Automatica 137 (2022) 110060

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Model-free LQR design by Q-function learning☆

Milad Farjadnasab[1], Maryam Babazadeh *,[1]

Sharif University of Technology, Iran

ARTICLE INFO

Article history:
Received 13 April 2021
Received in revised form 10 July 2021
Accepted 12 October 2021
Available online 21 December 2021

Keywords:
Convex optimization
Semi-definite programming (SDP)
Linear quadratic regulation (LQR)
Q-learning
Distributed control

ABSTRACT

Reinforcement learning methods such as Q-learning have shown promising results in the model-free design of linear quadratic regulator (LQR) controllers for linear time-invariant (LTI) systems. However, challenges such as sample-efficiency, sensitivity to hyper-parameters, and compatibility with classical control paradigms limit the integration of such algorithms in critical control applications. This paper aims to take some steps towards bridging the well-known classical control requirements and learning algorithms by using optimization frameworks and properties of conic constraints. Accordingly, a new off-policy model-free approach is proposed for learning the Q-function and designing the discrete-time LQR controller. The design procedure is based on non-iterative semi-definite programs (SDP) with linear matrix inequality (LMI) constraints. It is sample-efficient, inherently robust to model uncertainties, and does not require an initial stabilizing controller. The proposed model-free approach is extended to distributed control of interconnected systems, as well. The performance of the presented design is evaluated on several stable and unstable synthetic systems. The data-driven control scheme is also implemented on the IEEE 39-bus New England power grid. The results confirm optimality, sample-efficiency, and satisfactory performance of the proposed approach in centralized and distributed design.

1. Introduction

The theory of optimal control and reinforcement learning are historically intertwined (Sutton & Barto, 1998). Dynamic programming is recognized as a well-developed framework in optimal control that relies on solving the Hamilton–Jacobi–Bellman (HJB) equations (Bellman, 1957). Efficient approximation techniques have been proposed to solve the underlying problem, known as approximate dynamic programming (ADP) or reinforcement learning. They include actor-critic methods, policy gradient methods, and Q-learning (Buşoniu, de Bruin, Tolić, Kober, & Palunko, 2018).

For LTI systems, the HJB equation is reduced to an algebraic Riccati equation (ARE). Since ARE requires full knowledge of the system dynamics, model-free iterative algorithms such as policy iteration (PI) and value iteration (VI) have been developed for the canonical linear quadratic regulation (Lewis & Vrabie, 2009). Alternatively, optimal control problems such as LQR problem have been extensively studied in the context of convex optimization (Boyd, El Ghaoui, Feron, & Balakrishnan, 1994; Rami & Zhou, 2000) for known dynamics. To deal with the unknown dynamics, a straightforward approach is to fit a model of the system by using observations and treat the resulting model as the true model for control purposes. This procedure relies on the certainty equivalence principle which can be challenging due to the propagation of small modeling errors on long time horizons. In recent years, LQR control is considered as a standard benchmark for learning-based control of systems with unknown dynamics (Jha, Roy, & Bhasin, 2018; Tu & Recht, 2018; Umenberger & Schön, 2018). In Jha, Roy, and Bhasin (2017) a data-driven adaptive LQR is proposed for unknown dynamics. It utilizes past and current data together with standard gradient descent update laws and semi-global uniformly ultimately bounded (UUB) stability of the closed-loop system is established. In da Silva, Bazanella, Lorenzini, and Campestrini (2019) a data-driven method for computing the state feedback gains in discrete-time LQR control problem is proposed which converges to the optimal solution as the number of Markov parameters tends to infinity. Data-driven stabilizing policies by using LMIs are discussed in De Persis and Tesi (2019), where LMIs have shown their significant effect in the robustness of the closed-loop systems to different sources of uncertainty.

Recently, Lee and Hu (2019) has developed a Q-learning framework for LQR control based on an alternative optimization formulation of the problem. The proposed framework is then used

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Tongwen Chen under the direction of Editor Ian R. Petersen.

* Corresponding author.
E-mail addresses: mfarjadnasab@gmail.com (M. Farjadnasab), babazadeh@sharif.edu (M. Babazadeh).

[1] The authors are with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.

https://doi.org/10.1016/j.automatica.2021.110060
0005-1098/

to design a model-free Q-learning algorithm based on primal–dual updates. The algorithm in Lee and Hu (2019) assumes that an initial stabilizing controller is available and requires a collection of the state trajectories for a certain set of linearly independent initial conditions which can be prohibitive in online applications.

Although the results in model-free reinforcement learning and adaptive dynamic programming are very promising, their implementation in real-life applications has been subject to criticism, as the methods require too much data for a satisfactory level of performance. Moreover, some of them are too sensitive to the choice of hyper-parameters, and in many cases, the algorithms appear to be too fragile for incorporation in critical control systems (Mania, Guy, & Recht, 2018).

On the other hand, emerging complex interconnected systems have introduced the requirement of distributed control of the systems without direct and complete knowledge of all subsystems. However, the extension of the model-free LQR design to the distributed case is not straightforward from two main perspectives. Firstly, the distributed LQR problem with arbitrary control structure and known dynamics yields a non-deterministic polynomial-time (NP)-hard optimization problem with only sub-optimal solutions, provided by iterative convex and non-convex algorithms (Babazadeh, 2021; Babazadeh & Nobakhti, 2017; Lin, Fardad, & Jovanović, 2013). Secondly, by performing the design procedure in a completely distributed manner, the learning procedure relies on partial observations. Most of the existing results in this route are devoted to special cases such as the synchronization of subsystems whose dynamics are unknown (Azzollini, Yu, Yuan, & Baldi, 2020; Yu, DeLellis, Chen, Di Bernardo, & Kurths, 2012), or dynamically decoupled subsystems (Alemzadeh & Mesbahi, 2019). A distributed Q-learning approach is developed in Görges (2019), where the system graph and control graph are assumed to be the same. Although the closed-loop stability or convergence of the solution is not addressed in Görges (2019), it provides promising insights into the model-free distributed control of interconnected systems by learning appropriate Q functions.

This paper aims to contribute to this growing area of research by offering a new approach to find the Q-function in the Q-learning framework for the standard LQR problem. The proposed approach benefits from the following noteworthy features;

i It does not require an initial stabilizing policy,
ii Evaluation of the Q function is carried out in a single step, i.e., it is non-iterative in nature.
iii In this off-policy model-free algorithm, the Q-function parameters are computed by solving an SDP that only needs the system's state and input trajectories for a small number of samples.
iv It does not require hyper-parameter tuning.
v The proposed approach is versatile and extendable to model-free distributed control of interconnected systems with or without a central design entity.

Notation: The following notations are adopted in this paper: $\mathbb{N}$ and $\mathbb{N}_+$ represent the sets of non-negative and positive integers, respectively. $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ denote the sets of real numbers, non-negative real numbers and positive real numbers, respectively. $\mathbb{R}^n$ is the n-dimensional Euclidean space. $\mathbb{R}^{n\times m}$ is the set of all $n \times m$ real matrices. $\mathbb{S}^n$, $\mathbb{S}^n_+$, and $\mathbb{S}^n_{++}$ represent symmetric $n \times n$ matrices, cone of symmetric $n \times n$ positive semi-definite matrices, and symmetric $n \times n$ positive definite matrices, respectively. I_n is the $n \times n$ identity matrix. $A \succ 0, A \prec 0, A \succeq 0$, and $A \preceq 0$ describe symmetric positive definite, symmetric negative definite, symmetric positive semi-definite, and symmetric negative semi-definite matrices, respectively. $\mathcal{R}(A)$ and $\mathrm{eig}(A)$ denote the range-space and the set of eigenvalues of the matrix A, respectively.

2. Centralized problem setup

Consider the standard LQR problem for the discrete-time LTI system,

$$x(k+1) = Ax(k) + Bu(k), \quad x(0) = z \in \mathbb{R}^n, \ k \in \mathbb{N} \tag{1}$$

where $x(k)^T = [x_1(k)\ x_2(k)\ \cdots\ x_n(k)] \in \mathbb{R}^n$ is the state vector, $u(k)^T = [u_1(k)\ u_2(k)\ \cdots\ u_n(k)] \in \mathbb{R}^m$ is the input vector, and $z \in \mathbb{R}^n$ represents the initial state. Consider the linear state feedback control policy,

$$u(k) = Kx(k), \tag{2}$$

with $K \in \mathbb{R}^{m\times n}$. Following the notation of Lee and Hu (2019), the state trajectories of the closed-loop system with state feedback gain K, starting from $x(0) = z$ are denoted by $x(k; K, z)$. Consequently, the cost function for the standard LQR problem can be expressed as,

$$J(K, z) := \sum_{k=0}^{\infty} \begin{bmatrix} x(k;K,z) \\ Kx(k;K,z) \end{bmatrix}^T \Lambda \begin{bmatrix} x(k;K,z) \\ Kx(k;K,z) \end{bmatrix}, \tag{3}$$

where $\Lambda := \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \in \mathbb{S}^{(n+m)}_+$ is a block-diagonal matrix including the state and input weighting matrices $Q \in \mathbb{S}^n_+$ and $R \in \mathbb{S}^m_{++}$, respectively. The infinite horizon LQR problem can therefore be formulated as,

$$\min_{K\in\mathcal{K}} \sum_{i=1}^{r} J(K, z_i), \tag{4}$$

where $\mathcal{K}$ is the set of stabilizing state feedback gains for the pair (A, B), and $z_i \in \mathbb{R}^n$, $i \in \{1, 2, \ldots, r\}$ are chosen such that $\sum_{i=1}^r z_i z_i^T = Z \succ 0$ (Lee & Hu, 2019).

The cost function at time k for the standard LQR problem can be written as,

$$V\left(x(k)\right) = \sum_{i=k}^{\infty} r\left(x(i), u(i)\right), \tag{5}$$

with the stage cost,

$$r\left(x(i), u(i)\right) = x^T(i)Qx(i) + u^T(i)Ru(i). \tag{6}$$

It can be shown that the optimal value function is quadratic in the form of,

$$V^*\left(x(k)\right) = x^T(k)P^*x(k), \tag{7}$$

with the optimal policy,

$$u^*(k) = K^*x(k), \tag{8}$$

where $P^* \in \mathbb{S}^n_+$ and $K^* \in \mathbb{R}^{m\times n}$. It can also be shown that P^* is the solution to the discrete-time algebraic Riccati equation (DARE)

$$P = Q + A^TPA - \left(A^TPB\right)\left(R + B^TPB\right)^{-1}\left(B^TPA\right), \tag{9}$$

which can be solved by using the knowledge of the system dynamics A and B (Lewis & Vrabie, 2009).

Q-learning offers a model-free solution for solving the LQR problem. The Q-function is defined as (Bradtke, Ydstie, & Barto, 1994),

$$Q_K\left(x(k), u(k)\right) = r\left(x(k), u(k)\right) + V_K\left(x(k+1)\right), \tag{10}$$

which represents the value of taking action $u(k)$ from state $x(k)$ and following the policy (2) afterwards. For the case of LQR, the Q-function can be expressed as a quadratic form in both $x(k)$ and $u(k)$ (see Bradtke et al., 1994 for details),

$$Q_K\left(x(k), u(k)\right) = \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^T H \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}, \tag{11}$$

$$H = \begin{bmatrix} Q + A^T PA & A^T PB \\ B^T PA & R + B^T PB \end{bmatrix}. \tag{12}$$

Using the Kronecker product,

$$Q_K\left(x(k), u(k)\right) = vec(H)^T \left(\begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \otimes \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \right). \tag{13}$$

where $vec(H)$ denotes the vector formed by stacking the columns of H. Accordingly, a quadratic basis with state and input trajectories are constructed to learn the Q-function parameters H through Q-learning algorithm.

