2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Regret Analysis for Learning in a Multi-Agent Linear-Quadratic Control Problem

Seyed Mohammad Asghari, Mukul Gagrani, and Ashutosh Nayyar

***Abstract*— We consider a multi-agent Linear-Quadratic (LQ) reinforcement learning problem consisting of three systems, an unknown system and two known systems. In this problem, there are three agents – the actions of agent 1 can affect the unknown system as well as the two known systems while the actions of agents 2 and 3 can only affect their respective co-located known systems. Further, the unknown system's state can affect the known systems' state evolution. In this paper, we are interested in minimizing the infinite-horizon average cost. We propose a Thompson Sampling (TS)-based multi-agent learning algorithm where each agent learns the unknown system's dynamics independently. Our result indicates that the expected regret of our algorithm is upper bounded by $\tilde{O}(\sqrt{T})$ under certain assumptions, where $\tilde{O}(\cdot)$ hides constants and logarithmic factors. Numerical simulations are provided to illustrate the performance of our proposed algorithm.**

I. Introduction

Many modern control systems such as networked control systems and teams of autonomous systems consist of a group of agents acting in collaboration with each other to achieve a common goal under uncertainty [1]. Such systems have motivated the investigation of multi-agent (decentralized) control problems under various information structures [2–6]. Most of these works assume that the system model is known precisely to all the agents in the system. However, for most real-world systems the model and its parameters are often not known perfectly to the agents. Reinforcement learning provides a framework for controlling a dynamical system in the absence of perfect knowledge of system parameters. There exists a rich body of work in the field of multi-agent reinforcement learning where the system is usually modeled as a multi-agent Markov Decision Process (MDP) or a team Markov game [7–11]. However, these works mostly deal in a finite state space and action space setting and cannot be extended trivially to a system with continuous state/action space.

The adaptive control of a single-agent (centralized) linear quadratic (LQ) control problem has been well-studied [12], [13]. However, many of the available results are asymptotic in nature and do not take into account the performance during learning. Recently [14–17] have used online learning methods for single-agent LQ control problems which provide finite-time guarantees on the cost achieved by the learning algorithm. Among these is the idea of Thompson Sampling (TS) which has gained wide attention due to its computational efficiency and performance. TS based algorithms for single-agent LQ control problems have been proposed in [16–19] which achieve a regret of $\tilde{O}(\sqrt{T})$ over a time horizon of T. Here $\tilde{O}(\cdot)$ hides constants and logarithmic factors. This regret scaling is believed to be optimal for single-agent control LQ problems except for logarithmic factors.

We consider a multi-agent Linear-Quadratic (LQ) reinforcement learning problem consisting of three systems, an unknown system and two known systems. In this problem, there are three agents – the actions of agent 1 can affect the unknown system as well as the two known systems while the actions of agents 2 and 3 can only affect their respective co-located known systems. Further, the unknown system's state can affect the known systems' state evolution. In this paper, we are interested in minimizing the infinite-horizon average cost. Variations of this problem setting where the dynamics of all systems are known have been studied in the literature [20–22]. For our multi-agent learning problem, we propose a Thompson Sampling (TS)-based multi-agent learning algorithm where each agent learns the unknown system's dynamics independently. Our result indicates that the expected regret of our algorithm is upper bounded by $\tilde{O}(\sqrt{T})$ under certain assumptions, where $\tilde{O}(\cdot)$ hides constants and logarithmic factors. Numerical simulations are provided to illustrate the performance of our proposed algorithm.

A. Notation

The collection of matrices $A^1, A^2, \ldots, A^N$ (resp. vectors $X^1, X^2, \ldots, X^N$) is denoted as $A^{1:N}$ (resp. $X^{1:N}$). Given column vectors $X^1, X^2, \ldots, X^N$, the notation $\mathbf{vec}(X^{1:N})$ is used to denote the column vector formed by stacking vectors $X^1, X^2, \ldots, X^N$ on top of each other. For random variable/vector X, $\mathbb{E}[X]$ and $\mathbf{cov}(X)$ denote the expectation of X and the covariance matrix of X, respectively. For a strategy π, we use $\mathbb{E}^{\pi}[\cdot]$ to indicate that the expectation depends on the choice of π. We use $\mathbf{I}$ to denote the identity matrix and $\mathbf{0}$ to denote the zero matrix as well as the zero vector.

For two symmetric matrices A and B, $A \succeq B$ (resp. $A \succ B$) means that $(A - B)$ is positive semi-definite (PSD) (resp. positive definite (PD)). The transpose, trace, and spectral norm of matrix A are denoted by $A^{\intercal}$, $\mathbf{tr}(A)$, and $\|A\|$, respectively. For a block matrix A, we use $[A]_{m,n}$ to denote the block located at the m-th block row and n-th block column of A. Consider matrices P, Q, R, A, B of appropriate dimensions with P, Q being PSD matrices

S. M. Asghari, M. Gagrani, and A. Nayyar are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA. Email: asgharip@usc.edu; mgagrani@usc.edu; ashutosn@usc.edu.

This work was supported by NSF Grants ECCS 1509812 and ECCS 1750041.

