Primal-Dual Q-Learning Framework for LQR Design

Donghwan Lee and Jianghai Hu

Abstract—**Recently, reinforcement learning (RL) is receiving more and more attentions due to its successful demonstrations outperforming human performance in certain challenging tasks. The goal of this paper is to study a new optimization formulation of the linear quadratic regulator (LQR) problem via the Lagrangian duality theories in order to lay theoretical foundations of potentially effective RL algorithms. The new optimization problem includes the Q-function parameters so that it can be directly used to develop Q-learning algorithms, known to be one of the most popular RL algorithms. We prove relations between saddle-points of the Lagrangian function and the optimal solutions of the Bellman equation. As an example of its applications, we propose a model-free primal-dual Q-learning algorithm to solve the LQR problem and demonstrate its validity through examples.**

I. Introduction

The linear quadratic regulator (LQR) problem [1], [2] for linear time-invariant (LTI) systems has a long tradition, and is well understood nowadays. A standard approach is the dynamic programming [1] to solve the Bellman equation or algebraic Riccati equation (ARE). With the development of convex optimization [3] and semidefinite programming (SDP) techniques [4], the LQR problem has been revisited in terms of convex analysis and SDPs in many researches, e.g., [5]–[8]. Since the SDP is a convex optimization, standard Lagrangian duality results in [3], [4] can be used to formulate this as a saddle-point problem. Such connections have been comprehensively studied in [7]–[10].

On the other hand, reinforcement learning (RL) [11], [12] is a subfield of machine learning which addresses the problem of how an autonomous agent can learn an optimal policy to minimize long-term cumulative costs, while interacting with unknown environments. For LTI systems, RL was studied in [13], [14] to solve the LQR problem with the recursive least-square algorithm. Many classical RL algorithms, e.g., temporal difference methods [15], Q-learning [16], SARSA [17], are based on the sample-based stochastic dynamic programming to solve the Bellman equation, taking advantage of its contraction mapping or monotone property to guarantee their convergence. Despite the generality of RL frameworks, they are yet to directly handle constraints and various objectives. Therefore, the integration of the Bellman equation with optimization frameworks is worthwhile to study for more practical RL algorithms by leveraging the existing fruitful optimization algorithms and theories. Although RL for LQR design appears to be very well understood currently (e.g., [13], [14]), to the authors' knowledge, such optimization and duality interpretations of RL remain understudied so far. This situation motivates some questions: *how can we formulate an optimization of the LQR and understand its duality results for effective RL algorithms, especially, in terms of Q-learning?; how can we develop a Q-learning based on the optimization formulation?*

To answer the questions, we propose fundamental Lagrangian duality frameworks of the standard LQR associated with Q-learning [16], which is known to be one of the most popular RL algorithms. In particular, we derive a new optimization formulation of the LQR problem, which includes the Q-function parameters, and analyze its Lagrangian duality results [3] in Section III. We prove the relations among primal-dual solutions of the proposed optimization, the solution of the standard ARE, and the state-input trajectories. Some extensions to the present LQR results are pursued through the interplay between the system trajectories and the dual parameters. In Section IV, we propose a model-free primal-dual Q-learning algorithm that recovers an optimal policy using a collection of trajectories of an unknown system. Our results build upon the previous results [7], [8], [10], which studied dualities in terms of SDPs. However, we note that our optimization formulations are different from the existing ones in that ours include the Q-function [1] parameters, and thus, can be directly used to develop a new class of Q-learning algorithms based on primal-dual updates. The main results can be extended in several directions, e.g., input and energy constrained optimal control design and structured controller design, as discussed in the supplemental material [18]. We expect that this fundamental framework advances our understanding of RL and Q-learning for the LQR problem and will be useful to develop many primal-dual RL algorithms based on the SDP formulations [6].

Notation: The adopted notation is as follows: $\mathbb{N}$ and $\mathbb{N}_+$: sets of nonnegative and positive integers, respectively; $\mathbb{R}$: set of real numbers; $\mathbb{R}_+$: set of nonnegative real numbers; $\mathbb{R}_{++}$: set of positive real numbers; $\mathbb{R}^n$: n-dimensional Euclidean space; $\mathbb{R}^{n\times m}$: set of all $n\times m$ real matrices; A^T: transpose of matrix A; $A\succ 0$ ($A\prec 0$, $A\succeq 0$, and $A\preceq 0$, respectively): symmetric positive definite (negative definite, positive semi-definite, and negative semi-definite, respectively) matrix A; I_n: $n\times n$ identity matrix; $\mathbb{S}^n$: symmetric $n\times n$ matrices; $\mathbb{S}_+^n$: cone of symmetric $n\times n$ positive semi-definite matrices; $\mathbb{S}_{++}^n$: symmetric $n\times n$ positive definite matrices; $\mathbf{Tr}(A)$: trace of matrix A; $\rho(\cdot)$: spectral radius.

II. Problem Formulation and Preliminaries

A. Infinite-Horizon LQR Problem

Consider the LTI system

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

where $k\in\mathbb{N}$, $x(k)\in\mathbb{R}^n$ is the state vector, $u(k)\in\mathbb{R}^m$ is the input vector, and $z\in\mathbb{R}^n$ is the initial state.

Assuming the control $u(k)$ is given by a state-feedback control policy $u(k)=Fx(k)$, we denote by $x(k;F,z)$ the solution of (1) starting from $x(0)=z$. Under the state-feedback control policy, the cost function for the classical LQR problem is denoted by

$$J(F,z):=\sum_{k=0}^{\infty}\begin{bmatrix}x(k;F,z)\\Fx(k;F,z)\end{bmatrix}^T\Lambda\begin{bmatrix}x(k;F,z)\\Fx(k;F,z)\end{bmatrix}, \tag{2}$$

where $\Lambda:=\begin{bmatrix}Q & 0\\0 & R\end{bmatrix}\succeq 0$ is the weight matrix.

Remark 1. *In the discounted LQR problem, each term in (2) is multiplied by α^k with $\alpha\in(0,1)$, which is called the discount factor. Throughout the paper, we only consider the case $\alpha=1$ for simplicity, and all the results in this paper hold for the case $\alpha\in(0,1)$ by replacing (A,B) with $(\alpha^{1/2}A,\alpha^{1/2}B)$.*

*This material is based upon work supported by the National Science Foundation under Grant No. 1539527

D. Lee is with the Coordinated Science Laboratory (CSL), University of Illinois at Urbana-Champaign, IL 61801, USA donghwan@illinois.edu.

J. Hu is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA jianghai@purdue.edu.

By introducing the augmented state vector $v(k) := \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}$, we will consider the augmented system

$$v(k+1) = A_F v(k), \quad v(0) = v_0 \in \mathbb{R}^{n+m}, \tag{3}$$

where $A_F := \begin{bmatrix} A & B \\ FA & FB \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$. If $v_0 = \begin{bmatrix} z^T & z^T F^T \end{bmatrix}^T$, then the state and input parts of $v(k)$ are identical to $x(k)$ and $u(k)$ in (1). A useful property of A_F is that its spectral radius $\rho(A_F)$ is identical to that of $A + BF$.

Lemma 1. *$\rho(A+BF) = \rho(A_F)$ holds.*

Proof. Note that $\Omega := \begin{bmatrix} I_n & 0 \\ F & I_m \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$ is a nonsingular matrix with its inverse $\Omega^{-1} = \begin{bmatrix} I_n & 0 \\ -F & I_m \end{bmatrix}$. Then, we have

$$\rho(A_F) = \rho(\Omega^{-1} A_F \Omega) = \rho\left(\begin{bmatrix} A+BF & B \\ 0 & 0 \end{bmatrix}\right) = \rho(A+BF),$$

and the desired result follows. □

Define $\mathcal{F}$ as the set of all stabilizing state-feedback gains of system (A, B), i.e., $\mathcal{F} := \{F \in \mathbb{R}^{m\times n} : \rho(A+BF) < 1\}$. $\mathcal{F}$ is an open set, not necessarily convex [19, Lemma 2]; however, finding a state feedback gain $F \in \mathcal{F}$ can be reduced to a simple convex problem. In this paper, we study the infinite-horizon LQR problem.

