2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Thompson sampling for linear quadratic mean-field teams

Mukul Gagrani, Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar and Yi Ouyang

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683030

***Abstract*— We consider optimal control of an unknown multi-agent linear quadratic (LQ) system where the dynamics and the cost are coupled across the agents through the mean-field (i.e., empirical mean) of the states and controls. Directly using single-agent LQ learning algorithms in such models results in regret which increases polynomially with the number of agents. We propose a new Thompson sampling based learning algorithm which exploits the structure of the system model and show that the expected Bayesian regret of our proposed algorithm for a system with agents of $|M|$ different types at time horizon T is $\tilde{\mathcal{O}}(|M|^{1.5}\sqrt{T})$ irrespective of the total number of agents, where the $\tilde{\mathcal{O}}$ notation hides logarithmic factors in T. We present detailed numerical experiments to illustrate the salient features of the proposed algorithm.**

I. Introduction

Linear dynamical systems with a quadratic cost (henceforth referred to as LQ systems) are one of the most commonly used modeling framework in Systems and Control. Part of the appeal of LQ models is that the optimal control action in such models is a linear or affine function of the state; therefore, the optimal policy is easy to identify and easy to implement.

Broadly speaking, the regret of three classes of learning algorithms have been analyzed in the literature: Optimism in the face of uncertainty (OFU) based algorithms, certainty equivalence (CE) based algorithms, and Thompson sampling (TS) based algorithms.

OFU-based algorithms are inspired by the OFU principle for multi-armed bandits [1]. Starting with the work of [2], [3], most of the papers following this approach [4]–[6] provide a high probability bound on regret. As an illustrative example, it is shown in [6] that, with high probability, the regret of a OFU-based learning algorithm is $\tilde{\mathcal{O}}(d_x^{0.5}(d_x + d_u)\sqrt{T})$, where d_x is the dimension of the state, d_u is the dimension of the controls, T is the time horizon, and the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic terms in T.

Certainty equivalence (CE) is a classical adaptive control algorithm in Systems and Control [7], [8]. Most papers following this approach [9]–[12] also provide a high probability bound on regret. As an illustrative example, it is shown in [12] that, with high probability, the regret of a CE-based algorithm is $\tilde{\mathcal{O}}(d_x^{0.5}d_u\sqrt{T} + d_x^2)$.

Thompson sampling (TS) based algorithms are inspired by TS algorithm for multi-armed bandits [13]. Most papers following this approach [14]–[16] establish a bound on the expected Bayesian regret. As an illustrative example, [15] shows that the regret of a TS-based algorithm is $\tilde{\mathcal{O}}(d_x^{0.5}(d_x + d_u)\sqrt{T})$.

Two aspects of these regret bounds are important: the dependence on the time horizon T and the dependence on the dimensions (d_x, d_u) of the state and the controls. For all classes of algorithms mentioned above, the dependence on the time horizon is $\tilde{\mathcal{O}}(\sqrt{T})$. Moreover, there are multiple papers which show that, under different assumptions, the regret is lower bounded by $\Omega(\sqrt{T})$ [12], [17]. So, the time dependence in the available regret bounds is nearly order optimal. Similarly, even though the dependence of the regret bound on the dimensions of the state and the control varies slightly for each class of algorithms, [12] recently showed that the regret is lower bounded by $\tilde{\Omega}(d_x^{0.5}d_u\sqrt{T})$. So, there is only a small scope of improvement in the dimension dependence in the regret bounds.

The dependence of the regret bounds on the dimensions of the state and controls is critical for applications such as formation control of robotic swarms and demand response in power grids which have large numbers of agents (which can be of the order of 10^3 to 10^5). In such systems, the effective dimension of the state and the controls is nd_x and nd_u, where n is the number of agents and d_x and d_u are the dimensions of the state and controls of each agent. Therefore, if we take the regret bound of, say, the OFU algorithm proposed in [6], the regret is $\tilde{\mathcal{O}}(n^{1.5}d_x^{0.5}(d_x + d_u)\sqrt{T})$. Similar scaling with n holds for CE- and TS-based algorithms. The polynomial dependence on the number of agents is prohibitive and, because of it, the standard regret bounds are of limited value for large-scale systems.

There are many papers in the literature on the design of large-scale systems which exploit some structural property of the system to develop low-complexity design algorithms [18]–[23]. However, there has been very little investigation on the role of such structural properties in developing and analyzing learning algorithms.

Our main contribution is to show that by carefully exploiting the structure of the model, it is possible to design learning algorithms for large-scale LQ systems where the regret does not grow polynomially in the number of agents. In particular, we investigate mean-field coupled control systems, which have emerged as a popular modeling framework in multiple research communities including Control Systems, Economics,

M. Gagrani was with the Department of EE, USC. He is currently with Qualcomm AI Research, San Diego. Email: mgagrani@qti.qualcomm.com
S. Sudhakara and A. Nayyar are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA. Email: sagarsud@usc.edu; ashutosn@usc.edu.
A. Mahajan, Mcgill University, Canada. Email:aditya.mahajan@mcgill.ca
Yi Ouyang, Preferred Networks, USA. Email:ouyangyii@gmail.com
The research at USC was supported in part by NSF under grants ECCS 2025732 and ECCS 1750041 and by the Okawa Foundation Research Grant. The work of A. Mahajan was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-30.

Finance, and Statistical Physics [24]–[28]. These models are used in various applications ranging from demand response in smart grids, large scale communication networks, UAVs, finanancial markets, and many others. We refer the reader to [29] for a survey. There has been considerable interest in reinforcement learning for such models [30]–[35], but all of these papers focus on identifying asymptotically optimal policies and do not characterize regret.

Our main contribution is to design a TS-based algorithm for mean-field teams (which is a specific mean-field model proposed in [22], [23]) and show that the regret scales as $\tilde{\mathcal{O}}(|M|^{1.5} d_x^{0.5}(d_x + d_u)\sqrt{T})$, where $|M|$ is the number of types.

We would like to highlight that although we focus on a TS-based algorithm in the paper, it will be clear from the derivation that it is possible to develop OFU- and CE-based algorithms with similar regret bounds. Thus, the main takeaway message of our paper is that there is significant value in developing learning algorithms which exploit the structure of the model.

II. Background on mean-field teams

A. Mean-field teams model

We start by describing a slight generalization of the basic model of mean-field teams proposed in [22], [23]. Mean-field teams are also called cooperative mean-field games or mean-field control in the literature [36].

Consider a system with a large population of agents. The agents are heterogeneous and have multiple types. Let $M = \{1, \dots, |M|\}$ denote the set of types of agents, N^m, $m \in M$, denote the set of all agents of type m, and $N = \bigcup_{m \in M} N^m$ denote the set of all agents.

a) States, actions, and their mean-fields: Agents of the same type have the same state and action spaces. In particular, the state and control action of agents of type m take values in $\mathbb{R}^{d_x^m}$ and $\mathbb{R}^{d_u^m}$, respectively. For any generic agent $i \in N^m$ of type m, we use $x_t^i \in \mathbb{R}^{d_x^m}$ and $u_t^i \in \mathbb{R}^{d_u^m}$ to denote its state and control action at time t. We use $\boldsymbol{x}_t = \operatorname{vec}((x_t^i)_{i \in N})$ and $\boldsymbol{u}_t = \operatorname{vec}((u_t^i)_{i \in N})$ to denote the global state and control actions of the system at time t.