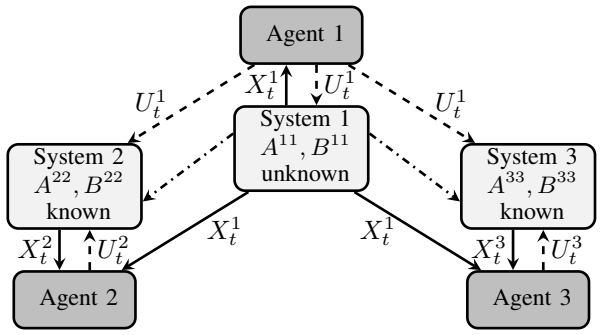

Fig. 1. Three-agent system model. Solid lines indicate communication links, dashed lines indicate control links, and dash-dot lines indicate that one system can affect another one.

and R being a PD matrix. We define $\Upsilon(P,Q,R,A,B)$ and $\Psi(P,R,A,B)$ as follows:

$$\begin{aligned}\Upsilon(P,Q,R,A,B) :=& Q + A^\intercal PA - \\ & A^\intercal PB(R+B^\intercal PB)^{-1}B^\intercal PA.\\ \Psi(P,R,A,B) :=& -(R+B^\intercal PB)^{-1}B^\intercal PA.\end{aligned}$$

Note that $P = \Upsilon(P,Q,R,A,B)$ is the discrete-time algebraic Riccati equation.

II. Problem Formulation

Consider a multi-agent linear system consisting of three systems as shown in Figure 1. The linear dynamics of system 1 are given by

$$X_{t+1}^1 = A^{11}X_t^1 + B^{11}U_t^1 + W_t^1, \tag{1}$$

and the linear dynamics of systems 2 and 3 are given by

$$\begin{aligned}X_{t+1}^2 =& A^{21}X_t^1 + A^{22}X_t^2 + B^{21}U_t^1 + B^{22}U_t^2 + W_t^2,\\ X_{t+1}^3 =& A^{31}X_t^1 + A^{33}X_t^3 + B^{31}U_t^1 + B^{33}U_t^3 + W_t^3,\end{aligned} \tag{2}$$

where, for $n \in \{1,2,3\}$, $X_t^n \in \mathbb{R}^{d_X^n}$ is the state of system n and $U_t^n \in \mathbb{R}^{d_U^n}$ is the action of agent n. The matrices $A^{n1}, A^{nn}, B^{n1}, B^{nn}$, $n \in \{2,3\}$, of systems 2 and 3 are known matrices with appropriate dimensions. However, $A^{11} \in \mathbb{R}^{d_X^1 \times d_X^1}$ and $B^{11} \in \mathbb{R}^{d_X^1 \times d_U^1}$ are *unknown* matrices of system 1. We assume that the initial states $X_1^{1:3}$ are zero and for $n \in \{1,2,3\}$, W_t^n, $t \geq 1$, is the system n-th noise which has zero-mean and covariance matrix $\mathbf{cov}(W_t^n) = \mathbf{I}$. Furthermore, the collection of random variables $W_{1:t}^{1:3}$, $t \geq 1$, are independent.

The overall system dynamics can be written as,

$$X_{t+1} = AX_t + BU_t + W_t \tag{3}$$

where we have defined

$$A = \begin{bmatrix} A^{11} & \mathbf{0} & \mathbf{0} \\ A^{21} & A^{22} & \mathbf{0} \\ A^{31} & \mathbf{0} & A^{33} \end{bmatrix}, \quad B = \begin{bmatrix} B^{11} & \mathbf{0} & \mathbf{0} \\ B^{21} & B^{22} & \mathbf{0} \\ B^{31} & \mathbf{0} & B^{33} \end{bmatrix}, \tag{4}$$

and $X_t = \mathbf{vec}(X_t^{1:3}), U_t = \mathbf{vec}(U_t^{1:3}), W_t = \mathbf{vec}(W_t^{1:3})$.

At each time t, the state X_t^1 of system 1 is directly observed by all the agents. Also, agents 2 and 3 perfectly observe the state of their respective co-located systems. Agent n's action U_t^n at time t is a function π_t^n of its information H_t^n, that is, $U_t^n = \pi_t^n(H_t^n)$ where

$$\begin{aligned}H_t^1 &= \{X_{1:t}^1, U_{1:t-1}^1\},\\ H_t^n &= \{X_{1:t}^1, X_{1:t}^n, U_{1:t-1}^1, U_{1:t-1}^n\}, \quad n \in \{2,3\}.\end{aligned} \tag{5}$$

Let $\pi = (\pi^1, \pi^2, \pi^3)$ where $\pi^n = (\pi_1^n, \pi_2^n, \ldots)$.