Problem 1 (Infinite-horizon LQR problem)**.** *Suppose that $z_i \in \mathbb{R}^n, i \in \{1, 2, \ldots, r\}$, are chosen such that $\sum_{i=1}^r z_i z_i^T = Z \succ 0$, where $r \in \mathbb{Z}_+$. Solve $F^* = \arg\min_{F\in\mathcal{F}} \sum_{i=1}^r J(F, z_i)$ if the optimal value of $\inf_{F\in\mathcal{F}} \sum_{i=1}^r J(F,\, z_i)$ exists and is attained, where $J(\cdot,\cdot)$ is defined in (2).*

Remark 2. *From the standard LQR theory, although $J^*(F, z)$ has different values for different $z \in \mathbb{R}^n$, the minimizer $F^* = \arg\min_{F\in\mathcal{F}} J(F, z)$ is not dependent on z. Therefore, it follows that $\arg\min_{F\in\mathcal{F}} J(F, z) = \arg\min_{F\in\mathcal{F}} \sum_{i=1}^r J(F, z_i)$ for any $z, z_i \in \mathbb{R}^n, i \in \{1, 2, \ldots, r\}$. For technical reasons that will become clear later, we solve $\arg\min_{F\in\mathcal{F}} \sum_{i=1}^r J(F, z_i)$ instead of $\arg\min_{F\in\mathcal{F}} J(F, z)$. Throughout the paper, we always assume that $Z \succ 0$.*

For a given $z \in \mathbb{R}^n$, if the optimal value of $\inf_{F\in\mathcal{F}} J(F, z)$ exists and is attained, then the optimal cost is denoted by $J^*(z) = J(F^*, z)$. Assumptions that will be used throughout the paper are summarized below.

Assumption 1.

- *$Q \succeq 0, R \succ 0$;*
- *(A, B) is stabilizable and Q can be written as $Q = C^T C$, where (A, C) is detectable.*

Under Assumption 1, the optimal value of $\inf_{F\in\mathcal{F}} J(F,\, z)$ exists, is attained, and $J^*(z)$ is a quadratic function, i.e., $J^*(z) = z^T X^* z$, where $X = X^*$ is the unique solution of the algebraic Riccati equation (ARE) [1, Proposition 4.4.1]

$$X = A^T X A - A^T X B (R + B^T X B)^{-1} B^T X A + Q, \quad X \succeq 0.$$

In this case, $J^*(z)$ as a function of $z \in \mathbb{R}^n$ is called the optimal value function.

The reader can refer to [1] and [20] for more details of the classical LQR results. The corresponding optimal control policy is $u^*(z) = F^* z$, where

$$F^* := -(R + B^T X^* B)^{-1} B^T X^* A \in \mathcal{F} \tag{4}$$

is the unique optimal gain. Alternatively, the Q-function [1] is defined as

$$Q^*(z, u) := z^T Q z + u^T R u + J^*(Az + Bu) = \begin{bmatrix} z \\ u \end{bmatrix}^T P^* \begin{bmatrix} z \\ u \end{bmatrix}, \tag{5}$$

where

$$P^* := \begin{bmatrix} Q + A^T X^* A & A^T X^* B \\ B^T X^* A & R + B^T X^* B \end{bmatrix}. \tag{6}$$

The optimal policy is then given by

$$u^*(z) = F^* z = \arg\min_{u\in\mathbb{R}^m} Q^*(z,\, u).$$

B. Useful Lemmas

Standard Lyapunov theorems for discrete-time LTI systems will be used extensively in this paper, which are listed below.

Lemma 2 (Lyapunov stability theorems [21, Chapter 3], [22, Theorem 5.D6])**.** *Let $A \in \mathbb{R}^{n\times n}$.*

1) *if $\rho(A) < 1$, then for any $Z \in \mathbb{S}_+^n$, $A^T P A + Z = P$ has a unique solution $P \in \mathbb{S}_+^n$.*
2) *$\rho(A) < 1$ if and only if for each given matrix $Z \in \mathbb{S}_{++}^n$, there exists $P \in \mathbb{S}_{++}^n$ such that $A^T P A + Z = P$. If such a P exists, then it is unique.*
3) *Suppose that (A, C) is observable (resp. detectable). Then, $\rho(A) < 1$ if and only if there exists $P \in \mathbb{S}_{++}^n$ (resp. $P \in \mathbb{S}_+^n$) such that $A^T P A + C^T C = P$. If such a P exists, then it is unique.*
4) *Suppose that (A, B) is reachable (resp. stabilizable). Then, $\rho(A) < 1$ if and only if there exists $P \in \mathbb{S}_{++}^n$ (resp. $P \in \mathbb{S}_+^n$) such that $APA^T + BB^T = P$. If such a P exists, then it is unique.*

Moreover, the Schur complement will be useful and its special form will be used in this paper.

Lemma 3 (Schur complement ([23, Theorem 1.12]))**.** *Let P be a symmetric matrix partitioned as $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}$, in which P_{22} is square and nonsingular. Then, $P \succeq 0$ if and only if $P_{22} \succ 0$ and $P_{11} - P_{12} P_{22}^{-1} P_{12}^T \succeq 0$.*

Throughout the paper, we will use the partition $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}$ for the matrix P, where $P_{11} \in \mathbb{S}^n, P_{12} \in \mathbb{R}^{n\times m}$, $P_{22} \in \mathbb{S}^m$. Similarly, we use the notation $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix}$ with $S_{11} \in \mathbb{S}^n, S_{22} \in \mathbb{S}^m, S_{12} \in \mathbb{R}^{n\times m}$.

III. Optimization Formulation and Duality

In this section, we propose a novel optimization formulation of the LQR problem. The new formulation includes the Q-function parameters in its dual form, and hence can be directly used to develop a new primal-dual Q-learning algorithm. Moreover, some important extensions to the present LQR theory can be successfully pursued through the interplay between the system trajectories and the dual parameters. Throughout the section, we will focus on the three optimization problems.

Problem 2 (Primal Problem I)**.** *Solve*

$$J_p := \inf_{S\in\mathbb{S}^{n+m}, F\in\mathbb{R}^{m\times n}} \mathbf{Tr}(\Lambda S)$$

$$\text{subject to} \quad F \in \mathcal{F}, \tag{7}$$

$$A_F S A_F^T + \begin{bmatrix} I_n \\ F \end{bmatrix} Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T = S. \tag{8}$$

Problem 3 (Primal Problem II)**.** *Solve*

$$J_p' := \inf_{P\in\mathbb{S}^{n+m},\, F\in\mathbb{R}^{m\times n}} \mathbf{Tr}\left(\begin{bmatrix}I_n\\F\end{bmatrix} Z \begin{bmatrix}I_n\\F\end{bmatrix}^T P\right)$$

$$\text{subject to} \quad F\in\mathcal{F},$$

$$A_F^T P A_F + \Lambda = P. \tag{9}$$

Problem 4 (Dual Problem)**.** *Solve*

$$J_d := \sup_{P\in\mathbb{S}^{n+m}} d(P) = \sup_{P\in\mathbb{S}^{n+m}} \inf_{S\in\mathbb{S}_+^{n+m}, F\in\mathcal{F}} L(P,F,S), \tag{10}$$

where

$$L(P,F,S) = \mathbf{Tr}(\Lambda S) + \mathbf{Tr}\left(\left(A_F S A_F^T - S + \begin{bmatrix}I_n\\F\end{bmatrix} Z \begin{bmatrix}I_n\\F\end{bmatrix}^T\right) P\right).$$

We first summarize overall flow of this section, which is visualized in Figure 1.

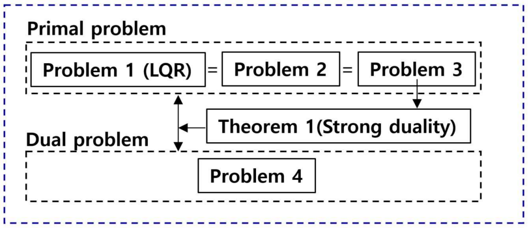

Fig. 1. Diagram for relations among the results.

