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

Weakly Coupled Constrained Markov Decision Processes in Borel Spaces

Mukul Gagrani and Ashutosh Nayyar

Abstract— **Consider a multi-agent stochastic control problem where the agents have decoupled system dynamics. Each agent has an associated cost function and a constraint function. The agents want to find a control strategy which minimizes their long term average cumulative cost function while keeping the long term average cumulative constraint function below a certain threshold. This problem is referred to as weakly coupled constrained Markov decision process (MDP). In this paper, we consider the problem of weakly coupled constrained MDP with Borel state and action spaces. We use the linear programming (LP) based approach of [1] to derive an occupation measure based LP to find the optimal decentralized control strategies for our problem. We show that randomized stationary policies are optimal for each agent under some assumptions on the transition kernels, cost and the constraint functions. We further consider the special case of multi-agent Linear Quadratic Gaussian (LQG) systems and show that the optimal control strategy could be obtained by solving a semi-definite program (SDP). We illustrate our results through numerical experiments.**

I. Introduction

Consider a multi-agent system with N agents that have decoupled dynamics, i.e, each agent's state evolution depends only on its own actions. Each agent has an associated cost function and a constraint function that depend on its local state and local action. The agents are coupled because the time-average of the total constraint function (summed over all agents and all times) must be kept below a threshold. Such multi-agent problems are referred to as Weakly coupled constrained Markov decision process (MDP) ([2], [3]). Weakly coupled MDP have been used as a model for online advertising [2], multi-server data center control [3], robotics [4] etc. In this paper, we study such problems with Borel state and action spaces. Our approach builds on the occupation measure based approaches for single agent constrained MDPs with Borel spaces [1].

For single-agent/centralized systems, constrained Markov decision process (CMDP) is a popular model for sequential decision making problems with constraints. The goal of the agent is to minimize its long term expected cost while keeping the constraint functions below a threshold. One approach to solve such single agent problems is based on the idea of occupation measures. These are joint measures on the state and action spaces that can be used to quantify the time-averaged cost and constraint values [1], [5–7]. Using such measures, the strategy design problem can be written as a linear program whose solution gives the optimal occupation measure. Linear programming (LP) based formulation are presented in [5], [6] for CMDPs with finite/countable state and action spaces. The idea of LP was extended for CMDPs with Borel state and action spaces in [1], [7], [8].

Weakly coupled MDP with finite state and action spaces have been studied in the literature. A resource allocation problem for multiple task completion was modeled as a weakly coupled MDP in [9]. Each individual task was modeled as an MDP with instantaneous resource constraints on control strategy. [2] considered the problem of budget allocation across independent MDPs. Optimal value functions are derived as a function of the available budget and the allocation problem is posed as multi-item, multiple choice knapsack problem for which a greedy algorithm is presented to determine budget allocation. A distributed online learning based algorithm was proposed for weakly coupled MDPs in [3] where the system dynamics were assumed to be unknown.

In this paper, we consider the problem of weakly coupled constrained MDP with Borel state and action spaces. We use the linear programming based approach of [1] to derive an occupation measure based LP to find the optimal decentralized control strategies for our problem. Our main contributions could be summarized as follows:

1) We formulate a LP to show that randomized stationary strategies are optimal for each agent under some assumptions on the transition kernels, cost and the constraint functions.
2) We consider the special case of multi-agent Linear Quadratic Gaussian (LQG) systems and show that the optimal control strategy could be obtained by solving a semi-definite program (SDP).

Finally, we also present some numerical experiments for a toy problem on multi-agent LQG. The following is the outline of the paper: We will start with problem formulation in section II and present the LP to solve the general Borel case in section III. We consider the multi-agent LQG case in section IV, provide numerical results in section V and conclude in section VI.

Notation

Random variables are denoted by upper case letters and their realizations by corresponding small letters. $X_{a:b}$ denotes the collection $(X_a, X_{a+1}, \cdots, X_b)$. Boldface letter $\mathbf{X}$ is used to denote the collection (X^1, X^2). $\mathbb{E}[\cdot]$ is the expectation of a random variable. For a collection of functions $\mathbf{g}$ and a probability distribution f, we use $\mathbb{E}_f^{\mathbf{g}}[\cdot]$ to denote that the expectation depends on the choice of functions in $\mathbf{g}$ and the distribution f. For any Borel space $\mathcal{S}$, let $\mathbb{B}(\mathcal{S})$ denote the set of all Borel sets of $\mathcal{S}$. $A \succeq B$ means that

Mukul Gagrani and Ashutosh Nayyar are with the Department of Electrical Engineering at the University of Southern California, Los Angeles, CA, USA. Email: mgagrani@usc.edu;ashutosn@usc.edu. This work was supported by NSF Grant ECCS 1509812 and ECCS 1750041.

$(A - B)$ is positive semi-definite. $\mathcal{N}(m, \Sigma)$ denotes the Gaussian distribution with mean m and covariance Σ.