At time t, the system incurs an instantaneous cost $c(X_t, U_t)$, which is a quadratic function given by

$$c(X_t, U_t) = X_t^\intercal Q X_t + U_t^\intercal R U_t, \tag{6}$$

where Q is a known symmetric positive semi-definite (PSD) matrix and R is a known symmetric positive definite (PD) matrix with the following structure,

$$Q = \begin{bmatrix} Q^{11} & Q^{12} & Q^{13} \\ Q^{21} & Q^{22} & Q^{23} \\ Q^{31} & Q^{32} & Q^{33} \end{bmatrix}, \quad R = \begin{bmatrix} R^{11} & R^{12} & R^{13} \\ R^{21} & R^{22} & R^{23} \\ R^{31} & R^{32} & R^{33} \end{bmatrix}. \tag{7}$$

A. The Optimal Multi-Agent Linear-Quadratic Problem

Let $\Theta := [A^{11}, B^{11}]$ be the dynamics parameter of the system 1. When Θ is known to the agents, minimizing the infinite horizon average cost is a multi-agent (decentralized) stochastic Linear-Quadratic (LQ) problem. Let $J(\Theta)$ be the optimal infinite horizon average cost under Θ, that is,

$$J(\Theta) = \inf_{\pi} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{\pi}[c(X_t, U_t)|\Theta]. \tag{8}$$

We make the following assumption about the multi-agent stochastic LQ problem.

Assumption 1. *$(A, Q^{1/2})$ is detectable and (A,B) is stabilizable. Furthermore, $(A^{nn}, (Q^{nn})^{1/2})$ is detectable and (A^{nn}, B^{nn}) is stabilizable for $n \in \{2,3\}$.*

Lemma 1. *Under Assumption 1, the optimal infinite horizon cost $J(\Theta)$ is given by*

$$J(\Theta) = \mathbf{tr}\left([P(\Theta)]_{1,1}\right) + \mathbf{tr}(\tilde{P}^2) + \mathbf{tr}(\tilde{P}^3), \tag{9}$$

where $P(\Theta)$, $\tilde{P}^2$, and $\tilde{P}^3$ are the unique PSD solutions to the following Ricatti equations:

$$P(\Theta) = \Upsilon(P(\Theta), Q, R, A, B), \tag{10}$$

$$\tilde{P}^n = \Upsilon(\tilde{P}^n, Q^{nn}, R^{nn}, A^{nn}, B^{nn}), \quad n \in \{2,3\}. \tag{11}$$

The optimal strategies π^* are given by

$$\begin{bmatrix} U_t^1 \\ U_t^2 \\ U_t^3 \end{bmatrix} = \begin{bmatrix} K^1(\Theta) \\ K^2(\Theta) \\ K^3(\Theta) \end{bmatrix} \begin{bmatrix} X_t^1 \\ \hat{X}_t^2 \\ \hat{X}_t^3 \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \tilde{K}^2(X_t^2 - \hat{X}_t^2) \\ \tilde{K}^3(X_t^3 - \hat{X}_t^3) \end{bmatrix}, \tag{12}$$

where the gain matrices $K(\Theta) := \begin{bmatrix} K^1(\Theta) \\ K^2(\Theta) \\ K^3(\Theta) \end{bmatrix}$, $\tilde{K}^2$, and $\tilde{K}^3$ are given by

$$K(\Theta) = \Psi(P(\Theta), R, A, B), \tag{13}$$

$$\tilde{K}^n = \Psi(\tilde{P}^n, R^{nn}, A^{nn}, B^{nn}), \quad n \in \{2,3\}. \tag{14}$$

Furthermore $\hat{X}_t^n = \mathbb{E}^{\pi^*}[X_t^n | H_t^1, \Theta]$, $n \in \{2,3\}$, *is the estimate (conditional expectation) of* X_t^n *given* H_t^1 *and* Θ. *The estimates* $\hat{X}_t^n$, $n \in \{2,3\}$, *can be computed recursively according to*

$$\hat{X}_1^n = \mathbf{0}, \quad \hat{X}_{t+1}^n = A^{n1} X_t^1 + A^{nn} \hat{X}_t^n + \Big(B^{n1} K^1(\Theta) + B^{nn} K^n(\Theta) \Big) \mathbf{vec}(X_t^1, \hat{X}_t^2, \hat{X}_t^3) \quad (15)$$

The proof is omitted due to the space limitation (see [23] for a proof).

B. The Multi-Agent Reinforcement Learning Problem

The problem we are interested in is to minimize the infinite horizon average cost when the matrices A^{11} and B^{11} of system 1 are unknown. In this case, the control problem can be seen as a Multi-Agent Reinforcement Learning (MARL) problem where all the three agents need to learn the system parameter $\Theta = [A^{11}, B^{11}]$ in order to minimize the infinite horizon average cost.

We adopt a Bayesian setting and assume that there is a prior distribution μ_1 for Θ. Since the actual parameter Θ is unknown, we define the expected regret of a (potentially randomized) policy $\pi = (\pi^1, \pi^2, \pi^3)$ up to time T as follows:

$$R(T, \pi) = \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(X_t, U_t) - T J(\Theta) \right], \quad (16)$$

which is the expected difference between the performance of the agents under policy π and the optimal infinite horizon cost under full information about the parameter Θ of system 1. Thus, the regret can be interpreted as a measure of the cost of not knowing system 1. The above expectation is with respect to the random noise of the overall system ($W_{1:T}$), the prior distribution μ_1, and randomization in the agents' strategies. The learning objective is to find a multi-agent strategy that minimizes the expected regret.