We will first prove that Problem 2 is an equivalent constrained optimization formulation of Problem 1. Now that an optimization formulation of Problem 1 has been obtained in Problem 2, it will be useful to study its dual problem. In particular, define the Lagrangian function of Problem 2 given in Problem 4, where $P\in\mathbb{S}^{n+m}$ is the Lagrange multiplier. In the Lagrangian formulation, the constraint $F\in\mathcal{F}$ does not appear because it is not an explicit equality or inequality constraint. Instead, we will treat $\mathcal{F}$ as the domain of the variable F. Moreover, since any feasible S satisfying Problem 2 is positive semidefinite, we set $\mathbb{S}_+^{n+m}$ to be the domain of the variable S as well. Rearranging some terms, the Lagrangian function can be written as

$$L(P,F,S) = \mathbf{Tr}\left(Z\begin{bmatrix}I_n\\F\end{bmatrix}^T P \begin{bmatrix}I_n\\F\end{bmatrix}\right) + \mathbf{Tr}((A_F^T P A_F - P + \Lambda)S). \tag{11}$$

The Lagrangian dual function is defined as $d(P) := \inf_{S\in\mathbb{S}_+^{n+m}, F\in\mathcal{F}} L(P,F,S)$. Then, the dual problem of Problem 2 is given by Problem 4. From the weak duality [3, Chapter 5], $J_d \le J_p$ holds with $J_p - J_d \ge 0$ being the duality gap. If the duality gap is zero, then it is said that the *strong duality* holds for the optimization. We will prove that this is indeed the case.

Theorem 1 (Strong duality)**.** *We have $J_p = J_d$.*

The main goal of this section is to prove Theorem 1. We note that Problem 2 is nonconvex because the set $\mathcal{F}$ in (7) is not convex and the equality constraint (8) is not linear. For general optimization problems, the strong duality holds if certain constraint qualifications hold, for instance, the Slater's condition [3, Chapter 5]. Unfortunately, since Problem 2 is nonconvex, it cannot be applied to our case, which makes the proof of the strong duality nontrivial. Instead, we will prove the strong duality by introducing Problem 3 which is equivalent to Problem 2. Throughout the paper, the following shorthands will be used:

$$J(F) := \sum_{i=1}^r J(F,z_i), \quad J^* := \sum_{i=1}^r J(F^*,z_i), \tag{12}$$

where F^* is the optimal gain in (4).

A. Properties of Problem 2

We summarize results of this subsection. All proofs are in Appendix.

Proposition 1. *The optimal solution of Problem 2 is attained at a unique point (S_p, F_p). In addition, Problem 2 is equivalent to Problem 1 in the sense that $J_p = J^*$, and $F_p = F^*$.*

Proposition 2. *Any feasible solution $(S,F)\in\mathbb{S}_+^{n+m}\times\mathbb{R}^{m\times n}$ of Problem 2 satisfies the followings:*

1) $(A+BF)S_{11}(A+BF)^T + Z = S_{11}$;

2) $\begin{bmatrix}I_n\\F\end{bmatrix} S_{11} \begin{bmatrix}I_n\\F\end{bmatrix}^T = S$;

3) $F = S_{12}^T S_{11}^{-1}$.

Proposition 3. *In Problem 2, the constraint $F\in\mathcal{F}$ can be replaced by $S\succeq 0$ without changing its optimal solution and optimal objective function value.*

Proposition 1 states that Problem 2 is equivalent to Problem 1. Additional properties of the solution of Problem 2 are summarized in Proposition 2. Later, we prove that if $(S,F)\in\mathbb{S}_+^{n+m}\times\mathbb{R}^{m\times n}$ is the solution of Problem 2, then S can be constructed from the trajectories generated under the policy $u(z) = Fz$. Therefor, the statement 3) provides a way to recover the state-feedback gain from the trajectories without the model knowledge. In Problem 2, (7) is a non-convex constraint. A natural question arises: is it possible to drop this constraint? Without (7), a possible solution of Problem 2 can have $S\not\succeq 0$ with $F\notin\mathcal{F}$. In particular, if $F\notin\mathcal{F}$, then $\rho(A_F)\ge 1$, and by the contraposition of 2) of Lemma 2, we have $S\notin\mathbb{S}_{++}^n$. With (7), any feasible solution S is guaranteed to be positive semidefinite. This implies that the constraint $S\succeq 0$ is implicitly imposed in Problem 2. More importantly, the constraint $F\in\mathcal{F}$ in Problem 2 can be replaced by $S\succeq 0$ without changing its optimal solution and optimal objective function value, which is the claim of Proposition 3.

B. Properties of Problem 3

We introduce two results for Problem 3 and defer all proofs to Appendix.

Proposition 4. *The optimal solution of Problem 3 is attained at the unique point (P_p', F_p'). In addition, Problem 3 is equivalent to Problem 1, i.e., $J_p' = J^*$, and $F_p' = F^*$.*

Proposition 5. *$P_p' = P^*$, and the optimal solution (P_p', F_p') of Problem 3 satisfies $P_p' \succeq 0$, $P_{p,22}' \succ 0$, and $F_p' = -(P_{p,22}')^{-1}(P_{p,12}')^T \in \mathcal{F}$.*

By Proposition 4, we have $F_p' = F^*$. However, it did not state that $P_p' = P^*$ holds as well, which is proved in Proposition 5. Overall, these results prove the equivalence of Problem 1 and Problem 3. In the next subsection, we provide a proof of Theorem 1 by using Problem 3.

C. *Proof of Theorem 1*

To prove Theorem 1, the following lemma will be used.

Lemma 4. *If $P \succeq 0, P_{22} \succ 0$, then*

$$\begin{bmatrix} I_n \\ F \end{bmatrix}^T P \begin{bmatrix} I_n \\ F \end{bmatrix} \succeq P_{11} - P_{12}P_{22}^{-1}P_{12}^T = \begin{bmatrix} I_n \\ P_{22}^{-1}P_{12}^T \end{bmatrix}^T P \begin{bmatrix} I_n \\ P_{22}^{-1}P_{12}^T \end{bmatrix},$$

$\forall F \in \mathbb{R}^{m\times n}$,

and the equality holds if and only if $F = -P_{22}^{-1}P_{12}^T$.

Proof. Noting $\begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix} = \Omega \begin{bmatrix} P_{11} - P_{12}P_{22}^{-1}P_{12}^T & 0 \\ 0 & P_{22} \end{bmatrix} \Omega^T$, where $\Omega = \begin{bmatrix} I_n & P_{12}P_{22}^{-1} \\ 0 & I_m \end{bmatrix}$, a direct calculation leads to

$$\begin{aligned} &\begin{bmatrix} I_n \\ F \end{bmatrix}^T P \begin{bmatrix} I_n \\ F \end{bmatrix} \\ &= P_{11} - P_{12}P_{22}^{-1}P_{12}^T + (P_{22}^{-1}P_{12}^T + F)P_{22}(P_{22}^{-1}P_{12}^T + F)^T \\ &\succeq P_{11} - P_{12}P_{22}^{-1}P_{12}^T, \end{aligned}$$

and the equality holds only if $F = -P_{22}^{-1}P_{12}^T$. □

Now, we are in position to prove Theorem 1.

Proof of Theorem 1. For a given $P \in \mathbb{S}^{n+m}$, the Lagrangian dual function $d(P)$ is

$$\begin{aligned} d(P) &= \inf_{F\in\mathcal{F}} \inf_{S\in\mathbb{S}_+^{n+m}} L(P,F,S) \\ &= \begin{cases} \inf_{F\in\mathcal{F}} \mathbf{Tr}\left(Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T P \begin{bmatrix} I_n \\ F \end{bmatrix} \right) & \text{if } P \in \mathcal{P} \\ -\infty \quad \text{otherwise} \end{cases} \end{aligned} \tag{13}$$

where $\mathcal{P} := \{P \in \mathbb{S}_+^{n+m} : A_F^T P A_F - P + \Lambda \succeq 0, \forall F \in \mathcal{F}\}$. We next show that the solution P_p' of Problem 3 is an element of $\mathcal{P}$, and thus, $\mathcal{P}$ is nonempty. By Proposition 5, $A_{F_p'}^T P_p' A_{F_p'} + \Lambda = P_p'$, where $F_p' = -(P_{p,22}')^{-1}(P_{p,12}')^T$. By Lemma 4, $A_F^T P_p' A_F + \Lambda \succeq A_{F_p'}^T P_p' A_{F_p'} + \Lambda = P_p'$ for all $F \in \mathbb{R}^{m\times n}$. Therefore, every $F \in \mathcal{F}$ satisfies $A_F^T P_p' A_F + \Lambda \succeq P_p'$, i.e., $P_p' \in \mathcal{P}$. Therefore, the dual problem is equivalent to

$$\sup_{P\in\mathbb{S}^{n+m}} d(P) = J_d = \sup_{P\in\mathcal{P}} \inf_{F\in\mathcal{F}} \mathbf{Tr}\left(Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T P \begin{bmatrix} I_n \\ F \end{bmatrix} \right). \tag{14}$$

