Learning convex bounds for linear quadratic control policy synthesis

Jack Umenberger  
Department of Information Technology  
Uppsala University  
Sweden  
jack.umenberger@it.uu.se

Thomas B. Schön  
Department of Information Technology  
Uppsala University  
Sweden  
thomas.schon@it.uu.se

Abstract

Learning to make decisions from observed data in dynamic environments remains a problem of fundamental importance in a number of fields, from artificial intelligence and robotics, to medicine and finance. This paper concerns the problem of learning control policies for unknown linear dynamical systems so as to maximize a quadratic reward function. We present a method to optimize the expected value of the reward over the posterior distribution of the unknown system parameters, given data. The algorithm involves sequential convex programming, and enjoys reliable local convergence and robust stability guarantees. Numerical simulations and stabilization of a real-world inverted pendulum are used to demonstrate the approach, with strong performance and robustness properties observed in both.

1 Introduction

Decision making for dynamical systems in the presence of uncertainty is a problem of great prevalence and importance, as well as considerable difficulty, especially when knowledge of the dynamics is available only via limited observations of system behavior. In machine learning, the data-driven search for a control policy to maximize the expected reward attained by a stochastic dynamic process is known as reinforcement learning (RL) [41]. Despite remarkable recent success in games [28, 39], a major obstacle to the deployment RL-based control on physical systems (e.g. robots and self-driving cars) is the issue of robustness, i.e., guaranteed safe and reliable operation. With the necessity of such guarantees widely acknowledged [2], so-called ‘safe RL’ remains an active area of research [18].

The problem of robust automatic decision making for uncertain dynamical systems has also been the subject of intense study in the area of robust control (RC) [49]. In RC, one works with a set of plausible models and seeks a control policy that is guaranteed to stabilize all models within the set. In addition, there is also a performance objective to optimize, i.e., a reward to be maximized, or equivalently, a cost to be minimized. Such cost functions are usually defined with reference to either a nominal model [17, 22] or the worst-case model [32] in the set. RC has been extremely successful in a number of engineering applications [34]; however, as has been noted, e.g., [43, 31], robustness may (understandably) come at the expense of performance, particularly for worst-case design.

The problem we address in this paper lies at the intersection of reinforcement learning and robust control, and can be summarized as follows: given observations from an unknown dynamical system, we seek a policy to optimize the expected cost (as in RL), subject to certain robust stability guarantees (as in RC). Specifically, we focus our attention on control of linear time-invariant dynamical systems, subject to Gaussian disturbances, with the goal of minimizing a quadratic function penalizing state deviations and control action. When the system is known, this is the classical linear quadratic regulator (LQR), a.k.a. $H_2$, optimal control problem [8]. We are interested in the setting in which the system is unknown, and knowledge of the dynamics must be inferred from observed data.
Contributions and paper structure  The principle contribution of this paper is an algorithm to optimize the expected value of the linear quadratic regulator reward/cost function, where the expectation is w.r.t. the posterior distribution of unknown system parameters, given observed data; c.f. Section 3 for a detailed problem formulation. Specifically, we construct a sequence of convex approximations (upper bounds) to the expected cost, that can be optimized via semidefinite programing 45. The algorithm, developed in Section 4 invokes the majorize-minimization (MM) principle 26, and consequently enjoys reliable convergence to local optima. An important part of our contribution lies in guarantees on the robust stability properties of the resulting control policies, c.f. Section 4.3. We demonstrate the proposed method via two experimental case studies: i) the benchmark problem on simulated systems considered in 14 43, and ii) stabilization of a real-world inverted pendulum. Strong performance and robustness properties are observed in both. Moving forward, from a machine learning perspective this work contributes to the growing body of research concerned with ensuring robustness in RL, c.f. Section 2. From a control perspective, this work appropriates cost functions more commonly found in RL (namely, expected reward) to a RC setting, with the objective of reducing conservatism of the resulting robust control policies.

2 Related work

Incorporating various notions of ‘robustness’ into RL has long been an area of active research 18. In so-called ‘safe RL’, one seeks to respect certain safety constraints during exploration and/or policy optimization, for example, avoiding undesirable regions of the state-action space 19 11. A related problem is addressed in ‘risk-sensitive RL’, in which the search for a policy takes both the expected value and variance of the reward into account 27 16. Recently, there has been an increased interest in notions of robustness more commonly considered in control theory, chiefly stability 31 8. Of particular relevance is the work of 3, which employs Lyapunov theory 24 to verify stability of learned policies. Like the present paper, 4 adopts a Bayesian framework; however, 4 makes use of Gaussian processes 45 to model the uncertain nonlinear dynamics, which are assumed to be deterministic. A major difference between 4 and our work is the cost function; in the former the policy is selected by optimizing for worst-case performance, whereas we optimize the expected cost. Robustness of data-driven control has also been the focus of a recently developed family of methods referred to as ‘coarse-ID control’, c.f., 12 14 7 40, in which finite-data bounds on the accuracy of the least squares estimator are combined with modern robust control tools, such as system level synthesis 47. Coarse-ID builds upon so-called ‘$H_{\infty}$ identification’ methods for learning models of dynamical systems, along with error bounds that are compatible with robust synthesis methods 23 11 10. $H_{\infty}$ identification assumes an adversarial (i.e. worst-case) disturbance model, whereas Coarse-ID is applicable to probabilistic models, such as those considered in the present paper. Of particular relevance to the present paper is 14, which provides sample complexity bounds on the performance of robust control synthesis for the infinite horizon LQR problem, when the true system is not known. Such bounds necessarily consider the worst-case model, given the observed data, where as we are concerned with expected cost over the posterior distribution of models. In closing, we briefly mention the so-called ‘Riemann-Stieltjes’ class of optimal control problems, for uncertain continuous-time dynamical systems, c.f., e.g., 37 36. Such problems often arise in aerospace applications (e.g. satellite control) where the objective is to design an open-loop control signal (e.g. for an orbital maneuver) rather than a feedback policy.

3 Problem formulation

In this section we describe in detail the specific problem that we address in this paper. The following notation is used: $\mathbb{S}^n$ denotes the set of $n \times n$ symmetric matrices; $\mathbb{S}^n_+$ (\(\mathbb{S}^n_{++}\)) denotes the cone of positive semidefinite (positive definite) matrices. $A \succeq B$ denotes $A - B \in \mathbb{S}^n_+$, similarly for $>$ and \(\mathbb{S}^n_{++}\). The trace of $A$ is denoted $\text{tr}\ A$. The transpose of $A$ is denoted $A'$. $|a|^2_q$ is shorthand for $a'Qa$. The convex hull of set $\Theta$ is denoted $\text{conv}\Theta$. The set of Schur stable matrices is denoted $\mathcal{S}$.

