# Optimistic robust linear quadratic dual control

**Jack Umenberger** UMNBRGR@MIT.EDU
*CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139*

**Thomas B. Schön** THOMAS.SCHON@IT.UU.SE
*Department of Information Technology, Uppsala University, Uppsala, 75236, Sweden*

**Editors:** A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

## Abstract

Recent work by Mania et al. (2019) has proved that certainty equivalent control achieves nearly optimal regret for linear systems with quadratic costs. However, when parameter uncertainty is large, certainty equivalence cannot be relied upon to stabilize the true, unknown system. In this paper, we present a dual control strategy that attempts to combine the performance of certainty equivalence, with the practical utility of robustness. The formulation preserves structure in the representation of parametric uncertainty, which allows the controller to target reduction of uncertainty in the parameters that 'matter most' for the control task, while robustly stabilizing the uncertain system. Control synthesis proceeds via convex optimization, and the method is illustrated on a numerical example.

**Keywords:** Dual control, linear systems, convex optimization

## 1. Introduction

Since the initial formulation of the 'dual control' problem by Feldbaum (1960) in the 1960s, learning to make decisions in uncertain and dynamic environments has remained a topic of sustained research activity. However, recent years have witnessed a resurgence of interest in such problems, inspired perhaps in part by the dramatic success of reinforcement learning, cf. Mnih et al. (2015); Silver et al. (2016). Specifically, linear systems with quadratic costs, a.k.a. 'the linear quadratic regulator', have been the subject of intense recent study, cf. Matni et al. (2019). Such research typically focuses on two main aspects: i) performance, usually measured in terms of bounds on regret, and ii) robustness, i.e., stability of the closed-loop system, which is often important in practical applications. Concerning the former, the work of Mania et al. (2019) has proved that 'certainty equivalent' (CE) control (nearly) achieves the optimal regret bound; provided that this controller stabilizes the system, which is the case when parameter uncertainty is sufficiently small. Inspired by this result, the present paper attempts to combine the performance of certainty equivalence with the practical advantages of robustness. Specifically, we propose a dual control strategy, for linear systems with quadratic costs, that optimizes for performance of the nominal, i.e., most likely, system (as in CE control), while robustly stabilizing the system in the presence of parametric uncertainty. The dual controller performs 'targeted exploration', attempting to reduce uncertainty in the parameters that 'matter most' for control, while balancing the exploration-exploitation tradeoff.

The contributions of this paper are twofold. In §3.2, we present a convex formulation of optimization of quadratic cost, for a nominal linear system, subject to robust stability guarantees under parametric uncertainty. This extends the existing system level synthesis (SLS) framework, cf. Wang et al. (2019b), by preserving structure in the representation of system uncertainty. In §3.3, we build upon this formulation to present an (approximate) dual control strategy, exploiting the preservation of structure to perform exploration that targets uncertainty reduction in the specific parameters that are 'preventing' certainty equivalent control from stabilizing the uncertain true system.

**Related work** Of greatest relevance to the present paper is the work of Mania et al. (2019), which proves that certainty equivalence, i.e., estimating model parameters via online least squares and then applying LQR, achieves (nearly optimal) $\tilde{\mathcal{O}}(\sqrt{T})$ regret. This result holds when the parameter error is sufficiently small so as to ensure closed-loop stability of the true system, which does not always hold in practice, e.g., Dean et al. (2017). Many recent works have addressed the issue of robustness in adaptive control, cf. Dean et al. (2017, 2018, 2019); Cohen et al. (2018). The work of Umenberger et al. (2019), cf. also Ferizbegovic et al. (2019); Iannelli et al. (2019), attempts to do 'targeted-exploration' by prioritizing uncertainty reduction in the system parameters to which performance is most sensitive. These methods consider worst-case costs to bound performance on the true system, and as such, can be conservative in practice. It is the ambition of this paper to combine the benefits of 'targeted exploration' with a more 'optimistic' CE strategy, that optimizes for performance of the nominal, rather than worst-case, system. Other recent work on adaptive linear quadratic control includes Thompson sampling (e.g. Ouyang et al. (2017); Abeille and Lazaric (2017, 2018)), model-free (e.g. Fazel et al. (2018); Malik et al. (2018)) and partially model-free (e.g. Agarwal et al. (2019a,b)) methods, as well as the 'optimism in the face of uncertainty' heuristic (e.g. Abbasi-Yadkori and Szepesvári (2011); Ibrahimi et al. (2012); Faradonbeh et al. (2019)).

## 2. Problem statement

In this section we describe in detail the problem addressed in this paper. Notation is largely standard. $\otimes$ denotes the Kronecker product. $\mathbb{S}^n$ denotes the space of $n \times n$ symmetric matrices. w.p. means 'with probability'. $\chi^2_n(p)$ denotes the value of the Chi-squared distribution with $n$ degrees of freedom and probability $p$. The space of real, proper (strictly proper) transfer matrices is denoted $\mathcal{RH}_\infty$ ($\frac{1}{z}\mathcal{RH}_\infty$). With some abuse of notation, $[t_1, t_2]$ for $t_1, t_2 \in \mathbb{N}$ denotes $\{t_1, \dots, t_2\}$.