III. Main Results

In this section, we propose the TS-MARL algorithm which is a Thompson Sampling (TS)-based algorithm for our multi-agent RL (MARL) problem. This algorithm is based on the algorithm proposed in [16] to minimize the regret in a single-agent LQ control problem.

Similar to [16], we make the following assumptions on the prior distribution μ_1.

Assumption 2. *Let* $\bar{\mu}_1$ *be a probability distribution on* $\mathbb{R}^{d_X^1 \times (d_X^1 + d_U^1)}$ *which is the product of independent distributions* $\bar{\mu}_1(i)$, $i = 1, \dots, d_X^1$. *We assume that* $\bar{\mu}_1(i)$, $i = 1, \dots, d_X^1$, *is Gaussian with mean* $\hat{\Theta}_1(i) \in \mathbb{R}^{d_X^1 + d_U^1}$ *and covariance matrix* $\Sigma_1 \in \mathbb{R}^{(d_X^1 + d_U^1) \times (d_X^1 + d_U^1)}$ *where* Σ_1 *is positive definite*[1]. *Then, the prior distribution* μ_1 *is the projection of* $\bar{\mu}_1$ *on a compact support* $\Omega_1 \subset \mathbb{R}^{d_X^1 \times (d_X^1 + d_U^1)}$.

Assumption 3. *For any* $\Theta \in \Omega_1$, *the Riccati equation* (10) *with* $[A^{11}, B^{11}] = \Theta$ *has a unique positive definite solution. Further, the projection set* Ω *is such that for any* $\Theta = [A^{11}, B^{11}]$ *and* $\tilde{\Theta}$ *in* Ω, *the closed-loop matrix* $A + BK(\tilde{\Theta})$ *has spectral norm less than* δ, *that is,* $\|A + BK(\tilde{\Theta})\| \leq \delta$ *where* $\delta < 1$ *is an initial parameter of the algorithm.*

[1]Note that all distributions $\bar{\mu}_1(i)$, $i = 1, \dots, d_X^1$, have the same covariance matrix Σ_1.

Note that Assumption 3 ensures that the closed-loop system is stable under the learning algorithm.

Similar to [16], we present the following result which provides an update rule for the posterior belief μ_t.

Lemma 2. *Let* μ_t *is the posterior belief on the unknown parameter* Θ *at time* t. *Then,* μ_t *is the projection of a distribution* $\bar{\mu}_t$ *on a compact support* $\Omega_1 \subset \mathbb{R}^{d_X^1 \times (d_X^1 + d_U^1)}$. $\bar{\mu}_t$ *is the product of independent Gaussian distributions* $\bar{\mu}_1(i)$, $i = 1, \dots, d_X^1$ *with mean* $\hat{\Theta}_t(i)$ *and covariance* Σ_t *that can be sequentially updated using observations as follows.*

$$\hat{\Theta}_{t+1}(i) = \hat{\Theta}_t(i) + \frac{Z_t^\intercal \Sigma_t (X_{t+1}^1(i) - \hat{\Theta}_t(i) Z_t)}{1 + Z_t^\intercal \Sigma_t Z_t} \quad (17)$$

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t Z_t Z_t^\intercal \Sigma_t}{1 + Z_t^\intercal \Sigma_t Z_t} \quad (18)$$

where $Z_t = \mathbf{vec}(X_t^1, U_t^1) \in \mathbb{R}^{d_X^1 + d_U^1}$.

Lemma 2 can be proved using arguments for the least square estimator (for example, see [24]).

Now, we introduce the TS-MARL algorithm. This algorithm is a multi-agent algorithm which is performed independently by all three agents. TS-MARL algorithm operates in episodes. Let t_k be start time of the k-th episode and $T_k = t_{k+1} - t_k$ be the length of this episode with the convention $T_0 = 1$. From the description of the algorithm, $t_1 = 1$ and $t_{k+1}, k \geq 1$, is given by

$$t_{k+1} = \min\{t > t_k : \quad t > t_k + T_{k-1} \text{ or } \det(\Sigma_t) < 0.5 \det(\Sigma_{t_k})\}. \quad (19)$$

At the beginning of episode k, each agent generates a random sample Θ_k from its posterior μ_{t_k} and computes the gain matrix $K(\Theta_k)$ from (13). Then, during episode k, agent n uses the gain matrix $K(\Theta_k)$ to compute its action U_t^n. Note that agents 2 and 3 need $\tilde{K}^2$ and $\tilde{K}^3$ respectively to calculate their actions U_t^2 and U_t^3. However, we know from (14) that $\tilde{K}^2$ and $\tilde{K}^3$ are independent of the unknown parameter Θ and hence, they can be calculated prior to the beginning of the algorithm. After the execution of the actions $U_t^{1:3}$ by the agents, all the agents observe the new state X_{t+1}^1 of the system 1 and the agents 2 and 3 further observe the new states X_{t+1}^2 and X_{t+1}^3 of their co-located systems, respectively. Then, each agent n independently uses the following equation to compute $\check{X}_{t+1}^2$ and $\check{X}_{t+1}^3$,

$$\check{X}_{t+1}^m = A^{m1} X_t^1 + A^{mm} \check{X}_t^m + \Big(B^{m1} K^1(\Theta_k) + B^{mm} K^m(\Theta_k) \Big) \mathbf{vec}(X_t^1, \check{X}_t^2, \check{X}_t^3) \quad m \in \{2,3\}. \quad (20)$$

Finally, each agent n independently uses its own gain matrix $K(\Theta_k)$ to update μ_{t+1} according to (17)-(18).