The empirical mean-field $(\bar{x}_t^m, \bar{u}_t^m)$ of agents of type m, $m \in M$, is defined as the empirical mean of the states and actions of all agents of that type, i.e.,

$$\bar{x}_t^m = \frac{1}{|N^m|} \sum_{i \in N^m} x_t^i \quad \text{and} \quad \bar{u}_t^m = \frac{1}{|N^m|} \sum_{i \in N^m} u_t^i.$$

The empirical mean-field $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$ of the entire population is given by

$$\bar{\boldsymbol{x}}_t = \operatorname{vec}(\bar{x}_t^1, \dots, \bar{x}_t^{|M|}) \quad \text{and} \quad \bar{\boldsymbol{u}}_t = \operatorname{vec}(\bar{u}_t^1, \dots, \bar{u}_t^{|M|}).$$

As an example, consider the temperature control of a multi-storied office building. In this case, N represents the set of rooms, M represents the set of floors, N^m represents all rooms in floor m, x_t^i represents the temperature in room i, $\bar{x}_t^m$ represents the average temperature in floor m, and $\bar{\boldsymbol{x}}_t$ represents the collection of average temperature in each floor. Similarly, u_t^i represents the heat exchanged by the air-conditioner in room i, $\bar{u}_t^m$ represents the average heat exchanged by the air-conditioners in floor m, and $\bar{\boldsymbol{u}}_t$ represents the collection of average heat exchanged in each floor of the building.

b) System dynamics and per-step cost: The system starts at a random initial state $x_1 = (x_1^i)_{i \in N}$, whose components are independent across agents. For agent i of type m, the initial state $x_1^i \sim \mathcal{N}(0, \mathsf{X}_1^i)$, and at time $t \geq 1$, the state evolves according to

$$x_{t+1}^i = \mathsf{A}^m x_t^i + \mathsf{B}^m u_t^i + \mathsf{D}^m \bar{\boldsymbol{x}}_t + \mathsf{E}^m \bar{\boldsymbol{u}}_t + w_t^i + v_t^m + \mathsf{F}^m v_t^0, \tag{1}$$

where A^m, B^m, D^m, E^m, F^m are matrices of appropriate dimensions, $\{w_t^i\}_{t \geq 1}$, $\{v_t^m\}_{t \geq 1}$, and $\{v_t^0\}_{t \geq 1}$ are i.i.d. zero-mean Gaussian processes which are independent of each other and the initial state. In particular, $w_t^i \in \mathbb{R}^{d_x^m}$, $v_t^m \in \mathbb{R}^{d_x^m}$, and $v_t^0 \in \mathbb{R}^{d_v^0}$, and $w_t^i \sim \mathcal{N}(0, \mathsf{W}^i)$, $v_t^m \sim \mathcal{N}(0, \mathsf{V}^m)$, and $v_t^0 \sim \mathcal{N}(0, \mathsf{V}^0)$.

Eq. (1) implies that all agents of type m have similar dynamical couplings. The next state of agent i of type m depends on its current local state and control action, the current mean-field of the states and control actions of the system, and is influenced by three independent noise processes: a local noise process $\{w_t^i\}_{t \geq 1}$, a noise process $\{v_t^m\}_{t \geq 1}$ which is common to all agents of type m, and a global noise process $\{v_t^0\}_{t \geq 1}$ which is common to all agents.

At each time-step, the system incurs a quadratic cost $c(\boldsymbol{x}_t, \boldsymbol{u}_t)$ given by

$$c(\boldsymbol{x}_t, \boldsymbol{u}_t) = \bar{\boldsymbol{x}}_t^{\mathsf{T}} \bar{\mathsf{Q}} \bar{\boldsymbol{x}}_t + \bar{\boldsymbol{u}}_t^{\mathsf{T}} \bar{\mathsf{R}} \bar{\boldsymbol{u}}_t + \sum_{m \in M} \frac{1}{|N^m|} \sum_{i \in N^m} \big[(x_t^i)^{\mathsf{T}} \mathsf{Q}^m x_t^i + (u_t^i)^{\mathsf{T}} \mathsf{R}^m u_t^i \big]. \tag{2}$$

Thus, there is a weak coupling in the cost of the agents through the mean-field.

c) Admissible policies and performance criterion: There is a system operator who has access to the states of all agents and control actions and chooses the control action according to a deterministic or randomized policy

$$\boldsymbol{u}_t = \pi_t(\boldsymbol{x}_{1:t}, \boldsymbol{u}_{1:t-1}). \tag{3}$$

Let $\boldsymbol{\theta} = (\theta^m)_{m \in M}$, where $(\theta^m)^{\mathsf{T}} = [\mathsf{A}^m, \mathsf{B}^m, \mathsf{D}^m, \mathsf{E}^m, \mathsf{F}^m]$, denotes the parameters of the system dynamics. The performance of any policy $\pi = (\pi_1, \pi_2, \dots)$ is given by

$$J(\pi; \boldsymbol{\theta}) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=1}^{T} c(\boldsymbol{x}_t, \boldsymbol{u}_t) \Big]. \tag{4}$$

Let $J(\boldsymbol{\theta})$ to denote the minimum of $J(\pi; \boldsymbol{\theta})$ over all policies.

We are interested in the setup where the system dynamics $\boldsymbol{\theta}$ are unknown and there is a prior p on $\boldsymbol{\theta}$. The Bayesian *regret* of a policy π operating for a horizon T is defined as

$$R(T; \pi) := \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - T J(\boldsymbol{\theta}) \right] \tag{5}$$

where the expectation is with respect to the prior on θ, the noise processes, the initial conditions, and the potential randomizations done by the policy π.

B. Planning solution for mean-field teams

In this section, we summarize the planning solution of mean-field teams presented in [22], [23] for a known system model.

Define the following matrices:

$$\bar{\mathsf{A}} = \operatorname{diag}(\mathsf{A}^1, \dots, \mathsf{A}^{|M|}) + \operatorname{rows}(\mathsf{D}^1, \dots, \mathsf{D}^{|M|}),$$
$$\bar{\mathsf{B}} = \operatorname{diag}(\mathsf{B}^1, \dots, \mathsf{B}^{|M|}) + \operatorname{rows}(\mathsf{E}^1, \dots, \mathsf{E}^{|M|}),$$

and let $\bar{\bar{\mathsf{Q}}} = \operatorname{diag}(\mathsf{Q}^1, \dots, \mathsf{Q}^{|M|}) + \bar{\mathsf{Q}}$ and $\bar{\bar{\mathsf{R}}} = \operatorname{diag}(\mathsf{R}^1, \dots, \mathsf{R}^{|M|}) + \bar{\mathsf{R}}$.

It is assumed that the system satisfies the following:

(A1) $\bar{\bar{\mathsf{Q}}} > 0$ and $\bar{\bar{\mathsf{R}}} > 0$. Moreover, for every $m \in M$, $\mathsf{Q}^m > 0$ and $\mathsf{R}^m > 0$.

(A2) The system $(\bar{\mathsf{A}}, \bar{\mathsf{B}})$ is stabilizable.[1] Moreover, for every $m \in M$, the system $(\mathsf{A}^m, \mathsf{B}^m)$ is stabilizable.

Now, consider the following $|M|+1$ discrete time algebraic Riccati equations (DARE):[2]

$$\breve{\mathsf{S}}^m = \operatorname{DARE}(\mathsf{A}^m, \mathsf{B}^m\mathsf{Q}^m, \mathsf{R}^m), \quad m \in M, \tag{6a}$$
$$\bar{\mathsf{S}} = \operatorname{DARE}(\bar{\mathsf{A}}, \bar{\mathsf{B}}, \bar{\bar{\mathsf{Q}}}, \bar{\bar{\mathsf{R}}}). \tag{6b}$$

Moreover, define

$$\breve{\mathsf{L}}^m = -\big((\mathsf{B}^m)^{\mathsf{T}}\breve{\mathsf{S}}^m\mathsf{B}^m + \mathsf{R}^m\big)^{-1}(\mathsf{B}^m)^{\mathsf{T}}\breve{\mathsf{S}}^m\mathsf{A}^m, \quad m \in M, \tag{7a}$$

$$\bar{\mathsf{L}} = -\big(\bar{\mathsf{B}}^{\mathsf{T}}\bar{\mathsf{S}}\bar{\mathsf{B}} + \bar{\bar{\mathsf{R}}}\big)^{-1}\bar{\mathsf{B}}^{\mathsf{T}}\bar{\mathsf{S}}\bar{\mathsf{A}}, \tag{7b}$$

and let $\operatorname{rows}(\bar{\mathsf{L}}^1, \dots, \bar{\mathsf{L}}^{|M|}) = \bar{\mathsf{L}}$.

