Partially Observable Markov Decision Processes for Planning in Uncertain Environments

Project Thesis
Lasse Peters

Institute of Mechanics and Ocean Engineering
Prof. Dr.-Ing. Robert Seifried
Hamburg University of Technology

Hybrid Systems Laboratory
Prof. Claire J. Tomlin
University of California at Berkeley

25.07.2019

Introduction

Types of Uncertainty

- state uncertainty
- outcome uncertainty

Introduction

Types of Uncertainty

- state uncertainty
- outcome uncertainty

Dealing with Uncertainty

- worst case disturbance sequences
- probabilistic reasoning

Introduction

Types of Uncertainty

- state uncertainty
- outcome uncertainty

Dealing with Uncertainty

- worst case disturbance sequences
- probabilistic reasoning

POMDPs

- general framework
- computationally demanding

Introduction

Types of Uncertainty

- state uncertainty
- outcome uncertainty

Dealing with Uncertainty

- worst case disturbance sequences
- probabilistic reasoning

POMDPs

- general framework
- computationally demanding
- recent research: faster solvers

Introduction

This Work

- problem specific approximations

Introduction

This Work

- problem specific approximations vs. full POMDP approaches

Introduction

This Work

- problem specific approximations vs. full POMDP approaches
- application domains:
  1. localization and planning
  2. motion planning with latent human intentions

Introduction

This Work

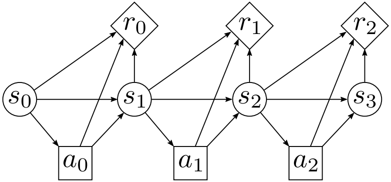
- problem specific approximations vs. full POMDP approaches
- application domains:
  1. localization and planning
  2. motion planning with latent human intentions

Outline

1. The Partially Observable Markov Decision Process
2. Approximate Online POMDP Solvers
    - DESPOT
    - POMCPOW
3. Application Domain: Localization and Planning
    - Baselines
    - POMDP Solvers
4. Conclusion and Future Work

The Partially Observable Markov Decision Process

Dynamic Decision Network



MDP

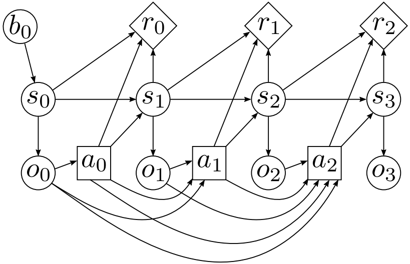
- state s
- action a
- reward r

Objective: Finding a policy π^* that maximizes

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right].$$

The Partially Observable Markov Decision Process

Dynamic Decision Network



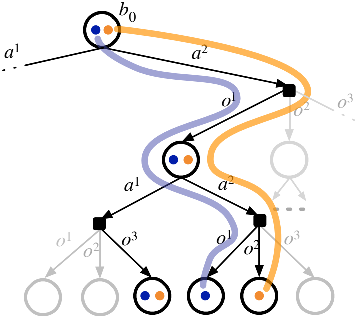
POMDP

- state s
- action a
- reward r
- observation o
- initial belief b_0

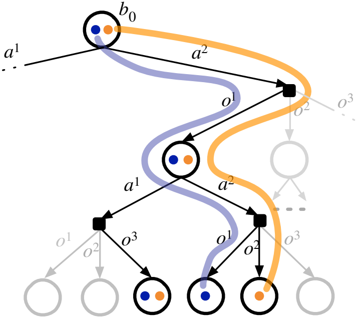
Objective: Finding a policy π^* that maximizes

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right].$$

Determinized Sparse Partially Observable Tree (DESPOT)



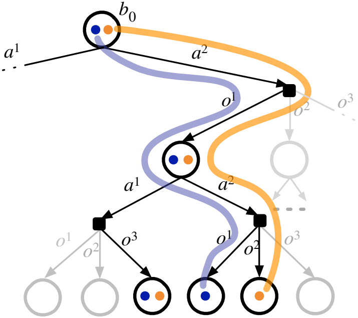
Determinized Sparse Partially Observable Tree (DESPOT)



High Level Idea

- incrementally construct sparse tree
- maintain bounds on *value* at nodes
- choose action with best lower bound

