

# Model-free stochastic linear quadratic control for discrete-time systems with multiplicative and additive noises via semidefinite programming<sup>★</sup>

Jing Guo <sup>a,b</sup>, Xiushan Jiang <sup>c,\*</sup>, Weihai Zhang <sup>a</sup>

<sup>a</sup>*College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China*

<sup>b</sup>*School of Mathematics and Statistics, Shandong University of Technology, Zibo 250049, China*

<sup>c</sup>*College of New Energy, China University of Petroleum (East China), Qingdao 266580, China*

---

## Abstract

This paper investigates a model-free solution to the stochastic linear quadratic regulation (LQR) problem for linear discrete-time systems with both multiplicative and additive noises. We formulate the stochastic LQR problem as a nonconvex optimization problem and rigorously analyze its dual problem structure. By exploiting the inherent convexity of the dual problem and analyzing Karush-Kuhn-Tucker conditions with respect to optimality in convex optimization, we establish an explicit relationship between the optimal point of the dual problem and the parameters of the associated Q-function. This theoretical insight, combined with the technique of the matrix direct sum, makes it possible to develop a novel model-free sample-efficient, non-iterative semidefinite programming algorithm that directly estimates optimal control gain without requiring an initial stabilizing controller, or noises measurability. The robustness of the model-free SDP method to errors is investigated. Our approach provides a new optimization-theoretic framework for understanding Q-learning algorithms while advancing the theoretical foundations of reinforcement learning in stochastic optimal control. Numerical validation on a pulse-width modulated inverter system demonstrates the algorithm's effectiveness, particularly in achieving a single-step non-iterative solution without hyper-parameter tuning.

**Key words:** Convex optimization; Model-free design; Multiplicative and additive noises; Reinforcement learning; Semidefinite programming.

---

## 1 Introduction

Optimal control (Lewis, Vrabie, & Syrmos, 2012; Yong & Zhou, 2012) is an important branch of modern control theory, which aims to find the optimal control policy for dynamic system by optimising the performance index. The linear quadratic regulation (LQR) problem, which has a wide engineering background, is a typical optimal control problem, and its quadratic performance index is actually a comprehensive consideration of the transient performance, steady state performance and control energy constraints of the system in classical control theory. Such a problem can be solved by such well-established

methods as Bellman's dynamic programming (Bellman, 1966; Bertsekas, 2019). With the development of convex optimization and semidefinite programming (SDP) (Boyd & Vandenberghe, 2004; Vandenberghe & Boyd, 1996), many research works have reinvestigated the LQR problem from the perspective of convex optimization, see (Yao, Zhang, & Zhou, 2001) and the references therein. The above methods for solving LQR usually require complete system information. And in fact, accurate dynamical models of most practical complex systems are difficult to obtain. To avoid the requirement of system dynamics in controller designs, reinforcement learning (RL) (Sutton & Barto, 1998; Bertsekas, 2019), a sub-field of machine learning that tackles the problem of how an agent learns optimal policy to minimize the cumulative cost by interacting with unknown environments, has been applied to the LQR problem for deterministic systems. Studies on the design of model-free controllers for LQR by means of the methods of RL, such as temporal

---

<sup>★</sup> This paper was not presented at any IFAC meeting.

\* Corresponding author.

Email addresses: [guojing8299@163.com](mailto:guojing8299@163.com) (Jing Guo), [x\\_sjiang@163.com](mailto:x_sjiang@163.com) (Xiushan Jiang), [w\\_hzhang@163.com](mailto:w_hzhang@163.com) (Weihai Zhang).

difference (Sutton & Barto, 1998), Q-learning (Watkins & Dayan, 1992) and so on, can be found in (Bradtko, Ydstie, & Barto, 1994; Lewis & Vrabie, 2009). However, most of the current RL methods are weakly extendible or lacks in theoretical assurances. Recently, many researchers have reconsidered the LQR problem from the viewpoint of optimization by connecting the RL method with classical convex optimization, for example, according to Lagrangian duality theory, a novel primal-dual Q-learning framework for LQR was established and a model-free algorithm to solve the LQR problem was designed in (Lee & Hu, 2018), from the viewpoint of primal-dual optimization. In (Farjadnasab & Babazadeh, 2022), a new model-free algorithm was proposed on the basis of the properties of optimization frameworks.

Due to the wide application of stochastic systems, many researchers have started to focus their attentions on applying RL methods to stochastic LQR problems. For linear discrete-time stochastic systems in the presence of additive noise, (Pang & Jiang, 2021) proposed an optimistic least squares policy iteration algorithm. (Li, Qin, Zheng, Wang, & Kang, 2022) studied the model-free design of stochastic LQR controllers from the perspective of primal-dual Q-learning optimization. For linear discrete-time systems with multiplicative noise, (Gravell, Esfahani, & Summers, 2020) proposed a policy gradient algorithm. For stochastic systems with both multiplicative and additive noises, (Lai, Xiong, & Shu, 2023) employed Q-learning to implement the model-free optimal control. Based on an unbiased estimator and an initial stabilizing controller, (Y. Jiang, Liu, & Feng, 2024) developed a data-driven value iteration algorithm using online data. In addition, many effective model-free RL algorithms have been proposed and convergence analyzed for stochastic optimal control problems associated with various dynamical systems, see (Pang & Jiang, 2023; Zhang, Guo, & Jiang, 2025; X. Jiang, Wang, Zhao, & Shi, 2024) for details.

Inspired by the above studies, this paper investigates the stochastic LQR problem for a class of discrete-time stochastic systems subject to both multiplicative and additive noises, and a novel model-free SDP method for solving the stochastic LQR problem is proposed. Compared with the existing work, the main contributions of this paper can be summarized as follows.

1) By using the recursive relationship satisfied by the vectorized covariance matrix of the augmented states, the stochastic LQR problem is equivalently formulated as a nonconvex optimization problem. Instead of proving strong duality, we start directly from the dual problem of the stochastic LQR problem, and then find out the relationship between the optimal point of the dual problem and the parameters of the Q-function by using the convexity of the dual problem and the Karush-Kuhn-Tucker (KKT) conditions.

2) By introducing auxiliary matrices and utilizing the technique of the matrix direct sum, the constraint of an optimization problem containing terms related to multiplicative noise is represented in the form of linear matrix inequality (LMI) constraint required for standard SDP. This cannot be achieved solely through the Schur complement property.

3) Based on Monte-Carlo method, a model-free implementation of the stochastic LQR controller design is given. A robust analysis of the model-free algorithm is presented by transforming errors into linear system transfer problems via dual variables. By integrating closed-loop matrix minimum singular value, the analysis unifies error boundedness and system stability, explicitly quantifying the relationship between key factors (such as sample complexity and data continuity) and errors.

The proposed model-free SDP algorithm has the following characteristics that are worth noting:

- It eliminates the need for an initial stabilizing controller – a common limitation in existing RL methods.
- It does not require hyper-parameter tuning.
- It does not require the multiplicative and additive noises to be measurable.
- The algorithm implementation procedure is done in a single step, but not in an iterative form.
- It only needs to collect input and state information over a short length of time.

Notations:  $\mathcal{R}$  is the set of real numbers and Euclidean spaces;  $\mathcal{Z}_+$  is the set of nonnegative integers;  $\mathcal{R}^n$  is the set of  $n$ -dimensional vectors;  $\mathcal{R}^{n \times m}$  is the set of  $n \times m$  real matrices.  $\delta_{ij}$  is the Kronecker delta function, i.e.,  $\delta_{ij} = 1$  when  $i = j$  and  $\delta_{ij} = 0$  when  $i \neq j$ .  $I_n$  denotes the identity matrix in  $\mathcal{R}^{n \times n}$ ,  $0$  denotes the zero vector or matrix with the appropriate dimension.  $\otimes$  denotes the Kronecker product.  $\|\cdot\|_F$  denotes the Frobenius norm;  $\|\cdot\|_2$  denotes the Euclidean norm for vectors and the spectral norm for matrices.  $\text{col}(x, y)$  denotes the column vector consisting of vectors  $x$  and  $y$ . The transpose of a matrix or vector  $M$  is represented by  $M^\top$ ; the trace of a square matrix  $M$  is denoted as  $\text{Tr}(M)$ ; spectral radius of matrix  $M$  is denoted as  $\rho(M)$ ; the minimum singular value of matrix  $M$  is denoted as  $\sigma_{\min}(M)$ ; the maximum and minimum eigenvalues of a real symmetric matrix  $M$  are denoted by  $\lambda_{\max}(M)$  and  $\lambda_{\min}(M)$ , respectively. For matrix  $M \in \mathcal{R}^{m \times n}$ , define  $\text{vec}(M) = \text{col}(M_1, M_2, \dots, M_n)$ , where  $M_j$  is the  $j$ th column of matrix  $M$ .  $\text{vec}^{-1}(\cdot)$  is an operation such that  $M = \text{vec}^{-1}(\text{vec}(M))$ . The direct sum of matrices  $M_1$  and  $M_2$  is defined as  $M_1 \oplus M_2 \triangleq \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$ .

$\mathbb{R}(M)$  denotes the range space of  $M$ ;  $\mathbb{N}(M)$  denotes the kernel space of  $M$ .  $\mathcal{S}^n$ ,  $\mathcal{S}_+^n$  and  $\mathcal{S}_{++}^n$  are the sets

of all  $n \times n$  symmetric, symmetric positive semidefinite and symmetric positive definite matrices, respectively; we write  $M \succeq 0$ , (resp.  $M \succ 0$ ) if  $M \in \mathcal{S}_+^n$ , (resp.  $M \in \mathcal{S}_{++}^n$ ). For  $H \in \mathcal{S}^n$ , define  $\text{vecs}(H) \triangleq [h_{11}, \sqrt{2}h_{12}, \dots, \sqrt{2}h_{1n}, h_{22}, \sqrt{2}h_{23}, \dots, \sqrt{2}h_{(n-1)n}, h_{nn}]^\top$ , where  $h_{ik}$  is the  $(i, k)$ th element of matrix  $H$ .

## 2 Problem formulation and preliminaries

Consider the following general linear discrete-time stochastic system with both multiplicative and additive noises

$$x_{k+1} = Ax_k + Bu_k + (A_1x_k + B_1u_k)v_k + w_k, \quad (1)$$

where  $k \in \mathcal{Z}_+$  is the discrete-time index,  $x_k \in \mathcal{R}^n$ ,  $u_k \in \mathcal{R}^m$ ,  $v_k \in \mathcal{R}^1$  and  $w_k \in \mathcal{R}^n$  are the system state, control input, system multiplicative noise and system additive noise, respectively.  $A, A_1 \in \mathcal{R}^{n \times n}$  and  $B, B_1 \in \mathcal{R}^{n \times m}$  are the system coefficient matrices. The system noise sequence  $\{(w_k, v_k), k \in \mathcal{Z}_+\}$  is defined on a complete probability space  $(\Omega, \mathcal{F}, \mathcal{P})$ . The initial state  $x_0$  is assumed to be a random variable with mean vector  $\mu_0$  and positive definite covariance matrix  $\Sigma_0$ . For convenience, it is further assumed that the following Assumption 1 holds.

