# Robust exploration in linear quadratic reinforcement learning

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

UPPSALA UNIVERSITET

KTH VETENSKAP OCH KONST

SWEDISH FOUNDATION *for* STRATEGIC RESEARCH

## Summary and contributions

This work is concerned with the problem of minimizing the worst-case quadratic cost for an uncertain linear dynamical system. We derive:

- a high-probability bound on the spectral norm of the system parameter estimation error
- exact convex formulation of worst-case infinite-horizon LQR
- a (convex) algorithm that balances the exploration/exploitation trade-off by performing robust, targeted exploration.

## Problem statement

We are concerned with control of linear time-invariant systems

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

The true parameters $\{A_{\text{tr}}, B_{\text{tr}}\}$ are unknown, and must be inferred from data, $\mathcal{D}_n := \{x_t, u_t\}_{t=1}^n$. We assume: i) that $\sigma_w$ is known, and ii) we have access to initial data $\mathcal{D}_0$, obtained, e.g. during a preliminary experiment.

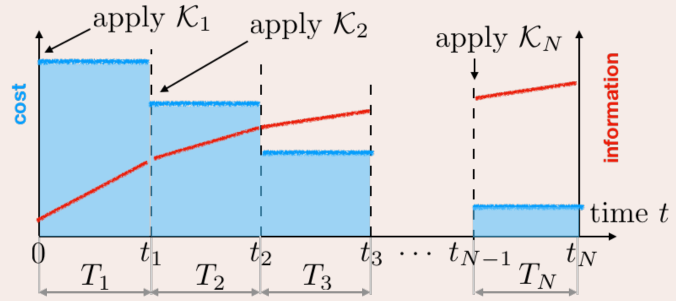
The posterior distribution $p(\theta|\mathcal{D}_n)$ is given by $\mathcal{N}(\mu_\theta, \Sigma_\theta)$, for $\theta = \text{vec}([A\ B])$ and a uniform prior $p(\theta) \propto 1$. This gives a high-probability elliptical credibility region:

$$\Theta_e(\mathcal{D}_n) := \{\theta \ : \ (\theta - \mu_\theta)^\top \Sigma_\theta{}^{-1}\theta - \mu_\theta) \le c_\delta\}. \qquad (2)$$

Static-gain policies: $u_t = Kx_t + \Sigma^{1/2} e_t$, where $e_t \sim \mathcal{N}(0, I)$.

'Robust reinforcement learning' (RRL) problem:

$$\min_{\{\mathcal{K}_i\}_{i=1}^N} \mathbb{E}\left[\sum_{t=0}^{T} \sup_{\{A_t, B_t\} \in \Theta_e(\mathcal{D}_t)} c(x_t, u_t)\right], \text{ s.t. } x_{t+1} = A_t x_t + B_t u_t + w_t, \qquad (3)$$




## Modeling uncertainty

We will work with models of the form $\mathcal{M}(\mathcal{D}) := \{\hat{A}, \hat{B}, D\}$ where $D \in \mathbb{S}^{n_x + n_u}$ specifies the following region centered about $\{\hat{A}, \hat{B}\}$:

$$\Theta_m(\mathcal{M}) := \{A, \ B \ : \ X^\top D X \preceq I, \ X = [\hat{A} - A, \ \hat{B} - B]^\top\} \qquad (4)$$

The following lemma suggests a specific means of constructing $D$, so as to ensure that $\Theta_m$ defines a high probability credibility region:

**Lemma 1.** *Given data $\mathcal{D}_n$ from (1), and $0 < \delta < 1$, let $D = \frac{1}{\sigma_w^2 c_\delta} \sum_{t=1}^{n-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$, with $c_\delta = \chi^2_{n_x^2 + n_x n_u}(\delta)$. Then $[A_{\text{tr}}, \ B_{\text{tr}}] \in \Theta_m(\mathcal{M})$ w.p. $1 - \delta$.*

## Approximate robust reinforcement learning problem

Consider the following approximation of (3),

$$\sum_{i=1}^{N} \sup_{\substack{\{A,B\} \in \\ \Theta_m(\mathcal{M}(\mathcal{D}_{t_i}))}} \mathbb{E}\left[\sum_{t=t_{i-1}}^{t_i} c(x_t, u_t)\right], \text{ s.t. } x_{t+1} = Ax_t + Bu_t + w_t, \ u_t = \mathcal{K}_i(x_t). \qquad (5)$$

- update the 'worst-case' model at the beginning of each epoch, when we deploy a new policy, rather than at each time step.
- select the worst-case model from $\Theta_m$ rather than $\Theta_e$.