II. Problem formulation

Consider a two-agent dynamical system with state process $\mathbf{X}_t = (X_t^1, X_t^2), t \geq 0$. $X_t^i \in \mathcal{X}^i$ is the state-component associated with agent i for $i \in \{1, 2\}$. The distribution of the initial state X_0^i is denoted by ν^i. X_0^1, X_0^2 are independent and let ν denotes the pair (ν^1, ν^2). At time t, agent i takes a control action $U_t^i \in \mathcal{U}^i$ and the states of the two agents evolve in a decoupled manner according to the following stochastic kernel:

$$X_{t+1}^1 \sim Q^1(\cdot|X_t^1, U_t^1), \tag{1}$$

$$X_{t+1}^2 \sim Q^2(\cdot|X_t^2, U_t^2). \tag{2}$$

Information and Strategies

Each agent can observe its component of the state perfectly at each time. Agents do not share any information with each other. Hence, the information available to agent i at time t is $\mathcal{I}_t^i = \{X_{1:t}^i, U_{1:t-1}^i\}$. Agent i maps its information to its corresponding action using a randomized strategy π_t^i as follows,

$$U_t^i \sim \pi_t^i(\cdot|\mathcal{I}_t^i).$$

where $\pi_t^i(\cdot|\mathcal{I}_t^i)$ is a probability distribution on the control space $\mathcal{U}^i$ of agent i. We allow for randomized strategies because we are considering a control problem with constraints [5]. The collection $\pi^i = \{\pi_t^i\}_{t\geq 0}$ denotes the control strategy of agent i and the pair $\pi = (\pi^1, \pi^2)$ is referred to as the joint control strategy of the agents.

Cost and Constraints

Agent i incurs an instantaneous cost $c^i(X_t^i, U_t^i)$ at each time t. In addition, agent i also has an associated constraint cost function $d^i(X_t^i, U_t^i)$ at time t. The long-term average cost function and constraint function under a joint strategy pair π and initial distribution pair ν is defined as:

$$J(\pi, \nu) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\nu^\pi \left[\sum_{t=0}^{T-1} c^1(X_t^1, U_t^1) + c^2(X_t^2, U_t^2)\right] \tag{3}$$

$$K(\pi, \nu) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\nu^\pi \left[\sum_{t=0}^{T-1} d^1(X_t^1, U_t^1) + d^2(X_t^2, U_t^2)\right] \tag{4}$$

The objective of the agents is to jointly minimize their long term average cost (3) while keeping the joint long term average constraint function (4) below a threshold k. We formally state the problem below.

Problem 1. *Find a joint control strategy π and initial distribution ν for the agents which minimizes the cost $J(\pi, \nu)$ subject to the constraint $K(\pi, \nu) \leq k$, i.e.,*

$$\inf_{\pi,\nu} J(\pi, \nu)$$
$$\textit{subject to } K(\pi, \nu) \leq k \tag{5}$$

Assumption 1. *Problem 1 is feasible i.e. there exists a pair (π, ν) such that $K(\pi, \nu) \leq k$ and $J(\pi, \nu) < \infty$.*

Remark 1. *Constrained problems which consider long term average cost and a fixed initial distribution have been studied for single agent systems in [3], [5], [6], [10]. These problems are referred to as "ergodic" problems in the literature. Problems which require the joint design of initial distribution and control strategy (as in Problem 1) are referred to as "minimum pair" problems. Such constrained problems have been considered in [1], [11–14] for centralized (single-agent) systems.*

A. Discussion

Problem 1 is an instance of constrained team decision problem with additive cost and constraint function. In the absence of the constraint (5), this problem can be decomposed into two single agent (centralized) decision problems, the solution to which can be obtained using Markov decision theory [15]. However, constraint (5) couples the decision making of the two agents. This is because the choice of control strategy for agent 2 can affect the choice of control strategy for agent 1 since (5) has to be satisfied jointly by the two agents. This coupling makes this problem non-trivial. Such problems are also referred to as weakly coupled Markov decision problems ([2], [3]) since the coupling among the agents is only due to the constraint (5).

The framework we discuss in this paper can be used to model problems where the agents are working as a team to achieve a common goal encoded by the constraint (5) while trying to minimize their cumulative individual costs. We give a few examples which can be posed as an instance of Problem 1.