The optimal Q-function $Q_K^*\left(x(k), u(k)\right)$ is regarded as the cost of executing the control signal $u(k)$ and then following the optimal policy K^* afterwards, i.e.,

$$\begin{aligned} Q_K^*\left(x(k), u(k)\right) &= r(x(k), u(k)) + V^*(x(k+1)) \\ &= \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^T H^* \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}, \end{aligned}$$

and the optimal control policy (8) is presented by,

$$\begin{aligned} u^*(k) &= \arg\min_u Q_K^*\left(x(k), u(k)\right) \\ &= -\left(R + B^T PB\right)^{-1} B^T PAx(k), \end{aligned}$$

with the optimal state feedback gain

$$K^* = -\left(R + B^T PB\right)^{-1} B^T PA.$$

3. SDP-based LQR control by learning the Q function: Centralized case

A non-iterative LMI-based optimization is proposed in Section 3.1 to find the optimal Q function parameters. The optimal Q function leads to the optimal policy that solves the LQR problem (4). By using properties of the LMI formulation, the design is extended to model-free learning of the Q function, and a completely data-driven optimal LQR control is presented in Section 3.2.

3.1. Model-based optimization of the Q-function in LQR control

Consider the alternative formulation of the infinite horizon LQR problem (4) presented in Lee and Hu (2019) and described by Problem 1.

Problem 1. Non-convex minimization with variables $S \in \mathbb{S}^{n+m}$ and $K \in \mathbb{R}^{m \times n}$.

$$\underset{S,K}{\text{Minimize}} \ \ trace(\Lambda S) \tag{14}$$

$$\text{Subject to } S \succeq 0, \quad A_K = \begin{bmatrix} A & B \\ KA & KB \end{bmatrix}$$

$$A_K S A_K^T + \begin{bmatrix} I_n \\ K \end{bmatrix} Z \begin{bmatrix} I_n \\ K \end{bmatrix}^T = S,$$

The optimal solution of Problem 1 is the optimal state feedback control in LQR problem (4). Moreover, the set of eigenvalues of A_K can be evaluated by,

$$\begin{aligned} &\det \begin{bmatrix} A - \lambda I & B \\ KA & KB - \lambda I \end{bmatrix} = \det \begin{bmatrix} A + BK - \lambda I & B \\ 0 & -\lambda I \end{bmatrix} \\ &= (-\lambda)^m \det(A + BK - \lambda I) = 0. \end{aligned}$$

Thus, $\text{eig}\,(A_K) = \text{eig}\,(A + BK) \cup \{\underbrace{0, \ldots, 0}_{\text{m zeros}}\}$. The constraint $K \in \mathcal{K}$ indicates that K is a stabilizing state feedback controller and the closed-loop state–space matrix $A + BK$ is schur stable, i.e., the spectrum of $A + BK$ is contained in the open unit disk in the complex plane. Spectrum analysis of A_K indicates that $A + BK$ is schur stable if and only if A_K is schur stable. According to the Lyapunov stability theorem, the matrix A_K (and as a result $A+BK$) is schur stable if and only if the Lyapunov matrix equality,

$$A_K S A_K^T + \begin{bmatrix} I_n \\ K \end{bmatrix} Z \begin{bmatrix} I_n \\ K \end{bmatrix}^T = S, \tag{15}$$

has a solution $S \succeq 0$ for each choice of the positive semi-definite matrix $\begin{bmatrix} I_n \\ K \end{bmatrix} Z \begin{bmatrix} I_n \\ K \end{bmatrix}^T$. Accordingly, $S \succeq 0$ is equivalent to the constraint $K \in \mathcal{K}$.

In this section, first, the dual problem associated with Problem 1 is represented as a semi-definite program whose dependence to the initial condition matrix is thoroughly captured in an affine objective function and the corresponding LMI constraints remain independent of the initial condition vectors. By using this structure, it is shown that the optimal solution of the dual problem is independent of the initial condition matrix Z. Moreover, the optimal solution of the corresponding semi-definite program is shown to be equivalent to the optimal Q-function parameters. Once the optimal Q-function is retrieved, the optimal control policy is derived.

Let the matrices G and H be the Lagrange multipliers associated with the inequality and equality constraints of Problem 1, respectively. The Lagrangian of the optimization Problem 1 is given by,

$$\begin{aligned} &\mathcal{L}(K, S, G, H) := trace(\Lambda S) - trace(GS) \\ &+ trace\left(\left(A_K S A_K^T + \begin{bmatrix} I_n \\ K \end{bmatrix} Z \begin{bmatrix} I_n \\ K \end{bmatrix}^T - S \right) H \right) \\ &= trace\left(\left(A_K^T H A_K - H - G + \Lambda \right) S \right) \\ &+ trace\left(Z \begin{bmatrix} I_n \\ K \end{bmatrix}^T H \begin{bmatrix} I_n \\ K \end{bmatrix} \right). \end{aligned} \tag{16}$$

The dual problem is given by,

$$\underset{G \succeq 0, H}{\text{Maximize}} \ \ g(G, H), \tag{17}$$

where the Lagrangian dual function $g(G, H)$ is,

$$g(G, H) := \inf_{K,S} \mathcal{L}(H, G, K, S). \tag{18}$$

The next theorem derives the dual problem as a standard convex optimization problem. Subsequently, it will be shown that its optimal solution is the same as the optimal parameters of the quadratic Q-function.

Theorem 1. *The dual problem (17) associated with the non-convex optimization in Problem 1 is representable as a convex optimization problem with conic constraints given by Problem 2.*

Problem 2. Convex minimization with variables $H \in \mathbb{S}^{n+m}$ and $W \in \mathbb{R}^{n \times n}$.

$$\underset{H,W}{\text{Maximize}} \ \ trace(ZW)$$

Subject to

$$\begin{aligned} &H_{11} - W - H_{12} H_{22}^{-1} H_{12}^T \succeq 0, \\ &\begin{bmatrix} A & B \end{bmatrix}^T (H_{11} - H_{12} H_{22}^{-1} H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda \succeq 0, \\ &H_{22} \succ 0. \end{aligned}$$

Proof. Since S and K are the minimizers of the Lagrangian function, boundedness of the first term in the Lagrangian (16)

from below requires that,

$$A_K^T H A_K - H - G + \Lambda \succeq 0. \tag{19}$$

The Lagrange multiplier G is positive semi-definite, and only appears in the first term of the Lagrangian (16). Thus, the constraint (19) is equivalent to,

$$A_K^T H A_K - H + \Lambda \succeq 0. \tag{20}$$

Let H be partitioned as $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}$, where $H_{11} \in \mathbb{S}^n$, $H_{22} \in \mathbb{S}^m$, and $H_{12} \in \mathbb{R}^{n\times m}$. The second term of the Lagrangian function (16) can be represented as,

$$trace\left(Z(K^T H_{22} K + K^T H_{12}^T + H_{12} K + H_{11})\right). \tag{21}$$

To construct the dual function, three different cases are considered:

Case 1: Let $H_{22} \succ 0$. In this case, the minimizer of the Lagrangian is obtained by taking the derivative of (21) with respect to K as follows,

$$\begin{aligned} &\frac{\partial}{\partial K}\left(trace\left(Z\left(K^T H_{22} K + K^T H_{12}^T + H_{12} K + H_{11}\right)\right)\right) \\ &= H_{22} K Z + H_{22}^T K Z^T + H_{12}^T Z + H_{12}^T Z^T = 0 \\ &\Rightarrow K^* = -H_{22}^{-1} H_{12}^T. \end{aligned}$$

Substituting K^* in the constraint (20) results in,

$$\begin{bmatrix} A & B \end{bmatrix}^T (H_{11} - H_{12} H_{22}^{-1} H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda \succeq 0. \tag{22}$$

Moreover, the objective function for the dual problem reduces to,

$$\underset{H}{\text{Maximize}}\ trace\left(Z\left(H_{11} - H_{12} H_{22}^{-1} H_{12}^T\right)\right). \tag{23}$$

Case 2: Let $H_{22} \succeq 0$ be a singular matrix, provided that $\mathcal{R}(H_{12}^T) \subseteq \mathcal{R}(H_{22})$. In this case, the Lagrangian would be still bounded. Let $H_{22} = U_r \Sigma_r U_r^T$ with $r < m$ be the singular value decomposition of the symmetric matrix H_{22}, where $\Sigma_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-zero eigenvalues of H_{22} on the main diagonal. Moreover, the columns of the unitary matrix $U_r \in \mathbb{R}^{m\times r}$ span the range-space of H_{22}. The pseudo-inverse of the singular matrix H_{22} can be represented as $H_{22}^\dagger = U_r \Sigma_r^{-1} U_r^T$. Following a procedure similar to the case 1, the dual objective function is derived as,

$$\underset{H}{\text{Maximize}}\ trace\left(Z\left(H_{11} - H_{12} H_{22}^\dagger H_{12}^T\right)\right), \tag{24}$$

with the constraint,

$$\begin{bmatrix} A & B \end{bmatrix}^T (H_{11} - H_{12} H_{22}^\dagger H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda \succeq 0. \tag{25}$$

However, every attainable objective value by the singular matrix $H_{22} \succeq 0$ with the condition $\mathcal{R}(H_{12}^T) \subseteq \mathcal{R}(H_{22})$ is alternatively achievable by $\tilde{H}_{22} \succ 0$ constructed as,

$$\tilde{H}_{22} = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & \Sigma_{m-r} \end{bmatrix} \begin{bmatrix} U_r^T \\ U_{m-r}^T \end{bmatrix}.$$

The columns of the unitary matrix U_{m-r} span the null-space of H_{12}. It is always possible to form U_{m-r}, because $\mathcal{R}(H_{12}^T) \subseteq \mathcal{R}(H_{22})$. The matrix Σ_{m-r} is an arbitrary diagonal matrix with positive diagonal elements. Note that,

$$\begin{aligned} &H_{11} - H_{12} \tilde{H}_{22}^{-1} H_{12}^T \\ &= H_{11} - H_{12} \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & \Sigma_{m-r}^{-1} \end{bmatrix} \begin{bmatrix} U_r^T \\ U_{m-r}^T \end{bmatrix} H_{12}^T \\ &= H_{11} - H_{12} U_r \Sigma_r^{-1} U_r^T H_{12}^T - \underbrace{H_{12} U_{m-r} \Sigma_{m-r}^{-1} U_{m-r}^T H_{12}^T}_{0} \\ &= H_{11} - H_{12} H_{22}^\dagger H_{12}^T. \end{aligned}$$

The last equality holds, because the columns of the unitary matrix U_{m-r} span the null-space of H_{12}.

Case 3: H_{22} contains at least one negative eigenvalue or it is positive semi-definite while $\mathcal{R}(H_{12}^T) \nsubseteq \mathcal{R}(H_{22})$. In this case, by minimizing (21) with respect to K, the Lagrangian would be unbounded from below.

In summary, boundedness of the Lagrangian from below requires that (i) $H_{22} \succ 0$, or (ii) $H_{22} \succeq 0$ with an additional constraint $\mathcal{R}(H_{12}^T) \subseteq \mathcal{R}(H_{22})$. Moreover, every permissible positive semi-definite candidate $H_{22} \succeq 0$ results in an objective value that can be alternatively derived by a positive-definite counterpart $\tilde{H}_{22} \succ 0$. Thus, without loss of generality we proceed by assuming,

$$H_{22} \succ 0. \tag{26}$$

Finally, by introducing the slack variable $W \in \mathbb{S}^n$ and the additional constraint,

$$H_{11} - W - H_{12} H_{22}^{-1} H_{12}^T \succeq 0, \tag{27}$$

the objective function (23) can be equivalently described by,

$$\underset{H,W}{\text{Maximize}}\ trace(ZW). \tag{28}$$

To see this, let W^* be the optimal solution of the reformulated objective function (28) and H^* be the optimal solution of the primary objective function (23). According to the constraint (27) and positive definiteness of Z,

$$trace\left(ZW^*\right) \le trace\left(Z\left(H_{11}^* - H_{12}^* (H_{22}^*)^{-1} (H_{12}^*)^T\right)\right).$$

Since W^* is the maximizer of the objective function (28), it follows that,

$$trace\left(ZW^*\right) \ge trace\left(Z\left(H_{11}^* - H_{12}^* (H_{22}^*)^{-1} (H_{12}^*)^T\right)\right),$$

and $trace\left(Z\left(W^* - H_{11}^* + H_{12}^* (H_{22}^*)^{-1} (H_{12}^*)^T\right)\right) = 0$. Consequently, all of the eigenvalues of the matrix $W^* - H_{11}^* + H_{12}^* (H_{22}^*)^{-1} (H_{12}^*)^T$ are equal to zero. Since W^* is symmetric, it follows that

$$W^* = H_{11}^* - H_{12}^* (H_{22}^*)^{-1} (H_{12}^*)^T,$$

which means that the optimization problem with the objective function (23) and the constraints (22) and (26) is equivalent to optimization Problem 2. □

The next theorem shows that the optimal solution of the dual problem is the same as the optimal parameters of the quadratic Q-function.