One important feature of TS−MARL algorithm is that its episode lengths are not fixed. The length T_k of each episode is dynamically determined according to two stopping criteria: (i) $t > t_k + T_{k-1}$, and (ii) $\det(\Sigma_t) < 0.5 \det(\Sigma_{t_k})$. The first stopping criterion provides that the episode length grows at a linear rate without triggering the second criterion. The second stopping criterion ensures that the determinant of sample covariance matrix during an episode should not be less than half of the determinant of sample covariance matrix at the beginning of this episode.

Algorithm TS−MARL for agent 1

Input: $\Omega_1; \hat{\theta}_1(1), \ldots, \hat{\theta}_1(d_X^1); \Sigma_1$
Initialization: $t \leftarrow 1$; $t_0 \leftarrow 0$;
$\hat{\Theta}_1(i) = \hat{\theta}_1(i), i = 1, \ldots, d_X^1$
for episodes $k = 1, 2, ...$
$t_k \leftarrow t$
$T_{k-1} \leftarrow t_k - t_{k-1}$
Generate $\Theta_k \sim \mu_{t_k}$
Compute $K(\Theta_k)$ from (13)
while $t \leq t_k + T_{k-1}$ and $\det(\Sigma_t) \geq 0.5 \det(\Sigma_{t_k})$
Apply $U_t^1 = K^1(\Theta_k)\, \mathbf{vec}(X_t^1, \check{X}_t^2, \check{X}_t^3)$
Observe new state X_{t+1}^1
Compute $\check{X}_{t+1}^2$ and $\check{X}_{t+1}^3$ using (20)
Compute $Z_t = \begin{bmatrix} \mathbf{I} & (K^1(\Theta_k))^\intercal \end{bmatrix}^\intercal X_t^1$
Use Z_t to update μ_{t+1} according to (17)-(18)
$t \leftarrow t + 1$

Algorithm TS−MARL for agent n, $n = 2, 3$

Input: $\Omega_1; \hat{\theta}_1(1), \ldots, \hat{\theta}_1(d_X^1); \Sigma_1$
Initialization: $t \leftarrow 1$; $t_0 \leftarrow 0$;
$\hat{\Theta}_1(i) = \hat{\theta}_1(i), i = 1, \ldots, d_X^1$
for episodes $k = 1, 2, ...$
$t_k \leftarrow t$
$T_{k-1} \leftarrow t_k - t_{k-1}$
Generate $\Theta_k \sim \mu_{t_k}$
Compute $K(\Theta_k)$ from (13)
while $t \leq t_k + T_{k-1}$ and $\det(\Sigma_t) \geq 0.5 \det(\Sigma_{t_k})$
Apply $U_t^n = K^n(\Theta_k)\, \mathbf{vec}(X_t^1, \check{X}_t^2, \check{X}_t^3)$
$+\tilde{K}^n(X_t^n - \check{X}_t^n)$
Observe new states X_{t+1}^1 and X_{t+1}^n
Compute $\check{X}_{t+1}^2$ and $\check{X}_{t+1}^3$ using (20)
Compute $Z_t = \begin{bmatrix} \mathbf{I} & (K^1(\Theta_k))^\intercal \end{bmatrix}^\intercal X_t^1$
Use Z_t to update μ_{t+1} according to (17)-(18)
$t \leftarrow t + 1$

Remark 1. *Note that $\check{X}_{t+1}^2$ and $\check{X}_{t+1}^3$ in the TS−MARL algorithm (given by* (20)*) are proxies for $\hat{X}_{t+1}^2$ and $\hat{X}_{t+1}^3$ of* (15) *where instead of the unknown parameter Θ, we have Θ_t.*

Remark 2. *Although all agents start with the same initial parameters (and hence the same prior μ_1), due to the independent execution of the TS−MARL algorithm, agents might generate different samples Θ_1 from μ_1. As a result, the computed gain matrices $K(\Theta_1)$ by the agents can be different. Since each agent uses its own $K(\Theta_1)$ to update the posterior belief, the new posterior μ_2 can be different among the agents. This difference in μ_2 among the agents will also lead to different μ_t, $t > 2$.*

In order to avoid issues pointed out in Remark 2, we make an assumption about how samples are generated by the agents.

Assumption 4. *All agents use the same sampling seed for generating samples from their posteriors μ_t.*

Now, we present our main result which is based on Assumption 4.

Theorem 1. *Under Assumptions 1-4, TS−MARL algorithm achieves a $\tilde{O}(\sqrt{T})$ regret for the MARL problem.*

Remark 3. *While the result of this paper has been presented for the case of 1 unknown system and 2 known systems, the result can be easily extended to the case of 1 unknown system and arbitrary number N of known systems.*

IV. Proof of Theorem 1

We first prove some preliminary results in the following lemmas which will be used in the proof of Theorem 1.

