Learning Robust LQ-Controllers Using Application Oriented Exploration

Mina Ferizbegovic, Jack Umenberger, Håkan Hjalmarsson, and Thomas B. Schön

***Abstract*—This letter concerns the problem of learning robust LQ-controllers, when the dynamics of the linear system are unknown. First, we propose a robust control synthesis method to minimize the worst-case LQ cost, with probability $1 - \delta$, given empirical observations of the system. Next, we propose an approximate dual controller that simultaneously regulates the system and reduces model uncertainty. The objective of the dual controller is to minimize the worst-case cost attained by a new robust controller, synthesized with the reduced model uncertainty. The dual controller is subject to an exploration budget in the sense that it has constraints on its worst-case cost with respect to the current model uncertainty. In our numerical experiments, we observe better performance of the proposed robust LQ regulator over the existing methods. Moreover, the dual control strategy gives promising results in comparison with the common greedy random exploration strategies.**

***Index Terms*—Identification, robust control, machine learning.**

I. Introduction

DECISION making in uncertain dynamic environments is a task of fundamental importance in a number of fields. Though the subject has received steady attention in control since the advent of 'dual control' in the 1960s [1], it has witnessed a resurgence in interest due to the recent success of reinforcement learning (RL), particularly in games [2], [3]. In such a setting, decisions are made with two objectives in mind. First, there is a goal to be achieved, typically a cost to be minimized. Second, due to the uncertainty there is a need to gather information about the environment, often referred to as 'learning' via 'exploration'. These two objectives are often competing, a fact known as the exploration/exploitation trade-off in RL.

In real-world applications, effective exploration should not compromise the safe and reliable operation of the system. Furthermore, exploration should be application specific; ideally, it should not excite the system arbitrarily, but rather in such a way that the information gathered subsequently is useful for achieving the goal (reducing the cost).

This letter is concerned with control of uncertain linear dynamical systems, with the goal of minimizing infinite horizon quadratic cost (on states and inputs). We present control policies that achieve robust, application-oriented exploration. For robustness, we bound (with high probability) the cost of the policy on the true, but unknown, system during exploration. By application-oriented, we mean that the policy is designed to excite the system so as to reduce uncertainty in such a way that, given this new information, the cost achieved by redesigning a robust controller is minimized.

A. Related Work

Study of simultaneous learning and control of dynamical systems began with the introduction of 'dual control' [1], [4] in the 1960s. Though the formulation was clear, synthesis via dynamic programming (DP) was intractable. Design of such controllers was restricted to special cases, e.g., linear systems with finite state/decision spaces [5], or relied on simplifying approximations to the problem [6]. Nevertheless, these early efforts [7] established the importance of balancing 'probing' (exploration) with 'caution' (robustness).

Effective exploration has strong connections to the topic of experiment design; in particular, the value of choosing inputs with consideration of the purpose of the model was recognized in [8]. Convex problem formulations [9], [10] ultimately led to the application-oriented and least-costly experiment design paradigms [11], [12], in which the objective is to reduce model uncertainty such that certain performance criteria can be achieved, while minimizing disruption to the system or experiment time. This paved the way for application-oriented experiment design approaches to dual control [13], see also [14], [15], [16], [17], [18] for adaptive, dual and data-driven model predictive control applications. However, with some exceptions, e.g., [19] (on which this letter builds), these methods do not consider robustness of the control strategies

Manuscript received March 1, 2019; revised May 3, 2019; accepted May 21, 2019. Date of publication June 6, 2019; date of current version June 21, 2019. This work was supported in part by NewLEADS—New Directions in Learning Dynamical Systems under Contract 2016-06079, and in part by the Wallenberg AI, Autonomous Systems and Software Program funded by Knut and Alice Wallenberg Foundation. Recommended by Senior Editor F. Dabbene. *(Corresponding author: Mina Ferizbegovic.)*

M. Ferizbegovic and H. Hjalmarsson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: minafe@kth.se; hjalmars@kth.se).

J. Umenberger and T. B. Schön are with the Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden (e-mail: jack.umenberger@it.uu.se; thomas.schon@it.uu.se).

Digital Object Identifier 10.1109/LCSYS.2019.2921512

to model uncertainty, and indeed assume that the true model parameters are known.

In contrast, we consider worst-case design with respect to a set of models for robustness. Another aspect of identification for control [20], [21] is concerned with exploration for reduced complexity modeling.

Spurred by the success of RL in games, learning for control has witnessed a resurgence of interest, with particular attention on linear quadratic control. In RL the goal is typically to minimize *regret*. Works such as [22], [23], [24] employ the so-called ‘optimism in the face of uncertainty’ (OFU) principle, which optimizes control actions for the ‘best-case’ model in the uncertain set. This leads to optimal regret but requires the solution of intractable non-convex optimization problems. Alternatively, the works of [25], [26] employ Thompson sampling, which optimizes the control action for a system drawn randomly from the posterior distribution over the set of uncertain models, given data. None of the works above consider robustness, which is essential for implementation on physical systems.

Robustness is studied in the so-called ‘coarse-ID’ family of methods, see [27], [28], [29]. In [29], sample convexity bounds are derived for LQR with unknown linear dynamics. This approach is extended to adaptive LQR in [27], however, the policies are not optimized for exploration and exploitation jointly; exploration is effectively random.