Theorem 2. *The optimal solution H^* of the optimization Problem 2, is independent of the initial condition matrix Z provided that $Z \succ 0$. In this case, H^* represents the optimal quadratic form of the Q-function expressed in (11).*

Proof. Since Problem 2 is convex, fulfillment of the Karush–Kuhn–Tucker (KKT) conditions is sufficient for optimality (Boyd & Vandenberghe, 2004). The Lagrangian associated with Problem 2 is described by,

$$\begin{aligned} &\mathcal{L}(H, W, M_i) = -trace\left(ZW\right) \\ &- trace\left(M_1\left(H_{11} - W - H_{12} H_{22}^{-1} H_{12}^T\right)\right) \\ &- trace\Big(M_2\Big(\begin{bmatrix} A & B \end{bmatrix}^T (H_{11} - H_{12} H_{22}^{-1} H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} \\ &- H + \Lambda\Big)\Big) - trace\left(M_3\left(H_{22}\right)\right) \end{aligned} \tag{29}$$

In addition to the feasibility conditions given in Problem 2, the optimality conditions require dual feasibility, complementary slackness, and stationarity conditions.

(I) Dual feasibility:

$$M_1 \succeq 0, \quad M_2 \succeq 0, \quad M_3 \succeq 0. \tag{30}$$

(II) Complementary slackness:

$$trace\left(M_1\left(H_{11} - W - H_{12}H_{22}^{-1}H_{12}^T\right)\right) = 0, \tag{31}$$

$$trace\Big(M_2\Big(\begin{bmatrix}A & B\end{bmatrix}^T\left(H_{11} - H_{12}H_{22}^{-1}H_{12}^T\right)\begin{bmatrix}A & B\end{bmatrix} - H + \Lambda\Big)\Big) = 0, \tag{32}$$

$$trace\left(M_3\left(H_{22}\right)\right) = 0. \tag{33}$$

(III) Stationarity conditions:

$$\nabla_W \mathcal{L} = 0, \quad \nabla_H \mathcal{L} = 0. \tag{34}$$

Now, consider the candidate for the optimal solution, $\hat{H} = \begin{bmatrix} Q + A^TPA & A^TPB \\ B^TPA & R + B^TPB \end{bmatrix}$, $\hat{W} = P$, in which $P \in \mathbb{S}^n_{++}$ is the solution to the DARE (9). Substituting the candidate $(\hat{H}, \hat{W})$ in the left-hand side of the inequality constraint (27) gives,

$$\begin{aligned}&\hat{H}_{11} - \hat{W} - \hat{H}_{12}\hat{H}_{22}^{-1}\hat{H}_{12}^T \\ &= Q + A^TPA - P - A^TPB\left(R + B^TPB\right)^{-1}B^TPA = 0.\end{aligned} \tag{35}$$

That is, the constraint (27) is always active and (31) is trivially satisfied without imposing any restriction on the Lagrange multiplier M_1. Moreover, substituting the candidate $\hat{H}$ in the left-hand side of (22) results in,

$$\begin{aligned}&\begin{bmatrix}A & B\end{bmatrix}^T(\hat{H}_{11} - \hat{H}_{12}\hat{H}_{22}^{-1}\hat{H}_{12}^T)\begin{bmatrix}A & B\end{bmatrix} - \hat{H} + \Lambda \\ &= \begin{bmatrix}A & B\end{bmatrix}^T P\begin{bmatrix}A & B\end{bmatrix} - \begin{bmatrix}A & B\end{bmatrix}^T P\begin{bmatrix}A & B\end{bmatrix} \\ &- \Lambda + \Lambda = 0.\end{aligned}$$

It can be observed that the constraint (22) is active, i.e., the constraint (32) is trivially satisfied without any requirement for the Lagrange multiplier M_2.

However, the constraint (26) is not active at the candidate point ($\hat{H}_{22} \succ 0$), which dictates the Lagrange multiplier M_3 to be equal to zero.

It remains to show that the stationary conditions can be satisfied at the candidate point $(\hat{H}, \hat{W})$ by appropriate selection of M_1 and M_2. To see this, we note that by using the matrices $E_1 = \begin{bmatrix}I_n & 0\end{bmatrix}^T \in \mathbb{R}^{(n+m)\times n}$ and $E_2 = \begin{bmatrix}0 & I_m\end{bmatrix}^T \in \mathbb{R}^{(n+m)\times m}$, the matrix variable H can be represented as,

$$H = E_1H_{11}E_1^T + E_1H_{22}E_2^T + E_2H_{12}^TE_1^T + E_2H_{22}E_2^T.$$

Then, the stationarity conditions (34), can be equivalently described by equalities (36)–(39).

$$\nabla_W \mathcal{L} = -Z^T + M_1^T = 0, \tag{36}$$

$$\nabla_{H_{11}} \mathcal{L} = -M_1 - \begin{bmatrix}A & B\end{bmatrix}M_2\begin{bmatrix}A & B\end{bmatrix}^T + E_1^TM_2E_1 = 0, \tag{37}$$

$$\begin{aligned}\nabla_{H_{12}} \mathcal{L} &= \begin{bmatrix}A & B\end{bmatrix}M_2\begin{bmatrix}A & B\end{bmatrix}^T\hat{H}_{12}\hat{H}_{22}^{-1} + M_1\hat{H}_{12}\hat{H}_{22}^{-1} \\ &+ E_1^TM_2E_2 = 0\end{aligned} \tag{38}$$

$$\begin{aligned}\nabla_{H_{22}} \mathcal{L} &= E_2^TM_2E_2 - \hat{H}_{22}^{-1}\Big(\hat{H}_{12}^TM_1\hat{H}_{12} \\ &+ \hat{H}_{12}^T\begin{bmatrix}A & B\end{bmatrix}M_2\begin{bmatrix}A & B\end{bmatrix}^T\hat{H}_{12}\Big)\hat{H}_{22}^{-1} = 0\end{aligned} \tag{39}$$

The Lagrange multiplier M_1 is chosen to be the same as (positive definite) initial condition matrix Z, i.e., $M_1 = Z \succ 0$, which satisfies the stationary condition (36). (Note that the complementary slackness poses no restriction on the selection of the Lagrange multiplier M_1.)

From (37) and (38), the Matrix M_2 should comply with the constraint,

$$E_1^TM_2E_1\hat{H}_{12}\hat{H}_{22}^{-1} + E_1^TM_2E_2 = 0, \tag{40}$$

Moreover, by using (37) and (39), the Matrix M_2 should satisfy the constraint,

$$\hat{H}_{12}^TE_1^TM_2E_1\hat{H}_{12} + \hat{H}_{22}E_2^TM_2E_2\hat{H}_{22} = 0. \tag{41}$$

By partitioning the Lagrange multiplier M_2 as $M_2 = \begin{bmatrix}M_{11} & M_{12} \\ M_{12}^T & M_{22}\end{bmatrix}$, the equalities (40) and (41) are represented as,

$$M_{11}\hat{H}_{12}\hat{H}_{22}^{-1} + M_{12} = 0, \tag{42}$$

$$\hat{H}_{12}^TM_{11}\hat{H}_{12} + M_{22} = 0. \tag{43}$$

By substituting (42) and (43) in (37), we have,

$$M_1 + \left[A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T\right]M_{11}\left[A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T\right]^T = M_{11}. \tag{44}$$

The existence of a feasible solution $M_{11} \succeq 0$ to assure (44) is equivalent to the existence of a Lyapunov matrix $M_{11} \succeq 0$ for a Lyapunov equality with dynamic system matrix $A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T$ and positive definite matrix $M_1 = Z \succ 0$. The system matrix $A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T$ can be regarded as the closed-loop dynamics of the discrete-time system (1) with a stabilizing state feedback control $K = K = \hat{H}_{22}^{-1}\hat{H}_{12}^T$, which is Hurwitz (Based on the preliminary argument in Section 2). Accordingly, the Lyapunov matrix $M_{11} \succeq 0$ exists. Finally, the Lagrange multiplier M_2 is constructed as follows,

$$\begin{aligned}M_2 &= \begin{bmatrix} M_{11} & -M_{11}\hat{H}_{12}\hat{H}_{22}^{-1} \\ \hat{H}_{22}^{-1}\hat{H}_{12}^TM_{11} & \hat{H}_{22}^{-1}\hat{H}_{12}^TM_{11}\hat{H}_{12}\hat{H}_{22}^{-1}\end{bmatrix} \\ &= \begin{bmatrix} I \\ -\hat{H}_{22}^{-1}\hat{H}_{12}^T\end{bmatrix}M_{11}\begin{bmatrix}I & -\hat{H}_{12}(\hat{H}_{22}^{-1})^T\end{bmatrix} \succeq 0.\end{aligned}$$

□

Remark 1. It should be noted that the dual problem is always convex, even if the primal problem is not convex (Boyd & Vandenberghe, 2004). For general convex primal problems, under some mild conditions such as Slater's condition (Boyd & Vandenberghe, 2004), strong duality holds, i.e., the optimal value of the primal and dual problems are equal. However, strong duality does not hold in general non-convex problems. Although the underlying primal problem in this paper is non-convex, we have shown that the optimal solution H^* of the dual problem represents the optimal parameters of the Q-function.

Theorem 2 implies that selection of positive definite matrix Z only affects the optimal value of the cost function and the optimal solution H^* would be independent of Z. Then, without loss of generality, the objective function in Problem 2 can be expressed as,

$$\underset{H,W}{\text{Maximize}}\ trace(W). \tag{45}$$

Moreover, as shown in proof of Theorem 2, the initial condition matrix Z is linearly affecting the Lagrange multipliers of both convex inequality constraints in Problem 2. That is,

$$\begin{aligned}&M_1 = Z \\ &M_2 = \begin{bmatrix} I \\ -\hat{H}_{22}^{-1}\hat{H}_{12}^T\end{bmatrix}M_{11}\begin{bmatrix}I & -\hat{H}_{12}(\hat{H}_{22}^{-1})^T\end{bmatrix} \\ &\left[A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T\right]M_{11}\left[A - B\hat{H}_{22}^{-1}\hat{H}_{12}^T\right]^T - M_{11} = -Z.\end{aligned}$$

Finally, (27) can be expressed in the form of an LMI by using the Schur complement property (Boyd et al., 1994) as $\begin{bmatrix} H_{11}-W & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \succeq 0$, which implicitly contains (26). Utilizing the Schur complement once more, (22) may also be rewritten as the following LMI,

$$\begin{bmatrix} \begin{bmatrix} A & B \end{bmatrix}^T H_{11} \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda & \begin{bmatrix} A & B \end{bmatrix}^T H_{12} \\ H_{12}^T \begin{bmatrix} A & B \end{bmatrix} & H_{22} \end{bmatrix} \succeq 0.$$

Consequently, Problem 1 is equivalently described by the semi-definite program in Problem 3, whose optimal point H^* constructs the optimal Q-function, according to Theorem 2.

Problem 3 (*Model-based SDP to Find the Q-function*)**.** A convex program with variables $H \in \mathbb{S}^{n+m}$ and $W \in \mathbb{S}^n$.

$$\underset{H,W}{\text{Maximize}} \ \ trace(W)$$
$$\text{Subject to} \ \begin{bmatrix} H_{11}-W & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \succeq 0,$$
$$\begin{bmatrix} \begin{bmatrix} A & B \end{bmatrix}^T H_{11} \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda & \begin{bmatrix} A & B \end{bmatrix}^T H_{12} \\ H_{12}^T \begin{bmatrix} A & B \end{bmatrix} & H_{22} \end{bmatrix} \succeq 0.$$

While the centralized LQR problem has a well-known convex representation for retrieving the state feedback controller, the semi-definite program derived in Problem 3 has some noteworthy properties. First, it can be modified in such a way that the requirement for model information is replaced by data samples of the system's input and state trajectories, which is elaborated in Section 3.2. Moreover, incorporation of the Q function parameters as the decision variables facilitates the extension of the model-based and model-free design to the distributed case where the problem is known to be non-convex even if the design topology is centralized. (See Section 4).