Dynamics, reward function and policies  We are concerned with control of discrete linear time-invariant dynamical systems of the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}(0, \Pi),$$

where $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$, and $w_t \in \mathbb{R}^{n_w}$ denote the state, input, and unobserved exogenous disturbance at time $t$, respectively. Let $\theta := \{A, B, \Pi\}$. Our objective is to design a feedback control
policy \( u_t = \phi(x_t) \) that minimizes the cost function \( \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} E [x'_t Q x_t + u'_t R u_t] \), where \( x_t \) evolves according to \( (1) \), and \( Q \geq 0 \) and \( R > 0 \) are user defined weight matrices. A number of different parametrizations of the policy \( \phi \) have been considered in the literature, from neural networks (popular in RL, e.g., \[4\]) to causal (typically linear) dynamical systems (common in RC, e.g., \[32\]). In this paper, we will restrict our attention to static-gain policies of the form \( u_t = K x_t \), where \( K \in \mathbb{R}^{n_u \times n_x} \) is constant. As noted in \[14\], controller synthesis and implementation, is simpler (and more computationally efficient) for such policies. When the parameters of the true system, denoted \( \theta_u := \{ A_u, B_u, \Pi_u \} \), are known this is the infinite horizon LQR problem, the optimal solution of which is well-known \[5\]. We assume that \( \theta_u \) is unknown; rather, our knowledge of the dynamics must be inferred from observed sequences of inputs and states.

**Observed data** We adopt the data-driven setup used in \[14\], and assume that \( \mathcal{D} := \{ x_{0:T}^i, u_{0:T}^i \}_{i=1}^N \), where \( x_{0:T}^i = \{ x_t^i \}_{t=0}^T \) is the observed state sequence attained by evolving the true system for \( T \) time steps, starting from an arbitrary \( x_0 \) and driven by arbitrary input \( u_{0:T}^i = \{ u_t^i \}_{t=0}^T \). Each of these \( N \) independent experiments is referred to as a rollout. We perform parameter inference in the offline/batch setting; i.e., all data \( \mathcal{D} \) is assumed to be available at the time of controller synthesis.

**Optimization objective** Given observed data and, possibly, prior knowledge of the system, we then have the posterior distribution over the model parameters denoted \( \pi(\theta) := p(\theta|\mathcal{D}) \), in place of the true parameters \( \theta_t \). The function that we seek to minimize is the expected cost w.r.t. the posterior distribution, i.e.,

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} E [x'_t Q x_t + u'_t R u_t] \mid x_{t+1}^i = Ax_t + Bu_t + w_t, w_t \sim \mathcal{N}(0, \Pi), \{ A, B, \Pi \} \sim \pi(\theta) \].

(2)

In practice, the support of \( \pi \) almost surely contains \( \{ A, B \} \) that are unstabilizable, which implies that (2) is infinite. Consequently, we shall consider averages over confidence regions w.r.t. \( \pi \). For convenience, let us denote the infinite horizon LQR cost, for given system parameters \( \theta \), by

\[
J(K|\theta) := \lim_{t \to \infty} E [x'_t (Q + K' R K) x_t] \mid x_{t+1} = (A + BK)x_t + w_t, w \sim \mathcal{N}(0, \Pi)]
\]

(3a)

\[
= \begin{cases} 
\text{tr } X \Pi \text{ with } X = (A + BK)' X (A + BK) + Q + K' R K, & A + BK \in \mathcal{S} \\
\infty, & \text{otherwise,}
\end{cases}
\]

(3b)

where the second equality follows from standard Gramian calculations, and \( \mathcal{S} \) denotes the set of Schur stable matrices. As an alternative to (2) we consider the cost function \( J^c(K) := \int_{\Theta_c} J(K|\theta) \pi(\theta) d\theta \), where \( \Theta_c \) denotes the c % confidence region of the parameter space w.r.t. the posterior \( \pi \). Though better suited to optimization than (2), which is almost surely infinite, this integral cannot be evaluated in closed form, due to the complexity of \( J(K|\theta) \) w.r.t. \( \theta \). Furthermore, there is still no guarantee that \( \Theta_c \) contain only stabilizable models. To circumvent both of these issues, we propose the following Monte Carlo approximation of \( J^c(K) \),

\[
J^c_M(K) := \frac{1}{M} \sum_{i=1}^{M} J(K|\theta_i), \quad \theta_i \sim \Theta_c \cap \bar{\mathcal{S}}, \quad i = 1, \ldots, M,
\]

(4)

where \( \bar{\mathcal{S}} \) denotes the set of stabilizable \( \{ A, B \} \).

**Posterior distribution** Given data \( \mathcal{D} \), the parameter posterior distribution is given by Bayes’ rule:

\[
\pi(\theta) := p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\mathcal{D}|\theta)p(\theta) \propto p(\mathcal{D}) \prod_{r=1}^{N} \prod_{t=1}^{T} p(x_t^r|x_{t-1}^r, u_{t-1}^r, \theta) =: \tilde{\pi}(\theta),
\]

(5)

where \( p(\mathcal{D}) \) denotes our prior belief on \( \theta, p(x_t^r|x_{t-1}^r, u_{t-1}^r, \theta) = \mathcal{N}(A x_{t-1}^r + B u_{t-1}^r, \Pi), \) and \( \tilde{\pi} = p(\mathcal{D})\pi \) denotes the unnormalized posterior. To sample from \( \pi \), we can distinguish between the different cases. First, consider the case when \( \Pi_u \) is known or can be reliably estimated independently of \( \{ A, B \} \). This is the setting in, e.g., \[14\]. In this case, the likelihood can be equivalently expressed as a Gaussian distribution over \( \{ A, B \} \). Then, when the prior \( p(A, B) \) is uniform (i.e. non-informative) or Gaussian (self-conjugate), the posterior \( p(A, B|\mathcal{D}, \Pi_u) \) is also Gaussian, c.f. Appendix A.1.1. Second, consider the general case in which \( \Pi_u \), along with \( \{ A, B \} \), is unknown. In this setting, one can select from a number of methods adapted for Bayesian inference in dynamical systems, such as Metropolis-Hastings \[29\], Hamiltonian Monte Carlo \[12\], and Gibbs sampling \[13\] \[49\]. When one
places a non-informative prior on $\Pi$ (e.g., $p(\Pi) \propto \det(\Pi)^{-\frac{n_u+1}{2}}$), each iteration of a Gibbs sampler targeting $\pi$ requires sampling from either a Gaussian or an inverse Wishart distribution, for which reliable numerical methods exist; c.f., Appendix A.1.2. In both of these cases we can sample from $\pi$ and evaluate $\bar{\pi}$ point-wise. To draw $\theta_i \sim \Theta^c \cap \bar{\mathcal{S}}$, as in (4), we can first draw a large number of samples from $\pi$, discard the (100-$c$)% of samples with the lowest unnormalized posterior values, and then further discard any samples that happen to be unstabilizable. For convenience, we define $\bar{\Theta}_M := \{\{\theta_i\}_{i=1}^M : \theta_i \sim \Theta^c \cap \bar{\mathcal{S}}, i = 1, \ldots, M\}$, which should be interpreted as a set of $M$ realizations of this procedure for sampling $\theta_i \sim \Theta^c \cap \bar{\mathcal{S}}$.