B. Contribution and Paper Structure

This letter presents robust dual control for linear systems, see Section II-A for a detailed problem formulation. By ‘robust’ we mean optimizing for the worst-case performance, which leads to bounds on performance for the true system (with high probability). Our specific contributions are as follows: i) we present a convex procedure for design of a robust controller that minimizes the worst-case cost on the true unknown system, with high probability, see Section III; ii) we present a convex procedure for design of a dual controller that minimizes the worst-case cost achieved by a robust controller redesigned with information gathered during exploration, subject to worst-case constraints on the cost of exploration, see Section IV. Performance of the proposed methods are investigated numerically in Section V.

II. Preliminaries

A. Problem Statement

We consider discrete linear time-invariant dynamical systems given by

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}\Big(0, \sigma_w^2 I\Big), \tag{1}$$

where $x_t \in^n$, $u_t \in^p$ and $\omega_t \in^n$ denote the state, input and process noise, respectively, at time t. We assume that x_t is directly measurable. Our objective is to minimize the expected infinite horizon linear-quadratic cost $\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=0}^{t=T} x_t^\top Q x_t + u_t^\top R u_t]$, where Q and R are user defined positive semidefinite matrices. When A and B are known, the optimal policy is a static state-feedback law $u_t = Kx_t$, where $K \in^{p\times n}$ is the solution of a discrete algebraic Riccati equation [30]. In this letter, A and B are unknown, so it shall be necessary to *learn* and *control* the system, simultaneously. Though not necessarily optimal, we will restrict our attention to static-gain policies of the form, $u_t = Kx_t + e_t$, where $e_t \sim \mathcal{N}(0, \Sigma)$ represent random excitations for the purpose of learning. A policy comprises $K \in^{p\times n}$ and $\Sigma \in \mathbb{S}_{++}^p$ (i.e., symmetric positive definite of dimension p), and is denoted $\mathcal{K} = \{K, \Sigma\}$. Despite the assumption of linear dynamics and observable states, this problem has attracted considerable recent attention, see, e.g., [27], [28], [29].

As the dynamics are unknown, all knowledge of the system comes from measured trajectories of (1). Given observations $\mathcal{D} = \{x_t, u_t\}_{t=1}^N$ of (1), we assume there exists a mapping to a model $\mathcal{M}(\mathcal{D}) = \{\hat{A}, \hat{B}, D\}$. Here $\hat{A}$ and $\hat{B}$ denote point estimates of A and B, respectively, and D quantifies the uncertainty associated with these estimates. Specifically, with the estimation error defined as

$$X(\mathcal{D}) := \begin{bmatrix} (\hat{A} - A)^\top \\ (\hat{B} - B)^\top \end{bmatrix} \tag{2}$$

$D(\mathcal{D})$ is such that with probability $1 - \delta$ we have

$$X(\mathcal{D})^\top D(\mathcal{D}) X(\mathcal{D}) \preceq I. \tag{3}$$

Explicit dependence of X and D on $\mathcal{D}$ is dropped from the notation when there is no ambiguity. Details on the mapping from data $\mathcal{D}$ to models $\mathcal{M}$ are provided in Section II-B.

Given this set-up, this letter seeks to solve two control problems. The first is a robust control problem: given data $\mathcal{D}_0$ from (1), find a policy $\mathcal{K}$ that minimizes the worst-case cost with probability $1 - \delta$. Denoting this worst-case cost by

$$\begin{aligned} J(\mathcal{K}, \mathcal{M}) &:= \max_{A,B} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\textstyle\sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t] \\ \text{s.t. } x_{t+1} &= Ax_t + Bu_t + w_t, \;\; u_t = Kx_t + e_t, \;\; X^\top D X \preceq I, \end{aligned} \tag{4}$$

the robust control problem can be expressed as

$$\min_{\mathcal{K}} \; J(\mathcal{K}, \mathcal{M}(\mathcal{D}_0)). \tag{5}$$

The second is a robust *dual control* problem: given data $\mathcal{D}_0$ from (1), collected during some arbitrary preliminary experiment, design a policy $\mathcal{K}_{\text{dc}}$ such that, i) after letting (1) evolve under $\mathcal{K}_{\text{dc}}$ for T_{dc} time steps and collecting observations $\mathcal{D}_{\text{dc}}$, the worst-case cost given by redesigning a robust controller with the new model $\mathcal{M}(\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}})$ is minimized, ii) subject to the constraint that the worst-case cost of $\mathcal{K}_{\text{dc}}$, given the model $\mathcal{M}(\mathcal{D}_0)$, must not exceed a user-specified ‘exploration budget’, J_{exp}. This problem can be expressed precisely as:

$$\begin{aligned} \min_{\mathcal{K}_{\text{dc}}, \mathcal{K}} \quad & \mathbb{E}\; J(\mathcal{K}, \mathcal{M}(\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}})) \\ \text{s.t.} \quad & J(\mathcal{K}_{\text{dc}}, \mathcal{M}(\mathcal{D}_0)) \le J_{\text{exp}}, \;\; \bar{x}_{t+1} = A\bar{x}_t + B\bar{u}_t + w_t, \\ & \bar{u}_t = K_{\text{dc}}\bar{x}_t + e_t, \;\; e_t \sim \mathcal{N}(0, \Sigma_{\text{dc}}), \mathcal{D}_{\text{dc}} = \{\bar{x}_t, \bar{u}_t\}_{t=1}^{T_{\text{dc}}}. \end{aligned} \tag{6}$$

The expectation in the objective of (6) is w.r.t. the distribution over $\mathcal{D}_{\text{dc}}$ which is a random variable.

