RL for Model-free Linear Quadratic Control with Process and Measurement Noises (r1)

Farnaz Adib Yaghmaie (farnaz.adib.yaghmaie@liu.se), Fredrik Gustafsson (fredrik.gustafsson@liu.se)

Motivation

Reinforcement Learning (RL) studies Learning approaches for model-free optimal control of dynamical systems. It shows impressive results but in general it is difficult to analyze. Here, we consider a Linear Quadratic optimal control problem which is

- theoretically tractable,
- practical in various engineering domains,
- possible to use Least Squares Temporal Difference Learning (LSTD).

Linear Quadratic Problem

Consider a linear Gaussian dynamical system

$$x_{k+1} = Ax_k + Ba_k + w_k, \tag{1}$$
$$y_k = x_k + v_k, \tag{2}$$

with $w_k \sim \mathcal{N}(\mathbf{o}, W_w),\ v_k \sim \mathcal{N}(\mathbf{o}, W_v)$

- *Differential value function* associated with a given policy π

$$V^\pi(y_k) = \mathbf{E}[\sum_{t=k}^{+\infty}(r(y_t, \pi(y_t)) - \lambda^\pi)|y_k]. \tag{3}$$

- *Average cost associated with the policy* $\pi(y_k)$

$$\lambda^\pi = \lim_{N\to\infty}\frac{1}{N}\mathbf{E}[\sum_{t=1}^{N} r(y_t, \pi(y_t))] \tag{4}$$

- *Quadratic running cost* with $R_y \geq 0$ and $R_a > 0$

$$r(y_k, a_k) = y_k^T R_y y_k + a_k^T R_a a_k \tag{5}$$

Lemma 1

The differential value function (3) associated with $\pi(y_k) = Ky_k$ is quadratic; i.e. $V^\pi(y_k) = y_k^T P^\pi y_k$

$$(A + BK)^T P^\pi (A + BK) - P^\pi + K^T R_a K + R_y = \mathbf{o}, \tag{6}$$

and

$$\lambda^\pi = Trace(K^T B^T P^\pi B K W_v) + Trace(P^\pi W_w) + Trace(P^\pi W_v) - Trace(L^T P^\pi L W_v). \tag{7}$$

Alg 1- Average Off-Policy Learning

1: **Initialize:** $K^{(0)}$ and set $k = 0$.

2: **repeat**

3: **Policy evaluation:** Let $\Phi_k := vecv(y_k)$, $r_k := r(y_k, \pi^i)$ and $\bar{\lambda}^i = \frac{1}{\tau}\sum_{t=1}^{\tau} r_t$. Estimate P^i from

$$vecs(\hat{P}^i) = (\sum_{t=0}^{\tau-1}\Phi_t(\Phi_t - \Phi_{t+1})^T)^{-1}(\sum_{t=0}^{\tau-1}\Phi_t(r_t - \lambda^i))$$

4: **Policy Improvement:** Let

$$c_k = y_k^T(R_y + K^{iT}R_aK^i - \hat{P}^i)y_k + y_{k+1}^T\hat{P}^i y_{k+1} - \bar{\lambda}^i,$$
$$\varphi_k = \begin{bmatrix} 2(a_k - K^i y_k) \otimes y_k \\ vecv(a_k) - vecv(K^i y_k)\end{bmatrix},$$
$$\xi^i = \begin{bmatrix} vec(A^T P^i B) \\ vecs(B^T P^i B)\end{bmatrix}, \tag{8}$$

- Estimation of some parameters

$$\hat{\xi}^i = (\sum_{t=0}^{\tau'-1}\varphi_t\varphi_t^T)^{-1}(\sum_{t=0}^{\tau'-1}\varphi_t c_t). \tag{9}$$

- Improved policy

$$K^{i+1} = -(\sum_{j=1}^{i}(\hat{N}^j + R_a))^{-1}(\sum_{j=1}^{i}\hat{H}^{jT}). \tag{10}$$

5: **until** Convergence.

Theorem 1

Assume that the estimated error is small enough. Then, Algorithm 1 produces stabilizing policy gains $K^{i+1}, i = 2..., I$; i.e. $\rho(A + BK^{i+1}) < 1$.

Simulation Results

A data center cooling with three sources coupled to their own cooling devices

$$x_{k+1} = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix} x_k + I_3 a_k + w_k$$

with $W_w = I_3,\ W_v = I_3$ and $r(y_k, a_k) = y_k^T 0.001 I_3 y_k + a_k^T I_3 a_k$.