Finally, define $\bar{w}_t^m = \frac{1}{|N^m|}\sum_{i\in N^m} w_t^i$, $\bar{\boldsymbol{w}}_t = \operatorname{vec}(\bar{w}_t^1, \dots, \bar{w}_t^{|M|})$ and $\bar{\boldsymbol{v}}_t = \operatorname{vec}(v_t^1, \dots, v_t^{|M|})$. Let $\breve{\mathsf{W}}^m = \frac{1}{|N^m|}\sum_{i\in N^m}\operatorname{var}(w_t^i - \bar{w}_t^m)$ and $\bar{\mathsf{W}} = \operatorname{var}(\bar{\boldsymbol{w}}_t) + \operatorname{diag}(\mathsf{V}^1, \dots, \mathsf{V}^{|M|}) + \operatorname{diag}(\mathsf{F}^1\mathsf{V}^0, \dots, \mathsf{F}^{|M|}\mathsf{V}^0)$. Note that since the noise processes are i.i.d., these covariances do not depend on time.

Now, split the state x_t^i of agent i of type m into two parts: the *mean-field* state $\bar{x}_t^m$ and the *relative* state $\breve{x}_t^i = x_t^i - \bar{x}_t^m$. Do a similar split of the controls: $u_t^i = \bar{u}_t^m + \breve{u}_t^i$. Since $\sum_{i\in N^m}\breve{x}_t^i = 0$ and $\sum_{i\in N^m}\breve{u}_t^i = 0$, the per-step cost (2) can be written as

$$c(\boldsymbol{x}_t, \boldsymbol{u}_t) = \bar{c}(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + \sum_{m\in M}\frac{1}{|N^m|}\sum_{i\in N^m}\breve{c}^m(\breve{x}_t^i, \breve{u}_t^i) \tag{8}$$

where $\bar{c}(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) = \bar{\boldsymbol{x}}_t^{\mathsf{T}}\bar{\bar{\mathsf{Q}}}\bar{\boldsymbol{x}}_t + \bar{\boldsymbol{u}}_t^{\mathsf{T}}\bar{\bar{\mathsf{R}}}\bar{\boldsymbol{u}}_t$ and $\breve{c}^m(\breve{x}_t^i, \breve{u}_t^i) = (\breve{x}_t^i)^{\mathsf{T}}\mathsf{Q}^m\breve{x}_t^i + (\breve{u}_t^i)^{\mathsf{T}}\mathsf{R}^m\breve{u}_t^i$. Moreover, the dynamics of mean-field and the relative components of the state are:

$$\bar{\boldsymbol{x}}_{t+1} = \bar{\mathsf{A}}\bar{\boldsymbol{x}}_t + \bar{\mathsf{B}}\bar{\boldsymbol{u}}_t + \bar{\boldsymbol{w}}_t + \bar{\boldsymbol{v}}_t + \bar{\mathsf{F}}v_t^0 \tag{9}$$

where $\bar{\mathsf{F}} = \operatorname{diag}(\mathsf{F}^1, \dots, \mathsf{F}^{|M|})$ and for any agent i of type m,

$$\breve{x}_t^i = \mathsf{A}_t^m\breve{x}_t^i + \mathsf{B}^m\breve{u}_t^i + \breve{w}_t^i, \tag{10}$$

where $\breve{w}_t^i = w_t^i - \bar{w}_t^m$.

The result below follows from [23, Theorem 6][3]:

Theorem 1 *Under assumptions* (A1) *and* (A2)*, the optimal policy for minimizing the cost* (4) *is given by*

$$u_t^i = \breve{\mathsf{L}}^m\breve{x}_t^i + \bar{\mathsf{L}}^m\bar{\boldsymbol{x}}_t. \tag{11}$$

Furthermore, the optimal performance is given by

$$J(\boldsymbol{\theta}) = \sum_{m\in M}\operatorname{Tr}(\breve{\mathsf{W}}^m\breve{\mathsf{S}}^m) + \operatorname{Tr}(\bar{\mathsf{W}}\bar{\mathsf{S}}). \tag{12}$$

a) Interpretation of the planning solution: Note that $\bar{\boldsymbol{u}}_t = \bar{\mathsf{L}}_t\bar{\boldsymbol{x}}_t$ is the optimal control for the mean-field system with dynamics (9) and per-step cost $\bar{c}(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$. Moreover, for agent i of type m, $\breve{u}_t^i = \breve{\mathsf{L}}_t^m\breve{x}_t^i$ is the optimal control for the relative system with dynamics (10) and per-step cost $\breve{c}^m(\breve{x}_t^i, \breve{u}_t^i)$. Theorem 1 shows that at every agent i of type m, we can consider the two decoupled systems—the mean-field system and the relative system— solve them separately, and then simply add their respective controls—$\bar{u}_t^m$ and $\breve{u}_t^i$—to obtain the optimal control action at agent i in the original mean-field team system. We will exploit this feature of the planning solution in order to develop a learning algorithm for mean-field teams.

III. Learning for mean-field teams

For the ease of exposition, we describe the algorithm for the special case when all types are of the same dimension (i.e., $d_x^m = d_x$ and $d_u^m = d_u$ for all $m \in M$) and the same number of agents (i.e., $|N^m| = n$ for all $m \in M$). We further assume that $d_v^0 = d_x$ and $\mathsf{F}^m = I$. Moreover, we assume noise covariances are given as $\mathsf{W}^i = \sigma_w^2 I$, $i \in N$, $\mathsf{V}^m = \sigma_v^2 I$, $m \in M$, and $\mathsf{V}^0 = \sigma_{v^0}^2 I$.

The above assumptions are not strictly needed for the analysis but we impose them because, under these assumptions, the covariance matrices $\bar{\Sigma}$ and $\breve{\Sigma}^m$ are scaled identity matrices. In particular, for any $m \in M$, $\breve{\Sigma}^m = (1 - \frac{1}{n})\sigma_w^2 I =: \breve{\sigma}^2 I$ and $\bar{\Sigma} = (\frac{\sigma_w^2}{n} + \sigma_v^2 + \sigma_{v^0}^2)I =: \bar{\sigma}^2 I$. This simpler form of the covariance matrices simplifies the description of the algorithm and the regret bounds.

Following the decomposition presented in Sec. II-B, we define $\bar{\theta}^{\mathsf{T}} = [\bar{\mathsf{A}}, \bar{\mathsf{B}}]$ to be the parameters of the mean-field dynamics (9) and $(\breve{\theta}^m)^{\mathsf{T}} = [\mathsf{A}^m, \mathsf{B}^m]$ to be the parameters of the relative dynamics (10). We let $\breve{\mathsf{S}}^m(\breve{\theta}^m)$ and $\bar{\mathsf{S}}(\bar{\theta})$ denote the solution to the Riccati equations (6) and $\breve{\mathsf{L}}^m(\breve{\theta}^m)$ and $\bar{\mathsf{L}}(\bar{\theta})$ denote the corresponding gains (7). Let $\breve{J}^m(\breve{\theta}^m) = \breve{\sigma}^2\operatorname{Tr}(\breve{\mathsf{S}}(\breve{\theta}^m))$ and $\bar{J}(\bar{\theta}) = \bar{\sigma}^2\operatorname{Tr}(\bar{\mathsf{S}}(\bar{\theta}))$ denote the performance of the m-th relative system and the mean-field system, respectively. As shown in Theorem 1,

$$J(\boldsymbol{\theta}) = \sum_{m\in M}\breve{J}^m(\breve{\theta}^m) + \bar{J}(\bar{\theta}). \tag{13}$$

[1] System matrices (A, B) are said to be stabilizable if there exists a gain matrix L such that all eigenvalues of $\mathsf{A} + \mathsf{B}\mathsf{L}$ are strictly inside the unit circle.

[2] For stabilizable (A, B) and $\mathsf{Q} > 0$, $\operatorname{DARE}(\mathsf{A}, \mathsf{B}, \mathsf{Q}, \mathsf{R})$ is the unique positive semidefinite solution of $\mathsf{S} = \mathsf{A}^{\mathsf{T}}\mathsf{S}\mathsf{A} - (\mathsf{A}^{\mathsf{T}}\mathsf{S}\mathsf{B})(\mathsf{R} + \mathsf{B}^{\mathsf{T}}\mathsf{S}\mathsf{B})^{-1}(\mathsf{A}^{\mathsf{T}}\mathsf{S}\mathsf{B}) + \mathsf{Q}$.