For $P_p' \in \mathcal{P}$, we have

$$d(P_p') = \inf_{F\in\mathcal{F}} \mathbf{Tr}\left(Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T P_p' \begin{bmatrix} I_n \\ F \end{bmatrix} \right). \tag{15}$$

Obviously, $d(P_p') \le J_d$. Since $P_p' \succeq 0$ is fixed and the objective function in (15) is quadratic with respect to F, the infimum in (15) is attained at $F = -(P_{p,22}')^{-1}(P_{p,12}')^T = F_p' \in \mathcal{F}$. Therefore, $J_p' = d(P_p')$, implying $J_p' \le J_d$. On the other hand, by the weak duality, $J_d \le J_p$. By Proposition 4, $J_p = J_p'$. Therefore, we have $J_p = J_d$. □

From the proof of Theorem 1, we have $J_p = J_d = J_p' = \mathbf{Tr}\left(Z \begin{bmatrix} I_n \\ F_p' \end{bmatrix}^T P_p' \begin{bmatrix} I_n \\ F_p' \end{bmatrix} \right)$. Therefore, we easily conclude that $P^* = P_p' \in \arg\sup_{P\in\mathbb{S}^{n+m}} d(P)$ and $(S^*, F^*) = (S_p, F_p) \in \arg\inf_{S\in\mathbb{S}_+^{n+m}, F\in\mathcal{F}} L(P^*, F, S)$. In summary, we conclude that the primal and dual optimal solutions consist of the solution of the ARE. Equivalently, (P^*, F^*, S^*) is a saddle point of the Lagrangian $L(\cdot,\cdot,\cdot)$, i.e., $L(P, F^*, S^*) \le L(P^*, F^*, S^*) \le L(P^*, F, S)$ for all $F \in \mathcal{F}, S \in \mathbb{S}_+^{n+m}, P \in \mathbb{S}^{n+m}$.

IV. Primal-Dual Adaptive LQR Design

In this section, we will study how to design the LQR policy without the knowledge of (A, B) as an application of the results in the previous section. The approach can be viewed as a version of the Q-learning algorithm in [13], [14]. We adopt the following assumptions.

Assumption 2.

1) *(A, B) is not known;*
2) *The input and state pair $(x(k), u(k))$, $k \in \mathbb{N}$, can be collected for different control policy $u(k) = Fx(k)$ and initial state $x(0) = z$ as many times as needed;*
3) *An initial state-feedback control gain $F_{\text{stab}} \in \mathcal{F}$ is known.*

Remark 3. *The proposed algorithm is an applications of the analysis given in the previous section. In particular, the proposed algorithm is a primal-dual algorithm [3], [24] to solve saddle point problems and constrained optimization problems. Beside the applicability of the algorithm, we introduce this algorithm to prove that the Q-learning algorithm in [13], [14] can be interpreted as a primal-dual procedure and to prove the connection between the duality analysis in the previous section and the Q-learning.*

In this section, we modify the LQR problem in Problem 1 to develop control design algorithms. Consider the augmented state vector $v(k) := \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}$ and assume that we know the initial state $v(0) = v_0 \in \mathbb{R}^{n+m}$. Denote by $v(k; F, v_0)$ the state trajectory of the augmented system (3) at time k starting from the initial augmented $\begin{bmatrix} x(0) \\ u(0) \end{bmatrix} = v_0$. In this section, we assume that $u(0)$ can be freely chosen, and the control policy $u(k) = Fx(k)$ is valid from $k = 1$. The cost corresponding to the control policy $u(z) = Fz$ is denoted by $\hat{J}(F, v_0) := \sum_{k=0}^{\infty} v(k; F, v_0)^T \Lambda v(k; F, v_0)$, where the control policy $u(k) = Fx(k)$ is applied from time $k = 1$.

Remark 4. *Compared to the cost in (2), only the difference of $\hat{J}(F, v_0)$ is in the degree of freedom in selecting the initial control input $u(0)$ of the augmented system (3). With appropriate selections of the initial state $x(0)$ and input $x(0)$, the second term $\begin{bmatrix} I_n \\ F \end{bmatrix} Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T$ on the left-hand side of (8) can be replaced by a strictly positive definite matrix Γ. Then, by 2) of Lemma 2, any feasible solution of (7) and (8) with $\begin{bmatrix} I_n \\ F \end{bmatrix} Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T$ replaced with $\Gamma \succ 0$ satisfies $S \succ 0$. This strict positive definiteness will bring some benefits in algorithmic developments in the sequel.*

In this case, one can choose $v_i \in \mathbb{R}^{n+m}, i \in \{1, 2, \ldots, r\}$, such that $\sum_{i=1}^{r} v_i v_i^T = \Gamma \succ 0$, where $\Gamma \in \mathbb{S}^{n+m}$. Define

$$\hat{F}^* = \arg\min_{F\in\mathcal{F}} \sum_{i=1}^{r} \hat{J}(F, v_i). \tag{16}$$

First, we claim that $\hat{F}^* = F^*$.

Proposition 6. *The optimal solution $\hat{F}^*$ of (16) is identical to F^*.*

Proof. Using the definition of $\hat{J}(\cdot,\cdot)$ and $J(\cdot,\cdot)$, an algebraic manipulation leads to

$$\sum_{i=1}^{r} \hat{J}(F,\, v_i) = \sum_{i=1}^{r} J(F,\, z_i) + \sum_{i=1}^{r} v_i^T \Lambda v_i, \tag{17}$$

where $z_i = \begin{bmatrix} A & B \end{bmatrix} v_i$. Since the last term on the right-hand side of the above equation is constant, the minimizer in (16) is equivalent to the minimizer of the first term on the right hand side of (17), which is identical to F^* by Remark 2. □

Following steps similar to Proposition 1, it can be proved that the problem in (16) can be converted to

$$\begin{aligned}\hat{J}_p := & \min_{S\in\mathbb{S}^{n+m},\,F\in\mathbb{R}^{m\times n}} \mathbf{Tr}(\Lambda S)\\ \text{subject to} \quad & A_F S A_F^T + \Gamma = S, \quad F \in \mathcal{F}.\end{aligned}$$

Since $\Gamma \succ 0$, $F \in \mathcal{F}$ can be replaced with $S \succ 0$ by 2) of Lemma 2, and we can obtain another equivalent primal problem.

Problem 5 (Primal problem)**.** *Solve*

$$\begin{aligned}\hat{J}_p := & \min_{S\in\mathbb{S}^{n+m},\,F\in\mathbb{R}^{m\times n}} \mathbf{Tr}(\Lambda S)\\ \text{subject to} \quad & A_F S A_F^T + \Gamma = S, \quad S \succ 0.\end{aligned}$$

Introduce a Lagrangian function for Problem 5, i.e., for any fixed $P \in \mathbb{S}^{n+m}$, $P_0 \in \mathbb{S}_+^{n+m}$, define $\hat{L}(P, P_0, F, S) := \mathbf{Tr}(\Lambda S) + \mathbf{Tr}((A_F S A_F^T + \Gamma - S)P) + \mathbf{Tr}(-SP_0)$. The corresponding dual problem is

$$\hat{J}_d := \sup_{P\in\mathbb{S}^{n+m}, P_0\in\mathbb{S}_+^{n+m}} \inf_{S\in\mathbb{S}^{n+m}, F\in\mathbb{R}^{m\times n}} \hat{L}(P, P_0, F, S).$$

Following similar lines as in the previous section, we can prove that the strong duality holds, i.e., $\hat{J}_p = \hat{J}_d$, and the primal and dual optimal points for (P, F) are identical to (P^*, F^*). In other words, (P^*, F^*) is a saddle point of the Lagrangian function $L(P, P_0, F, S)$ with some $P_0 = P_0^*$. It is also known that the saddle point should satisfies the KKT condition [3]. In the following proposition, we derive a KKT condition of Problem 5, which is satisfied by the saddle point (P^*, F^*).