**Dynamics and modeling** We are concerned with control of linear time-invariant systems

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

where $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$ and $w_t \in \mathbb{R}^n$ denote the state (which is assumed to be directly measurable), input and process noise, respectively, at time $t$. We assume that the true parameters $\{A_{\text{tr}}, B_{\text{tr}}\}$ are unknown; as such, all knowledge about the true system dynamics must be inferred from observed data, $\mathcal{D}_n := \{x_t, u_t\}_{t=1}^n$. We assume that $\sigma_w$ is known, or has been estimated, and that we have access to initial data, denoted (with slight notational abuse) $\mathcal{D}_0$, obtained, e.g. during a preliminary experiment. Given data $\mathcal{D}_n$ we define a model $\mathcal{M}_\delta(\mathcal{D}_n) = \{\hat{A}, \hat{B}, D\}$, where $(\hat{A}, \hat{B}) := \arg\min_{A,B} \sum_{t=1}^{n-1} |x_{t+1} - Ax_t - Bu_t|^2$ denote *nominal* parameters given by the ordinary least squares estimates of $(A_{\text{tr}}, B_{\text{tr}})$, and $D \in \mathbb{S}^{n_x+n_u}$ is a matrix that quantifies the uncertainty in our nominal parameter estimate. Specifically, given a user-specified tolerance $0 < \delta < 1$, $D := \frac{1}{\sigma_w^2 c_\delta} \sum_{t=1}^{n-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$, where $c_\delta = \chi^2_{n_x^2+n_x n_u}(\delta)$, defines a $1-\delta$ probability credibility region for the true parameters as follows:

**Lemma 1 (Umenberger et al. (2019))** *Given data $\mathcal{D}_n$ from (1) with true parameters $A = A_{\text{tr}}$ and $B = B_{\text{tr}}$, and a user-specified $0 < \delta < 1$, define the set*

$$\Theta_m(\mathcal{M}_\delta(\mathcal{D}_n)) := \{A,\ B\ :\ X^\top D X \preceq I,\ X = [\hat{A} - A,\ \hat{B} - B]^\top\}. \tag{2}$$

*where $\{\hat{A}, \hat{B}, D\} = \mathcal{M}_\delta(\mathcal{D}_n)$. Then $\{A_{\text{tr}},\ B_{\text{tr}}\} \in \Theta_m(\mathcal{M}_\delta)$ w.p. $1 - \delta$.*

Lemma 1 is a consequence of the fact the posterior distribution of parameters $A, B$ (for a uniform prior) is Gaussian for models of the form (1), cf. e.g. Umenberger and Schön (2018). Similar credibility regions have been attained using results from high-dimensional statistics in recent works such as Dean et al. (2017).

**Control objective** Our objective is to design a feedback control policy $u_t = \phi(x_{1:t}, u_{1:t-1})$ so as to minimize the cost function $\sum_{t=1}^{T} c(x_t, u_t)$, where $c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$ for user-specified positive semidefinite matrices $Q$ and $R$. When the parameters of the true system, $\{A_{\text{tr}}, B_{\text{tr}}\}$, are known this is the well-known LQR problem. As discussed, we do not assume knowledge of the true parameters; as such, the controller must regulate and learn the system simultaneously. To this end, we partition the total 'control time' $[1, T]$ into two intervals: $[1, T_e]$ and $[T_e + 1, T]$ for $T_e \in \mathbb{N}$. At time $t = 1$, given initial data $\mathcal{D}_0$, a policy $\phi_1$ is designed and applied to the system for $t \in [1, T_e]$. Then, at time $t = T_e + 1$, a new policy $\phi_2$ is designed, based on $\mathcal{D}_0$ and data $\mathcal{D}_{T_e}$ collected under $\phi_1$. Policy $\phi_2$ is then applied for $t \in [T_e + 1, T]$. We can write this control task as:

$$\min_{\phi_1,\phi_2} \quad \mathbb{E}\sum_{t=1}^{T} c(x_t, u_t), \quad \text{s.t. dynamics in (1) with } A = A_{\text{tr}},\ B = B_{\text{tr}}, \tag{3}$$
$$u_t = \phi_1(\cdot),\ t = 1, \dots, T_e,\ u_t = \phi_2(\cdot),\ t = T_e + 1, \dots, T.$$

One can think of the first interval, $[1, T_e]$, as an 'exploration' or 'learning' period where the data collected is used to design an improved controller, $\phi_2$, applied during the second interval $[T_e+1, T]$, which could be considered an 'exploitation' period. However, the task is to minimize the total cost therefore, it is important to balance exploration and exploitation. The decision to nominate a specific time, $T_e$, at which the control policy will be 'updated' requires some justification. A more natural formulation might update the controller whenever new data becomes available. We shall discuss this aspect of the formulation in more detail in §3.4, where alternative formulations are considered. For now, suffice to say that this formulation, i) simplifies the presentation of the technical developments to follow, and ii) still captures the importance of balancing 'exploration' with 'exploitation.'