We approximate the above with the infinite-horizon cost, appropriately scaled:

$$\mathbb{E}\left[\sum_{i=1}^{N} T_i \times J_\infty(\mathcal{K}_i, \Theta_m(\mathcal{M}(\mathcal{D}_{t_i})))\right]. \qquad (6)$$

## Convex optimization of infinite horizon cost

The infinite horizon cost can be expressed as

$$\lim_{\tau\to\infty} \frac{1}{\tau}\mathbb{E}\left[\sum_{t=1}^{\tau} x_t^\top Q x_t + u_t^\top R u_t\right] = \text{tr}\left(\begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}\left[\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top\right]\right). \qquad (7)$$

For known $A$ and $B$ the covariance $W = \mathbb{E}[x_t x_t^\top]$ satisfies:

$$W \succeq [A\ B] \begin{bmatrix} W & WK^\top \\ KW & KWK^\top + \Sigma \end{bmatrix} [A\ B]^\top + \sigma_w^2 I. \qquad (8)$$

We introduce the change of variables $Z = WK^\top$ and $Y = KWK^\top + \Sigma$, collated in the variable $\Xi = \begin{bmatrix} W & Z \\ Z^\top & Y \end{bmatrix}$. With this change of variables, minimizing (7) subject to (8) is a convex program. To ensure that (8) holds for all $\{A, B\} \in \Theta_m(\mathcal{M})$ we have:

$$S(\lambda, \Xi, \hat{A}, \hat{B}, D) := \begin{bmatrix} I & \sigma_w I & 0 \\ \sigma_w I & W - [\hat{A}\ \hat{B}]\Xi[\hat{A}\ \hat{B}]^\top - \lambda I & [\hat{A}\ \hat{B}]\Xi^\top \\ 0 & \Xi[\hat{A}\ \hat{B}]^\top & \lambda D - \Xi \end{bmatrix} \succeq 0. \qquad (9)$$

**Theorem 1.** *The problem $\min_{\mathcal{K}} J_\infty(\mathcal{K}, \Theta_m(\mathcal{M}))$ can be solved by the convex SDP:*

$$\min_{\lambda, \Xi} \ \texttt{tr}\ (\texttt{blkdiag}(Q, R)\Xi), \text{ s.t. } S(\lambda, \Xi, \hat{A}, \hat{B}, D) \succeq 0, \ \lambda \ge 0, \qquad (10)$$

*with the optimal policy given by $\mathcal{K} = \{Z^\top W^{-1}, \ Y - Z^\top W^{-1} Z\}$.*

## Propagating uncertainty

Define the approximate model, at time $t = t_j$ given data $\mathcal{D}_{t_i}$, by

$$\tilde{\mathcal{M}}_j(\mathcal{D}_{t_i}) := \{\tilde{A}_{j|i}, \tilde{B}_{j|i}, \tilde{D}_{j|i}\} \approx \mathbb{E}[\mathcal{M}(\mathcal{D}_{t_j})|\mathcal{D}_{t_i}].$$

Recall that: $D_{i+1} = D_i + \frac{1}{\sigma_w^2 c_\delta}\sum_{t=t_i}^{t_{i+1}} \begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$. We use the approximation:

$$\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} \begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top\right] \approx T_{i+1}\begin{bmatrix} W_i & W_i K_i^\top \\ K_i^\top W_i & K_i W_i K_i^\top + \Sigma_i \end{bmatrix} = T_{i+1}\Xi_i. \qquad (11)$$

To preserve convexity in our formulation, we approximate future nominal parameter estimates with the current estimates: given data $\mathcal{D}_{t_i}$ we set $\tilde{A}_{j|i} = \hat{A}_i$ and $\tilde{B}_{j|i} = \hat{B}_i$.