**Summary** We seek the solution of the optimization problem $\min_K J_M^c(K)$ for $K \in \mathbb{R}^{n_u \times n_u}$.

### 4 Solution via semidefinite programing

In this section we present the principle contribution of this paper: a method for solving $\min_K J_M^c(K)$ via convex (semidefinite) programing (SDP). It is convenient to consider an equivalent representation

\[
\min_{K, \{X_i\}_{i=1}^M} \frac{1}{M} \sum_{i=1}^M \text{tr} \, X_i \Pi_i,
\]

subject to

\[
X_i \succeq (A_i + B_i K)'X_i(A_i + B_i K) + Q + K'R_K, \quad \theta_i \in \bar{\Theta}_M^c,
\]

where the Comparison Lemma [30, Lecture 2] has been used to replace the equality in (3b) with the inequality in (6b). We introduce the notation $\mathcal{S}_+ := \{S \in \mathbb{S}_+^n : S \succeq cI, S \preceq cI\}$, where $c$ and $c$ are arbitrarily small and large positive constants, respectively. $\mathcal{S}_+^c$ serves as a compact approximation of $\mathcal{S}_+^c$, suitable for use with SDP solvers, i.e., $S \in \mathcal{S}_+^c \implies S \in \mathcal{S}_+^c$.

#### 4.1 Common Lyapunov relaxation

The principle challenge in solving (6) is that the constraint (6b) is not jointly convex in $K$ and $X'$. The usual approach to circumventing this nonconvexity is to first apply the Schur complement (6b), and then conjugate by the matrix $\text{diag}(X_i^{-1}, I, I, I)$, which leads to the equivalent constraint

\[
\begin{bmatrix}
X_i^{-1} & (A_i + B_i K)'X_i^{-1} & X_i^{-1}Q^{1/2} & X_i^{-1}K' \\
(A_i + B_i K)X_i^{-1} & X_i^{-1} & 0 & 0 \\
Q^{1/2}X_i^{-1} & 0 & I & 0 \\
KX_i^{-1} & 0 & 0 & R^{-1}
\end{bmatrix} \succeq 0.
\]

With the change of variables $Y_i = X_i^{-1}$ and $L_i = KX_i^{-1}$, (7) becomes an LMI, in $Y_i$ and $L_i$. This approach is effective when $M = 1$ (i.e., we have a single nominal system, as in standard LQR). However, when $M > 1$ we cannot introduce a new $Y_i$ for each $X_i^{-1}$, as we lose uniqueness of the controller $K$ in $L_i = KX_i^{-1}$, i.e., in general $L_iY_i^{-1} \neq L_jY_j^{-1}$ for $i \neq j$. One strategy (prevalent in robust control, e.g., [14, §C]) is to employ a ‘common Lyapunov function’, i.e., $Y = X_i^{-1}$ for all $i = 1, \ldots, M$. This gives the following convex relaxation (upper bound) of problem (6).

\[
\min_{K, Y} \text{tr} \, Z_i,
\]

subject to

\[
\begin{bmatrix}
Z_i & G_i' \\
G_i & Y
\end{bmatrix} \succeq 0,
\]

\[
\begin{bmatrix}
Y & YA_i' + L'B_i' & YQ^{1/2} & L' \\
A_iY + B_i' & Y & 0 & 0 \\
Q^{1/2}Y & 0 & I & 0 \\
L & 0 & 0 & R^{-1}
\end{bmatrix} \succeq 0, \quad \theta_i \in \bar{\Theta}_M^c.
\]

where $G_i$ denotes the Cholesky factorization of $\Pi_i$, i.e., $\Pi_i = G_iG_i'$, and $\{Z_i\}_{i=1}^M$ are slack variables used to encode the cost (6a) with the change of variables, i.e.,

\[
\min_{Y} \text{tr} \, Y^{-1}\Pi_i \leq \min_{Z_i} \text{tr} \, Z_i \text{ s.t. } Z_i \succeq G_i'Y^{-1}G_i \iff \min_{Z_i} \text{tr} \, Z_i \text{ s.t. } \begin{bmatrix} Z_i & G_i' \\
G_i & Y
\end{bmatrix} \succeq 0.
\]

The approximation in (8) is highly conservative, which motivates the iterative local optimization method presented in Section 4.2. Nevertheless, (8) provides a principled way (i.e., a one-shot convex program) to initialize the iterative search method derived in Section 4.2.
4.2 Iterative improvement by sequential semidefinite programming

To develop this iterative search method first consider an equivalent representation of \( J(K|\theta_i) \),

\[
J(K|\theta_i) = \min_{X_i \in \mathbb{S}_+^n} \text{tr} \ X_i \Pi_i \quad \text{s.t.} \quad \left[ \begin{array}{c} X_i - Q_k \quad (A_i + B_i K)' \quad K' \\ A_i + B_i K \quad X_i^{-1} \quad 0 \\ K \quad 0 \quad R^{-1} \end{array} \right] \succeq 0, \tag{9a}
\]

This representation highlights the nonconvexity of \( J(K|\theta_i) \) due to the \( X_i^{-1} \) term, which was addressed (in the usual way) by a change of variables in Section 4.1. In this section, we will instead replace \( X_i^{-1} \) with a linear approximation and prove that this leads to a tight convex upper bound. Given \( S \in \mathbb{S}_+^n \), let \( T(S, S_0) \) denote the first order (i.e., linear) Taylor series approximation of \( S^{-1} \) about some nominal \( S_0 \in \mathbb{S}_+^n \), i.e., \( T(S, S_0) := S_0^{-1} + \frac{\partial S^{-1}}{\partial S} \bigg|_{S=S_0} (S-S_0) = S_0^{-1} - S_0^{-1} (S-S_0) S_0^{-1} \).

