Information Theoretic Regret Bounds for Online Nonlinear Control

Sham Kakade, Akshay Krishnamurthy,
Kendall Lowrey, Motoya Ohnishi, Wen Sun

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING · Microsoft · Cornell University

NEURAL INFORMATION PROCESSING SYSTEMS

Model-based Reinforcement Learning
- The Kernelized Nonlinear Regulator Model

This work studies the following nonlinear control problem:

$$x_{h+1} = f(x_h, u_h) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2 I), \qquad (1)$$

where $h \in \{0, 1, \ldots, H-1\}$; the state $x_h \in \mathbb{R}^{d_{\mathcal{X}}}$; the control $u_h \in \mathcal{U}$ where $\mathcal{U}$ may be an arbitrary set; $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$.

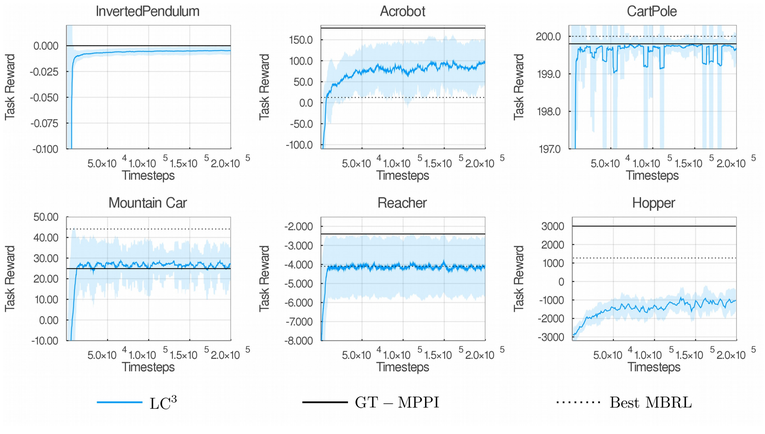
- Assume that f lives in a known RKHS.
- Equivalently: $f(x, u) = W^* \phi(x, u)$.
  W^*: Unknown linear mapping, $\phi : \mathcal{X} \times \mathcal{U} \to \mathcal{V}$: Known function.
  $\mathcal{V}$: Either finite or countably infinite dimensional Euclidean vector space.

- Main result: a regret bound $\widetilde{O}(\tilde{d}_{\text{effective}}\sqrt{HT})$. ($T$:#episodes)
- Proof techniques: a novel self-bounding regret lemma.

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



Theory & Algorithm

Lower Confidence-based Continuous Control (LC^3)

1: Initialize BALL^0 to be any set containing W^*
2: **for** $t = 0 \ldots T$ **do**
3: $\quad \pi^t = \text{argmin}_{\pi \in \Pi} \min_{W \in \text{BALL}^t} J^\pi(x_0; c^t, W)$
4: $\quad$ Execute π^t to sample $\tau^t := \{x_h^t, u_h^t, c_h^t, x_{h+1}^t\}_{h=0}^{H-1}$
5: $\quad$ Update BALL^t
6: **end for**

- Π: Policy class
- BALL^t: Confidence ball of weight matrices W at tth episode
- c^t: Immediate cost function at tth episode
- $J^\pi(x_0; c^t, W)$: Expected total cost of π under the dynamics $W\phi(x, u) + \epsilon$

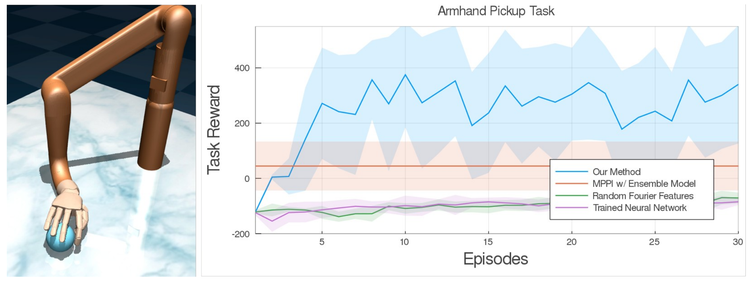
Assumption 1
We have access to an oracle that implements Line 3 of Algorithm
– One may use some effective heuristics such as DDP, MPPI etc.

Assumption 2
For every episode, the cost function is non-negative and the realized cumulative cost has uniformly bounded second moments, i.e.,

$$\sup_{\pi \in \Pi} \mathbb{E}\left[\left(\sum_{h=0}^{H-1} c^t(x_h, u_h)\right)^2 \,\middle|\, x_0, \pi\right] \le V_{\max}.$$

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



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:

$$J^\pi(x_0; c, W^\star) - J^\pi(x_0; c, W)$$
$$\le \sqrt{HV^\pi(x_0; c, W^\star)} \sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1} \min\left\{\frac{1}{\sigma^2}\left\|(W^\star - W)\,\phi(x_h, u_h)\right\|_2^2, 1\right\}\right]}.$$

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