

# Structured exploration in the finite horizon linear quadratic dual control problem <sup>★</sup>

Andrea Iannelli,<sup>\*</sup> Mohammad Khosravi,<sup>\*</sup> Roy S. Smith<sup>\*</sup>

<sup>\*</sup> *Automatic Control Lab, ETH, Zürich 8092, Switzerland  
(e-mail: iannelli/khosravm/rsmith@control.ee.ethz.ch).*

---

**Abstract:** This paper presents a novel approach to synthesize dual controllers for unknown linear time-invariant systems with the tasks of optimizing a quadratic cost while reducing the uncertainty. To this end, a synthesis problem is defined where the feedback law has to simultaneously gain knowledge of the system and robustly optimize the cost. By framing the problem in a finite horizon setting, the trade-offs arising when the tasks include both identification and control are formally captured in the optimization problem. Results show that efficient exploration strategies are achieved when the structure of the problem is exploited.

*Keywords:* Robustness, Linear Quadratic Regulation, Experiment design, Exploration

---

## 1. INTRODUCTION

One of the most widespread approaches in control is the Linear Quadratic (LQ) regulator, whereby the goal is to design a feedback law which minimizes deviations of the states from a desired reference trajectory (e.g. the origin) while keeping as small as possible the necessary action. In the full state information case (standard LQR), when the dynamics is *exactly* known the problem has a well known optimal solution (see Bertsekas (2005)). In the infinite horizon case, that is when transient features are negligible and the problem is approximated in an infinitely long time window, this consists of a static gain matrix associated with the solution of an Algebraic Riccati Equation (ARE). When the problem is studied on a finite horizon, the optimal feedback law is time-varying and is associated with a Riccati difference (or differential) equation (DRDE).

Despite important control theoretic works on robust  $\mathcal{H}_2$  analysis and filtering problems (Sznaier et al., 2002; Sun and Packard, 2005), the solution of the LQ control problem when the dynamics is unknown is far less understood. Notably, this has been used in the last few years as case study to show possible complementarities of Reinforcement Learning (RL) and control theory-based approaches for the fundamental problem of optimally manipulating an unknown system by using the information carried in the collected data (Recht, 2019). Bridging these two communities has been the effort of many recent works, see e.g. (Cohen et al., 2018; Dean et al., 2019; Matni et al., 2019), but despite the variety of techniques considered, no strategies which allow for an easy implementation on one hand, and provide optimal cost guarantees on the other, have been found (Recht, 2019). Moreover, a fundamental unsolved problem is what is the best strategy to extract information about the system such that the performance can be improved while preserving at the same time safety.

In other words, borrowing terminology from the reinforcement learning community (e.g. multi-armed bandits, Sutton and Barto (2018)), specifying an optimal *policy* (control law) that robustly balances *exploration* (acquiring knowledge of the system by testing and identification) and *exploitation* (operating the system to maximize the *reward*, or performance).

The approach considered here owes to the long and rich tradition of dual control (Åström and Wittenmark, 1971), where the problem of simultaneously identifying and controlling a system was first formalized, and experiment design, whereby one attempts to determine the most suitable inputs in order to extract information from the unknown plant (Annergren et al., 2017). The material presented here is also inspired by a recent publication (Ferizbegovic et al. (2020)) where the unknown LQ problem is framed in a dual control setting. Specifically, given an initial estimate of the dynamics in the form of nominal state matrices and an ellipsoidal uncertainty set, the joint optimization of two robust feedback laws  $G_{K1}$  and  $G_{K2}$  is proposed therein considering two distinct infinite horizon problems.

While retaining the same application-oriented philosophy, i.e. promoting a reduction of the uncertainty which is beneficial for the purpose of minimizing a given cost, the work here substantially changes the synthesis approach by framing the problem in a finite horizon setting. This is motivated by the fact that the dual control problem is more realistically described in a finite time window, due to the importance of transient features. The design of two robust static feedback laws tasked with different goals is thus shifted to that of a single time-varying law  $G_K$ , responsible for dealing simultaneously with the two tasks. From an optimal control perspective, but in a dual control setting now, the problem is formulated as the solution of an DRDE rather than of an ARE. The main advantage is that by framing the problem in a finite horizon setting the different trade-offs between exploration and exploitation are captured and can be optimized over. New insights

---

<sup>★</sup> This work is supported by the Swiss National Science Foundation under grant no. 200021\_178890 and by the Swiss Competence Center for Energy Research FEEB&D.

into these trade-offs are in turn believed to help gaining a deeper understanding of the unknown LQ problem.

The main technical contribution of the paper is the formulation of a Semidefinite program (SDP) to solve the robust dual LQR control problem in the finite horizon setting optimally balancing exploration and exploitation. This is presented in Section 3, where also the corresponding programs for the nominal and robust (but with fixed uncertainty, i.e. without exploration) problems are derived. The other important contribution is gathered in Section 4, where features of the synthesised policies are shown through numerical examples. The application-oriented nature of the exploration strategy proposed here is termed *structured*, in order to emphasize the ability to exploit specific properties of the analysed system.

