A multiobjective LQR synthesis approach to dual control for uncertain plants

Andrea Iannelli and Roy S. Smith

***Abstract*—The paper proposes a dual control finite horizon LQR synthesis procedure for unknown systems characterized by mean and covariance estimates. The optimized policy comprises time-varying state-feedback and dithering components, and the control problem is framed as a multiobjective synthesis which seeks a balance between exploitation and exploration costs. It is shown that classic experiment design problems can be recast in this framework by replacing the exploitation cost with an information reward. Numerical examples demonstrate the different dual control trade-offs on plants with different properties.**

***Index Terms*—Optimal control, Robust control, Identification for control, Uncertain systems, LMIs**

I. Introduction

DATA-driven methods have recently received great interest, owing to their potential to provide feedback laws without requiring accurate models of the plant. This is indeed a favourable feature in scenarios where systems are increasingly complex and data are seemingly unlimited. Using data to infer properties of dynamical systems requires, however, particular care. Informativity plays a crucial role [1], as the collected data should be useful (in a sense to be formally stated) for making decisions. In addition, data obtained from physical systems come with a cost, which encompasses different aspects from resource use to safety. Formulating systematic approaches to gather data such that these aspects are fulfilled is thus central to enable effective data-driven methods, and has been an active topic of research in different communities.

Taking the viewpoint of system identification, experiment design proposed strategies to account for informativity criteria in the selection of the input signal [2]. In application-oriented experiment design [3], the aim is specifically to design the least-costly experiment delivering a model that, when later used to synthesize a controller, provides satisfactory performance [4]. By simultaneously considering input design and performance aspects in the synthesis, the problem originally investigated by dual control is recovered [5]. When parametric models are used for identification, the information requirement can be enforced by minimizing a measure of the error associated with the estimator [6], e.g. for prediction error methods this can be approximated by the Fisher information matrix (FIM) [7]. An alternative approach used to promote informativity in dual controllers is to prescribe persistent excitation, which ensures identifiability and favours convergence of the identification algorithm [8].

In Reinforcement Learning (RL), this problem is framed as the search for a policy that balances exploration and exploitation [9], which are conceptually related to the dual control tasks of identifying the system while minimizing a predefined cost. The typical approach, e.g. Q-learning and actor-critic, is to calibrate exploration by optimizing regret bounds, which quantify the merits of a policy by comparing its running cost with that incurred by an expert baseline policy [10]. Examples are forced-exploration schemes [9] and optimism in the face of uncertainty [11].

Drawbacks of the previous approaches are that the exploratory actions are generically aimed at reducing the *amount* of uncertainty, are fixed a priori (e.g. using regrets to prescribe a rate of decay of random actions), and do not directly leverage the properties of the system. The concept of structured exploration, inspired by ideas initially presented in [12], was recently proposed in [13] to address some of these aspects. The problem is framed within the dual Linear Quadratic Regulator (LQR) setting, commonly used as a baseline for data-driven algorithms [14], [15], [10], [16], [11]. Starting from a Coarse-ID [15] description of the plant (i.e. nominal estimate of the state-space matrices and ellipsoidal confidence region with finite samples), a robust time-varying policy is designed for a finite horizon optimal control problem to simultaneously gather information of the plant and regulate it.

However, an important feature in [12], [13] is that the *worst-case* system's response is used to update the ellipsoidal uncertainty set. This means that, in practice, the uncertainty reduction for which the dual policy is optimized can only be achieved if the true system coincides with the worst-case one in the ellipsoidal set. To overcome this limitation, this work proposes a more realistic scenario where the *expected* system's response is used for the uncertainty set update. Two distinct costs are defined to formally state the problem. The first, namely the *exploration* cost, is the LQR cost incurred by the expected system, whose covariance is used for the uncertainty set update. The second, namely the *exploitation* cost, is the cost that, due to uncertainty in the plant, a robust policy would incur in operation. The latter cost is thus a functional of the uncertainty set, and measures the quality of the data that are generated by the dual policy. In the experiment design literature, this functional is commonly expressed as an *information* reward (e.g. related to the FIM), and it will be shown that the proposed framework can easily

This work is supported by the Swiss National Science Foundation under grant no. 200021_178890.

The authors are with the Department of Electrical Engineering, Automatic Control Lab, ETH, Zürich 8092, Switzerland iannelli/rsmith@control.ee.ethz.ch

accommodate this alternative choice. The dual controller thus results from the joint minimization of these two costs. This problem can be approached from a multiobjective control perspective [17], where the costs usually represent different systems' norms. Here both costs represent $\mathcal{H}_2$ norms, but have the different meaning discussed above. The time-varying policy is obtained by solving a Semidefinite program (SDP), whose structure is quite flexible and allows different definitions of the costs. One of the main benefit of the approach is that, by linking the policy with the reduction in the uncertainty in a finite horizon setting where transient effects are captured, trade-offs classically arising in the dual control problem could be optimized over. Numerical examples illustrate interesting features of the optimized policies, and show good results in simulation, also for off-nominal scenarios.

While the method proposed here can be sequentially applied to identify the system from measurements obtained with the dual policy, and then robustly control the uncertain system, future efforts will consider the problem of using the collected data to directly update the policy, in the spirit of data-driven methods [10], [16], [18], [14].