We now define the function

\[
\bar{J}_c(K_\star, \bar{K}) := \min_{X_i \in \mathbb{S}_+^n} \text{tr} \ X_i \Pi_i \quad \text{s.t.} \quad \left[ \begin{array}{c} X_i - Q_k \quad (A_i + B_i K)' \quad K' \\ A_i + B_i K \quad T(X_i, \bar{X}_i) \quad 0 \\ K \quad 0 \quad R^{-1} \end{array} \right] \succeq 0, \tag{10a}
\]

where \( \bar{X}_i \) is any \( X_i \in \mathbb{S}_+^n \) that achieves the minimum in (9a), with \( K = \bar{K} \) for some nominal \( K \), i.e., \( J(K|\theta_i) = \text{tr} \ X_i \Pi_i \). Analogously to (9b), we define

\[
\bar{J}_c^c(K, \bar{K}) := \frac{1}{M} \sum_{\theta_i \in \Theta_M} \bar{J}(K, \bar{K}|\theta_i). \tag{11}
\]

We now show that \( \bar{J}_c^c(K, \bar{K}) \) is a convex upper bound on \( J_c^c(K) \), which is tight at \( K = \bar{K} \). The proof is given in Appendix A.2.2 and makes use of the following technical lemma (c.f. A.2.1 for proof).

**Lemma 4.1.** \( T(S, S_0) \preceq S^{-1} \) for all \( S, S_0 \in \mathbb{S}_+^n \), where \( T(S, S_0) \) denotes the first-order Taylor series expansion of \( S^{-1} \) about \( S_0 \).

**Theorem 4.1.** Let \( \bar{J}_c^c(K, \bar{K}) \) be defined as in (11), with \( \bar{K} \) such that \( \bar{J}_c^c(K, \bar{K}) \) is finite. Then \( \bar{J}_c^c(K, \bar{K}) \) is a convex upper bound on \( J_c^c(K) \), i.e., \( \bar{J}_c^c(K, \bar{K}) \geq J_c^c(K) \) \( \forall K \). Furthermore, the bound is ‘tight’ at \( \bar{K} \), i.e., \( \bar{J}_c^c(K, \bar{K}) = J_c^c(K) \).

**Iterative algorithm** To improve upon the common Lyapunov solution given by (8), we can solve a sequence of convex optimization problems: \( K^{(k+1)} = \arg\min_K \bar{J}_c^c(K, K^{(k)}) \), c.f. Algorithm 1 for details. This procedure of optimizing tight surrogate functions in lieu of the actual objective function is an example of the ‘majorize-minimization (MM) principle’, a.k.a. optimization transfer [26]. MM algorithms enjoy good numerical robustness, and (with the exception of some pathological cases) reliable convergence to local minima [44]. Indeed, it is readily verified that \( \bar{J}_c^c(K^{(k)}) = \bar{J}_c^c(K^{(k)}, K^{(k)}) \geq \bar{J}_c^c(K^{(k+1)}, K^{(k)}) \geq J_c^c(K^{(k+1)}) \), where equality follows from tightness of the bound, and the second inequality is due to the fact that \( \bar{J}_c^c(K, K^{(k)}) \) is an upper bound. This implies that \( \{\bar{J}_c^c(K^{(k)})\}_{k=1}^{\infty} \) is a converging sequence.

**Remark 4.1.** This sequential SDP approach can be applied in other robust control settings, e.g., mixed \( H_2/H_\infty \) [17], to improve on the common Lyapunov solution, c.f., Section 5.1 for an illustration.

4.3 Robustness

Hitherto, we have considered the performance component of the robust control problem, namely minimization of the expected cost; we now address the robust stability requirement. It is desirable for the learned policy to stabilize every model in the confidence region \( \Theta^c \); in fact, this is necessary for the cost \( J^c(K) \) to be finite. Algorithm 1 ensures stability of each of the \( M \) sampled systems from \( \tilde{\Theta}_M \), which implies that \( \phi \) stabilizes the entire region as \( M \to \infty \). However, we would like to be able to say something about robustness for finite \( M \). To this end, we make two remarks. First, if closed-loop stability of each sampled model is verified with a common Lyapunov function, then the policy stabilizes the convex hull of the sampled systems:
Suppose there exists $\Theta^c \cap S$, i.e., $\Theta^c_M$, using the appropriate Bayesian inference method from Section 5.

Theorem 4.2. Suppose there exists $K \in \mathbb{R}^{n_s \times n_u}$ such that $(A_i + B_i K)'X(A_i + B_i K) - X < 0$ for $X > 0$ and all $\Theta = \{A_i, B_i\}_{i=1}^N$. Then $(A + BK)'X(A + BK) - X < 0$ for all $(A, B) \in \text{conv} \Theta$, where $\text{conv} \Theta$ denotes the convex hull of $\Theta$.

The proof of Theorem 4.2 is given in [14, 2]. The conditions of Theorem 4.2 hold for the common Lyapunov approach in [14, 2] and can be made to hold for Algorithm 1 by introducing an additional Lyapunov stability constraint (with common Lyapunov function) for each sampled system, at the expense of some conservatism. Second, we observe empirically that Algorithm 1 returns policies that very nearly stabilize the entire region $\Theta^c$, despite a very modest number of samples $M$ relative to the dimension of the parameter space, c.f., Section 5.1 in particular Figure 2. A number of recent papers have investigated sampling (or grid) based approaches to stability verification of control policies, e.g., [40, 14, 6]. Understanding why policies from Algorithm 1 generalize effectively to the entire region $\Theta^c$ is an interesting topic of future research.