3.2. Model-free learning of the Q-function parameters

The optimization problem in Section 3.1 requires knowledge of the system dynamics, i.e. the matrices A and B. In this subsection, we try to circumvent this requirement by using the system's input and state data. Define $D \in \mathbb{R}^{(n+m)\times l}$ with input and state trajectories as,

$$D := \begin{bmatrix} x(k_0) & x(k_0+1) & \cdots & x(k_0+l-1) \\ u(k_0) & u(k_0+1) & \cdots & u(k_0+l-1) \end{bmatrix}, \tag{46}$$

where $x(k) = [x_1(k) \ \ x_2(k) \ \ \cdots \ \ x_n(k)]^T \in \mathbb{R}^n$, $u(k) = [u_1(k) \ \ u_2(k) \ \ \cdots \ \ u_m(k)]^T \in \mathbb{R}^m$, and k_0 is the index of the first sample to be collected. Define $X_D \in \mathbb{R}^{n\times l}$ as,

$$\begin{aligned} X_D &:= \begin{bmatrix} A & B \end{bmatrix} D \\ &= \begin{bmatrix} Ax(k_0)+Bu(k_0) & Ax(k_0+1)+Bu(k_0+1) \\ & \cdots \ Ax(k_0+l-1)+Bu(k_0+l-1) \end{bmatrix} \\ &= \begin{bmatrix} x(k_0+1) & x(k_0+2) & \cdots & x(k_0+l) \end{bmatrix}. \end{aligned} \tag{47}$$

According to the properties of matrix congruence (Strang, 1968), as long as the matrix D has column rank of $m+n$, it is possible to multiply the inequality constraint,

$$\begin{bmatrix} A & B \end{bmatrix}^T (H_{11} - H_{12}H_{22}^{-1}H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} - H + \Lambda \succeq 0. \tag{48}$$

from left and right by D^T and D, to derive the equivalent constraint,

$$\begin{aligned} D^T \begin{bmatrix} A & B \end{bmatrix}^T & (H_{11} - H_{12}H_{22}^{-1}H_{12}^T) \begin{bmatrix} A & B \end{bmatrix} D \\ & - D^T(H-\Lambda)D \succeq 0 \end{aligned} \tag{49}$$

$$\Rightarrow X_D^T(H_{11} - H_{12}H_{22}^{-1}H_{12}^T)X_D - D^T(H-\Lambda)D \succeq 0. \tag{50}$$

To ensure that the matrix D has a column rank of $m+n$, the number of samples l must be more than or equal to the sum of the number of inputs and states. Therefore, it suffices to collect $l+1 = m+n+1$ samples. Moreover, due to the linear dependency of the states and inputs on the past states and inputs, the input signal within the interval $[k_0, k_0+l-1]$ must be persistently exciting (PE) of order L with $L \geq n$ (Willems, Rapisarda, Markovsky, & De Moor, 2005). The formal definition of persistent excitation for the signal u within the range $[k_0, k_0+l-1]$ is given in the following.

Definition 1 (*Willems et al., 2005*)**.** The signal $u_{[k_0,k_0+l-1]}$ (restriction of the sequence u to the interval $[k_0, k_0+l-1]$) is persistently exciting of order L, if the matrix $\mathcal{H}_L(u)$ has full row rank.

$$\mathcal{H}_L(u) = \begin{bmatrix} u(k_0) & u(k_0+1) & \cdots & u(k_0+l-L) \\ u(k_0+1) & u(k_0+2) & \cdots & u(k_0+l-L+1) \\ \vdots & \vdots & \ddots & \vdots \\ u(k_0+L-1) & u(k_0+L) & \cdots & u(k_0+l-1) \end{bmatrix} \tag{51}$$

In practice, a small probing noise composed of a Gaussian white noise or a sum of sinusoids must be added to the control signal to ensure the PE condition.

Finally, since $H_{22} \succ 0$, by direct application of Schur complement (Boyd et al., 1994), the constraint (50) is equivalent to,

$$\begin{bmatrix} X_D^T H_{11} X_D - D^T(H-\Lambda)D & X_D^T H_{12} \\ H_{12}^T X_D & H_{22} \end{bmatrix} \succeq 0. \tag{52}$$

Theorem 3 introduces an SDP for solving the LQR problem (4) without any information of the system dynamics nor an initial admissible control policy, i.e., an initial stabilizing controller. This means that the optimal LQR controller may be found for an initially open-loop unstable system as long as $l+1$ samples could be collected to form the matrices D and X_D.

Theorem 3. *The optimal state feedback controller $K^* \in \mathbb{R}^{m\times n}$ associated with the LQR problem (4) is,*

$$K^* = -(H_{22}^*)^{-1}(H_{12}^*)^T, \tag{53}$$

where $H^* = \begin{bmatrix} H_{11}^* & H_{12}^* \\ (H_{12}^*)^T & H_{22}^* \end{bmatrix} \in \mathbb{R}^{l\times l}$ is the solution of the SDP given in Problem 4, and the matrices D and X_D are formed by the trajectories as (46) and (47), respectively.

Problem 4 (*Model-free SDP to Derive the Q-function*)**.** A convex program with variables $H \in \mathbb{S}^{n+m}$ and $W \in \mathbb{S}^n$.

$$\underset{H,W}{\text{Maximize}} \ \ trace(W)$$
$$\text{subject to} \ \begin{bmatrix} H_{11}-W & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \succeq 0,$$
$$\begin{bmatrix} X_D^T H_{11} X_D - D^T(H-\Lambda)D & X_D^T H_{12} \\ H_{12}^T X_D & H_{22} \end{bmatrix} \succeq 0.$$

Proof. It is already shown that according to the properties of matrix congruence, the constraint (50) is equivalent to the second constraint in Problem 4. Based on Theorem 2, the exact Q-function is retrieved by solving Problem 2 which is equivalently described as Problem 4. Once the optimal Q-function is derived, Q-learning offers the optimal state feedback controller as $K^* = -(H_{22}^*)^{-1}(H_{12}^*)^T$. □

It should be noted that model-free iterative algorithms such as policy iteration and Q-learning are also capable of solving the linear quadratic regulation problem with the guarantee of convergence to the optimal policy as the number of iterations goes to infinity. However, the proposed off-policy approach is non-iterative and derives the optimal policy in a single step by using only $l+1=m+n+1$ samples. It does not require the availability of an initial stabilizing controller, or hyper-parameter tuning. It is sample-efficient and as a known feature of LMIs, it is inherently robust to model uncertainties.

Moreover, it is known that the extension of the model-free LQR design to the distributed case is not straightforward. The distributed optimal problem with arbitrary control structure results in an NP-hard problem, even if the dynamics are known. In the model-free distributed design, only partial observation is available which faces the learning procedure with serious challenges. One of the noteworthy features of the proposed method is that it is amenable to be extended to model-free distributed control of interconnected systems at the expense of introducing some levels of sub-optimality.

4. Optimal control of interconnected systems

Emerging complex interconnected systems have introduced the requirement of distributed control of the systems without direct and complete knowledge of all subsystems. In this section, the results of Section 3 will be extended to the case where the controller is implemented in a distributed manner. Consider an interconnected system consisting of N subsystems. The dynamic equation of each subsystem is given by,

$$x_i(k+1)=\sum_{j=1}^{N}A_{ij}x_j(k)+B_iu_i(k), \tag{54}$$

where $x_i\in\mathbb{R}^{n_i}$ is the vector of subsystem's states, $u_i\in\mathbb{R}^{m_i}$ is the vector of subsystem's inputs, $A_{ii}\in\mathbb{R}^{n_i\times n_i}$ and $B_i\in\mathbb{R}^{m_i}$ are the subsystem's internal dynamics, and $A_{ij}\in\mathbb{R}^{n_i\times n_j}$ determines the influence of subsystem j on subsystem i. The overall interconnected system is described as,

$$x(k+1)=Ax(k)+Bu(k),\quad x(0)=z\in\mathbb{R}^n, \tag{55}$$

$$A=\begin{bmatrix}A_{11}&A_{12}&\cdots&A_{1N}\\A_{21}&A_{22}&\cdots&A_{2N}\\\vdots&\vdots&\ddots&\vdots\\A_{N1}&A_{N2}&\cdots&A_{NN}\end{bmatrix},\quad B=\begin{bmatrix}B_1&0&\cdots&0\\0&B_2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&B_N\end{bmatrix}.$$

Moreover, $n=\sum_{i=1}^{N}n_i$, and $m=\sum_{i=1}^{N}m_i$.

The control signal for the subsystem i is constructed by a state feedback policy,

$$u_i(k)=K_{\mathcal{N}_i}x_{\mathcal{N}_i}, \tag{56}$$

where $K_{\mathcal{N}_i}\in\mathbb{R}^{m_i\times n_{\mathcal{N}_i}}$ is the local controller gain, and $x_{\mathcal{N}_i}$ is constructed by stacking the state vectors x_j for $j\in\mathcal{N}_i$. The subsystem j is referred to as a neighbor of the subsystem i ($j\in\mathcal{N}_i$) if the subsystem i has access to the states of subsystem j, i.e., x_j.

The cost function for subsystem i is defined as,

$$V_i\left(x(k)\right)=\sum_{l=k}^{\infty}r_i\left(x(l),u(l)\right),$$

where the stage cost is

$$r_i\left(x(l),u(l)\right)=\frac{1}{N}x^T(k)Qx(k)+u_i^T(l)R_iu_i^T(l).$$

$Q\in\mathbb{S}_+^n$ and $R\in\mathbb{S}_{++}^m$ are the weighting matrices for the overall performance defined as,

$$Q=\begin{bmatrix}Q_{11}&Q_{12}&\cdots&Q_{1N}\\Q_{21}&Q_{22}&\cdots&Q_{2N}\\\vdots&\vdots&\ddots&\vdots\\Q_{N1}&Q_{N2}&\cdots&Q_{NN}\end{bmatrix},\quad R=\begin{bmatrix}R_1&0&\cdots&0\\0&R_2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&R_N\end{bmatrix},$$

where $Q_{ij}\in\mathbb{R}^{n_i\times n_j}$. The matrix $R_i\in\mathbb{S}_{++}^{m_i}$ represents the input weighting matrix for the subsystem i. The optimal control problem for an interconnected system is expressed as,

$$\min_{K_i}\sum_{i=1}^{N}V_i\left(x_i(0)\right)\quad\text{subject to (54), (56).} \tag{57}$$

Let the structured identity $I_{\mathcal{S}}$ be defined as,

$$I_{\mathcal{S}}=\begin{bmatrix}I_{\mathcal{S}_{11}}&I_{\mathcal{S}_{12}}&\cdots&I_{\mathcal{S}_{1N}}\\I_{\mathcal{S}_{21}}&I_{\mathcal{S}_{22}}&\cdots&I_{\mathcal{S}_{2N}}\\\vdots&\vdots&\ddots&\vdots\\I_{\mathcal{S}_{N1}}&I_{\mathcal{S}_{N2}}&\cdots&I_{\mathcal{S}_{NN}}\end{bmatrix}\in\mathbb{R}^{m\times n},$$

where $I_{\mathcal{S}_{ij}}=\begin{cases}J_{i,j}, & \text{if } j\in\mathcal{N}_i\\ 0_{i,j}, & \text{otherwise}\end{cases}$ $J_{i,j}\in\mathbb{R}^{m_i\times n_j}$ and $0_{i,j}\in\mathbb{R}^{m_i\times n_j}$ are the matrices of ones and zeros with the specified dimensions, respectively. Moreover, consider the block diagonal matrix

$$I_{\mathcal{D}}=\begin{bmatrix}J_{11}&0&\cdots&0\\0&J_{22}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&J_{NN}\end{bmatrix}\in\mathbb{R}^{m\times m}.$$

The set $\mathcal{S}$ is defined as $\mathcal{S}:=\{K\in\mathbb{R}^{m\times n}:K\circ I_{\mathcal{S}}=K\}$, where $\circ$ represents the element-wise multiplication of matrices. The state feedback gain for the overall system is a distributed control strategy in which the inputs of all subsystems adopt the state feedback policy $u_i(k)=K_{\mathcal{N}_i}x_{\mathcal{N}_i}$, and the structural constraint $K\in\mathcal{S}$ is satisfied. It is known that the optimal design of such state feedback gains is NP-hard even if the system dynamics are known.

