Partially Observable Markov Decision Processes (POMDPs)

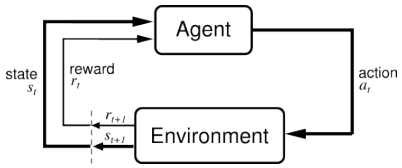
Pieter Abbeel
UC Berkeley EECS

Many slides adapted from Jur van den Berg

Outline

- POMDPs
- Separation Principle / Certainty Equivalence
- Locally Optimal Solutions for POMDPs

Markov Decision Process (S, A, T, R, H)



Given

- S: set of states
- A: set of actions
- T: S x A x S x {0,1,…,H} → [0,1], $T_t(s,a,s') = P(s_{t+1} = s' \mid s_t = s, a_t = a)$
- R: S x A x S x {0, 1, …, H} → $\Re$ $R_t(s,a,s')$ = reward for $(s_{t+1} = s', s_t = s, a_t = a)$
- H: horizon over which the agent will act

Goal:

- Find π : S x {0, 1, …, H} → A that maximizes expected sum of rewards, i.e.,

$$\pi^* = \arg\max_{\pi} \mathrm{E}[\sum_{t=0}^{H} R_t(S_t, A_t, S_{t+1})|\pi]$$

POMDP

= MDP

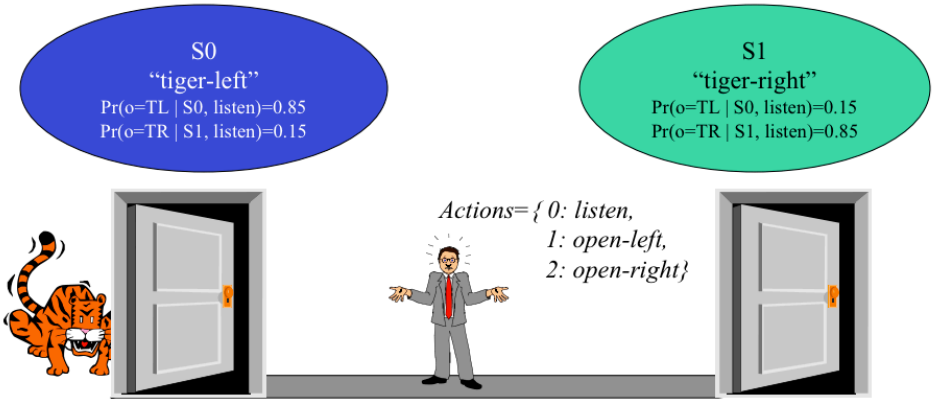
BUT

don't get to observe the state itself, instead get sensory measurements

We know how to compute a probability distribution over the state from past weeks.

Now: what action to take given current probability distribution rather than given current state.

POMDPs: Tiger Example



Reward Function

- *Penalty for wrong opening: -100*
- *Reward for correct opening: +10*
- *Cost for listening action: -1*

Observations

- *to hear the tiger on the left (TL)*
- *to hear the tiger on the right(TR)*

POMDPs

- Canonical solution method 1:
  - Run value iteration, but now the state space is the space of probability distributions
  - → value and optimal action for every possible probability distribution
  - → will automatically trade off information gathering actions versus actions that affect the underlying state
- Canonical solution method 2:
  - Search over sequences of actions with limited lookahead
  - Branching over actions and observations
- Canonical solution method 3:
  - Plan in the MDP
  - Run probabilistic inference (filtering) to track probability distribution
  - Choose optimal action for MDP for what is currently the most likely state

POMDPs

- Bad news: generally intractable (Papadimitriou and Tsitsiklis, 1987)

- Readings:
  - Planning and acting in partially Observable stochastic domains, Leslie P. Kaelbling
  - Optimal Policies for partially Observable Markov Decision Processes, Anthony R.Cassandra 1995
  - Hierarchical Methods for Planning under Uncertainty Joelle Pineau
  - POMDP's tutorial in Tony's POMDP Page:
    - http://www.cs.brown.edu/research/ai/pomdp/