II. Problem statement

Consider the discrete linear time-invariant system:

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

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 a normally distributed process noise with zero mean and covariance $\sigma_w^2 I_{n_x}$. The objective is to optimally control the plant, according to a quadratic cost J introduced later, in a given finite horizon $t \in [1, T]$, $T \in \mathbb{N}$. The matrices A and B are unknown and estimated from measurements of x and u through the Coarse-ID approach [15]. Given a dataset $\mathcal{G} = \{(x_t, u_t) : 1 \leq t \leq N\}$, a nominal estimate is computed as:

$$(\hat{A}, \hat{B}) = \arg\min_{\bar{A},\bar{B}} \sum_{t=1}^{N-1} \left\| -x_{t+1} + \bar{A}x_t + \bar{B}u_t \right\|_2^2, \tag{2}$$

and the true dynamics is assumed to belong to the set:

$$\Omega(X, D_N) = \{X : X^\top D_N X \preceq I\},\ D_N \in \mathbb{S}^{n_x+n_u}, \tag{3a}$$

$$X = \begin{bmatrix} (\hat{A} - A)^\top \\ (\hat{B} - B)^\top \end{bmatrix}, \quad X \in \mathbb{R}^{(n_x+n_u)\times n_x}. \tag{3b}$$

Therefore, D_N defines an ellipsoidal uncertainty set which holds with probability 1-δ if chosen as:

$$D_N = \frac{1}{c_\chi \sigma_w^2} \sum_{t=1}^{N} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top, \quad c_\chi = \chi^2_{n_x^2+n_x n_u}(\delta), \tag{4}$$

where $\chi_n^2(\delta)$ is the critical value for a Chi-squared distribution with n degrees of freedom and probability level δ. See [15] for less conservative coefficients than c_χ. An interpretation of D_N is as the covariance of the posterior distribution of $\theta = \text{vec}([A\ B])$ given $\mathcal{G}$ [19], i.e.:

$$\theta \sim p(\theta|\mathcal{G}) = \mathcal{N}\left(\text{vec}([\hat{A}\ \hat{B}]), (c_\chi D_N)^{-1} \otimes I_{n_x}\right). \tag{5}$$

The probabilistic plant's description (5) is then replaced by the (high-probability) deterministic one (3a) by using the Chi-squared distribution.

We consider a time-varying policy $\pi(K_t, S_t)$:

$$u_t = K_t x_t + e_t, \qquad e_t \sim \mathcal{N}(0, S_t), \quad S_t \in \mathbb{S}^{n_u}. \tag{6}$$

This consists of a state-feedback part K_t and white noise (dithering) e_t, normally distributed with covariance S_t. Given cost matrices $Q \succeq 0$ and $R \succeq 0$, the expected finite horizon quadratic cost J in $[1, T]$ is defined as:

$$J = \mathbb{E}\left[\sum_{t=1}^{T-1} \left(x_t^\top Q x_t + u_t^\top R u_t\right) + x_T^\top Q x_T\right], \tag{7}$$

where the expectation is with respect to w and e. The following closed-loop costs can be defined from J:

$$\mathcal{J}_1^\pi(\hat{A}, \hat{B}) = \min_\pi J, \tag{8a}$$

$$\mathcal{J}_2^\pi(\hat{A}, \hat{B}, D_t) = \min_\pi \max_{(A,B)\in\Omega(X,D_t)} J. \tag{8b}$$

Eq. (8a) defines the optimal LQR cost for the expected plant, which is achieved by $\pi_{\text{DRDE}}(K_t^{\text{DRDE}}, 0)$ where K_t^{DRDE} is associated with the stabilizing solution of the Riccati difference equation (DRDE). Eq. (8b) defines the worst-case LQR cost among all the plants in $\Omega(X, D_t)$. Note that Ω can be a time-varying set if D_t changes within the horizon (as pointed out by the subscript).

It is important to recognize that, given a plant (the expected plant $(\hat{A}, \hat{B})$ will always be considered here) and an horizon $[1, T]$, there exists a mapping from a policy π to the uncertainty set D_t. Consider the time-varying policy $\pi(K_t, S_t)$ acting on the given plant through the dynamics (1). By recalling the definition of D_N in (3), the output of the aforementioned mapping at $t \in [1, T]$ is:

$$D_t = \frac{1}{c_\chi \sigma_w^2} \sum_{l=1}^{t} \begin{bmatrix} P_l & P_l K_l^\top \\ * & K_l P_l K_l^\top + S_l \end{bmatrix}, \tag{9}$$

where P_l denotes the state covariance matrix at timestep l, i.e. $P_l = \mathbb{E}\left[x_l x_l^\top\right] \in \mathbb{S}^{n_x}$. The mapping (9), which associates a policy $\pi(K_t, S_t)$ with an ellipsoid at each time inside the horizon, will be denoted by $\mathcal{D}(\pi) : T\ \mathbb{R}^{n_u \times n_x} \times T\ \mathbb{S}^{n_u} \to T\ \mathbb{S}^{n_x+n_u}$.