1) Resource constrained problems: Consider a problem where the agents are sharing resources (e.g. control resources, energy resources) with each other. The goal of the agents is to minimize their total costs with a constraint on the resource utilization. Problems of such type can be modeled using the framework of Problem 1 where $d^i(X_t^i, U_t^i)$ and $c^i(X_t^i, U_t^i)$ is the resource consumption and the cost respectively for agent i. For example, consider a smart building with two air conditioning systems which are maintaining temperatures of two different rooms while sharing a common power supply. The state X_t^i denotes the temperature of room i while U_t^i denotes the amount of power consumed by air conditioner i. Suppose the temperature of room i evolves as follows: $X_{t+1}^i = A^i X_t^i + B^i U_t^i + W_t^i$, where W_t^i is random noise. The objective of the agents is to minimize the deviation of the room temperature around a nominal value while keeping the total power consumed below a certain threshold i.e.

$$\min \ \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\nu^\pi \left[\sum_{t=0}^{T-1}\sum_{i=1}^{2} ||X_t^i - X_{nom}^i||^2\right]$$
$$\text{subject to } \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\nu^\pi \left[\sum_{t=0}^{T-1}\sum_{i=1}^{2} ||U_t^i||^2\right] \leq k$$

2) Remote estimation: Consider an estimation problem with multiple estimators. Estimator i wants to form an estimate $\hat{X}_t^i$ of a corresponding Markov source X_t^i at each time t. The sources are being observed by a shared sensor. In order to compute the estimate, estimator i can request the sensor to transmit X_t^i at time t using the decision variable $A_t^i \in \{0,1\}$. $A_t^i = 1$ indicates that estimator i has requested an observation. Due to limited power supply, the sensor can handle a limited number of observation requests on average. The objective of the agents is to minimize the cumulative estimation error such that the average cumulative number of requested observations is below a certain threshold. This problem can be easily modeled using the framework of Problem 1 with X_t^i as the state and $(\hat{X}_t^i, A_t^i)$ as the action of agent i.

3) Mean-field constraint: Consider a two-agent problem where the state space $\mathcal{X}^i = \{0,1\}$. Each agent has a control cost given by $c^i(U_t^i)$. The agents goal is to minimize the time-averaged cost while keeping the time-averaged fraction of agents in state 1 below a threshold, i.e.,

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1}{2}\sum_{i=1}^{2}\mathbb{I}(X_t^i = 1)\right] \leq k$$

III. Optimal Strategies

We are going to restrict our attention to the case when the state space and the control space $\mathcal{X}^i, \mathcal{U}^i$ are Borel spaces [1] (e.g. Euclidean space). Single agent constrained Markov decision process in Borel spaces can be solved using infinite dimensional linear programming approach [1]. In this approach an optimal occupation measure (joint probability measure) of the state and control is computed using a linear program. The optimal pair of control strategy and an initial distribution is obtained using the optimal occupation measure.

Building on the single-agent solution, we will provide an infinite dimensional LP which will characterize the solution to our *multi-agent* problem described in Problem 1. To do so, we will need the following definitions.

Definition 1.

1) *Let* $w^i(x^i,u^i) := 1 + c^i(x^i,u^i)$ *and* $\hat{w}^i(x^i) := \inf_{u^i\in\mathcal{U}^i} w^i(x^i,u^i)$.
2) *Let* $\mathcal{F}^i(\mathcal{X}^i \times \mathcal{U}^i)$ *be the vector space of measurable functions from* $\mathcal{X}^i \times \mathcal{U}^i$ *to* $\mathbb{R}$ *with finite* w^i *norm. That is,* $f^i \in \mathcal{F}^i(\mathcal{X}^i \times \mathcal{U}^i)$ *if*

$$\sup_{(x^i,u^i)\in\mathcal{X}^i\times\mathcal{U}^i} \frac{|f^i(x^i,u^i)|}{w^i(x^i,u^i)} < \infty$$

.
3) *Let* $\mathcal{M}_+^i(\mathcal{X}^i \times \mathcal{U}^i)$ *be the vector space of positive measures on* $\mathcal{X}^i \times \mathcal{U}^i$ *with finite* w^i *variations. That is,* $\mu^i \in \mathcal{M}_+^i(\mathcal{X}^i \times \mathcal{U}^i)$ *if*

$$\int_{\mathcal{X}^i\times\mathcal{U}^i} w^i(x^i,u^i)\mu^i\left(dx^i,du^i\right) < \infty.$$

[1]A Borel space is a Borel subset of complete and separable metric space

4) *Define the bilinear form* $\langle f^i,\mu^i\rangle$ *for* $f^i \in \mathcal{F}^i(\mathcal{X}^i\times\mathcal{U}^i)$, $\mu^i \in \mathcal{M}_+^i(\mathcal{X}^i\times\mathcal{U}^i)$ *as follows:*

$$\langle f^i,\mu^i\rangle := \int_{\mathcal{X}^i\times\mathcal{U}^i} f^i(x^i,u^i)\mu^i\left(dx^i,du^i\right)$$

Let $\mu^i \in \mathcal{M}_+^i(\mathcal{X}^i \times \mathcal{U}^i)$ be a probability measure on the joint state and action space. Note that any distribution $\mu^i \in \mathcal{M}_+^i(\mathcal{X}^i \times \mathcal{U}^i)$ can be decomposed in terms of its marginal on $\mathcal{X}^i$ and a conditional distribution over the control space $\phi^i(\cdot|x^i)$ such that

$$\mu^i(B^i,C^i) = \int_{B^i} \phi^i(C^i|x^i)\hat{\mu}^i(dx^i), \forall B^i \in \mathbb{B}(\mathcal{X}^i), C^i \in \mathbb{B}(\mathcal{U}^i) \tag{6}$$

where $\mu^i(B^i,C^i)$ denotes the measure of the rectangle $B^i\times C^i$ and $\hat{\mu}^i(B^i) := \mu^i(B^i,\mathcal{U}^i)$ for all $B^i \in \mathbb{B}(\mathcal{X}^i)$ is the marginal of μ^i on $\mathcal{X}^i$. We will write the measure $\mu^i = \hat{\mu}^i\cdot\phi^i$ when the corresponding decomposition is as in (6).