Proposition 7. *Suppose that $(\hat{S}, \hat{F})$ is the primal optimal point and $(\hat{P}, \hat{P}_0)$ is the dual optimal point of Problem 5. Then, $(\hat{S}, \hat{F}, \hat{P})$ satisfies the KKT condition for (S, F, P)*

$$A_F S A_F^T + \Gamma - S = 0, \tag{18}$$

$$S \succ 0, \tag{19}$$

$$A_F^T P A_F - P + \Lambda = 0, \tag{20}$$

$$2(P_{12}^T + P_{22}F)\begin{bmatrix} A & B\end{bmatrix} S \begin{bmatrix} A & B\end{bmatrix}^T = 0. \tag{21}$$

Proof. By 2) of Lemma 2, $\Gamma \succ 0$ guarantees $S \succ 0$. From the KKT condition of the generalized inequality constrained optimization in [3, chapter 5.9.2], the KKT condition of Problem 5 can be summarized as the primal feasibility condition $A_F S A_F^T + \Gamma - S = 0, S \succ 0$, the complementary slackness condition $\mathbf{Tr}(SP_0) = 0$, the dual feasibility condition $P_0 \succeq 0$, and

$$\frac{\partial \hat{L}(P, P_0, S, F)}{\partial S} = A_F^T P A_F - P + \Lambda - P_0 = 0,$$

$$\frac{\partial \hat{L}(P, P_0, S, F)}{\partial F} = 2(P_{12}^T + P_{22}F)\begin{bmatrix} A & B\end{bmatrix} S \begin{bmatrix} A & B\end{bmatrix}^T = 0.$$

Since $S \succ 0$ is guaranteed, the only solution P_0 that satisfies $\mathbf{Tr}(SP_0) = 0$ is $P_0 = 0$. Therefore, the KKT condition in (18)-(21) is obtained. According to [3, Section 5.5.3, pp. 243], the strong duality ensures that any pair of primal and dual optimal points (saddle points) must satisfy the KKT condition. This completes the proof. □

In what follows, we will study procedures to solve the KKT condition. Algorithm 1 iteratively solves (20) and (21). In particular, we will prove that (P_t, F_t) in Algorithm 1 converges to $(\hat{P}, \hat{F})$ that solves (20) and (21), and they are identical to (P^*, F^*) defined in (6) and (4). To prove this, we use the following lemma.

Lemma 5. *Assume $F \in \mathcal{F}$ and define the mapping $\mathcal{T}(P) := A_F^T P A_F + \Lambda$. Then, the following properties hold:*

1) $\mathcal{T}$ is $\mathbb{S}_+^{n+m}$-monotone, i.e., $P \succeq P' \Rightarrow \mathcal{T}(P) \succeq \mathcal{T}(P')$.

2) There exists a matrix norm $\|\cdot\|$ such that $\mathcal{T}$ is a contraction mapping.

3) $\mathcal{T}$ has a unique fixed point $\bar{P} \in \mathbb{S}^{n+m}$ such that $\mathcal{T}(\bar{P}) = \bar{P}$.

Algorithm 1 Primal-Dual Algorithm

```
Initialize $F_0 \in \mathcal{F}$, $\varepsilon > 0$, and set $t = 0$.
repeat
    Dual Update: Solve for $P_t$ from the equation $A_{F_t}^T P_t A_{F_t} + \Lambda = P_t$.
    Primal Update: $F_{t+1} = -(P_{22})_t^{-1}(P_{12})_t^T$.
    $t \leftarrow t + 1$
until $||F_t - F_{t+1}|| \le \varepsilon$.
```

Proof. The proof of 1) is straightforward. For 2), consider any matrix norm $\|\cdot\|$ and any two matrices $P, P' \in \mathbb{S}_+^{n+m}$. We have $\|\mathcal{T}(P - P')\| \le \|P - P'\|\|A_F\|^2$. For any $\varepsilon > 0$, there exists a matrix norm $||\cdot||$ such that $\|A_F\| \le \rho(A_F) + \varepsilon$ by [25, Theorem 4.2.1], and since $\rho(A_F) < 1$, we can find a matrix norm $\|\cdot\|$ such that $\|\mathcal{T}(P - P')\| \le \|P - P'\|(\rho(A_F) + \varepsilon)^2$, where $(\rho(A_F) + \varepsilon)^2 < 1$. Therefore, $\mathcal{T}$ is a contraction mapping with respect to the norm. Using this norm, define the metric $d(P, P') = \|P - P'\|$ and consider the metric space $(\mathbb{S}_+^{n+m}, d)$. This metric space is complete [26, Definition 3.12]. This is because by [26, pp.54], $(\mathbb{R}^{n+m}, d)$ being a complete metric space and $\mathbb{S}_+^{n+m}$ being a closed subset of $\mathbb{R}^{n+m}$ imply that $(\mathbb{S}_+^{n+m}, d)$ is complete. Then, by the Banach's fixed point theorem [26, Theorem 9.23], $\mathcal{T}$ has a unique fixed point. □

Proposition 8. *In Algorithm 1 without the stopping criterion, $\lim_{t\to\infty} P_t = P^*$ and $\lim_{t\to\infty} F_t = F^*$, where P^* and F^* are defined in (6) and (4), respectively.*

Proof. See Appendix F. □

Remark 5. *Algorithm 1 can be interpreted as a policy iteration of the standard dynamic programming [1]. In particular, pre- and post-multiplying $A_{F_t}^T P_{t+1} A_{F_t} + \Lambda = P_{t+1}$ by $\begin{bmatrix} x^T & u^T\end{bmatrix}$ and its transpose, we have*

$$\begin{aligned}&Q_{t+1}((A + BF_t)x, F_t(A + BF_t)x) + x^T Q x + u^T R u\\ &= Q_{t+1}(x, u), \quad \forall x \in \mathbb{R}^n, u \in \mathbb{R}^m,\end{aligned} \tag{22}$$

where $Q_{t+1}(x, u) := \begin{bmatrix} x\\ u\end{bmatrix}^T P_{t+1} \begin{bmatrix} x\\ u\end{bmatrix}$, which corresponds to the Bellman equation for the Q-function. The update $F_{t+1} = -(P_{22})_t^{-1}(P_{12})_t^T$ in Algorithm 1 can be expressed as $F_{t+1}x = \arg\min_{u\in\mathbb{R}^m} Q_{t+1}(x, u), \forall x \in \mathbb{R}^n$. In this respect, Algorithm 1 is equivalent to the policy iteration for Q-functions in [2, Section 2.3]. It is known that Q_t iteration converges to the Q-function in (5). Therefore, the policy iteration can be interpreted as an algorithm solving the saddle point problem.

Remark 6. *If the linear equation $P_t = A_{F_t}^T P_t A_{F_t} + \Lambda$ in Algorithm 1 is replaced by $P_{t+1} = A_{F_t}^T P_t A_{F_T} + \Lambda$ and the primal update is modified to $F_{t+1} = -(P_{22}^{t+1})^{-1}(P_{22}^{t+1})^T$, its convergence can be proved using a similar argument of the proof of Proposition 8. In addition, it can be proved that the iteration is equivalent to the value iteration or the Riccati recursion $X_{k+1} = A^T X_k A - A^T X_k B(R + B^T X_k B)^{-1} B^T X_k A + Q$ with any $X_0 \in \mathbb{S}_+^n$.*

Note that Algorithm 1 iteratively solves the KKT condition in Proposition 7, and the other primal variable S is not used. However, one can develop an algorithm that does not require the knowledge of the system by using variables corresponding to the primal variable S. Firstly, for any $F_t \in \mathcal{F}$, define $S(F_t) \in \mathbb{S}_+^{n+m}$ as a solution of (8), i.e., $A_{F_t} S(F_t) A_{F_t}^T + \begin{bmatrix} I_n \\ F_t \end{bmatrix} Z \begin{bmatrix} I_n \\ F_t \end{bmatrix}^T = S(F_t)$. $S(F_t)$ can be viewed as the primal variable corresponding to S in Problem 5 for fixed $F = F_t$. Such a $S(F_t)$ is unique by 1) of Lemma 2. As shown in the proof of Proposition 1, $S(F_t)$ can be described by

$$\begin{aligned} S(F_t) &= \sum_{k=0}^{\infty} (A_{F_t}^k) \begin{bmatrix} I_n \\ F_t \end{bmatrix} Z \begin{bmatrix} I_n \\ F_t \end{bmatrix}^T (A_{F_t}^k)^T \\ &= \sum_{i=1}^{r} \sum_{k=0}^{\infty} \begin{bmatrix} x(k; F_t, z_i) \\ F x(k; F_t, z_i) \end{bmatrix} \begin{bmatrix} x(k; F_t, z_i) \\ F x(k; F_t, z_i) \end{bmatrix}^T, \end{aligned}$$

which can be approximated from the observation of state trajectories.