[3] The model considered in [23] did not include common noise, but it is easy to verify that their results continue to hold for models with common noise.

a) Prior and posterior beliefs:: We assume that the unknown parameters $\breve{\theta}^m$, $m \in M$, lie in compact subsets $\breve{\Theta}^m$ of $\mathbb{R}^{(d_x+d_u)\times d_x}$. Similarly, $\bar{\theta}$ lies in a compact subset $\bar{\Theta}$ of $\mathbb{R}^{|M|(d_x+d_u)\times|M|d_x}$. Let $\breve{\theta}^m(\ell)$ denote the ℓ-th column of $\breve{\theta}^m$. Thus $\breve{\theta}^m = \mathrm{cols}(\breve{\theta}^m(1), \dots, \breve{\theta}^m(d_x))$. Similarly, let $\bar{\theta}(\ell)$ denote the ℓ-th column of $\bar{\theta}$. Thus, $\bar{\theta} = \mathrm{cols}(\bar{\theta}(1), \dots, \bar{\theta}(|M|d_x))$.

We use $\mathcal{N}(\mu, \Sigma)$ to denotes the Gaussian distribution with mean μ and covariance Σ and $p\big|_\Theta$ to denote the projection of probability distribution p on the set Θ.

We assume that the priors $\bar{p}_1$ and $\breve{p}_1^m, m \in M$, on $\bar{\theta}$ and $\breve{\theta}^m, m \in M$, respectively, satisfy the following:

(A3) $\bar{p}_1$ is given as:

$$\bar{p}_1(\bar{\theta}) = \big[\prod_{\ell=1}^{|M|d_x} \bar{\lambda}_1^\ell(\bar{\theta}(\ell)) \big]\Big|_{\bar{\Theta}}$$

where for $\ell \in \{1, \dots, |M|d_x\}$, $\bar{\lambda}_1^\ell = \mathcal{N}(\bar{\mu}_1(\ell), \bar{\Sigma}_1)$ with mean $\bar{\mu}_1(\ell) \in \mathbb{R}^{|M|(d_x+d_u)}$ and positive-definite covariance $\bar{\Sigma}_1 \in \mathbb{R}^{|M|(d_x+d_u)\times|M|(d_x+d_u)}$.

(A4) $\breve{p}_1^m$ is given as:

$$\breve{p}_1^m(\breve{\theta}^m) = \big[\prod_{\ell=1}^{d_x} \breve{\lambda}_1^{m,\ell}(\breve{\theta}^m(\ell)) \big]\Big|_{\breve{\Theta}^m}$$

where for $\ell \in \{1, \dots, d_x\}$, $\breve{\lambda}_1^{m,\ell} = \mathcal{N}(\breve{\mu}_1^m(\ell), \breve{\Sigma}_1^m)$ with mean $\breve{\mu}_1^m(\ell) \in \mathbb{R}^{d_x+d_u}$ and positive-definite covariance $\breve{\Sigma}_1^m \in \mathbb{R}^{(d_x+d_u)\times(d_x+d_u)}$.

These assumptions are similar to the assumptions on the prior in the recent literature on TS for LQ systems [14], [15].

Following the discussion after Theorem 1, we maintain separate posterior distributions on $\bar{\theta}$ and $\breve{\theta}^m$, $m \in M$. In particular, we maintain a posterior distribution $\bar{p}_t$ on $\bar{\theta}$ based on the mean-field state and action history as follows: for any Borel subset B of $\mathbb{R}^{|M|(d_x+d_u)\times|M|d_x}$,

$$\bar{p}_t(B) = \mathbb{P}(\bar{\theta} \in B \mid \bar{\boldsymbol{x}}_{1:t}, \bar{\boldsymbol{u}}_{1:t-1}). \tag{14}$$

For every $m \in M$, we also maintain a separate posterior distribution $\breve{p}_t^m$ on $\breve{\theta}^m$ as follows. At each time $t > 1$, we select an agent $j_{t-1}^m \in N^m$ as $\arg\max_{i\in N^m} (\breve{z}_{t-1}^i)^\intercal \breve{\Sigma}_{t-1}^m \breve{z}_{t-1}^i$, where $\breve{\Sigma}_{t-1}^m$ is a covariance matrix defined recursively by (18b). Then, for any Borel subset B of $\mathbb{R}^{(d_x+d_u)\times d_x}$,

$$\breve{p}_t^m(B) = \mathbb{P}(\breve{\theta}^m \in B \mid \{\breve{x}_s^{j_s^m}, \breve{u}_s^{j_s^m}, \breve{x}_{s+1}^{j_s^m}\}_{1\le s<t}\}), \tag{15}$$

See the supplementary file of [37] for a discussion on the rule to select j_{t-1}^m.

For the ease of notation, we use $\bar{\boldsymbol{z}}_t = \mathrm{vec}(\bar{z}_t^1, \dots, \bar{z}_t^{|M|})$, where $\bar{z}_t^m = \mathrm{vec}(\bar{x}_t^m, \bar{u}_t^m)$, and $\breve{z}_t^i = \mathrm{vec}(\breve{x}_t^i, \breve{u}_t^i)$. Then, we can write the dynamics (9)–(10) of the mean-field and the relative systems as

$$\bar{\boldsymbol{x}}_{t+1} = \bar{\theta}^\intercal \bar{\boldsymbol{z}}_t + \bar{\boldsymbol{w}}_t + \bar{\boldsymbol{v}}_t + v_t^0, \tag{16a}$$

$$\breve{x}_{t+1}^i = (\breve{\theta}^m)^\intercal \breve{z}_t^i + \breve{w}_t^i, \quad \forall i \in N^m, m \in M. \tag{16b}$$

Recall that $\bar{\sigma}^2 = \sigma_w^2/n + \sigma_v^2 + \sigma_{v^0}^2$ and $\breve{\sigma}^2 = (1 - \frac{1}{n})\sigma_w^2$.

Lemma 1 *The posterior distributions are as follows:*

1) *The posterior on $\bar{\theta}$ is*

$$\bar{p}_t = \big[\prod_{\ell=1}^{|M|d_x} \bar{\lambda}_t^\ell(\bar{\theta}(\ell)) \big]\Big|_{\bar{\Theta}},$$

where for $\ell \in \{1, \dots, |M|d_x\}$, $\bar{\lambda}_t^\ell = \mathcal{N}(\bar{\mu}_t(\ell), \bar{\Sigma}_t)$, and

$$\bar{\mu}_{t+1}(\ell) = \bar{\mu}_t(\ell) + \frac{\bar{\Sigma}_t \bar{\boldsymbol{z}}_t \big(\bar{\boldsymbol{x}}_{t+1}(\ell) - \bar{\mu}_t(\ell)^\intercal \bar{\boldsymbol{z}}_t\big)}{\bar{\sigma}^2 + (\bar{\boldsymbol{z}}_t)^\intercal \bar{\Sigma}_t \bar{\boldsymbol{z}}_t}, \tag{17a}$$

$$\bar{\Sigma}_{t+1}^{-1} = \bar{\Sigma}_t^{-1} + \frac{1}{\bar{\sigma}^2} \bar{\boldsymbol{z}}_t \bar{\boldsymbol{z}}_t^\intercal. \tag{17b}$$

2) *The posterior on $\breve{\theta}^m$, $m \in M$, at time t is*

$$\breve{p}_t^m(\breve{\theta}^m) = \big[\prod_{\ell=1}^{d_x} \breve{\lambda}_t^{m,\ell}(\breve{\theta}^m(\ell)) \big]\Big|_{\breve{\Theta}^m},$$

where for $\ell \in \{1, \dots, d_x\}$, $\breve{\lambda}_t^{m,\ell} = \mathcal{N}(\breve{\mu}_t^m(\ell), \breve{\Sigma}_t^m)$, and

$$\breve{\mu}_{t+1}^m(\ell) = \breve{\mu}_t^m(\ell) + \frac{\breve{\Sigma}_t^m \breve{z}_t^{j_t^m} \big(\breve{x}_{t+1}^{j_t^m}(\ell) - \breve{\mu}_t^m(\ell)^\intercal \breve{z}_t^{j_t^m}\big)}{\breve{\sigma}^2 + (\breve{z}_t^{j_t^m})^\intercal \breve{\Sigma}_t^m \breve{z}_t^{j_t^m}}, \tag{18a}$$

$$(\breve{\Sigma}_{t+1}^m)^{-1} = (\breve{\Sigma}_t^m)^{-1} + \frac{1}{\breve{\sigma}^2} \breve{z}_t^{j_t^m} (\breve{z}_t^{j_t^m})^\intercal. \tag{18b}$$

PROOF Note that the dynamics of $\bar{\boldsymbol{x}}_t$ and $\breve{x}_t^i$ in (16) are linear and the noises $\bar{\boldsymbol{w}}_t + \bar{\boldsymbol{v}}_t + v_t^0$ and $\breve{w}_t^i$ are Gaussian. Therefore, the result follows from standard results in Gaussian linear regression [38]. ∎

b) The Thompson sampling algorithm:: We propose a Thompson sampling algorithm referred to as TSDE-MF which is inspired by the TSDE (Thompson sampling with dynamic episodes) algorithm proposed in [14], [15] and the structure of the optimal planning solution for the mean-field teams described in Sec. II-B.