In the rest of this section, two alternative cases are investigated. In Section 4.1, the entity that designs the distributed controller has access to the state and input data from every subsystem, while the implementation of the controller is distributed. Section 4.2 describes the case where each subsystem offers the model-free design of its own local controller with only receiving data from its neighboring subsystems.

4.1. Model-free distributed LQR control: Centralized design

Consider again the model-based SDP provided in Problem 3. It is known that incorporation of an arbitrary structural constraint $K\in\mathcal{S}$ makes the optimal control design non-convex. Consider Problem 3 with two additional affine constraints on the Q-function parameters, presented as Problem 5.

Problem 5 (*Model-based SDP for Centralized Design of Distributed Controllers*)**.**

$$\underset{H,W}{\text{Maximize}}\ \ trace(W)$$

$$\text{subject to}\ \begin{bmatrix}H_{11}-W&H_{12}\\H_{12}^T&H_{22}\end{bmatrix}\succeq 0,$$

$$H_{12}^T\in\mathcal{S},\quad \mathcal{S}=\{K\in\mathbb{R}^{m\times n}:K\circ I_{\mathcal{S}}=K\},$$

$$H_{22}\in\mathcal{D},\quad \mathcal{D}=\{P\in\mathbb{R}^{m\times m}:P\circ I_{\mathcal{D}}=P\},$$

$$\begin{bmatrix}\begin{bmatrix}A&B\end{bmatrix}^T H_{11}\begin{bmatrix}A&B\end{bmatrix}-H+\Lambda&\begin{bmatrix}A&B\end{bmatrix}^T H_{12}\\H_{12}^T\begin{bmatrix}A&B\end{bmatrix}&H_{22}\end{bmatrix}\succeq 0.$$

Then, it is straightforward to see that the sub-optimal structured controller given by $K^* = -(H_{22}^*)^{-1}(H_{12}^*)^T$, satisfies the required structure, i.e., $K \in \mathcal{S}$. where H_{12}^* and H_{22}^* are the solution of the SDP in Problem 5.

Moreover, similar to the procedure described in Section 3.2, it is possible to multiply the second matrix inequality from the left and right side by D^T and D respectively, which results in an equivalent inequality given in Problem 6.

Problem 6 (*Model-free SDP for Centralized Design of Distributed Controllers*)**.**

$$\underset{H,W}{\text{Maximize}} \ \ trace(W)$$

$$\text{subject to} \ \begin{bmatrix} H_{11} - W & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \succeq 0,$$

$$H_{12}^T \in \mathcal{S}, \quad \mathcal{S} = \{K \in \mathbb{R}^{m\times n} : K \circ I_{\mathcal{S}} = K\},$$

$$H_{22} \in \mathcal{D}, \quad \mathcal{D} = \{P \in \mathbb{R}^{m\times m} : P \circ I_{\mathcal{D}} = P\},$$

$$\begin{bmatrix} X_D^T H_{11} X_D - D^T(H - \Lambda)D & X_D^T H_{12} \\ H_{12}^T X_D & H_{22} \end{bmatrix} \succeq 0.$$

It should be noted that the replacement of the non-convex constraint $K \in \mathcal{S}$ by the two convex constraints in Problem 6 introduces some levels of conservatism to the problem by limiting the feasible set of the Q-function parameters. As illustrated in Section 5, empirical studies confirm a satisfactory level of performance for this data-driven approach in most cases and show promising directions toward optimal model-free distributed control of LTI systems.

In this setting, a central processing unit collects the required data and constructs the matrices D and X_D according to (46) and (47) by receiving input and state trajectories from all the subsystems. Then, the sub-optimal controller K^* is calculated by (53), and the matrices $K_{\mathcal{N}_i}^*$ are transmitted to their respective subsystems.

4.2. Model-free distributed LQR control: Distributed design

Design of distributed controllers for interconnected systems is confronted by difficulties such as computational limitations and privacy concerns. Accordingly, it would be beneficial to perform the model-free design procedure in a distributed manner, where no central entity is utilized. In this section, we address the model-free design procedure by offering a conic constraint optimization for each subsystem which makes the routine design distributed, that is, the ith subsystem designs its local state feedback gains $K_{\mathcal{N}_i} \in \mathbb{R}^{m_i \times n_{\mathcal{N}_i}}$ for the state feedback policy (56), only using information received from neighboring subsystems $j \in \mathcal{N}_i$.

In the case of model-based distributed design, it is assumed that subsystem i has access to the matrices $\bar{A}_{\mathcal{N}_i} \in \mathbb{R}^{n_{\mathcal{N}_i}\times n}$ and $\bar{B}_{\mathcal{N}_i} \in \mathbb{R}^{n_{\mathcal{N}_i}\times m}$ as,

$$\bar{A}_{\mathcal{N}_i} = \begin{bmatrix} A_{j1} & A_{j2} & \cdots & A_{jN} \end{bmatrix}, \quad \text{for } j \in \mathcal{N}_i,$$

$$\bar{B}_{\mathcal{N}_i} = \begin{bmatrix} 0 & \cdots & 0 & B_j & 0 & \cdots & 0 \end{bmatrix}, \quad \text{for } j \in \mathcal{N}_i,$$

where $n_{\mathcal{N}_i} = \sum_{j\in\mathcal{N}_i} n_j$.

Consider again the Q-function (11). Lack of access to the states and control signals of subsystems $j \notin \mathcal{N}_i$ can be circumvented by enforcing the corresponding rows and columns in H to be zero. Otherwise, direct access to non-neighboring subsystems (model or data) would be necessary. Since additional constraints are added to the problem, the controller must be sub-optimal as a result of limited access to system information. It is straightforward to show that for the LMI representation of the constraints, one may eliminate the corresponding zero rows/columns associated with the subsystems $j \notin \mathcal{N}_i$, and use the more compact form of the Q-function given by,

$$H_{\mathcal{N}_i} = \begin{bmatrix} H_{\mathcal{N}_i,11} & H_{\mathcal{N}_i,12} \\ H_{\mathcal{N}_i,12}^T & H_{\mathcal{N}_i,22} \end{bmatrix} \in \mathbb{R}^{l_{\mathcal{N}_i}\times l_{\mathcal{N}_i}}, \tag{58}$$

$$H_{\mathcal{N}_i,11} \in \mathbb{R}^{n_{\mathcal{N}_i}\times n_{\mathcal{N}_i}}, \quad H_{\mathcal{N}_i,12} \in \mathbb{R}^{n_{\mathcal{N}_i}\times m_{\mathcal{N}_i}},$$

$$H_{\mathcal{N}_i,22} \in \mathbb{R}^{m_{\mathcal{N}_i}\times m_{\mathcal{N}_i}}, \quad W_{\mathcal{N}_i} \in \mathbb{R}^{n_{\mathcal{N}_i}\times n_{\mathcal{N}_i}},$$

where, $m_{\mathcal{N}_i} = \sum_{j\in\mathcal{N}_i} m_j$ and $l_{\mathcal{N}_i} = n_{\mathcal{N}_i} + m_{\mathcal{N}_i}$.

The customized model-based distributed design of the sub-optimal state feedback controllers for subsystem i is described in Problem 7.

Problem 7 (*Model-based SDP for Distributed Design of Distributed Controllers*)**.**

$$\underset{H_{\mathcal{N}_i},W_{\mathcal{N}_i}}{\text{Maximize}} \ \ trace(W_{\mathcal{N}_i})$$

$$\text{subject to} \ \begin{bmatrix} H_{\mathcal{N}_i,11} - W_{\mathcal{N}_i} & H_{\mathcal{N}_i,12} \\ H_{\mathcal{N}_i,12}^T & H_{\mathcal{N}_i,22} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} \begin{bmatrix} \bar{A}_{\mathcal{N}_i} & \bar{B}_{\mathcal{N}_i} \end{bmatrix}^T H_{\mathcal{N}_i,11} \begin{bmatrix} \bar{A}_{\mathcal{N}_i} & \bar{B}_{\mathcal{N}_i} \end{bmatrix} - \bar{H}_{\mathcal{N}_i} + \Lambda & \\ H_{\mathcal{N}_i,12}^T \begin{bmatrix} \bar{A}_{\mathcal{N}_i} & \bar{B}_{\mathcal{N}_i} \end{bmatrix} & \end{bmatrix}$$

$$\begin{bmatrix} \bar{A}_{\mathcal{N}_i} & \bar{B}_{\mathcal{N}_i} \end{bmatrix}^T H_{\mathcal{N}_i,12} \\ H_{\mathcal{N}_i,22} \end{bmatrix} \succeq 0.$$

By solving Problem 7, the sub-optimal controller gain matrix for the neighborhood $\mathcal{N}_i$ is obtained as,

$$\bar{K}_{\mathcal{N}_i} = -H_{\mathcal{N}_i,22}^{-1} H_{\mathcal{N}_i,12}^T \in \mathbb{R}^{m_{\mathcal{N}_i}\times n_{\mathcal{N}_i}}. \tag{59}$$

The rows of $\bar{K}_{\mathcal{N}_i}$ corresponding to the control signals of subsystem i are then restored as the controller gain matrix $K_{\mathcal{N}_i} \in \mathbb{R}^{m_i \times n_{\mathcal{N}_i}}$.

To extend this approach to the model-free distributed design, note that the subsystem i does not have access to the states and control signals of the subsystems $j \notin \mathcal{N}_i$. Accordingly, let,

$$D_{\mathcal{N}_i} := \begin{bmatrix} x_{\mathcal{N}_i}(k_0) & x_{\mathcal{N}_i}(k_0+1) & \cdots & x_{\mathcal{N}_i}(k_0 + l_{\mathcal{N}_i} - 1) \\ u_{\mathcal{N}_i}(k_0) & u_{\mathcal{N}_i}(k_0+1) & \cdots & u_{\mathcal{N}_i}(k_0 + l_{\mathcal{N}_i} - 1) \end{bmatrix},$$

in which $x_{\mathcal{N}_i}(k)$ is already defined and $u_{\mathcal{N}_i}(k)$ is formed by stacking the input vectors $u_j(k)$ for $j \in \mathcal{N}_i$.

Similar to Section 3.2, by incorporation of state and input trajectories, the subsystem i faces the matrix inequality constraint,

$$\begin{aligned} &D_{\mathcal{N}_i}^T \begin{bmatrix} A_{\mathcal{N}_i} & B_{\mathcal{N}_i} \end{bmatrix}^T (H_{\mathcal{N}_i,11} - H_{\mathcal{N}_i,12} H_{\mathcal{N}_i,22}^{-1} H_{\mathcal{N}_i,12}^T) \\ &\begin{bmatrix} A_{\mathcal{N}_i} & B_{\mathcal{N}_i} \end{bmatrix} D_{\mathcal{N}_i} - D_{\mathcal{N}_i}^T (H_{\mathcal{N}_i} - \Lambda_{\mathcal{N}_i}) D_{\mathcal{N}_i} \succeq 0, \end{aligned} \tag{60}$$

$$A_{\mathcal{N}_i} = \begin{bmatrix} A_{ij} \end{bmatrix} \in \mathbb{R}^{n_{\mathcal{N}_i}\times n_{\mathcal{N}_i}}, i,j \in \mathcal{N}_i,$$

$$B_{\mathcal{N}_i} = \begin{bmatrix} B_j \end{bmatrix} \in \mathbb{R}^{n_{\mathcal{N}_i}\times m_{\mathcal{N}_i}}, j \in \mathcal{N}_i,$$

$$\Lambda_{\mathcal{N}_i} = \begin{bmatrix} Q_{\mathcal{N}_i} & 0 \\ 0 & R_{\mathcal{N}_i} \end{bmatrix} \in \mathbb{R}^{l_{\mathcal{N}_i}\times l_{\mathcal{N}_i}},$$

$$Q_{\mathcal{N}_i} = \begin{bmatrix} Q_{ij} \end{bmatrix} \in \mathbb{R}^{n_{\mathcal{N}_i}\times n_{\mathcal{N}_i}}, i,j \in \mathcal{N}_i,$$

$$R_{\mathcal{N}_i} = \begin{bmatrix} R_j \end{bmatrix} \in \mathbb{R}^{m_{\mathcal{N}_i}\times m_{\mathcal{N}_i}}, j \in \mathcal{N}_i,$$