1) **Primal Update**: Let $F_t \in \mathcal{F}$. A primal feasible solution of (18) and (19) is approximated by

$$\tilde{S}(F_t) := \sum_{i=1}^{r} \sum_{k=0}^{M} v(k; F_t, v_i) v(k; F_t, v_i)^T. \quad (23)$$

2) **Dual Update:** Since $\tilde{S}(F_t) \succ 0$, the dual feasibility condition (20) holds if and only if $\tilde{S}(F_t)(A_{F_t}^T P A_{F_t} - P + \Lambda)\tilde{S}(F_t) = 0$, which can be rewritten by the linear matrix equation

$$W(F_t)^T P W(F_t) + \tilde{S}(F_t)(\Lambda - P)\tilde{S}(F_t) = 0, \quad (24)$$

where $W(F_t) := \sum_{i=1}^{r} \sum_{k=0}^{M} v(k+1; F_t, v_i) v(k; F_t, v_i)^T$. Note that both $\tilde{S}(F_t)$ and $W(F_t)$ can be recursively computed from the input and state at every time step. Therefore, for a fixed stabilizing $F_t \in \mathcal{F}$, the corresponding dual variable P can be computed by solving the linear matrix equation (24). Note that since $\Gamma \succ 0$ and $\rho(A_{F_t}) < 1$, P is positive definite.

3) **Policy Update:** Since $P \succ 0$, the primal update step $F = -P_{22}^{-1} P_{12}^T$ can be performed directly.

Remark 7. *Note that the number $M \geq 1$ can be taken arbitrarily. This is because $\tilde{S}(F_0) \succ 0$ for all $M \geq 1$, and once $\tilde{S}(F_0)$ is nonsingular, the selection of M does not affect the solution of (24). However, for small M, $\tilde{S}(F_0) \succ 0$ may be close to a singular matrix in some cases, and this results in ill-conditioning problems when solving (24) using linear equation solvers. Therefore, an appropriately selected M is helpful to avoid the numerical problems.*

The overall algorithm is given in Algorithm 2.

Algorithm 2 Model-Free Primal-Dual Algorithm

```
1: Initialize $F_0 \in \mathcal{F}$ and set $t = 0$.
2: repeat
3:     $S_t = \tilde{S}(F_t)$ in (23)
4:     Solve for $P$ from the equation (24), and set $P_t = P$.
5:     $F_{t+1} = -(P_{22})_t^{-1} (P_{12})_t^T$.
6:     $t \leftarrow t + 1$
7: until a certain stopping criterion is satisfied.
```

Remark 8. *Algorithm 2 can be interpreted as a version of the Q-learning in [13]. The algorithm in [13] uses a recursive least square algorithm to solve the Bellman equation (22). In [13], artificial disturbances are injected into the control input, and it is assumed that the collected state-input data guarantees the uniqueness of the least-square solution, which is called the persistent excitation assumption. On the other hand, Algorithm 2 uses a matrix equation which exactly characterizes the Bellman equation under the assumption that we can obtain the state-input trajectories of a certain set of initial vectors which are linearly independent in $\mathbb{R}^n$. Although our assumption is stronger than the persistent excitation assumption, Algorithm 2 solves the Bellman equation exactly for a given state-feedback gain F_t. If our assumption is met, then Algorithm 2 quickly converges to the optimal solution as will be illustrated in the subsequent example.*

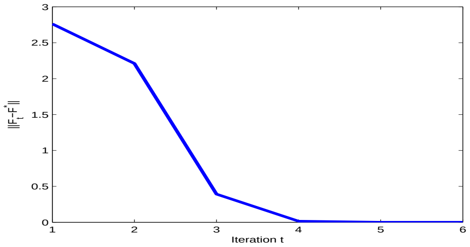

Fig. 2. Evolution of $||F_t - F^*||$.

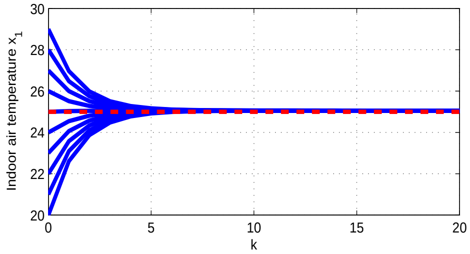

Fig. 3. Trajectories of $x_1(k)$ (indoor air temperature, blue lines) under the designed LQR control policy.

Example 1. *Consider a room's thermal dynamic model expressed as (1) with*

$$A = \begin{bmatrix} 0.9500 & 0.0250 & 0.0250 & 0 \\ 0.0250 & 0.9750 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0.0250 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where $x_1(k)$ is the indoor air temperature ($^\circ C$), $x_2(k)$ is the wall temperature ($^\circ C$), $x_3(k)$ is the outdoor air temperature ($^\circ C$), $x_4(k)$ is the reference temperature ($^\circ C$). The outdoor air temperature and reference temperature are kept constants ($30^\circ C$ and $25^\circ C$, respectively) over time. We want to design an LQR control policy with the discount factor $\alpha = 0.9$, $Q = \begin{bmatrix} 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -1 \end{bmatrix}^T$, and $R = 0.1$. The cost function enforces the indoor temperature to track the desired reference temperature. Suppose that the system matrices are unknown, but we guess that a stabilizing state-feedback gain (for $(\alpha^{1/2} A, \alpha^{1/2} B)$) is $F_{\rm stab} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$. The optimal LQR gain is $F^* = \begin{bmatrix} -1.8631 & -0.0855 & -0.0873 & 2.0359 \end{bmatrix}$. We select the set of initial state vectors

$$V := \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \\ v_5^T \end{bmatrix} = \begin{bmatrix} 25 & 27 & 30 & 25 & 5 \\ 25 & 27 & 30 & 27 & 0 \\ 25 & 27 & 30 & 25 & 0 \\ 27 & 27 & 30 & 25 & 0 \\ 25 & 25 & 30 & 25 & 5 \end{bmatrix},$$

where $v_i \in \mathbb{R}^{n+m}, i \in \{1, 2, \ldots, 5\}$, are initial states of the augmented state vector $v(k) := \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}$. Since V has full rank, $V^T V = \Gamma \succ 0$. Algorithm 2 is applied with $M = 10$, where M is the time horizon used to approximate the primal variable in (23). Figure 2 depicts the evolution of $\|F_t - F^*\|$, which becomes close to zero within 5 iterations. Figure 3 illustrates the trajectories of $x_1(k)$ (indoor air temperature) under the LQR control policy obtained using Algorithm 2 and the initial conditions $x(0) = \begin{bmatrix} x_1(0) & 29 & 30 & 25 \end{bmatrix}^T, x_1(0) \in \{20, 21, \ldots, 29\}$.

Remark 9. *Additional potential applications of the proposed analysis can be summarized as follows. Various SDP formulations of Problem 5 or Problem 2 can be derived, and they can be used to develop new analysis and control design approaches. For example, an SDP-based optimal control design with energy and input constraints can be derived. Another direction is algorithms for structured controller designs [19], [27]. These approaches are included in the supplemental material [18].*

Conclusion

In this paper, we have studied connections among the Lagrangian duality of the LQR problem, the corresponding KKT condition, ARE, and value/policy iterations for the Q-function. We have proved that the LQR problem can be converted to a nonconvex optimization problem which has the zero duality gap and derived its exact dual problem. We also prove that the Q-function is constructed from the dual variables and prove that the dynamic programming and Q-learning are primal-dual update procedures. As an application of our analysis, a model-free LQR design algorithm is also developed. The algorithm can be improved in many directions, for instance, finding an initial stabilizing gain without knowledge of the system and generalizing to the linear quadratic Gaussian (LQG) design problems. A possible extension is to release the requirements of R being positive definite and Λ being a block-diagonal matrix as in [28]. Another direction is to study the Kalman filtering problem from the duality perspective as discussed in [29].