5 Experimental results

5.1 Numerical simulations using synthetic systems

In this section, we study the infinite horizon LQR problem specified by

$$A_{tr} = \text{toeplitz}(a, a'), a = [1.01, 0.01, 0, \ldots, 0] \in \mathbb{R}^{n_s}, B_{tr} = I, \Pi_{tr} = I, Q = 10^{-3}I, R = I,$$

where $\text{toeplitz}(r, c)$ denotes the Toeplitz matrix with first row $r$ and first column $c$. This is the same problem studied in [14, §6] (for $n_s = 3$), where it is noted that such dynamics naturally arise in consensus and distributed averaging problems. To obtain problem data $D$, each rollout involves simulating [14], with the true parameters, for $T = 6$ time steps, excited by $u_t \sim N(0, I)$ with $x_0 = 0$. Note: to facilitate comparison with [14], we too shall assume that $\Pi_{tr}$ is known. Furthermore, for all experiments $\Theta^c$ will denote the 95% confidence region, as in [14]. We compare the following methods of control synthesis: existing methods: (i) nominal: standard LQR using the nominal model from the least squares, i.e., $\{A_{ls}, B_{ls}\} := \arg \min_{A,B} \sum_{s=1}^N \sum_{t=1}^T (x_t - Ax_{t-1} - Bu_{t-1})^2$; (ii) worst-case: optimize for worst-case model (95% confidence) s.t. robust stability constraints, i.e., the method of [14, §5.2]; (iii) $H_2/H_\infty$: enforce stability constraint from [14, §5.2], but optimize performance for the nominal model $\{A_{ls}, B_{ls}\}$; proposed method(s): (iv) CL: common Lyapunov relaxation of [8]; (v) proposed: the method proposed in this paper, i.e., Algorithm 1; additional new methods: (vi) alternate-r: initialize with the $H_2/H_\infty$ solution, and apply the iterative optimization method proposed in Section 4.2 c.f., Remark 4.1; (vii) alternate-s: optimize for the nominal model $\{A_{ls}, B_{ls}\}$, enforce stability for the sampled systems in $\tilde{\Theta}_M$. Before proceeding, we wish to emphasize that the different control synthesis methods have different objectives; a lower cost does not mean that the associated method is ‘better’. This is particularly true for worst-case which seeks to optimize performance for the worst possible model so as to bound the cost on the true system.

To evaluate performance, we compare the cost of applying a learned policy $K$ to the true system $\theta = \{A_{ls}, B_{ls}\}$, to the optimal cost achieved by the optimal controller $K_{lb}$ (designed using $\theta$), i.e., $J(K_\theta)/J(K_{lb})$. We refer to this as ‘LQR suboptimality.’ In Figure 1 we plot LQR suboptimality is shown as a function of the number of rollouts $N$, for $n_s = 3$. We make the following observations. Foremost, the methods that enforce stability ‘stochastically’ (i.e. point-wise), namely proposed and alternate-s, attain significantly lower costs than the methods that enforce stability
Table 1: Median % of unstable closed-loop models, with open-loop models sampled from the 95% confidence region of the posterior, for system of varying dimension $n_x$; c.f. Section 5.1 for details. Parenthesized quantities denote the % of cases for which the policy synthesis optimization problem was infeasible (i.e. no policy was returned). 50 experiments were conducted, with $N = 50$. $H_2/H_\infty$ and alternate-r have the same robustness guarantees as worst-case, and are omitted.

| $n_x$ | optimal | nominal | worst-case | CL | proposed | alternate-s |
|-------|---------|---------|------------|----|----------|-------------|
| 3     | 61.6 (0)| 28.75 (0)| 0 (0)      | 0.10 (0)| 1.35 (0) |
| 6     | 95.37 (0)| 58.41 (0)| 0 (0)      | 0.18 (0)| 1.76 (0) |
| 9     | 99.6 (0)| 81.9 (0) | 0 (0)      | 0.24 (0)| 1.40 (0) |
| 12    | 100 (0)| 94.28 (0)| 0 (46.0)  | 0 (0) | 1.27 (0) |

It is natural to ask whether the reduction in cost exhibited by proposed (and alternate-s) come at the expense of robustness, namely, the ability to stabilize a large region of the parameter space. Empirical results suggest that this is not the case. To investigate this we sample 5000 fresh (i.e. not used for learning) models from $\Theta \cap \bar{S}$ and check closed-loop stability of each; this is repeated for 50 independent experiments with varying $n_x$ and $N = 50$. The median percentage of models that were unstable in closed-loop is recorded in Table 1. We make two observations: (i) the proposed method exhibits strong robustness. Even for $n_x = 12$ (i.e. 288-dim parameter space), it stabilizes > 99% of samples from the confidence region, with only $M = 100$ MC samples. (ii) when the robust methods (worst-case, $H_2/H_\infty$, alternate-r) are feasible, the resulting policies were found to stabilize 100% of samples; however, for $n_x = 12$, the methods were infeasible almost half the time, whereas proposed always returned a policy. Further evidence is provided in Figure 2, which plots robustness and performance as a function of the number of MC samples, $M$. For $n_x = 3$ and $M \geq 800$, the entire confidence region is stabilized with very high probability, suggesting that $M \to \infty$ is not required for robust stability in practice.

5.2 Real-world experiments on a rotary inverted pendulum

We now apply the proposed algorithm to the classic problem of stabilizing a (rotary) inverted pendulum, on real (i.e. physical) hardware (Quanser QUBE 2), c.f. Section A.3 for details. To generate
Figure 2: (left) Median % of unstable closed-loop models, with open-loop models sampled from the 95% confidence region of the posterior, for \( n_x = 3 \) and \( N = 15 \), as a function of the number of samples \( M \) used in the MC approximation \( 4 \). (right) LQR suboptimality as a function of \( M \). 50 experiments were conducted, c.f. Section 5.1 for details. Shaded regions cover the interquartile range.

Figure 3: (a) LQR cost on real-world pendulum experiment, as a function of the number of rollouts. \( \infty \) cost denotes controllers that resulted in instability during testing. n/a denotes cases in which the synthesis problem was infeasible. (b) Pendulum angle and control signal recorded after 10 rollouts.

Data-inefficiency is a well-known limitation of RL; understanding and mitigating this inefficiency is the subject of considerable research \([14, 15, 18, 20, 21]\). Investigating the role that a Bayesian approach to uncertainty quantification plays in the apparent sample-efficiency of the proposed method is an interesting topic for further inquiry.
Acknowledgments

This research was financially supported by the Swedish Foundation for Strategic Research (SSF) via the project ASSEMBLE (contract number: RIT15-0012) and via the projects Learning flexible models for nonlinear dynamics (contract number: 2017-03807) and NewLEADS - New Directions in Learning Dynamical Systems (contract number: 621-2016-06079), both funded by the Swedish Research Council.

References

[1] P. Abbeel and A. Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In Proceedings of the 22nd international conference on Machine learning, pages 1–8. ACM, 2005.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

[3] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin. Provably safe and robust learning-based model predictive control. Automatica, 49(5):1216–1226, 2013.