We can now describe the linear program that characterizes the optimal control strategies. Let $\mu^1 \in \mathcal{M}_+^1(\mathcal{X}^1\times\mathcal{U}^1)$ and $\mu^2 \in \mathcal{M}_+^2(\mathcal{X}^2\times\mathcal{U}^2)$. Consider the following linear program:

$$\text{LP-1: } \min_{\mu^1,\mu^2} \langle\mu^1,c^1\rangle + \langle\mu^2,c^2\rangle$$

$$\text{subject to: } \langle\mu^1,d^1\rangle + \langle\mu^2,d^2\rangle \leq k \tag{7}$$

$$\mu^i(B,\mathcal{U}^i) = \int_{\mathcal{X}^i\times\mathcal{U}^i} Q^i(B|x^i,u^i)\mu^i\left(dx^i,du^i\right), \quad \forall i, B \in \mathbb{B}(\mathcal{X}^i) \tag{8}$$

$$\mu^i(\mathcal{X}^i,\mathcal{U}^i) = 1, \mu^i \in \mathcal{M}_+^i(\mathcal{X}^i\times\mathcal{U}^i) \tag{9}$$

LP-1 is an infinite dimensional linear program whose solution consists of a probability measure on the state and action space for each agent. Theorem 1 characterizes the solution to Problem 1 in terms of the solution to the LP-1 under the following assumption

Assumption 2.

1) $c^i(x^i,u^i)$ *is non-negative and inf-compact,* $d^i(x^i,u^i)$ *is non-negative and lower semi continuous* $\forall i$.
2) *The transition kernel* Q^i *is weakly continuous* $\forall i$.
3) $d^i \in \mathcal{F}^i(\mathcal{X}^i\times\mathcal{U}^i), \quad \forall i$.
4) $\int_{\mathcal{X}^i} \hat{w}^i(y^i)Q^i(dy^i|\cdot) \in \mathcal{F}^i(\mathcal{X}^i\times\mathcal{U}^i), \quad \forall i$

Assumption 2 ensures that there exists a solution to LP-1. Similar assumption has been made in the analysis of single agent constrained MDP (see [1]). We are now ready to state our main result.

Theorem 1. *Under Assumption 1 and 2 there exists* μ_*^1, μ_*^2 *that achieve the optimal value of LP-1. Let* $\mu_*^i = \hat{\mu}_*^i\cdot\phi_*^i, i \in \{1,2\}$ *be the decomposition of* μ_i^* *into the marginal and conditional distribution as in* (6)*. Then, an optimal control strategy for agent* i *in Problem 1 is a randomized stationary strategy* $\phi_*^i(\cdot|x^i)$ *and the optimal initial distribution is* $\hat{\mu}_*^i(\cdot)$*. Moreover, the optimal cost achieved under* (ϕ_*^1,ϕ_*^2) *when the initial state distribution is* $(\hat{\mu}_*^1,\hat{\mu}_*^2)$ *is* $\langle\mu_*^1,c^1\rangle + \langle\mu_*^2,c^2\rangle$.

Proof outline. The proof follows by considering a centralized problem where a single agent knows the entire state and

action history and takes both actions. The optimal cost of the centralized problem serves as a lower bound for Problem 1. We then establish that this lower bound is achieved under the control strategy and initial distribution described in Theorem 1. Note that Theorem 1 implies that each agent's optimal strategy is a stationary Markov strategy since the distribution of U_t^i depends only on X_t^i. □

Remark 2. *The results obtained in this section hold true when the state space $\mathcal{X}^i$ and action space $\mathcal{U}^i$ are finite. In this case, the infinite dimensional linear program LP-1 simplifies to the following finite dimensional linear program:*

$$\min_{\mu^1,\mu^2} \sum_{i=1}^{2} \sum_{x^i,u^i} \mu^i(x^i,u^i)c^i(x^i,u^i)$$

$$\text{subject to } \sum_{i=1}^{2} \sum_{x^i,u^i} \mu^i(x^i,u^i)d^i(x^i,u^i) \leq k$$

$$\sum_{u^i} \mu^i(x^i,u^i) = \sum_{y^i,u^i} Q^i(x^i|y^i,u^i)\mu^i(y^i,u^i) \quad \forall i, x^i$$

$$\sum_{x^i,u^i} \mu^i(x^i,u^i) = 1 \text{ and } \mu^i(x^i,u^i) \geq 0 \quad \forall i, x^i, u^i$$