Appendix A
Proof of Proposition 1

For any $F \in \mathcal{F}$, the objective function of Problem 1 can be written as

$$\sum_{i=1}^{r} J(F, z_i) = \sum_{i=1}^{r}\sum_{k=0}^{\infty} \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix}^T \Lambda \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix}$$
$$= \mathbf{Tr}(\Lambda S),$$

where $S := \sum_{i=1}^{r}\sum_{k=0}^{\infty} \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix} \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix}^T$. Observe that S can be represented by

$$S = \sum_{k=0}^{\infty} (A_F^k) \begin{bmatrix} I_n \\ F \end{bmatrix} Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T (A_F^k)^T, \tag{25}$$

which satisfies the Lyapunov equation (8). Moreover, since $F \in \mathcal{F}$, $\rho(A_F) < 1$ by Lemma 1. By 1) of Lemma 2, S in (25) is the unique solution. Since (S, F) above can be an arbitrary feasible point of Problem 2, this shows that the optimal value of Problem 2, J_p, is lower bounded by the optimal value of Problem 1, $J(F^*)$. Noting that (S^*, F^*) where S^* is the unique solution of (8) with $F = F^*$ is a feasible point of Problem 2 whose objective function is exactly $J(F^*)$, we conclude that (S^*, F^*) is the unique solution of Problem 2 and $J_p = J(F^*)$.

Appendix B
Proof of Proposition 2

By expanding (8) in Problem 2 and comparing the first $n \times n$ block diagonal matrix, we obtain

$$\begin{bmatrix} A & B \end{bmatrix} S \begin{bmatrix} A^T \\ B^T \end{bmatrix} + Z = S_{11}. \tag{26}$$

Plugging the left-hand side of (8) into S in (26) and using (26) again, we have

$$(A+BF)S_{11}(A+BF)^T + Z = S_{11}. \tag{27}$$

In addition, noticing that (8) can be written as $\begin{bmatrix} I_n \\ F \end{bmatrix} \left(\begin{bmatrix} A & B \end{bmatrix} S \begin{bmatrix} A^T \\ B^T \end{bmatrix} + Z \right) \begin{bmatrix} I_n \\ F \end{bmatrix}^T = S$, and combining (26) with the above equation yield the second statement. Comparing both sides of the equation in 2) results in $S_{11}F^T = S_{12}, FS_{11}F^T = S_{22}$. Since $S_{11} \succeq Z \succ 0$, solving $S_{11}F^T = S_{12}$ leads to the third statement.

Appendix C
Proof of Proposition 3

Note that since $Z \succ 0$, there exists nonsingular $M \in \mathbb{R}^{n\times n}$ such that $Z = M^T M$. Then, the pair $\left(A_F, \begin{bmatrix} I_n \\ F \end{bmatrix} M^T\right)$ is stabilizable for any $F \in \mathbb{R}^{m\times n}$ because the state-feedback gain $K = -M^{-T}\begin{bmatrix} A & B \end{bmatrix}$ satisfies $A_F + \begin{bmatrix} I_n \\ F \end{bmatrix} M^T K = 0$. By the forth statement of Lemma 2, $\rho(A_F) < 1$, i.e., $F \in \mathcal{F}$, if and only if (8) in Problem 2 has a solution $S \succeq 0$. This shows that we can equivalently replace the constraint (7) with $S \succeq 0$.

Appendix D
Proof of Proposition 4

For any $F \in \mathcal{F}$, the objective function of Problem 1 is

$$J(F) = \sum_{i=1}^{r}\sum_{k=0}^{\infty} \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix}^T \Lambda \begin{bmatrix} x(k;F,z_i) \\ Fx(k;F,z_i) \end{bmatrix}$$
$$= \sum_{i=1}^{r}\sum_{k=0}^{\infty} z_i^T \begin{bmatrix} I_n \\ F \end{bmatrix}^T (A_F^T)^k \Lambda A_F^k \begin{bmatrix} I_n \\ F \end{bmatrix} z_i$$
$$= \mathbf{Tr}\left(\begin{bmatrix} I_n \\ F \end{bmatrix} Z \begin{bmatrix} I_n \\ F \end{bmatrix}^T P \right), \tag{28}$$

where P is a solution of the Lyapunov equation (9) corresponding to the given $F \in \mathcal{F}$. Moreover, since $F \in \mathcal{F}$, $\rho(A_F) < 1$ by Lemma 1. By 1) of Lemma 2, P is the unique solution. Since (P, F) above is an arbitrary feasible point of Problem 3, this shows that the optimal value of Problem 3, J_p', is lower bounded by the optimal value of Problem 1, $J(F^*)$. Noting that (P^*, F^*) where P^* is the unique solution of (9) with $F = F^*$ is a feasible point of Problem 3 whose objective function is exactly $J(F^*)$, we conclude that (P^*, F^*) is the unique solution of Problem 3 and $J_p' = J(F^*)$.

Appendix E
Proof of Proposition 5

By direct calculations with definitions in (4) and (6), we have $\begin{bmatrix} I_n \\ F^* \end{bmatrix}^T P^* \begin{bmatrix} I_n \\ F^* \end{bmatrix} = X^*$. Then, from the definition (6), it follows that $P^* = \begin{bmatrix} A & B \end{bmatrix}^T X^* \begin{bmatrix} A & B \end{bmatrix} + \Lambda = \begin{bmatrix} A & B \end{bmatrix}^T \begin{bmatrix} I_n \\ F^* \end{bmatrix}^T P^* \begin{bmatrix} I_n \\ F^* \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix} + \Lambda = A_{F^*}^T P^* A_{F^*} + \Lambda$. By Proposition 4, $F_p' = F^*$, and P^* uniquely solves the Lyapunov equation (9). This implies $P^* = P_p'$. The second statement is directly proved by using the definitions of P^* and F^* in (6) and (4), respectively.

Appendix F
Proof of Proposition 8

We first prove $F_t \in \mathcal{F} \Rightarrow F_{t+1} \in \mathcal{F}$. Since $F_t \in \mathcal{F}$ and $\Lambda \succeq 0$, $A_{F_t}^T P_t A_{F_t} + \Lambda = P_t$ admits a unique solution $P_t \succeq 0$ by 1) of Lemma 2. Moreover, $\Lambda_{22} \succ 0$ implies that $(P_{22})_t \succ 0$. If $F_{t+1} = -(P_{22})_t^{-1}(P_{12})_t^T$, then Lemma 4 leads to

$$P_t = A_{F_t}^T P_t A_{F_t} + \Lambda$$

$$
\begin{aligned}
&= \begin{bmatrix} A & B \end{bmatrix}^T \begin{bmatrix} I_n \\ F_t \end{bmatrix}^T P_t \begin{bmatrix} I_n \\ F_t \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix} + \Lambda \\
&\succeq \begin{bmatrix} A & B \end{bmatrix}^T (P_{11} - P_{12}P_{22}^{-1}P_{12}^T) \begin{bmatrix} A & B \end{bmatrix} + \Lambda \\
&= \begin{bmatrix} A & B \end{bmatrix}^T \begin{bmatrix} I_n \\ F_{t+1} \end{bmatrix}^T P_t \begin{bmatrix} I_n \\ F_{t+1} \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix} + \Lambda \\
&= A_{F_{t+1}}^T P_t A_{F_{t+1}} + \Lambda.
\end{aligned}
$$