The objective of the work is to design a dual control policy π_{Dc} which balances exploration and exploitation in the LQR problem with uncertain plants. The problem is framed as the joint optimization of the exploration cost $\mathcal{J}_1^{\pi_1}$ and the exploitation cost $\mathcal{J}_2^{\pi_2}$, coupled via $\mathcal{D}(\pi_1)$. The cost $\mathcal{J}_1^{\pi_1}$ accounts for the exploration side of the problem, since the associated policy contributes to the reduction of the uncertainty. The cost $\mathcal{J}_2^{\pi_2}$ is a functional of the uncertainty set, and provides a characterization of the cost to pay for robust operation of the plant, thus accounts for the exploitation side. The following optimization problem can then be defined (the superscripts of the costs denote the associated optimized policy):

$$\min_{\pi_1,\pi_2} \quad \mathcal{J}_2^{\pi_2}(\hat{A}, \hat{B}, D_0 + \mathcal{D}(\pi_1)), \tag{10a}$$

$$\text{such that:} \quad \mathcal{J}_1^{\pi_1}(\hat{A}, \hat{B}) < \alpha \mathcal{J}_1^{\pi_{\text{DRDE}}}(\hat{A}, \hat{B}), \tag{10b}$$

where D_0 is the initial estimate's ellipse, $\alpha > 1$, and $\mathcal{J}_1^{\pi_{\mathrm{DRDE}}}$ is the cost associated with π_{DRDE}. The two costs are coupled via $\mathcal{D}(\pi_1)$, which maps the action of π_1 into the uncertainty set to which $\mathcal{J}_2^{\pi_2}$ must be robust. The solution of the multiobjective synthesis problem in Eq. (10) provides a dual policy $\pi_{\mathrm{Dc}} \equiv \pi_1$ which balances exploration of the expected system (according to $\mathcal{J}_1^{\pi_1}$), while targeting future exploitation by minimizing the worst-case cost $\mathcal{J}_2^{\pi_2}$ incurred to robustly operate the plant. The trade-off between the costs is condensed in the parameter α, which accounts for the multiobjective problem by characterizing the sublevel set of policies π_1 with respect to the cost achieved by the greediest one, i.e. π_{DRDE}. It is noted that $\pi_2(K_t^2, 0)$ does not represent the robust policy redesigned with the new uncertain model, as this would be computed a-posteriori by solving problem (8b) with the new uncertainty set, and is only a fictitious policy introduced to allow the definition of the worst-case cost $\mathcal{J}_2^{\pi_2}$ as a functional of the uncertainty. While inspired by their application-oriented strategy, the problem in (10) is distinct from those solved in [12], [13]. Differences, which force the problem to be solved with a different approach, include: the finite horizon setting (also used in [13]); the multiobjective formulation; and the use of the expected system response to update the uncertainty. The importance of the latter feature, which is deemed a more realistic working assumption, will be evident in Section IV, where simulation results show that the predicted uncertainty reduction is effectively achieved in practice, and at the same time the dual policy exhibits a certain degree of robustness to mismatches between the true and expected plants.

III. Convex optimization formulation

A. Multiobjective synthesis approach

The starting point is Lemma 1 from [13].

Lemma 1: Define $\bar{Q}_t := \begin{bmatrix} Q^{\frac{1}{2}} \\ R^{\frac{1}{2}} K_t \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times n_x}$, and $\bar{R} := \begin{bmatrix} 0 \\ R^{\frac{1}{2}} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times n_x}$. The cost J defined in Eq. (7) is equivalent to:

$$J = \mathrm{Tr}\left(\sum_{t=1}^{T-1}\left(\bar{Q}_t P_t \bar{Q}_t^\top + \bar{R}_t S_t \bar{R}_t^\top\right) + Q P_T Q^\top\right). \tag{11}$$

Lemma 1 gives an expression of J which depends separately on P_t and S_t (via a summation of terms). This is similar to the infinite horizon case (where only one term would feature), for which an SDP formulation exists [17]. From this premise, in [13] the two separate synthesis problems in Eq. (8) were solved by using standard Linear Matrix Inequalities (LMIs) for the $\mathcal{H}_2$ norm and, in the case of (8b), combining them with a matrix version of the S-procedure [20] to enforce the robust constraint. The SDP formulation for problem (8a) is given below as it will be useful for later discussion. Its solution is the optimal LQR cost $\mathcal{J}_1^{\pi_{\mathrm{DRDE}}}$ appearing in Eq. (10b), and the optimizer is the Riccati controller $\pi_{\mathrm{DRDE}}(K_t^{\mathrm{DRDE}}, 0)$.

Program 1: SDP for $\mathcal{J}_1^{\pi}$ (8a).

$$\min_{Y_t, P_t, Z_t, S_t} \mathrm{Tr}\left(\sum_{t=1}^{T-1} Y_t + Q P_T Q^\top\right) = \mathcal{J}_1^{\pi}, \tag{12a}$$

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

$$\begin{bmatrix} P_t & P_t \hat{A}^\top + Z_t \hat{B}^\top \\ * & P_{t+1} - \sigma_w^2 I - \hat{B} S_t \hat{B}^\top \end{bmatrix} \succeq 0, \tag{12c}$$

$$Y_t \succeq 0, S_t \succeq 0, P_1 \succeq 0, P_{t+1} \succeq 0,\ \forall t \in [1, T-1], \tag{12d}$$

where $Z_t := P_t K_t^\top$.

