Near-Optimal Regret for the Safe Learning-based Control of the Constrained Linear Quadratic Regulator

Spencer Hutchinson
Coauthors: Nanfei Jiang, Mahnoosh Alizadeh

Department of Electrical and Computer Engineering
University of California, Santa Barbara

April 24, 2026

Warm Up: Unconstrained Adaptive LQR

Unconstrained Adaptive LQR with Regret

In each time step t:

- Observe state x_t.
- Choose input u_t.
- Incur cost $\ell(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$.
- State evolves: $x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \text{iid}$

System (A, B) is *unknown*. Aim to minimize *regret* against best linear controller,

$$R_T = \sum_{t=1}^{T} \ell(x_t, u_t) - J^\star,$$

State-of-the-art is $\tilde{\mathcal{O}}(\sqrt{T})$ regret bounds in this setting.[1]

[1] Abbasi-Yadkori & Szepesvari (2011), Simchowitz & Foster (2020)

Prior Work on Constrained Adaptive Control

Adaptive MPC[2]

- Optimizes policy subject to constraints over prediction horizon.
- Theoretical guarantees include constraint satisfaction (via recursive feasibility) and *asymptoticasymptotic* cost bounds.

Constrained LQR with Regret Bounds[3]

- $\tilde{\mathcal{O}}(T^{2/3})$ regret and satisfaction of *robust constraints*:

$$\alpha_j^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant \beta, \quad \forall (w_\tau)_{\tau=1}^t \in \mathcal{W}, \quad \forall t \in [T]$$

Can we get $\tilde{\mathcal{O}}(\sqrt{T})$ ***regret for the constrained LQR?***

We show $\tilde{\mathcal{O}}(\sqrt{T})$ regret and satisfaction of *chance-constraints*:

$$\mathbb{P}\left(\alpha_j^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant \beta\right) \geqslant 1 - \delta, \quad \forall t \in [T]$$

[2] e.g. Genceli & Nikolaou (1996), Aswani et al. (2013), Tanaskovic et al. (2014)

[3] Dean et al. (2019), Li et al. (2021)

Adaptive LQR with Chance-constraints

Interaction Model

In each time step $t \in [T]$:

- Observe state x_t.
- Choose input u_t.
- Incur cost $\ell(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$.
- State evolves: $x_{t+1} = Ax_t + Bu_t + w_t, \;\; w_t \sim \mathcal{N}(\mathbf{0}, W)$

System (A, B) is *unknown*.

Chance-constraints

$$\mathbb{P}\left(\alpha_j^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant \beta\right) \geqslant 1 - \delta \quad \forall j \in [J], \forall t \in [T]$$

Main assumption Known *baseline policy K* ensures that,

$$\mathbb{P}\left(\alpha_j^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant \beta - \epsilon\right) \geqslant 1 - \delta \quad \forall j \in [J], \forall t \in [T]$$

Regret against best linear controller

Evaluate performance with *regret* against a benchmark cost $J^\star$.

$$R_T = \sum_{t=1}^{T} \ell(x_t, u_t) - J^\star$$

Take $J^\star$ to be best cost attainable by a linear controller that satisfies the chance-constraints.

$$\begin{aligned} J^\star = \min_K \ & \mathbb{E} \sum_{t=1}^{T} \ell(x_t, u_t) \\ \text{s.t.} \ & u_t = K x_t \\ & K \text{ is stabilizing} \\ & \mathbb{P}\left(\alpha_j^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant \beta \right) \geqslant 1 - \delta, \quad \forall j \in [J], \forall t \in [T] \end{aligned}$$

Covariance Viewpoint

View the problem as choosing the target steady-state covariance:

$$\Sigma_t = \begin{bmatrix} \Sigma_{t,xx} & \Sigma_{t,xu} \\ \Sigma_{t,ux} & \Sigma_{t,uu} \end{bmatrix} \approx \operatorname{cov}\begin{pmatrix} x_t \\ u_t \end{pmatrix}$$

