KTH ROYAL INSTITUTE
OF TECHNOLOGY

Safe learning for control:
Combining disturbance estimation, reachability analysis and reinforcement learning with systematic exploration

Caroline Heidenreich

June 1, 2017

Motivational Example

- Autonomous vehicle with partly known model
- Task: find optimal control without driving off the road
- To simplify, we only look at the truck's position

Motivation

How can we find the optimal control?

1. Model-based control:
   - Not possible without physical insight.

2. Learn a policy with Reinforcement Learning (RL):
   - Directly or indirectly.
   - Requires to visit all (safe) states.

Motivation

How can we make sure to stay on the road?

- RL algorithms not designed for satisfying constraints.
- We need an additional safety-preserving controller.

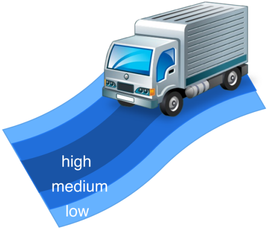
$\Rightarrow$ Safe Learning Control

Algorithm

Markov Decision Process

- Discretise states and actions.
- Assign rewards to each state-action pair.
- Determine objective that agent should maximise.

Reward function



Objective function

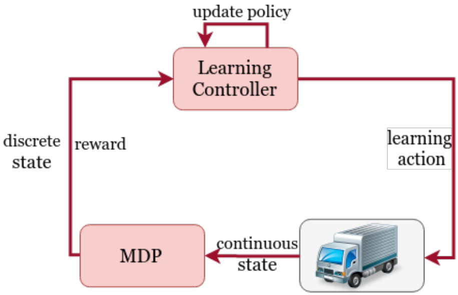
$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

Algorithm

Reinforcement Learning

Finding optimal policy by

- play action
- receive reward
- update policy



✓ There are algorithms that converge to the optimal policy.
✗ No safety guarantees.

Algorithm

Safe Set Calculation

How can we ensure safety with uncertain dynamics?

- Treat the unknown dynamics as bounded disturbance.
- Determine for each state if our control manages to keep us on the road for all disturbances.

Algorithm

Safe Learning

- At the borders of the safe set: Apply safe control.
- Within the safe set: Reinforcement learning.

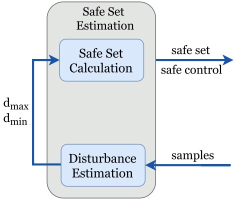
✓ Learn a control without leaving the road.

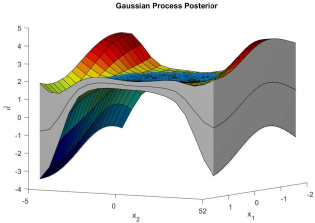
✗ Small safe set due to conservative disturbance range.

Algorithm

Disturbance Estimation with Gaussian Processes

- Update the disturbance range with measured data.
- Gaussian Process regression: Non-parametric regression method that gives:
  - **a.** an estimate for the disturbance.
  - **b.** a measure how certain this estimate is.





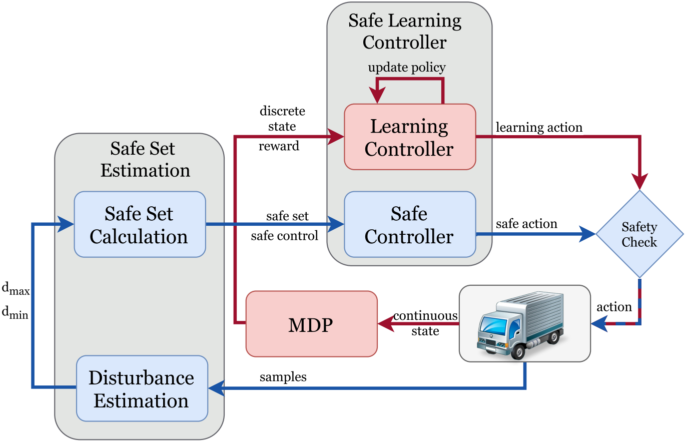
Algorithm

Exploration

- Trade-off between exploration and exploitation.
- Need for visiting the whole safe set i.o. to learn policy.
- Chosen method: Promote state-action pairs that have not been visited often.