The problem in Eq. (10) differs from those in (8) because of the multiobjective feature, i.e. $\mathcal{J}_1^{\pi_1}$ and $\mathcal{J}_2^{\pi_2}$ are jointly optimized. A first important aspect concerns the state covariance update for the expected plant, which, for a policy $\pi(K_t, S_t)$, must follow the Stein equation [21]:

$$P_{t+1} = (\hat{A} + \hat{B} K_t) P_t (\hat{A} + \hat{B} K_t)^\top + \sigma_w^2 I + \hat{B} S_t \hat{B}^\top. \tag{13}$$

This is typically enforced with the LMI (12c), which however only guarantees (by Schur complement) that P_{t+1} is greater than the right-hand side of (13), i.e. it gives a lower bound for the updated covariance. In problem (10), the state covariances also influences, via the mapping $\mathcal{D}$, the cost $\mathcal{J}_2^{\pi}$, thus a lower bound on P_{t+1} does not generally guarantee that (13) holds. To this end, the following upper bound LMI at each t is proposed:

$$U_t(P_t, Z_t, S_t, \bar{V}_t) + U_t^\top(P_t, Z_t, S_t, \bar{V}_t) \preceq 0, \tag{14}$$

$$U_t(P_t, Z_t, S_t, \bar{V}_t) = \begin{bmatrix} -P_t & -F_t \\ F_t^\top & P_{t+1} - \sigma_w^2 I - \hat{B} S_t \hat{B}^\top - \bar{V}_t \end{bmatrix},$$

$$F_t = P_t \hat{A}^\top + Z_t \hat{B}^\top,$$

where $\bar{V}_t \in \mathbb{S}^{n_x}$ are given positive definite matrices. By Schur complement, LMI (14) implies that at each t:

$$P_{t+1} \preceq (\hat{A}+\hat{B}K_t) P_t (\hat{A}+\hat{B}K_t)^\top + \sigma_w^2 I + \hat{B} S_t \hat{B}^\top + \bar{V}_t. \tag{15}$$

Clearly, the quality of the upper bound depends on $\bar{V}_t$. The following Program provides tight matrices $\bar{V}_t$ for the upper bound LMI (14).

Program 2: SDP for the matrices $\bar{V}_t$ in (14).

$$\min_{V_t} \quad \mathrm{Tr}\left(\sum_{t=1}^{T-1} V_t\right),$$
$$U_t(P_t^{\mathrm{DRDE}}, Z_t^{\mathrm{DRDE}}, 0, V_t) + U_t(P_t^{\mathrm{DRDE}}, Z_t^{\mathrm{DRDE}}, 0, V_t)^\top \preceq 0,$$
$$V_t \succeq 0, \qquad \forall t \in [1, T-1].$$

The optimizers of Program 2 are the tightest (as measured by their traces) matrices $\bar{V}_t$ for which the upper bound LMI constraint (14) is satisfied by the solution of Program 1, for which Eq. (13) is always guaranteed to hold. This provides a systematic and effective way of pre-computing the matrices $\bar{V}_t$ to use in constraint (14).

The other instrumental step consists of formulating a convex relaxation of the bilinear term in the lower diagonal block of the mapping $\mathcal{D}$ (9):

$$K_l P_l K_l^\top = Z_l^\top P_l^{-1} Z_l \succeq Z_l^\top \bar{K}_l^\top + \bar{K}_l Z_l - \bar{K}_l P \bar{K}_l^\top, \quad (17)$$

This bound, derived for the static case in ([12], Lemma 1), is tight when $\bar{K}_l = K_l$. Here it will be used $\bar{K}_l = K_l^{\text{DRDE}}$. Eq. (17) allows D_t to be lower bounded by:

$$D_t \succeq \hat{D}_t = \frac{1}{c_\chi \sigma_w^2} \sum_{l=1}^{t} \begin{bmatrix} P_l & Z_l \\ * & Z_l^\top \bar{K}_l^\top + \bar{K}_l Z_l - \bar{K}_l P_l \bar{K}_l^\top + S_l \end{bmatrix} \quad (18)$$

This lower bound is important since by using $\hat{D}_t$ in the place of D_t uncertainty is underestimated in the optimization, i.e. in practice its reduction will be larger.

Assume now that from prior knowledge, or an experimental dataset $\mathcal{G}$, initial estimates for $\hat{A}$, $\hat{B}$, and D_0 are available such that $(A, B) \in \Omega(X, D_0)$. The dual control synthesis problem, stated in Eq. (10), is solved with Program 3. The superscripts 2 and 1 of the optimization variables emphasize their connection with $\mathcal{J}_2^{\pi_2}$ and $\mathcal{J}_1^{\pi_1}$, respectively.