Outline

- POMDPs
- Separation Principle / Certainty Equivalence
- Locally Optimal Solutions for POMDPs

Separation Principle

- Assume: $x_{t+1} = Ax_t + Bu_t + w_t \quad w_t \sim \mathcal{N}(0, Q_t)$
  $z_t = Cx_t + v_t \quad v_t \sim \mathcal{N}(0, R_t)$
- Goal: $\text{minimize } \mathrm{E}\left[\sum_{t=0}^{H} u_t^\top U_t u_t + x_t^\top X_t x_t\right]$
- Then, optimal control policy consists of:

  1. Offline/Ahead of time: Run LQR to find optimal control policy for fully observed case, which gives sequence of feedback matrices $K_1, K_2, \ldots$

  2. Online: Run Kalman filter to estimate state, and apply control
  $u_t = K_t \mu_{t|0:t}$

Separation Principle

- Intuition: consider Kalman filter equations:

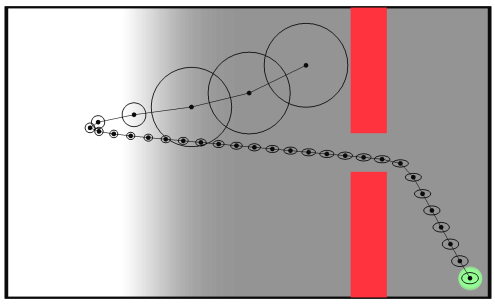
$$\begin{aligned}
\mu_{t+1|0:t} &= A_t \mu_{t|0:t} + B_t u_t \\
\Sigma_{t+1|0:t} &= A_t \Sigma_{t|0:t} A_t^\top + Q_t \\
K_{t+1} &= \Sigma_{t+1|0:t} C_{t+1}^\top (C_{t+1} \Sigma_{t+1|0:t} C_{t+1}^\top + R_{t+1})^{-1} \\
\mu_{t+1|0:t+1} &= \mu_{t+1|0:t} + K_{t+1}(z_{t+1} - (C_{t+1}\mu_{t+1|0:t} + d)) \\
\Sigma_{t+1|0:t+1} &= (I - K_{t+1} C_{t+1}) \Sigma_{t+1|0:t}
\end{aligned}$$

→ no way to influence the uncertainty through choice of actions

- More intricate math: can show that cost consists of 3 contributions:
  - 1. Cost for no-noise, fully observed case
  - 2. Additional cost due to dynamics noise [we saw this]
  - 3. Additional cost due to uncertainty about state [try to work out that this third term decomposes and is not affected by control inputs]