Determinized Sparse Partially Observable Tree (DESPOT)



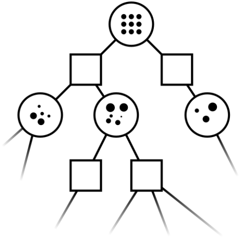
High Level Idea

- incrementally construct sparse tree
- maintain bounds on *value* at nodes
- choose action with best lower bound

Domain Knowledge

- user-defined initial bound estimates

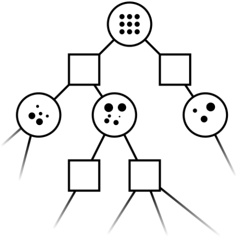
Partially Observable Monte-Carlo Planning with Observation Widening (POMCPOW)



High Level Idea

- Monte-Carlo tree search
- locally approximate the *value function* through Monte-Carlo simulations
- choose action with highest value

Partially Observable Monte-Carlo Planning with Observation Widening (POMCPOW)



High Level Idea

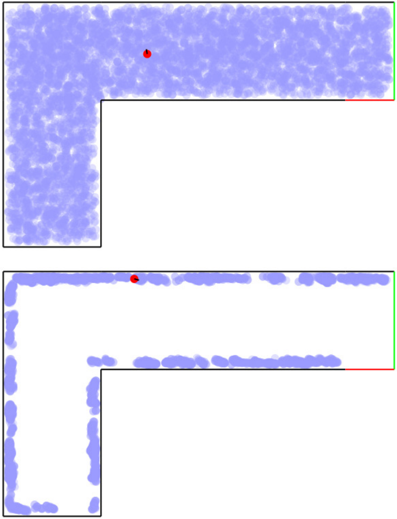
- Monte-Carlo tree search
- locally approximate the *value function* through Monte-Carlo simulations
- choose action with highest value

Domain Knowledge

- user-defined value estimate

Localization and Planning

Screenshot of the Simulator

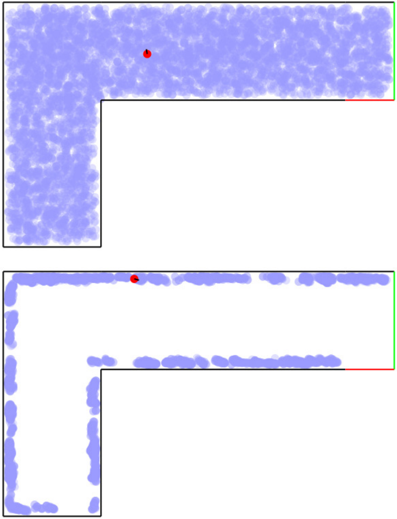


Problem Details

- *known room* but *unknown location*

Localization and Planning

Screenshot of the Simulator

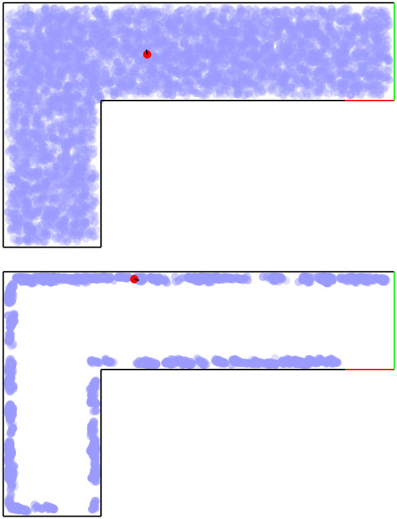


Problem Details

- *known room* but *unknown location*
- objective:
  - success: leave room at exit (green)
  - failure: falling down stairs (red)
  - penalties: time and collisions
- observations: collision sensor
- dynamics: noisy differential drive
  - actions: left, right, straight

Localization and Planning

Screenshot of the Simulator



Problem Details

- *known room* but *unknown location*
- objective:
  - success: leave room at exit (green)
  - failure: falling down stairs (red)
  - penalties: time and collisions
- observations: collision sensor
- dynamics: noisy differential drive
  - actions: left, right, straight

POMDP

- continuous state space
- discrete action and observation space

Baselines

Mode Control

- use mode of belief to approximate the state
- treat POMDP as fully observable problem

Baselines

Mode Control

- use mode of belief to approximate the state
- treat POMDP as fully observable problem