The TSDE-MF algorithm consists of a coordinator $\mathcal{C}$ and $|M|+1$ *actors*: a mean-field actor $\bar{\mathcal{A}}$ and a relative actor $\breve{\mathcal{A}}^m$, for each $m \in M$. These actors are described below while the whole algorithm is presented in Algorithm 1.

- At each time, the coordinator $\mathcal{C}$ observes the current global state $(x_t^i)_{i\in N}$, computes the mean-field state $\bar{\boldsymbol{x}}_t$ and the relative states $(\breve{x}_t^i)_{i\in N}$, and sends the mean-field state $\bar{\boldsymbol{x}}_t$ to be the mean-field actor $\bar{\mathcal{A}}$ and the relative states $\breve{\boldsymbol{x}}_t^m = (\breve{x}_t^i)_{i\in N^m}$ of the all the agents of type m to the relative actor $\breve{\mathcal{A}}^m$. The mean-field actor $\bar{\mathcal{A}}$ computes the mean-field control $\bar{\boldsymbol{u}}_t$ and the relative actor $\breve{\mathcal{A}}^m$ computes the relative control $\breve{\boldsymbol{u}}_t^m = (\breve{u}_t^i)_{i\in N^m}$ (as per the details presented below) and sends it back to the coordinator $\mathcal{C}$. The coordinator then computes and executes the control action $u_t^i = \bar{u}_t^m + \breve{u}_t^i$ for each agent i of type m.
- The mean-field actor $\bar{\mathcal{A}}$ maintains the posterior $\bar{p}_t$ on $\bar{\theta}$ according to (17). The actor works in episodes of dynamic length. Let $\bar{t}_k$ and $\bar{T}_k$ denote the start and the length of episode k, respectively. Episode k ends if the determinant of covariance $\bar{\Sigma}_t$ falls below half of its

value at the beginning of the episode (i.e., $\det(\bar{\Sigma}_t) < 0.5\det(\bar{\Sigma}_{\bar{t}_k})$) or if the length of the episode is one more than the length of the previous episode (i.e., $t - \bar{t}_k > \bar{T}_{k-1}$). Thus,

$$\bar{t}_{k+1} = \min\{t > t_k : \det(\bar{\Sigma}_t) < 0.5\det(\bar{\Sigma}_{t_k}) \text{ or } t - \bar{t}_k > \bar{T}_{k-1}\}. \quad (19)$$

At the beginning of episode k, the mean-field actor $\bar{\mathcal{A}}$ samples a parameter $\bar{\theta}_k$ from the posterior distribution $\bar{p}_t$. During episode k, the mean-field actor $\bar{\mathcal{A}}$ generates the mean-field controls using the samples $\bar{\theta}_k$, i.e., $\bar{\boldsymbol{u}}_t = \bar{\mathsf{L}}(\bar{\theta}_k)\bar{\boldsymbol{x}}_t$.

- Each relative actor $\breve{\mathcal{A}}^m$ is similar to the mean-field actor. Actor $\breve{\mathcal{A}}^m$ maintains the posterior $\breve{p}^m$ on $\breve{\theta}^m$ according to (18). The actor works in episodes of dynamic length. The episodes of each relative actor $\breve{\mathcal{A}}^m$ and the mean-field actor $\bar{\mathcal{A}}$ are separate from each other.[4] Let $\breve{t}^m_k$ and $\breve{T}^m_k$ denote the start and length of episode k, respectively. The termination condition for each episode is similar to that of the mean-field actor $\bar{\mathcal{A}}$. In particular,

$$\breve{t}^m_{k+1} = \min\{t > t^m_k : \det(\breve{\Sigma}^m_t) < 0.5\det(\breve{\Sigma}^m_{t^m_k}) \text{ or } t - \breve{t}^m_k > \breve{T}^m_{k-1}\}. \quad (20)$$

At the beginning of episode k, the relative actor $\breve{\mathcal{A}}^m$ samples a parameter $\breve{\theta}^m_k$ from the posterior distribution $\breve{p}^m_t$. During episode k, the relative actor $\breve{\mathcal{A}}^m$ generates the relative controls using the sample $\breve{\theta}^m_k$, i.e., $\breve{\boldsymbol{u}}^m_t = (\breve{\mathsf{L}}^m(\breve{\theta}^m_k)\breve{x}^i_t)_{i\in N^m}$.

Note that the algorithm does not depend on the horizon T. A partially distributed version of the algorithm is presented in the conclusion.

c) Regret bounds:: We make the following assumption to ensure that the closed loop dynamics of the mean field state and the relative states of each agent are stable. We use the notation $\|\cdot\|$ to denote the induced norm of a matrix.

(A5) There exists $\delta \in (0,1)$ such that

- for any $\bar{\theta}, \bar{\phi} \in \bar{\Theta}$ where $\bar{\theta}^{\mathsf{T}} = [\bar{\mathsf{A}}_{\bar{\theta}}, \bar{\mathsf{B}}_{\bar{\theta}}]$, we have $\|\bar{\mathsf{A}}_{\bar{\theta}} + \bar{\mathsf{B}}_{\bar{\theta}}\bar{\mathsf{L}}(\bar{\phi})\| \le \delta$.
- for any $m \in M$, $\breve{\theta}^m, \breve{\phi}^m \in \breve{\Theta}^m$, where $(\breve{\theta}^m)^{\mathsf{T}} = [\mathsf{A}_{\breve{\theta}^m}, \mathsf{B}_{\breve{\theta}^m}]$, we have $\|\mathsf{A}_{\breve{\theta}^m} + \mathsf{B}_{\breve{\theta}^m}\breve{\mathsf{L}}(\breve{\phi}^m)\| \le \delta$.

This assumption is similar to an assumption imposed in the literature on TS for LQ systems [15]. According to Theorem 11 in [12], the assumption is satisfied if

$$\bar{\Theta} = \{(\bar{\mathsf{A}}, \bar{\mathsf{B}}) : \|\bar{\mathsf{A}} - \bar{\mathsf{A}}_0\| \le \bar{\epsilon}, \|\bar{\mathsf{B}} - \bar{\mathsf{B}}_0\| \le \bar{\epsilon}\}$$
$$\breve{\Theta}^m = \{(\breve{\mathsf{A}}^m, \breve{\mathsf{B}}^m) : ||\breve{\mathsf{A}}^m - \breve{\mathsf{A}}^m_0|| \le \breve{\epsilon}^m, \|\breve{\mathsf{B}}^m - \breve{\mathsf{B}}^m_0\| \le \breve{\epsilon}^m\}$$

for stabilizable $(\bar{\mathsf{A}}^m_0, \bar{\mathsf{B}}^m_0)$ and $(\breve{\mathsf{A}}^m_0, \breve{\mathsf{B}}^m_0)$, and small constants $\bar{\epsilon}$ $\breve{\epsilon}^m$ depending on the choice of $(\bar{A}^m_0, \bar{B}^m_0)$ and $(\breve{\mathsf{A}}^m_0, \breve{\mathsf{B}}^m_0)$. In other words, the assumption holds when the true system is in a small neighborhood of a known nominal system, and the small neighborhood can be learned with high probability by running some stabilizing procedure [12].

[4]We use the episode count k as a local variable which is different for each actor.