Lemma 3. *Under Assumption 4, at each time t, $\check{X}_t^2$, $\check{X}_t^3$, and μ_t calculated independently by the agents are all equal.*

Lemma 4. *Let S_t^n be a random process that evolves as*

$$S_{t+1}^n = C^n S_t^n + W_t^n, \quad S_1^n = \mathbf{0}, \tag{21}$$

where $C^n = A^{nn} + B^{nn}\tilde{K}^n$. Define $\Delta_t^n = \mathbf{cov}(S_t^n)$, then the sequence of matrices Δ_t^n, $t \geq 1$, is increasing[2] and it converges to a PSD matrix Δ^n as $t \to \infty$.

Our approach for the proof of Theorem 1 is to construct an auxiliary Single-Agent Reinforcement Learning (SARL) problem based on the MARL problem of Section II. This auxiliary SARL problem is used for the regret analysis of the TS−MARL algorithm. We proceed in three steps:

- Step 1: Constructing an auxiliary SARL problem
- Step 2: Showing the connection between the auxiliary SARL problem and the MARL problem
- Step 3: Using the SARL problem to bound the regret of the MARL problem

Step 1: Constructing an auxiliary SARL problem

Consider a single-agent system with dynamics

$$X_{t+1}^\diamond = AX_t^\diamond + BU_t^\diamond + \mathbf{vec}(W_t^1, \mathbf{0}, \mathbf{0}), \tag{22}$$

where $X_t^\diamond \in \mathbb{R}^{d_X^1 + d_X^2 + d_X^3}$ is the state of the system, $U_t^\diamond \in \mathbb{R}^{d_U^1 + d_U^2 + d_U^3}$ is the action of the auxiliary agent, W_t^1 is the noise vector of system 1 defined in (1), and matrices A and B are as defined in (4). The initial state $X_1^\diamond$ is assumed to be zero. The action $U_t^\diamond = \pi_t^\diamond(H_t^\diamond)$ at time t is a function of the history of observations $H_t^\diamond = \{X_{1:t}^\diamond, U_{1:t-1}^\diamond\}$. The auxiliary agent's strategy is denoted by $\pi^\diamond = (\pi_1^\diamond, \pi_2^\diamond, \ldots)$. The

[2]Note that increasing is in the sense of partial order $\succeq$, that is, $\Delta_1^n \preceq \Delta_2^n \preceq \Delta_3^n \preceq \ldots$

instantaneous cost $c(X_t^\diamond, U_t^\diamond)$ of the system is a quadratic function given by

$$c(X_t^\diamond, U_t^\diamond) = (X_t^\diamond)^\intercal Q X_t^\diamond + (U_t^\diamond)^\intercal R U_t^\diamond, \tag{23}$$

where matrices Q and R are as defined in (6).

When $\Theta = [A^{11}, B^{11}]$ (note that A^{11} and B^{11} are sub-block matrices of A and B as described in (4)) is known to the auxiliary agent, minimizing the infinite horizon average cost is a single-agent stochastic Linear-Quadratic (LQ) control problem. Let $J^\diamond(\Theta)$ be the optimal infinite horizon average cost under Θ, that is,

$$J^\diamond(\Theta) = \inf_{\pi^\diamond} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{\pi^\diamond}[c(X_t^\diamond, U_t^\diamond)|\Theta]. \tag{24}$$

The above single-agent stochastic LQ control problem has been widely studied in the literature. It is well-known that under Assumption 1, the optimal infinite horizon cost $J^\diamond(\Theta)$ is given by $J^\diamond(\Theta) = \mathbf{tr}([P(\Theta)]_{1,1})$ where $P(\Theta)$ is as defined in (10). Furthermore, the optimal strategy $\pi^{\diamond *}$ is given by $U_t^\diamond = K(\Theta) X_t^\diamond$ where $K(\Theta)$ is as defined in (13).

When Θ is unknown, this single-agent stochastic LQ control problem becomes a Single-Agent Reinforcement Learning (SARL) problem. We define the expected regret of a policy $\pi^\diamond$ up to time T compared with the optimal infinite horizon cost $J^\diamond(\Theta)$ to be

$$R^\diamond(T, \pi^\diamond) = \mathbb{E}^{\pi^\diamond}\left[\sum_{t=1}^{T} c(X_t^\diamond, U_t^\diamond) - T J^\diamond(\Theta)\right]. \tag{25}$$

The above expectation is with respect to the random noise of system 1 ($W_{1:T}^1$), the prior distribution μ_1, and randomization in the auxiliary agent's strategy.

For this SARL problem, the TS-based algorithm of [16], referred to as TS-SARL algorithm hereafter, achieves a $\tilde{O}(\sqrt{T})$ expected regret for the SARL problem, that is,

$$R^\diamond(T, \texttt{TS-SARL}) \leq \tilde{O}(\sqrt{T}), \tag{26}$$

where $\tilde{O}(\cdot)$ hides constants and logarithmic factors.

Step 2: Showing the connection between the auxiliary SARL problem and the MARL problem

We present the following two lemmas that show the connection between the auxiliary SARL problem and the MARL problem.