Most Likely Reflex Agent (MLRA)

- handcrafted feedback policy
- P-controller tracking preset path

Baselines

Mode Control

- use mode of belief to approximate the state
- treat POMDP as fully observable problem

Most Likely Reflex Agent (MLRA)

- handcrafted feedback policy
- P-controller tracking preset path

Most Likely Model Predictive Control (MLMPC)

- use MPC to plan a path
- cost function: negative reward

$\Rightarrow$ MLMPC is an optimal policy for the fully observed problem (MDP)

Integrating Domain Knowledge for POMDP Solvers

Value Estimates

1. Rollout Value Estimate
    - simulate default policy: “always straight”
2. Analytic Value Estimate
    - estimate remaining steps from distance to goal
    - approximate value by cumulative living penalty

Integrating Domain Knowledge for POMDP Solvers

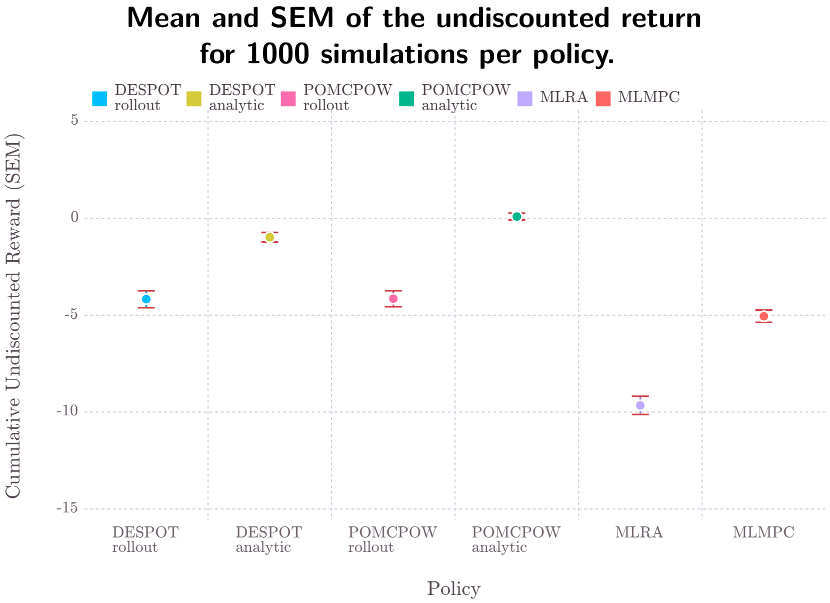
Value Estimates

1. Rollout Value Estimate
    - simulate default policy: “always straight”
2. Analytic Value Estimate
    - estimate remaining steps from distance to goal
    - approximate value by cumulative living penalty

POMDP Solver Setups

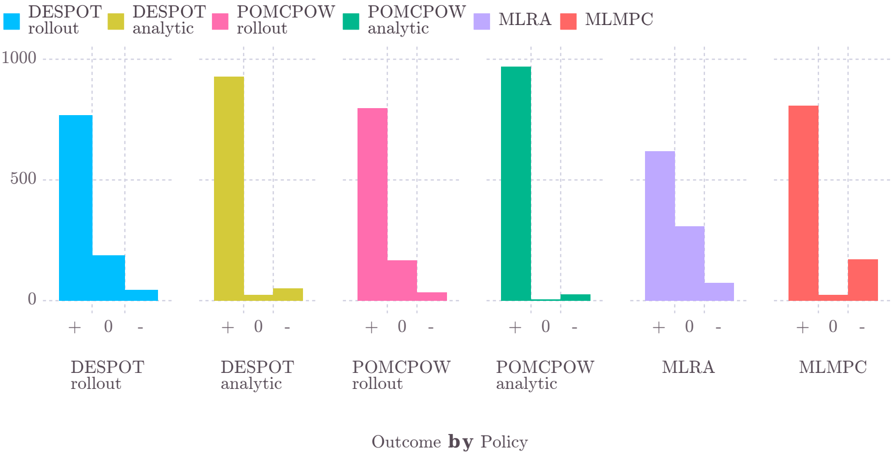
1. DESPOT-rollout
2. DESPOT-analytic
3. POMCPOW-rollout
4. POMCPOW-analytic

Evaluation – Undiscounted Return



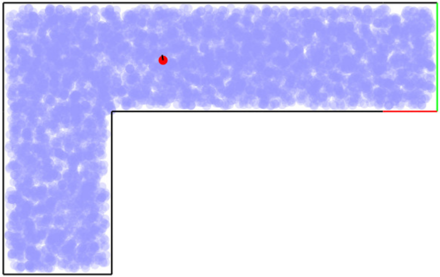
Evaluation – Outcome Distribution

Histogram of outcome frequencies grouped by policy.