Program 3: SDP for the dual control problem (10).

$$\min_{Y_t^2, P_t^2, Z_t^2, p_t, Y_t^1, P_t^1, Z_t^1, S_t^1} \quad \operatorname{Tr}\left(\sum_{t=1}^{T-1} Y_t^2 + Q P_T^2 Q^\top \right) = \mathcal{J}_2^{\pi_2}, \quad (19a)$$

$$\begin{bmatrix} Y_t^2 & \begin{bmatrix} Q^{\frac{1}{2}} P_t^2 \\ R^{\frac{1}{2}} Z_t^{2^\top} \end{bmatrix} \\ * & P_t^2 \end{bmatrix} \succeq 0, \quad (19b)$$

$$\begin{bmatrix} P_t^2 & -\begin{bmatrix} P_t^2 & Z_t^2 \end{bmatrix} \begin{bmatrix} \hat{A}^\top \\ \hat{B}^\top \end{bmatrix} & \begin{bmatrix} P_t^2 & Z_t^2 \end{bmatrix} \\ * & P_{t+1}^2 - \sigma_w^2 I - p_t I & 0 \\ * & * & p_t (D_0 + \hat{D}_t(P^1, Z^1, S^1)) \end{bmatrix} \succeq 0, \quad (19c)$$

$$Y_t^2 \succeq 0, P_1^2 \succeq 0, P_{t+1}^2 \succeq 0, p_t \geq 0, \quad (19d)$$

$$\operatorname{Tr}\left(\sum_{t=1}^{T-1} Y_t^1 + Q P_T^1 Q^\top \right) = \mathcal{J}_1^{\pi_1} \leq \alpha \mathcal{J}_1^{\pi_{\text{DRDE}}}(\hat{A}, \hat{B}), \quad (19e)$$

$$\begin{bmatrix} Y_t^1 - \bar{R} S_t^1 \bar{R}^\top & \begin{bmatrix} Q^{\frac{1}{2}} P_t^1 \\ R^{\frac{1}{2}} Z_t^{1^\top} \end{bmatrix} \\ * & P_t^1 \end{bmatrix} \succeq 0, \quad (19f)$$

$$\begin{bmatrix} P_t^1 & P_t^1 \hat{A}^\top + Z_t^1 \hat{B}^\top \\ * & P_{t+1}^1 - \sigma_w^2 I - \hat{B} S_t^1 \hat{B}^\top \end{bmatrix} \succeq 0, \quad (19g)$$

$$U_t(P_t^1, Z_t^1, S_t^1, \bar{V}_t) + U_t^\top (P_t^1, Z_t^1, S_t^1, \bar{V}_t) \preceq 0, \quad (19h)$$

$$Y_t^1 \succeq 0, S_t^1 \succeq 0, P_1^1 \succeq 0, P_{t+1}^1 \succeq 0, \quad \forall t \in [1, T-1]. \quad (19i)$$

where $Y_t^2 \in \mathbb{S}^{n_u+n_x}, P_t^2 \in \mathbb{S}^{n_x}, Z_t^2 \in \mathbb{R}^{n_x \times n_u}, p_t, Y_t^1 \in \mathbb{S}^{n_u+n_x}, P_t^1 \in \mathbb{S}^{n_x}, Z_t^1 \in \mathbb{R}^{n_x \times n_u}, S_t^{1 \in \mathbb{S}^{n_u}}$ and the total number of decision variables is $T(n_x^2 + \frac{3}{2} n_u^2 + 4 n_x n_u + 2 n_x + \frac{3}{2} n_u)$. LMIs (19b)-(19c)-(19d) provide the constraints required to express $\mathcal{J}_2^{\pi_2}$ by the linear cost in Eq. (19a). This is achieved by first writing the true system matrices A and B as a function of X, $\hat{A}$, and $\hat{B}$ (3b). The S-lemma (with multipliers p_t) [20] is then employed to transform the conditions in Program 1, valid for one specific plant, into a robust optimization problem which holds for all the plants in $\Omega(X, D_0 + \hat{D}_t)$. A line search on p_t is used to overcome the bilinearity between p_t and $\hat{D}_t$. LMIs (19f)-(19g)-(19h)-(19i) provide the constraints required to enforce that $\mathcal{J}_1^{\pi_1}$ remains smaller than $\alpha \mathcal{J}_1^{\pi_{\text{DRDE}}}$. Additional convex constraints could also be appended (e.g. norm bounds on S_t^1 or P_t^1). The coupling between the two problems is given by $\hat{D}_t$, which, as highlighted by its argument in (19c), depends on the variables with superscript 1. The sought time-varying dual policy is $\pi_{\text{Dc}}(K_t^{\text{Dc}}, S_t^{\text{Dc}}) \equiv \pi_1(\bar{Z}_t^{1^\top} \bar{P}_t^{1^{-1}}, \bar{S}_t^1)$. Both its state-feedback (which has a dual purpose) and dithering (whose purpose is purely exploration) components are optimized in Program 3, and thus will be computed according to the defined multiobjective criterion (10).

B. Interpretation and connections with input design

By using π_{Dc} to control system (1) in the finite horizon $[1, T]$, the information gathered in expectation from the system's response is such that the worst-case LQR cost is minimized. The expectation is with respect to w, e, and the true dynamics, that at the beginning of the experiment is inside $\Omega(X, D_0)$, and for which the mean $(\hat{A}, \hat{B})$ is used. This choice is supported by recent studies on the favourable sub-optimality gap properties of the certainty equivalence assumption [22].

