Learning convex bounds for linear quadratic control policy synthesis

Jack Umenberger and Thomas B. Schön

UPPSALA UNIVERSITET

Swedish Foundation *for* Strategic Research

Summary and contributions

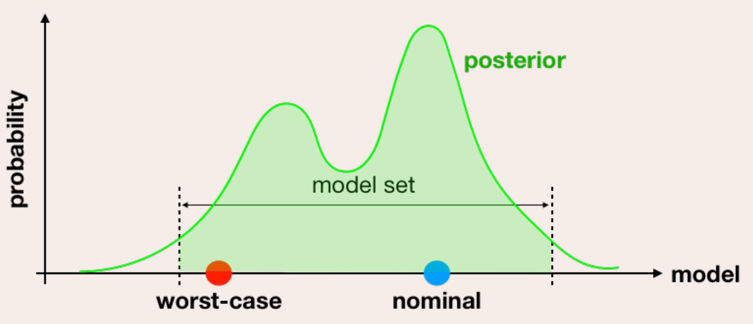
This work concerns the problem of learning control policies for unknown linear dynamical systems so as to optimize a quadratic reward.

We present a method to optimize the expected value of the reward over the posterior distribution of the unknown system parameters, given data.

- we build convex upper bounds on the expected cost.
- algorithm proceeds via sequential convex programing.
- strong performance and robustness properties are observed during numerical simulations and stabilization of a real-world inverted pendulum.

Background

Given (i) a cost function to minimize and (ii) data from an unknown dynamical system there are a number of ways to design a control policy.



- **certainty equivalence:** fit a nominal model to the data, and solve the problem as if the true system behaved exactly as the model.
- **robust control:** design a controller to stabilize a set of models; optimize performance for nominal or worst-case model.
- **probabilistic robust control:** optimize for expected performance given a posterior belief over models.

Problem setup

Dynamics and cost

We consider linear time-invariant dynamics:

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

Let $\theta := \{A, B, \Pi\}$.

The parameters θ are unknown.

We seek a static state-feedback policy $u_t = Kx_t$ that minimizes the cost function $\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\left[x_t'Qx_t + u_t'Ru_t\right]$ for given Q and R.

Observed data

We assume access to observed trajectories from the true system:

$$\mathcal{D} := \{x^r_{0:T}, u^r_{0:T}\}_{r=1}^{N}$$

Each of the N independent experiments is referred to as a rollout.

Parameter posterior

Given data $\mathcal{D}$ and a prior over parameters $p(\theta)$, the posterior distribution can be expressed by Bayes' rule:

$$\pi(\theta) := p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\mathcal{D}|\theta)p(\theta)$$

$$\propto p(\theta)\prod_{r=1}^{N}\prod_{t=1}^{T} p(x_t^r | x_{t-1}^r, u_{t-1}^r, \theta)$$

Sampling from posterior

Known Π and non-informative or Gaussian prior $\rightarrow$ posterior $p(\theta|\mathcal{D})$ is also Gaussian.

Unknown Π $\rightarrow$ posterior lacks a 'convenient' closed form.

We can generate samples from $p(\theta|\mathcal{D})$ using Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling, which alternates between:

$$\{A_k, B_k\} \sim p(A, B|\Pi_{k-1}, \mathcal{D}),$$
$$\Pi_k \sim p(\Pi|A_k, B_k, \mathcal{D})$$

The distribution $p(A, B|\Pi_{k-1}, \mathcal{D})$ is Gaussian $\rightarrow$ sampling is straightforward.

$p(\Pi|A, B, \mathcal{D})$ is an inverse Wishart distribution $\rightarrow$ sampling is straightforward.

Optimization objective

We seek to minimize the expected cost w.r.t. the posterior distribution,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\left[x_t'Qx_t + u_t'Ru_t \mid x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0,\Pi),\ \{A, B, \Pi\} \sim \pi(\theta)\right].$$

For convenience: denote the infinite horizon LQR cost, for given system parameters θ, by

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

$$= \begin{cases} \text{tr } X\Pi \text{ with } X = (A+BK)'X(A+BK) + Q + K'RK, & A + BK \text{ stable} \\ \infty, & \text{otherwise,} \end{cases}$$

More appropriate: integrate over some c % confidence region Θ^c of the posterior:

$$J^c(K) := \int_{\Theta^c} J(K|\theta)\pi(\theta)d\theta.$$

We approximate this integral with Monte Carlo:

$$J^c_M(K) := \frac{1}{M}\sum_{i=1}^{M} J(K|\theta_i), \qquad \{\theta_i\}_{i=1}^{M} \sim \Theta^c,$$