Lemma 5. *Let $J^\diamond(\Theta)$ be the optimal infinite horizon cost of the auxiliary SARL problem when Θ is known, $J(\Theta)$ be the optimal infinite horizon cost of the MARL problem when Θ is known, and Δ^n, $n \in \{2, 3\}$, be as defined in Lemma 4. Then,*

$$J(\Theta) = J^\diamond(\Theta) + \mathbf{tr}(D^2 \Delta^2) + \mathbf{tr}(D^3 \Delta^3), \tag{27}$$

where we have defined $D^n := Q^{nn} + (\tilde{K}^n)^\intercal R^{nn} \tilde{K}^n$ for $n \in \{2, 3\}$.

Lemma 6. *At each time t, the following equality holds between the expected cost under the policies of the TS-SARL and the TS-MARL algorithms,*

$$\mathbb{E}^{\textit{TS-MARL}}[c(X_t, U_t)] = \mathbb{E}^{\textit{TS-SARL}}[c(X_t^\diamond, U_t^\diamond)] + \mathbf{tr}(D^2 \Delta_t^2) + \mathbf{tr}(D^3 \Delta_t^3). \tag{28}$$

Step 3: Using the SARL problem to bound the regret of the MARL problem

In this step, we use the connections between the auxiliary SARL problem and our MARL problem, which was established in Step 2, to prove Theorem 1. Note that from the definition of the expected regret in the the MARL problem given by (16), we have,

$$\begin{aligned}
&R(T, \texttt{TS-MARL}) = \mathbb{E}^{\texttt{TS-MARL}}\left[\sum_{t=1}^{T} c(X_t, U_t) - T J(\Theta)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}^{\texttt{TS-SARL}}[c(X_t^\diamond, U_t^\diamond)] + \sum_{t=1}^{T} [\mathbf{tr}(D^2 \Delta_t^2) + \mathbf{tr}(D^3 \Delta_t^3)] \\
&- T\,\mathbb{E}[J^\diamond(\Theta)] - T\,\mathbf{tr}(D^2 \Delta^2) - T\,\mathbf{tr}(D^3 \Delta^3) \\
&= \sum_{t=1}^{T} \mathbb{E}^{\texttt{TS-SARL}}[c(X_t^\diamond, U_t^\diamond) - T J^\diamond(\Theta)] \\
&+ \sum_{t=1}^{T} [\mathbf{tr}(D^2(\Delta_t^2 - \Delta^2)) + \mathbf{tr}(D^3(\Delta_t^3 - \Delta^3))] \\
&= R^\diamond(T, \texttt{TS-SARL}) + \sum_{t=1}^{T} [\mathbf{tr}(D^2(\Delta_t^2 - \Delta^2)) \\
&+ \mathbf{tr}(D^3(\Delta_t^3 - \Delta^3))] \leq R^\diamond(T, \texttt{TS-SARL}) \leq \tilde{O}(\sqrt{T}),
\end{aligned} \tag{29}$$

where the second equality is correct because of Lemma 5, Lemma 6, and the fact that $J(\Theta)$ is independent of the policy of the TS-MARL algorithm, that is, $\mathbb{E}^{\texttt{TS-MARL}}[J(\Theta)] = \mathbb{E}[J(\Theta)]$. Furthermore, the third equality is correct due to the fact that $J^\diamond(\Theta)$ is independent of the policy of the TS-SARL algorithm, that is, $\mathbb{E}[J^\diamond(\Theta)] = \mathbb{E}^{\texttt{TS-SARL}}[J^\diamond(\Theta)]$, the fourth equality is correct by definition of the expected regret in the SARL problem, and the penultimate inequality is correct because from Lemma 4, the sequence of matrices Δ_t^n is increasing, that is, $\Delta^n - \Delta_t^n \succeq \mathbf{0}$ and D^n is positive semi-definite, and consequently, $\mathbf{tr}(D^n(\Delta_t^n - \Delta^n)) \leq 0$, $n \in \{2, 3\}$. Finally, the last inequality is correct because of (26). This proves the statement of Theorem 1.

V. Experiments

In this section, we illustrate the performance of the TS-MARL algorithm through numerical experiments.

We consider an instance of the MARL problem where system 1 (which is unknown to the agents in our problem), has the following parameters (which are the same as the model studied in [25]) with $d_X^1 = d_U^1 = 3$,

$$A^{11} = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B^{11} = \mathbf{I}_3, \tag{30}$$

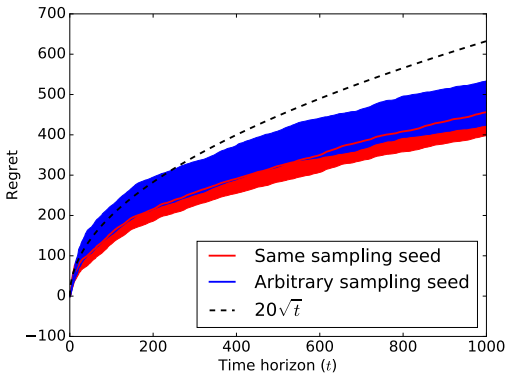

Fig. 2. Expected regret under the TS-MARL algorithm

and systems 2 and 3 are one-dimensional, that is, $d_X^2 = d_X^3 = d_U^2 = d_U^3 = 1$, with the following parameters,

$$A^{21} = B^{21} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \quad A^{22} = 1.01, \quad B^{22} = 1 \quad (31)$$

$$A^{31} = B^{31} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}, \quad A^{33} = 1.01, \quad B^{33} = 1. \quad (32)$$