Using Theorem 1, the optimal control strategy is the conditional distribution of the action obtained from μ_*^i as follows:

$$\phi_*^i(u^i|x^i) := \frac{\mu_*^i(x^i,u^i)}{\sum_{\tilde{u}^i} \mu_*^i(x^i,\tilde{u}^i)} \tag{10}$$

In the finite case, it can be established that the optimal cost is independent of the initial state distribution. Moreover, the optimal control strategy is given by (10) *for any initial state distribution. Similar observations were made in [3].*

Remark 3. *Consider the case when the system has $N > 2$ agents under the same assumptions. In addition, the agents have to satisfy multiple constraints of the form in* (5)*. The results obtained in this section can be easily generalized to handle this case. We can write down the LP-1 in which each agent has an associated measure μ^i and add a constraint in the linear program of the form in* (7) *corresponding to each joint constraint of the form in* (5)*.*

Theorem 1 applies to arbitrary dynamics, cost and constraint functions as described in (1)-(4). When the dynamics and cost are specialized, the infinite dimensional linear program may be reducible to more tractable optimization problems. We demonstrate this for the linear quadratic systems in the next section.

IV. Constrained Linear Quadratic Systems

In this section, we consider an instance of Problem 1 when the system dynamics are linear, cost and constraint function have a quadratic form and the disturbances are Gaussian. We refer to such systems as the constrained Linear Quadratic Gaussian (LQG) multi-agent systems.

The state $X_t^i \in \mathbb{R}^{n_i}$ of agent i evolves according to the following linear dynamics:

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

where $W_t^i \sim \mathcal{N}(0, I)$, $U_t^i \in \mathbb{R}^{m_i}$ and A^i, B^i are matrices of appropriate dimensions. The instantaneous cost and constraint function are given as follows:

$$c^i(X_t^i, U_t^i) = (X_t^i)' Q^i X_t^i + (U_t^i)' R^i U_t^i, \tag{12}$$

$$d^i(X_t^i, U_t^i) = (X_t^i)' M^i X_t^i + (U_t^i)' N^i U_t^i, \tag{13}$$

where Q^i, M^i, R^i, N^i are symmetric positive definite matrices for $i \in \{1,2\}$. This problem can be seen as a special case of Problem 1 where the state and action spaces are Borel spaces since $\mathcal{X}^i = \mathbb{R}^{n_i}, \mathcal{U}^i = \mathbb{R}^{m_i}$. It can be verified easily that Assumption 2 holds true for this problem and hence we can obtain the optimal control strategy by solving the LP-1 and using Theorem 1. For that purpose, we define the following moments associated with a measure μ^i on $\mathcal{X}^i \times \mathcal{U}^i$,

$$m_x^i = \int_{\mathbb{R}^{n_i}} x\mu^i(dx, \mathcal{U}^i), \Sigma_{xx}^i = \int_{\mathbb{R}^{n_i}} xx'\mu^i(dx, \mathcal{U}^i)$$

$$m_u^i = \int_{\mathbb{R}^{m_i}} u\mu^i(\mathcal{X}^i, du), \Sigma_{uu}^i = \int_{\mathbb{R}^{m_i}} uu'\mu^i(\mathcal{X}^i, du)$$

$$\Sigma_{xu}^i = \int_{\mathbb{R}^{n_i} \times \mathbb{R}^{m_i}} xu'\mu^i(dx, du)$$

The next theorem shows that in the case of LQG systems the infinite dimensional linear program (LP-1) can be reduced to a finite dimensional semi-definite program (SDP).

Theorem 2. *Consider the following SDP:*

$$\textit{LQG-SDP} : \min_{\Sigma_{xx}^i, \Sigma_{uu}^i, \Sigma_{xu}^i} \sum_{i=1}^{2} Tr(Q^i \Sigma_{xx}^i) + Tr(R^i \Sigma_{uu}^i)$$

$$\textit{subject to} : \sum_{i=1}^{2} Tr(M^i \Sigma_{xx}^i) + Tr(N^i \Sigma_{uu}^i) \leq k \tag{14}$$

$$\Sigma_{xx}^i = A^i \Sigma_{xx}^i (A^i)' + A^i \Sigma_{xu}^i (B^i)' + B^i \Sigma_{ux}^i (A^i)' + B^i \Sigma_{uu}^i (B^i)' + I \tag{15}$$

$$\begin{bmatrix} \Sigma_{xx}^i & \Sigma_{xu}^i \\ (\Sigma_{xu}^i)' & \Sigma_{uu}^i \end{bmatrix} \succeq 0 \tag{16}$$

