Model-based Reinforcement Learning with Parametrized Physical Models and Optimism-Driven Exploration

Berkeley

Artificial Intelligence Research Laboratory

Christopher Xie, Sachin Patil, Teodor Moldovan, Sergey Levine, Pieter Abeel

University of Washington, UC Berkeley

Problem:

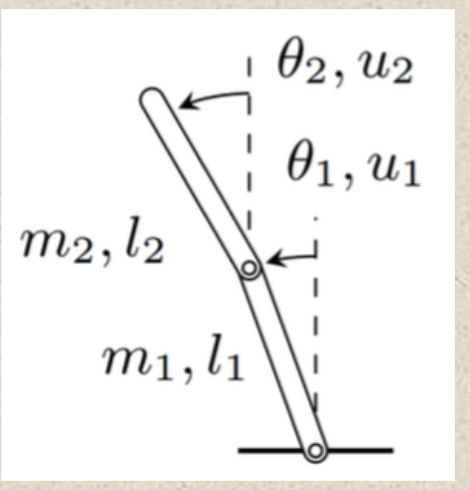
- Complete a specific robotic task without prior knowledge of the dynamics of the system

Approach:

- Employ optimistic exploration based MPC along with a simple least squares regression model that can be updated continuously with new data in real time

Assumptions:

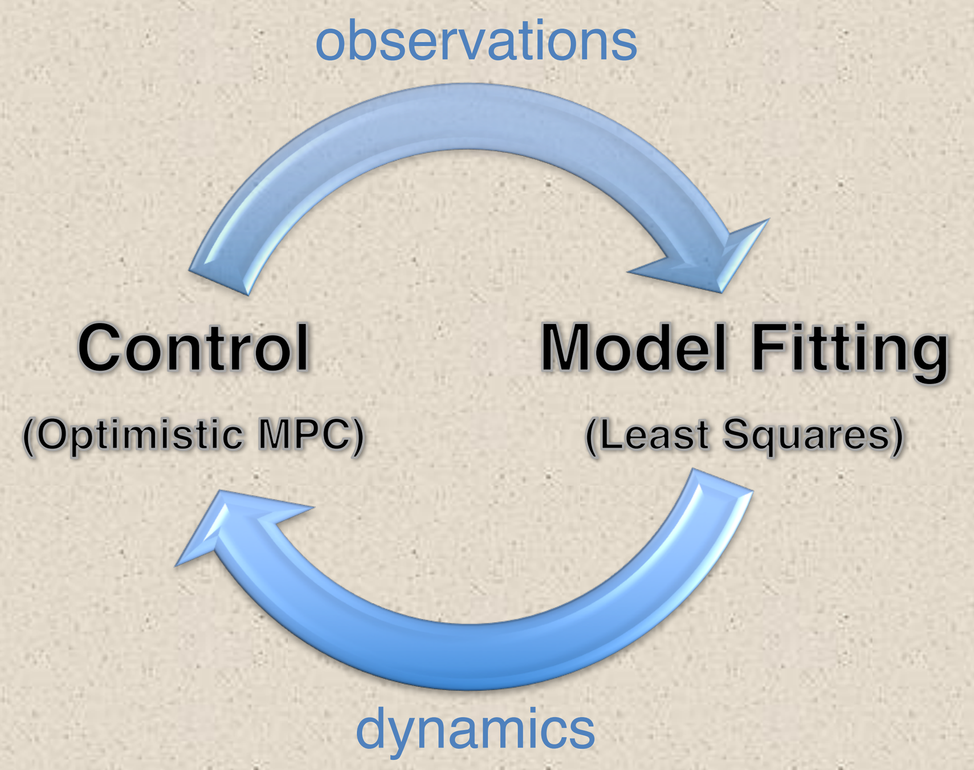
- Robot is an open-chain manipulator
- Robot morphology is known, but NOT physical parameters such as mass and length of links



Related Work:

- PILCO [Deisenroth et al. 2011]
- Optimism-driven exploration for nonlinear systems [Moldovan et al. 2015]
- Approximate real-time optimal control based on sparse Gaussian process models [Boedecker et al. 2014]

Algorithm:



Dynamics Model:

- Transform nonlinear dynamics into a simple linear model

$$M(q)\ddot{q} + C(q,\dot{q}) + g(q) = \tau$$

$$H(q,\dot{q},\ddot{q}) \cdot \Delta = \tau$$

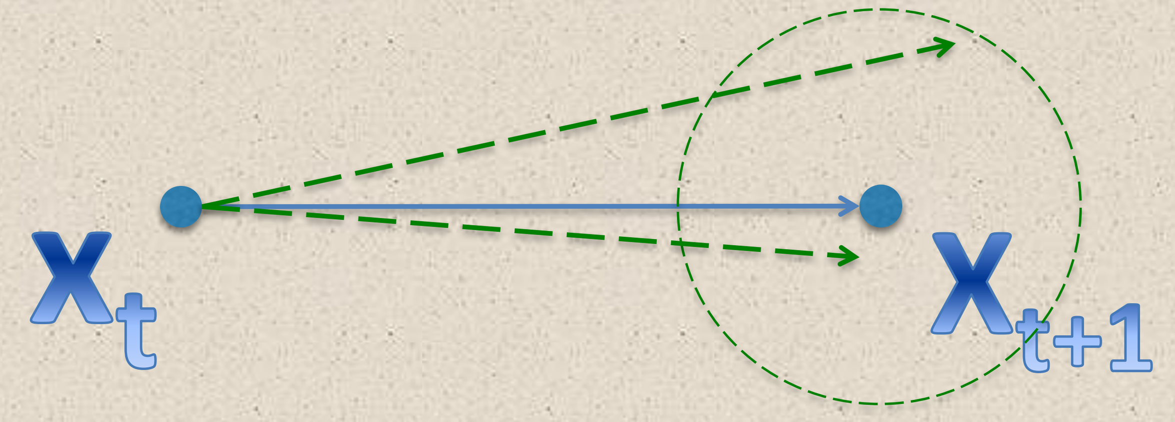
- System Identification is least squares

$$\hat{\Delta} = \underset{\Delta}{\operatorname{argmin}} \|A\Delta - b\|_2^2 \qquad A = \begin{bmatrix} H(q_1,\dot{q}_1,\ddot{q}_1) \\ \vdots \\ H(q_N,\dot{q}_N,\ddot{q}_N) \end{bmatrix} \quad b = \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_N \end{bmatrix}$$

- Optimistic exploration encoded by dynamics slack variables

$$\ddot{q} = \hat{f}_\Delta(q,\dot{q},\tau) = M_\Delta^{-1}(q)\left(\tau - C_\Delta(q,\dot{q}) - g_\Delta(q)\right)$$

$$\ddot{q}_t = \hat{f}_\Delta(q_t,\dot{q}_t,\tau_t) + \xi_t = \tilde{f}_\Delta(q_t,\dot{q}_t,\tau_t,\xi_t)$$

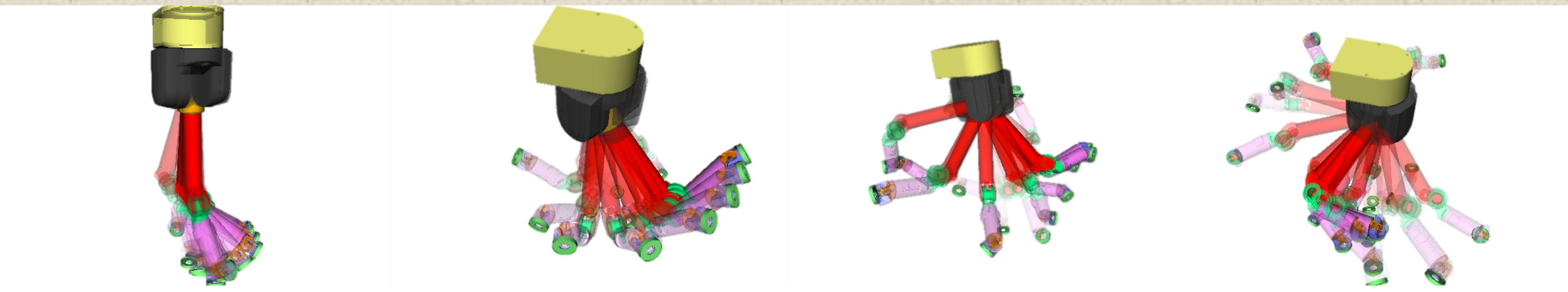


Experiments: Benchmarks

	pendulum	cartpole	double pendulum
DDP with known dynamics	3.04 ± 0.89s	7.44 ± 3.26s	3.7 ± 0.89s
our method	3.28 ± 1.17s	8.31 ± 3.15s	4.98 ± 1.83s
optimism-driven exploration [21]	3.9 ± 1s	10 ± 3s	17 ± 7s
Boedecker et al. [9]	—	12-18s	—
PILCO [13]	12s	17.5s	50s

- Interaction time to successfully complete each benchmark task
- Our method is pretty competitive compared to lower bound

Experiments: 7 DOF Barrett Arm



target pose:	1	2	3	4	5
DDP with known dynamics	1.43 ± 0.03s	1.64 ± 0.02s	1.34 ± 0.02s	2.68 ± 0.84s	1.57 ± 0.03s
our method	5.84 ± 2.76s	9.11 ± 3.4s	10.9 ± 4.62s	9.14 ± 6.22s	3.61 ± 1.12s
target pose:	6	7	8	9	10
DDP with known dynamics	2.05 ± 0.0s	0.35 ± 0.09s	1.9 ± 0.0s	2.65 ± 0.0s	4.98 ± 3.32s
our method	6.15 ± 2.64s	4.6 ± 2.35s	3.71 ± 1.34s	7.77 ± 2.36s	9.99 ± 4.49s

- Interaction time to successfully reach each target pose
- Again, our method is competitive compared to lower bound

Future Work:

- Evaluate our approach on a real system with unmodeled effects
- Combine our linear model with more sophisticated statistical models