Algorithm

Summary of Approach

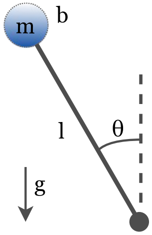


Implementation

Reinforcement Learning	Version of Delayed Q-learning
Disturbance Estimation	Gaussian Processes
Safe Set	Hamilton-Jacobi-Isaacs Reachability
Exploration	Incremental Q-learning

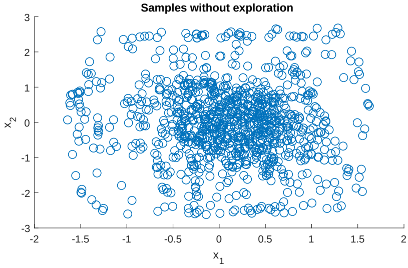
Evaluate approach on:

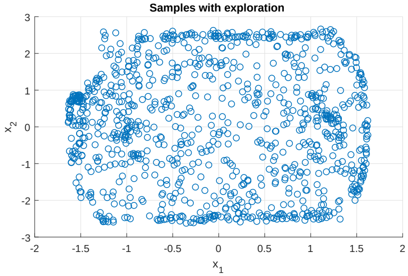
- Inverted Pendulum System.
- Two states: position and angular velocity.
- Four iterations with 10,000 learning steps.