Common Lyapunov relaxation

By the Schur complement, $J(K|\theta_i)$ can be expressed as:

$$J(K|\theta_i) = \min_{X_i \in \mathbb{S}^{n_x}_{\epsilon}} \text{ tr } X_i\Pi_i$$

$$\text{s.t.} \begin{bmatrix} X_i^{-1} & X_i^{-1}(A_i + B_iK)' & X_i^{-1}Q^{1/2} & X_i^{-1}K' \\ (A_i + B_iK)X_i^{-1} & X_i^{-1} & 0 & \\ Q^{1/2}X_i^{-1} & 0 & I & 0 \\ KX_i^{-1} & 0 & 0 & R^{-1} \end{bmatrix} \succeq 0.$$

For $M = 1$ (one system) the usual trick is a change of variables $Y_i = X_i^{-1}$ and $L_i = KX_i^{-1}$.

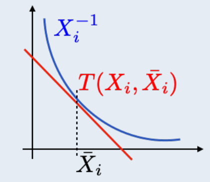
When $M > 1$ this not effective as we lose uniqueness of the controller K in $L_i = KX_i^{-1}$.

Convex upper bound

The cost for a single model θ_i is also given by:

$$J(K|\theta_i) = \min_{X_i \in \mathbb{S}^{n_x}_{\epsilon}} \text{ tr } X_i\Pi_i$$

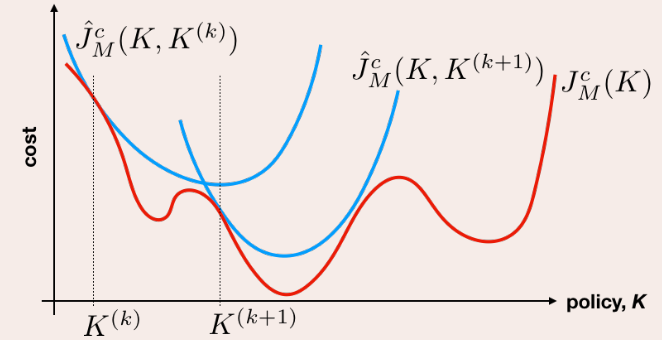
$$\text{s.t.} \begin{bmatrix} X_i - Q & (A_i + B_iK)' & K' \\ A_i + B_iK & X_i^{-1} & 0 \\ K & 0 & R^{-1} \end{bmatrix} \succeq 0.$$



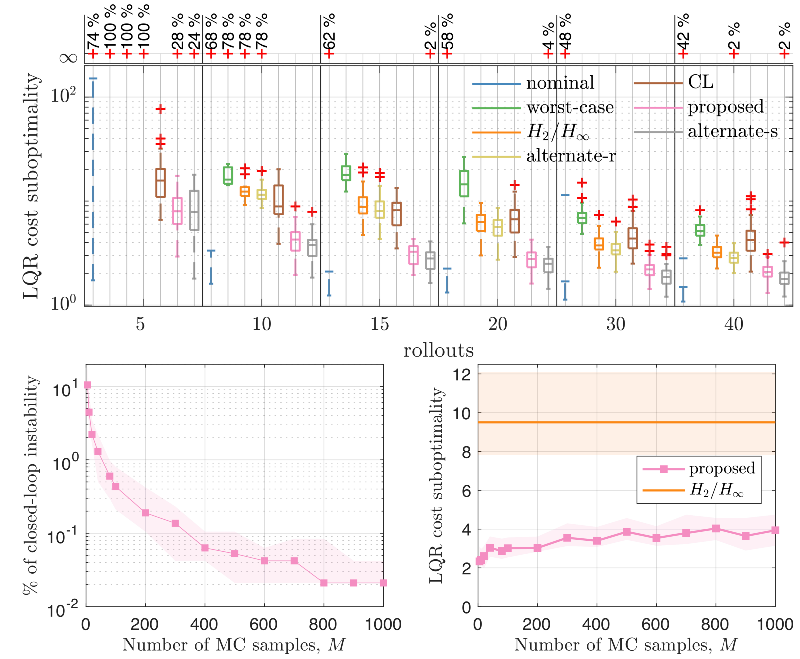
Substituting $T(X_i, \bar{X}_i)$ into $J(K|\theta_i)$ gives $\hat{J}(K, \bar{K}|\theta_i)$.

Theorem: $\hat{J}(K, \bar{K}|\theta_i)$ is a convex upper bound on $J(K|\theta_i)$, tight at $K = \bar{K}$.

Iterative algorithm



Simulation studies



Control of inverted pendulum