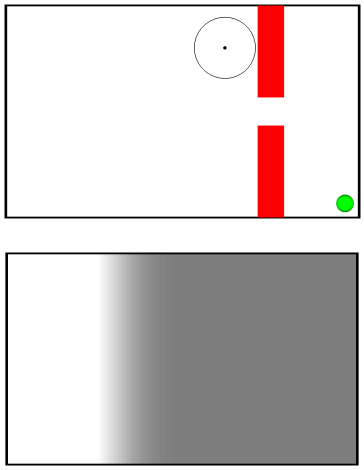
Outline

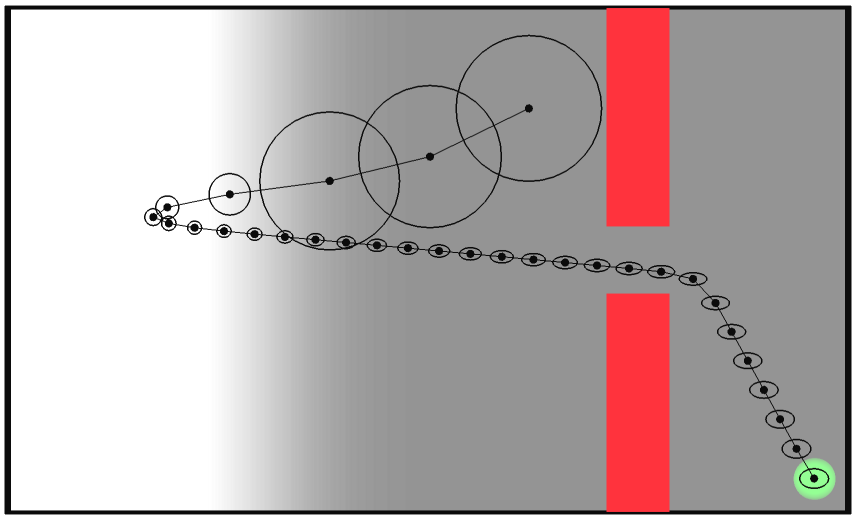
- POMDPs
- Separation Principle / Certainty Equivalence
- Locally Optimal Solutions for POMDPs



Motion Planning under Uncertainty

- Example
  - Initial uncertain state
  - Reach goal region, avoid obstacles
  - Motion and sensing uncertainty
  - Minimize expected cost
    - Deviation from goal
    - Control effort
    - Probability of collision





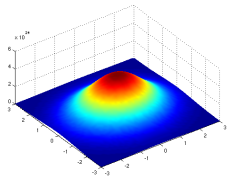
Motion Planning under Uncertainty

- Stochastic Motion and Observation Model

$$\mathbf{x}_{t+1} = \mathbf{f}[\mathbf{x}_t, \mathbf{u}_t, \mathbf{m}_t], \qquad \mathbf{m}_t \sim \mathcal{N}[\mathbf{0}, I],$$
$$\mathbf{z}_t = \mathbf{h}[\mathbf{x}_t, \mathbf{n}_t], \qquad \mathbf{n}_t \sim \mathcal{N}[\mathbf{0}, I],$$

  - Non-linear
- User-defined objective / cost function
- Plan path (or control policy) for robot that minimizes expected cost

General POMDPs



- Belief: $b[\mathbf{x}_t] = p[\mathbf{x}_t | \mathbf{u}_0, \ldots, \mathbf{u}_{t-1}, \mathbf{z}_1, \ldots, \mathbf{z}_t].$
- Bayesian Filter / Belief Dynamics:

$$b[\mathbf{x}_{t+1}] = \eta\, p[\mathbf{z}_{t+1} | \mathbf{x}_{t+1}] \int p[\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{u}_t]\, b[\mathbf{x}_t]\, d\mathbf{x}_t, \qquad \mathbf{b}_{t+1} = \beta[\mathbf{b}_t, \mathbf{u}_t, \mathbf{z}_{t+1}].$$

- Find policy minimizing cost function:

$$\mathbf{u}_t = \pi_t(\mathbf{b}_t) \qquad \mathop{\mathrm{E}}_{\mathbf{z}_1, \ldots, \mathbf{z}_\ell} \left[c_\ell[\mathbf{b}_\ell] + \sum_{t=0}^{\ell-1} c_t[\mathbf{b}_t, \mathbf{u}_t] \right],$$

- Value iteration:

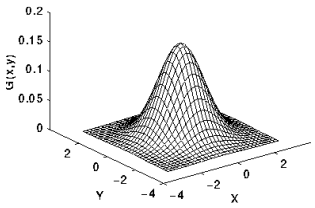
$$v_\ell[\mathbf{b}_\ell] = c_\ell[\mathbf{b}_\ell]$$
$$v_t[\mathbf{b}_t] = \min_{\mathbf{u}_t} (c_t[\mathbf{b}_t, \mathbf{u}_t] + \mathop{\mathrm{E}}_{\mathbf{z}_{t+1}} [v_{t+1}[\beta[\mathbf{b}_t, \mathbf{u}_t, \mathbf{z}_{t+1}]]])$$
$$\pi_t[\mathbf{b}_t] = \mathop{\mathrm{argmin}}_{\mathbf{u}_t} (c_t[\mathbf{b}_t, \mathbf{u}_t] + \mathop{\mathrm{E}}_{\mathbf{z}_{t+1}} [v_{t+1}[\beta[\mathbf{b}_t, \mathbf{u}_t, \mathbf{z}_{t+1}]]]),$$