[4] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In Advances in Neural Information Processing Systems (NIPS), 2017.

[5] D. P. Bertsekas. Dynamic programming and optimal control. Belmont, MA: Athena Scientific, 1995.

[6] R. Bobiti and M. Lazar. A sampling approach to finding lyapunov functions for nonlinear discrete-time systems. In Control Conference (ECC), 2016 European, pages 561–566. IEEE, 2016.

[7] R. Boczar, N. Matni, and B. Recht. Finite-data performance guarantees for the output-feedback control of an unknown system. arXiv preprint arXiv:1803.09186, 2018.

[8] J. B. Burl. Linear optimal control: $H_2$ and $H_\infty$ methods. Addison-Wesley Longman Publishing Co., Inc., 1998.

[9] C. K. Carter and R. Kohn. On gibbs sampling for state space models. Biometrika, 81(3):541–553, 1994.

[10] J. Chen and G. Gu. Control-oriented system identification: an $H_\infty$ approach, volume 19. Wiley-Interscience, 2000.

[11] J. Chen and C. N. Nett. The Caratheodory-Fejer problem and $H_2/H_\infty$ identification: a time domain approach. In Proceedings of the 32nd IEEE Conference on Decision and Control (CDC), pages 68–73, 1993.

[12] S. H. Cheung and J. L. Beck. Bayesian model updating using hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters. Journal of engineering mechanics, 135(4):243–255, 2009.

[13] J. Ching, M. Muto, and J. Beck. Bayesian linear structural model updating using Gibbs sampler with modal data. In Proceedings of the 9th International Conference on Structural Safety and Reliability, pages 2609–2616. Millpress, 2005.

[14] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. arXiv preprint arXiv:1701.01688, 2017.

[15] M. Deisenroth and C. E. Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In Proceedings of the 28th International Conference on machine learning (ICML-11), pages 465–472, 2011.

[16] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. arXiv:1710.07283, 2017.

[17] J. Doyle, K. Zhou, K. Glover, and B. Bodenheimer. Mixed $H_2$ and $H_{\infty}$ performance objectives. II. Optimal control. IEEE Transactions on Automatic Control, 39(8):1575–1587, 1994.

[18] J. Garcia and F. Fernández. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1):1437–1480, 2015.

[19] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. Journal of Artificial Intelligence Research, 24:81–108, 2005.
[20] S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, and S. Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. In International Conference on Learning Representations, 2017.

[21] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine. Continuous deep Q-learning with model-based acceleration. In International Conference on Machine Learning, pages 2829–2838, 2016.

[22] W. M. Haddad, D. S. Bernstein, and D. Mustafa. Mixed-norm $H_2/H_\infty$ regulation and estimation: The discrete-time case. Systems & Control Letters, 16(4):235–247, 1991.

[23] A. J. Helmicki, C. A. Jacobson, and C. N. Nett. Control oriented system identification: a worst-case/deterministic approach in $H_\infty$. IEEE Transactions on Automatic Control, 36(10):1163–1176, 1991.

[24] H. K. Khalil. Nonlinear systems. Prentice-Hall, New Jersey, 2(5):5–1, 1996.

[25] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. Journal of machine learning research, 4(Dec):1107–1149, 2003.

[26] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. Journal of computational and graphical statistics, 9(1):1–20, 2000.

[27] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. Machine learning, 49(2-3):267–290, 2002.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

[29] B. Ninness and S. Henriksen. Bayesian system identification via Markov chain Monte Carlo techniques. Automatica, 46(1):40–51, 2010.

[30] Oliveira, Mauricio de. MAE 280B: Linear Control Design., 2009.

[31] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot. Robust Constrained Learning-based NMPC enabling reliable mobile robot path tracking. The International Journal of Robotics Research, 35(13):1547–1563, 2016.

[32] I. R. Petersen, M. R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. IEEE Transactions on Automatic Control, 45(3):398–412, 2000.

[33] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook., 2012.

[34] I. Postlethwaite, M. C. Turner, and G. Herrmann. Robust control applications. Annual Reviews in Control, 31(1):27–39, 2007.

[35] C. E. Rasmussen. Gaussian processes in machine learning. In Advanced lectures on machine learning, pages 63–71. Springer, 2004.

[36] I. M. Ross, R. J. Proulx, and M. Karpenko. Unscented optimal control for space flight. In 24th International Symposium on Space Flight Dynamics, 2014.

[37] I. M. Ross, R. J. Proulx, M. Karpenko, and Q. Gong. Riemann–stieljes optimal control problems for uncertain dynamic systems. Journal of Guidance, Control, and Dynamics, 38(7):1251–1263, 2015.

[38] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.

[39] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Pannemerschelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. nature, 529(7587):484–489, 2016.

[40] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. arXiv preprint arXiv:1802.08334, 2018.

[41] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[42] S. Tu, R. Boczar, A. Packard, and B. Recht. Non-asymptotic analysis of robust control from coarse-grained identification. arXiv preprint arXiv:1707.04791, 2017.

[43] S. Tu and B. Recht. Least-squares temporal difference learning for the linear quadratic regulator. arXiv preprint arXiv:1712.08642, 2017.
[44] F. Vaida. Parameter convergence for EM and MM algorithms. *Statistica Sinica*, pages 831–840, 2005.

[45] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

[46] J. Vinogradska, B. Bischoff, D. Nguyen-Tuong, A. Romer, H. Schmidt, and J. Peters. Stability of controllers for gaussian process forward models. In *International Conference on Machine Learning*, pages 545–554, 2016.

[47] Y.-S. Wang, N. Matni, and J. C. Doyle. A system level approach to controller synthesis. *arXiv preprint arXiv:1610.04815*, 2016.

[48] A. Wills, T. B. Schön, F. Lindsten, and B. Ninness. Estimation of linear systems using a Gibbs sampler. *IFAC Proceedings Volumes*, 45(16):203–208, 2012.