If the interaction between the non-neighboring subsystems is not significant, one may neglect the effects of subsystems $j \notin \mathcal{N}_i$ on subsystems $k \in \mathcal{N}_i$ and derive the following approximation,

$$\begin{aligned} &\begin{bmatrix} A_{\mathcal{N}_i} & B_{\mathcal{N}_i} \end{bmatrix} D_{\mathcal{N}_i} = \\ &\begin{bmatrix} A_{\mathcal{N}_i} & B_{\mathcal{N}_i} \end{bmatrix} \begin{bmatrix} x_{\mathcal{N}_i}(k_0) & x_{\mathcal{N}_i}(k_0+1) & \cdots & x_{\mathcal{N}_i}(k_0 + l_{\mathcal{N}_i} - 1) \\ u_{\mathcal{N}_i}(k_0) & u_{\mathcal{N}_i}(k_0+1) & \cdots & u_{\mathcal{N}_i}(k_0 + l_{\mathcal{N}_i} - 1) \end{bmatrix} \\ &\approx \begin{bmatrix} x_{\mathcal{N}_i}(k_0+1) & x_{\mathcal{N}_i}(k_0+2) & \cdots & x_{\mathcal{N}_i}(k_0 + l_{\mathcal{N}_i}) \end{bmatrix} := X_{\mathcal{N}_i}. \end{aligned}$$

Then, the constraint (60) can be represented as

$$X_{\mathcal{N}_i}^T(H_{\mathcal{N}_{i,11}} - H_{\mathcal{N}_{i,12}}H_{\mathcal{N}_{i,22}}^{-1}H_{\mathcal{N}_{i,12}}^T)X_{\mathcal{N}_i} - D_{\mathcal{N}_i}^T(H_{\mathcal{N}_i} - \Lambda_{\mathcal{N}_i})D_{\mathcal{N}_i} \succeq 0,$$

which can be equivalently described as the LMI,

$$\begin{bmatrix} X_{\mathcal{N}_i}^T H_{\mathcal{N}_{i,11}} X_{\mathcal{N}_i} - D_{\mathcal{N}_i}^T(H_{\mathcal{N}_i} - \Lambda_{\mathcal{N}_i})D_{\mathcal{N}_i} & X_{\mathcal{N}_i}^T H_{\mathcal{N}_{i,12}} \\ H_{\mathcal{N}_{i,12}}^T X_{\mathcal{N}_i} & H_{\mathcal{N}_{i,22}} \end{bmatrix} \succeq 0.$$

As a direct result, Problem 8 is introduced to derive the Q-function parameters for subsystem i in a distributed manner.

Problem 8 (*Model-free SDP for Distributed Design of Distributed Controllers*)**.**

$$\underset{H_{\mathcal{N}_i}, W_{\mathcal{N}_i}}{\text{Maximize}} \ trace(W_{\mathcal{N}_i})$$

$$\text{subject to} \begin{bmatrix} H_{\mathcal{N}_{i,11}} - W_{\mathcal{N}_i} & H_{\mathcal{N}_{i,12}} \\ H_{\mathcal{N}_{i,12}}^T & H_{\mathcal{N}_{i,22}} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} X_{\mathcal{N}_i}^T H_{\mathcal{N}_{i,11}} X_{\mathcal{N}_i} - D_{\mathcal{N}_i}^T(H_{\mathcal{N}_i} - \Lambda_{\mathcal{N}_i})D_{\mathcal{N}_i} & X_{\mathcal{N}_i}^T H_{\mathcal{N}_{i,12}} \\ H_{\mathcal{N}_{i,12}}^T X_{\mathcal{N}_i} & H_{\mathcal{N}_{i,22}} \end{bmatrix} \succeq 0.$$

Problem 8 would be solved by subsystem i, and the sub-optimal controller for this subsystem is obtained as,

$$K_{\mathcal{N}_i} = -H_{\mathcal{N}_{i,22}}^{-1}H_{\mathcal{N}_{i,12}}^T.$$

5. Simulation results

In this section, the performance of the proposed model-free approach is investigated and compared to the available design methods. In Section 5.1, three different model-free control schemes, centralized control, centralized design of distributed control, and distributed design of distributed local controllers are evaluated. In Section 5.2, the efficacy of the proposed approach is tested on the LQR control of the swing dynamics of a large-scale power network. Numerical tests are implemented in CVX for Matlab (Grant & Boyd, 2014) by SDPT3 version 4.0 as the core solver.

5.1. Synthetic examples

To evaluate the performance of the proposed approach, different systems including stable and unstable dynamics, with different cost functions, and different levels of interaction between non-neighboring subsystems are tested. Nine different cases are considered whose properties are summarized in Table 1. The numerical values of the state–space matrix A, B, Q, and R in each of the nine cases are given in Appendix.

5.1.1. Centralized control

For each of the systems in Table 1, the centralized controller is designed by the model-free SDP in Problem 4 (K_{Q-LMI}), by solving the DARE using the system model ($K^*_{Ricatti}$), and by the policy iteration algorithm in Bradtke et al. (1994) (K_{Q-PI}). The performance of the centralized control in each case is evaluated by averaging the total cost for 20 different initial conditions chosen from a Gaussian distribution with zero mean and variance equal to $5I$. The results are reported in Table 2.

In the absence of any structural constraint, the model-based solution from Ricatti equation reveals the minimum achievable cost in each case. The results confirm that for the centralized case, the proposed model-free approach yields the globally optimal solution by using only a limited number of samples. It should be noted that the PI algorithm needs an initial stabilizing controller and converges after about 700 samples, while our proposed approach does not require the system to be initially stable and obtains the controller after only 9 samples. Moreover, due to the sensitivity of the PI algorithm to parameter tuning, approximate values are reported for the K_{Q-PI} controller. In theory, the state-feedback gains must reach their optimal values as the number of iterations approach infinity (Bradtke et al., 1994).

To illustrate the performance of the proposed method, consider the 9th case in Table 1 with the matrices A, B, Q, and R given in Appendix. This system is open-loop unstable with $\lambda_{max} = 1.1267$. The LQR cost involves inter-system costs and thus the matrix Q is not diagonal. The system starts from the initial condition $x(0) = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}^T$ without incorporation of a state feedback control and $l + 1 = 9$ samples are collected to build the matrices D and X_D. A very small Gaussian probing noise is added to the input signal as discussed in Section 3.2. After collection of 9 samples, Problem 4 is solved by using CVX and the optimal controller gain is updated. The state trajectories during the model-free design procedure are illustrated in Fig. 1. The optimal data-driven controller is applied at time step 10 and regulates the trajectories in an optimal manner.

Table 1
Properties of the interconnected systems in Section Section 5.1.

Case	Open-loop Stability	LQR Cost	Non-neighboring Interaction
1	Unstable	Intra-system costs	None
2			Moderate
3			Significant
4	Stable		None
5			Moderate
6			Significant
7	Unstable	inter-system costs	None
8			Moderate
9			Significant

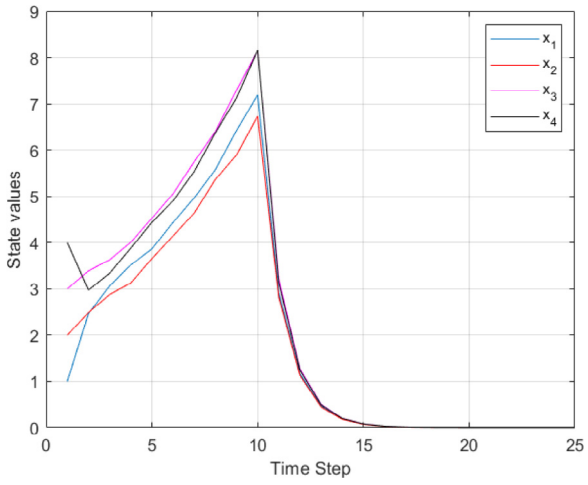

Fig. 1. State trajectories during the online centralized design process of the centralized controller.

5.1.2. Centralized design of distributed controllers

For each of the systems in Table 1, a distributed control system is designed by the model-based optimization in Problem 5

Table 2
Performance of the centralized controllers by the model-free Problem 4 (K_{Q-LMI}), the DARE ($K^*_{Ricatti}$), and the adaptive PI algorithm (Bradtke et al., 1994) (K_{Q-PI}).

Case	K_{Q-LMI}	$K^*_{Ricatti}$	K_{Q-PI}
1	180.5313	180.5313	≈ 199
2	157.0710	157.0710	≈ 166
3	152.4436	152.4436	≈ 159
4	149.6543	149.6543	≈ 156
5	149.2763	149.2763	≈ 152
6	145.3031	145.3031	≈ 149
7	489.2740	489.2740	≈ 502
8	457.7161	457.7161	≈ 465
9	450.7323	450.7323	≈ 457

Table 3
Performance of the distributed controllers by Problem 5 (K^*_{Di}), Problem 6 ($K_{DiCQLMI}$), and the optimal centralized controller ($K^*_{Ricatti}$).

Case	K^*_{Di}	$K_{DiCQLMI}$	$K^*_{Ricatti}$
1	180.7893	180.7894	180.5313
2	163.0718	163.0723	157.0710
3	166.5080	166.5280	152.4436
4	149.6770	149.6771	149.6543
5	152.3677	152.3673	149.2763
6	150.5624	150.5625	145.3031
7	490.6976	490.7150	489.2740
8	461.1610	461.1039	457.7161
9	465.9966	465.9563	450.7323

(K^*_{Di}) and model-free optimization in Problem 6 ($K_{DiCQLMI}$), respectively. The distributed control topology is shown in Fig. 2. The performance of the distributed controllers for K^*_{Di} and $K_{DiCQLMI}$ is evaluated by averaging the total cost for 20 different initial conditions chosen from a Gaussian distribution with zero mean and variance equal to $5I$. The results are reported in Table 3 and also compared to the centralized best achievable performance of Riccati equations reported in the last column ($K^*_{Ricatti}$). The results indicate that the proposed model-based and model-free approaches offer the same levels of performance, which confirms the theoretical argument in Section 4. Moreover, while 8 communication links out of 12 possible links between the subsystems are eliminated, the performance of the distributed control is similar to that of the optimal centralized controller with 12 communication links in all cases. According to Table 3, the data-driven distributed controllers in cases 1, 4, and 7 yield nearly the same level of performance as their centralized counterparts. The main feature of these three cases is that the matrix A and the control system have the same structure. The most deviation from the centralized performance belongs to the distributed controllers in cases 3, 6, 9, where the dynamic interactions between non-neighboring subsystems are significant.

It should be noted that the arbitrary structure of the controller makes the problem NP-hard even when the dynamics of the systems are completely known. Again, our proposed model-free approach does not require the system to be initially stable and obtains the sub-optimal controller after only 9 samples.

To illustrate the performance of the distributed controllers, consider again the 9th case in Table 1. As reported in Table 3, the 9*th* case corresponds to the most deviation of performance with respect to the centralized controllers. Its dynamic interactions between non-neighboring subsystems are significant and the matrix

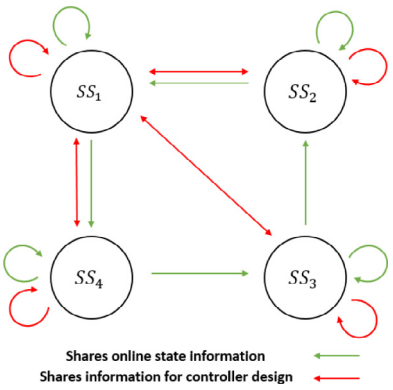

Fig. 2. Distributed control and centralized design in Section 5.1.2.

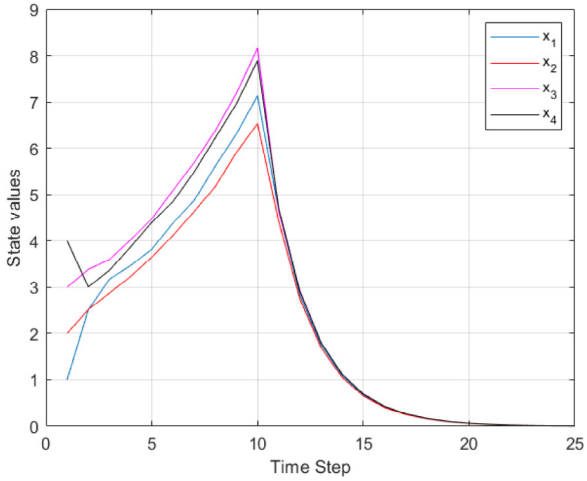

Fig. 3. State trajectories during the online centralized design process of the distributed controller.