B. Uncertainty Quantification

The problem formulated in Section II-A assumes the existence of a mapping from data $\mathcal{D}$ to a model $\mathcal{M}(\mathcal{D})$, complete with bounds on the spectral properties of the estimation error, see (3). In this section we provide one concrete suggestion for such a mapping, which makes use of the following result from [29].

Proposition 1: Assume we have N independent samples $(y^{(l)}, x^{(l)}, u^{(l)})$ such that $y^{(l)} = Ax^{(l)} + Bu^{(l)} + w^{(l)}$, where $w^{(l)}$ are i.i.d. with $\mathcal{N}\big(0, \sigma_w^2 I\big)$ and are independent from $x^{(l)}$ and $u^{(l)}$. Also, let us assume that $N \geq n + p$. Let $\hat{A}$ and $\hat{B}$ denote the least squares estimates given by

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{k=1}^{N} \left\| y^{(l)} - Ax^{(l)} - Bu^{(l)} \right\|_2^2.$$

With estimation error X defined as in (2), with probability $1 - \delta$, we have that

$$XX^\top \preceq \underbrace{\bar{C}(n, p, \delta)\left(\sum_{l=1}^{N}\begin{bmatrix}x^{(l)}\\ u^{(l)}\end{bmatrix}\begin{bmatrix}x^{(l)}\\ u^{(l)}\end{bmatrix}^\top\right)^{-1}}_{D^{-1}}, \tag{7}$$

where $\bar{C}(n, p, \delta) = \sigma_w^2(\sqrt{n+p} + \sqrt{n} + \sqrt{2\log\frac{1}{\delta}}])^2$.

Given Proposition 1, a suitable model $\mathcal{M}(\mathcal{D}) = \{\hat{A}, \hat{B}, D\}$ is given by taking $\{\hat{A}, \hat{B}\}$ as the least squares estimates, and D as defined in (7), as two applications of the Schur complement yields

$$XX^\top \preceq D^{-1} \iff \begin{bmatrix} I & X^\top \\ X & D^{-1}\end{bmatrix} \succeq 0 \iff X^\top DX \preceq I.$$

To apply Proposition 1, the data $\mathcal{D}$ must be comprised of independent samples $\{y^{(l)}, x^{(l)}, u^{(l)}\}_{l=1}^{N}$. One way to satisfy this independence assumption is to collect data in the form of *rollouts*: one can evolve system (1) forward for T_r, initialized at $x_0^{(l)} = 0$ and excited by arbitrary inputs $u_t^{(l)}$, and then take the final state-transition as the observed data, i.e., $y^{(l)} = x_{T_\mathrm{r}}^{(l)}$, $x^{(l)} = x_{T_\mathrm{r}-1}^{(l)}$, $u^{(l)} = u_{T_\mathrm{r}-1}^{(l)}$. If, as is the case in the dual control setting, only a single trajectory can be observed, rather than multiple independent rollouts, then one can *subsample* the data; i.e., take every T_ss-th data point, with T_ss sufficiently large such that the samples $y^{(l)} = x_{T_\mathrm{ss}l}$, $x^{(l)} = x_{T_\mathrm{ss}l-1}$, $u^{(l)} = u_{T_\mathrm{ss}l-1}$, are approximately independent.

The above strategies are conservative, in that they do not make use of all the available data. Alternative strategies for bounding the spectral error that may reduce conservatism (e.g., bootstrap methods [31]) can be found in [29], [32]. The focus of this letter is on control synthesis given a mapping from data to models with bounds on uncertainty; the user is free to employ models with tighter bounds on estimation error if preferred.

III. Designing a Robust Controller

In this section we present a solution to the robust control problem specified in (5). For this problem, exploration is unnecessary, and so we consider policies of the form $u_t = Kx_t$. The closed-loop system can be expressed as

$$x_{t+1} = (A + BK)x_t + w_t, \ \ y_t = \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}}K \end{bmatrix} x_t. \tag{8a}$$

The $\mathcal{H}_2$ norm of the system can be computed as:

$$\min_{W} \ \mathrm{tr} \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}}K \end{bmatrix} W \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}}K \end{bmatrix}^\top \tag{9a}$$

$$\text{s.t.} \quad (A + BK)W(A + BK)^\top - W + \sigma_w^2 I \preceq 0, \tag{9b}$$

where W is the controllability Gramian.

As the system dynamics (i.e., A and B) and the controller K are unknown, we will circumvent the non-convexity in (9) via the usual change of variables. The cost function (9a) is given by tr Y subject to the LMI

$$S_1(W, Y, Z) := \left[\begin{array}{cc|c} \multicolumn{2}{c|}{Y} & \begin{matrix} Q^{\frac{1}{2}}W \\ R^{\frac{1}{2}}Z^\top \end{matrix} \\ \hline WQ^{\frac{1}{2}} & ZR^{\frac{1}{2}} & W \end{array}\right] \succeq 0,$$

with the change of variables $Z = WK^\top$. After applying the Schur complement, the Lyapunov condition (9b) can be written as

$$\begin{bmatrix} W & W(A + BK)^\top \\ (A + BK)W & W - \sigma_w^2 I \end{bmatrix} \succeq 0. \tag{10}$$

The real system dynamics (A, B) are unknown. In their place, we have an approximate model $\mathcal{M}(\mathcal{D}_0) = \{\hat{A}, \hat{B}, D_0\}$ obtained from data $\mathcal{D}_0$, see Section II-B. We can rewrite the elements in the LMI (10) containing A and B as:

$$W(A + BK)^\top = \underbrace{W\hat{A}^\top + Z\hat{B}^\top}_{F} + \underbrace{[-W \ \ -Z]}_{G}X, \tag{11}$$

where X is defined in (2). Now, we can rewrite (10) as

$$\begin{bmatrix} H & F + GX \\ (F + GX)^\top & C \end{bmatrix} \succeq 0, \tag{12}$$

where $H = W$ and $C = W - \sigma_w^2 I$. We require (12) to hold for the 'worst-case' (A,B) (in the confidence region); as a sufficient condition, we enforce that (12) holds for all (A,B) with the $1 - \delta$ confidence region given by (3). To ensure this, we will use the following theorem from [33].