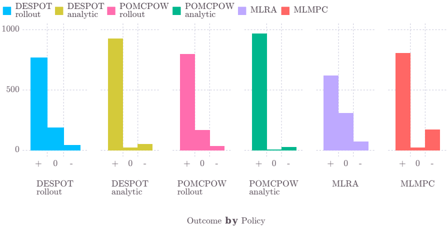


Evaluation – Outcome Distribution

Initial Conditions

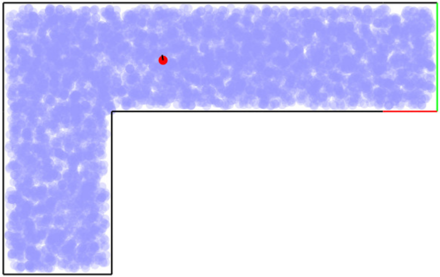


Histogram of outcome frequencies grouped by policy.

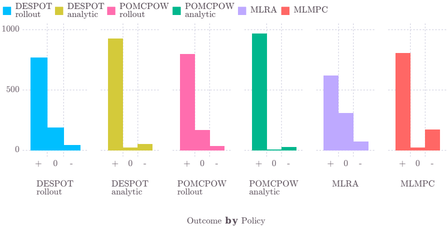


Evaluation – Outcome Distribution

Initial Conditions



Histogram of outcome frequencies grouped by policy.



⇒ POMCPOW-analytic near optimal with respect to safety

Evaluation – Qualitative Analysis

MLMPC

POMCPOW-Analytic

Evaluation – Qualitative Analysis

MLMPC

- mode approximation compromises safety
- passive information gathering
- fails for highly symmetric beliefs

POMCPOW-Analytic

Evaluation – Qualitative Analysis

MLMPC

- mode approximation compromises safety
- passive information gathering
- fails for highly symmetric beliefs

POMCPOW-Analytic

- active information gathering
- reliably reduces uncertainty
- safe and efficient behaviors

Conclusion and Future Work

Conclusions

1. safe and efficient behaviors
   ⇒ suitable approach for safety vs. efficiency tradoff
2. near real-time planning capabilities
   ⇒ already a useful high-level planner for moderately sized problems
3. designing suitable heuristic guidance hard but cruitial for performance

Conclusion and Future Work

Conclusions

1. safe and efficient behaviors
   ⇒ suitable approach for safety vs. efficiency tradoff
2. near real-time planning capabilities
   ⇒ already a useful high-level planner for moderately sized problems
3. designing suitable heuristic guidance hard but cruitial for performance

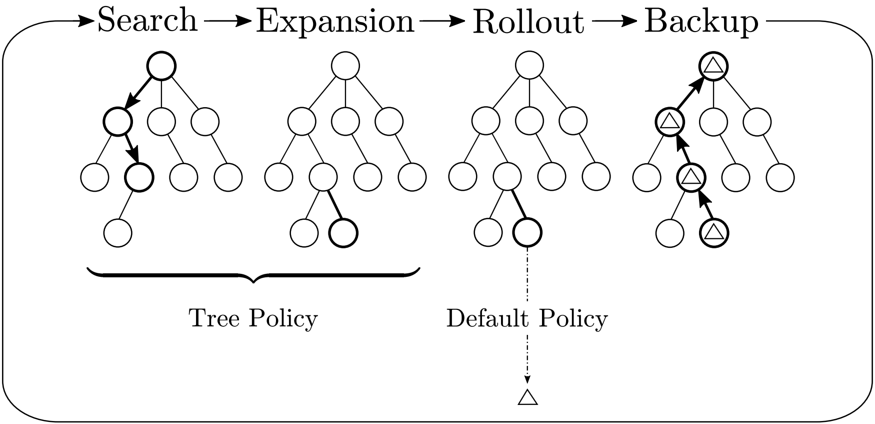
Future Work

1. theory for a-priory safety assurances
2. improving scalability by using GPU
3. learning for designing of heuristic guidance