Further, we consider the following matrices (with the same structure as the model in [25]) for the cost function,

$$Q = 10^{-3}\mathbf{I}_5, \quad R = \mathbf{I}_5. \quad (33)$$

The prior distribution used in TS-MARL algorithm is set according to Assumptions 2 and 3 with $\hat{\Theta}_1(i)$, $i = 1, 2, 3$, to be an all-one vector, $\Sigma_1 = \mathbf{I}_6$, and $\Omega_1 = \{\tilde{\Theta} : \|A + BK(\tilde{\Theta})\| \leq \delta\}$ where we use $\delta = 0.99$ for the simulations.

While the theoretical result of Theorem 1 required the same sampling seed among the agents (i.e., Assumption 4), we consider both cases of same sampling seed and arbitrary sampling seed for the experiments. We ran 50 simulations and show the mean of regret with the 95% confidence interval for each scenario.

As it can be seen from Figure 2, for both of theses cases, our proposed TS-MARL algorithm achieves a $\tilde{O}(\sqrt{T})$ regret for our MARL problem, which matches the theoretical results of Theorem 1.

VI. Conclusion

In this paper, we studied a multi-agent Linear-Quadratic (LQ) reinforcement learning problem consisting of three systems, an unknown system and two known systems, and three agents. The goal was to minimize the infinite-horizon average cost. We proposed a Thompson Sampling (TS)-based multi-agent learning algorithm where each agent learns the unknown system's dynamics independently. We showed that the expected regret of our algorithm is upper bounded by $\tilde{O}(\sqrt{T})$ under certain assumptions where $\tilde{O}(\cdot)$ hides constants and logarithmic factors.

References

[1] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview," *Information Sciences*, vol. 380, pp. 117–131, 2017.

[2] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Transactions on Automatic Control*, vol. 50, no. 12, pp. 1984–1996, 2005.

[3] A. Lamperski and J. C. Doyle, "The $\mathcal{H}_2$ control problem for quadratically invariant systems with delays," *IEEE Transactions on Automatic Control*, vol. 60, pp. 1945–1950, July 2015.

[4] A. Mishra, C. Langbort, and G. E. Dullerud, "Team optimal control of stochastically switched systems with local parameter knowledge," *IEEE Transactions on Automatic Control*, vol. 60, pp. 2086–2101, Aug 2015.

[5] A. Mahajan and A. Nayyar, "Sufficient statistics for linear control strategies in decentralized systems with partial history sharing," *IEEE Transactions on Automatic Control*, vol. 60, no. 8, pp. 2046–2056, 2015.

[6] S. M. Asghari and A. Nayyar, "Dynamic teams and decentralized control problems with substitutable actions," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5302–5309, 2017.

[7] S. Kar, J. M. F. Moura, and H. V. Poor, "Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus innovations," *IEEE Transactions on Signal Processing*, vol. 61, pp. 1848–1862, April 2013.

[8] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*, pp. 157–163, Elsevier, 1994.

[9] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 2137–2145, 2016.

[10] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," *arXiv preprint arXiv:1802.08757*, 2018.

[11] M. Gagrani and A. Nayyar, "Thompson sampling for some decentralized control problems," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 1053–1058, IEEE, 2018.

[12] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*, vol. 75. SIAM, 2015.

[13] M. C. Campi and P. Kumar, "Adaptive linear quadratic gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[14] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

[15] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, pp. 4192–4201, 2018.

[16] Y. Ouyang, M. Gagrani, and R. Jain, "Control of unknown linear systems with thompson sampling," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1198–1205, IEEE, 2017.

[17] M. Abeille and A. Lazaric, "Improved regret bounds for thompson sampling in linear quadratic control problems," in *International Conference on Machine Learning*, pp. 1–9, 2018.

[18] Y. Abbasi-Yadkori and C. Szepesvári, "Bayesian optimal control of smoothly parameterized systems," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 2–11, AUAI Press, 2015.

[19] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On optimality of adaptive linear-quadratic regulators," *arXiv preprint arXiv:1806.10749*, 2018.

[20] L. Lessard, "Decentralized lqg control of systems with a broadcast architecture," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 6241–6246, IEEE, 2012.

[21] S. M. Asghari, Y. Ouyang, and A. Nayyar, "Optimal local and remote controllers with unreliable uplink channels," *IEEE Transactions on Automatic Control*, pp. 1–1, 2018.

[22] Y. Ouyang, S. M. Asghari, and A. Nayyar, "Optimal infinite horizon decentralized networked controllers with unreliable communication," *arXiv preprint arXiv:1806.06497*, 2018.

[23] S. M. Asghari, *Team decision theory and decentralized stochastic control*. PhD thesis, University of Southern California, 2019.

[24] J. Sternby, "On consistency for the method of least squares using martingale theory," *IEEE Transactions on Automatic Control*, vol. 22, no. 3, pp. 346–352, 1977.

[25] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.