Observe that the control task in (3) depends on the true, but unknown, system parameters $A_{\text{tr}}, B_{\text{tr}}$, as we want to optimize for performance on the true system. In place of the true system parameters, we will optimize for our 'best guess' of the parameters, i.e., the nominal parameters $\hat{A}, \hat{B}$ from least squares corresponding to the mode of the posterior distribution. To ensure reasonable behavior of the true system in closed-loop, we also require the controllers to stabilize the true system with high probability. Let $\mathcal{S}_{\text{CL}}(A, B, \phi)$ denote the closed loop system formed by combining (1) with the policy $\phi$. The problem addressed in this paper is as follows:

$$\min_{\phi_1,\phi_2} \quad \mathbb{E}\sum_{t=1}^{T} c(x_t, u_t) \tag{4a}$$
$$\text{s.t. } x_{t+1} = \hat{A}_1 x_t + \hat{B}_1 u_t + w_t,\ \{\hat{A}_1, \hat{B}_1\} = \mathcal{M}_\delta(\mathcal{D}_0),\ u_t = \phi_1(\cdot),\ t \in [1, T_e] \tag{4b}$$
$$x_{t+1} = \hat{A}_2 x_t + \hat{B}_2 u_t + w_t,\ \{\hat{A}_2, \hat{B}_2\} = \mathcal{M}_\delta(\mathcal{D}_0 \cup \mathcal{D}_{T_e}),\ u_t = \phi_2(\cdot),\ t \in [T_e + 1, T], \tag{4c}$$
$$\mathcal{S}_{\text{CL}}(A, B, \phi_1) \text{ is stable } \forall \{A, B\} \in \Theta_m(\mathcal{M}_\delta(\mathcal{D}_0)) \tag{4d}$$
$$\mathcal{S}_{\text{CL}}(A, B, \phi_2) \text{ is stable } \forall \{A, B\} \in \Theta_m(\mathcal{M}_\delta(\mathcal{D}_0 \cup \mathcal{D}_{T_e})),\ w_t \sim \mathcal{N}\left(0, \sigma_w^2 I_{n_x}\right) \forall t. \tag{4e}$$

## 3. Controller synthesis

In this section, we present an approximate solution to the problem presented in (4). In what follows, **bold** symbols denote the z-transform of time domain signals, e.g., the z-transform of $x$ is denoted $\mathbf{x}$.

### 3.1. Preliminary results from System Level Synthesis

In this section we review some essential results from the System Level Synthesis (SLS) framework proposed by Wang et al. (2019b); for a comprehensive tutorial, cf. Anderson et al. (2019). Consider the closed-loop behavior of (1) under the stabilizing controller $\mathbf{u} = \mathbf{K}\mathbf{x}$; in particular, consider the transfer functions $\mathbf{\Phi_x}$ and $\mathbf{\Phi_u}$ from disturbance $\mathbf{w}$ to state $\mathbf{x}$ and control $\mathbf{u}$, respectively. This can be expressed as $\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} \mathbf{w}$. Following in the spirit of the Youla parameterization, rather than designing the controller $\mathbf{K}$ to obtain the closed-loop responses $\mathbf{\Phi_x} = (zI - A - B\mathbf{K})^{-1}$ and $\mathbf{\Phi_u} = \mathbf{K}\mathbf{\Phi_x}$, in SLS one designs the closed-loop responses directly, and then recovers the controller as $\mathbf{K} = \mathbf{\Phi_u}\mathbf{\Phi_x}^{-1}$. The following theorem characterizes the space of all closed-loop responses achievable by a stabilizing controller.

**Theorem 2 (Anderson et al. (2019), Theorem 4.1)** *The affine subspace defined by*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} = I, \quad \mathbf{\Phi_x}, \mathbf{\Phi_u} \in \frac{1}{z}\mathcal{RH}_\infty, \tag{5}$$

*parametrizes all closed-loop responses achievable by a stabilizing controller. Further, the response is achieved by the controller $\boldsymbol{K} = \mathbf{\Phi_u}\mathbf{\Phi_x}^{-1}$.*

The following theorem considers a 'perturbed' version of the constraints in (5), that is useful for synthesizing robust controllers, e.g., when the system parameters $A$,$B$ are uncertain.

**Theorem 3 (Anderson et al. (2019), Theorem 4.3)** *Suppose that $\mathbf{\Phi_x}, \mathbf{\Phi_u}, \mathbf{\Delta}$ satisfy*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} = I + \mathbf{\Delta}, \quad \mathbf{\Phi_x}, \mathbf{\Phi_u} \in \frac{1}{z}\mathcal{RH}_\infty. \tag{6}$$

*Then the controller $\boldsymbol{K} = \mathbf{\Phi_u}\mathbf{\Phi_x}^{-1}$ stabilizes system (1) with parameters $A, B$ if and only if $(I + \mathbf{\Delta})^{-1}$ is stable.*

While the preceding theorems define affine subspaces (i.e. (5) and (6)) that are convenient to optimize over, the decision variables $\mathbf{\Phi_x}$ and $\mathbf{\Phi_u}$ are infinite dimensional transfer matrices. As is common in the SLS framework, we will work with finite impulse response (FIR) approximations:

$$\mathbf{\Phi_x}(z) = \sum\nolimits_{k=0}^{F} \Phi_x^k z^{-k}, \quad \mathbf{\Phi_u}(z) = \sum\nolimits_{k=0}^{F} \Phi_u^k z^{-k}. \tag{7}$$

Henceforth, we will restrict our attention to policies of the form $\phi(z) = \mathbf{\Phi_u}(z)\mathbf{\Phi_x}(z)^{-1}$.