The policy can thus be seen as the solution of an input design problem where the goal is to minimize the uncertainty for a subsequent robust design [2], [3]. In fact, Program 3 can easily accommodate experiment design objectives by replacing the exploitation cost $\mathcal{J}_2^{\pi_2}$ with an information reward. Since the ellipsoidal matrix D_t represents the system's matrices error covariance at time t (recall Eq. 5), it can be interpreted as the FIM for the Coarse-ID setting. The following costs, all convex in the optimization variables, could then be considered [23]:

- D-optimality: $\mathcal{J}_{\text{FIM}}(\hat{D}_t) = -\sum_{t=1}^{T} \left(\det \hat{D}_t \right)^{\frac{1}{n_x+n_u}}$,
- E-optimality: $\mathcal{J}_{\text{FIM}}(\hat{D}_t) = -\sum_{t=1}^{T} \min \ \operatorname{eig}(\hat{D}_t)$,
- A-optimality: $\mathcal{J}_{\text{FIM}}(\hat{D}_t) = -\sum_{t=1}^{T} \operatorname{Tr}(\hat{D}_t)^{-1}$,

leading to different dual control synthesis problems.

Program 4: SDP for dual control with FIM objective.

$$\min_{Y_t^1, P_t^1, Z_t^1, S_t^1} \quad \mathcal{J}_{\text{FIM}}(\hat{D}_t(P_t^1, Z_t^1, S_t^1)), \quad (20a)$$

$$\operatorname{Tr}\left(\sum_{t=1}^{T-1} Y_t^1 + Q P_T^1 Q^\top \right) = \mathcal{J}_1^{\pi_1} \leq \alpha \mathcal{J}_1^{\pi_{\text{DRDE}}}(\hat{A}, \hat{B}), \quad (20b)$$

$$\text{LMI (19f)}, \text{LMI (19g)}, \text{LMI (19h)}, \text{LMI (19i)}. \quad (20c)$$

Program 4 presents the same LMIs constraints needed for the exploration cost $\mathcal{J}_1^{\pi_1}$, and replaces $\mathcal{J}_2^{\pi_2}$ with the information objective $\mathcal{J}_{\text{FIM}}$.

It is noted that $\mathcal{J}_2^{\pi_2}$ and $\mathcal{J}_{\text{FIM}}$ consist of a summation of terms from each time step inside the horizon. If this method was used in conjunction with an *explore-then-commit* strategy [24], whereby data, once the horizon is concluded, are used to update the system's estimate, then these exploitation costs could be defined using only the contribution at the final time T. However, in future applications of this method we envisage updating the dual policy π_{Dc} on-line, enabling in this way a data-driven method which takes informativity into account. For this, it is paramount that exploratory actions are performed throughout the horizon, and the present definition of $\mathcal{J}_2^{\pi_2}$ capture these fundamental transient features.

IV. Illustrative examples

Consider the two plants:

$$A_1 = \begin{bmatrix} 0.9 & 0.5 & 0 \\ 0 & 0.9 & 0.2 \\ 0 & -0.2 & 0.8 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & .1 \\ 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \tag{21a}$$

$$A_2 = \begin{bmatrix} 0.1 & 2 & 2 \\ 0 & 0.1 & 2 \\ 0 & 0 & 0.1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & .6 \\ 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \tag{21b}$$

with cost matrices $Q = I_{n_x}$ and R=blkdiag(10,1), and σ_w^2=0.5, T=100, δ=0.05. Plant 1 is lightly damped and the least damped mode is close to uncontrollable. Conversely, the upper triangular structure in Plant 2 shows that it is highly damped and controllable. For each plant an estimate $(\hat{A}, \hat{B}, D_0)$ is obtained by simulating one trajectory of length N=100 with $u_t \sim \mathcal{N}(0, \sigma_u^2 I_{n_u})$, $\sigma_u^2 = 3$. The SDPs are solved with MOSEK [25].

Figure 1 shows a comparison between the costs achieved by the policies π_1 obtained with Program 3 (Dc) and Program 4 (FIM criteria).

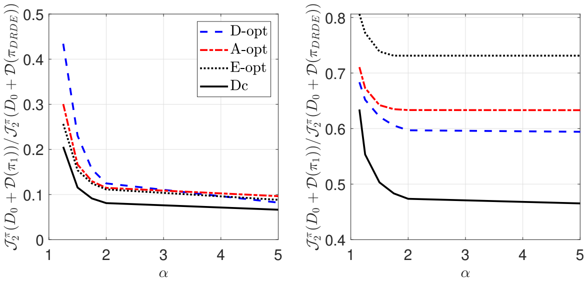

Fig. 1: Comparison of costs for plant 1 (left) and 2 (right).