Theorem 1 [33]: The data matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{P}, \mathcal{F}, \mathcal{G}, \mathcal{H})$ satisfy the robust fractional quadratic matrix inequality

$$\begin{bmatrix} \mathcal{H} & \mathcal{F} + \mathcal{G}X \\ (\mathcal{F} + \mathcal{G}X)^\top & \mathcal{C} + X^\top\mathcal{B} + \mathcal{B}^\top X + X^\top\mathcal{A}X \end{bmatrix} \succeq 0,$$

for all X with $I - X^\top\mathcal{P}X \succeq 0$, if and only if there is $t \geq 0$ such that

$$\begin{bmatrix} \mathcal{H} & \mathcal{F} & \mathcal{G} \\ \mathcal{F}^\top & \mathcal{C} & \mathcal{B}^\top \\ \mathcal{G}^\top & \mathcal{B} & \mathcal{A} \end{bmatrix} - t\begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & -\mathcal{P} \end{bmatrix} \succeq 0. \tag{13}$$

By Theorem 1, (12) holds for all $X^\top DX \preceq I$ iff

$$S_2(t, Z, W, D, \hat{A}, \hat{B}) := \begin{bmatrix} H & F & G \\ F^\top & C - tI & 0 \\ G^\top & 0 & tD \end{bmatrix} \succeq 0, \tag{14}$$

which is simply (13) with the substitutions $\mathcal{A}, \mathcal{B} = 0$, $\mathcal{C} = C$ and $\mathcal{H} = H$ as defined in (12), $\mathcal{F} = F$ and $\mathcal{G} = G$ as defined in (11), and $\mathcal{P} = D$. We can then write the robust control problem (5) as the following semidefinite program

$$\begin{aligned} \min_{t,Z,Y,W} \ \operatorname{tr} Y \quad \text{s.t.} \quad & S_1(W, Y, Z) \succeq 0, \ \ t \geq 0, \\ & S_2(t, Z, W, D_0, \hat{A}, \hat{B}) \succeq 0. \end{aligned} \tag{15}$$

IV. Designing a Dual Controller

In this section we present an approximate solution to the robust dual control (see [1], [4]) problem specified in (6).

A. Robust Exploration

In this subsection, we present a convex formulation of the search for policies $\mathcal{K} = \{K, \Sigma\}$ of the form $u_t = Kx_t + e_t$, where K is robustly stabilizing and $e_t \sim \mathcal{N}(0, \Sigma)$. The developments below closely follow those of Section III. For such policies, the Lyapunov condition for the controllability Gramian W becomes

$$(A + BK)W(A + BK)^\top - W + B\Sigma B^\top + \sigma_w^2 I \preceq 0. \tag{16}$$

By two applications of the Schur complement, (16) becomes

$$\begin{bmatrix} W & 0 & W(A+BK)^\top \\ 0 & \Sigma & \Sigma B^\top \\ (A+BK)W & B\Sigma & W - \sigma_w^2 I \end{bmatrix} \succeq 0. \tag{17}$$

Now, similarly to (12) we can write (17) as

$$\begin{bmatrix} H_{\text{dc}} & F_{\text{dc}} + G_{\text{dc}}X \\ (F_{\text{dc}} + G_{\text{dc}}X)^\top & C_{\text{dc}} \end{bmatrix} \succeq 0, \tag{18}$$

with $H_{\text{dc}} = \texttt{blkdiag}(W, \Sigma)$, where $\texttt{blkdiag}$ is the block-diagonal operator, $C_{\text{dc}} = W - \sigma_w^2 I$, $Z = WK^\top$,

$$F_{\text{dc}} = \begin{bmatrix} W\hat{A}^\top + Z\hat{B}^\top \\ \Sigma\hat{B}^\top \end{bmatrix}, \quad G_{\text{dc}} = \begin{bmatrix} -W & -Z \\ 0 & -\Sigma \end{bmatrix}.$$

Again, by Theorem 1, (18) holds for all $X^\top D X \preceq I$ iff

$$S_{\text{dc}}(t, Z, W, \Sigma, D, \hat{A}, \hat{B}) := \begin{bmatrix} H_{\text{dc}} & F_{\text{dc}} & G_{\text{dc}} \\ F_{\text{dc}}^\top & C_{\text{dc}} - tI & 0 \\ G_{\text{dc}}^\top & 0 & tD \end{bmatrix} \succeq 0, \tag{19}$$

which is (13) with the substitutions $\mathcal{A}, \mathcal{B} = 0$, $\mathcal{C} = C_{\text{dc}}$ and $\mathcal{H} = H_{\text{dc}}$, $\mathcal{F} = F_{\text{dc}}$ and $\mathcal{G} = G_{\text{dc}}$, and $\mathcal{P} = D$.

