *parallelization tool*. In this case, all the local POMDPs have equal model parameters, but different behavior policies independent of the main target policy. It has been found that the algorithm enables a more efficient exploration of the state space and reduction of variance compared to the single agent case (similar effects as in $A2C$ or $A3C$ algorithms [5]).

VI. Simulation

In this section we shall illustrate the proposed algorithm using simulations. The underlying POMDP is assumed to belong to a class of Boyan chains [27] depicted in Fig. 1. POMDP is modeled as in [2], [20], additionally assuming that the states are not directly observable.

In each state with odd number there are two possible actions: either a^h (staying on the current main route) or a^{exit} (exiting and using an alternative route). The goal state, where the process ends, is state 15. Opting for action a^{exit} results in a reward of -2.5 across all the states, with a probability of 0.2 ($p^{\text{exit}}_{\text{stuck}}$) of remaining in the same state. Action a^h yields a reward of -1 for any state transition, and the probability of remaining in the same state increases as $1 - \frac{1}{s}$. The control policy to be optimized is determined by maximizing the expectation of the current Q-function estimate on the belief state distribution as has been specified in (11). The discount factor $\gamma = 0.9$ is used. The feature vectors utilized in the Q-function approximation are 14-dimensional, with 7 dimensions allocated for each of the two potential actions. Each dimension is represented using Gaussian radial basis functions defined as $e^{-\frac{(s-z_i)^2}{2\sigma^2}}$, where i ranges from 1 to 7, z_i takes values from the set $1, 3, 5, 7, 9, 11, 13$, and $\sigma^2 = 2$.

Simulations are conducted across multiple episodes since state 15 is absorbing.

We analyzed the performance of the proposed algorithm (9), (13) with 10 agents, assuming identical POMDPs and communications according to a sparse time-invariant communication graph. We have assumed that all the parameters in θ^i are involved in the consensus scheme (13), i.e. $\theta^i = \bar{\theta}^i$ for all $i = 1, ..., 10$. The agents use different stationary behavior policies (exit probabilities for each observation) as follows: $\pi_b(a^{\text{exit}}|y) = [0.15, 0.24, 0.13, 0.38, 0.55, 0.89, 0.64, 0.97, 0.75, 0.69]$. The HMM parameter estimation is performed according to Section III. The step sizes in (9) are set to $\mu = 0.001$ and $\nu = 1.5$ (respecting the need for two time-scales). In Fig. 2 we compare two cases: 1) The agents cannot perfectly observe the state (location); the observation space is assumed to be the same as the state space but an agent can wrongly observe one of the two neighboring locations with 0.25 probability each, except in state 1 in which state 2 is observed with probability 0.5; and 2) The state is fully observable. The figure shows the evolution of the value of the policy (11) found at every 100 steps by calculating the mean discounted rewards received under 500 episodes (with the policy fixed by the choice of the Q estimate at the corresponding step). The figure also shows the true optimal value function under full observability and full knowledge of the underlying MDP model. In the case of full observability, as expected, we have convergence close to the true optimal value/policy (exact convergence to the optimum cannot be achieved due to the Q-function parametrization), while in the case of partial observability we also achieve very good performance. It has been observed that larger network connectivity increases the rate of convergence and noise immunity of the scheme.

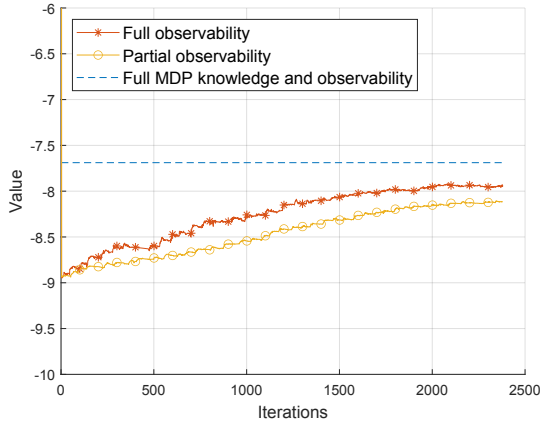

Fig. 2. Evolution of the mean value functions corresponding to the agents' policy estimates under the two described scenarios.

REFERENCES

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

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

[3] M. S. Stanković, M. Beko, N. Ilić, and S. S. Stanković, "Distributed consensus-based multi-agent temporal-difference learning," *Automatica*, vol. 151, p. 110922, 2023.

[4] 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.

[5] V. Mnih, K. Karavukcuoglu, D. Silver, A. Rusu, J. Veness, J. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–535, 2022.

[6] H. Yu and D. P. Bertsekas, "Learning policies for partially observable environments: scaling up," in *Proc. Conf. on Uncertainty in AI*, 2004, pp. 619–627.

[7] M. Kayaalp, F. Ghadieh, and A. H. Sayed, "Policy evaluation in decentralized POMDPs with belief sharing," *IEEE Open Journal on Control Systems*, vol. 2, pp. 125–145, 2023.

[8] D. Yoon, H. Lee, and N. Hovakimyan, "Hidden markov model estimation-based q-larning for partially observable decision process," in *IEEE Conf. Decision and Control*, 2018, pp. 1967–1972.

[9] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: scaling up," in *Machine Learning Proceedings 1995*, 1995, pp. 362–370.

[10] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *25th Intern. Conf. Machine Learning*, 2008.

[11] H. R. Maei, C. Szepesvari, S. Bhatnagar, and R. S. Sutton, "Toward off policy learning control with function approximation," in *Proc. Intern. Conf. Machine Learning*, 2010, pp. 719–726.

[12] M. S. Stanković, M. Beko, and S. S. Stanković, "Distributed multi-agent gradient based Q-learning with linear function approximation," in *2024 European Control Conference (ECC)*, 2024, pp. 2500–2505.

[13] 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.

[14] O. Sener and V. Koltun, "Multi task learning as multi objective optimization," in *Proc. 32nd Conf. Neural Inf. Proc. Sys.*, 2018.

[15] F. LeGland and L. Mevel, "Recursive estimation of hidden Markov models," in *Proc. Conf. Decision and Control*, 1997, pp. 3468–3473.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

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

[18] ——, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM J. Control Optim.*, vol. 25, pp. 1266–1290, 1987.

[19] V. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control And Optimization*, vol. 38, pp. 447–469, 2000.

[20] M. S. Stanković, M. Beko, N. Ilić, and S. S. Stanković, "Multi-agent off-policy actor-critic algorithm for distributed multi-task reinforcement learning," *European Journal of Control*, vol. 74, p. 100853, 2023.

[21] 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.

[22] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: a survey," *J. of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[23] Y. W. Teh, V. Bapst, W. M. Czarnecki, J. Quan, J. Kirkpatrick, R. Hadsell, N. Heess, and R. Pascanu, "Distral: Robust multitask reinforcement learning," *arXiv:1707.04175*, 2017.

[24] S. El-Bsat, H. Bou-Ammar, and M. E. Taylor, "Scalable multitask policy gradient reinforcement learning," in *Proc. AAAI Conf. on Artificial Intelligence*, 2017, pp. 1847–1853.

[25] S. Ruder, "An overview of multitask learning in deep neural networks," *arXiv:1706.05098*, 2017.

[26] V. Krishnamurthy and G. G. Yin, "Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime," *IEEE Trans. Information Theory*, vol. 48, pp. 458–476, 2002.

[27] J. A. Boyan, "Technical update: Least-squares temporal difference learning," *Machine learning*, vol. 49, pp. 233–246, 2002.