## 2. PROBLEM DESCRIPTION

### 2.1 Background

Consider the discrete linear time-invariant system:

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}(0, \sigma_w^2 I_{n_x}), \quad x_0 = 0, \quad (1)$$

where  $x_t \in \mathbb{R}^{n_x}$  is the (measured) state,  $u_t \in \mathbb{R}^{n_u}$  is the control input, and  $w_t \in \mathbb{R}^{n_x}$  is the normally distributed process noise with zero mean and covariance  $\sigma_w^2 I_{n_x}$ . Given cost matrices  $Q \succeq 0$  and  $R \succeq 0$ , the objective is to design a feedback law minimizing the expected finite horizon quadratic cost  $J$  in  $[1, T]$  (with  $1 < T < \infty$ ):

$$J = \mathbb{E} \left[ \sum_{t=1}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t) + x_T^\top Q x_T \right], \quad (2)$$

where the expectation is with respect to  $w$ . When  $A$  and  $B$  are known, the optimal input is given by the time-varying state-feedback law  $u_t = K_t^{\text{DRDE}} x_t$ , where  $K_t^{\text{DRDE}}$  is associated with the stabilizing solution of the discrete time Riccati difference equation (DRDE) for (1).

The case of unknown  $A$  and  $B$ , where the only access to information on (1) is through measurements of  $x$  and  $u$ , is considered here. An estimation of the unknown dynamics is obtained through the so-called Coarse-ID approach (Dean et al. (2019)). Given a dataset made of  $N$  samples  $\mathcal{S} = \{(x_t, u_t) : 1 \leq t \leq N\}$ , the *nominal dynamics* is estimated through the least squares problem:

$$(\hat{A}, \hat{B}) = \arg \min_{A, B} \sum_{t=1}^{N-1} \|-x_{t+1} + Ax_t + Bu_t\|_2^2, \quad (3)$$

and the *true dynamics* belongs to the ellipsoidal set  $\Omega$ :

$$\Omega(X, D) = \{X : X^\top D X \preceq I\}, \quad D \in \mathbb{S}^{n_x + n_u}, \quad (4a)$$

$$X = \begin{bmatrix} (\hat{A} - A)^\top \\ (\hat{B} - B)^\top \end{bmatrix}, \quad X \in \mathbb{R}^{(n_x + n_u) \times n_x}, \quad (4b)$$

where  $D$  defines the uncertainty. A possible estimate for  $D$  can be obtained by making use of an empirical Bayes argument and taking it as the variance of the posterior distribution of  $(A, B)$  given  $\mathcal{S}$  (Umenberger et al. (2019)). Precisely, (4a) holds with probability  $1 - \delta$  for:

$$D = \frac{1}{c_\chi \sigma_w^2} \sum_{t=1}^{N-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top, \quad c_\chi = \chi_{(n_x^2 + n_x n_u), \delta}^2, \quad (5)$$

where  $\chi_{n, \delta}^2$  is the critical value for a Chi-squared distribution with  $n$  degrees of freedom and probability level  $\delta$ .

Given this uncertainty description, the objective is to synthesize a policy  $G_K$  that minimizes the worst-case  $J$ :

$$J_{\text{WC}} = \min_{G_K} \max_{(A, B) \in \Omega(X, D)} J. \quad (6)$$

To this end,  $G_K(K_t, S_t)$  is parametrized as:

$$u_t = K_t x_t + e_t, \quad e_t \sim \mathcal{N}(0, S_t), \quad S_t \in \mathbb{S}^{n_u}. \quad (7)$$

The law consists of a time-varying state-feedback part, and a random excitation input  $e_t$  with time-varying covariance for the purpose of exploration. The time-varying formulation of  $S_t$  captures the fact that, as knowledge of the system is acquired, the random part of the excitation should decrease. This aspect is also found in several methods proposed in the RL community, e.g. the concept of exploration rate in  $\epsilon$ -greedy algorithms (Sutton and Barto, 2018). Formal ways for adapting rates to the learning progress have been proposed in the learning literature (Auer et al., 2002; Cesa-Bianchi et al., 2017), e.g. leveraging the concept of regret bounds (Matni et al., 2019). In this formulation,  $S_t$  will be an optimization variable and thus its value will depend, among other things, on the properties of the system to be identified.

### 2.2 Motivating example

The importance of formulating the dual control problem in a finite horizon, rather than in an infinite one as proposed in (Ferizbegovic et al., 2020) and in general in the recent learning-LQR literature (Cohen et al., 2018; Dean et al., 2019), is discussed here.