### 3.2. Robust control formulation

In this section, we present a convex formulation of the following problem: (approximately) optimize the infinite-horizon quadratic cost, for a given nominal model $\{\hat{A}, \hat{B}\}$, while robustly stabilizing all models $\{A, B\}$ in the model set $\Theta_m(\mathcal{M}_\delta)$. This result extends existing SLS formulations, by preserving the structure in the representation of uncertainty captured by $D$.

Following straightforward calculations, cf. e.g. (Anderson et al., 2019, §2.2.2), the infinite horizon cost function can be written as

$$\lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}[c(x_t, u_t)] = \mathbb{E} \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} \right\|_F^2 = \sigma_w^2 \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} \right\|_{\mathcal{H}_2}^2 \tag{8}$$

subject to the affine constraints in (5) with the nominal parameters $\hat{A}, \hat{B}$, i.e.,

$$\begin{bmatrix} zI - \hat{A} & -\hat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} = I. \tag{9}$$

By making use of the FIR approximation in (7), the rightmost side of (8) can be written as:

$$J_{\mathcal{H}_2} = \sigma_w^2 \left\| \begin{bmatrix} Q^{\frac{1}{2}} \otimes I_F & 0 \\ 0 & R^{\frac{1}{2}} \otimes I_F \end{bmatrix} \begin{bmatrix} \bar{\mathbf{\Phi}}_\mathbf{x} \\ \bar{\mathbf{\Phi}}_\mathbf{u} \end{bmatrix} \right\|_F^2, \tag{10}$$

where $\bar{\mathbf{\Phi}}_\mathbf{x} = \left[\Phi_x^\top(0) \dots \Phi_x^\top(F)\right]^\top$ and $\bar{\mathbf{\Phi}}_\mathbf{u} = \left[\Phi_u^\top(0) \dots \Phi_u^\top(F)\right]^\top$, denote the FIR parameters from (7), stacked vertically. Note that (10) is a convex quadratic function of $\bar{\mathbf{\Phi}}_\mathbf{x}$ and $\bar{\mathbf{\Phi}}_\mathbf{u}$.

To ensure robustness of the policy on the true, unknown system (as in, e.g., (4d)), we make use of Theorem 3. Specifically, as $\mathbf{\Phi_x}$ and $\mathbf{\Phi_u}$ are constrained to satisfy (9), we can express $\mathbf{\Delta}$ in (6) as $\mathbf{\Delta} = (\hat{A} - A_{\text{tr}})\mathbf{\Phi_x} + (\hat{B} - B_{\text{tr}})\mathbf{\Phi_u}$, by substituting (9) into (6), with $A = A_{\text{tr}}$ and $B = B_{\text{tr}}$. By Theorem 3, the controller $\mathbf{\Phi_u}\mathbf{\Phi_x}^{-1}$ will stabilize the true system, if and only if $(I + \mathbf{\Delta})^{-1}$ is stable. Of course, $\mathbf{\Delta}$ is defined in terms of $A_{\text{tr}}, B_{\text{tr}}$, which are unknown; however, by Lemma 1, they are known to lie in $\Theta_m(\mathcal{M}_\delta)$ with high-probability. A sufficient condition for stability of $(I + \mathbf{\Delta})^{-1}$ is given by the small gain theorem: $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \leq 1$ implies stability of $(I + \mathbf{\Delta})^{-1}$. The $\mathcal{H}_\infty$-norm of a transfer matrix can be computed/constrained as follows:

**Lemma 4 (Dumitrescu (2007))** *Let $\boldsymbol{H}(z) = \sum_{i=0}^{F} H(i) z^{-i}$, with $H \in \mathbb{R}^{p \times m}$ and $\bar{H} = \left[H(0)^\top \dots H(F)^\top\right]^\top$. Then $\|\boldsymbol{H}\|_{\mathcal{H}_\infty} \leq \gamma$ iff there exists $P \in \mathbb{S}^{p(F+1)}$ satisfying*

$$P = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0F} \\ \star & P_{11} & \cdots & P_{1F} \\ \star & \star & \ddots & \vdots \\ \star & \star & \star & P_{FF} \end{bmatrix}, \sum_{i=0}^{F} P_{ii} = \gamma I, \sum_{i=0}^{F-k} P_{i(i+k)} = 0, k = 1, \dots, F, \tag{11a}$$

$$\begin{bmatrix} P & \bar{H} \\ \bar{H}^\top & I \end{bmatrix} \succeq 0. \tag{11b}$$

We can now present the main contribution of this section.