**Assumption 1** Assume that

- (i)  $\{(w_k, v_k), k \in \mathcal{Z}_+\}$  is independent of  $x_0$ ;
- (ii) The sequence of random vectors  $\{w_k, k \in \mathcal{Z}_+\}$  is independent and identically distributed (i.i.d.) with mean zero and covariance matrix  $\Sigma$ , where  $\Sigma$  is bounded and positive definite;
- (iii) The sequence of random variables  $\{v_k, k \in \mathcal{Z}_+\}$  is i.i.d. with mean zero and variance  $\sigma$ , where  $0 < \sigma < \bar{\sigma}$ ;
- (iv)  $\{w_k, k \in \mathcal{Z}_+\}$  and  $\{v_k, k \in \mathcal{Z}_+\}$  are mutually independent.

**Definition 1** (Kubrusly & Costa, 1985) System (1) with  $u_k \equiv 0$  is called mean square stable (MSS) if for any initial state  $x_0$  and system noise sequence  $\{(w_k, v_k), k \in \mathcal{Z}_+\}$  satisfying Assumption 1, there exist  $\mu \in \mathcal{R}^n$  and  $X \in \mathcal{R}^{n \times n}$  which are independent of  $x_0$ , such that

- (i)  $\lim_{k \rightarrow \infty} \|\mathbb{E}(x_k) - \mu\|_2 = 0$ ,
- (ii)  $\lim_{k \rightarrow \infty} \|\mathbb{E}(x_k x_k^\top) - X\|_2 = 0$ .

**Definition 2** System (1) with the feedback control policy  $u_k = Lx_k$  is called stabilizable, if the closed-loop system  $x_{k+1} = (A + BL)x_k + (A_1 + B_1L)x_kv_k + w_k$  is MSS. In this case, the feedback control policy  $u_k = Lx_k$  is called to be admissible,  $L \in \mathcal{R}^{m \times n}$  is called a (mean square) stabilization gain of system (1), and the corresponding set of all (mean square) stabilizing state feedback gains is denoted as  $\mathcal{L}$ .

Consequently, the cost functional can be defined as

$$J(L, x_0) \triangleq \sum_{k=0}^{\infty} \alpha^k \mathbb{E} \begin{bmatrix} x_k \\ Lx_k \end{bmatrix}^\top W \begin{bmatrix} x_k \\ Lx_k \end{bmatrix}, \quad (2)$$

where  $\alpha \in (0, 1)$  is the discount factor and  $W \triangleq Q \oplus R$  is a block-diagonal matrix including the state and input weighting matrices  $Q \in \mathcal{S}_+^n$  and  $R \in \mathcal{S}_{++}^m$ , respectively.

In this paper, we consider the infinite-horizon stochastic LQR problem.

### Problem 1 (Stochastic LQR Problem)

Find an optimal feedback gain  $L^* \in \mathcal{L}$ , if it exists, that minimizes the cost functional (2). That is,

$$L^* \triangleq \arg \min_{L \in \mathcal{L}} J(L, x_0).$$

The following assumptions are necessary to ensure that an optimal feedback gain  $L^*$  exists.

**Assumption 2** Assume that

- (i) System (1) is (mean square) stabilizable.
- (ii)  $(A, A_1 | \sqrt{Q})$  is exactly detectable (Zhang, Xie, & Chen, 2017).

Under Assumption 2, the optimal value of  $\inf_{L \in \mathcal{L}} J(L, x_0)$  exists and is attained, and the corresponding optimal cost  $J(L^*, x_0)$ , which is abbreviated as  $J^*$ , is given by (Lai et al., 2023)

$$J^* = \mathbb{E} (x_0^\top P^* x_0) + \frac{\alpha}{1 - \alpha} \text{Tr}(P^* \Sigma),$$

where  $P^* \succeq 0$  is the unique solution to the following discrete-time generalized algebraic Riccati equation (DGARE):

$$P = \mathcal{R}(P), \quad (3)$$

where

$$\begin{aligned} & \mathcal{R}(P) \\ & \triangleq Q + \alpha A^\top PA + \alpha \sigma A_1^\top PA_1 - \alpha^2 (A^\top PB + \sigma A_1^\top PB_1) \\ & \quad (R + \alpha B^\top PB + \alpha \sigma B_1^\top PB_1)^{-1} (B^\top PA + \sigma B_1^\top PA_1). \end{aligned}$$

The corresponding optimal control gain  $L^*$  for the stochastic LQR problem is given by

$$\begin{aligned} L^* = & -\alpha (R + \alpha B^\top P^* B + \alpha \sigma B_1^\top P^* B_1)^{-1} \\ & \times (B^\top P^* A + \sigma B_1^\top P^* A_1). \end{aligned}$$

It is clear from the above that the stochastic LQR problem can be solved by using the knowledge of the system dynamics. The following Q-learning offers a model-free solution for solving the stochastic LQR problem. The Q-function is defined as (see (Lai et al., 2023) for more details),

$$Q(x_k, u_k) \triangleq \mathbb{E}(U(x_k, u_k)) + \alpha V(L, x_{k+1}),$$

where the value function

$$V(L, x_k) \triangleq \mathbb{E} \sum_{i=k}^{\infty} \alpha^{i-k} U(x_i, Lx_i),$$

with  $U(x_i, u_i) \triangleq x_i^\top Q x_i + u_i^\top R u_i$ .

Denote  $V^*(x_k)$  as the optimal value function generated by the optimal admissible control policy. That is,

$$V^*(x_k) \triangleq V(L^*, x_k) = \underset{L \in \mathcal{L}}{\text{Minimize}} V(L, x_k).$$

For the case of stochastic LQR, we have

**Lemma 1** (Lai et al., 2023) The optimal Q-function  $Q^*(x_k, u_k) \triangleq \mathbb{E}(U(x_k, u_k)) + \alpha V^*(x_{k+1})$  can be expressed as

$$\begin{aligned} Q^*(x_k, u_k) &= \mathbb{E} \left[ [x_k^\top, u_k^\top] H^* [x_k^\top, u_k^\top]^\top \right] \\ &\quad + \frac{\alpha}{1-\alpha} \text{Tr}(P^* \Sigma), \end{aligned} \quad (4)$$

where  $P^* \succeq 0$  is the unique solution to DGARE (3),  $H^* = \begin{bmatrix} H_{11}^* & H_{12}^* \\ H_{21}^* & H_{22}^* \end{bmatrix}$  with  $H_{11}^* = Q + \alpha A^\top P^* A + \alpha \sigma A_1^\top P^* A_1, H_{12}^* = \alpha A^\top P^* B + \alpha \sigma A_1^\top P^* B_1 = (H_{21}^*)^\top, H_{22}^* = R + \alpha B^\top P^* B + \alpha \sigma B_1^\top P^* B_1$ .

Furthermore, the optimal control policy is presented by

$$u_k^* = L^* x_k = \arg \min_{u_k} Q^*(x_k, u_k)$$

with the optimal state feedback gain

$$L^* = -(H_{22}^*)^{-1} (H_{12}^*)^\top. \quad (5)$$

The following lemmas give necessary and sufficient conditions for the feedback control policy to be admissible, which will be used extensively in this paper.

**Lemma 2** (Lai et al., 2023)

The following statements are equivalent:

- (i) The feedback control policy  $u_k = Lx_k$  is admissible for system (1);
- (ii)  $\rho(C_L) < 1$ , where  $C_L \triangleq (A + BL) \otimes (A + BL) + \sigma(A_1 + B_1L) \otimes (A_1 + B_1L)$ ;
- (iii) for each given  $Y \in \mathcal{S}_{++}^n$ , there exists a unique  $G \in \mathcal{S}_{++}^n$ , such that  $(A + BL)^\top G (A + BL) + \sigma(A_1 + B_1L)^\top G (A_1 + B_1L) + Y = G$ ;
- (iv)  $\rho(\overline{C_L}) < 1$ , where  $\overline{C_L} \triangleq \overline{A_L} \otimes \overline{A_L} + \sigma \overline{A_L^1} \otimes \overline{A_L^1}$  with  $\overline{A_L} \triangleq \begin{bmatrix} A & B \\ LA & LB \end{bmatrix}$  and  $\overline{A_L^1} \triangleq \begin{bmatrix} A_1 & B_1 \\ LA_1 & LB_1 \end{bmatrix}$ ;
- (v) for each given  $Z \in \mathcal{S}_{++}^{n+m}$ , there exists a unique  $S \in \mathcal{S}_{++}^{n+m}$ , such that  $\overline{A_L} S \overline{A_L}^\top + \sigma \overline{A_L^1} S \overline{A_L^1}^\top + Z = S$ .

**Lemma 3** (Li et al., 2022) Let  $M \in \mathcal{R}^{n \times n}$ . Then, when  $\alpha > 1 - \frac{\lambda_{\min}(X)}{\lambda_{\max}(M^\top Y M)}$ ,  $\rho(M) < 1$  if and only if for each given  $X \in \mathcal{S}_{++}^n$ , there exists a unique  $Y \in \mathcal{S}_{++}^n$ , such that  $\alpha M^\top Y M + X = Y$ .

### 3 Model-based stochastic LQR via SDP

The stochastic LQR problem and the Lagrange dual problem associated with the stochastic LQR problem are first formulated as a nonconvex optimization problem and an SDP problem, respectively. Then using the convexity of the dual problem and the KKT conditions, the relationship between the optimal point of the dual problem and the parameters of the Q-function is found. Finally, the above relationship is employed to present a novel non-iterative model-based SDP algorithm for estimating the optimal control gain.

#### 3.1 Problem reformulation

By introducing the augmented system, which is described by

$$z_{k+1} = \overline{A_L} z_k + \overline{A_L^1} z_k v_k + \overline{L} w_k,$$

where the augmented state vector is defined as  $z_k \triangleq \text{col}(x_k, u_k)$ , with the initial augmented state vector  $z_0 \triangleq \text{col}(x_0, Lx_0)$ , and  $\overline{L} \triangleq \begin{bmatrix} I_n \\ L \end{bmatrix} \in \mathcal{R}^{(n+m) \times n}$ , we can derive an equivalent optimization reformulation of Problem 1 (stochastic LQR problem) as follows.

**Problem 2** (Primal Problem) Nonconvex optimization

with optimization variables  $L \in \mathcal{R}^{m \times n}$  and  $S \in \mathcal{S}^{n+m}$ .

$$J_p \triangleq \underset{L \in \mathcal{R}^{m \times n}, S \in \mathcal{S}^{n+m}}{\text{Minimize}} \text{Tr}(WS), \quad (6)$$

$$\text{subject to } S \succ 0, \quad (7)$$

$$\begin{aligned} & \alpha \left( \overline{A_L} S \overline{A_L}^\top + \sigma \overline{A_L^1} S \overline{A_L^1}^\top \right) + \overline{L} X_0 \overline{L}^\top \\ & + \frac{\alpha}{1-\alpha} \overline{L} \Sigma \overline{L}^\top = S, \end{aligned} \quad (8)$$

where  $X_0 \triangleq \mathbb{E}[x_0 x_0^\top]$ .

Note that Problem 2 is nonconvex since the equality constraint (8) is not linear.

**Proposition 1** Problem 2 is equivalent to Problem 1 in the sense that  $J_p = J^*$  and  $L_p = L^*$ , where  $(S_p, L_p)$  is the optimal point of Problem 2. In addition,  $(S_p, L_p)$  is unique.