Consider the scenario where the unknown system (1) has to be operated over a certain horizon  $[1, T]$  while minimizing (2). Since the dynamics are not known, a simple strategy consists of choosing an intermediate time  $T_{sw} < T$ , and dividing the horizon into two phases. In the first (exploration, or *ID-phase*), the system is excited with random input  $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$  and the measured response (e.g. in the form of  $\mathcal{S}$ ) is used to identify the nominal matrices through (3). In the second (exploitation, or *K-phase*), a controller which optimizes (2) for the identified nominal matrices is synthesised. One possible option is the use of time-varying feedback  $K_t^{\text{DRDE}}$  on the remaining horizon  $[T_{sw}, T]$ . That is, a pure exploration phase is followed by a pure exploitation phase (often termed *explore-then-commit* in the RL literature (Garivier et al., 2016)). This clearly leads to a trade-off between the *duration* of these two phases, where for high  $T_{sw}$  the benefit of estimating the model more accurately contrasts with the disadvantage of optimally controlling the plant for a shorter time, and viceversa.

In order to exemplify this aspect, an experiment is performed on a horizon of length  $T=100$  with two randomly generated stable plants having  $n_x=3$ ,  $n_u=2$ , and using  $\sigma_u=1$ ,  $\sigma_w=0.5$ . Figure 1 shows the total expected cost  $J_{\text{tot}}$  (obtained by averaging over 100 realizations of noise and random excitation) as a function of  $T_{sw}$ . The total cost can be broken down into  $J_{\text{ID}}$  and  $J_K$ , associated respectively with the identification part in the horizon  $[1, T_{sw}]$  and with the deployment of the controller in  $[T_{sw}, T]$ .

It can be observed that there exists an *optimal* switching time where the benefit of further exploring the unknown dynamics is overcome by the cost of exploration. This

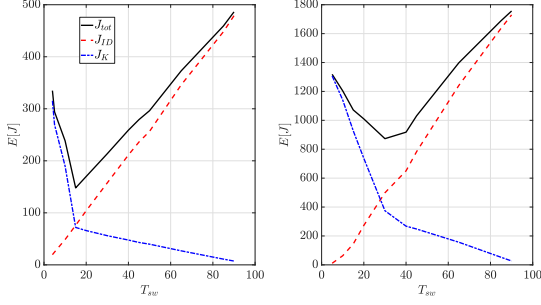


Fig. 1. Expected costs for the two unknown plants as a function of the switching time  $T_{sw}$ .

trade-off depends on the unknown true system, and it can only be captured in a finite horizon setting, where distinctive transient features of the *ID-phase* and *K-phase* are retained. While the explore-then-commit strategy tested here captures fundamental conflicting aspects arising in the dual control problem, it has important limitations, among which are robustness ( $K_t^{\text{DRDE}}$  is not robust to estimation errors) and optimality (the interplay between identification and control is not exploited since exploration and exploitation are sequentially applied). The next section addresses these fundamental aspects of the dual control problem by proposing a novel synthesis strategy.

### 3. SEMIDEFINITE PROGRAMS FORMULATIONS

The goal is to derive a convex formulation for synthesising a feedback law  $G_K$  (7) that optimizes  $J_{\text{WC}}$  (6). In order to clearly present the steps involved and highlight their meaning, the presentation is broken down into 3 parts. Section 3.1 deals with the nominal case (where the estimated system coincides with the true one, i.e.  $A = \hat{A}$  and  $B = \hat{B}$ ), conceptually equivalent to solving the associated DRDE. Section 3.2 considers a worst-case design where the set of uncertainty is fixed throughout the horizon, which thus is a robust version of the DRDE. Finally, Section 3.3 establishes the dual control formulation, where exploration is promoted and thus the uncertainty of the system can be reduced while robustly controlling the plant.

The key idea is to use the application-oriented approach first introduced in (Ferizbegovic et al., 2020), with the important differences that the problem is formulated in the finite horizon and a single (time-varying) policy responsible for simultaneously exploring and controlling is synthesized. Exploration is used only to update the uncertainty matrix  $D$  (5), while the nominal matrices  $\hat{A}$  and  $\hat{B}$  are kept fixed. This is in line with related works (Ferizbegovic et al., 2020; Umenberger et al., 2019) that make use of the certainty equivalence assumption.

The first step consists of deriving an expression for the cost  $J$  which can be used in the robust optimization problem (6). Let us begin by denoting by  $P_t$  the covariance matrix of the state at timestep  $t$ :

$$P_t = \mathbb{E} [x_t x_t^\top] \in \mathbb{S}^{n_x}. \quad (8)$$

Define also  $\bar{Q}_t := \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}} K_t \end{bmatrix} \in \mathbb{R}^{(n_x+n_u) \times n_x}$ ,  $\bar{R} := \begin{bmatrix} 0 \\ R^{\frac{1}{2}} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u) \times n_x}$ . Then the following result, proved in the Appendix, holds.

*Lemma 1.* The cost  $J$  in (2) is equivalent to:

$$J = \text{Tr} \left( \sum_{t=1}^{T-1} (\bar{Q}_t P_t \bar{Q}_t^\top + \bar{R} S_t \bar{R}^\top) + Q P_T Q^\top \right). \quad (9)$$

The benefit of Lemma 1 is that it allows the finite horizon cost to be rewritten as a function of the covariances  $P_t$ , as it is the case for the infinite horizon cost, whose minimization in turn is equivalent to the computation of the  $\mathcal{H}_2$  norm of (1).