Q is not diagonal. Similar to the previous case study, the unstable system starts from the initial condition $x(0) = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}^T$ without incorporation of a state feedback control and $l+1=9$ samples are collected by a central unit (subsystem 1 is regarded as the central entity in this example). Problem 6 is solved and the sub-optimal control gains are sent to each subsystem. Fig. 3 shows the state trajectories during the online model-free design of the sub-optimal controllers for each subsystem. Comparing the responses of the controlled system in Fig. 3 with the centralized controller in Fig. 1 confirms superior performance of the centralized control at the expense of 8 additional communication links.

5.1.3. Distributed design of distributed control

Again, for each of the systems in Table 1, a distributed control system is designed based on the control and design structure given by I_S. In this case, as shown in Fig. 4, each subsystem design its own controller by using the state and input data shared by the neighboring subsystems.

In each case, the distributed controller is designed by the model-free optimization in Problem 8 ($K_{DiCQLMI}$) and also using a distributed PI algorithm recently introduced in Görges (2019) (K_{DiALQR}). The performance of the distributed controllers is evaluated by averaging the total cost for 20 different initial conditions chosen from a Gaussian distribution with zero mean and variance equal to $5I$. The results are reported in Table 4 and also compared to the centralized best achievable performance of Riccati equations reported in the last column ($K^*_{Ricatti}$). Unlike our proposed

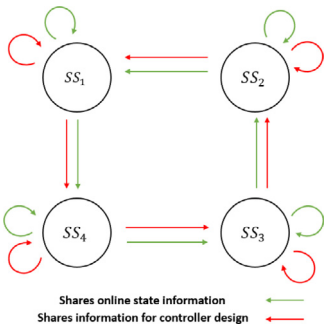

Fig. 4. Distributed design and distributed control topology in Section 5.1.3.

Table 4
Comparison between distributed controllers obtained by Problem 8 (K_{DiQLMI}), the distributed adaptive LQR controller introduced in Görges (2019), (K_{DiALQR}), and the optimal centralized controller ($K^*_{Ricatti}$), and.

Case	K_{DiQLMI}	K_{DiALQR}	$K^*_{Ricatti}$
1	≈ 205	≈ 215	180.5313
2	≈ 175	≈ 190	157.0710
3	≈ 220	≈ 205	152.4436
4	≈ 160	≈ 160	149.6543
5	≈ 190	≈ 165	149.2763
6	≈ 175	≈ 155	145.3031
7	≈ 540	inf	489.2740
8	≈ 515	inf	457.7161
9	≈ 600	inf	450.7323

controller (K_{DiQLMI}), K_{DiALQR} could not converge to a controller gain for systems 7, 8, and 9 where the LQR problem consists of inter-subsystem costs. In other cases, the results for K_{DiALQR} indicate some levels of sensitivity to the learning parameters. For this reason, we have performed learning parameter tuning to achieve reliable results.

Once again, the performance of the distributed design of distributed controllers is investigated for the 9th case in Table 1. The system starts from the initial condition $x(0) = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}^T$ without incorporation of a state feedback control and $l_{\mathcal{N}_i} + 1 = 5$ samples are collected by each subsystem. Problem 8 is solved by each subsystem individually and the sub-optimal control gains are updated.

Fig. 5 shows the state trajectories during an attempt to design the distributed controller using Problem 8 in each subsystem. Comparing the state trajectories of the controlled system in Fig. 5 with Figs. 1 and 3 indicates mild performance degradation when the design procedure is performed in a completely distributed manner, and design of each sub-controller relies on only partial observations.

5.2. IEEE 39-bus New England power grid

The proposed approach is utilized for the LQR control of the swing dynamics of a large-scale power network. The system under study is the IEEE 39-bus New England power grid model. By removing higher order generator dynamics the classical synchronous machine model for each node of the grid is described by the swing equation, Sauer, Pai, and Chow (2017), $m_i\ddot{\theta} + d_i\dot{\theta} = P_{mi} - P_{ei}$, where m_i and d_i denote the inertia and damping coefficients of the generator at the ith bus, respectively. θ_i is the rotor angle of the ith generator. P_{mi} and P_{ei} represent the mechanical power and active power associated with generator i. The active

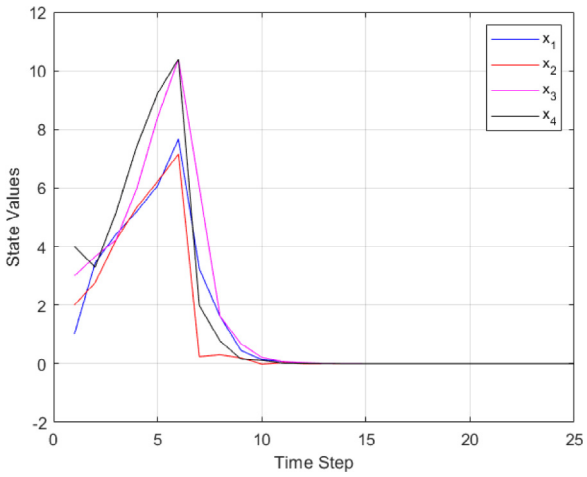

Fig. 5. State trajectories during the model-free distributed design of the distributed controllers.

power flow P_{ei} can be expressed in terms of neighboring phase angles as,

$$P_{ei} = \sum_{j=1}^{n} k_{ij} \sin(\theta_i - \theta_j), \tag{61}$$

in which k_{ij} is the susceptance of the connecting line between buses i and j. Assuming that the phase angles are sufficiently close, it is conventional to use the linearized approximation of the dynamic system (61) represented as $m_i\ddot{\theta} + d_i\dot{\theta} = -\sum_{j=1}^{n} k_{ij}(\theta_i - \theta_j) + u_i$. u_i is the control signal, i.e., the mechanical power applied to the generator. The set of these linear equations can be written as the more compact form $M\ddot{\theta} + D\dot{\theta} + L\theta = u$, where, M and D are the diagonal matrices of inertia and damping coefficients, and L is the Kron-reduced Laplacian matrix with off-diagonals. The Laplacian matrix L shows all of the interactions among generator subsystems. Accordingly, the state–space description of the generator swing dynamics are obtained using Power Systems Toolbox (Chow & Cheung, 1992) as

$$\dot{x} = Ax(t) + Bu(t), \tag{62}$$

in which $x(t) = \begin{bmatrix}\theta_1 & \dot{\theta}_1 & \cdots & \theta_9 & \dot{\theta}_9\end{bmatrix}^T \in \mathbb{R}^{18}$ represents the states of the interconnected system and $u(t) = \begin{bmatrix}u_1 & u_2 & \cdots & u_9\end{bmatrix} \in \mathbb{R}^9$ denotes the generators' mechanical power. The system dynamic is discretized with a sample time of 0.05 s. The goal of LQR control is to remove the effects of input disturbances on the rotor's angle and frequency of all the generators in the interconnected system. The weighting matrices $Q = I_{18}$ and $R = I_9$ serve as the LQR cost matrices, and a step disturbance with a large amplitude of 10 p.u. is added to the first generator's mechanical power input at time step 90 for 10 time instances. The performances of the three control schemes proposed in this paper are investigated.

5.2.1. Centralized control

In this case, it is assumed that all of the subsystems are capable of online communications with each other to share their state information for control. Each generator uses a local controller to regulate its frequency. In the learning phase, every subsystem records its state and input trajectories for $l + 1 = 28$ samples and sends them to the central processor. A very small Gaussian probing noise is added to the input of each generator during this period. In this scenario, a centralized controller is designed using the model-free optimization-based Q-learning algorithm in Section 3.2 and is used at every subsystem afterward.

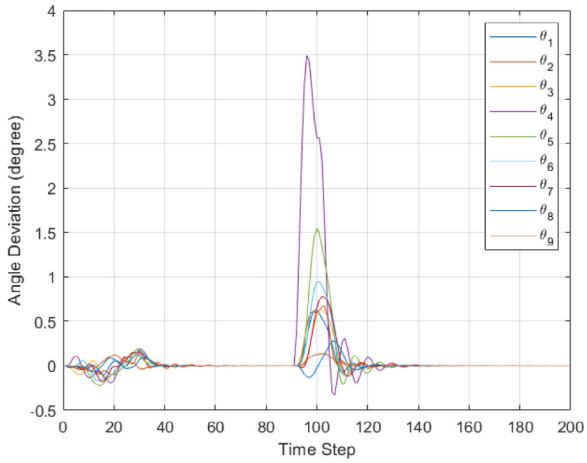

Fig. 6. Rotor angle deviation in centralized control.

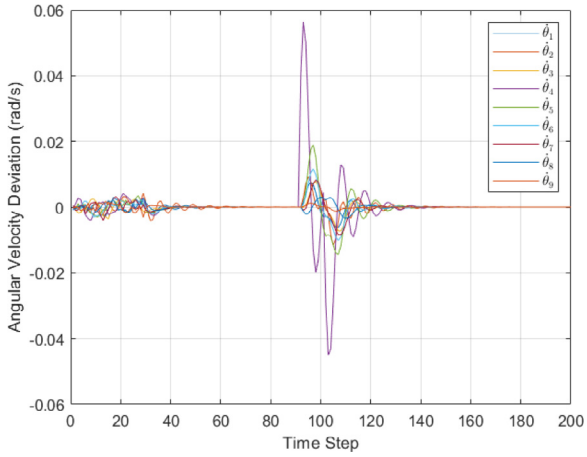

Fig. 7. Rotor angular velocity deviation in centralized control.

Figs. 6 and 7 illustrate the angle and angular velocity trajectories of all generators, respectively. The results show that the additive disturbance at time step 90 is successfully rejected by the centralized data-driven LQR control.

5.2.2. Centralized design of distributed control

In this scenario, the subsystems are assumed to have access to the states of their neighboring subsystems according to Fig. 8 while a central system has access to the state and input trajectories of all subsystems for model-free design. $l + 1 = 28$ samples are required to be collected by the central processor for the distributed controller design. The subsystems use this optimal controller after that point. Figs. 9 and 10 illustrate the angle and angular velocity of all generators, respectively. It can be seen that the rotors experience more fluctuations and regulation requires more time as compared to the centralized controller, which is expected as the detrimental effect of elimination of some feedback links in the control structure.

5.2.3. Distributed design of distributed control

Fig. 11 depicts the control and design topology of the third scenario, where each subsystem is required to design its own controller by using the state and input trajectories of its neighboring subsystems. Every subsystem i uses a local stabilizing controller at first. After collecting $l_{\mathcal{N}_i}$ data samples of input and state trajectories of subsystems $j \in \mathcal{N}_i$, subsystem i solves Problem 8 and

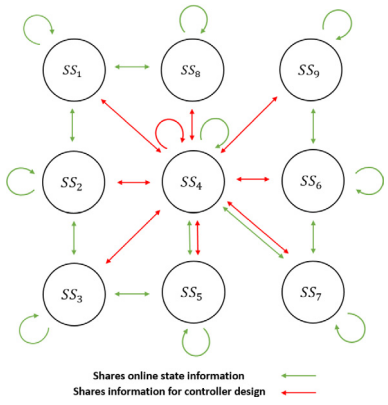

Fig. 8. Centralized design and distributed control topology for the New England power network.

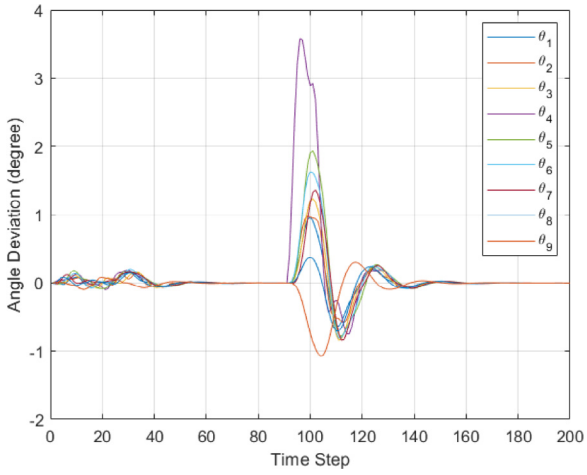

Fig. 9. Rotor angle deviation for the centralized design of distributed control.