Can then choose input that (approximately) induces that covariance:

$$u_t = \mathcal{N}(K_t x_t, U_t), \quad K_t = \Sigma_{t,ux}\Sigma_{t,xx}^{-1}, \quad U = \Sigma_{t,uu} - K_t \Sigma_{t,xx} K^\top$$

Advantages of this perspective:

- Space of steady-state covariance is convex.
- Expected cost is linear in covariance at steady-state.
- Constraint is linear in covariance at steady-state.
- Existing tools from unconstrained adaptive LQR.[4]

$$\Sigma_{xx} = [A\ B]\Sigma[A\ B]^\top + W, \qquad \Sigma \succeq 0$$

$$\mathbb{E}[\ell(x_t, u_t)] = \left\langle \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}, \Sigma \right\rangle$$

$$\mathbb{P}\left(\alpha^\top \begin{bmatrix} x_t \\ \end{bmatrix} \leqslant \beta\right) \geqslant 1 - \delta \qquad \Longleftrightarrow \qquad \langle \alpha\alpha^\top, \Sigma \rangle \leqslant \frac{\beta^2}{\quad}$$

Unconstrained Optimistic SDP[5]

Optimism in the face of uncertainty is a well-known paradigm for decision-making under uncertainty.

> *"Behave as though unknown is as favorable as reasonably possible"*

When choosing Σ_t, the "unknown" is the system (A, B) in the steady-state condition:

$$\Sigma_{xx} = [A\ B]\Sigma[A\ B]^\top + W$$

Error bounds on the true system in terms of the estimate $(\hat{A}, \hat{B})$:

$$\|[\hat{A}\ \hat{B}]\Sigma[\hat{A}\ \hat{B}]^\top - [A\ B]\Sigma[A\ B]^\top\| \leqslant \eta\langle \bar{V}^{-1}, \Sigma\rangle,$$

Relax the steady-state condition accordingly:

$$\Sigma_{xx} \succeq [\hat{A}\ \hat{B}]\Sigma[\hat{A}\ \hat{B}]^\top + W - \eta\langle \bar{V}^{-1}, \Sigma\rangle I$$

[5] Cohen, Koren & Mansour, 2019

Unconstrained Optimistic SDP[6]

Optimism in the face of uncertainty is a well-known paradigm for decision-making under uncertainty.

> *"Behave as though unknown is as favorable as reasonably possible"*

In unconstrained setting, efficient policy can then be computed with an *optimistic* SDP:

$$\begin{aligned} \Sigma_t = \quad & \underset{\Sigma=\left[\begin{smallmatrix}\Sigma_{xx} & \Sigma_{xu}\\ \Sigma_{ux} & \Sigma_{uu}\end{smallmatrix}\right]}{\arg\min} \quad \left\langle \left[\begin{smallmatrix} Q & 0\\ 0 & R\end{smallmatrix}\right], \Sigma \right\rangle \\ & \text{s.t. } \Sigma_{xx} \succeq [\hat{A}\ \hat{B}]\Sigma[\hat{A}\ \hat{B}]^\top + W - \eta\langle \bar{V}^{-1}, \Sigma\rangle I \\ & \Sigma \succeq 0 \end{aligned}$$