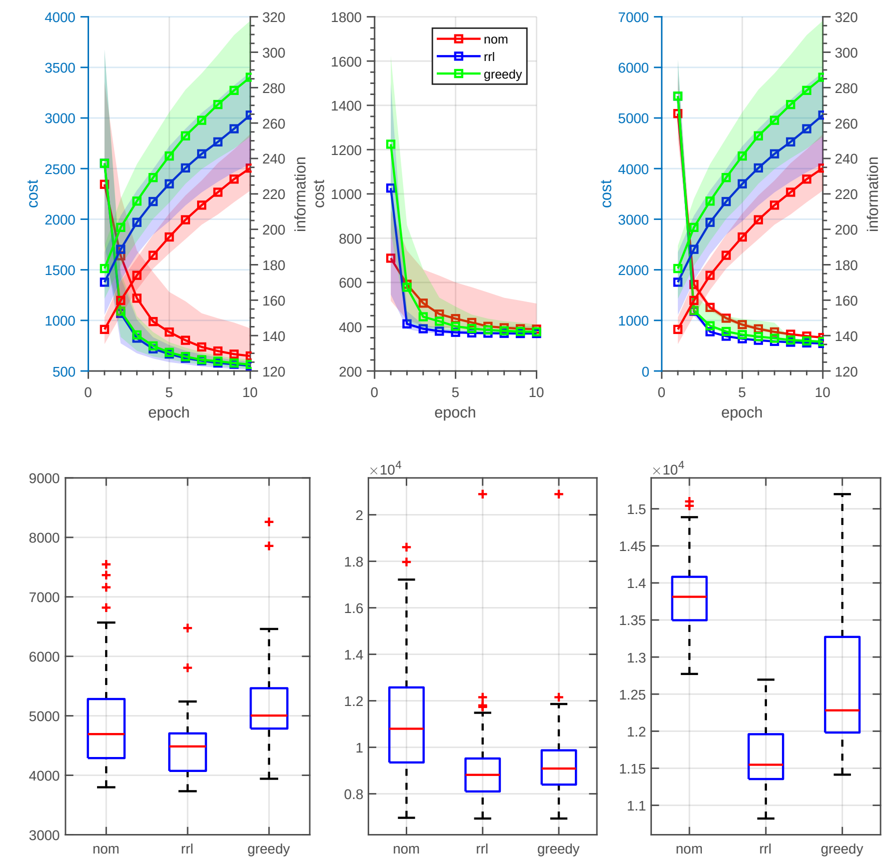
## Algorithm

**Algorithm 1** Receding horizon application to true system

1: **Input:** initial data $\mathcal{D}_0$, confidence $\delta$, LQR cost matrices $Q$ and $R$, epochs $\{t_i\}_{i=1}^N$.
2: **for** $i = 1 : N$ **do**
3: Compute/update nominal model $\mathcal{M}(\mathcal{D}_{t_{i-1}})$.
4: Solve convex program.
5: Recover policy $\mathcal{K}_i$: $K_i = Z_i^\top W_i^{-1}$ and $\Sigma_i = Y_i - Z_i^\top W_i^{-1} Z_i$.
6: Apply policy to true system for $t_{i-1} < t \le t_i$, which evolves according to (1) with $u_t = K_i x_t + \Sigma_i^{1/2} e_t$.
7: Form $\mathcal{D}_{t_i} = \mathcal{D}_{t_{i-1}} \cup \{x_{t_{i-1}:t_i}, u_{t_{i-1}:t_i}\}$ based on newly observed data.
8: **end for**

## Numerical simulations

$$A_{\text{tr}} = \begin{bmatrix} 1.1 & 0.5 & 0 \\ 0 & 0.9 & 0.1 \\ 0 & -0.2 & 0.8 \end{bmatrix}, \ B_{\text{tr}} = \begin{bmatrix} 0 & 1 \\ 0.1 & 0 \\ 0 & 2 \end{bmatrix}, \ \sigma_w = 0.5.$$

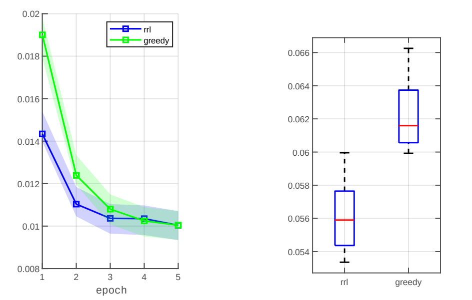



Information is defined as $1/\lambda_{\max}(D_i^{-1})$, at the $i$th epoch, which is the (inverse) of the 2-norm of parameter error, cf. (4). The larger the information, the more certain the system (in an absolute sense).

## Hardware-in-the-loop experiment

Interconnection of:

- a physical servo-mechanism (Quanser Qube)
- a synthetic (simulated) LTI system.




**Paper on arxiv:** Robust exploration in linear quadratic reinforcement learning.