**Proof.** Using the properties of matrix trace, we can rewrite the objective function of Problem 1 as

$$J(L, x_0) = \text{Tr}(WS),$$

where

$$S \triangleq \sum_{k=0}^{\infty} \alpha^k \mathbb{E} \begin{bmatrix} x_k \\ Lx_k \end{bmatrix} \begin{bmatrix} x_k \\ Lx_k \end{bmatrix}^\top.$$

Using Assumption 1, we can obtain the following recursive relationship equation

$$\text{vec}(\mathbb{E}[z_{k+1} z_{k+1}^\top]) = \overline{C_L} \text{vec}(\mathbb{E}[z_k z_k^\top]) + \text{vec}(\overline{L} \Sigma \overline{L}^\top).$$

Thus, for  $k \geq 1$ , we have

$$\begin{aligned} & \text{vec}(\mathbb{E}[z_k z_k^\top]) \\ & = \overline{C_L}^k \text{vec}(\overline{L} X_0 \overline{L}^\top) + \sum_{j=0}^{k-1} \overline{C_L}^j \text{vec}(\overline{L} \Sigma \overline{L}^\top), \end{aligned}$$

where  $\overline{C_L} \triangleq \overline{A_L} \otimes \overline{A_L} + \sigma \overline{A_L^1} \otimes \overline{A_L^1}$ . Hence,  $\text{vec}(S)$  becomes

$$\begin{aligned} & \text{vec}(S) \\ & = \sum_{k=0}^{\infty} \alpha^k \overline{C_L}^k \text{vec}(\overline{L} X_0 \overline{L}^\top) \\ & + \sum_{k=1}^{\infty} \alpha^k \sum_{j=0}^{k-1} \overline{C_L}^j \text{vec}(\overline{L} \Sigma \overline{L}^\top). \end{aligned}$$

By an algebraic manipulation, one can obtain that

$$\begin{aligned} & \text{vec}(S) - \alpha \overline{C_L} \text{vec}(S) \\ & = \text{vec}(\overline{L} X_0 \overline{L}^\top) + \frac{\alpha}{1-\alpha} \text{vec}(\overline{L} \Sigma \overline{L}^\top), \end{aligned}$$

which implies that  $S$  satisfies (8).

Since  $X_0 \triangleq \mathbb{E}[x_0 x_0^\top] = \Sigma_0 + \mu_0 \mu_0^\top \succ 0$  leads to  $\overline{L} (X_0 + \frac{\alpha}{1-\alpha} \Sigma) \overline{L}^\top \succ 0$ . Then, when  $\alpha > 1 - \frac{\lambda_{\min}(\overline{L}(X_0 + \frac{\alpha}{1-\alpha} \Sigma) \overline{L}^\top)}{\lambda_{\max}(\overline{A_L} S \overline{A_L}^\top + \sigma \overline{A_L^1} S \overline{A_L^1}^\top)}$ , one can obtain from Lemma 5 in (Lai et al., 2023), Lemma 2 and Lemma 3 that  $\rho(\overline{C_L}) < 1$  if and only if (8) has a unique solution  $S \succ 0$ . It follows from Lemma 2 that  $L \in \mathcal{L}$ . Therefore, we can replace the constraint  $L \in \mathcal{L}$  in Problem 1 by  $S \succ 0$  without changing its optimal point.

If  $(S_p, L_p)$  is the optimal point of Problem 2 and the corresponding optimal value is  $J_p$ , then  $S_p \succ 0$  and  $\rho(\overline{C_{L_p}}) < 1$ . Thus  $L_p \in \mathcal{L}$  is a feasible point of Problem 1, thereby,  $J_p \geq J^*$ . Furthermore, if  $S_p$  is the unique solution of (8) with  $L_p = L^* \in \mathcal{L}$ , then the resulting objective function of Problem 2 is  $J_p = J^*$ . Thus, we can conclude that  $J_p = J^*$ . The uniqueness of  $(S_p, L_p)$  can be shown by the uniqueness of  $L^*$ .  $\square$

**Remark 1** Unlike the system studied in (Li et al., 2022), which contains only additive noise, the system in this paper is more general and contains both multiplicative and additive noises. The presence of multiplicative noise makes the generalized expression of the augmented system more complex and difficult to obtain. To address this challenge, unlike the approach in (Li et al., 2022) which establishes equations satisfied by  $S$  using the derived general term formula for augmented system  $z_{k+1}$ , we directly focus on the term  $\mathbb{E}[z_{k+1} z_{k+1}^\top]$ , utilize the technique of vectorization to obtain the constraint satisfied by  $\text{vec}(S)$  directly from the recursive relationship satisfied by  $\text{vec}(\mathbb{E}[z_{k+1} z_{k+1}^\top])$ , and then use the properties of the Kronecker product and  $S = \text{vec}^{-1}(\text{vec}(S))$  to obtain the constraint satisfied by  $S$ .

### 3.2 Model-based stochastic LQR via SDP

We begin by representing the Lagrange dual problem associated with the nonconvex optimization in Problem 2 as a standard convex optimization problem as shown in Problem 3.

**Problem 3** (Dual Problem) Convex optimization with optimization variables  $M \in \mathcal{R}^{n \times n}$  and  $F = \begin{bmatrix} F_{11} & F_{12} \\ F_{12}^\top & F_{22} \end{bmatrix} \in \mathcal{S}^{n+m}$  with  $F_{11} \in \mathcal{S}^n$ ,  $F_{22} \in \mathcal{S}^m$ , and

$F_{12} \in \mathcal{R}^{n \times m}$ .

$$\underset{F,M}{\text{Maximize}} \quad \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) M \right), \quad (9)$$

$$\text{subject to } F_{22} \succ 0, \quad (10)$$

$$\mathcal{P}(F) - M \succeq 0, \quad (11)$$

$$\begin{aligned} & \alpha [A \ B]^\top \mathcal{P}(F) [A \ B] - F + W \\ & + \alpha \sigma [A_1 \ B_1]^\top \mathcal{P}(F) [A_1 \ B_1] \succeq 0, \end{aligned} \quad (12)$$

where  $\mathcal{P}(F) \triangleq F_{11} - F_{12} F_{22}^{-1} F_{12}^\top$ .

**Theorem 1** Problem 3 is an equivalent constrained convex optimization formulation of the Lagrange dual problem associated with Problem 2.

**Proof.** The Lagrangian associated with Problem 2 is

$$\begin{aligned} & \mathcal{L}_1(S, L, F_0, F) \\ & \triangleq \text{Tr}(WS) - \text{Tr}(F_0 S) \\ & + \text{Tr} \left( \left( \alpha (\overline{A_L} S \overline{A_L}^\top + \sigma \overline{A_L^1} S \overline{A_L^1}^\top) + \overline{L} X_0 \overline{L}^\top \right. \right. \\ & \left. \left. + \frac{\alpha}{1-\alpha} \overline{L} \Sigma \overline{L}^\top - S \right) F \right) \\ & = \text{Tr} \left( \left( \alpha \overline{A_L}^\top \overline{F A_L} + \alpha \sigma \overline{A_L^1}^\top \overline{F A_L^1} - F - F_0 + W \right) S \right) \\ & + \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \overline{L}^\top F \overline{L} \right), \end{aligned}$$

where the matrices  $F_0 \in \mathcal{S}_+^{n+m}$  and  $F \in \mathcal{S}^{n+m}$  are the Lagrange multipliers associated with the inequality and equality constraints of Problem 2, respectively.

The dual problem is  $\underset{F_0, F}{\text{Maximize}} \mathcal{D}_1(F_0, F)$ , where the Lagrange dual function  $\mathcal{D}_1(F_0, F)$  is defined as

$$\mathcal{D}_1(F_0, F) \triangleq \inf_L \inf_{S \succeq 0} \mathcal{L}_1(S, L, F_0, F).$$

When the Lagrangian  $\mathcal{L}_1(S, L, F_0, F)$  is unbounded from below in  $S$  and  $L$ , the dual function  $\mathcal{D}_1(F_0, F)$  takes on the value  $-\infty$ . Thus for the Lagrange dual problem:

$$\underset{F_0, F}{\text{Maximize}} \mathcal{D}_1(F_0, F),$$

the dual feasibility requires that  $F_0 \succeq 0$  and that  $\mathcal{D}_1(F_0, F) > -\infty$ . For given  $F_0$  and  $F$ , since  $S \succ 0$ , the Lagrange dual function  $\mathcal{D}_1(F_0, F)$  becomes

$$\begin{aligned} & \mathcal{D}_1(F_0, F) \\ & = \begin{cases} \inf_L \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \overline{L}^\top F \overline{L} \right), & \text{if } (F_0, F) \in \mathcal{F} \\ -\infty, & \text{otherwise} \end{cases} \end{aligned}$$

where

$$\begin{aligned} \mathcal{F} \triangleq & \{(F_0, F) : \\ & \alpha \overline{A_L}^\top \overline{F A_L} + \alpha \sigma \overline{A_L^1}^\top \overline{F A_L^1} - F - F_0 + W \succeq 0, \forall L \in \mathcal{L}\}. \end{aligned}$$

Using the block-diagonal matrix representation of the Lagrange multiplier  $F$ ,

$$\mathcal{L}_2(L, F) \triangleq \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \overline{L}^\top F \overline{L} \right)$$

can be represented as

$$\begin{aligned} \mathcal{L}_2(L, F) = & \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \right. \\ & \left. (L^\top F_{22} L + L^\top F_{12}^\top + F_{12} L + F_{11}) \right). \end{aligned}$$

Since the Lagrange multiplier  $F_0$  only appears in the constraint condition:

$$\alpha \overline{A_L}^\top \overline{F A_L} + \alpha \sigma \overline{A_L^1}^\top \overline{F A_L^1} - F - F_0 + W \succeq 0,$$

the dual feasibility is equivalent to

$$\alpha \overline{A_L}^\top \overline{F A_L} + \alpha \sigma \overline{A_L^1}^\top \overline{F A_L^1} - F + W \succeq 0.$$

Therefore, the Lagrange dual problem associated with Problem 2 is equivalent to

$$\underset{F}{\text{Maximize}} \inf_L \mathcal{L}_2(L, F), \quad (13)$$

$$\text{subject to } \alpha \overline{A_L}^\top \overline{F A_L} + \alpha \sigma \overline{A_L^1}^\top \overline{F A_L^1} - F + W \succeq 0. \quad (14)$$

In order to obtain an explicit expression for the function  $\mathcal{D}_2(F) \triangleq \inf_L \mathcal{L}_2(L, F)$ , consider the following three different cases:

Case 1: Let  $F_{22} \succ 0$ . In this case, let  $\frac{\partial \mathcal{L}_2(L, F)}{\partial L} = 0$ , one can obtain

$$\begin{aligned} & F_{22} L \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) + F_{22}^\top L \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right)^\top \\ & + F_{12}^\top \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) + F_{12}^\top \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right)^\top = 0. \end{aligned}$$

By solving the above equation, the optimal point of the function  $\mathcal{L}_2(L, F)$  can be solved to be  $L^* = -F_{22}^{-1} F_{12}^\top$ , so the infimum of  $\mathcal{L}_2(L, F)$  is attained at  $L^*$ , and hence

$$\begin{aligned} \mathcal{D}_2(F) = & \mathcal{L}_2(L^*, F) \\ = & \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \mathcal{P}(F) \right). \end{aligned} \quad (15)$$

Substituting  $L^*$  into the constraint (14) and combining it with the objective function in (15) yields an alternative equivalent representation of the dual problem (13)–(14):

$$\begin{aligned} & \text{Maximize}_F \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \mathcal{P}(F) \right), \\ & \text{subject to } \alpha [A \ B]^\top \mathcal{P}(F) [A \ B] - F + W \\ & \quad + \alpha \sigma [A_1 \ B_1]^\top \mathcal{P}(F) [A_1 \ B_1] \succeq 0. \end{aligned}$$