#### 3.1 Nominal design

The solution to the nominal problem, as in the SDP-based computation of the  $\mathcal{H}_2$  norm, can be obtained by minimizing (9) while constraining the covariance  $P_t$  to satisfy  $\forall t \in [1, T-1]$  the discrete time Lyapunov inequalities associated with the closed loop (1)-(7):

$$\min_{G_K} \text{Tr} \left( \sum_{t=1}^{T-1} (\bar{Q}_t P_t \bar{Q}_t^\top + \bar{R} S_t \bar{R}^\top) + Q P_T Q^\top \right), \quad (10a)$$

$$P_{t+1} \succeq (A + B K_t) P_t (A + B K_t)^\top + \sigma_w^2 I + B S_t B^\top, \quad (10b)$$

$$P_1 \succeq \sigma_w^2 I, \quad (10c)$$

where (10c) comes from the assumed zero initial condition on the state (that is,  $P_0 \equiv 0$ ). For generality, and to make more clear the differences among the 3 synthesis approaches, the generic policy  $G_K(K_t, S_t)$  is considered. However, as intuitive and confirmed later by the results, the random excitation part  $S_t$  will be zero in this case.

The program in (10) is convex and can be recast as an SDP with well known Linear Matrix Inequalities (LMI) manipulations (Scherer and Weiland, 2000). First, (10a) is upper bounded by replacing the argument of the summation in (10a) with  $Y_t \in \mathbb{S}^{n_x+n_u} \succeq 0$ , and the new objective function (10a) is written by using Schur complement as:

$$\min_{G_K} \text{Tr} \left( \sum_{t=1}^{T-1} Y_t + Q P_T Q^\top \right), \quad (11)$$

$$\begin{bmatrix} Y_t - \bar{R} S_t \bar{R}^\top & \bar{Q}_t P_t \\ P_t \bar{Q}_t^\top & P_t \end{bmatrix} \succeq 0, \quad \forall t \in [1, T-1].$$

Note that  $\bar{Q}_t$  and  $P_t$  give rise to bilinear terms, thus the auxiliary variable  $Z_t = P_t K_t^\top$  is defined. As for the inequalities (10b), they can be recast as coupled LMIs by application of Schur complement. This leads to:

*Program 1.* Nominal design

$$\min_{G_K} J = \min_{Y_t, P_t, Z_t, S_t} \text{Tr} \left( \sum_{t=1}^{T-1} Y_t + Q P_T Q^\top \right), \quad (12a)$$

$$\begin{bmatrix} Y_t - \bar{R} S_t \bar{R}^\top & \begin{bmatrix} Q^{\frac{1}{2}} P_t \\ R^{\frac{1}{2}} Z_t^\top \end{bmatrix} \\ * & P_t \end{bmatrix} \succeq 0, \quad (12b)$$

$$\begin{bmatrix} P_t & F_t \\ * & P_{t+1} - \sigma_w^2 I - B S_t B^\top \end{bmatrix} \succeq 0, \quad (12c)$$

$$Y_t \succeq 0, S_t \succeq 0, P_{t+1} \succeq 0, \quad \forall t \in [1, T-1],$$

$$P_1 \succeq \sigma_w^2 I,$$

where  $F_t := P_t A^\top + Z_t B^\top$ . Solving Program 1 is conceptually equivalent to solving the DRDE and leads to identical

results (see Fig. 2). The advantage of this formulation is that it allows robustness constraints to be enforced and the effect of exploration to be included.

### 3.2 Robust control design

In the unknown dynamics case, the only knowledge is that  $(A, B) \in \Omega(X, D_0)$ , where  $\hat{A}$ ,  $\hat{B}$ , and  $D_0$  are assumed to be available from prior experiments or approximate knowledge of the system. Therefore, the LMIs (12c) have to be hold for all possible  $(A, B)$  inside  $\Omega$ . To solve this robust optimization problem,  $A$  and  $B$  are written as a function of  $X$ ,  $\hat{A}$ , and  $\hat{B}$  using definition (4b) and a Schur complement is applied to overcome the nonlinearity arising from  $BS_tB^\top$ . Then, since  $X$  has to lie inside an ellipsoidal set (4a), an extension of the S-lemma to the matrix case, proposed in (Luo et al., 2004), is employed and the following program is proposed.