[6] Cohen, Koren & Mansour, 2019

Constrained Optimistic SDP

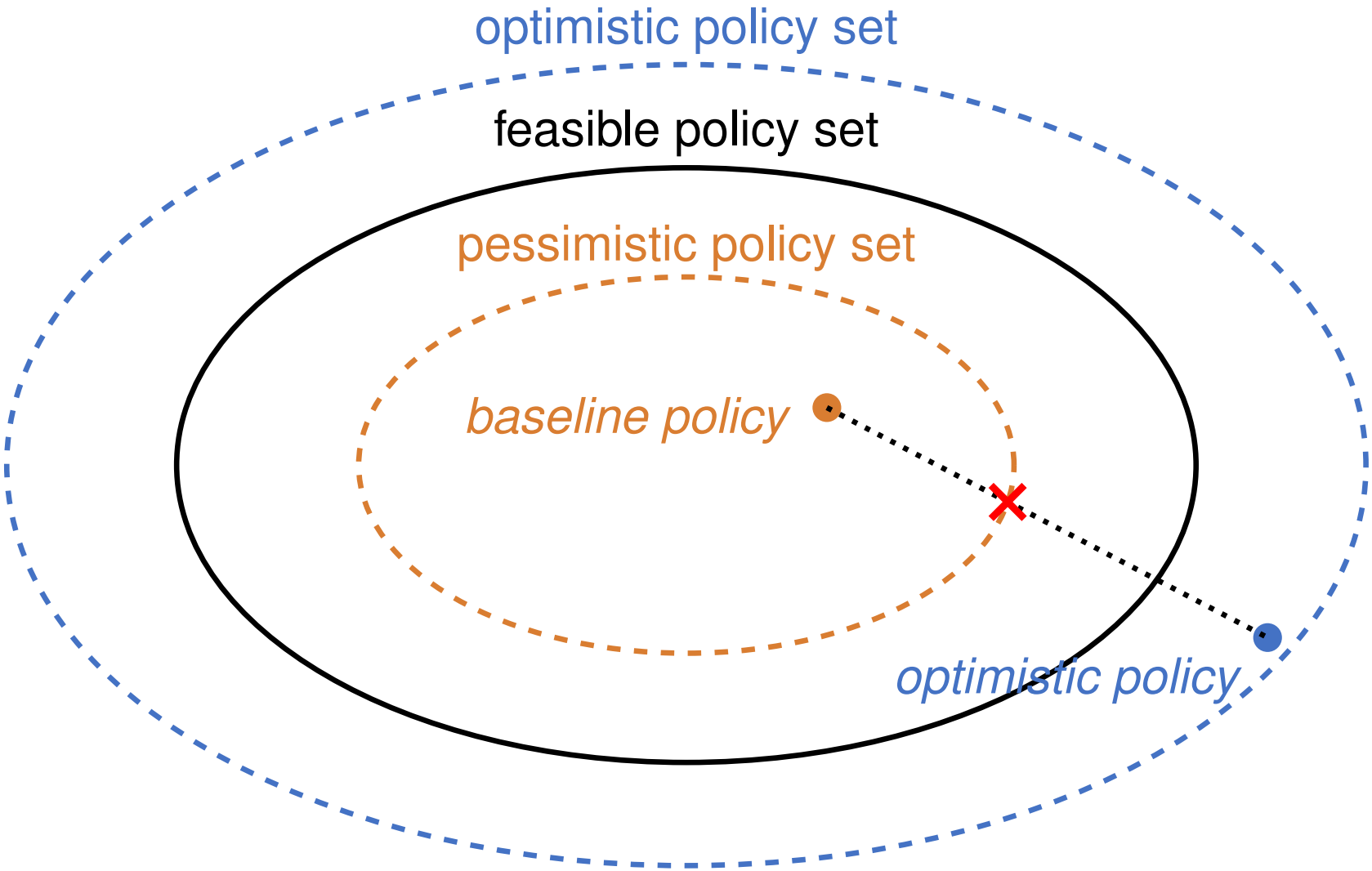
We incorporate constraints in to optimistic SDP:

$$\begin{aligned}\Sigma^o = \underset{\Sigma=\left[\begin{smallmatrix}\Sigma_{xx} & \Sigma_{xu}\\ \Sigma_{ux} & \Sigma_{uu}\end{smallmatrix}\right]}{\arg\min} \quad & \left\langle \left[\begin{smallmatrix} Q & 0\\ 0 & R\end{smallmatrix}\right], \Sigma \right\rangle \\ \text{s.t. } & \Sigma_{xx} \succeq [\hat{A}\ \hat{B}]\Sigma[\hat{A}\ \hat{B}]^\top + W - \eta\langle \bar{V}^{-1}, \Sigma\rangle I \\ & \langle \alpha\alpha^\top, \Sigma\rangle \leqslant \beta^2/\Phi^{-1}(1-\delta)^2 \\ & \Sigma \succeq 0\end{aligned}$$

The resulting policy will have low regret, but is *not* guaranteed to satisfy the constraints due to the relaxed steady-state condition.