Case 2: Let  $F_{22} \succeq 0$  be a singular matrix, suppose that  $\mathbb{R}(F_{12}^\top) \subseteq \mathbb{R}(F_{22})$ . In this case, the infimum of  $\mathcal{L}_2(L, F)$  still exists. Suppose  $\text{rank}(F_{22}) = q < m$ , then  $F_{22}$  can be factored as  $F_{22} = U_q \Lambda_q U_q^\top$ , where  $U_q \in \mathcal{R}^{m \times q}$  satisfies  $U_q^\top U_q = I_q$  and  $\Lambda_q = \text{diag}(\lambda_1, \dots, \lambda_q)$  with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q > 0$ . Moreover, the columns of  $U_q$  span  $\mathbb{R}(F_{22})$ . The pseudo-inverse of the singular matrix  $F_{22}$  can be represented as  $F_{22}^\dagger = U_q \Lambda_q^{-1} U_q^\top$ . Following a procedure similar to the Case 1, the Lagrange dual problem associated with Problem 2 is derived as

$$\begin{aligned} & \text{Maximize}_F \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) (F_{11} - F_{12} F_{22}^\dagger F_{12}^\top) \right), \\ & \text{subject to } \alpha [A \ B]^\top (F_{11} - F_{12} F_{22}^\dagger F_{12}^\top) [A \ B] \\ & \quad + \alpha \sigma [A_1 \ B_1]^\top (F_{11} - F_{12} F_{22}^\dagger F_{12}^\top) [A_1 \ B_1] \\ & \quad - F + W \succeq 0. \end{aligned}$$

Since  $\mathbb{R}(F_{12}^\top) \subseteq \mathbb{R}(F_{22})$ , there exists an unitary matrix  $U_{m-q}$  such that the columns of  $U_{m-q}$  span  $\mathbb{N}(F_{12})$ . Let  $\Lambda_{m-q}$  be an arbitrary diagonal matrix with positive diagonal elements, and construct a matrix  $\widehat{F}_{22} \succ 0$  as

$$\widehat{F}_{22} = [U_q \ U_{m-q}] \begin{bmatrix} \Lambda_q & 0 \\ 0 & \Lambda_{m-q} \end{bmatrix} [U_q \ U_{m-q}]^\top.$$

Noting that

$$\begin{aligned} & F_{12} \widehat{F}_{22}^{-1} F_{12}^\top \\ &= F_{12} [U_q \ U_{m-q}] \begin{bmatrix} \Lambda_q^{-1} & 0 \\ 0 & \Lambda_{m-q}^{-1} \end{bmatrix} \begin{bmatrix} U_q^\top \\ U_{m-q}^\top \end{bmatrix} F_{12}^\top \\ &= F_{12} U_q \Lambda_q^{-1} U_q^\top F_{12}^\top - \underbrace{F_{12} U_{m-q} \Lambda_{m-q}^{-1} U_{m-q}^\top F_{12}^\top}_0 \\ &= F_{12} F_{22}^\dagger F_{12}^\top, \end{aligned}$$

it can be concluded that every admissible candidate  $F_{22} \succeq 0$  leads to an objective value which can be obtained by replacing a counterpart  $\widehat{F}_{22} \succ 0$ .

Case 3:  $F_{22}$  has at least one negative eigenvalue or  $F_{22} \succeq 0$  while  $\mathbb{R}(F_{12}^\top) \not\subseteq \mathbb{R}(F_{22})$ . In this case, the Lagrangian

is unbounded from below through minimizing  $\mathcal{L}_2(L, F)$  in terms of  $L$ .

To summarize, the existence of  $\inf_L \mathcal{L}_2(L, F)$  requires that (i)  $F_{22} \succ 0$  or (ii)  $F_{22} \succeq 0$  with an additional constraint  $\mathbb{R}(F_{12}^\top) \subseteq \mathbb{R}(F_{22})$ . Moreover, every permissible  $F_{22} \succeq 0$  results in an objective value that can be alternatively derived by a counterpart  $\widehat{F}_{22} \succ 0$ . Thus, without loss of generality, we follow up with the assumption that  $F_{22} \succ 0$ . Therefore, the dual problem (13)–(14) is equivalent to

$$\text{Maximize}_F \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \mathcal{P}(F) \right), \quad (16)$$

$$\text{subject to } F_{22} \succ 0, \quad (17)$$

$$\begin{aligned} & \alpha [A \ B]^\top \mathcal{P}(F) [A \ B] - F + W \\ & + \alpha \sigma [A_1 \ B_1]^\top \mathcal{P}(F) [A_1 \ B_1] \succeq 0. \end{aligned} \quad (18)$$

Finally, by introducing the slack variable  $M \in \mathcal{S}^n$  and the additional constraint (11), the Lagrange dual problem (16)–(18) can be equivalently described by Problem 3. In fact, let  $F^*$  be the optimal point of the dual problem (16)–(18) and  $(F^*, M^*)$  be the optimal point of Problem 3. According to the constraint (11) and positive definiteness of  $X_0 + \frac{\alpha}{1-\alpha} \Sigma$ ,

$$\begin{aligned} & \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) M^* \right) \\ & \leq \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \mathcal{P}(F^*) \right). \end{aligned}$$

Since  $M^*$  is the maximum point of the objective function (11),

$$\begin{aligned} & \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) M^* \right) \\ & \geq \text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) \mathcal{P}(F^*) \right), \end{aligned}$$

and hence

$$\text{Tr} \left( \left( X_0 + \frac{\alpha}{1-\alpha} \Sigma \right) (M^* - \mathcal{P}(F^*)) \right) = 0.$$

Consequently, all of the eigenvalues of the matrix  $M^* - \mathcal{P}(F^*)$  are equal to zero. Since  $M^*$  is symmetric, it follows that

$$M^* = \mathcal{P}(F^*),$$

which means that the dual problem (16)–(18) is equivalent to Problem 3.  $\square$

The next theorem gives the relationship between the optimal point to the dual problem associated with the

stochastic LQR problem and the parameters of the optimal Q-function.

**Theorem 2** The optimal point of Problem 3, which is denoted by  $(F^*, M^*)$ , is independent of  $X_0 + \frac{\alpha}{1-\alpha}\Sigma$ . Furthermore,  $F^* = H^*$  and  $M^* = P^*$ , where  $H^*$  is the parameter of the optimal Q-function expressed in (4) and  $P^*$  is the solution to DGARE (3).

**Proof.** By introducing the Lagrange multipliers  $G_1 \in \mathcal{S}_+^{n+m}$ ,  $G_2 \in \mathcal{S}_+^m$  and  $G_3 \in \mathcal{S}_+^n$  associated with the inequality constraints of Problem 3, the Lagrangian associated with Problem 3 is described by

$$\begin{aligned} \mathcal{L}_3(F, M, G_1, G_2, G_3) &= -\text{Tr}\left(\left(X_0 + \frac{\alpha}{1-\alpha}\Sigma\right)M\right) \\ &\quad - \text{Tr}\left(G_1\left(\alpha[A\ B]^\top \mathcal{P}(F)[A\ B]\right.\right. \\ &\quad \left.\left.+ \alpha\sigma[A_1\ B_1]^\top \mathcal{P}(F)[A_1\ B_1] - F + W\right)\right) \\ &\quad - \text{Tr}(G_2F_{22}) - \text{Tr}(G_3(\mathcal{P}(F) - M)). \end{aligned}$$

Since Problem 3 is convex, the KKT conditions are sufficient for the points to be primal and dual optimal (Boyd & Vandenberghe, 2004). In other words, if  $\tilde{F}, \tilde{M}, \tilde{G}_1, \tilde{G}_2$  and  $\tilde{G}_3$  are any points that satisfy the following KKT conditions of Problem 3:

- (i) Primal feasibility: (10)–(12);
- (ii) Dual feasibility:

$$G_1 \succeq 0, \quad G_2 \succeq 0, \quad G_3 \succeq 0;$$

- (iii) Complementary slackness:

$$\begin{aligned} \text{Tr}\left(G_1\left(\alpha[A\ B]^\top \mathcal{P}(F)[A\ B] - F + W\right.\right. \\ \left.\left.+ \alpha\sigma[A_1\ B_1]^\top \mathcal{P}(F)[A_1\ B_1]\right)\right) = 0, \end{aligned} \quad (19)$$

$$\text{Tr}(G_2F_{22}) = 0, \quad (20)$$

$$\text{Tr}(G_3(\mathcal{P}(F) - M)) = 0; \quad (21)$$

- (iv) Stationarity conditions:

$$\frac{\partial \mathcal{L}_3}{\partial M} = 0, \quad \frac{\partial \mathcal{L}_3}{\partial F} = 0, \quad (22)$$

then  $(\tilde{F}, \tilde{M})$  and  $(\tilde{G}_1, \tilde{G}_2, \tilde{G}_3)$  are primal and dual optimal points, respectively.

Choose  $\tilde{F} = H^*$  and  $\tilde{M} = P^*$ . Substituting  $\tilde{F} = H^*$  into the left-hand side of (12) results in the left-hand side of (12) to be equal to 0, from which it follows that the constraint (12) is active, i.e., the constraint (19) is trivially satisfied without any requirement for  $\tilde{G}_1$ . The constraint (10) is not active at the candidate point  $(\tilde{F}, \tilde{M}) =$

$(H^*, P^*)$  due to  $\tilde{F}_{22} = H_{22}^* \succ 0$ , which dictates  $\tilde{G}_2 = 0$ . Moreover, substituting  $(\tilde{F}, \tilde{M}) = (H^*, P^*)$  into the left-hand side of the inequality constraint (11) results in the left-hand side of (11) to be equal to 0. That is, the constraint (11) is active and (21) is trivially satisfied without imposing any restriction on  $\tilde{G}_3$ .

It is shown below that the stationary conditions (22) can be satisfied at the point  $(\tilde{F}, \tilde{M})$  by appropriate selection of  $\tilde{G}_1$  and  $\tilde{G}_3$ . To this end, the matrix variable  $F$  is reformulated as

$$F = E_1 F_{11} E_1^\top + E_1 F_{12} E_2^\top + E_2 F_{12}^\top E_1^\top + E_2 F_{22} E_2^\top$$

by using the matrices  $E_1 = [I_n \ 0]^\top \in \mathcal{R}^{(n+m) \times n}$  and  $E_2 = [0 \ I_m]^\top \in \mathcal{R}^{(n+m) \times m}$ . Then, the stationarity conditions (22), can be equivalently described by (23)–(26).

$$\frac{\partial \mathcal{L}_3}{\partial M} = -\left(X_0 + \frac{\alpha}{1-\alpha}\Sigma\right)^\top + G_3^\top = 0, \quad (23)$$

$$\begin{aligned} \frac{\partial \mathcal{L}_3}{\partial F_{11}} &= -G_3 + E_1^\top G_1 E_1 - \alpha[A\ B]G_1[A\ B]^\top \\ &\quad - \alpha\sigma[A_1\ B_1]G_1[A_1\ B_1]^\top = 0, \end{aligned} \quad (24)$$

$$\begin{aligned} \frac{\partial \mathcal{L}_3}{\partial F_{12}} &= 2\alpha\left([A\ B]G_1[A\ B]^\top\right. \\ &\quad \left.+\sigma[A_1\ B_1]G_1[A_1\ B_1]^\top\right)\tilde{F}_{12}\tilde{F}_{22}^{-1} \\ &\quad + 2G_3\tilde{F}_{12}\tilde{F}_{22}^{-1} + 2E_1^\top G_1 E_2 = 0, \end{aligned} \quad (25)$$

$$\begin{aligned} \frac{\partial \mathcal{L}_3}{\partial F_{22}} &= -\tilde{F}_{22}^{-1}F_{12}^\top\left(G_3 + \alpha[A\ B]G_1[A\ B]^\top\right. \\ &\quad \left.+\alpha\sigma[A_1\ B_1]G_1[A_1\ B_1]^\top\right)\tilde{F}_{12}\tilde{F}_{22}^{-1} \\ &\quad + E_2^\top G_1 E_2 = 0. \end{aligned} \quad (26)$$

Choose

$$\tilde{G}_3 = X_0 + \frac{\alpha}{1-\alpha}\Sigma \succ 0, \quad (27)$$

which satisfies the stationary condition (23).