B. Updating the Model After Dual Control

To recapitulate, our objective is to search for a robustly stabilizing policy $\mathcal{K}_{\text{dc}}$ so that the cost of a robust controller redesigned with data $\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}}$ is minimized, where $\mathcal{D}_{\text{dc}}$ is the data observed while $\mathcal{K}_{\text{dc}}$ is applied. To achieve this, we need to approximate the model $\mathcal{M}(\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}})$ as a function of $\mathcal{K}_{\text{dc}} = \{K_{\text{dc}}, \Sigma\}$.

Recall, see (7), that the uncertainty $D(\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}})$ is given by

$$D(\mathcal{D}_0) + \frac{1}{\bar{C}} \sum_{l=1}^{N} \begin{bmatrix} \bar{x}^l \\ \bar{u}^l \end{bmatrix} \begin{bmatrix} \bar{x}^l \\ \bar{u}^l \end{bmatrix}^\top \tag{20}$$

where $N = T_{\text{dc}}/T_{\text{ss}}$ is the number of approximately uncorrelated samples obtained by taking every T_{ss}-th sample from $\{\bar{x}_t \bar{u}_t\}_{t=1}^{T_{\text{dc}}}$, the trajectory of (1) evolving under $\mathcal{K}_{\text{dc}}$. For sufficiently large N we can approximate the empirical covariance by its stationary distribution,

$$\sum_{l=1}^{N} \begin{bmatrix} \bar{x}^l \\ \bar{u}^l \end{bmatrix} \begin{bmatrix} \bar{x}^l \\ \bar{u}^l \end{bmatrix}^\top \approx N \begin{bmatrix} \Sigma_{xx} & \Sigma_{xx} K_{\text{dc}}^\top \\ K_{\text{dc}} \Sigma_{xx} & K_{\text{dc}} \Sigma_{xx} K_{\text{dc}}^\top + \Sigma \end{bmatrix} \tag{21}$$

where $\Sigma_{xx} = \mathbb{E}[\bar{x}\bar{x}^\top]$, i.e., the stationary state-covariance. As the true values of (A, B) are unknown, we cannot compute Σ_{xx}. We choose to approximate Σ_{xx} with the worst-case state-covariance, i.e., W that satisfies (19). Then we can define

$$D_{\text{dc}} := D_0 + \frac{T_{\text{dc}}}{T_{\text{ss}}} \underbrace{\frac{1}{\bar{C}} \begin{bmatrix} W_{\text{dc}} & Z_{\text{dc}} \\ Z_{\text{dc}}^\top & Z_{\text{dc}}^\top W_{\text{dc}}^{-1} Z_{\text{dc}} + \Sigma \end{bmatrix}}_{\tilde{D}_{\text{dc}}}, \tag{22}$$

where $Z_{\text{dc}} = W_{\text{dc}} K_{\text{dc}}^\top$, as the 'worst-case' uncertainty resulting from application of $\mathcal{K}_{\text{dc}}$ for T_{dc} timesteps.

Ideally, we would also take into account the effect of the data $\mathcal{D}_{\text{dc}}$ on our point estimates $\hat{A}, \hat{B}$, for the purpose of dual-control synthesis. To preserve convexity, we will instead use the initial point estimates based on $\mathcal{D}_0$. To summarize, our updated model after running the dual controller can be approximated by $\mathcal{M}(\mathcal{D}_0 \cup \mathcal{D}_{\text{dc}}) \approx \{\hat{A}(\mathcal{D}_0), \hat{B}(\mathcal{D}_0), D_{\text{dc}}\}$.

C. Convex Relaxation

During synthesis of $\mathcal{K}_{\text{dc}}$, the search for the redesigned policy $K = Z^\top W^{-1}$ is constrained by, see (14),

$$S_2(t, Z, W, D_0 + \frac{T_{\text{dc}}}{T_{\text{ss}}} \tilde{D}_{\text{dc}}, \hat{A}, \hat{B}) \succeq 0. \tag{23}$$

For fixed t, this is an LMI in D_{dc}; however, D_{dc} is a nonlinear function of Z_{dc} and W_{dc}, see (22). In what follows, we derive an affine approximation of D_{dc}. As increasing D_{dc} enlarges the feasible set defined by (23), it is desirable for this approximation to lower bound D_{dc}. We make use of the following result.

Lemma 1: The inequality $U^\top M^{-1} U \succeq U^\top V + V^\top U - U^\top M U$ holds for every $M \succ 0$, U and V.

Proof: We have for every $M \succ 0$ that $U^\top M^{-1} U - U^\top V - V^\top U + U^\top M U = \|U - MV\|_{M^{-1}}^2 \succeq 0$. ■

By choosing $M = W_{\text{dc}}$ and $U = \begin{bmatrix} W_{\text{dc}} & Z_{\text{dc}} \end{bmatrix}$ we can lower bound the nonlinear term in (23) as follows:

$$\begin{bmatrix} W_{\text{dc}} & Z_{\text{dc}} \\ Z_{\text{dc}}^\top & Z_{\text{dc}}^\top W_{\text{dc}}^{-1} Z_{\text{dc}} \end{bmatrix} \succeq \begin{bmatrix} W_{\text{dc}} \\ Z_{\text{dc}}^\top \end{bmatrix} V + V^\top \begin{bmatrix} W_{\text{dc}} \\ Z_{\text{dc}}^\top \end{bmatrix}^\top - V^\top W_{\text{dc}} V.$$

This gives an affine lower bound on $\tilde{D}_{\text{dc}}$ for fixed V. The bound is tight when $U = MV$, and so the optimal choice is $V = W_{\text{dc}}^{-1} \begin{bmatrix} W_{\text{dc}} & Z_{\text{dc}} \end{bmatrix} = \begin{bmatrix} I & K_{\text{dc}}^\top \end{bmatrix}$. As K_{dc} is unknown, we instead choose the robust controller for the nominal model $\mathcal{M}(\mathcal{D}_0)$, i.e., $K_0 = \arg\min_K J(K, \mathcal{M}(\mathcal{D}_0))$. With this choice, the bound is tight for $K_{\text{dc}} = K_0$.

