ESE 680-004: Learning and Control **Fall 2019**

Lecture 12: Learning to Control Markov Decision Processes

Lecturer: Nikolai Matni *Scribes: Klayton Wittler*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1 Introduction

Previously we were able to show with system level synthesis and robust control bounds [1] that end-to-end guarantees of performance for a controller

$$\frac{\hat{J} - J_*}{J_*} \leq C(robustness, excitability)\sqrt{\frac{(d+p)\log\frac{1}{\delta}}{N}} \tag{1}$$

with probability of $1-\delta$ for sufficiently large N. Here $\hat{J}$ represents the learned controller on the true system, J_* is the optimal performance, d is the number of states and p is the number of inputs, and C is a constant depending on the true system.

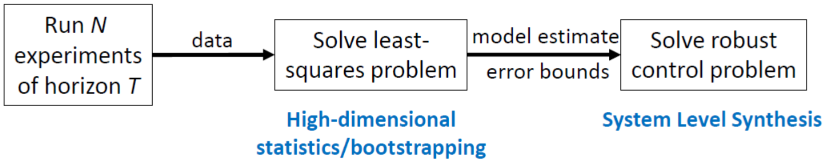
Run *N* experiments of horizon *T*
data
Solve least-squares problem
model estimate
error bounds
Solve robust control problem
High-dimensional statistics/bootstrapping
System Level Synthesis

Figure 1: Learning and control pipeline

However to understand how good this guarantee is we need to understand different performance metrics for learned control policies.

2 MDP

2.1 Finite MDP

In finite MDP we consider the following equation where T is the horizon length.

$$\begin{aligned} &\min_{\pi} \mathbb{E}[\sum_{t=0}^{T-1} c_t(x_t, u_t) + c_T(x_T)] \\ &\text{s.t. } x_{t+1} = f_t(x_t, u_t, w_t) \\ &\quad\; u_t = \pi_t(x_{0:t}, u_{0:t-1}) \end{aligned} \tag{2}$$

Here $x_t \in \mathbb{R}^{n_x}$ is the state, $u_t \in \mathbb{R}^{n_u}$ the control input, and ${}_t \in \mathbb{R}^{n_w}$ the the state transition noise. The control policy is $\pi = \{\pi_1, ..., \pi_t, .., \pi_{T-1}\}$ where a particular π_t maps the current state and previous inputs to the current input and is possibly a random mapping.

If dynamic transition functions and cost functions are known and Markovian we can restrict the policy search to $u_t = \pi_t(x_t)$ and solve using dynamic programming via Bellman iteration on the value function.

$$\begin{aligned} V_T(x_T) &= \mathbb{E}[c_T(x_T)] \\ V_t(x_t) &= \min_{u_t} \mathbb{E}[c_t(x_t, u_t) + V_{t+1}(f_t(x_t, u_t, w_t))] \end{aligned} \tag{3}$$

In the example of the Linear Quadratic Regulator (LQR) with time invariant dynamics and cost, $Q \succ 0$, and $R \succ 0$.

$$\begin{aligned} &\min_{\pi} \mathbb{E}[\sum_{t=0}^{T-1} x_t^T Q x_t + u_t^T R u_t + x_T^T Q_T x_T] \\ &\text{s.t. } \; x_{t+1} = A x_t + B u_t, x_0 = \zeta \\ &\qquad u_t = \pi_t(x_t) \end{aligned} \tag{4}$$

where $V_T(x) = x_T^T Q_T x_T$ and assume $V_{t+1} = x_{t+1}^T P_{t+1} x_{t+1}$ to solve the recursion on

$$\begin{aligned} V_t(z) &= \min_u z^T Q z + u^T R u + V_{t+1}(Az + Bu) \\ &= \min_u z^T Q z + u^T R u + (Az + Bu)^T P_{t+1}(Az + Bu) \\ &= z^T(Q + A^T P_{t+1} A - A^T P_{t+1} B (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A) z \end{aligned} \tag{5}$$

with

$$\begin{aligned} u_t^* &= K x_t \\ K &= -(R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A \\ P_t &= Q + A^T P_{t+1} A - A^T P_{t+1} B (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A \\ P_T &= Q_T \end{aligned} \tag{6}$$

2.2 Infinite MDP

Moving to the infinite horizon setting with static cost and dynamics functions then different infinite horizon costs can be formulated.