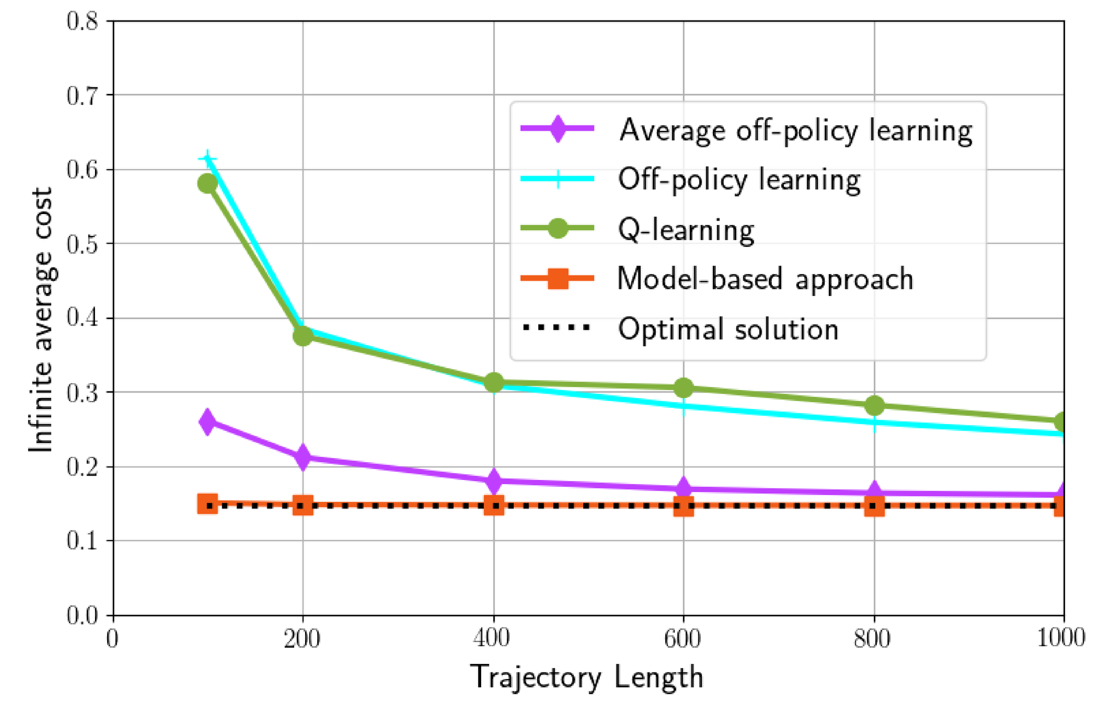

Figure 1-Median of infinite average cost for 100 stable trajectories

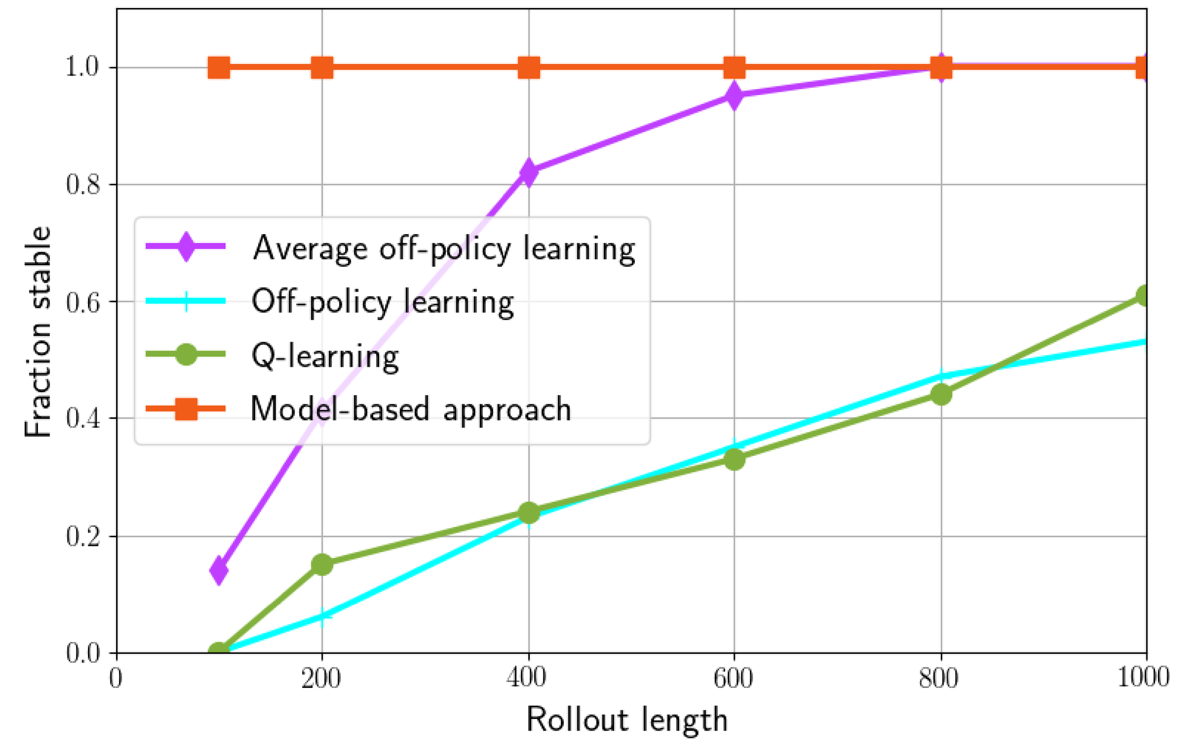

Figure 2- The fraction of stable policy gains generated by each algorithm in all iterations

Conclusions

- We have considered both process and measurement noises.
- We have proposed a model -free algorithm that outperforms the classical Q- and off--policy learning.

References

[r1] F. Adib Yaghmaie and F. Gustafsson "Using Reinforcement Learning for Model-free Linear Quadratic Control with Process and Measurement Noises", In 2019 Decision and Control, IEEE 58th Conference on, 2019, pp. 6510-6517.

LINKÖPING UNIVERSITY

LINK-SIC

LINKÖPING UNIVERSITY
Division of Automatic Control