From (24) and (25), the matrix  $G_1$  should be subject to the following constraint:

$$E_1^\top G_1 E_1 \tilde{F}_{12} \tilde{F}_{22}^{-1} + E_1^\top G_1 E_2 = 0. \quad (28)$$

Furthermore, from (24) and (26),  $G_1$  should also be subject to the following constraint:

$$-\tilde{F}_{12}^\top E_1^\top G_1 E_1 \tilde{F}_{12} + \tilde{F}_{22} E_2^\top G_1 E_2 \tilde{F}_{22} = 0. \quad (29)$$

By partitioning the Lagrange multiplier  $G_1$  as  $G_1 = \begin{bmatrix} G_{11}^1 & G_{12}^1 \\ (G_{12}^1)^\top & G_{22}^1 \end{bmatrix}$ , the equalities (28) and (29) are repre-

sented as

$$G_{11}^1 \tilde{F}_{12} \tilde{F}_{22}^{-1} + G_{12}^1 = 0, \quad (30)$$

$$-\tilde{F}_{12}^\top G_{11}^1 \tilde{F}_{12} + \tilde{F}_{22} G_{22}^1 \tilde{F}_{22} = 0. \quad (31)$$

By substituting (27), (30) and (31) into (24), we have

$$\begin{aligned} & \tilde{G}_3 + \alpha \left( A - B \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \right) G_{11}^1 \left( A - B \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \right)^\top \\ & + \alpha \sigma \left( A_1 - B_1 \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \right) G_{11}^1 \left( A_1 - B_1 \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \right)^\top \\ & = G_{11}^1. \end{aligned} \quad (32)$$

As we choose to let  $\tilde{F} = H^*$ , based on the optimal control gain representation given by (5),  $A - B \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top$  and  $A_1 - B_1 \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top$  in (32) become  $\mathcal{A} \triangleq A + BL^*$  and  $\mathcal{A}_1 \triangleq A_1 + B_1 L^*$ , respectively. Since  $L^* \in \mathcal{L}$  leads to  $\rho(C_{L^*}) < 1$ , then when

$$\alpha > 1 - \frac{\lambda_{\min}(\tilde{G}_3)}{\lambda_{\max}(\mathcal{A}G_{11}^1 \mathcal{A}^\top + \sigma \mathcal{A}_1 G_{11}^1 \mathcal{A}_1^\top)},$$

one can obtain from Lemma 2 and Lemma 3 that for given  $\tilde{G}_3 = X_0 + \frac{\alpha}{1-\alpha} \Sigma \succ 0$ , there exists a unique  $\tilde{G}_{11}^1 \succ 0$  such that (32) holds. Accordingly, construct  $\tilde{G}_1$  as

$$\begin{aligned} \tilde{G}_1 &= \begin{bmatrix} \tilde{G}_{11}^1 & -\tilde{G}_{11}^1 \tilde{F}_{12} \tilde{F}_{22}^{-1} \\ -\tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \tilde{G}_{11}^1 & \tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \tilde{G}_{11}^1 \tilde{F}_{12} \tilde{F}_{22}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I_n \\ -\tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \end{bmatrix} \tilde{G}_{11}^1 \begin{bmatrix} I_n \\ -\tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \end{bmatrix}^\top \succeq 0. \end{aligned}$$

In summary, it can be concluded that  $(H^*, P^*)$  and

$$\left( \begin{bmatrix} I_n \\ -\tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \end{bmatrix} \tilde{G}_{11}^1 \begin{bmatrix} I_n \\ -\tilde{F}_{22}^{-1} \tilde{F}_{12}^\top \end{bmatrix}^\top, 0, X_0 + \frac{\alpha}{1-\alpha} \Sigma_0 \right)$$

are primal and dual optimal points, respectively. Therefore,  $F^* = H^*$  and  $M^* = P^*$ . Moreover, according to the expressions of  $H^*$  and  $P^*$ ,  $(H^*, P^*)$  is independent of  $X_0 + \frac{\alpha}{1-\alpha} \Sigma$ .  $\square$

**Remark 2** The primal problem in this paper is nonconvex, which makes it nontrivial to prove that strong duality holds. Instead of proving that strong duality holds, this paper proceeds directly from the dual problem and finds the relationship between the optimal point of the dual problem and the parameters in Q-learning by using the convexity of the dual problem and the conclusion that the KKT conditions in convex optimization are sufficient for optimality (Boyd & Vandenberghe, 2004).

Theorem 2 shows that the term  $X_0 + \frac{\alpha}{1-\alpha} \Sigma$  only affects the optimal value of the objective function (9) in Problem 3, and that optimal point  $(F^*, M^*)$  is independent of  $X_0 + \frac{\alpha}{1-\alpha} \Sigma$ . Then, without loss of generality, the objective function (9) in Problem 3 can be expressed as

$$\underset{F, M}{\text{Maximize}} \text{Tr}(M).$$

Using the Schur complement property, the constraint (11) in Problem 3 can be rewritten in the following linear matrix inequality (LMI) form:

$$\begin{bmatrix} F_{11} - M & F_{12} \\ F_{12}^\top & F_{22} \end{bmatrix} \succeq 0. \quad (33)$$

Moreover, the constraint (10) in Problem 3 is implicitly contained in (33). Introducing the auxiliary matrix  $\bar{I} \triangleq [I_{n+m} \ I_{n+m}]$  and  $\bar{D} \triangleq [A \ B] \oplus [A_1 \ B_1]$ , and then using the technique of the matrix direct sum and the Schur complement property once more, the constraint (12) in Problem 3 can also be rewritten in the LMI form described by

$$\begin{bmatrix} \alpha \bar{I} \bar{D}^\top \bar{F}_{11}^2(\sigma) \bar{D} \bar{I}^\top - F + W \alpha^{\frac{1}{2}} \bar{I} \bar{D}^\top \bar{F}_{12}^2(\sigma) \\ \alpha^{\frac{1}{2}} \bar{F}_{12}^2(\sigma)^\top \bar{D} \bar{I}^\top & \bar{F}_{22}^{-2} \end{bmatrix} \succeq 0,$$

where  $\bar{F}_{ij}^2(\sigma) \triangleq F_{ij} \oplus \sigma F_{ij}$ ,  $i, j = 1, 2$ , and  $\bar{F}_{22}^{-2} \triangleq F_{22} \oplus F_{22}$ . Consequently, Problem 3 is equivalently described by a standard SDP problem in Problem 4.

**Problem 4** SDP with optimization variables  $F \in \mathcal{S}^{n+m}$  and  $M \in \mathcal{S}^n$ .

$$\underset{F, M}{\text{Maximize}} \text{Tr}(M),$$

$$\text{subject to } \begin{bmatrix} F_{11} - M & F_{12} \\ F_{12}^\top & F_{22} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} \alpha \bar{I} \bar{D}^\top \bar{F}_{11}^2(\sigma) \bar{D} \bar{I}^\top - F + W \alpha^{\frac{1}{2}} \bar{I} \bar{D}^\top \bar{F}_{12}^2(\sigma) \\ \alpha^{\frac{1}{2}} \bar{F}_{12}^2(\sigma)^\top \bar{D} \bar{I}^\top & \bar{F}_{22}^{-2} \end{bmatrix} \succeq 0.$$

**Remark 3** Unlike the deterministic LQ problem studied in (Farjadianasab & Babazadeh, 2022), this paper investigates the stochastic LQR problem for stochastic systems subject to multiplicative and additive noises. Consequently, the third constraint (12) in the dual problem differs from its counterpart in (Farjadianasab & Babazadeh, 2022) not only by adding a discount factor coefficient  $\alpha$  but also by including a term related to multiplicative noise:  $\alpha \sigma [A_1 \ B_1]^\top \mathcal{P}(F) [A_1 \ B_1]$ . The presence of this term makes it impossible to rewrite (12) into LMI form solely by exploiting the Schur complement property. To address this, auxiliary matrices  $\bar{I}$

and  $\bar{D}$  are introduced, and the technique of direct sum is employed to achieve the desired transformation. This technique is also utilized in the subsequent reformulation of Problem 5.

With the above preparation, we can now present the model-based SDP algorithm.

---

**Algorithm 1** Model-Based SDP Algorithm

---

- 1: Solve SDP in Problem 4 for  $(F^*, M^*)$ .
  - 2: Obtain the solution  $P^*$  of the DGARE (3) as  $P^* = M^*$ , and the optimal state feedback gain  $L^* = -(F_{22}^*)^{-1}(F_{12}^*)^\top$ .
- 

#### 4 Model-free implementation of model-based SDP algorithm

In this section, based on Monte-Carlo method, a model-free implementation of the SDP controller design is given, and a completely data-driven optimal stochastic LQR control is presented by using the technique of the matrix direct sum.

Assume that the triplets  $(x_k^{(i)}, u_k^{(i)}, x_{k+1}^{(i)})$ ,  $i = 1, \dots, N$ ,  $k = 0, \dots, K-1$  are available, where  $N$  denotes the number of sampling experiments implemented and  $K$  denotes the length of the sampling time horizon in the data collection phase. Let  $Z^{(1)}, Z^{(2)}, \dots, Z^{(N)}$  be a set of samples from  $Z \triangleq [z_0 \ z_1 \ \dots \ z_{K-1}]$  and  $Y^{(1)}, Y^{(2)}, \dots, Y^{(N)}$  be a set of samples from  $Y \triangleq [x_1 \ x_2 \ \dots \ x_K]$ . According to (1), we have

$$Y = [A \ B]Z + [A_1 \ B_1]Z\mathcal{V} + \mathcal{W}, \quad (34)$$

where  $\mathcal{V} \triangleq v_0 \oplus v_1 \oplus \dots \oplus v_{K-1}$ ,  $\mathcal{W} \triangleq [w_0 \ w_1 \ \dots \ w_{K-1}]$ . Noting (34) and using Assumption 1, one has

$$\begin{aligned} & \mathbb{E}[Y^\top \mathcal{P}(F)Y] \\ &= \mathbb{E}\left[Z^\top [A \ B]^\top \mathcal{P}(F)[A \ B]Z\right] \\ &\quad + \sigma \mathbb{E}\left[Z^\top [A_1 \ B_1]^\top \mathcal{P}(F)[A_1 \ B_1]Z\right] \\ &\quad + \mathbb{E}[\mathcal{W}^\top \mathcal{P}(F)\mathcal{W}]. \end{aligned} \quad (35)$$