Experimental Results

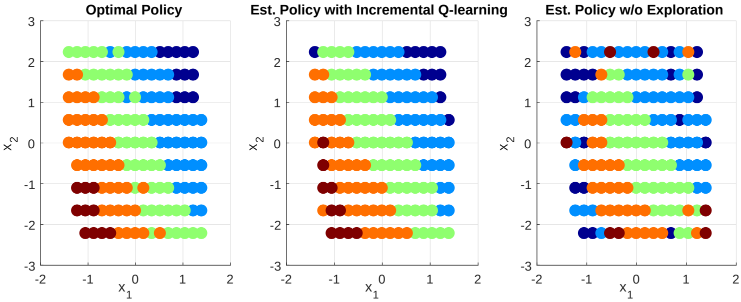
Exploration





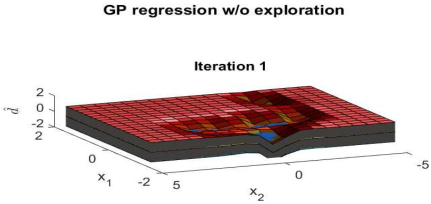
Experimental Results

Policy Estimation



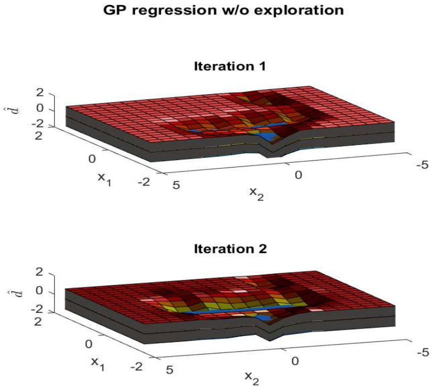
Experimental Results

Disturbance Estimation



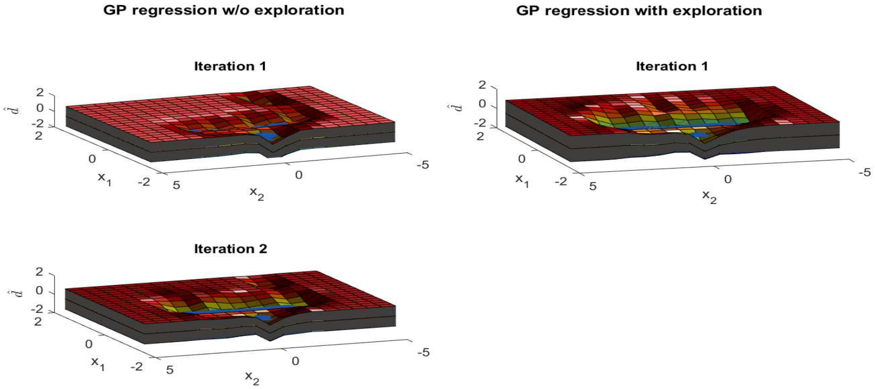
Experimental Results

Disturbance Estimation



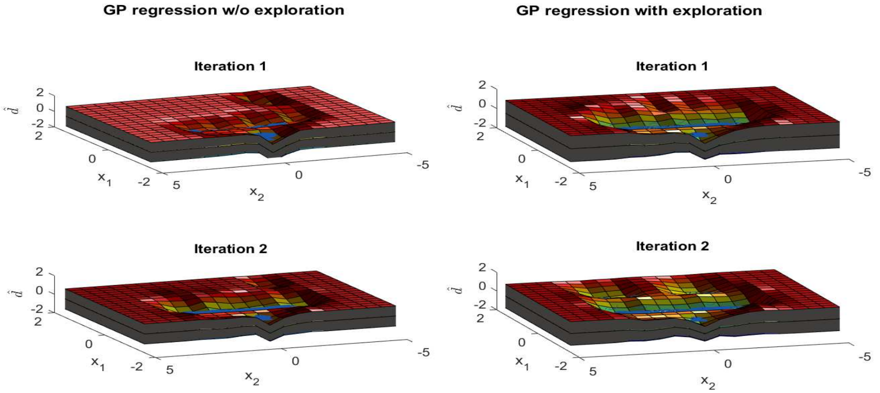
Experimental Results

Disturbance Estimation



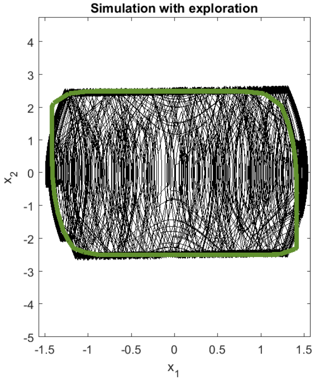
Experimental Results

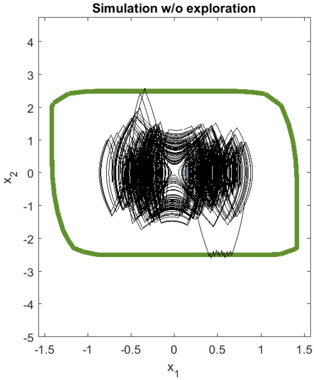
Disturbance Estimation



Experimental Results

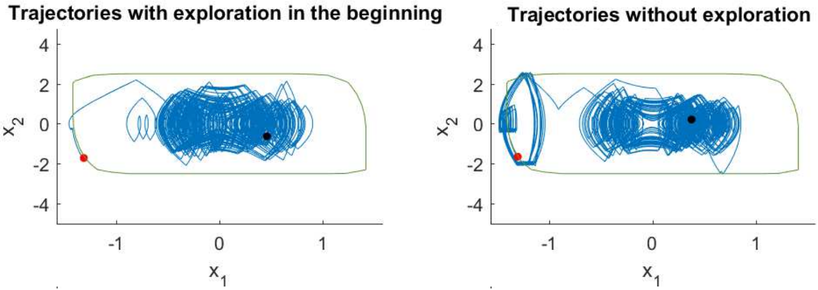
Trajectories





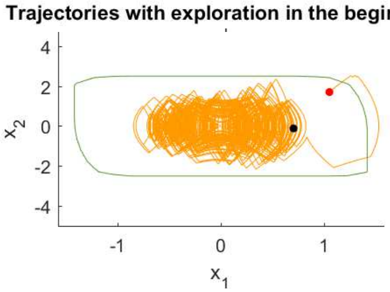
Experimental Results

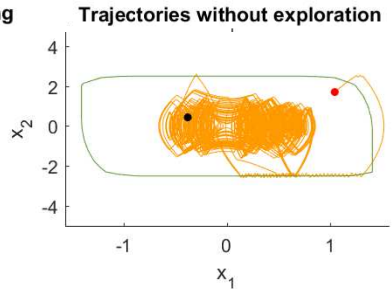
Trajectories



Experimental Results

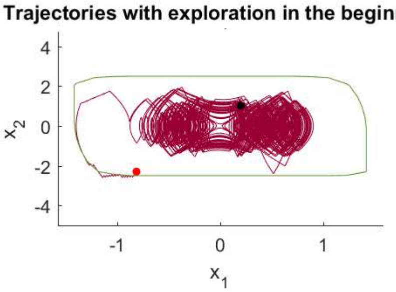
Trajectories

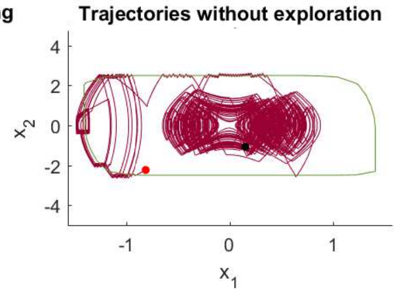




Experimental Results

Trajectories





Conclusions

- We manage to learn an accurate policy for inverted pendulum.
- System can always be brought back to safety.
- Considerably better results by incorporating exploration.

Future Work

Some theoretical & practical challenges remain:

- Joint design of safety and learning loop.
- Recursive estimation of disturbance bounds.
- Formal guarantees for the whole algorithm.

$\left.\begin{array}{l}\text{Recursive estimation of disturbance bounds.}\\ \text{Formal guarantees for the whole algorithm.}\end{array}\right\}$ theoretically challenging

Thank you for listening!

Questions?