For each value of the parameter α, the corresponding exploitation cost $\mathcal{J}_2^{\pi}(\hat{A}, \hat{B}, D_0 + \mathcal{D}(\pi_1))$ divided by $\mathcal{J}_2^{\pi}(\hat{A}, \hat{B}, D_0 + \mathcal{D}(\pi_{\mathrm{DRDE}}))$ is plotted. This ratio indicates the cost incurred by a robust design using the information extracted by π_1 compared to a design based on the information obtained by the optimal LQR controller. For a given ellipsoidal set, cost $\mathcal{J}_2^{\pi}$ (8b) is computed with the SDP formulation from [13]. While it is expected that the uncertainty reduction achieved with π_{Dc} is associated with the smallest $\mathcal{J}_2^{\pi}$ since this is the objective in (10), it is interesting to observe that the gap between the curves is more pronounced when a small exploratory budget is available and depends on the unknown plant.

In Fig. 2 the policies π_{Dc} for the two plants (*Dc*-solid line) are compared with those obtained with the dual control algorithm from [13] (*Wc*-dashed line), which uses the worst-case state covariance to update the system's response, by showing the feedback gains (for reference, K_t^{DRDE} is also reported with dotted lines) and the dithering covariance S_t. In the bottom plots, a break down of $\mathcal{J}_1^{\pi_{\mathrm{Dc}}}$ in terms of $J_t^x = \mathbb{E}\left[x_t^\top Q x_t\right]$, $J_t^u = \mathbb{E}\left[x_t^\top K_t^\top R K_t x_t\right]$ and $J_t^e = \mathbb{E}\left[e_t^\top R e_t\right]$ is displayed. The total number of decision variables in Program 3 for this case is 4800 and the run time for a fixed value of p is 29s.

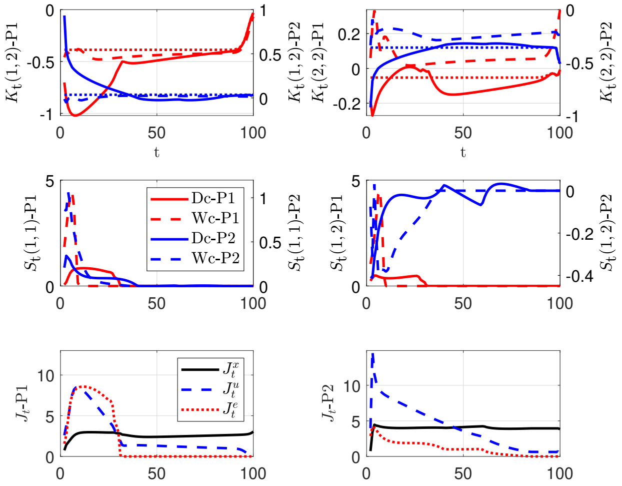

Fig. 2: Policies and costs for the two plants (α=2).

The comparison with *Wc* reveals that π_{Dc} uses larger state-feedback exploratory actions (since it can only leverage the expected state covariance) and is able to use smaller dithering actions. The comparison of π_{Dc} between the two plants shows that in Plant 1 state-feedback and dithering are both used to excite the system. Conversely, for Plant 2 almost only state-feedback is used, and for a longer time (see also the cost breakdown). This can be interpreted observing its higher controllability, which favours exploratory actions carried out using feedback. Interestingly, the importance of controllability properties for informative experiments was analytically found in [15] with respect to the spectral norm of Coarse-ID estimates.

The results obtained by drawing 500 samples of w_t and e_t from their distributions and simulating the expected estimate $(\hat{A}, \hat{B})$ of Plant 1 under the dual policy are shown in the first two plots of Fig. 3. The averaged quantities of interest (*Empirical*) are compared with those obtained from Program 3 (*Optimized*). The upper right box shows a perfect match of the cost $\mathcal{J}_1$ incurred throughout the horizon. This confirms that the dynamic of the state-covariance in simulation is the same as the optimized one, i.e. the upper bound (18) on P_{t+1} correctly enforces the state-covariance update. The singular values (sv) of the matrix D_t also reveal interesting features. The top left box confirms that the ellipse computed from data is larger, but with a very small gap, than the optimized one (i.e. Eq. 9 is a tight lower bound). Moreover, the comparison of sv in the bottom left plot shows that the ellipse associated with the D-optimality is smaller across the horizon. Nonetheless, as seen in Fig. 1, it results in a 50% higher exploitation cost. This points out the key difference between the proposed dual policy, which excites the system so that the most valuable information for the intended application of the model is extracted, and the experiment design policies, where only the size of the uncertainty (defined according to a certain geometric criterion) is targeted. The bottom right plot of Fig. 3 investigates the robustness of π_{Dc} by analyzing its action on 2000 randomly drawn plants in the initial Coarse-ID set $\Omega(X, D_0)$ (3) and showing the ratios between the resulting expected exploration ($\mathcal{J}_1$, on the left) and exploitation ($\mathcal{J}_2$, on the right) costs and the corresponding costs for the nominal plant considered in the previous analyses (and for which π_{Dc} was designed). The covariance σ_u^2 of the white-input signal employed in the Coarse-ID is a measure of the size of the set Ω (note that $\frac{\sigma_u}{\sigma_w}$ can be interpreted as signal-to-noise ratio of the experiment). The results show that the policy allows efficient exploration also in off-nominal conditions (the ratio $\mathcal{J}_2$ is always smaller than 1), at the price of a higher exploration cost (the ratio $\mathcal{J}_1$ is always greater than 1). This aspect can be mitigated by refining the Coarse-ID estimate (i.e. by increasing σ_u), and points out a potential advantage of employing this scheme in an on-line setting where the dual policy can be updated.