*Program 2.* Robust control design

$$J_{\text{WC}} = \min_{Y_t, P_t, Z_t, S_t, p_t} \text{Tr} \left( \sum_{t=1}^{T-1} Y_t + QP_TQ^\top \right), \quad (13a)$$

$$\begin{bmatrix} Y_t - \bar{R}S_t\bar{R}^\top & \begin{bmatrix} Q^{\frac{1}{2}}P_t \\ R^{\frac{1}{2}}Z_t^\top \end{bmatrix} \\ * & P_t \end{bmatrix} \succeq 0, \quad (13b)$$

$$\begin{bmatrix} \begin{bmatrix} P_t & 0 \\ * & S_t \end{bmatrix} & H_t & G_t \\ * & P_{t+1} - \sigma_w^2 I - p_t I & 0 \\ * & * & p_t D_0 \end{bmatrix} \succeq 0, \quad (13c)$$

$$Y_t \succeq 0, S_t \succeq 0, P_{t+1} \succeq 0, p_t \geq 0, \forall t \in [1, T-1], \\ P_1 \succeq \sigma_w^2 I,$$

where  $G_t := -\begin{bmatrix} P_t & Z_t \\ 0 & S_t \end{bmatrix}$ ,  $H_t := -G_t \begin{bmatrix} \hat{A}^\top \\ \hat{B}^\top \end{bmatrix}$ , and  $p_t$  is a multiplier from the S-lemma. The crucial feature of Program 2 is that the ellipsoid  $\Omega(X, D_0)$  defining the uncertainty is fixed throughout the horizon. The consequence of this is that exploration is not encouraged, since it has an associated cost for which is not rewarded. In other words, the generation of control inputs with a different goal than just minimizing the performance objective will inevitably incur in a higher cost (or *regret*). This argument has a clear interpretation for  $S_t$ , where the LMIs (12b)-(13b) show that non-zero  $S_t$  always determine an additional contribution to the cost via  $Y_t$ . In fact, the random excitation part  $S_t$  will be zero here (as it was commented on earlier for the nominal case). As for  $K_t$ , this will correspond to the stabilizing solution of the DRDE formed by taking at each time-step the worst-case matrices (which are in principle time-varying), i.e. it is the robust optimal policy.

### 3.3 Robust dual control design

In order to promote exploration, it is necessary to describe how the feedback law contributes to obtain knowledge of the system. More formally, a mapping between  $G_K$  and the uncertainty  $D_t$  at a given time  $t$  has to be formulated. From its definition in (4), the following is proposed:

$$D_t = \frac{1}{c_X \sigma_w^2} \sum_{l=1}^t \begin{bmatrix} P_l & Z_l \\ * & (Z_l^\top P_l^{-1} Z_l) + S_l \end{bmatrix}. \quad (14)$$

Note the explicit influence of  $S_t$  and  $K_t$  (via  $Z_t = P_t K_t^\top$ ) on  $D_t$ , with the policy also having an indirect effect on  $P_t$ .

Due to the nonlinearity involving  $Z_l$  and  $P_l$  in the lower diagonal block, (14) cannot be readily used and thus a convex relaxation is sought. To this end, the matrix inequality in (Ferizbegovic et al. (2020), Lemma 1) is employed here to formulate, for a given matrix  $\bar{K} \in \mathbb{R}^{n_u \times n_x}$ , the following lower bound on  $D_t$ :

$$D_t \succeq \hat{D}_t = \frac{1}{c_X \sigma_w^2} \sum_{l=1}^t \begin{bmatrix} P_l & Z_l \\ * & Z_l^\top \bar{K}^\top + \bar{K} Z_l - \bar{K} P_l \bar{K}^\top + S_l \end{bmatrix}. \quad (15)$$

The bound is tight when  $\bar{K}_t = K_t$ . In this work  $\bar{K}_l$  is chosen as the nominal controller from Program 1. The following dual control design problem is then proposed.

*Program 3.* Robust dual control design

$$J_{\text{WC}} = \min_{Y_t, P_t, Z_t, S_t, p_t} \text{Tr} \left( \sum_{t=1}^{T-1} Y_t + QP_TQ^\top \right), \quad (16a)$$

$$\begin{bmatrix} Y_t - \bar{R}S_t\bar{R}^\top & \begin{bmatrix} Q^{\frac{1}{2}}P_t \\ R^{\frac{1}{2}}Z_t^\top \end{bmatrix} \\ * & P_t \end{bmatrix} \succeq 0, \quad (16b)$$

$$\begin{bmatrix} \begin{bmatrix} P_t & 0 \\ * & S_t \end{bmatrix} & H_t & G_t \\ * & P_{t+1} - \sigma_w^2 I - p_t I & 0 \\ * & * & p_t (D_0 + \hat{D}_t) \end{bmatrix} \succeq 0, \quad (16c)$$

$$Y_t \succeq 0, S_t \succeq 0, P_{t+1} \succeq 0, p_t \geq 0, \forall t \in [1, T], \\ P_1 \succeq \sigma_w^2 I.$$

The bilinearity between  $p_t$  and  $\hat{D}_t$  is overcome by using a line search on  $p_t$  (assumed constant for simplicity, but allowed to be time-varying).