**Theorem 5** *Given a model $\mathcal{M}_\delta = \{\hat{A}, \hat{B}, D\}$, a convex upper bound for $\min_\phi \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}[c(x_t, u_t)]$ for a system (1) with parameters $A = \hat{A}$ and $B = \hat{B}$, subject to the constraint that the controller $\phi$ stabilizes all models in $\Theta_m(\mathcal{M}_\delta)$, is given by $J_\infty^*(\mathcal{M}_\delta) = \min_\phi J_\infty(\phi, \mathcal{M}_\delta)$, where*

$$J_\infty(\phi, \mathcal{M}_\delta) := \left\{ J_{\mathcal{H}_2}(\mathbf{\Phi_x}, \mathbf{\Phi_u}) \mid (9),\ \exists P, \lambda \in \mathbb{R} \text{ s.t. } (11a), \begin{bmatrix} P & 0 & \bar{\mathbf{\Phi}} \\ 0 & (1-\lambda)I & 0 \\ \bar{\mathbf{\Phi}}^\top & 0 & \lambda D \end{bmatrix} \succeq 0 \right\} \tag{12}$$

*where $\bar{\mathbf{\Phi}} = \begin{bmatrix} \Phi_x(0) & \dots & \Phi_x(F) \\ \Phi_u(0) & \dots & \Phi_u(F) \end{bmatrix}^\top$, and $\phi$ can be realized as $\phi = \mathbf{\Phi_u}\mathbf{\Phi_x}^{-1}$.*

**Proof** The convex program $\min_{\mathbf{\Phi_x},\mathbf{\Phi_u},P}\ J_{\mathcal{H}_2}(\mathbf{\Phi_x},\mathbf{\Phi_u})$ s.t. (9) optimizes the infinite-horizon cost. All that remains is to ensure robust stability: a sufficient condition is $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \leq 1$. This sufficient, but not necessary, condition is the source of the conservatism, i.e., the reason we only have an upper bound. $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \leq 1$ can be enforced by combining Lemma 4 with the following lemma:

**Lemma 6 (Luo et al. (2004))** *The data matrices $(\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{P},\mathcal{F},\mathcal{G},\mathcal{H})$ satisfy, for all $X$ with $I - X^\top \mathcal{P} X \succeq 0$, the robust fractional quadratic matrix inequality*

$$\begin{bmatrix} \mathcal{H} & \mathcal{F}+\mathcal{G}X \\ (\mathcal{F}+\mathcal{G}X)^\top & \mathcal{C}+X^\top\mathcal{B}+\mathcal{B}^\top X+X^\top\mathcal{A}X \end{bmatrix} \succeq 0, \text{ iff } \begin{bmatrix} \mathcal{H} & \mathcal{F} & \mathcal{G} \\ \mathcal{F}^\top & \mathcal{C}-\lambda I & \mathcal{B}^\top \\ \mathcal{G}^\top & \mathcal{B} & \mathcal{A}+\lambda\mathcal{P} \end{bmatrix} \succeq 0, \tag{13}$$

*for some $\lambda \geq 0$.*

Note that $\mathbf{\Delta}^\top = [\mathbf{\Phi_x}^\top,\ \mathbf{\Phi_u}^\top][\hat{A}-A_{\text{tr}},\ \hat{B}-B_{\text{tr}}]^\top = [\mathbf{\Phi_x}^\top,\ \mathbf{\Phi_u}^\top]X$ where $X$ is defined in (2). Furthermore, $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \leq 1 \iff \left\|\mathbf{\Delta}^\top\right\|_{\mathcal{H}_\infty} \leq 1$. Let $\mathbf{H}(z)$ (from Lemma 4) denote $\mathbf{\Delta}^\top$, then the (11b) can be put in the form of the MI on the left in (13) by choosing $\mathcal{H}=P$, $\mathcal{G}=\bar{\mathbf{\Phi}}$, $\mathcal{C}=I$, and $\mathcal{A},\mathcal{B},\mathcal{F}$ all zero. Further, by choosing $\mathcal{P}=D$ in Lemma 6, the condition $I \succeq X^\top\mathcal{P}X$ is equivalent to $\{A_{\text{tr}},B_{\text{tr}}\} \in \Theta_m(\mathcal{M}_\delta)$, cf. (2). The final constraint (LMI) in (12) is then equivalent to the second LMI in (13), which implies that $\left\|\mathbf{\Delta}^\top\right\|_{\mathcal{H}_\infty} \leq 1$ for the true model parameters w.p. $1-\delta$. ■

Observe that this formulation preserves the structure in the uncertainty representation, as encoded in the matrix $D$. This is in contrast to other SLS methods, e.g., Dean et al. (2017, 2019), that reduce model uncertainty to a single (or at most two) scalar quantities, e.g., $\|\hat{A}-A_{\text{tr}}\|_2 \leq \epsilon_A$.

### 3.3. Robust dual control formulation

In this section we return to the dual control problem outlined in (4), namely: minimize cost over $[1,T]$, via an initial policy $\phi_1$, designed using data $\mathcal{D}_0$, and applied for $t \in [1,T_e]$, followed by a second policy $\phi_2$, designed using additional data $\mathcal{D}_{T_e}$, and applied for $t \in [T_e+1,T]$. The key idea is dual control: $\phi_1$ affects not only the cost, but also the data $\mathcal{D}_{T_e}$ available for the design of $\phi_2$.