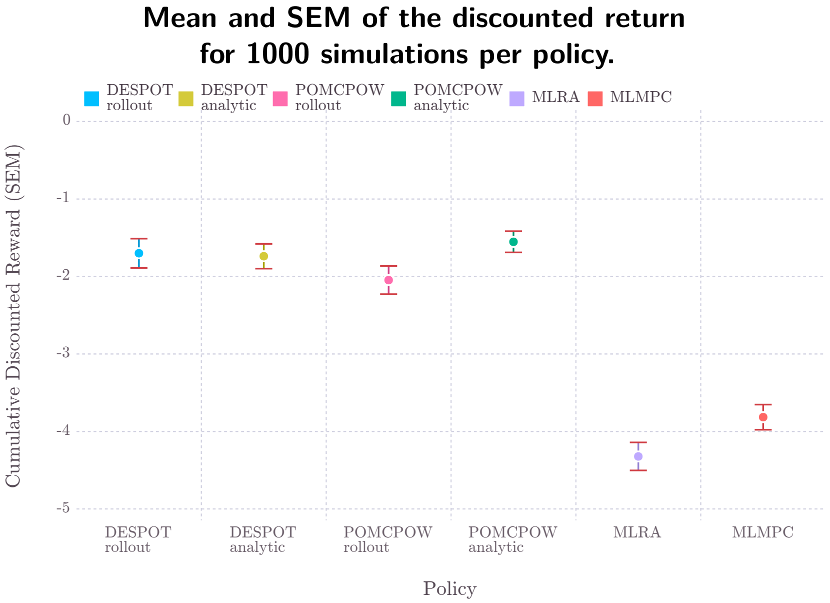
End

End

Monte-Carlo Tree Search

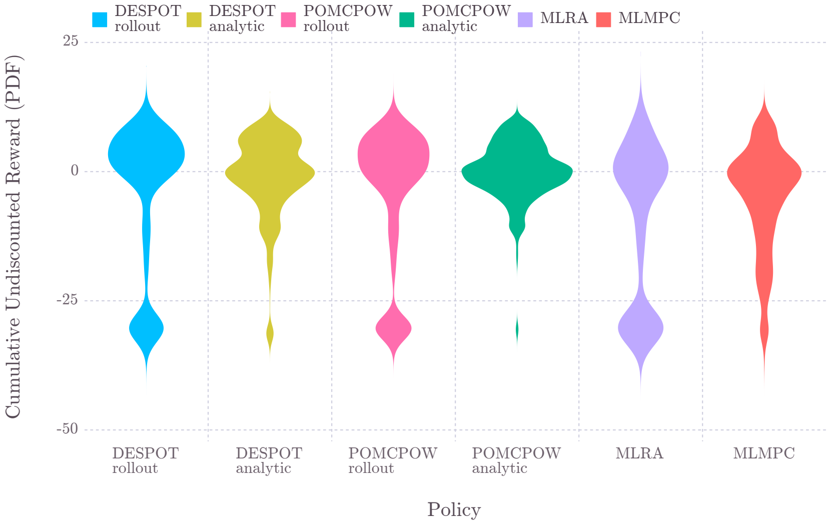


Evaluation: Localization and Planning – Objective Value



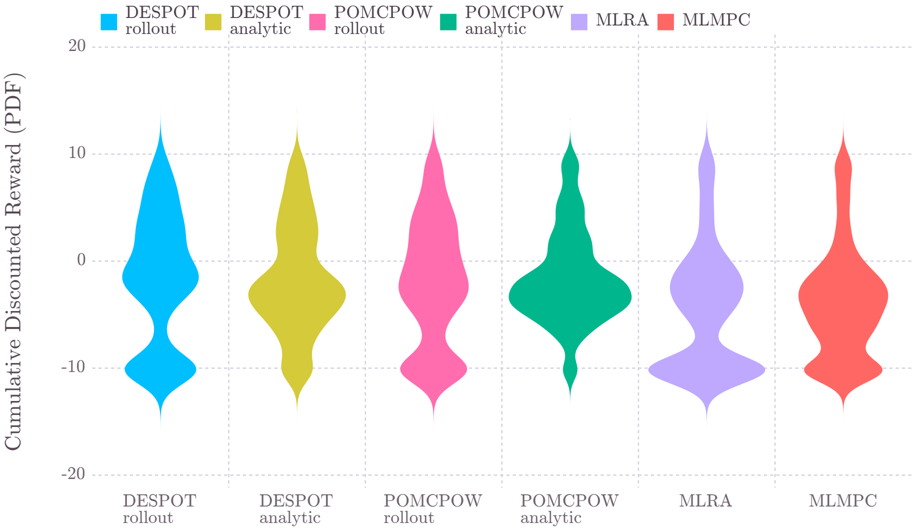
Evaluation – Undiscounted Return

Distribution of the undiscounted return for 1000 simulations per policy.

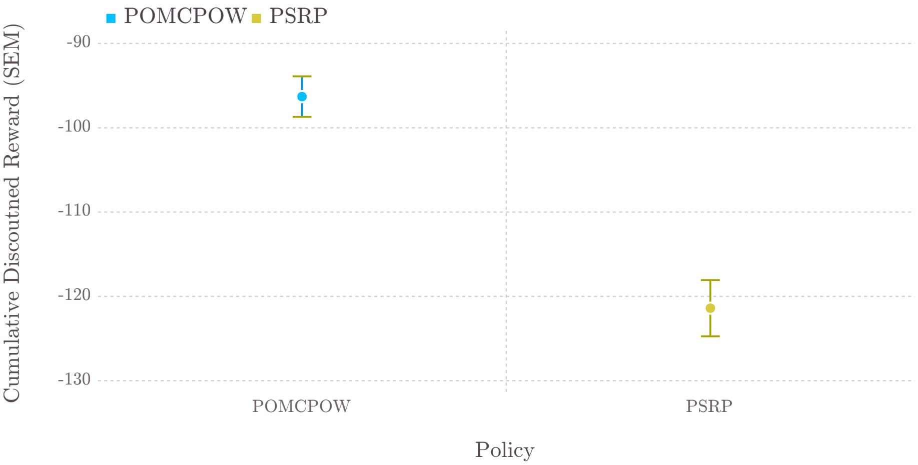