Program 3 clearly shows that the policy  $G_K$  can now perform application-oriented exploration. The key enabler is the policy-dependent and time-varying upper bound on the true uncertainty  $\hat{D}_t$  in (16c). The feedback law is indeed optimized so that the system's response will allow the worst-case matrices  $A$  and  $B$  to be eliminated from the uncertainty set, to the benefit of the feasibility of the LMIs (16c) and in turn of the achievable cost. Exploration itself, however, has a cost and thus trade-offs will arise. The cost associated with  $S_t$  is seen directly in (16b), while that related to  $K_t$  can be interpreted as due to the deviation of  $K_t$  from the robust optimal policy. The first trade-off is on which part of the policy  $G_K(K_t, S_t)$  should be used for exploration, whether the state-feedback, the random excitation or both. Another trade-off is on which portion of the horizon exploration should be pursued (reminiscent of the scenario in Figure 1). It is important in this regard to note that a conceptually similar (convex) formulation for the mapping (15) between the policy and the uncertainty  $D_t$  was proposed in (Ferizbegovic et al., 2020). However, while here the cost to pay to keep adding contributions to  $\hat{D}_t$  is well captured in Program 3, it is not clear how this can be accounted for in an infinite horizon setting, where  $J$  is effectively an averaged cost and thus does not depend on how many terms (i.e. samples) are featured in the summation leading to  $\hat{D}_t$ .

#### 4. NUMERICAL EXAMPLES

Consider the following system:

$$A = \begin{bmatrix} 0.18 & 0.1 & 0 \\ 0 & 0.18 & 0.04 \\ 0 & -0.04 & 0.16 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}. \quad (17)$$

with cost matrices  $Q = I_{n_x}$  and  $R = \text{blkdiag}(10,1)$ ,  $\sigma_w^2 = 0.5$ , and  $\delta = 0.05$ . First, a Coarse-ID estimate of the system is obtained through 100 simulated roll outs (each of length  $T_r = 5$ ) with  $u_t \sim \mathcal{N}(0, \sigma_u^2 I_{n_x})$ ,  $\sigma_u^2 = 1$ . Figure 2 shows the time-varying gains of the feedback matrix  $K_t$  optimized on the horizon  $[1, 100]$  using different design schemes:  $K^{\text{DRDE}}$  by solving the DRDE associated with  $(\hat{A}, \hat{B})$  (3);  $K^{\text{Nom}}$  by solving Program 1 for  $(\hat{A}, \hat{B})$ ;  $K^{\text{Rob}}$  by solving Program 2 for  $(\hat{A}, \hat{B}, D_0)$ . The SDP programs are solved using YALMIP (Löfberg, 2004) in conjunction with the solver SDPT3 (Tutuncu et al., 2003).

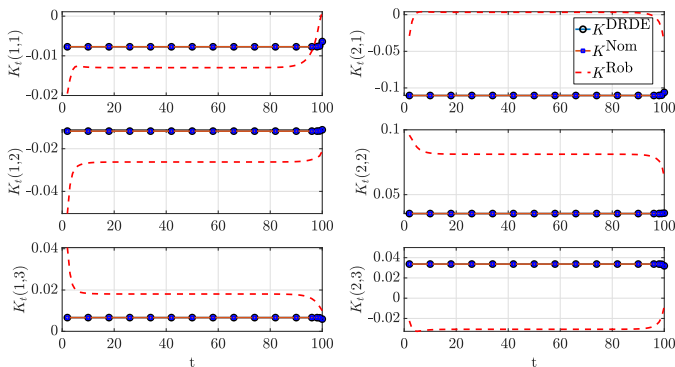


Fig. 2. Optimal controllers for the nominal and robust (fixed uncertainty) problem.

The first observation is that, as expected,  $K^{\text{DRDE}}$  coincides with  $K^{\text{Nom}}$ . Moreover  $J_{\text{DRDE}} \cong J_{\text{Nom}} \cong 80$ , where the former was obtained from the known closed form solution  $J_{\text{DRDE}} = \sum_{t=1}^{T-1} \mathbb{E} [w_t^\top X_{t+1} w_t]$  (where  $X_t$  is the stabilizing solution of the DRDE, Bertsekas (2005)), while  $J_{\text{Nom}}$  was directly provided by Program 1. As for the robust design, which achieved  $J_{\text{Rob}} \cong 240$ , note that  $K^{\text{Rob}}$  is generally far from the optimal controller for the nominal plant because of the requirement to guarantee robustness (at the expense of nominal performance).

The dual control policy, designed using Program 3, is illustrated in Figure 3 by comparing the state-feedback dual controller  $K^{\text{Dc}}$  with the nominal  $K^{\text{Nom}}$ , and also by reporting the covariance  $S_t$  of the excitation input. The *timestep* cost  $J_t^{\text{tot}}$ , together with its two contributions  $J_t^x = \mathbb{E} [x_t^\top (Q + K_t^\top R K_t) x_t]$  and  $J_t^e = \mathbb{E} [e_t^\top R e_t]$ , is finally shown in the bottom right plot.

There is a clear exploration action taking place in the first part of the finite horizon, performed only by the state-feedback  $K_t$ , while the covariance  $S_t$  is practically zero. This can also be appreciated from the plot with the costs where  $J_t^e \cong 0$ ,  $J_t^{\text{tot}} = J_t^x$ , and the latter has an initially increasing and later decreasing trend, before achieving a constant value. Indeed, since the cost would increase linearly in the optimal finite horizon problem, this can be read as a qualitative indication that, after approximately 20 timesteps,  $K_t$  has stopped exploring and is only devoted to (robust) exploitation.