Suppose $\Sigma_{xx}^{i,*}, \Sigma_{uu}^{i,*}, \Sigma_{xu}^{i,*}$ is the solution of the LQG-SDP. Then, the Gaussian measure on $\mathcal{X}^i \times \mathcal{U}^i$ with mean 0 and second moments $\Sigma_{xx}^{i,*}, \Sigma_{uu}^{i,*}, \Sigma_{xu}^{i,*}$ is optimal for LP-1. Moreover, the optimal control strategy for agent i is a Gaussian stationary randomized policy given as:

$$\phi_*^i(U^i|X^i) \sim \mathcal{N}\left(m_{u|x}^i, \Sigma_{u|x}^i\right) \tag{17}$$

where $m_{u|x}^i = \Sigma_{ux}^{i,*}(\Sigma_{xx}^{i,*})^{-1} X^i$ and $\Sigma_{u|x}^i = \Sigma_{uu}^{i,*} - \Sigma_{ux}^{i,*}(\Sigma_{xx}^{i,*})^{-1}\Sigma_{xu}^{i,*}$ for $i \in \{1,2\}$. Also, the corresponding optimal initial distribution ν_*^i for agent i is $\mathcal{N}(0, \Sigma_{xx}^{i,*})$.

Proof. We will first show that it is sufficient to consider Gaussian measures for LP-1. Consider a measure μ^i on $\mathcal{X}^i \times \mathcal{U}^i$ with means m_x^i, m_u^i and second moment matrix $\begin{bmatrix} \Sigma_{xx}^i & \Sigma_{xu}^i \\ (\Sigma_{xu}^i)' & \Sigma_{uu}^i \end{bmatrix}$. Now, observe that,

$$\langle \mu^i, c^i \rangle = Tr(Q^i \Sigma_{xx}^i) + Tr(R^i \Sigma_{uu}^i),$$

$$\langle \mu^i, d^i \rangle = Tr(M^i \Sigma_{xx}^i) + Tr(N^i \Sigma_{uu}^i).$$

Suppose $(X^i, U^i) \sim \mu^i$ and let $\tilde{X}^i = A^i X^i + B^i U^i + W^i$ be the next state. Then, (8) encodes the constraint that distribution of X^i and $\tilde{X}^i$ should be the same. Let μ^i be a feasible measure for LP-1 which satisfies (8). This means that the first and the second moment of X^i and $\tilde{X}^i$ should match when $(X^i, U^i) \sim \mu^i$, i.e.,

$$m_x^i = A^i m_x^i + B^i m_u^i \tag{18}$$

$$\Sigma_{xx}^i = A^i \Sigma_{xx}^i (A^i)' + A^i \Sigma_{xu}^i (B^i)' + B^i \Sigma_{ux}^i (A^i)' + B^i \Sigma_{uu}^i (B^i)' + I \tag{19}$$

Now, consider a Gaussian measure μ_g^i which has the same 1^{st} and 2^{nd} moments as μ^i. If $(X^i, U^i) \sim \mu_g^i$ are jointly Gaussian then $\tilde{X}^i = A^i X^i + B^i U^i + W^i$ is also Gaussian with mean and covariance given by the right hand side of (18) and (19) above. Thus, $X^i, \tilde{X}^i$ are both Gaussian with same mean and covariance since (18),(19) holds true for the moments of μ_g^i. Hence, μ_g^i satisfies (8). Also, $\langle \mu_g^i, c^i \rangle = \langle \mu^i, c^i \rangle$ and $\langle \mu_g^i, d^i \rangle = \langle \mu^i, d^i \rangle$ as μ^i, μ_g^i have the same second moments. Thus, for any feasible μ^i there exists a feasible Gaussian measure μ_g^i which achieves the same value of the linear program. Hence, it is sufficient to consider the class of Gaussian measures in LP-1. Since a Gaussian measure can be characterized only by the first and the second moments, we can reduce LP-1 to the SDP presented in the lemma by setting $m_x^i = m_u^i = 0$ without loss of generality.

Finally, using theorem 1 and the fact that optimal μ_*^i is Gaussian, it can be easily shown using that the optimal control strategy is Gaussian with mean $m_{u|x}^i, \Sigma_{u|x}^i$ as defined in the lemma. □

Based on the optimal randomized strategy in Theorem 2 (see (17)), one can write the optimal action of agent i as follows:

$$U_t^{i,*} = K_*^i X_t^i + V_t^i,$$

where $K_*^i := \Sigma_{ux}^{i,*} (\Sigma_{xx}^{i,*})^{-1}$ and $V_t^i \sim \mathcal{N}(0, \Sigma_{u|x}^i)$. Note that agent i is using its local state in a linear fashion.

As noted earlier, in the absence of the constraint (5), Problem 1 would decompose into two single agent unconstrained LQG control problem. This would imply that the optimal unconstrained controller for each agent is also a linear function of its local state. However, the gain matrix in the unconstrained problem may be different from that obtained in Theorem 2. Also, the optimal constrained controller obtained via Theorem 2 has a noise term V_t^i in contrast with the deterministic linear controller in the unconstrained case. In the next lemma, we show that the agents can in fact ignore the control noise and use a deterministic linear control strategy.