**Infinite-horizon approximation** In what follows, we will approximate the cost $\sum_{t=1}^{N}\mathbb{E}c(x_t,u_t)$ by the infinite-horizon (i.e. stationary value) $N \times \lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}c(x_t,u_t)$. Such an approximation can be expected to be valid when the horizon $N$ is sufficiently long, so as to allow the system to reach the stationary distribution. Alternatives to this approximation are discussed in §3.4. With the infinite horizon approximation, we can express (4) as

$$\min_{\phi_1}\ T_e \times J_\infty(\phi_1,\mathcal{M}_\delta(\mathcal{D}_0)) + (T-T_e)\times J_\infty^*(\mathcal{M}_\delta(\mathcal{D}_0\cup\mathcal{D}_{T_e})). \tag{14}$$

Note the dependence of the cost during $[T_e,T]$ on $\mathcal{D}_{T_e}$. Note also that $\phi_2$ is defined implicitly by $J_\infty^*$, cf. (12). Problem (14) cannot be solved exactly, as it depends on $\mathcal{D}_{T_e}$ which is not available at time $t=1$. As such, we must predict the influence that $\phi_1$ will have on 'future' data $\mathcal{D}_{T_e}$.

**Propagating uncertainty** To approximately solve (14), we require an approximation of $\mathcal{M}_\delta(\mathcal{D}_0\cup\mathcal{D}_{T_e})$, i.e., $\tilde{\mathcal{M}}(\phi_1) \approx \mathbb{E}[\mathcal{M}_\delta(\mathcal{D}_0\cup\mathcal{D}_{T_e})|\mathcal{D}_0,\phi_1]$. Let $\mathcal{M}_\delta(\mathcal{D}_0)=\{\hat{A}_1,\hat{B}_1,D_1\}$. We then define the approximate model as $\tilde{\mathcal{M}}(\phi_1) =: \{\hat{A}_1,\hat{B}_1,\tilde{D}\}$. First, note that we approximate the predicted nominal parameters by the current estimates. Updating these estimates based on the expected value of

future data involves difficult integrals that must be computed numerically, which would destroy convexity, cf. Lobo and Boyd (1999). The predicted 'uncertainty matrix' $\tilde{D}$ is defined as follows. Given $\mathcal{D}_{T_e}$, $D$ at time $T_e$ can be computed as $D_1 + \frac{1}{\sigma_w^2 c_\delta}\sum_{t=1}^{T_e}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$. We can approximate the empirical covariance with stationary distribution over $x$ and $u$; i.e., as with the stationary approximation of the cost in (8), we have: $\sum_{t=1}^{T_e}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top \approx T_e\mathbb{E}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top = T_e\sigma_w^2\bar{\mathbf{\Phi}}^\top\bar{\mathbf{\Phi}}$. We then define $\tilde{D} := D_1 + \frac{T_e}{c_\delta}\bar{\mathbf{\Phi}}^\top\bar{\mathbf{\Phi}}$.

**Convex relaxation of dual control** By substituting the approximate model $\tilde{\mathcal{M}}(\phi_1)$ for $\mathcal{M}_\delta(\mathcal{D}_0 \cup \mathcal{D}_{T_e})$ in (14), we can remove the dependence of the cost on unknown future data. Furthermore, observe that (for fixed $\lambda$) the LMI constraint in (12) is linear in $D$. This implies that we can optimize over $D$ and $\phi$ jointly, as a convex program. Unfortunately, $\tilde{D}$ in $\tilde{\mathcal{M}}(\phi_1)$ is a quadratic function of $\phi_1$, and so directly substituting $D \to \tilde{D}$ results in a non-convex matrix inequality.

To circumvent this difficulty, we introduce the following linear approximation of $\tilde{D}$:

$$D_\ell := D_1 + \frac{T_e}{c_\delta}\left(\bar{\mathbf{\Phi}}^\top\bar{\mathbf{\Phi}}_{\text{nom}} + \bar{\mathbf{\Phi}}_{\text{nom}}^\top\bar{\mathbf{\Phi}} - \bar{\mathbf{\Phi}}_{\text{nom}}^\top\bar{\mathbf{\Phi}}_{\text{nom}}\right), \quad \bar{\mathbf{\Phi}}_{\text{nom}} = \arg\min_{\mathcal{K}} J_\infty(\mathcal{K}, \mathcal{M}_\delta(\mathcal{D}_0)). \tag{15}$$

$D_\ell$ is nothing more than a first-order approximation of $\tilde{D}$, linearized at $\bar{\mathbf{\Phi}}_{\text{nom}}$. We could linearize at any arbitrary point; however, the solution $\bar{\mathbf{\Phi}}_{\text{nom}}$ to the nominal robust control problem given $\mathcal{D}_0$ is a natural choice. Furthermore, $D_\ell \preceq \tilde{D}$, as $(\bar{\mathbf{\Phi}} - \bar{\mathbf{\Phi}}_{\text{nom}})^\top(\bar{\mathbf{\Phi}} - \bar{\mathbf{\Phi}}_{\text{nom}}) \succeq 0 \iff \bar{\mathbf{\Phi}}^\top\bar{\mathbf{\Phi}} \succeq \bar{\mathbf{\Phi}}^\top\bar{\mathbf{\Phi}}_{\text{nom}} + \bar{\mathbf{\Phi}}_{\text{nom}}^\top\bar{\mathbf{\Phi}} - \bar{\mathbf{\Phi}}_{\text{nom}}^\top\bar{\mathbf{\Phi}}_{\text{nom}}$. We now present the main contribution of this section: given the initial model $\mathcal{M}_\delta(\mathcal{D}_0) = \{\hat{A}_1, \hat{B}_1, D_1\}$, the problem $\min_{\phi_1}\ T_e \times J_\infty(\phi_1, \mathcal{M}_\delta(\mathcal{D}_0)) + (T - T_e) \times J_\infty^*(\tilde{\mathcal{M}}(\phi_1))$ admits the following convex upper bound:

$$\min_{\mathbf{\Phi_x}^1, \mathbf{\Phi_u}^1, \mathbf{\Phi_x}^2, \mathbf{\Phi_u}^2, P^1, P^2, \lambda^1} \quad T_e \times J_{\mathcal{H}_2}(\mathbf{\Phi_x}^1, \mathbf{\Phi_u}^1) + (T - T_e) \times J_{\mathcal{H}_2}(\mathbf{\Phi_x}^2, \mathbf{\Phi_u}^2) \tag{16a}$$

$$\text{s.t.} \quad \{\mathbf{\Phi_x}^1, \mathbf{\Phi_u}^1\} \text{ and } \{\mathbf{\Phi_x}^2, \mathbf{\Phi_u}^2\} \text{ each satisfy (9) with } A = \hat{A}_1, B = \hat{B}_1 \tag{16b}$$

$$P^1, P^2 \text{ each satisfy (11a)} \tag{16c}$$

$$\begin{bmatrix} P^1 & 0 & \bar{\mathbf{\Phi}}^1 \\ 0 & (1-\lambda^1)I & 0 \\ (\bar{\mathbf{\Phi}}^1)^\top & 0 & \lambda^1 D_1 \end{bmatrix} \succeq 0, \begin{bmatrix} P^2 & 0 & \bar{\mathbf{\Phi}}^2 \\ 0 & (1-\lambda^2)I & 0 \\ (\bar{\mathbf{\Phi}}^2)^\top & 0 & \lambda^2 D_\ell \end{bmatrix} \succeq 0 \tag{16d}$$

where $\bar{\mathbf{\Phi}}^i$ is defined analogously to $\bar{\mathbf{\Phi}}$ in Theorem 5, for $i = 1, 2$. As $D_\ell \preceq \tilde{D}$, the feasible set for (16) with $D_\ell$ (in (16d)) is smaller than that of the program with (quadratic) $\tilde{D}$. Therefore, (16) constitutes an upper bound. Notice that (16) is only convex for fixed $\lambda^2$, due to bilinearity with $D_\ell$. In practice, one has to grid search over the scalar parameter $\lambda_2$.

### 3.4. Discussion

In this section, we discuss the partitioning of $[1, T]$ into two sub-intervals. A downside of this approach is that it requires the user to explicitly specify the 'exploration' period, i.e., select $T_e$. The proposed approach could also be considered a 'one-step-look-ahead' dual control, as there is only a single period of exploration ($[1, T_e]$), before a single period of exploitation ($[T_e, T]$). Such a drawback can be partially mitigated by adopting a 'multistep-look-ahead' strategy, as in Umenberger et al. (2019). In such a framework, the current period of exploration is followed by further exploration, rather than pure exploitation. This approach also requires the user to select 'epoch times',

at which the controller will be updated. To circumvent this, one could adopt a model predictive control (MPC) strategy, with SLS as in Wang et al. (2019a), but exploiting the dual control effect, as in Lobo and Boyd (1999). The major obstacle to extending the proposed approach to the multistep or MPC setting is the need to search over more than one multiplier, $\lambda_2$, cf. (16d). Such extensions represent interesting directions for future research. At any rate, the method proposed in this paper could be used (with fixed multipliers) as a convex means of providing sub-optimal initialization for local search methods.

## 4. Numerical illustration

In this section we illustrate the proposed method with a numerical example. Consider the linear control problem with parameters:

$$A_{\text{tr}} = \begin{bmatrix} 0.5 & 1.1 \\ 0 & 0.8 \end{bmatrix}, B_{\text{tr}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, Q = \begin{bmatrix} 1 & 0 \\ 0 & 0.001 \end{bmatrix}, R = \begin{bmatrix} 10^3 & 0 \\ 0 & 10^3 \end{bmatrix}, \sigma_w = 1, \delta = 0.1$$

Initial data $\mathcal{D}_0$ is generated by simulating (1) open-loop with $u_t \sim \mathcal{N}(0, I)$ for $t = 6$ timesteps; this is repeated 10 times. For control, $T = 100$, $T_e = 10$, and $F = 12$. We compare three methods: i) nominal control: $\phi_1 = \arg\min_\phi J_\infty(\phi, \mathcal{M}_\delta(\mathcal{D}_0))$ and $\phi_2 = \arg\min_\phi J_\infty(\phi, \mathcal{M}_\delta(\mathcal{D}_0 \cup \mathcal{D}_{T_e}))$, i.e., no explicit exploration; ii) dual control: as proposed in this paper; iii) greedy control: $\phi_1$ is given by $u = \mathbf{K}x + e$ where $\mathbf{K}$ is the nominal controller (i), and $e \sim \mathcal{N}(0, \sigma I)$. $\sigma$ is tuned (a posteriori) to give the same exploration cost as dual control on the true system, and represents a 'greedy strategy' of injecting as much 'exploration signal' $e$ into the input as possible, as opposed to targeting uncertainty reduction in specific parameters. $\phi_2$ is given by the nominal control, synthesized with the additional data collected during exploration. This experiment is repeated 1000 times, with the results presented in Fig. 1. The greedy strategy performs slightly better than the nominal controller during exploitation, but this cannot offset the increased cost of exploration, leading to worse performance in terms of total cost (total cost is the sum of cost during exploration and exploitation). Dual control balances exploration and exploitation to achieve the lowest total cost among all methods compared: although it incurs the same exploration cost as the greedy strategy (due to the tuning of $\sigma$ in the greedy algorithm), it achieves significantly lower exploitation cost, as the exploration is targeted towards reducing uncertainty in the parameters that 'matter most' for the task at hand.