How do we ensure constraint satisfaction while maintaining efficiency?

Inspiration: Scaled-back Optimism[7]



How do we construct the pessimistic policy set?

[7] Hutchinson, Turan & Alizadeh, 2024

Bounding the System Covariance

To construct the pessimistic policy set, need to understand how the actual system covariance $\mathrm{cov}\left(\begin{smallmatrix} x_t \\ u_t \end{smallmatrix}\right)$ relates to the chosen policy.

Lemma

Let the following hold:

- Σ_t *is* $\mathcal{F}_{t-\rho}$ *measurable where* $\rho \asymp \log(T)$.
- $\|\Sigma_t - \Sigma_{t+1}\| \leqslant \zeta$
- $\Sigma_{t,xx} \succeq [A\ B]\Sigma_t[A\ B]^\top + W - \eta$

Then,

$$\mathrm{cov}\left(\begin{smallmatrix} x_t \\ u_t \end{smallmatrix} | \mathcal{F}_{t-\rho}\right) - \Sigma_t \preceq \tilde{\mathcal{O}}(1)\,(1/T + \zeta + \eta)\,I \tag{1}$$

Algorithm design techniques:

- Delay state information by ρ steps before using it for estimation.
- Slowly vary policy.

Choose pessimistic set such that constraints are satisfied given (1).

Algorithm

Algorithm 1: Safe Adaptive LQR

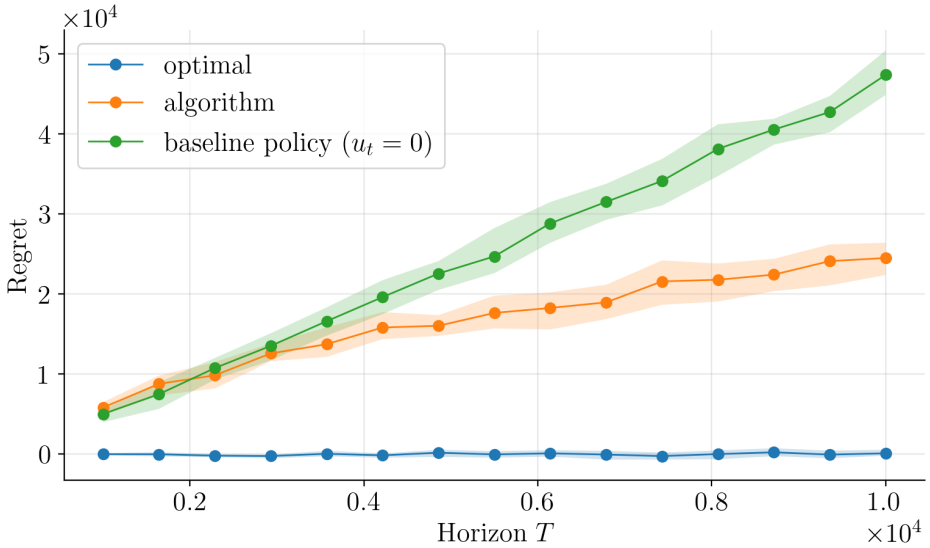
Initialization phase for τ_1 time steps.
for $t = \tau_1$ **to** T **do**
  if *substantial new information* **then**
    $(\hat{A}, \hat{B}) \leftarrow$ estimate system with delayed information
    $\Sigma^o \leftarrow$ constrained optimistic SDP
    $\Sigma^{\text{base}} \leftarrow$ estimate covariance of baseline policy
    $\mathcal{E}^p \leftarrow$ pessimistic set
    $\phi = \max\{\phi \in [0, 1] : \phi\Sigma^o + (1 - \phi)\Sigma^{\text{base}} \in \mathcal{E}^p\}$
    $\bar{\Sigma} = \phi\Sigma^o + (1 - \phi)\Sigma^{\text{base}}$
  end
  $\Sigma_t = \Sigma_{t-1} + (\bar{\Sigma} - \Sigma_{t-1})\zeta$
  $u_t \leftarrow$ extract policy from Σ_t
end