Evaluation: Localization and Planning – Objective Value

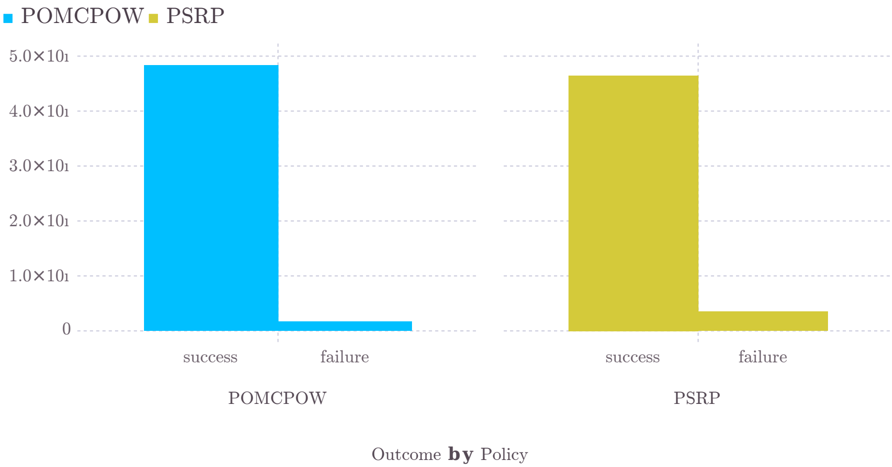
Distribution of the return for 1000 simulations per policy:



Evaluation: Motion Planning with Latent Human Intentions



Evaluation: Motion Planning with Latent Human Intentions


■ POMCPOW ■ PSRP
5.0×10^{1}
4.0×10^{1}
3.0×10^{1}
2.0×10^{1}
1.0×10^{1}
success
failure
POMCPOW
success
failure
PSRP
Outcome **by** Policy