The discounted cost setting:

$$\begin{aligned} &\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(x_t, u_t)] \\ &\gamma \in (0, 1] \end{aligned} \tag{7}$$

If $c(x_t, u_t)$ is bounded almost surely and $\gamma < 0$ then $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(x_t, u_t)] < \infty$ which is easy to work with theoretically. The optimal cost-to-go and policy can also be obtained from the Bellman Equation

$$V_*(x) = \min_u \mathbb{E}[c(x, u) + \gamma V_*(f(x, u, w))] \tag{8}$$

which is a model-based approach using the value function and the model free approach being

$$\begin{aligned} Q_*(x, u) &= \min_v \mathbb{E}[c(x, u) + \gamma Q_*(f(x, u, w), v)] \\ \pi_*(x) &= \operatorname*{argmin}_u Q_*(x, u) \end{aligned} \tag{9}$$

The downside being for control the bounded discounted cost does not guarantee stability.

Asymptotic average cost:

$$\begin{aligned} &\mathbb{E}[\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)] \\ &\gamma \in (0, 1] \end{aligned} \tag{10}$$

Care must be taken to make sure the limit converges but if the closed loop system is stable then it will converge. It is also appropriate for stability in stochastic control, but is difficult to work with theoretically. One application is in stochastic LQR below.

$$\begin{aligned}
&\min_{\pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=0}^{T-1} x_t^T Q x_t + u_t^T R u_t + x_T^T Q_T x_T] \\
&\text{s.t. } x_{t+1} = Ax_t + Bu_t \\
&\qquad u_t = \pi_t(x_t)
\end{aligned} \tag{11}$$

which can be solved with Discrete Algebraic Riccati (DAR) recursion similar to as before. Also if A, B, and $Q^{\frac{1}{2}}$ are stabilizable and detectable then the closed loop system is stable and converges to a stationary distribution and static policy.

3 MDP with unknown dynamics

Now if the dynamic transition functions are unknown but the cost functions are known and time invariant, then learning can be broken into two categories episodic tasks and single-trajectory tasks. In episodic tasks data is collected over a finite horizon, updates are made, and the system is reset to begin the next episode. In single-trajectory tasks a system is evaluated under a policy during a single evolution.

There exist a tension between identifying an unknown system and controlling it known as the exploration vs. exploitation tradeoff. Exploration requires sufficient excitation of a system to achieve an accurate model but this degrades performance and if an incorrect model is exploited then a system is left with sub-optimal performance. To quantify this tradeoff, there are two main performance metrics Probably Approximately correct (PAC) bounds and Regret bounds.

3.1 PAC

Tasks are performed over a horizon H which may be infinite but the system can be reset after some time H_r. The optimal achievable cost is V_* and the number of episodes for which the policy π is not ϵ-optimal ($V_\pi > V_* + \epsilon$) is N_ϵ. A policy is said to be episodic (ϵ,δ)-PAC if after T episodes it satisfies

$$\mathbb{P}[N_\epsilon > (n_x, n_u, H. \frac{1}{\epsilon}, \frac{1}{\delta})] \leq \delta \tag{12}$$

which guarantees the chosen policy is ϵ-optimal on all but a number of episodes polynomial in the parameters with probability at least $1-\delta$. The algorithm normally operates in two phases, pure exploration to approximate the system model then a exploitation where the model is used to create a control policy. Informally, PAC can be thought of the number of episodes required to have an ϵ-optimal policy.