[49] K. Zhou and J. C. Doyle. *Essentials of robust control*, volume 104. Prentice Hall Upper Saddle River, NJ, 1998.
A Supplementary material

A.1 Sampling from the posterior distribution

A.1.1 Case I: \( \Pi_{tr} \) known

First, consider the case where \( \Pi_{tr} \) is known; i.e., \( \theta = \{ A, B \} \). From Bayes’ rule, we have

\[
\pi(\theta) := p(\theta|D) = \frac{1}{p(D)} p(D|\theta) p(\theta) \propto p(\theta) \prod_{i=1}^{N} p(x^i,|x^i, u^i, \theta) =: \hat{\pi}(\theta),
\]

where \( D = \{ x^i_+, x^i_-, u^i \}_{i=1}^{N} \) and \( p(x^i_+, x^i_-, u^i, \theta) = \mathcal{N}(x^i_+ - Ax^i_- - Bu^i, \Pi) \). When \( \Pi \) is known, we can express the likelihood \( p(D|\theta) \) in a form equivalent to an (un-normalized) Gaussian distribution over \( \theta \), i.e.,

\[
\prod_{i=1}^{N} p(x^i_+|x^i_-, u^i, \theta) \quad (13a)
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_i |x^i_+ - Ax^i_- - Bu^i|^2_{\Pi^{-1}} \right)
\]

\[
= \exp \left( -\frac{1}{2} \sum_i |x^i_+ - D^i_{x,u} \theta_{AB}|^2_{\Pi^{-1}} \right)
\]

\[
= \exp \left( -\frac{1}{2} \left( \theta_{AB} \left( \sum_i D^i_{x,u} \Pi^{-1} D^i_{x,u} \right) \theta_{AB} - 2\theta_{AB} \sum_i D^i_{x,u} \Pi^{-1} x^i_+ + \sum_i x^i_+ \Pi^{-1} x^i_+ \right) \right)
\]

\[
\propto \mathcal{N}(\mu, \Sigma),
\]

where \( \Sigma = \left( \sum_i D^i_{x,u} \Pi^{-1} D^i_{x,u} \right)^{-1} \), \( \mu = \Sigma \left( \sum_i D^i_{x,u} \Pi^{-1} x^i_+ \right) \), \( \theta_{AB} = \text{vec}(A', B') \), and \( D^i_{x,u} = I_{n_x} \otimes [x^i_+, u^i] \). This implies that \( \pi(\theta) = \mathcal{N}(\mu, \Sigma) p(\theta) \). Therefore, when the prior \( p(\theta) \) is non-informative (\( p(\theta) \propto 1 \)) or Gaussian (self-conjugate), the posterior is also Gaussian.

A.1.2 Case II: \( \Pi_{tr} \) unknown

Next, consider the generic case in which all parameters are unknown. Then \( \theta = \{ A, B, \Pi \} \). One approach to sampling from the posterior involves Gibbs sampling [9], i.e., alternating between the following two sampling steps:

\[
\{ A_k, B_k \} \sim p(A, B|\Pi_{k-1}, D),
\]

\[
\Pi_k \sim p(\Pi|A_k, B_k, D)
\]

to form the Markov Chain \( \{ A_k, B_k, \Pi_k \}_{k=1}^{\infty} \). As demonstrated in [A.1], the distribution \( p(A, B|\Pi_{k-1}, D) \) is Gaussian, so sampling is straightforward. To sample from \( p(\Pi|A_k, B_k, D) \), first note

\[
p(\Pi|A, B, D) \propto p(D|A, B, \Pi)p(\Pi).
\]

Observe that

\[
p(D|A, B, \Pi) \propto \frac{1}{\det(\Pi)^{\frac{n}{2}}} \exp \left( -\frac{1}{2} \sum_i |x^i_+ - Ax^i_- - Bu^i|^2_{\Pi^{-1}} \right)
\]

\[
= \frac{1}{\det(\Pi)^{\frac{n}{2}}} \exp \left( -\frac{1}{2} \text{tr} \Phi \Pi^{-1} \right), \quad \Phi := \sum_i (x^i_+ - Ax^i_- - Bu^i)(x^i_+ - Ax^i_- - Bu^i)'
\]

\[
\propto \mathcal{W}^{-1}(\Phi, \nu), \nu = N - n_x - 1,
\]

where \( \mathcal{W}^{-1}(\cdot, \cdot) \) denotes the inverse Wishart distribution. Note, if \( N \leq n_x + 1 \) then \( \nu \) is not valid. However, we may consider a prior \( p(\Pi) \) such as \( p(\Pi) \propto \det(\Pi)^{\frac{-n_x-1}{2}} \) (Jeffreys’ noninformative prior) which means

\[
p(\Pi|A, B, D) \propto \frac{1}{\det(\Pi)^{\frac{n + \nu}{2}}} \exp \left( -\frac{1}{2} \text{tr} \Phi \Pi^{-1} \right) \propto \mathcal{W}^{-1}(\Phi, \nu),
\]

where \( \nu = N > 0 \). This is a well-defined inverse Wishart distribution, sampling from which is straightforward.

A.2 Proofs

A.2.1 Proof of Lemma [4, 1]

Lemma. \( T(S, S_0) \preceq S^{-1} \) for all \( S, S_0 \in \mathbb{S}^n_{++} \), where \( T(S, S_0) \) denotes the first-order Taylor series expansion of \( S^{-1} \) about \( S_0 \).
Proof. Let \( D = S - S_0 = L'L \), i.e., \( L \) is the Cholesky factorization of \( D \). Then,

\[
S_0 > 0 \iff S_0^{-1} > 0 \iff LS_0^{-1}L' \geq 0 \iff I + LS_0^{-1}L' \geq I \iff (I + LS_0^{-1}L')^{-1} \geq I
\]

\[
S_0^{-1}L'(I + LS_0^{-1}L')^{-1}LS_0^{-1} \equiv S_0^{-1}L'LS_0^{-1}
\]

\[
(\hat{S}_0 + L'L)^{-1} \geq S_0^{-1} - S_0^{-1}(S - S_0)S_0^{-1} \iff S^{-1} \geq T(S, S_0),
\]

where the penultimate implication invokes the Woodbury matrix inversion identity [33, eq. 159].

A.2.2 Proof of Theorem 4.1

Theorem. Let \( \hat{J}_M(K, \hat{K}) \) be defined as in (11), with \( \hat{K} \) such that \( \hat{J}_M(K) \) is finite. Then \( \hat{J}_M(K, \hat{K}) \) is a convex upper bound on \( J_M(K) \), i.e., \( \hat{J}_M(K, \hat{K}) \geq J_M(K) \forall K \). Furthermore, the bound is ‘tight’ at \( \hat{K} \), i.e., \( \hat{J}_M(K, \hat{K}) = J_M(K) \).