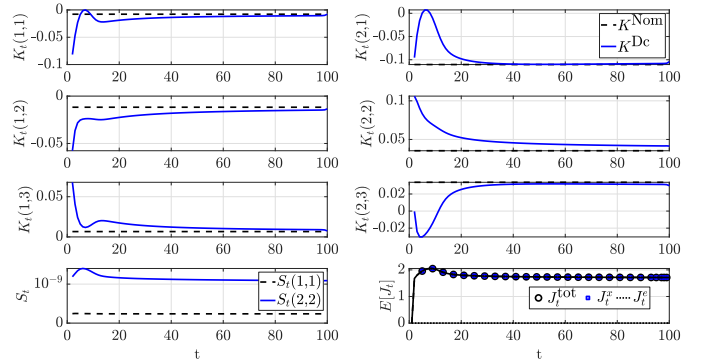


Fig. 3. Dual controller and cost for system (17).

In order to emphasize the *structured* property of the exploration actions, a different system is considered next:

$$A = \begin{bmatrix} 0.9 & 0.5 & 0 \\ 0 & 0.9 & 0.2 \\ 0 & -0.2 & 0.8 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & .1 \\ 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}. \quad (18)$$

Note that the system has now all its eigenvalues very close to the unit disk, and that the least damped mode is close to become uncontrollable.

Coarse-ID estimates of the system are obtained as for (17), and the dual control policy is shown in Figure 4.

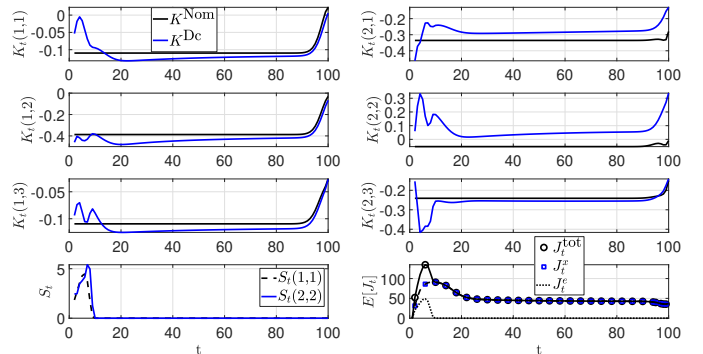


Fig. 4. Dual controller and cost for system (18).

Exploration actions can again be detected in the first part of the horizon, however this time they are performed by both the state-feedback  $K_t$  and the covariance  $S_t$ . Two types of trade-offs arising in the dual control problem can be appreciated by comparing the different trends in Figs. 3-4. The first is the one between exploration and exploitation, captured by the fact that the former only lasts for a certain fraction of the total mission. The second trade-off concerns the choice, for the purpose of exploration, between  $K_t$  and  $S_t$ . It is indeed observed, as expected, that whenever it is possible to explore in a *controlled* manner (i.e. without resorting to random excitation), this is the preferred way. The best exploration strategy inevitably depends on the true (unknown) plant to control, for example its controllability and margin of stability. Figs. 3-4 also show that, unlike the nominal case where there is no sensible variation of the optimized controller within the horizon (except for the very last timesteps, recall Fig. 2), the dual task for which the policy  $G_K(K_t, S_t)$  is designed makes the most important features of the problem (e.g.  $K_t$ ,  $S_t$ , timestep costs  $J_t^{\text{tot}}$ ) inherently time-varying and thus this type of dual control problem should be studied in a finite horizon setting.

## 5. CONCLUSIONS

The paper proposes a dual control synthesis approach for the finite horizon LQ problem. A control law is designed with the twofold objective of minimizing the worst-case quadratic cost in the face of an ellipsoidal uncertainty set while reducing it based on the system response. This is achieved by formulating an application-oriented, since the effect of the policy on the ellipsoidal set is captured in the optimization problem, and safe, since the designed controller is robust, exploration action. SDP programs to solve the nominal, robust (with fixed uncertainty) and dual control problems are proposed, and their application is shown. The resulting exploration encompasses different types of trade-offs and shows how the optimal actions depend on the features of the true plant.