Utilizing PAC bounds for LQR can be done for both an episodic set up as well as in the single-trajectory case. Previously in [2] its been shown that LQR with asymptotic average cost is episodic PAC-learnable via injecting white in time Gaussian noise to the open-loop system over at most $(n_x, n_u, H_r, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ episodes followed by least-squares system identification and the robust synthesis method to guarantee

$$\mathbb{P}[V_\pi - V_* \geq \epsilon] \leq \delta \tag{13}$$

which can be restated in the robust synthesis framework as $\hat{J} - J_* \lesssim J_* O(\epsilon)$ with probability at least $1-\delta$ as long as $N \gtrsim \frac{\sigma_w^2 (n+p)\log\frac{1}{\delta}}{\lambda_{min}(\Lambda_C)\epsilon^2}$. Where as defined earlier $\hat{J}$ represents the learned controller on the true system, J_* is the optimal performance, d is the number of states and p is the number of inputs, and

Similarly, in a single-trajectory over an infinite horizon a policy π can be characterized as PAC-learnable if

$$\mathbb{P}[N_\epsilon > (n_x, n_u, \frac{1}{\epsilon}, \frac{1}{\delta})] \leq \delta \tag{14}$$

Which guarantees the number of time-steps, N_ϵ, where $V_\pi(x_t) > V_*(x_t) + \epsilon$ to be less than $poly(\frac{1}{\epsilon}, \log\frac{1}{\delta})$ with probability $1-\delta$. Where $V_\pi(x_t)$ represents the cost-to-go from state x_t achieved by a policy and V_* is the optimal cost-to-go that can be achieved.

The limits of PAC bounds is that it is only penalized for sub-optimality above ϵ and is not guaranteed to converge to optimal since it can ceases learning after it is ϵ-optimal.

3.2 Regret

Regret bounds evaluate the quality of an adaptive policy by comparing its running cost to a baseline. Here b_T is the baseline cost at time T to be compared to the regret incurred by a policy.

$$R^\pi(T) \coloneqq \sum_{t=0}^{T} c_t(x_T, \pi_t(x_{0:t}, u_{0:t-1})) - b_T \tag{15}$$

The two most common regret guarantees are expected regret

$$\mathbb{E}[R^\pi(T) \leq poly(n_x, n_u, T)] \tag{16}$$

and high probability regret.

$$\mathbb{P}[R^\pi(T) \geq poly(n_x, n_u, T, \frac{1}{\delta})] \leq \delta \tag{17}$$

Regret bound for LQR are in the form of

$$R^\pi(T) \coloneqq \sum_{t=0}^{T} x_t^T Q x_t + u_t^T R u_t - TV_* \tag{18}$$

with V_* being the optimal asymptotic average cost achieved by the true optimal LQR controller. The policy from this can be shown to be $u_t = \hat{K}x_t + \eta_t$, where $\hat{K} = dlqr(\hat{A}, \hat{B}, Q, R)$ is the solution to LQR with estimated dynamics and has exploration in the form of $\eta_t \sim \mathcal{N}(0, \sigma_{\eta,t}^2 I)$ which injects some noise into the control input. This achieves a regret bound of

$$R^\pi(T) \leq poly(n_x, n_u, \log(\frac{1}{\delta}))O(T^{\frac{1}{2}}) \tag{19}$$

Similarly, moving this again to the robust synthesis framework $\hat{J}_T - TJ_* \leq \tilde{O}(T^{2/3})$ for moderate uncertainty [3] and $\hat{J}_T - TJ_* \leq \tilde{O}(T^{1/2})$ for small uncertainty [4].

The limits of regret bounds is that it has no worst-case guarantee because it only tracking the integral of sub-optimal behaviour and cannot distinguish between a few severe mistakes and many small ones. However it is different from PAC in that all sub-optimal behaviour is tracked so a balance between exploration and exploitation must be made.

3.3 Uniform-PAC

To handle the downsides of both PAC and Regret bounds a new framework was proposed in [5] to satisfy both PAC and high probability regret bounds. This is done by simultaneously for all $\epsilon > 0$ selecting an ϵ-optimal policy on all episodes except for a number that scales polynomially with $\frac{1}{\epsilon}$ with high probability. The key insight into proving these bounds is to leverage time-uniform concentration bounds such as the finite-time versions of the law of iterated logarithm which gives horizon-dependent confidence levels.

References

[1] James Anderson, John C. Doyle, Steven Low, and Nikolai Matni. System level synthesis, 2019.

[2] N. Matni S. Dean, H. Mania. On the sample complexity of the linear quadratic regulator, 2017.

[3] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator, 2018.

[4] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control, 2019.

[5] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning, 2017.