We now employ (35) to provide the model-free implementation of Algorithm 1. Assuming that  $Z^{(i)}$  has full row rank for each  $i = 1, \dots, N$ , then by the properties of matrix congruence, if the constraint (12) in Problem 3 is left-multiplied by  $Z^{(i)\top}$  and right-multiplied by  $Z^{(i)}$ ,

we obtain the equivalent constraint of (12) as follows:

$$\begin{aligned} & \alpha Z^{(i)\top} [A \ B]^\top \mathcal{P}(F)[A \ B]Z^{(i)} \\ &+ \alpha \sigma Z^{(i)\top} [A_1 \ B_1]^\top \mathcal{P}(F)[A_1 \ B_1]Z^{(i)} \\ &- Z^{(i)\top} (F - W) Z^{(i)} \succeq 0. \end{aligned}$$

Using the properties of positive semidefinite matrices and  $N > 0$ , one obtains

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \left[ \alpha Z^{(i)\top} [A \ B]^\top \mathcal{P}(F)[A \ B]Z^{(i)} \right] \\ &+ \frac{1}{N} \sum_{i=1}^N \left[ \alpha \sigma Z^{(i)\top} [A_1 \ B_1]^\top \mathcal{P}(F)[A_1 \ B_1]Z^{(i)} \right] \\ &- \frac{1}{N} \sum_{i=1}^N \left[ Z^{(i)\top} (F - W) Z^{(i)} \right] \succeq 0. \end{aligned} \quad (36)$$

Based on the Monte-Carlo method, we can approximate the mathematical expectation using the numerical average of  $N$  sample paths. According to (35), one has

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \left[ \alpha Z^{(i)\top} [A \ B]^\top \mathcal{P}(F)[A \ B]Z^{(i)} \right] \\ &+ \frac{1}{N} \sum_{i=1}^N \left[ \alpha \sigma Z^{(i)\top} [A_1 \ B_1]^\top \mathcal{P}(F)[A_1 \ B_1]Z^{(i)} \right] \\ &\approx \frac{1}{N} \sum_{i=1}^N \left[ \alpha Y^{(i)\top} \mathcal{P}(F)Y^{(i)} \right] - \alpha \mathbb{E}[\mathcal{W}^\top \mathcal{P}(F)\mathcal{W}]. \end{aligned} \quad (37)$$

By substituting (37) into (36) and noting that

$$\mathbb{E}[\mathcal{W}^\top \mathcal{P}(F)\mathcal{W}] \succeq 0,$$

we have

$$\frac{1}{N} \sum_{i=1}^N \left[ \alpha Y^{(i)\top} \mathcal{P}(F)Y^{(i)} - Z^{(i)\top} (F - W) Z^{(i)} \right] \succeq 0. \quad (38)$$

Since  $F_{22} \succ 0$ , using the Schur complement and the technique of matrix direct sum, the constraint (38) can be equivalently expressed as

$$\begin{aligned} & \begin{bmatrix} \alpha \bar{I}_K^{-N} \bar{Y}^{N\top} \bar{F}_{11}^{-N} \bar{Y}^N \bar{I}_K^{-N\top} - \bar{Z}^N(F) \alpha^{\frac{1}{2}} \bar{I}_K^{-N} \bar{Y}^{N\top} \bar{F}_{12}^{-N} \\ \alpha^{\frac{1}{2}} \bar{F}_{12}^{-N\top} \bar{Y}^N \bar{I}_K^{-N\top} & \bar{F}_{22}^{-N} \end{bmatrix} \\ & \succeq 0, \end{aligned} \quad (39)$$

where

$$\begin{aligned}\overline{F_{ij}}^N &\triangleq \underbrace{F_{ij} \oplus F_{ij} \oplus \cdots \oplus F_{ij}}_N, i, j = 1, 2, \\ \overline{I_K}^N &\triangleq \left[ \underbrace{I_K \ I_K \ \cdots \ I_K}_N \right], \\ \overline{Y}^N &\triangleq Y^{(1)} \oplus Y^{(2)} \oplus \cdots \oplus Y^{(N)}, \\ \overline{Z(F)}^N &\triangleq \sum_{i=1}^N \left[ Z^{(i)\top} (F - W) Z^{(i)} \right].\end{aligned}$$

Based on the above analysis, Theorem 3 introduces an SDP in Problem 5 for solving the stochastic LQR problem without any information about the system dynamics and without requiring an initial stabilizing control policy.

**Problem 5** SDP with optimization variables  $F \in \mathcal{S}^{n+m}$  and  $M \in \mathcal{S}^n$ .

Maximize  $\text{Tr}(M)$ ,

$$\begin{aligned}&\text{subject to } \begin{bmatrix} F_{11} - M & F_{12} \\ F_{12}^\top & F_{22} \end{bmatrix} \succeq 0, \\ &\left[ \begin{array}{cc} \alpha \overline{I_K}^N \overline{Y}^{N\top} \overline{F_{11}}^N \overline{Y}^N \overline{I_K}^{N\top} - \overline{Z}^N(F) \alpha^{\frac{1}{2}} \overline{I_K}^N \overline{Y}^{N\top} \overline{F_{12}}^N \\ \alpha^{\frac{1}{2}} \overline{F_{12}}^{N\top} \overline{Y}^N \overline{I_K}^{N\top} & \overline{F_{22}}^N \end{array} \right] \\ &\succeq 0.\end{aligned}$$

**Theorem 3** Suppose that  $Z^{(i)}$  has full row rank for each  $i = 1, \dots, N$ . Then the optimal state feedback gain associated with the stochastic LQR problem is approximated by

$$\widehat{L}^* = -\left(\widehat{F}_{22}\right)^{-1} \left(\widehat{F}_{12}\right)^\top, \quad (40)$$

where  $(\widehat{F}, \widehat{M})$  is the optimal point of Problem 5 with

$$\widehat{F} = \begin{bmatrix} \widehat{F}_{11} & \widehat{F}_{12} \\ (\widehat{F}_{12})^\top & \widehat{F}_{22} \end{bmatrix}.$$

**Proof.** It is already shown that according to the Monte-Carlo method and the properties of matrix congruence and the positive semidefinite matrices, the constraint (12) in Problem 3 can be approximated by the constraint (38), and the constraint (38) is equivalent to the constraint (39). Based on Theorem 2, the exact parameter  $H^*$  of the optimal Q-function is obtained by solving Problem 3, which is approximately equivalent to Problem 5, and then Q-learning offers an estimate of the optimal state feedback gain, as shown in (40).  $\square$

It should be noted that, to ensure the rank condition in Theorem 3, the length of the sampling time horizon  $K$  must be more than or equal to the sum of the dimensions of the input and the state. The rank condition in Theorem 3 is aimed at allowing the system to oscillate sufficiently during the learning process to produce enough data to fully learn the system information. To this end, in practice, a small exploration noise consisting of a Gaussian white noise or a sum of sinusoids must be added to the control signal.

**Remark 4** The model-free implementation process in this paper differs from that in (Farjadnasab & Babazadeh, 2022). Due to the presence of multiplicative and additive noise, the next-time-step state  $x_{k+1}$  in this paper becomes a random variable, which is fundamentally different from the state-determined representation in (Farjadnasab & Babazadeh, 2022). Therefore, the model-free implementation method from (Farjadnasab & Babazadeh, 2022) is no longer applicable to this paper. To address this challenge, we introduce a set of  $N$  samples for the next-time-step state vector  $Y$  and the augmented system state vector  $Z$  for  $k = 0, \dots, K-1$ . Based on the Monte-Carlo method, we approximate the mathematical expectation using the sample mean to derive an approximate equivalent description of the constraints. This enables the estimation of the optimal feedback gain.

In view of the results given above, Algorithm 2 allows one to solve the stochastic LQR problem on the basis of the available data.

---

#### Algorithm 2 Model-Free SDP Algorithm

---

- 1: Choose the initial state  $x_0$  from a distribution with the constant mean vector and positive definite covariance matrix.
  - 2: Apply  $u_k = u_k^{(i)} + d_k$  to system (1) for  $K$  time steps, where  $d_k$  is the exploration noise with zero mean and covariance matrix  $\Sigma_d \succ 0$ , and collect the input and state data to construct the data matrices  $Z^{(i)}$  and  $Y^{(i)}$  for each  $i = 1, \dots, N$  until the rank condition in Theorem 3 is satisfied.
  - 3: Solve SDP in Problem 5 for  $(\widehat{F}, \widehat{M})$ .
  - 4: Obtain the estimate of  $P^*$  given by  $\widehat{P}^* = \widehat{M}$ , and the estimate of the optimal state feedback gain  $\widehat{L}^*$  associated with stochastic LQR problem described by (40).
- 

It is worth noting that model-free iterative algorithms, such as policy iteration, value iteration, and Q-learning, can also solve the stochastic LQR problem and guarantee the convergence to the optimal policy as the number of iterations goes to infinity. However, the proposed off-policy algorithm is non-iterative and computes the estimate of the optimal policy in a single step by utilizing only the sampled data with time-domain length

$K$  ( $K \geq n + m$ ). It neither requires an initial stabilizing controller to be provided, nor does it require hyper-parameter tuning. It is characterized by high sampling efficiency and, as a known feature of LMIs, it is inherently robust to model uncertainty. Furthermore, since the SDP in Problem 5 does not contain terms associated with multiplicative and additive noises, the implementation of the proposed model-free SDP algorithm does not require the multiplicative and additive noises to be measurable.

**Remark 5** The multiplicative noise is set to be a scalar  $v_k$  just for simplicity of presentation, and the results of this paper can be generalized to the following system:

$$x_{k+1} = Ax_k + Bu_k + \sum_{l=1}^M (A_l x_k + B_l u_k) v_k^l + w_k,$$

where  $\vec{v}_k = (v_k^1, v_k^2, \dots, v_k^M) \in \mathcal{R}^M$  is system multidimensional multiplicative noise. The following assumptions are made regarding the aforementioned system: The noises  $v_k^1, v_k^2, \dots, v_k^M$  are i.i.d. (not necessarily Gaussian distributed) satisfying  $\mathbb{E}(v_k^l) = 0$  and  $\mathbb{E}[v_k^l v_k^{l\top}] = \delta_{ij}\sigma$ , where  $\sigma > 0$  and  $\delta_{ij}$  is a Kronecker function. Moreover,  $\{w_k, k \in \mathcal{Z}_+\}$  and  $\{\vec{v}_k, k \in \mathcal{Z}_+\}$  are mutually independent. If multiplicative noise is multidimensional, e.g.,  $\vec{v}_k = (v_k^1, v_k^2, \dots, v_k^M)$ , the augmented system becomes

$$z_{k+1} = \overline{A_L} z_k + \sum_{l=1}^M \overline{A_L^l} z_k v_k^l + \overline{L} w_k,$$

where  $\overline{A_L^l} \triangleq \begin{bmatrix} A_l & B_l \\ L A_l & L B_l \end{bmatrix} \in \mathcal{R}^{(n+m) \times (n+m)}$ , the remaining symbols are the same as when  $v_k$  a scalar. Accordingly, Equation (8) satisfied by  $S$  in this paper becomes

$$\alpha \left( \overline{A_L S \overline{A_L}^\top} + \sigma \sum_{l=1}^M \overline{A_L^l} S \overline{A_L^l}^\top \right) + \overline{L} X_0 \overline{L}^\top + \frac{\alpha}{1-\alpha} \overline{L} \Sigma \overline{L}^\top = S.$$