## REFERENCES

- Annergren, M., Larsson, C.A., Hjalmarsson, H., Bombois, X., and Wahlberg, B. (2017). Application-oriented input design in system identification: Optimal input design for control. *IEEE Control Systems Magazine*, 37(2), 31–56.
- Åström, K. and Wittenmark, B. (1971). Problems of identification and control. *J. Math. Anal. Appl.*, 34(1), 90–113.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2), 235–256.
- Bertsekas, D.P. (2005). *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmont, MA, USA, 3rd edition.
- Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. (2017). Boltzmann exploration done right. In *Proceedings of NIPS*.
- Cohen, A., Hassidim, A., Koren, T., Lazić, N., Mansour, Y., and Talwar, K. (2018). Online linear quadratic control. In *35th ICML*.
- Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2019). On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*.
- Ferizbegovic, M., Umenberger, J., Hjalmarsson, H., and Schön, T.B. (2020). Learning Robust LQ-Controllers Using Application Oriented Exploration. *IEEE Control Systems Letters*, 4(1), 19–24.
- Garivier, A., Kaufmann, E., and Lattimore, T. (2016). On explore-then-commit strategies. In *Proceedings of NIPS*.
- Löfberg, J. (2004). YALMIP : A Toolbox for Modeling and Optimization in MATLAB. In *IEEE International Symposium on Computer-Aided Control System Design*.
- Luo, Z., Sturm, J., and Zhang, S. (2004). Multivariate nonnegative quadratic mappings. *SIAM Journal on Optimization*, 14(4), 1140–1162.
- Matni, N., Proutiere, A., Rantzer, A., and Tu, S. (2019). From self-tuning regulators to reinforcement learning and back again. In *arXiv*.
- Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1), 253–279.
- Scherer, C. and Weiland, S. (2000). *Linear Matrix Inequalities in Control*. Lecture Notes, Dutch Institute for Systems and Control.
- Sun, K. and Packard, A. (2005). Robust  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  Filters for Uncertain LFT Systems. *IEEE Transactions on Automatic Control*, 50, 715–720.
- Sutton, R.S. and Barto, A.G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition.
- Sznaier, M., Amishima, T., Parrilo, P., and Tierno, J. (2002). A convex approach to robust  $\mathcal{H}_2$  performance analysis. *Automatica*, 38, 957–966.
- Tutuncu, R., Toh, K., and Todd, M. (2003). Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95.
- Umenberger, J., Ferizbegovic, M., Schön, T.B., and Hjalmarsson, H. (2019). Robust exploration in linear quadratic reinforcement learning. In *Advances in NIPS*.

## APPENDIX

**Proof of Lemma 1.** Recall from Section 3 the definitions:  $P_t = \mathbb{E}[x_t x_t^\top]$ ,  $\bar{Q}_t := \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}} K_t \end{bmatrix}$ ,  $\bar{R} := \begin{bmatrix} 0 \\ R^{\frac{1}{2}} \end{bmatrix}$ . Define  $M_t = Q + K_t^\top R K_t$ . By virtue of the chosen policy (7),  $J$  can be rewritten as:

$$J = \mathbb{E} \left[ \sum_{t=1}^{T-1} (x_t^\top M_t x_t + e_t^\top R e_t) + x_T^\top Q x_T \right] \quad (19)$$

Consider the first term in the summation (i.e. the one that depends on  $x_t$ ). Simple matrix manipulations give it as:

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^{T-1} x_t^\top M_t x_t \right] &= \mathbb{E} [x^\top (I_{T-1} \otimes M_t) x] \\ &= \mathbb{E} [\text{Tr}(x x^\top I_{T-1} \otimes M_t)] = \text{Tr}(\mathbb{E}[x x^\top] I_{T-1} \otimes M_t) \end{aligned} \quad (20)$$

where  $x \in \mathbb{R}^{(T-1)n_x}$  denotes the vector obtained stacking the states  $x_t$ ,  $\mathcal{W} = \mathbb{E}[x x^\top] \in \mathbb{S}^{(T-1)n_x}$  denotes the covariance matrix of the state over the horizon and  $\otimes$  is the Kronecker product. It follows that:

$$\text{Tr}(\mathcal{W} I_{T-1} \otimes M_t) = \text{Tr} \left( (I_{T-1} \otimes \bar{Q}_t) \mathcal{W} (I_{T-1} \otimes \bar{Q}_t)^\top \right) \quad (21)$$

Due to the block diagonal structure of  $(I_{T-1} \otimes \bar{Q}_t)$ , it follows that:

$$\text{Tr} \left( (I_{T-1} \otimes \bar{Q}_t) \mathcal{W} (I_{T-1} \otimes \bar{Q}_t)^\top \right) = \text{Tr} \left( \sum_{t=1}^{T-1} (\bar{Q}_t P_t \bar{Q}_t^\top) \right) \quad (22)$$

The contribution to the cost only depends on the *diagonal* terms of  $\mathcal{W}$ , which are the covariance matrices  $P_t$  at the various timesteps:

$$\mathbb{E} \left[ \sum_{t=1}^{T-1} x_t^\top M_t x_t \right] = \text{Tr} \left( \sum_{t=1}^{T-1} (\bar{Q}_t P_t \bar{Q}_t^\top) \right) \quad (23)$$

The contribution of the state to the cost at  $t = T$  directly follows from (23) specializing it to the case when  $u$  is not penalized and thus  $\bar{Q}_T \equiv Q$ . Therefore:

$$\mathbb{E} [Q P_T Q^\top] = \text{Tr}(Q P_T Q^\top) \quad (24)$$

Finally, the proof for the term depending on  $e_t$  in the cost (19) follows along the same lines. This is further simplified by the fact that  $e_t$  are uncorrelated for different times, and thus their covariance matrix in the horizon has the block diagonal structure  $(I_{T-1} \otimes S_t)$ .