We are now in a position to present an approximate convex solution to (6). In what follows, $\mathcal{M}(\mathcal{D}_0) = \{\hat{A}, \hat{B}, D_0\}$.

Consider the following program:

$$
\begin{aligned}
\min_{\substack{t_{\mathrm{dc}}, Z_{\mathrm{dc}}, W_{\mathrm{dc}}, \Sigma \\ t, Z, W, Y, \bar{D}_{\mathrm{dc}}}} \quad & \operatorname{tr} Y && \text{(24a)}\\
\text{s.t.} \quad & t \geq 0,\ t_{\mathrm{dc}} \geq 0,\ S_1(W, Y, Z) \succeq 0, && \text{(24b)}\\
& S_2(t, Z, W, D_0 + \frac{T_{\mathrm{dc}}}{T_{\mathrm{ss}}} \bar{D}_{\mathrm{dc}}, \hat{A}, \hat{B}) \succeq 0, && \text{(24c)}\\
& \operatorname{tr} Y_{\mathrm{dc}} \leq J_{\mathrm{exp}},\ S_1(W_{\mathrm{dc}}, Y_{\mathrm{dc}}, Z_{\mathrm{dc}}) \succeq 0, && \text{(24d)}\\
& S_{\mathrm{dc}}(t_{\mathrm{dc}}, Z_{\mathrm{dc}}, W_{\mathrm{dc}}, \Sigma, D_0, \hat{A}, \hat{B}) \succeq 0, && \text{(24e)}\\
\bar{D}_{\mathrm{dc}} \preceq \frac{1}{\bar{C}} & \begin{bmatrix} W_{\mathrm{dc}} & Z_{\mathrm{dc}} \\ Z_{\mathrm{dc}}^\top & Z_{\mathrm{dc}}^\top K_0^\top + K_0 Z_{\mathrm{dc}} - K_0 W_{\mathrm{dc}} K_0^\top + \Sigma \end{bmatrix} && \text{(24f)}
\end{aligned}
$$

For fixed t, (24f) is an SDP. To circumvent the non-convexity of the product between t and $\bar{D}_{\mathrm{dc}}$ in (24c) we can grid over the scalar variable t. The computational complexity scales linearly in the number of grid points.

To summarize, in (24) we approximated the robust dual control problem given by (6) as follows: i) we approximated D_{dc} by its stationary distribution, see (21), and later approximated Σ_{xx} with the worst-case state-covariance, see (22); ii) we derived an affine approximation of D_{dc} (see Lemma 2); iii) we did not consider the effect of the data $\mathcal{D}_{\mathrm{dc}}$ on the point estimates $\hat{A}$, $\hat{B}$, in order to preserve convexity.

V. Numerical Experiments

We now illustrate our results on robust and dual control with numerical experiments in MATLAB using YALMIP for formulation [34] with MOSEK [35] as the solver. We compared our robust controller to the robust controller designed in [29]. We also compared our dual controller with a greedy random exploration strategy. For these comparisons, we randomly generated 100 systems of the form (1). Each entry in $A \in^{3\times 3}$ and $B \in^{3\times 2}$ was sampled from $\mathcal{N}(0, 1)$. A was scaled so as to have spectral radius of 1.05, and controllability of each system was verified. The methods were also compared on the particular system

$$
A = \begin{bmatrix} 1.1 & 0.5 & 0 \\ 0 & 0.9 & 0.2 \\ 0 & -0.2 & 0.8 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 0.1 & 0 \\ 0 & 2 \end{bmatrix}.
$$

In both cases $\sigma_w = 0.5$ and the user defined matrices for calculating the cost were $Q = I$ and $R = \operatorname{diag}(10, 1)$.

A. Comparison of Robust Controller With [29]

For each trial, $\mathcal{D}_0$ was constructed by running $N = 500$ rollouts, each of length $T_{\mathrm{r}} = 6$. During each rollout, white noise with unit variance σ_u was used as the input u. Given $\mathcal{D}_0$, and $\delta = 0.05$, two robust controllers were synthesized using: i) the method proposed in Section III (rc-prop); ii) the synthesis method proposed in [29] with the confidence intervals ϵ_A and ϵ_B taken as the square root of the maximum eigenvalues of diagonal sub-blocks (of appropriate dimension) of D_0 (rc-sls). This process was repeated for 100 randomly generated systems, see Figure 1, and for 100 Monte Carlo trials with the particular system above, see Figure 2. The red line in the middle of each box denotes the median. The tops and bottoms of each box are the 25th and 75th percentiles, and observations beyond 1.5 times the interquartile range are marked as outliers by red crosses. We evaluate performance via two metrics: i) the infinite horizon LQ-cost when the controller is applied to the true system, ii) the theoretical worst-case cost given $\mathcal{D}_0$, i.e., $J(K, \mathcal{M}(\mathcal{D}_0))$ in (4). The proposed controller demonstrates superior performance in all cases. Notice that rc-sls and rc-prop are synthesized using the same model $\mathcal{M}(\mathcal{D}_0)$; however, rc-prop takes into account the structure of D_0, while rc-sls makes use of the spectral norm only.

