Information Theoretic Regret Bounds for Online Nonlinear Control

Model-based Reinforcement Learning - The Kernelized Nonlinear Regulator Model

This work studies the following nonlinear control problem:

\[ x_{t+1} = f(x_t, u_t) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2 I), \]  

where \( h \in \{0, 1, \ldots, H - 1\} \); the state \( x_t \in \mathbb{R}^d \); the control \( u_t \in \mathbb{U} \) where \( \mathbb{U} \) may be an arbitrary set; \( f : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{X} \).

- Assumed that \( f \) lives in a known RKHS.
- Equivalently, \( f(x, u) = W^\top \phi(x, u) \).
- \( W^* \): Unknown linear mapping, \( \phi : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{V} \): Known function. \( \mathbb{V} \): Either finite or countably infinite dimensional Euclidean vector space.

Main result: a regret bound \( \tilde{O}(d_{\text{effective}} \sqrt{HT}) \). \( (T:=\text{#episodes}) \)


Theory & Algorithm

Lower Confidence-based Continuous Control (LC^3)

1. Initialize \( \text{BALL}^0 \) to be any set containing \( W^* \)
2. for \( t = 0 \ldots T \) do
3. \( \pi_t^* = \arg\min_{\pi \in \Pi} \min_{W \in \text{BALL}^t} J(x_t, c_t^*, W) \)
4. Execute \( \pi_t^* \) to sample \( c_t^* := (x_t^*, u_t^*, c_t^*, x_{t+1}^*)_{h=0}^{H-1} \)
5. Update \( \text{BALL}^{t+1} \)
6. end for

- \( \Pi \): Policy class
- \( \text{BALL}^t \): Confidence ball of weight matrices \( W \) at 4th episode
- \( \epsilon^t \): Immediate cost function at 4th episode
- \( J(x_t, c_t^*, W) \): Expected total cost of \( r \) under the dynamics \( W(\cdot, u) + \epsilon \)

Assumption 1

We have access to an oracle that implements Line 3 of Algorithm 3 - One may use some effective heuristics such as DDP, MPPI etc.

Assumption 2

For every episode, the cost function is non-negative and the realized cumulative cost has uniformly bounded second moments, i.e.,

\[ \mathbb{E}_{\pi} \left( \sum_{t=0}^{H-1} J(x_t, u_t) \right)^2 \leq \mathbb{V}_{\text{max}}. \]

Armhand Robotics System

- 33 degree of freedom robotic arm and hand system tasked with picking up an spherical object
- Learning model dynamics for the real world applications such as robotics requires sufficiently complex features
- We test our method with features created by six ensemble of MuJoCo models
- Each element of the ensemble is unable to complete the task in isolation

Experiments

Benchmark Tasks with Random Features

- Benchmark tasks including MuJoCo environments from OpenAI Gym
- We use Random Fourier Features
- It is observed that the Thompson Sampling variant of our proposed algorithm with RFFs quickly increased reward in early stages, indicating low sample complexities empirically
- Our algorithm consistently performs well on simple continuous control tasks

We do NOT require bounded state space, bounded cost function or bounded feature map.

1. We develop a stopping time martingale to handle the unbounded nature of the (realized) cumulative costs
2. We develop a novel way to handle Gaussian smoothing through the chi-squared distance function between two distributions
3. We utilize methods developed for the analysis of linear bandits and Gaussian process bandits (e.g. maximum information gain)

We address multi-step extension to RL settings

We prove a “self-bounding” regret bound that relates the instantaneous regret to the second moment of the stochastic process

Self-Bounding, Simulation Lemma

For any policy, model parameterization, and non-negative cost, and for any initial state, we have:

\[ J^*(x_0; W^*) - J^*(x_0; W) \leq \sqrt{H^2 \mathbb{V}_{\text{cost}}(x_0; W) \mathbb{E} \sum_{t=1}^{H} \min_{\mathcal{M}_{\beta}} \left\{ \frac{1}{2} \| (W - W^*) \phi(x_t, u_t) \|_2^2 \right\}}. \]

Maze

- We construct a continuous maze environment, where an agent plans for continuous actions with MPPI
- Exploration is necessary to find the goal
- We compare random walk exploration to our method, with and without posterior sampling