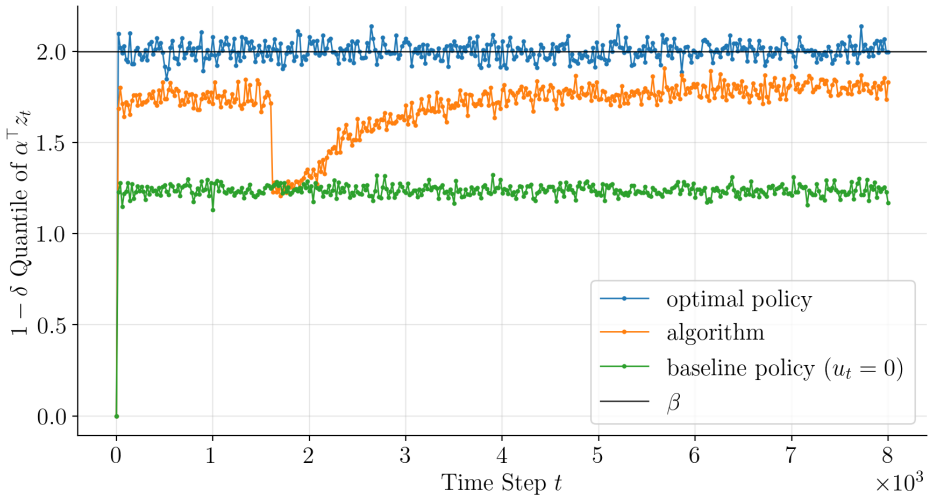
Numerical Experiments

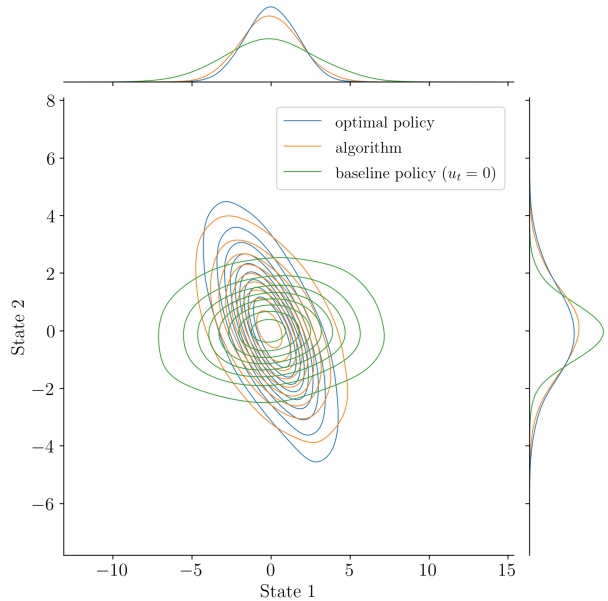
$$A = \begin{bmatrix} 0.95 & 0.1 \\ 0 & 0.5 \end{bmatrix}, \; B = \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}$$

$$Q = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 0.1 \end{bmatrix}, \; R = 1, \; W = 0.7I$$

$$\mathbb{P}\left(\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \leqslant 2 \right) \geqslant 0.9$$







Conclusion

Contributions

- $\tilde{\mathcal{O}}(\sqrt{T})$ regret for LQR with chance-constraints
- Proposed algorithm combines optimistic SDP with "scaled-back optimism"

Future work

- Extend to non-Gaussian (e.g. via Cantelli's inequality).
- Does certainty equivalence work in this setting?
- What can we say about robust constraints?

Thank you!

shutchinson@ucsb.edu