Consider the mapping $\mathcal{T}(P_t) := A_{F_{t+1}}^T P_t A_{F_{t+1}} + \Lambda$ in Lemma 5 with $F = F_{t+1}$. Then, the last inequality can be compactly written as $P_t \succeq \mathcal{T}(P_t)$. Since $\mathcal{T}$ is $\mathbb{S}_+^{n+m}$-monotone, applying repeatedly $\mathcal{T}$ on both sides of $P_t \succeq \mathcal{T}(P_t)$ leads to an $\mathbb{S}_+^{m+n}$-monotonically nonincreasing sequence $\mathcal{T}^i(P_t), i = 1, 2, \ldots$ in the positive semidefinite cone $\mathbb{S}_+^{n+m}$ that is bounded from below. Thus, the limit $\lim_{i\to\infty} \mathcal{T}^i(P_t) =: P_{t+1}$ exists and solves the Lyapunov equation $\mathcal{T}(P_{t+1}) = P_{t+1}$. We will now prove that this implies $\rho(A_{F_{t+1}}) < 1$, and equivalently, $F_{t+1} \in \mathcal{F}$ by Lemma 1. since $R \succ 0$ by Assumption 1, there exists a nonsingular M such that $R = M^T M$. Then, Λ can be expressed as $\Lambda = U^T U$, where $U := \begin{bmatrix} C & 0 \\ 0 & M \end{bmatrix}$ and C is defined in Assumption 1. We next show that $(A_{F_{t+1}}, U)$ is detectable for any $F_{t+1} \in \mathbb{R}^{m\times n}$. By Assumption 1, (A, C) is detectable. Thus, there exists an observer gain $L \in \mathbb{R}^{n\times m}$ such that $\rho(A + LC) < 1$. Construct the observer gain $W = \begin{bmatrix} L & -BM^{-1} \\ 0 & -F_{t+1}BM^{-1} \end{bmatrix}$ so that $\rho(A_{F_{t+1}} + WU) = \rho\left(\begin{bmatrix} A + LC & 0 \\ F_{t+1}A & 0 \end{bmatrix}\right) = \rho(A + LC) < 1$. Therefore, $(A_{F_{t+1}}, U)$ is detectable for any $F_{t+1} \in \mathbb{R}^{m\times n}$. By the third statement of Lemma 2, $\rho(A_{F_{t+1}}) < 1$ (i.e., $F_{t+1} \in \mathcal{F}$) if and only if there exists a positive semidefinite solution $P \succeq 0$ of $A_{F_{t+1}}^T P A_{F_{t+1}} - P + \Lambda = 0$. Such $P \succeq 0$ exists since $\mathcal{T}(P_{t+1}) = P_{t+1}$. Therefore, we have $\rho(A_{F_{t+1}}) < 1$.

Moreover, $P_t \succeq P_{t+1}$ monotonically nonincreasing, for all $t \in \mathbb{N}$, we obtain a sequence $(P_t)_{t=0}^\infty$ that is $\mathbb{S}_+^{n+m}$-monotonically nonincreasing, and hence, converges, i.e., $\lim_{t\to\infty} P_t =: \bar{P}$. By examining the second $m \times m$ block-diagonal matrices, $\bar{P}_{22} \succ 0$. Accordingly, $\lim_{t\to\infty} F_t = -\bar{P}_{22}^{-1}\bar{P}_{12}^T =: \bar{F}$ holds. Since $A_{\bar{F}}^T \bar{P} A_{\bar{F}} - \bar{P} + \Lambda = 0$, one has $\bar{F} \in \mathcal{F}$ by 3) of Lemma 2. The primal matrix variable $\bar{S}$ that solves $A_{\bar{F}} S A_{\bar{F}}^T + \Gamma - S = 0$ for S in (18) exists, is unique, and positive definite by 2) of Lemma 2. The triplet $(\bar{P}, \bar{F}, \bar{S})$ is a solution of the KKT condition in Proposition 7. According to [3, Section 5.5.3, pp. 243], the strong duality ensures that any pair of primal and dual optimal points must satisfy the KKT conditions, while a solution of the KKT condition may not correspond to the primal and dual optimal points.

Next, we prove that $(\bar{P}, \bar{F})$ satisfying $A_{\bar{F}}^T \bar{P} A_{\bar{F}} - \bar{P} + \Lambda = 0$, $\bar{F} = -\bar{P}_{22}^{-1}\bar{P}_{12}^T$, and $\bar{P} \succeq 0$ is identical to (P^*, F^*). Plugging $-\bar{P}_{22}^{-1}\bar{P}_{12}^T$ into $\bar{F}$ of $A_{\bar{F}}^T \bar{P} A_{\bar{F}} - \bar{P} + \Lambda = 0$ results in $\begin{bmatrix} A & B \end{bmatrix}^T (\bar{P}_p - \bar{P}_{12}(\bar{P}_{22})^{-1}\bar{P}_{12}^T) \begin{bmatrix} A & B \end{bmatrix} + \Lambda = \bar{P}$. Introducing the auxiliary matrix variable $\bar{X} = \bar{P}_{11} - \bar{P}_{12}\bar{P}_{22}^{-1}\bar{P}_{12}^T$, which satisfies $\bar{X} \succeq 0$ by Lemma 3, one has $\begin{bmatrix} A & B \end{bmatrix}^T \bar{X} \begin{bmatrix} A & B \end{bmatrix} + \Lambda = \bar{P}$, which is written as $\begin{bmatrix} A^T\bar{X}A + Q & A^T\bar{X}B \\ B^T\bar{X}A & B^T\bar{X}B + R \end{bmatrix} = \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{12}^T & \bar{P}_{22} \end{bmatrix}$. Plugging the expressions $\bar{P}_{11} = A^T\bar{X}A + Q$, $\bar{P}_{12} = A^T\bar{X}B$, and $\bar{P}_{22} = B^T\bar{X}B + R$ into $\bar{X} = \bar{P}_{11} - \bar{P}_{12}\bar{P}_{22}^{-1}\bar{P}_{12}^T$ yields

$$Q + A^T\bar{X}A - A^T\bar{X}B(R + B^T\bar{X}B)^{-1}B^T\bar{X}A = \bar{X}, \tag{29}$$

which is exactly the ARE. Thus, under Assumption 1, we must have $\bar{X} = X^*$. The desired result follows from the definition of (P^*, F^*) in (6) and (4).

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. Nashua, MA: Athena Scientific, 2005, vol. 1.

[2] ——, *Dynamic Programming and Optimal Control*, 4th ed. Nashua, MA: Athena Scientific, 2005, vol. 2.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[4] L. Vandenberghe and S. Boyd, “Semidefinite programming,” *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.

[5] J. Willems, “Least squares stationary optimal control and the algebraic riccati equation,” *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 621–634, 1971.

[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*. Philadelphia, PA: SIAM, 1994.

[7] D. D. Yao, S. Zhang, and X. Y. Zhou, “Stochastic linear-quadratic control via semidefinite programming,” *SIAM Journal on Control and Optimization*, vol. 40, no. 3, pp. 801–823, 2001.

[8] M. A. Rami and X. Y. Zhou, “Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls,” *Automatic Control, IEEE Transactions on*, vol. 45, no. 6, pp. 1131–1143, 2000.

[9] V. Balakrishnan and L. Vandenberghe, “Semidefinite programming duality and linear time-invariant systems,” *Automatic Control, IEEE Transactions on*, vol. 48, no. 1, pp. 30–41, 2003.

[10] A. Gattami, “Generalized linear quadratic control,” *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 131–136, 2010.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.

[12] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.

[13] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, “Adaptive linear quadratic control using policy iteration,” in *American Control Conference, 1994*, vol. 3, 1994, pp. 3475–3479.

[14] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *Circuits and Systems Magazine, IEEE*, vol. 9, no. 3, pp. 32–50, 2009.

[15] R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[16] C. J. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[17] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, England, 1994, vol. 37.

[18] D. Lee and J. Hu, “Supplementlal material for “primal-dual q-learning framework for LQR design”,” *arXiv:1811.08475, https://arxiv.org/abs/1811.08475*, 2018.

[19] J. C. Geromel, C. De Souza, and R. Skelton, “Static output feedback controllers: Stability and convexity,” *IEEE Transactions on Automatic Control*, vol. 43, no. 1, pp. 120–125, 1998.

[20] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. Wiley-Interscience New York, 1972.

[21] G. Gu, *Discrete-Time Linear Systems: Theory and Design with Applications*. Springer Science & Business Media, 2012.

[22] C.-T. Chen, *Linear System Theory and Design*. Oxford University Press, Inc., 1995.

[23] F. Zhang, *The Schur complement and its applications*. Springer Science & Business Media, 2006, vol. 4.

[24] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[25] D. Serre, *Matrices: Theory and Applications*. New York: SpringerVerlag, 2010.

[26] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[27] F. Lin, M. Fardad, and M. R. Jovanovic, “Augmented lagrangian approach to design of structured optimal state feedback gains,” *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2923–2929, 2011.

[28] A. Ferrante and L. Ntogramatzidis, “On the generalized algebraic riccati equations,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 9555–9560, 2017.

[29] M. Zorzi, “Robust kalman filtering under model perturbations,” *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2902–2907, 2017.