The corresponding results in this paper can be proved in the same way for the multidimensional case. For example: Problem 3 needs only one modification, i.e., Equation (12) is changed to be

$$\alpha [A \ B]^\top \mathcal{P}(F) [A \ B] + \alpha \sigma \sum_{l=1}^M [A_l \ B_l]^\top \mathcal{P}(F) [A_l \ B_l] - F + W \succeq 0,$$

and the second constraint in Problem 4 is modified to

$$\mathcal{P}(F) \triangleq \begin{bmatrix} \mathcal{P}(F)_{11} & \mathcal{P}(F)_{12} \\ \mathcal{P}(F)_{12}^\top & \mathcal{P}(F)_{22} \end{bmatrix} \succeq 0,$$

where

$$\begin{aligned} \mathcal{P}(F)_{11} &\triangleq \alpha \overline{I_{n+m}^{M+1}} \overline{D^{M+1}}^\top \overline{F_{11}^{M+1}}(\sigma) D^{M+1} \overline{I_{n+m}^{M+1}}^\top - F + W, \\ \mathcal{P}(F)_{12} &\triangleq \alpha^{\frac{1}{2}} \overline{I_{n+m}^{M+1}} \overline{D^{M+1}}^\top \overline{F_{12}^{M+1}}(\sigma), \\ \mathcal{P}(F)_{22} &\triangleq \overline{F_{22}^{M+1}} \end{aligned}$$

with

$$\begin{aligned} \overline{I_{n+m}^{M+1}} &\triangleq \underbrace{\begin{bmatrix} I_{n+m} & I_{n+m} & \cdots & I_{n+m} \end{bmatrix}}_{M+1}, \\ \overline{D^{M+1}} &\triangleq [A \ B] \oplus_{l=1}^M [A_l \ B_l], \\ \overline{F_{ij}^{M+1}}(\sigma) &\triangleq F_{ij} \oplus_{l=1}^M (\sigma F_{ij}), i, j = 1, 2, \\ \overline{F_{22}^{M+1}} &\triangleq \oplus_{l=1}^{M+1} F_{22}. \end{aligned}$$

We introduce five key optimization problems (Problems 1 to 5) in this paper. To improve readability, we visually illustrate their relationships in Fig. 1.

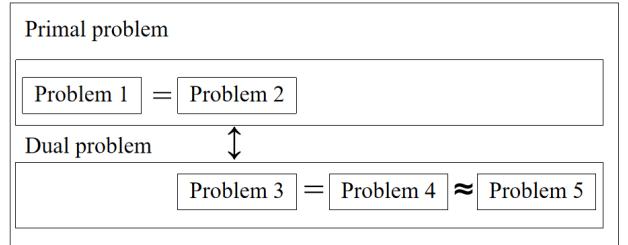


Fig. 1. Diagram for relations among the results.

In the aforementioned model-free implementation, replacing the mathematical expectation with sample averaging introduces error. In fact, due to the presence of stochastic noise, error is unavoidable. Beyond Monte Carlo approximation error, sources of error include but are not limited to the impact of probing noise in data collection, the sample estimation error and numerical errors in SDP solution. Taking into account the presence of errors, it is necessary to study the robustness of the model-free algorithm to errors in the next section.

## 5 Robustness analysis

In this section, the robustness of the model-free SDP algorithm to errors in the learning process is studied.

The estimation error is shown to be bounded and the estimation of the optimal control policy is proven to be admissible under mild conditions.

**Theorem 4** Assume that the data matrices  $Z^{(i)} \in \mathbb{R}^{(n+m) \times K}$  (for  $i = 1, 2, \dots, N$ ) constructed from system interaction data are full-row rank, with  $\sigma_{\min}(Z^{(i)}) \geq \gamma > 0$  and  $K \geq n + m$ , and that there exists a constant  $\underline{\sigma}$  such that  $\sigma_{\min}(I - \alpha C_{\widehat{L}^*}) \geq \underline{\sigma} > 0$ , where

$$C_{\widehat{L}^*} = (A + B\widehat{L}^*) \otimes (A + B\widehat{L}^*) + \sigma(A_1 + B_1\widehat{L}^*) \otimes (A_1 + B_1\widehat{L}^*).$$

Then, for any confidence level  $1 - \beta$  ( $\beta \in (0, 0.1)$ ), the estimation error satisfies

$$\mathbb{P}\left(\|\widehat{L}^* - L^*\|_F \leq \widetilde{O}\left(\frac{(n+m)(1 + \|F^*\|_F)}{\sqrt{NK} \cdot \sigma_d \cdot \underline{\sigma}}\right)\right) \geq 1 - \delta,$$

where  $\sigma_d > 0$  is the minimum intensity of the exploration noise.

**Proof.** The proof proceeds in four key steps, leveraging convex optimization duality, matrix perturbation theory, and probability tail bounds.

#### Step 1: Probability Bound of Sample Average Deviation

Define the sample average deviation as

$$\Delta_N = \frac{1}{N} \sum_{i=1}^N \left[ \alpha Y^{(i)\top} \mathcal{P}(\widehat{F}) Y^{(i)} - Z^{(i)\top} (\widehat{F} - W) Z^{(i)} \right] - \mathbb{E} \left[ \alpha(Y^{(i)})^\top \mathcal{P}(\widehat{F}) Y^{(i)} - (Z^{(i)})^\top (\widehat{F} - W) Z^{(i)} \right].$$

By Assumption 1,  $Y^{(i)}$  and  $Z^{(i)}$  are composed of independent and identically distributed random variables, so the elements of  $\Delta_N$  are sub-exponential random variables. By the matrix-valued sub-exponential tail bound (Vershynin, 2018) and Union Bound, we have

$$\mathbb{P}(\|\Delta_N\|_F \leq \epsilon_1) \geq 1 - \frac{\beta}{2},$$

where  $\epsilon_1 = c_1 \cdot \frac{(n+m)\ln(2/\beta)}{\sqrt{NK} \cdot \sigma_d \cdot \gamma}$  for some constant  $c_1 > 0$  dependent on the noise covariance matrices  $\Sigma$  and  $\sigma$ . The term  $1/\sqrt{NK}$  arises from the law of large numbers,  $(n+m)$  corrects for the dimension of  $Z^{(i)}$ , and  $\sigma_d \cdot \gamma$  reflects the persistent excitation of the data.

By the expectation decomposition (35), ignoring the non-negative additive noise expectation term  $\mathbb{E}[W^\top \mathcal{P}(F) W]$ , the sample average approximates the expectation, ensuring the covariance of  $\Delta_N$  is bounded by  $\sigma_d^2 K$ , i.e.,  $\text{Tr}(\text{Cov}(\Delta_N)) \leq \sigma_d^2 K$ . (For the definition

of the covariance matrix of a random variable with matrix values, please refer to (Vershynin, 2018).)

#### Step 2: Linear Relationship Between $\Delta F$ and $\Delta_N$

Let  $\Delta F = \widehat{F} - F^*$  denote the deviation between the model-free optimal dual variable  $\widehat{F}$  and the true optimal dual variable  $F^* = H^*$ . Define the expectation term as

$$G(F) = \mathbb{E} \left[ \alpha(Y^{(i)})^\top \mathcal{P}(F) Y^{(i)} - (Z^{(i)})^\top (F - W) Z^{(i)} \right].$$

Since  $F^*$  is the optimal solution of Problem 4, it satisfies the stationary condition  $G(F^*) = 0$ . For small  $\Delta F$  (valid when  $N, K$  are sufficiently large), we perform a first-order Taylor expansion of the expectation term around  $F^*$ :

$$G(F)_{F=\widehat{F}} \approx \nabla_F G(F)_{F=F^*} \cdot \Delta F.$$

By the definition of  $\Delta_N$  and the law of large numbers, substituting the Taylor expansion gives

$$\Delta_N \approx \nabla_F G(F)_{F=F^*} \cdot \Delta F.$$

Split  $G(F)$  into quadratic and linear terms. Taking the partial derivatives of the quadratic terms in blocks and combining them with the derivatives of the linear terms, the gradient  $\nabla_F G(F)$  is exactly the linear combination of the covariance evolution terms, leading to:

$$\nabla_F G(F)_{F=F^*} = I - \alpha C_{\widehat{L}^*}.$$

Thus, the linear equation holds:

$$(I - \alpha C_{\widehat{L}^*}) \cdot \Delta F = \Delta_N.$$

#### Step 3: Bound of Dual Variable Deviation $\|\Delta F\|_F$

Taking the Frobenius norm of both sides of the linear equation and using the matrix norm inequality  $\|AX\|_F \geq \sigma_{\min}(A) \cdot \|X\|_F$ , we have:

$$\sigma_{\min}(I - \alpha C_{\widehat{L}^*}) \cdot \|\Delta F\|_F \leq \|\Delta_N\|_F.$$

By the stability condition  $\sigma_{\min}(I - \alpha C_{\widehat{L}^*}) \geq \underline{\sigma} > 0$ , rearranging gives:

$$\|\Delta F\|_F \leq \frac{\|\Delta_N\|_F}{\underline{\sigma}}.$$

Combining with the probability bound of  $\Delta_N$  from Step 1, the Union Bound yields

$$\mathbb{P}\left(\|\Delta F\|_F \leq \frac{\epsilon_1}{\underline{\sigma}}\right) \geq 1 - \beta.$$

#### Step 4: Bound of Gain Estimation Error $\|\Delta L\|_F$

Let  $\Delta L = \widehat{L}^* - L^*$ , with  $L^* = -(F_{22}^*)^{-1}(F_{12}^*)^\top$  and  $\widehat{L}^* = -(\hat{F}_{22})^{-1}(\hat{F}_{12})^\top$ . Let  $\Delta F_{12} = \hat{F}_{12} - F_{12}^*$  and  $\Delta F_{22} = \hat{F}_{22} - F_{22}^*$ . Expanding  $\Delta L$  using the matrix inverse perturbation formula gives

$$\Delta L = -(\hat{F}_{22})^{-1}\Delta F_{12}^\top + (\hat{F}_{22})^{-1}\Delta F_{22} \cdot (F_{22}^*)^{-1}(F_{12}^*)^\top.$$

Since  $F_{22}^* \succ 0$ , for sufficiently small  $\|\Delta F_{22}\|_F$ ,  $\hat{F}_{22}$  is invertible with  $\|(\hat{F}_{22})^{-1}\|_2 \leq \frac{2}{\sigma_{\min}(F_{22}^*)}$ . Using the triangle inequality  $\|A+B\|_F \leq \|A\|_F + \|B\|_F$  and operator norm inequality  $\|AB\|_F \leq \|A\|_2\|B\|_F$ , we obtain

$$\|\Delta L\|_F \leq c \cdot \|\Delta F\|_F, \quad (41)$$

where  $c = \frac{2(1+2\|L^*\|_2/\sigma_{\min}(F_{22}^*))}{\sigma_{\min}(F_{22}^*)}$  is a constant independent of  $N, K$ .

Substituting the bound of  $\|\Delta F\|_F$  from Step 3 and suppressing the logarithmic factor  $\ln(2/\beta)$  with  $\tilde{O}$ , we get

$$\|\widehat{L}^* - L^*\|_F \leq \tilde{O} \left( \frac{(n+m)(1+\|F^*\|_F)}{\sqrt{NK} \cdot \sigma_d \cdot \underline{\sigma}} \right).$$

This completes the proof.

**Remark 6** The error bound shows that the estimation error decays at a rate of  $\mathcal{O}(1/\sqrt{NK})$ , which is consistent with the statistical efficiency of data-driven control algorithms (Lai et al., 2023). Moreover, based on the bounds for the estimation error given in Theorem 4, the sample complexity of Algorithm 2 can be analyzed, i.e., the total sample count  $s = NK$  satisfies  $s = \tilde{O} \left( \frac{(n+m)^2}{\varepsilon^2} \right)$ . Unlike (Lai et al., 2023), the proof method in this paper transforms the error into a linear system transfer problem via dual variables. It unifies error boundedness and system stability analysis through the closed-loop matrix minimum singular value, explicitly quantifying the relationship between key factors (such as sample complexity and data sustained excitation) and the error. This approach eliminates reliance on Q-function modeling.