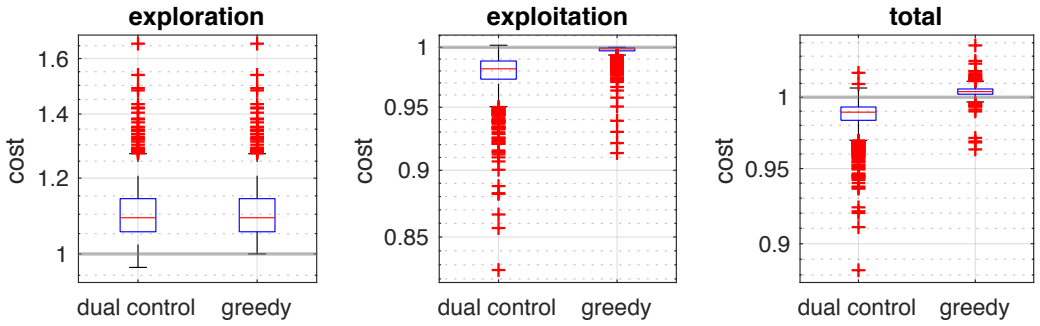


Figure 1: Costs during exploration ($t \in [1, T_e]$), exploitation ($t \in [T_e, T]$), and the total cost (exploration + exploitation). Costs are normalized by the cost of the nominal control (i.e. unity implies the same cost as the nominal control). Dual control exhibits best performance.

## Acknowledgments

We would like to thank the anonymous reviewers for their many useful comments in improving the quality of this manuscript. We regret that we were not able to incorporate all their suggestions, due to space restrictions. This research was financially supported by the National Science Foundation (Award No. EFMA-1830901), the Department of Navy, Office of Naval Research (Award No. N00014-18-1-2210), the Swedish Foundation for Strategic Research (via the project *ASSEMBLE*, Contract No. RIT15-0012) and the Swedish Research Council (via the project *NewLEADS - New Directions in Learning Dynamical Systems*, Contract No. 621-2016-06079).

## References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Marc Abeille and Alessandro Lazaric. Thompson Sampling for linear-quadratic control problems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M Kakade, and Karan Singh. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

James Anderson, John C Doyle, Steven H Low, and Nikolai Matni. System level synthesis. *Annual Reviews in Control*, 2019.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *to appear in Foundations of Computational Mathematics (arXiv:1710.01688)*, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4192–4201, 2018.

Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. *to appear at American Control Conference (arXiv:1809.10121)*, 2019.

Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*, volume 103. Springer, 2007.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time adaptive stabilization of LQ systems. *IEEE Transactions on Automatic Control*, 2019. accepted.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

AA Feldbaum. Dual control theory. i. *Avtomatika i Telemekhanika*, 21(9):1240–1249, 1960.

Mina Ferizbegovic, Jack Umenberger, Håkan Hjalmarsson, and Thomas B Schön. Learning robust lq-controllers using application oriented exploration. *IEEE Control Systems Letters*, 4(1):19–24, 2019.

Andrea Iannelli, Mohammad Khosravi, and Roy S Smith. Structured exploration in the finite horizon linear quadratic dual control problem. *arXiv preprint arXiv:1910.14492*, 2019.

Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

Miguel Sousa Lobo and Stephen Boyd. Policies for simultaneous estimation and optimization. In *Proceedings of the 1999 American Control Conference (Cat. No. 99CH36251)*, volume 2, pages 958–964. IEEE, 1999.

Zhi-Quan Luo, Jos F Sturm, and Shuzhong Zhang. Multivariate nonnegative quadratic mappings. *SIAM Journal on Optimization*, 14(4):1140–1162, 2004.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*, 2018.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. *arXiv preprint arXiv:1906.11392*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *submitted to IEEE Transactions on Automatic Control (arXiv:1709.04047)*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Jack Umenberger and Thomas B Schön. Learning convex bounds for linear quadratic control policy synthesis. In *Advances in Neural Information Processing Systems*, pages 9584–9595, 2018.

Jack Umenberger, Mina Ferizbegovic, Thomas B Schön, and Håkan Hjalmarsson. Robust exploration in linear quadratic reinforcement learning. *arXiv preprint arXiv:1906.01584*, 2019.

Han Wang, Shaoru Chen, Victor M Preciado, and Nikolai Matni. Robust model predictive control via system level synthesis. *arXiv preprint arXiv:1911.06842*, 2019a.

Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. A system level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 2019b.