Algorithm 1 TSDE-MF

initialize mean-field actor: $\bar{\Theta}$, $(\bar{\mu}_1, \bar{\Sigma}_1)$, $\bar{t}_0 = 0$, $\bar{T}_{-1} = 0$, $k = 0$
initialize relative-actor-m: $\breve{\Theta}^m$, $(\breve{\mu}^m_1, \breve{\Sigma}^m_1)$, $\breve{t}^m_0 = 0$, $\breve{T}^m_{-1} = 0$, $k = 0$
for $t = 1, 2, \ldots$ **do**
  observe $(x^i_t)_{i\in N}$
  compute $\bar{\boldsymbol{x}}_t$, $(\breve{\boldsymbol{x}}^m_t)_{m\in M}$
  $\bar{\boldsymbol{u}}_t \leftarrow$ MEAN-FIELD-ACTOR$(\bar{\boldsymbol{x}}_t)$
  for $m \in M$ **do**
    $\breve{\boldsymbol{u}}^m_t \leftarrow$ RELATIVE-ACTOR-$m(\breve{\boldsymbol{x}}^m_t)$
    for $i \in N^m$ **do**
      agent i applies control $u^i_t = \bar{u}^m_t + \breve{u}^i_t$
    end for
  end for
end for

function MEAN-FIELD-ACTOR$(\bar{\boldsymbol{x}}_t)$
  global var t
  Update $\bar{p}_t$ according (17)
  if $t - \bar{t}_k > \bar{T}_{k-1}$ or $\det(\bar{\Sigma}_t) < 0.5\det(\bar{\Sigma}_k)$ **then**
    $T_k \leftarrow t - \bar{t}_k$, $k \leftarrow k+1$, $\bar{t}_k \leftarrow t$
    sample $\bar{\theta}_k \sim \bar{p}_t$
    $\bar{\mathsf{L}} \leftarrow \bar{\mathsf{L}}(\bar{\theta}_k)$
  end if
  return $\bar{\mathsf{L}}\bar{\boldsymbol{x}}_t$
end function

function RELATIVE-ACTOR-$m((\breve{x}^i_t)_{i\in N^m})$
  global var t
  Update $\breve{p}^m_t$ according (18)
  if $t - \breve{t}^m_k > \breve{T}^m_{k-1}$ or $\det(\breve{\Sigma}^m_t) < 0.5\det(\breve{\Sigma}^m_k)$ **then**
    $T^m_k \leftarrow t - \breve{t}^m_k$, $k \leftarrow k+1$, $\breve{t}^m_k \leftarrow t$
    sample $\breve{\theta}^m_k \sim \breve{p}^m_t$
    $\breve{\mathsf{L}}^m \leftarrow \breve{\mathsf{L}}^m(\breve{\theta}^m_k)$
  end if
  return $(\breve{\mathsf{L}}^m\breve{x}^i_t)_{i\in N^m}$
end function

Theorem 2 *Under* (A1)–(A5)*, the regret of* TSDE-MF *is upper bounded as follows:*

$$R(T; \text{TSDE-MF}) \le \tilde{\mathcal{O}}\big((\bar{\sigma}^2|M|^{1.5} + \breve{\sigma}^2|M|)d_x^{0.5}(d_x+d_u)\sqrt{T}\big).$$

Recall that $\bar{\sigma}^2 = \sigma_w^2/n + \sigma_v^2 + \sigma_{v^0}^2$ and $\breve{\sigma}^2 = (1 - \frac{1}{n})\sigma_w^2$. So, we can say that $R(T; \text{TDSE-MF}) \le \tilde{\mathcal{O}}(\bar{\sigma}^2|M|^{1.5}d_x^{0.5}(d_x+d_u)\sqrt{T})$. Compared with the original TSDE regret $\tilde{\mathcal{O}}(n^{1.5}|M|^{1.5}\sqrt{T})$ which scales superlinear with the number of agents, the regret of the proposed algorithm is bounded by $\tilde{\mathcal{O}}(|M|^{1.5}\sqrt{T})$ irrespective of the total number of agents.

The following special cases are of interest:

- In the absence of common noises (i.e., $\sigma_v^2 = \sigma_{y^0}^2 = 0$), and when $n \gg |M|$, $R(T; \texttt{TDSE-MF}) \le \tilde{\mathcal{O}}(\breve{\sigma}^2 |M| d_x^{0.5}(d_x + d_u)\sqrt{T})$.
- For homogeneous systems (i.e., $|M| = 1$), we have $R(T; \texttt{TDSE-MF}) \le \tilde{\mathcal{O}}((\bar{\sigma}^2 + \breve{\sigma}^2) d_x^{0.5}(d_x + d_u)\sqrt{T})$. Thus, the scaling with the number of agents is $\tilde{\mathcal{O}}((1 + \frac{1}{n})\sqrt{T})$.

Note that these results show that in mean-field systems with common noise regret scales as $\mathcal{O}(|M|^{1.5})$ in the number of types, while in mean-field systems without common noise, the regret scales as $\mathcal{O}(|M|)$. *Thus, the presence of common noise fundamentally changes the scaling of the learning algorithm.*

IV. Regret analysis

For the ease of notation, we simply use $R(T)$ instead of $R(T; \texttt{TSDE-MF})$ in this section. Eq. (13) and (8) imply that the regret may be decomposed as

$$R(T) = \bar{R}(T) + \sum_{m \in M} \frac{1}{n} \sum_{i \in N^m} \breve{R}^{i,m}(T) \tag{21}$$

where

$$\bar{R}(T) := \mathbb{E}\Bigg[\sum_{t=1}^{T} \bar{c}(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) - T\bar{J}(\bar{\theta})\Bigg],$$

$$\breve{R}^{i,m}(T) := \mathbb{E}\Bigg[\sum_{t=1}^{T} \breve{c}^m(\breve{x}_t^i, \breve{u}_t^i) - T\breve{J}(\theta^{\breve{m}})\Bigg].$$

Note that $\bar{R}(T)$ is the regret associated with the mean-field system and $\breve{R}^{i,m}(T)$ is the regret of the i-th relative system of type m. Observe that for the mean-field actor in our algorithm is essentially implementing the TSDE algorithm of [14], [15] for the mean-field system with dynamics (9) and per-step cost $\bar{c}(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$. This is because:

1) As mentioned in the discussion after Theorem 1, we can view $\bar{\boldsymbol{u}}_t = \bar{\mathsf{L}}(\bar{\theta})\bar{\boldsymbol{x}}_t$ as the optimal control action of the mean-field system.
2) The posterior distribution $\bar{p}_t$ on $\bar{\theta}$ depends only on $(\bar{\boldsymbol{x}}_{1:t}, \bar{\boldsymbol{u}}_{1:t-1})$.

Thus, $\bar{R}(T)$ is precisely the regret of the TSDE algorithm analyzed in [15]. Therefore, we have the following.

Lemma 2 *For the mean-field system,*

$$\bar{R}(T) \le \tilde{\mathcal{O}}\big(\bar{\sigma}^2 |M|^{1.5} d_x^{0.5}(d_x + d_u)\sqrt{T}\big). \tag{22}$$

Unfortunately, we cannot use the same argument to bound $\breve{R}^{i,m}(T)$. Even though we can view $\breve{u}_t^i = \breve{\mathsf{L}}^m(\breve{\theta}^m)\breve{x}_t^i$ as the optimal control action of the LQ system with dynamics (10), the posterior $\breve{p}_t^m$ on $\breve{\theta}^m$ depends on terms other than $(\breve{x}_{1:t}^i, \breve{u}_{1:t-1}^i)$. Therefore, we cannot directly use the results of [15] to bound $\breve{R}^{i,m}(T)$. In the rest of this section, we present a bound on $\breve{R}^{i,m}(T)$.