**Theorem 5** Assume that the error between the estimated dual optimal solution  $\hat{F}$  and the true dual optimal solution  $F^* = H^*$  satisfies  $\|\hat{F} - F^*\|_F \leq \varepsilon$ , where  $\varepsilon > 0$  is a sufficiently small constant depending on  $N$  and  $K$  (i.e.,  $\varepsilon \rightarrow 0$  as  $N, K \rightarrow \infty$ ), then the control policy  $u_k = \widehat{L}^* x_k$  is admissible.

**Proof.** For the true optimal gain  $L^*$ , since  $L^* \in \mathcal{L}$ , (Lai et al., 2023, Lemma 1) implies that  $\rho(C_{L^*}) < 1$ .

For the estimated gain  $\widehat{L}^*$ , according to (41), the estimation error  $\|\hat{F} - F^*\|_F \leq \varepsilon$  implies

$$\|\widehat{L}^* - L^*\|_F \leq c\varepsilon.$$

That is, there exists a  $\delta = c\varepsilon$  such that  $\widehat{L}^* \in \mathcal{B}_\delta(L^*) \triangleq \{L \in \mathcal{R}^{m \times n} \mid \|L - L^*\|_F \leq \delta\}$ , where  $c$  is a positive constant given in Theorem 4. From the continuity of matrix operations and the continuity of the spectral radius in (Boyd & Vandenberghe, 2004), we obtain  $\rho(C_{\widehat{L}^*}) < 1$ , for all  $\widehat{L}^* \in \mathcal{B}_\delta(L^*)$ . According to (Lai et al., 2023, Lemma 1) and Definition 2, the control policy  $u_k = \widehat{L}^* x_k$  is admissible. This completes the proof.  $\square$

**Remark 7** Theorem 5 aligns with Theorem 3 in (Lai et al., 2023) in core logic: both rely on the continuity of stability with respect to control gain and the boundedness of estimation error. The key distinction lies in the gain estimation source— $\widehat{L}^*$  in this paper is derived from non-iterative SDP optimization rather than iterative RL, but the admissibility proof leverages the same equivalence between spectral radius condition and MSS, and the small estimation error assumption. This confirms that Algorithm 2 not only achieves high estimation accuracy but also preserves closed-loop stability, ensuring practical applicability.

## 6 Simulation

Consider the pulse-width modulated inverter system used in the uninterruptible power supply studied in (Jung & Tzou, 1996), and assume that it is perturbed by both multiplicative and additive noises. Then the discrete-time dynamical system is obtained via Euler discretization with step size  $\delta t = 0.1s$ , which can be described by system (1) with  $A = \begin{bmatrix} 0.6929 & 8.6545 \\ -0.0241 & 0.8603 \end{bmatrix}$ ,

$A_1 = \begin{bmatrix} 0.01 & 0.02 \\ -0.001 & 0.05 \end{bmatrix}$ ,  $B = \begin{bmatrix} 0.1290 & 0.0267 \end{bmatrix}^\top$ , and  $B_1 = \begin{bmatrix} -0.02 & 0.005 \end{bmatrix}^\top$ . Assume that  $v_k$  and  $w_k$  follow the standard Gaussian distributions. Select  $Q = I$ ,  $R = 0.00001$  in the cost functional (2). Choose  $\alpha = 0.5$ . Then, the optimal feedback control gain  $L^*$  can be computed by the model-based SDP Algorithm 1 as  $L^* = \begin{bmatrix} -4.8599 & -64.0491 \end{bmatrix}$ , and the solution  $P^*$  of DGARE (3) can be obtained by  $P^* = M^*$  in Theorem 2 as  $P^* = \begin{bmatrix} 1.0215 & 0.1206 \\ 0.1206 & 1.6917 \end{bmatrix}$ , with the residual  $\|e(P^*)\|_F = 5.8025 \times 10^{-4}$ , where  $e(P) \triangleq P - \mathcal{R}(P)$ .

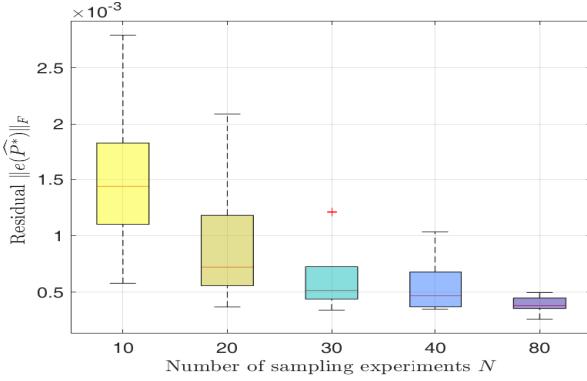


Fig. 2. Residuals  $\|e(\widehat{P}^*)\|_F$  of Algorithm 2 with different values of  $N$ .

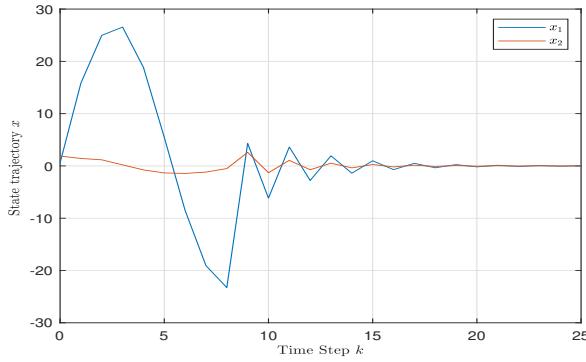


Fig. 3. A numerical average of state trajectories in data collection phase and after learning.

The model-free Algorithm 2 is employed to estimate the optimal feedback gain in the steps that follow. Choose the initial state  $x_0$  from a Gaussian distribution with mean vector  $\mu_0 = [1, 2]^\top$  and covariance matrix  $\Sigma_0 = 5I_2 \succ 0$ . Let  $K = 9$ . Select 5 values of  $N$ , that is, 10, 20, 30, 40 and 80. For each value of  $N$ , Algorithm 2 is run for 10 times, in which  $w_k$  and  $v_k$  are generated randomly for each time. Then, CVX is used to solve Problem 5 each time, and the estimates of  $L^*$  and  $P^*$  can be obtained, which are denoted as  $\widehat{L}^*$  and  $\widehat{P}^*$ , respectively. Fig. 2 shows the residual  $\|e(\widehat{P}^*)\|_F$  of Algorithm 2 with different values of  $N$ . One can observe that the average of  $\|e(\widehat{P}^*)\|_F$  decreases as the value of  $N$  increases. When  $N = 20$ , the average of  $\|e(\widehat{P}^*)\|_F$  can achieve an order of  $10^{-4}$ . One can observe from the simulation results that, although Algorithm 2 does not know the information about the system dynamics, the estimate of the optimal control gain obtained is very close to the optimal control gain obtained from the model-based SDP Algorithm 1, and the model-free Algorithm 2 performs very effectively. Fig. 3 shows the state trajectories derived from the numerical average of the data collected from  $N = 80$  sampling experiments during the model-free design approach, where the trajectories in time domains  $\{0, \dots, 8\}$  are the state trajectories gathered dur-

ing the data collection phase of the reinforcement learning process, and starting at moment 9, the system begins to apply the model-free optimal controller to generate optimal state trajectories. The results confirm that the proposed model-free algorithm is able to efficiently learn the optimal control gain and stabilize the system.

## 7 Conclusions

In this paper, a novel model-free SDP method for solving the stochastic LQR problem has been proposed. The controller design scheme is based on an SDP with LMI constraints, which is sample-efficient and non-iterative. Moreover, it neither requires an initial stabilizing controller to be provided, nor does it require multiplicative and additive noises to be measurable. The robustness of the model-free SDP method to errors has been investigated. In addition, the proposed algorithm has been verified on a pulse-width modulated inverter system. The results confirm that the proposed algorithm is able to efficiently learn the optimal control gains and stabilize the systems.

## References

- Bellman, R. (1966). Dynamic programming. *Science*, 153(3731), 34–37.
- Bertsekas, D. (2019). *Reinforcement learning and optimal control*. Athena Scientific.
- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.
- Bradtko, S. J., Ydstie, B. E., & Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American control conference* (Vol. 3, pp. 3475–3479).
- Farjadnasab, M., & Babazadeh, M. (2022). Model-free LQR design by Q-function learning. *Automatica*, 137, 110060.
- Gravell, B., Esfahani, P. M., & Summers, T. (2020). Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11), 5283–5298.
- Jiang, X., Wang, Y., Zhao, D., & Shi, L. (2024). Online Pareto optimal control of mean-field stochastic multi-player systems using policy iteration. *Science China Information Sciences*, 67(4), 140202.
- Jiang, Y., Liu, L., & Feng, G. (2024). Adaptive linear quadratic control for stochastic discrete-time linear systems with unmeasurable multiplicative and additive noises. *IEEE Transactions on Automatic Control*, 69(11), 7808–7815.
- Jung, S.-L., & Tzou, Y.-Y. (1996). Discrete sliding-mode control of a PWM inverter for sinusoidal output waveform synthesis with optimal sliding curve. *IEEE Transactions on Power Electronics*, 11(4), 567–577.

- Kubrusly, C., & Costa, O. (1985). Mean square stability conditions for discrete stochastic bilinear systems. *IEEE Transactions on Automatic Control*, 30(11), 1082–1087.
- Lai, J., Xiong, J., & Shu, Z. (2023). Model-free optimal control of discrete-time systems with additive and multiplicative noises. *Automatica*, 147, 110685.
- Lee, D., & Hu, J. (2018). Primal-dual Q-learning framework for LQR design. *IEEE Transactions on Automatic Control*, 64(9), 3756–3763.
- Lewis, F. L., & Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3), 32–50.
- Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal control*. John Wiley & Sons.
- Li, M., Qin, J., Zheng, W. X., Wang, Y., & Kang, Y. (2022). Model-free design of stochastic LQR controller from a primal-dual optimization perspective. *Automatica*, 140, 110253.
- Pang, B., & Jiang, Z.-P. (2021). Robust reinforcement learning: A case study in linear quadratic regulation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 9303–9311).
- Pang, B., & Jiang, Z.-P. (2023). Reinforcement learning for adaptive optimal stationary control of linear stochastic systems. *IEEE Transactions on Automatic Control*, 68(4), 2383–2390.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.
- Vandenberghe, L., & Boyd, S. (1996). Semidefinite programming. *SIAM Review*, 38(1), 49–95.
- Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8, 279–292.
- Yao, D. D., Zhang, S., & Zhou, X. Y. (2001). Stochastic linear-quadratic control via semidefinite programming. *SIAM Journal on Control and Optimization*, 40(3), 801–823.
- Yong, J., & Zhou, X. Y. (2012). *Stochastic controls: Hamiltonian systems and HJB equations* (Vol. 43). Springer Science & Business Media.
- Zhang, W., Guo, J., & Jiang, X. (2025). Model-free  $H_\infty$  control of Itô stochastic system via off-policy reinforcement learning. *Automatica*, 174, 112144.
- Zhang, W., Xie, L., & Chen, B.-S. (2017). *Stochastic  $H_2/H_\infty$  control: A Nash game approach*. CRC Press.