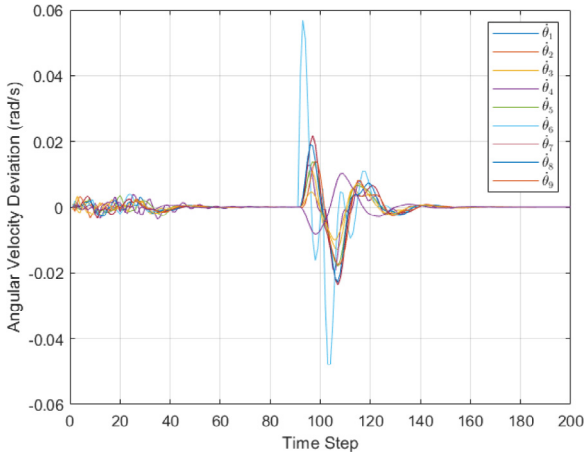

Fig. 10. Rotor angular velocity for the centralized design of distributed control.

uses the derived sub-optimal controller afterward. Figs. 12 and 13 illustrate the angle deviation and angular velocity deviation of the generators in the network, respectively. Comparing the responses of the controlled system with two previous designs indicates noticeable performance degradation when the design procedure is performed in a completely distributed manner.

Table 5
The matrix A for each of the nine synthetic examples in Section 5.1.

Case	1	2	3	4	5
A	$\begin{bmatrix} 0.2 & 0.7 & 0 & 0 \\ 0 & 0.4 & 0.7 & 0 \\ 0 & 0 & 0.5 & 0.8 \\ 0.7 & 0 & 0 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0.46 & 0.11 & 0.12 \\ 0.11 & 0.36 & 0.35 & 0.12 \\ 0.13 & 0.13 & 0.48 & 0.49 \\ 0.47 & 0.12 & 0.12 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.2 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.3 & 0.4 & 0 & 0 \\ 0 & 0.3 & 0.3 & 0 \\ 0 & 0 & 0.4 & 0.4 \\ 0.4 & 0 & 0 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.3 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.4 & 0.1 & 0.1 & 0.4 \end{bmatrix}$

Case	6	7	8	9
A	$\begin{bmatrix} 0.25 & 0.35 & 0.15 & 0.14 \\ 0.15 & 0.25 & 0.25 & 0.15 \\ 0.14 & 0.14 & 0.34 & 0.33 \\ 0.34 & 0.14 & 0.14 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.2 & 0.7 & 0 & 0 \\ 0 & 0.4 & 0.7 & 0 \\ 0 & 0 & 0.5 & 0.8 \\ 0.7 & 0 & 0 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0.46 & 0.11 & 0.12 \\ 0.11 & 0.36 & 0.35 & 0.12 \\ 0.13 & 0.13 & 0.48 & 0.49 \\ 0.47 & 0.12 & 0.12 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.2 & 0.4 \end{bmatrix}$

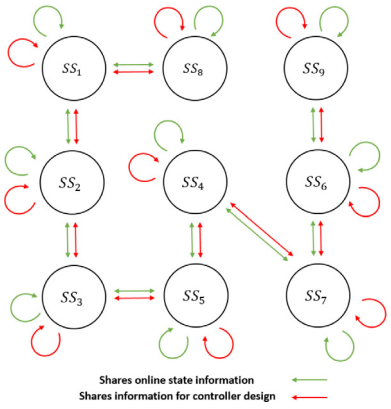

Fig. 11. Distributed design and distributed control topology for the New England power network.

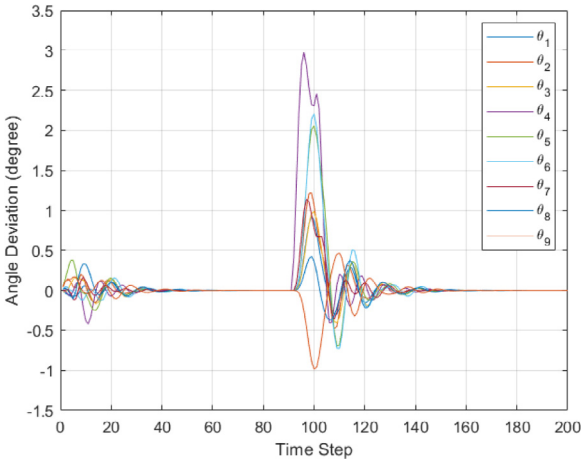

Fig. 12. Rotor angle deviation for the distributed design of distributed control.

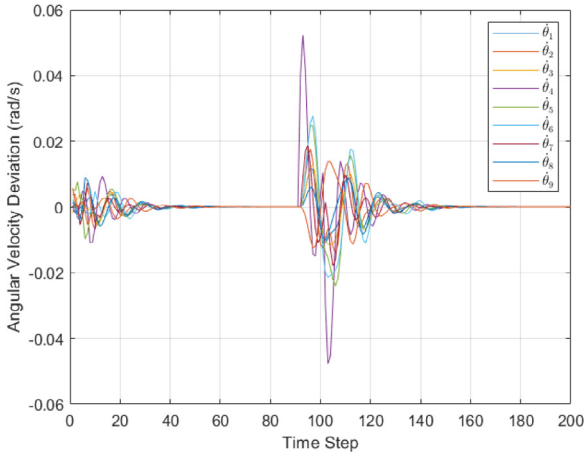

Fig. 13. Rotor angular velocity deviation for the distributed design of distributed control.

6. Conclusion

In this paper, a new off-policy model-free approach for solving the LQR problem is developed. The design is based on non-iterative semi-definite programs with linear matrix inequality constraints. It is sample-efficient, inherently robust to model uncertainties, and does not require an initial stabilizing controller. The proposed model-free approach is amenable to extend to distributed control of interconnected systems, as well. The performances of centralized and distributed strategies are evaluated based on the total cost of the LQR problem for a set of synthetic interconnected systems including stable and unstable dynamics, with different cost functions, and different levels of interaction between non-neighboring subsystems. Moreover, the proposed method is evaluated for the New England IEEE 39-bus system. The results confirm that the proposed method delivers optimality of the centralized model-free controller and the satisfactory performance of the distributed controllers with much better consistency. Future works focus on using the proposed framework for the distributed learning of controllers by using dual decomposition, analysis of stability in distributed model-free design, and extending the proposed approach to the relevant criteria in classic control theory such as H_2 and H_∞ performance indices.

Appendix

The state–space matrix A for each of the nine synthetic cases in Section 5.1 are provided in Table 5. The state–space matrix B and the weighting matrix R in all cases are $B = I_4$, and $R = I_4$. The weighting matrix Q for the first six cases are $Q = I_4$, while for the last three cases (with inter-system costs) the Q matrix is given as,

$$Q = \begin{bmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}.$$

References

Alemzadeh, Siavash, & Mesbahi, Mehran (2019). Distributed Q-learning for dynamically decoupled systems. In *2019 American control conference (ACC)* (pp. 772–777).

Azzollini, Ilario Antonio, Yu, Wenwu, Yuan, Shuai, & Baldi, Simone (2020). Adaptive leader-follower synchronization over heterogeneous and uncertain networks of linear systems without distributed observer. *IEEE Transactions on Automatic Control*.

Babazadeh, Maryam (2021). Regularization for optimal sparse control structures: a primal-dual framework. In *2021 American control conference (ACC)*. IEEE.

Babazadeh, Maryam, & Nobakhti, Amin (2017). Sparsity promotion in state feedback controller design. *IEEE Transactions on Automatic Control*, *62*(8), 4066–4072.

Bellman, Richard (1957). *Dynamic programming*. Princeton University Press.

Boyd, Stephen, El Ghaoui, Laurent, Feron, Eric, & Balakrishnan, Venkataramanan (1994). *Linear matrix inequalities in system and control theory*. SIAM.

Boyd, Stephen P., & Vandenberghe, Lieven (2004). *Convex optimization*. Cambridge University Press.

Bradtke, Steven J., Ydstie, B. Erik, & Barto, Andrew G. (1994). Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American control conference, Vol. 3* (pp. 3475–3479).

Buşoniu, Lucian, de Bruin, Tim, Tolić, Domagoj, Kober, Jens, & Palunko, Ivana (2018). Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, *46*, 8–28.

Chow, Joe H., & Cheung, Kwok W. (1992). A toolbox for power system dynamics and control engineering education and research. *IEEE Transactions on Power Systems*, *7*(4), 1559–1564.

De Persis, Claudio, & Tesi, Pietro (2019). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, *65*(3), 909–924.

Görges, Daniel (2019). Distributed adaptive linear quadratic control using distributed reinforcement learning. *IFAC-PapersOnLine*, *52*(11), 218–223.

Grant, Michael, & Boyd, Stephen (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Jha, Sumit Kumar, Roy, Sayan Basu, & Bhasin, Shubhendu (2017). Data-driven adaptive LQR for completely unknown LTI systems. *IFAC-PapersOnLine*, *50*(1), 4156–4161.

Jha, Sumit Kumar, Roy, Sayan Basu, & Bhasin, Shubhendu (2018). Direct adaptive optimal control for uncertain continuous-time LTI systems without persistence of excitation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *65*(12), 1993–1997.

Lee, Donghwan, & Hu, Jianghai (2019). Primal-dual Q-learning framework for LQR design. *IEEE Transactions on Automatic Control*, 3756–3763.

Lewis, Frank L., & Vrabie, Draguna (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, *9*, 32–50.

Lin, Fu, Fardad, Makan, & Jovanović, Mihailo R. (2013). Design of optimal sparse feedback gains via the alternating direction method of multipliers. *IEEE Transactions on Automatic Control*, *58*(9), 2426–2431.

Mania, Horia, Guy, Aurelia, & Recht, Benjamin (2018). Simple random search provides a competitive approach to reinforcement learning. arXiv preprint arXiv:1803.07055.

Rami, Mustapha Ait, & Zhou, Xun Yu (2000). Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. *IEEE Transactions on Automatic Control*, *45*(6), 1131–1143.

Sauer, Peter W., Pai, M. A., & Chow, Joe H. (2017). *Power systems dynamics and stability*. Wiley-IEEE Press.

da Silva, Gustavo R. Gonçalves, Bazanella, Alexandre S., Lorenzini, Charles, & Campestrini, Lucíola (2019). Data-driven LQR control design. *IEEE Control Systems Letters*, *3*(1), 180–185.

Strang, Gilbert (1968). *Linear algebra and its applications*.

Sutton, Richard S., & Barto, Andrew G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Tu, Stephen, & Recht, Benjamin (2018). Least-squares temporal difference learning for the linear quadratic regulator. In *International conference on machine learning* (pp. 5005–5014).

Umenberger, Jack, & Schön, Thomas B. (2018). Learning convex bounds for linear quadratic control policy synthesis. In *Advances in neural information processing systems* (pp. 9561–9572).

Willems, Jan C., Rapisarda, Paolo, Markovsky, Ivan, & De Moor, Bart L. M. (2005). A note on persistency of excitation. *Systems & Control Letters*, *54*(4), 325–329.

Yu, Wenwu, DeLellis, Pietro, Chen, Guanrong, Di Bernardo, Mario, & Kurths, Jürgen (2012). Distributed adaptive control of synchronization in complex networks. *IEEE Transactions on Automatic Control*, *57*(8), 2153–2158.

Milad Farjadnasab received the B.Sc. in Electrical Engineering in 2018, and M.Sc. in Control Engineering in 2020, both from Sharif University of Technology, Iran. He is currently a Ph.D. student at the Department of Electrical and Computer Engineering at McMaster University, where his research is focused on cognitive control systems for mobile robots.

Maryam Babazadeh received her B.Sc., M.Sc. and Ph.D. degrees from Electrical Engineering Department, Sharif University of Technology, Iran, in the field of Control Systems in 2009, 2011 and 2016, respectively. Since 2017 she has been a faculty member in the Department of Electrical Engineering at Sharif university of Technology. Her research interests focus on control structures, distributed control, reinforcement learning and optimization algorithms.