For the ease of notation, for any episode k, we use $\breve{\mathsf{L}}_k^m$ and $\breve{\mathsf{S}}_k^m$ to denote $\breve{\mathsf{L}}^m(\breve{\theta}_k^m)$ and $\breve{\mathsf{S}}^m(\breve{\theta}_k^m)$. Recall that the relative value function for average cost LQ problem is $x^{\mathsf{T}}\mathsf{S}x$, where S is the solution to DARE. Therefore, at any time t, episode k, agent i of type m, and state $\breve{x}_t^i \in \mathbb{R}^{d_x}$, with $\breve{u}_t^i = \breve{\mathsf{L}}_k^m \breve{x}_t^i$ and $\breve{z}_t^i = \mathrm{vec}(\breve{x}_1^i, \breve{u}_t^i)$, the average cost Bellman equation is

$$\begin{aligned}\breve{J}^m(\breve{\theta}_k^m) + (\breve{x}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_t^i = \breve{c}^m(\breve{x}_t^i, \breve{u}_t^i) \\ + \mathbb{E}\big[((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i + \breve{w}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i + \breve{w}_t^i)\big].\end{aligned}$$

Adding and subtracting $\mathbb{E}[(\breve{x}_{t+1}^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_{t+1}^i \mid \breve{z}_t^i]$ and noting that $\breve{x}_{t+1}^i = (\breve{\theta}^m)^{\mathsf{T}}\breve{z}_t^i + \breve{w}_t^i$, we get that

$$\begin{aligned}\breve{c}^m(\breve{x}_t^i, \breve{u}_t^i) = \breve{J}^m(\breve{\theta}_k^m) + (\breve{x}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_t^i - \mathbb{E}[(\breve{x}_{t+1}^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_{t+1}^i | \breve{z}_t^i] \\ + ((\breve{\theta}^m)^{\mathsf{T}}\breve{z}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m((\breve{\theta}^m)^{\mathsf{T}}\breve{z}_t^i) - ((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i).\end{aligned} \tag{23}$$

Let $\breve{K}_T^m$ denote the number of episodes of the relative systems of type m until the horizon T. For each $k > \breve{K}_T^m$, we define $\breve{t}_k^m$ to be $T+1$. Then, using (23), we have that for any agent i of type m,

$$\begin{aligned}\breve{R}^{i,m}(T) = \underbrace{\mathbb{E}\Bigg[\sum_{k=1}^{\breve{K}_T^m} \breve{T}_k^m \breve{J}^m(\breve{\theta}_k^m) - T\breve{J}^m(\breve{\theta}^m)\Bigg]}_{\text{regret due to sampling error} =: \breve{R}_0^{i,m}(T)} \\ + \underbrace{\mathbb{E}\Bigg[\sum_{k=1}^{\breve{K}_T^m} \sum_{t=\breve{t}_k^m}^{\breve{t}_{k+1}^m - 1} \big[(\breve{x}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_t^i - (\breve{x}_{t+1}^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m \breve{x}_{t+1}^i\big]\Bigg]}_{\text{regret due to time-varying controller} =: \breve{R}_1^{i,m}(T)} \\ + \underbrace{\mathbb{E}\Bigg[\sum_{k=1}^{\breve{K}_T^m} \sum_{t=\breve{t}_k^m}^{\breve{t}_{k+1}^m - 1} \big[((\breve{\theta}^m)^{\mathsf{T}}\breve{z}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m((\breve{\theta}^m)^{\mathsf{T}}\breve{z}_t^i) \\ - ((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i)^{\mathsf{T}}\breve{\mathsf{S}}_k^m((\breve{\theta}_k^m)^{\mathsf{T}}\breve{z}_t^i)\big]\Bigg]}_{\text{regret due to model mismatch} =: \breve{R}_2^{i,m}(T)}.\end{aligned} \tag{24}$$

Lemma 3 *The terms in* (24) *are bounded as follows:*

1) $\breve{R}_0^{i,m}(T) \le \tilde{\mathcal{O}}(\breve{\sigma}^2\sqrt{(d_x + d_u)T})$.
2) $\breve{R}_1^{i,m}(T) \le \tilde{\mathcal{O}}(\breve{\sigma}^2\sqrt{(d_x + d_u)T})$.
3) $\breve{R}_2^{i,m}(T) \le \tilde{\mathcal{O}}(\breve{\sigma}^2(d_x + d_u)\sqrt{d_x T})$.

Proof We provide an outline of the proof. See the supplementary file of [37] for complete details.

The first term $\breve{R}_0^{i,m}(T)$ can be bounded using the basic property of Thompson sampling: for any measurable function f, $\mathbb{E}[f(\breve{\theta}_k^m)] = \mathbb{E}[f(\breve{\theta}^m)]$ because $\breve{\theta}_k^m$ is a sample from the posterior distribution on $\breve{\theta}^m$.

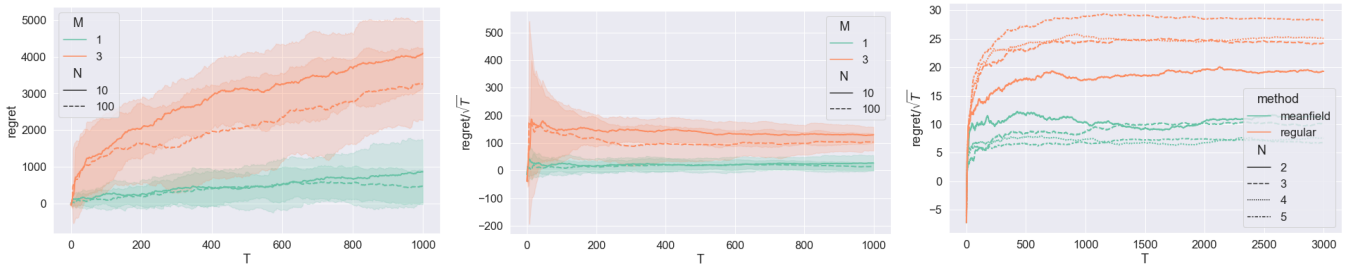
Note that the second term $\breve{R}_1^{i,m}(T)$ is a telescopic sum, which we can simplify to establish

$$\breve{R}_1^{i,m}(T) \le \mathcal{O}\big(\mathbb{E}[\breve{K}_T^m(\breve{X}_T^i)^2]\big),$$

where $\breve{X}_t^i = \max_{1 \le t \le T} \|\breve{x}_t^i\|$ is the maximum norm of the relative state along the entire trajectory. The final bound on $\breve{R}_1^{i,m}(T)$ can be obtained by bounding $\breve{K}_T^m$ and $\mathbb{E}[(\breve{X}_T^i)^2]$.

Using the sampling condition for $\breve{p}_t^m$ and an existing bound in the literature, we first establish that

$$\breve{R}_2^i(T) \le \sqrt{\mathbb{E}\big[(\breve{X}_T^i)^2 \textstyle\sum_{t=1}^{T} (\breve{z}_t^i)^{\mathsf{T}}\breve{\Sigma}_t^m \breve{z}_t^i\big]} \times \tilde{\mathcal{O}}(\sqrt{T})$$



(a) $R(T)$ vs T for TSDE−MF (b) $R(T)/\sqrt{T}$ vs T for TSDE−MF (c) TSDE−MF vs TSDE

Fig. 1: Expected regret vs time.

Then, we upper bound $(\breve{z}_t^i)^\intercal \breve{\Sigma}_t^m \breve{z}_t^i$ by $(\breve{z}_t^{j_t^m})^\intercal \breve{\Sigma}_t^m \breve{z}_t^{j_t^m}$ which follows from the definition of j_t^m. Finally, we show that $\mathbb{E}[(\breve{X}_T^i)^2 \sum_t (\breve{z}_t^{j_t^m})^\intercal \breve{\Sigma}_t^m \breve{z}_t^{j_t^m}]$ is $\tilde{\mathcal{O}}(1)$ using the fact that $(\breve{\Sigma}_t^m)^{-1}$ is obtained by linearly combining $\{\breve{z}_s^{j_s^m} (\breve{z}_s^{j_s^m})^\intercal\}_{1 \le s < t}$ as in (18b). ∎

Combining the three bounds in Lemma 3, we get that

$$\breve{R}^{i,m}(T) \le \tilde{\mathcal{O}}(\breve{\sigma}^2 d_x^{0.5}(d_x + d_u)\sqrt{T}). \tag{25}$$

By subsituting (22) and (25) in (21), we get the result of Theorem 2.

V. Numerical Experiments

In this section, we illustrate the performance of TSDE−MF for a homogeneous (i.e., $|M| = 1$) mean-field LQ system for different values of the number n of agents, with $\mathsf{A} = 1, \mathsf{B} = 0.3, \mathsf{D} = 0.5, \mathsf{E} = 0.2, \mathsf{Q} = 1, \bar{\mathsf{Q}} = 1, \mathsf{R} = 1$,and $\bar{\mathsf{R}} = 0.5$. We set the local noise variance $\sigma_w^2 = 1$.