Lemma 1. *Let g_*^i be the following deterministic stationary linear controller:*

$$g_*^i(X_t^i) := \Sigma_{ux}^{i,*} (\Sigma_{xx}^{i,*})^{-1} X_t^i. \tag{20}$$

where $\Sigma_{ux}^{i,*}, \Sigma_{xx}^{i,*}$ are obtained from the SDP in Theorem 2. Then, g_*^i is an optimal control strategy for agent i.

Proof Outline. It can be shown, using an induction argument, that the expected instantaneous cost and constraint under the optimal policy $\phi_* = (\phi_*^1, \phi_*^2)$ from theorem 2 is lower bounded by the expected instantaneous cost and constraint under $g_* = (g_*^1, g_*^2)$ when the initial distribution is $\hat{\mu}_*$. That is, $\mathbb{E}_{\hat{\mu}_*}^{\phi_*}[c(\mathbf{X}_t, \mathbf{U}_t)] \geq \mathbb{E}_{\hat{\mu}_*}^{g_*}[c(\mathbf{X}_t, \mathbf{U}_t)]$ and $\mathbb{E}_{\hat{\mu}_*}^{\phi_*}[d(\mathbf{X}_t, \mathbf{U}_t)] \geq \mathbb{E}_{\hat{\mu}_*}^{g_*}[d(\mathbf{X}_t, \mathbf{U}_t)]$ for all time t. Therefore, the average cost and constraint function achieved under the pair $(g_*, \hat{\mu}_*)$ is not more than the average cost and constraint function achieved under the pair $(\phi_*, \hat{\mu}_*)$. Hence, g_*^i is also an optimal control strategy for agent i. □

So far we assumed that the noise in the system dynamics W_t^i was Gaussian. The following extends our results to non-Gaussian noise.

Lemma 2. *Suppose the system dynamics are as in* (11) *and the noise W_t^i is non-Gaussian with mean* 0 *and covariance matrix I. The results of Theorem 2 and Lemma 1 hold true for this case.*

Proof Outline. It can be shown that LQG-SDP is a relaxation of LP-1 when the system dynamics are as in (11) with non-Gaussian noise and the cost, constraint function have the quadratic form in (12),(13). Therefore, the optimal value of this SDP is a lower bound for the optimal value of LP-1. We can further show that this lower bound is achieved if the agents follow the control strategy g_*^i as described in Lemma 1. Therefore, g_*^i is optimal in the non-Gaussian case as well. □

V. Numerical Experiments

In this section, we will present numerical experiments for a multi-agent LQ problem with constraints. Consider a two agent system where $X_t^i \in \mathbb{R}^2$ and $U_t^i \in \mathbb{R}^2$ for $i = 1, 2$. The system dynamics is characterized by the following matrices:

$$A^1 = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}, B^1 = \begin{bmatrix} 3 & 1 \\ 2 & -1 \end{bmatrix},$$
$$A^2 = A^1, B^2 = B^1.$$

The cost matrices are given as follows:

$$Q^1 = \begin{bmatrix} 4 & 2.8 \\ 2.8 & 2 \end{bmatrix}, R^1 = \begin{bmatrix} 14.5 & 3.4 \\ 3.4 & 0.8 \end{bmatrix},$$
$$Q^2 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}, R^2 = \begin{bmatrix} 1.3 & 1.2 \\ 1.2 & 1.2 \end{bmatrix}.$$

The constraint matrices are set to the following:

$$M^1 = \begin{bmatrix} 1.1 & 0.9 \\ 0.9 & 0.75 \end{bmatrix}, N^1 = \begin{bmatrix} 0.1 & 0.3 \\ 0.3 & 1.1 \end{bmatrix},$$
$$M^2 = \begin{bmatrix} 0.35 & 1.2 \\ 1.2 & 4.4 \end{bmatrix}, N^2 = \begin{bmatrix} 0.15 & 0.15 \\ 0.15 & 0.18 \end{bmatrix}.$$

Let $J(t) := \frac{1}{t} \sum_{s=0}^{t-1} \sum_{i=1}^{2} (X_s^i)' Q^i X_s^i + (U_s^i)' R^i U_s^i$ be the running average cost and similarly $K(t)$ be the running average constraint function. We will compare these running averages under the optimal constrained controller obtained from the SDP in Theorem 2 with the optimal unconstrained