Proof. We will prove that \( J(K, \hat{K}; \theta_i) \) is a tight convex bound on \( J(K; \theta_i) \), as this implies that \( \hat{J}_M(K, \hat{K}) := \frac{1}{2} \sum_i J(K, \hat{K}; \theta_i) \) is a tight convex bound on \( J_M(K) := \frac{1}{2} \sum_i J(K; \theta_i) \). First, we prove convexity. \( J(K, \hat{K}; \theta_i) \) is defined as the supremum over an infinite family of convex functions over a compact convex set, and is therefore itself a convex function. Note: the minimum of a linear function, e.g. \( \min_{X_i \in \mathbb{S}^n_{++}} \text{tr} \ X_i G_i G_i' \) can be trivially expressed as the supremum of a convex function, i.e., \( \sup_{X_i \in \mathbb{S}^n_{++}} - \text{tr} \ X_i G_i G_i' \). Next, we prove the upper bound. From Lemma 4.1, \( X_i^{-1} \geq T_i(X, \bar{X}) \) for all \( X_i \in \mathbb{S}^n_{++} \). Therefore, any \( X_i \in \mathbb{S}^n_{++} \) that satisfies (10b) also satisfies (9b). This means the feasible set of (10) is a subset of the feasible set of (9), hence \( J(K, \hat{K}; \theta_i) \geq J(K; \theta_i) \). Finally, we prove tightness. As we have already proved \( J(K, \hat{K}; \theta_i) \geq J(K; \theta_i) \), it suffices to prove that \( \bar{X}_i \) is a feasible solution to (10b). As \( T(\bar{X}_i, \bar{X}) = \bar{X}_i^{-1} \), for \( K = \hat{K} \) and \( X_i = \bar{X}_i \), (10b) is equivalent to (9b), which is feasible by definition of \( \bar{X}_i \). Hence, \( \bar{X}_i \) is a feasible solution of (10), that achieves \( \text{tr} \ X_i G_i G_i' = J(K; \theta_i) \), by definition of \( \bar{X}_i \).

A.2.3 Proof of Theorem 4.2

Theorem. Suppose there exists \( K \in \mathbb{R}^{n \times n} \) such that \( (A_i + B_i K)'X(A_i + B_i K) - X < 0 \) for \( X > 0 \) and all \( \Theta = \{A_i, B_i\}_{i=1}^{n} \). Then \( (A + BK)'X(A + BK) - X < 0 \) for all \( \{A, B\} \in \text{conv}\Theta \), where \( \text{conv}\Theta \) denotes the convex hull of \( \Theta \).

Proof. It is sufficient to show that \( (A + BK)'X(A + BK) - X < 0 \) defines a convex set in terms of \( (A, B) \). By the Schur complement,

\[
(A + BK)'X(A + BK) - X < 0 \iff \begin{bmatrix} X & (A + BK)' \\ (A + BK) & X^{-1} \end{bmatrix} > 0,
\]

which is convex in \( A, B \) for given (i.e. fixed) \( K \) and \( X \).

A.3 Additional material for experiments on the rotary inverted pendulum

A.3.1 System description

The Quanser QUBE-Servo 2 inverted pendulum is depicted in Figure 4. The system consists of an actuated arm that rotates in the horizontal plane; actuation is provided by an electrical motor. Attached to the end of the rotary arm is an un-actuated pendulum, which is free to rotate. The voltage applied to the electric motor constitutes the control input \( u \) for the LQR problem. Rotary encoders record the angular position of the rotary arm and pendulum, denoted \( \alpha_u \) and \( \alpha_p \), respectively. These angular positions are fed through a high-pass-filter to provide angular velocity estimates, \( (\dot{\alpha}_u, \dot{\alpha}_p) \). The observed state is then given by \( \mathbf{x} = [\alpha_u, \alpha_p, \dot{\alpha}_u, \dot{\alpha}_p] \).

A.3.2 Experimental procedure

To generate one rollout of problem data, we first swing-up the pendulum to the inverted position, stabilized by an LQR designed using a physics-based model of the system. Then we apply the voltage signal \( v_t = K x_t + \sin(\omega t) + w_t \) and record the resulting state evolution (sampled at 100Hz), until the pendulum angle \( \alpha_p \) exceeds \( 20^\circ \), or the rotary arm angle \( \alpha_u \) exceeds \( 50^\circ \). Here \( K = [1, 10, 1, -3] \) constitutes a state feedback policy that does not stabilize the system, but does keep the pendulum upright from slightly longer than if it were absent. This extends the typical rollout duration to around 3-5 seconds, before the pendulum angle exceeds \( 20^\circ \). The angular frequency \( \omega \) is randomly sampled each rollout, \( \omega \sim U[20, 35] \), where \( U[a, b] \) denotes the uniform distribution over the interval \( [a, b] \). Finally, \( w_t \) denotes band-limited white noise, with a sampling time
of $T_s = 0.01$, and a gain of 0.05. $w_t$ represents additional exogenous disturbances that we artificially introduce to the system.

Training data then consists of $\{x_t, u_t\}_{t=0}^T$, where $x_t$ denotes the recorded state sequence, and $u_t = Kx_t + \sin(\omega t)$, i.e., the disturbance $w_t$ is not observed for learning. $T$ is truncated to 500 (i.e. 5 seconds) in the event that the rollout lasts this long. A typical rollout is depicted in Figure 5.

Figure 4: The Quanser QUBE-Servo 2 rotary pendulum, in the inverted position. Photo: www.quanser.com/products/qube-servo-2.

A.3.3 Bounds for robust methods

The robust synthesis methods worst-case, $H_2/H_\infty$, and alternate-$r$ require bounds on the error of the least squares estimate, i.e., $\epsilon_A \geq \|A_{ls} - A_{tr}\|_2$ and $\epsilon_B \geq \|B_{ls} - B_{tr}\|_2$. In [14], these bounds are estimated, with a specific confidence level, via a Bootstrap algorithm, assuming that the covariance is known. In our setting, we estimate these bounds as described in Section 5.2, i.e., by sampling from the 95% confidence region of the posterior distribution. This ensures a fair comparison between the methods, as they are, in essence, required to stabilize the same region of the parameter space. We observed, however, that these bounds on the least squares error were too conservative; the magnitude of the uncertainty was too large, and the control synthesis optimization problems were infeasible. The experiments presented in Figure 3 were attained by scaling down these error bounds by a factor of 100. A number of scaling factors were tested, but 100 was found to achieve a reasonable trade-off between robustness and feasibility. It is worth emphasizing that the proposed method used the samples from 95% confidence region without any such scaling.
Figure 5: A typical rollout from the experimental procedure in A.3.2.