For the regret plots in Figure 1a,1b, we set the common noise variance to $\sigma_v^2 + \sigma_{v^0}^2 = 1$. The prior distribution used in the simulation are set according to **(A3)** and **(A4)** with $\breve{\mu}(\ell) = [1, 1]$, $\bar{\mu}(\ell) = [1, 1]$, $\bar{\Sigma}_1 = I$, and $\breve{\Sigma}_1 = I$, $\breve{\Theta} = \{\breve{\theta} : \mathsf{A} + \mathsf{B}\breve{\mathsf{L}}(\breve{\theta}) \le \delta\}$, $\bar{\Theta} = \{\bar{\theta} : \mathsf{A} + \mathsf{D} + (\mathsf{B} + \mathsf{E})\bar{\mathsf{L}}(\bar{\theta}) \le \delta\}$ and $\delta = 0.99$.

In the comparison of TSDE−MF method with TSDE in Figure 1c, we consider the same dynamics and cost parameters as above but without common noise (i.e. $\sigma_v^2 + \sigma_{v^0}^2 = 0$).

a) Empirical evaluation of regret:: We run the system for 500 different sample paths and plot the mean and standard deviation of the expected regret $R(T)$ for $T = 5000$. The regret for different values of n is shown in 1a–1b. As seen from the plots, the regret reduces with the number of agents and $R(T)/\sqrt{T}$ converges to a constant. Thus, the empirical regret matches the upper bound of $\tilde{\mathcal{O}}((1 + \frac{1}{n})\sqrt{T})$ obtained in Theorem 2.

b) Comparison with naive TSDE algorithm:: We compare the performance of TSDE−MF with that of directly using the TSDE algorithm presented in [14], [15] for different values of n. The results are shown in Fig. 1c. As seen from the plots, the regret of TSDE−MF is smaller than TSDE but more importantly, the regret of TSDE−MF reduces with n while that of TSDE increases with n. This matches their respective upper bounds of $\tilde{\mathcal{O}}((1+\frac{1}{n})\sqrt{T})$ and $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$. These plots clearly illustrate the significance of our results even for small values of n.

VI. Conclusion

We consider the problem of controlling an unknown LQ mean-field team. The planning solution (i.e., when the model is known) for mean-field teams is obtained by solving the mean-field system and the relative systems separately. Inspired by this feature, we propose a TS-based learning algorithm TSDE−MF which separately tracks the parameters $\bar{\theta}$ and $\breve{\theta}^m$ of the mean-field and the relative systems, respectively. The part of the TSDE−MF algorithm that learns the mean-field system is similar to the TSDE algorithm for single agent LQ systems proposed in [14], [15] and its regret can be bounded using the results of [14], [15]. However, the part of the TSDE−MF algorithm that learns the relative component is different and we cannot directly use the results of [14], [15] to bound its regret. Our main technical contribution is to provide a bound on the regret on the relative system, which allows us to bound the total regret under TSDE−MF.

a) Distributed implementation of the algorithm:: It is possible to implement Algorithm 1 in a distributed manner as follows. Instead of a centralized coordinator which collects all the observations and computes all the controls, we can consider an alternative implementation in which there is an actor $\mathcal{A}^m$ associated with type m and a mean-field actor $\bar{\mathcal{A}}$. Each agent observes its local state and action. The actor $\mathcal{A}^m$ for type m computes $(j_t^m, \bar{x}_t^m)$ using a distributed algorithm, sends $\bar{x}_t^m$ to the mean-field actor, and locally computes $\breve{\mathsf{L}}^m(\breve{\theta}_k)$. The mean-field actor computes $\bar{\mathsf{L}}(\bar{\theta}_k)$ and sends the m-th block column $\bar{\mathsf{L}}^m(\bar{\theta}_k)$ to actors $\mathcal{A}^m$. Each actor $\mathcal{A}^m$ then sends $(\bar{x}_t^m, \bar{\mathsf{L}}^m(\bar{\theta}_k), \breve{\mathsf{L}}^m(\breve{\theta}_k))$ to each agent of type m using a distributed algorithm. Each agent then applies the control law (11).

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[2] 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.

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

[4] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time analysis of optimal adaptive policies for linear-quadratic systems." arXiv:1711.07230, 2017.

[5] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *International Conference on Machine Learning*, pp. 1300–1309, PMLR, 2019.

[6] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation," in *International Conference on Machine Learning*, pp. 23–31, PMLR, 2020.

[7] G. Goodwin, P. Ramadge, and P. Caines, "Discrete time stochastic multivariable adaptive control," *IEEE Transactions on Automatic Control*, vol. 19, pp. 449–456, June 1980.

[8] G. Goodwin, P. Ramadge, and P. Caines, "Discrete time stochastic adaptive control," *SIAM J. Control and Optimization*, vol. 19, pp. 829–853, Nov. 1981.

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

[10] H. Mania, S. Tu, and B. Recht, "Certainty equivalent control of LQR is efficient." arXiv:1902.07826, 2019.

[11] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[12] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*, pp. 8937–8948, PMLR, 2020.

[13] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, 2012.

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

[15] Y. Ouyang, M. Gagrani, and R. Jain, "Posterior sampling-based reinforcement learning for control of unknown linear systems," *IEEE Transactions on Automatic Control*, 2019.

[16] 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.

[17] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *International Conference on Machine Learning*, pp. 1328–1337, PMLR, 2020.

[18] J. Lunze, "Dynamics of strongly coupled symmetric composite systems," *International Journal of Control*, vol. 44, no. 6, pp. 1617–1640, 1986.

[19] M. K. Sundareshan and R. M. Elbanna, "Qualitative analysis and decentralized controller synthesis for a class of large-scale systems with symmetrically interconnected subsystems," *Automatica*, vol. 27, no. 2, pp. 383–388, 1991.

[20] G.-H. Yang and S.-Y. Zhang, "Structural properties of large-scale systems possessing similar structures," *Automatica*, vol. 31, no. 7, pp. 1011–1017, 1995.

[21] S. C. Hamilton and M. E. Broucke, "Patterned linear systems," *Automatica*, vol. 48, no. 2, pp. 263–272, 2012.

[22] J. Arabneydi and A. Mahajan, "Team-optimal solution of finite number of mean-field coupled lqg subsystems," in *Conf. Decision and Control*, (Kyoto, Japan), Dec. 2015.

[23] J. Arabneydi and A. Mahajan, "Linear Quadratic Mean Field Teams: Optimal and Approximately Optimal Decentralized Solutions," 2016. arXiv:1609.00056.

[24] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[25] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized epsilon-Nash equilibria," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1560–1571, 2007.

[26] M. Huang, P. E. Caines, and R. P. Malhamé, "Social optima in mean field LQG control: centralized and decentralized strategies," *IEEE Transactions on Automatic Control*, vol. 57, no. 7, pp. 1736–1751, 2012.

[27] G. Y. Weintraub, C. L. Benkard, and B. V. Roy, "Oblivious Equilibrium: A Mean Field Approximation for Large-Scale Dynamic Games," in *Neural Information Processing Systems*, pp. 1489–1496, Dec. 2005.

[28] G. Y. Weintraub, C. L. Benkard, and B. Van Roy, "Markov perfect industry dynamics with many firms," *Econometrica*, vol. 76, no. 6, pp. 1375–1411, 2008.

[29] D. A. Gomes and J. Saúde, "Mean field games models—a brief survey," *Dynamic Games and Applications*, vol. 4, no. 2, pp. 110–154, 2014.

[30] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International Conference on Machine Learning*, pp. 5567–5576, Jul 2018.

[31] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games," in *International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 251–259, 2019.

[32] N. Tiwari, A. Ghosh, and V. Aggarwal, "Reinforcement learning for mean field game," *arXiv preprint arXiv:1905.13357*, 2019.

[33] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," in *Neural Information Processing Systems*, pp. 4966–4976, 2019.

[34] S. G. Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, "Multi type mean field reinforcement learning," *arXiv preprint arXiv:2002.02513*, 2020.

[35] M. A. uz Zaman, K. Zhang, E. Miehling, and T. Başar, "Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2278–2284, IEEE, 2020.

[36] A. Angiuli, J.-P. Fouque, and M. Laurière, "Unified reinforcement Q-learning for mean field game and control problems." arXiv:2006.13912, 2020.

[37] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "Thompson sampling for linear quadratic mean-field teams." arXiv preprint arXiv:2011.04686, 2020.

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