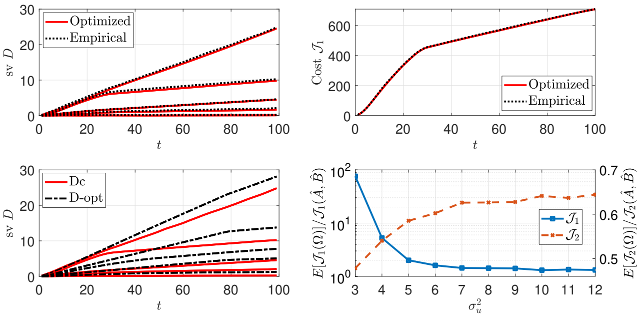

Fig. 3: Simulated results for Plant 1 (α=2).

V. Conclusions

A convex program to synthesize dual controllers for the finite horizon Linear Quadratic Regulator is proposed. The problem is formulated as the joint minimization of two costs corresponding to exploration and exploitation objectives. The designed policy excites the system in such a way that estimates computed from the response of the expected plant in the initial uncertainty set lead to the smallest worst-case cost. Results show that the dual policy captures dual control trade-offs including duration of the exploratory actions and their distribution between state-feedback and dithering components.

Acknowledgements

The authors would like to thank Nikolai Matni for suggestions and discussions on the topic.

References

[1] H. J. Van Waarde, J. Eising, H. L. Trentelman, and M. K. Camlibel, "Data informativity: a new perspective on data-driven analysis and control," *IEEE Transactions on Automatic Control*, vol. 54, pp. 2828–2840, 2020.

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

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

[4] C. R. Rojas, J. C. Agüero, J. S. Welsh, , and G. C. Goodwin, "On the equivalence of least costly and traditional experiment design for control," *Automatica*, vol. 47, pp. 1938–1948, 2008.

[5] K. Åström and B. Wittenmark, "Problems of identification and control," *J. Math. Anal. Appl.*, vol. 34, pp. 90–113, 1971.

[6] J. Rathouský and V. Havlena, "MPC-based approximate dual controller by information matrix maximization," *Int. J. Adapt. Control Signal Process*, vol. 27, pp. 974–999, 2013.

[7] M. Gevers, A. Sanfelice Bazanella, X. Bombois, and L. Miskovic, "Identification and the information matrix: How to get just sufficiently rich?" *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2828 – 2840, 2009.

[8] G. Marafioti, R. R. Bitmead, and M. Hovd, "Persistently exciting model predictive control," *International Journal of Adaptive Control and Signal Processing*, vol. 28, no. 6, pp. 536–552, 2014.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[10] N. Matni, A. Proutiere, A. Rantzer, and S. Tu, "From self-tuning regulators to reinforcement learning and back again," in *arXiv*, 2019.

[11] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. of the Annual Conference on Learning Theory*, 2011.

[12] M. Ferizbegovic, J. Umenberger, H. Hjalmarsson, and T. B. Schön, "Learning Robust LQ-Controllers Using Application Oriented Exploration," *IEEE Control Systems Letters*, vol. 4, pp. 19–24, 2020.

[13] A. Iannelli, M. Khosravi, and R. S. Smith, "Structured exploration in the finite horizon linear quadratic dual control problem," in *arXiv:1910.14492*, 2019.

[14] A. Cohen, A. Hassidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar, "Online linear quadratic control," in *35th International Conference on Machine Learning*, 2018.

[15] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, Aug 2019.

[16] G. R. Gonçalves da Silva, A. S. Bazanella, C. Lorenzini, and L. Campestrini, "Data-Driven LQR Control Design," *IEEE Control Systems Letters*, vol. 3, pp. 180–185, 2019.

[17] C. Scherer and S. Weiland, *Linear Matrix Inequalities in Control*. Lecture Notes, 2000.

[18] H. Hjalmarsson, M. Gevers, S. Gunnarsson, and O. Lequin, "Iterative feedback tuning: theory and applications," *IEEE Control Systems Magazine*, vol. 18, no. 4, pp. 26–41, 1998.

[19] J. Umenberger, M. Ferizbegovic, T. B. Schön, and H. Hjalmarsson, "Robust exploration in linear quadratic reinforcement learning," in *Advances in NIPS*, 2019.

[20] Z. Luo, J. Sturm, and S. Zhang, "Multivariate nonnegative quadratic mappings," *SIAM Journal on Optimization*, vol. 14, pp. 1140–1162, 2004.

[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA, USA: Athena Scientific, 2005, vol. I.

[22] H. Mania, S. Tu, and B. Recht, "Certainty Equivalence is Efficient for Linear Quadratic Control," in *Advances in NIPS*, 2019.

[23] I. Manchester, "Input Design for System Identification via Convex Relaxation," in *IEEE CDC*, 2010.

[24] A. Garivier, E. Kaufmann, and T. Lattimore, "On explore-then-commit strategies," in *Advances in NIPS*, 2016.

[25] *The MOSEK Optimization Toolbox for MATLAB manual (Release 9.0.104)*, MOSEK ApS, 2019.