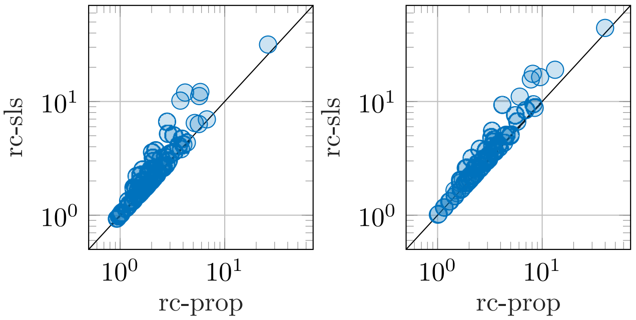

Fig. 1. Comparison of performance of the proposed robust controller and the robust controller in [29] on the randomly generated systems using: i) true system dynamics (left), ii) the worst-case cost (right).

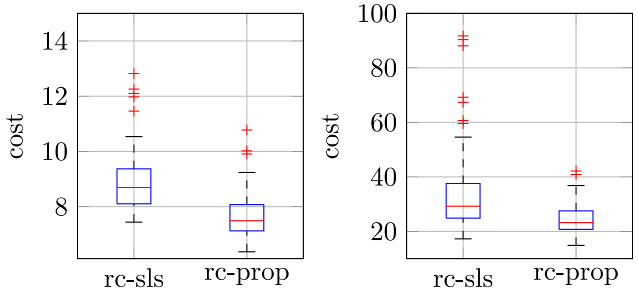

Fig. 2. Comparison of performance of the proposed robust controller and the robust controller in [29] on the particular system using true system dynamics (left), and the worst-case cost (right).

B. Comparison of Dual Controller With Greedy Exploration

For each trial, the initial dataset $\mathcal{D}_0$ is comprised of $N = 500$ rollouts of length $T = 6$, with white noise excitation u of variance $\sigma_u = 1$. Given $\mathcal{D}_0$, and $\delta = 0.05$, our objective is to solve the dual control problem (6) with $T_{\mathrm{dc}} = 1000$. We compare the following methods: i) dc: the dual control strategy proposed in Section IV. For the grid we used 20 points which are equally spaced between 0.05 and $2t_{nom}$, where t_{nom} is attained by solving (15) with $\mathcal{D}_0$; ii) exp: a greedy exploration strategy with policy $u_t = K_0 + e_t$, where $K_0 = \arg\min_K J(\{K, 0\}, \mathcal{M}(\mathcal{D}_0))$ and $e_t \sim \mathcal{N}(0, \Sigma_e)$ with Σ_e tuned so that the theoretical worst-case cost $J(\{K_0, \Sigma_e\}, \mathcal{M}(\mathcal{D}_0))$ is equal to the exploration budget J_{exp}. The exploration budget was set to $1.2J(\{K_0, 0\}, \mathcal{M}(\mathcal{D}_0))$, though the results are qualitatively insensitive to the exploration budget. The two exploration methods are evaluated via the following metrics, see Figure 3. First, we compare the worst-case cost achieved by the redesigned robust controllers, after exploration, using the theoretical worst-case uncertainty reduction, see (22). From Figure 3, we see that the proposed method results in significantly lower cost after redesign, compared to the greedy strategy. Next, we compare the empirical cost of exploration

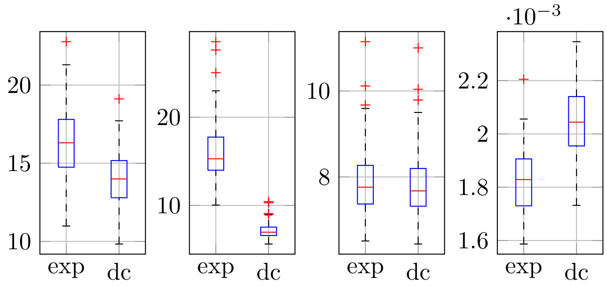

Fig. 3. From left to right: Comparison of i) worst-case cost of controller after redesign using theoretical exploration, ii) empirical cost of exploration, iii) empirical cost of controller on true system after redesign, iv) 2-norm of worst-case inverse uncertainty matrices, i.e., $\|D^{-1}\|_2$.

when the exploration policies are applied to the true system. Interestingly, though both policies are tuned to achieve the same worst-case exploration cost (given by the budget) the proposed approach delivers significantly lower exploration cost on the true system. Next, we apply the redesigned controllers to the true system, and observe comparable cost for each. Finally, we investigate the nature of the uncertainty reduction achieved by each exploration policy. In Figure 3 we plot $\|D^{-1}\|_2$ for each method, where D denotes the 'worst-case uncertainty matrix' given in (22). This is a scalar measure of the magnitude of uncertainty. Notice that the absolute uncertainty achieved by the proposed method is larger than that achieved by greedy strategy; however, the performance of the re-designed controller is much lower. This suggests that the proposed method is reducing the uncertainty in a structured way, targeting uncertainty reduction in the parameters that 'matter most for control'.

A standard laptop computer runs the simulations. Finding the robust controller which involves solving the SDP problem takes about 0.05 s, and finding the approximate dual controller for 20 grid points takes about 2.5 s.