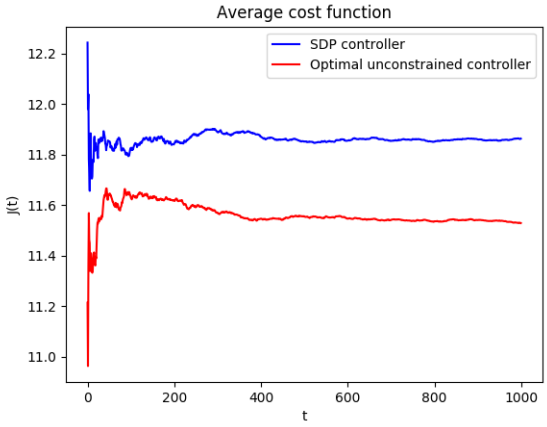

Fig. 1: Trajectory of the running average cost

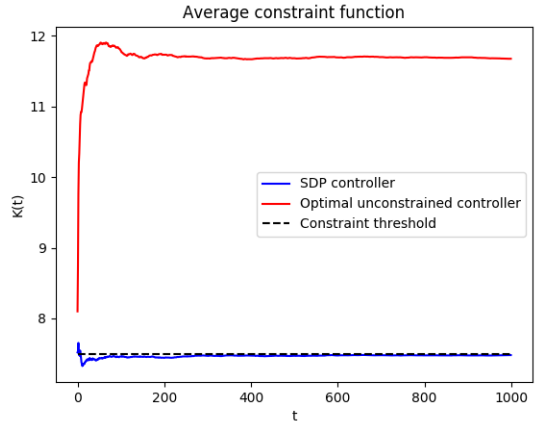

Fig. 2: Trajectory of the running average constraint function

controllers for each agent. The optimal unconstrained controllers can be obtained by solving the discrete Riccati equation for each agent [16].

Figure 1 shows the average cost $J(t)$ as a function of time for the optimal constrained controller (referred to as SDP controller in the figure) and the optimal unconstrained controller. It can be seen that the optimal constrained controller performs worse than the optimal unconstrained controller in terms of the achieved average cost. Figure 2 shows the average constraint $K(t)$ as a function of time for the optimal constrained and unconstrained controller when the constraint threshold is set to $k = 7.6$. It can be observed that the controller obtained via the SDP is able to satisfy the constraint threshold while the unconstrained controller could not. Thus, the optimal constrained controller is able to meet the constraint at the expense of higher cost compared to the optimal unconstrained controller.

VI. Conclusion

We considered the problem of weakly coupled constrained MDP with Borel state and action spaces. We showed that randomized stationary policies are optimal for each agent under some assumptions on the transition kernels, cost and the constraint functions. Our approach was to consider a centralized problem where a single agent knows the entire state and action history and takes both the actions. We solve the centralized problem using the occupation measure based LP of [1] and established that the obtained solution is optimal for our original problem. Further, we considered the case of multi-agent LQG and showed that the infinite dimensional LP can be simplified to a SDP for obtaining the optimal control strategy. Finally, we illustrated our results through some numerical experiments.

References

[1] O. Hernández-Lerma, J. González-Hernández, and R. R. López-Martínez, "Constrained average cost markov control processes in borel spaces," *SIAM Journal on Control and Optimization*, vol. 42, no. 2, pp. 442–468, 2003.

[2] C. Boutilier and T. Lu, "Budget allocation using weakly coupled, constrained markov decision processes," 2016.

[3] X. Wei, H. Yu, and M. J. Neely, "Online learning in weakly coupled markov decision processes: A convergence time study," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 1, p. 12, 2018.

[4] D. A. Dolgov and E. H. Durfee, "Optimal resource allocation and policy formulation in loosely-coupled markov decision processes." in *ICAPS*, 2004, pp. 315–324.

[5] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[6] A. Piunovskiy, *Optimal control of random sequences in problems with constraints*. Springer Science & Business Media, 2012, vol. 410.

[7] O. Hernández-Lerma and J. González-Hernández, "Constrained markov control processes in borel spaces: the discounted case," *Mathematical Methods of Operations Research*, vol. 52, no. 2, pp. 271–285, 2000.

[8] M. Kamgarpour and T. Summers, "On infinite dimensional linear programming approach to stochastic control," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6148–6153, 2017.

[9] N. Meuleau, M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. L. Dean, and C. Boutilier, "Solving very large weakly coupled markov decision processes," in *AAAI/IAAI*, 1998, pp. 165–172.

[10] L. I. Sennott, "Constrained average cost markov decision chains," *Probability in the Engineering and Informational Sciences*, vol. 7, no. 1, pp. 69–83, 1993.

[11] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*. Springer Science & Business Media, 2012, vol. 30.

[12] M. Kurano, "The existence of a minimum pair of state and policy for markov decision processes under the hypothesis of doeblin," *SIAM journal on control and optimization*, vol. 27, no. 2, pp. 296–307, 1989.

[13] M. Kurano, J.-I. Nakagami, and Y. Huang, "Constrained markov decision processes with compact state and action spaces: the average case," *Optimization*, vol. 48, no. 2, pp. 255–269, 2000.

[14] O. Hernandez-Lerma and J. Gonzalez-Hernandez, "Infinite linear programming and multichain markov control processes in uncountable spaces," *SIAM journal on control and optimization*, vol. 36, no. 1, pp. 313–335, 1998.

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

[16] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, Belmont, MA, 2012, vol. 2.