VI. Conclusion

In this letter, we considered a design of robust LQ controllers with unknown dynamics. First, we designed a nominal robust controller using initial data uncertainty obtained from Proposition 1. Next, we designed a robust dual controller. The objective of the robust dual controller is to minimize the worst-case cost attained by a new robust controller, synthesized with the reduced model uncertainty, subject to constraints on the exploration cost. Numerical simulations show that our nominal controller has lower worst-case cost than the robust controller proposed in [29] for both true system and the worst-case cost. This implies that our robust controller is less conservative. Moreover, the worst-case cost of the proposed dual controller was reduced in comparison to a greedy random exploration strategy.

References

[1] A. A. Feldbaum, "Dual control theory. I," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

[2] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[3] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484, 2016.

[4] A. A. Feldbaum, "Dual control theory. II," *Avtomatika i Telemekhanika*, vol. 21, no. 11, pp. 1453–1464, 1960.

[5] K. J. Åström and B. Wittenmark, "Problems of identification and control," *J. Math. Anal. Appl.*, vol. 34, no. 1, pp. 90–113, 1971.

[6] Y. Bar-Shalom, "Stochastic dynamic programming: Caution and probing," *IEEE Trans. Autom. Control*, vol. AC-26, no. 5, pp. 1184–1195, Oct. 1981.

[7] Y. Bar-Shalom and E. Tse, "Caution, probing, and the value of information in the control of uncertain systems," *Ann. Econ. Soc. Meas.*, vol. 5, no. 3, pp. 323–337, 1976.

[8] M. Gevers and L. Ljung, "Optimal experiment designs with respect to the intended model application," *Automatica*, vol. 22, no. 5, pp. 543–554, 1986.

[9] H. Jansson and H. Hjalmarsson, "Input design via LMIs admitting frequency-wise model specifications in confidence regions," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1534–1549, Oct. 2005.

[10] H. Hjalmarsson and H. Jansson, "Closed loop experiment design for linear time invariant dynamical systems via LMIs," *Automatica*, vol. 44, no. 3, pp. 623–636, 2008.

[11] H. Hjalmarsson, "System identification of complex and structured systems," *Eur. J. Control*, vol. 15, nos. 3–4, pp. 275–310, 2009.

[12] M. Annergren, C. A. Larsson, H. Hjalmarsson, X. Bombois, and B. Wahlberg, "Application-oriented input design in system identification: Optimal input design for control [applications of control]," *IEEE Control Syst. Mag.*, vol. 37, no. 2, pp. 31–56, Apr. 2017.

[13] C. A. Larsson, C. R. Rojas, X. Bombois, and H. Hjalmarsson, "Experimental evaluation of model predictive control with excitation (MPC-X) on an industrial depropanizer," *J. Process Control*, vol. 31, pp. 1–16, Jul. 2015.

[14] C. A. Larsson, A. Ebadat, C. R. Rojas, X. Bombois, and H. Hjalmarsson, "An application-oriented approach to dual control with excitation for closed-loop identification," *Eur. J. Control*, vol. 29, pp. 1–16, May 2016.

[15] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013.

[16] T. A. N. Heirung, B. E. Ydstie, and B. Foss, "Dual adaptive model predictive control," *Automatica*, vol. 80, pp. 340–348, Jun. 2017.

[17] M. Tanaskovic, L. Fagiano, R. Smith, and M. Morari, "Adaptive receding horizon control for constrained MIMO systems," *Automatica*, vol. 50, no. 12, pp. 3019–3029, 2014.

[18] M. Lorenzen, M. Cannon, and F. Allgöwer, "Robust MPC with recursive model update," *Automatica*, vol. 103, pp. 461–471, May 2019.

[19] M. Barenthin and H. Hjalmarsson, "Identification and control: Joint input design and H_∞ state feedback with ellipsoidal parametric uncertainty via LMIs," *Automatica*, vol. 44, no. 2, pp. 543–551, 2008.

[20] M. Gevers, "Towards a joint design of identification and control?" in *Essays on Control*. Boston, MA, USA: Springer, 1993, pp. 111–151.

[21] H. Hjalmarsson, "From experiment design to closed-loop control," *Automatica*, vol. 41, no. 3, pp. 393–438, 2005.

[22] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.

[23] M. Ibrahimi, A. Javanmard, and B. V. Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2636–2644.

[24] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time adaptive stabilization of LQ systems," *IEEE Trans. Autom. Control*, to be published.

[25] Y. Ouyang, M. Gagrani, and R. Jain, "Learning-based control of unknown linear systems with Thompson sampling," *IEEE Trans. Autom. Control*, to be published.

[26] M. Abeille and A. Lazaric, "Thompson sampling for linear-quadratic control problems," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1246–1254.

[27] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4192–4201.

[28] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *Proc. Amer. Control Conf.*, 2019.

[29] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Found. Comput. Math.*, 2017.

[30] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Sci., 1995.

[31] B. Efron, "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 569–593.

[32] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proc. Conf. Learn. Theory*, 2018, pp. 1–35.

[33] Z.-Q. Luo, J. F. Sturm, and S. Zhang, "Multivariate nonnegative quadratic mappings," *SIAM J. Optim.*, vol. 14, no. 4, pp. 1140–1162, 2004.

[34] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, vol. 3. New Orleans, LA, USA, 2004, pp. 284–289.

[35] *The MOSEK Optimization Toolbox for MATLAB Manual*, MOSEK ApS, Copenhagen, Denmark, 2015.