Literature

- Bry and Roy
- Du Toit and Burdick
- Li and Todorov
- Platt, Tedrake, Kaelbling, Lozano-Perez
- van den Berg, Patil, Alterovitz, Goldberg, Abbeel

Locally Optimal Solutions

- Belief is Gaussian
  - $\mathbf{b}_t = (\hat{\mathbf{x}}_t, \Sigma_t)$,
- Belief dynamics
  - Extended Kalman Filter



Extended Kalman Filter

- Standard form:

$$\begin{aligned}\mu_{t+1|0:t} &= A_t\mu_{t|0:t} + B_t u_t + w_t \\ \Sigma_{t+1|0:t} &= A_t\Sigma_{t|0:t}A_t^\top + Q_t\end{aligned}$$

$$\begin{aligned}K_{t+1} &= \Sigma_{t+1|0:t}C_{t+1}^\top(C_{t+1}\Sigma_{t+1|0:t}C_{t+1}^\top + R_{t+1})^{-1} \\ \mu_{t+1|0:t+1} &= \mu_{t+1|0:t} + K_{t+1}(z_{t+1} - (C_{t+1}\mu_{t+1|0:t} + d)) \\ \Sigma_{t+1|0:t+1} &= (I - K_{t+1}C_{t+1})\Sigma_{t+1|0:t}\end{aligned}$$

- More compactly:

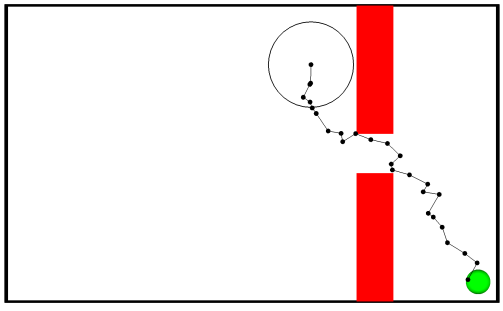
$$(\mu_{t+1|0:t+1}, \Sigma_{t+1|0:t+1}) = f(\mu_{t|0:t}, \Sigma_{t|0:t}, w_t, v_t)$$

  - Or:

$$b_{t+1} = f(b_t, w_t, v_t)$$

Value Function

- Quadratic value function $v_t[\mathbf{b}] = \frac{1}{2}\mathbf{b}^T S_t \mathbf{b} + \mathbf{b}^T \mathbf{s}_t + s_t,$
  - Approximately valid around given nominal trajectory in belief space $(\bar{\mathbf{b}}_0, \bar{\mathbf{u}}_0, \ldots, \bar{\mathbf{b}}_\ell, \bar{\mathbf{u}}_\ell),$



Solution Methods

- Iterative LQR: (most common in literature)
  - Quadratize $c_t[\mathbf{b}_t]$ around nominal belief $\underline{\mathbf{b}}_t$
  - Linearize belief dynamics around nominal belief $\underline{\mathbf{b}}_t$
  - Backward recursion

$$v_\ell[\mathbf{b}_\ell] = c_\ell[\mathbf{b}_\ell]$$
$$v_t[\mathbf{b}_t] = \min_{\mathbf{u}_t}(c_t[\mathbf{b}_t, \mathbf{u}_t] + \mathop{\mathrm{E}}_{\mathbf{z}_{t+1}}\left[v_{t+1}[\beta[\mathbf{b}_t, \mathbf{u}_t, \mathbf{z}_{t+1}]]\right])$$

  - Execute current control policy
  - Iterate!

- Just as well: nonlinear optimization through sequential convex programming

Extended Kalman Filter

- Standard form:

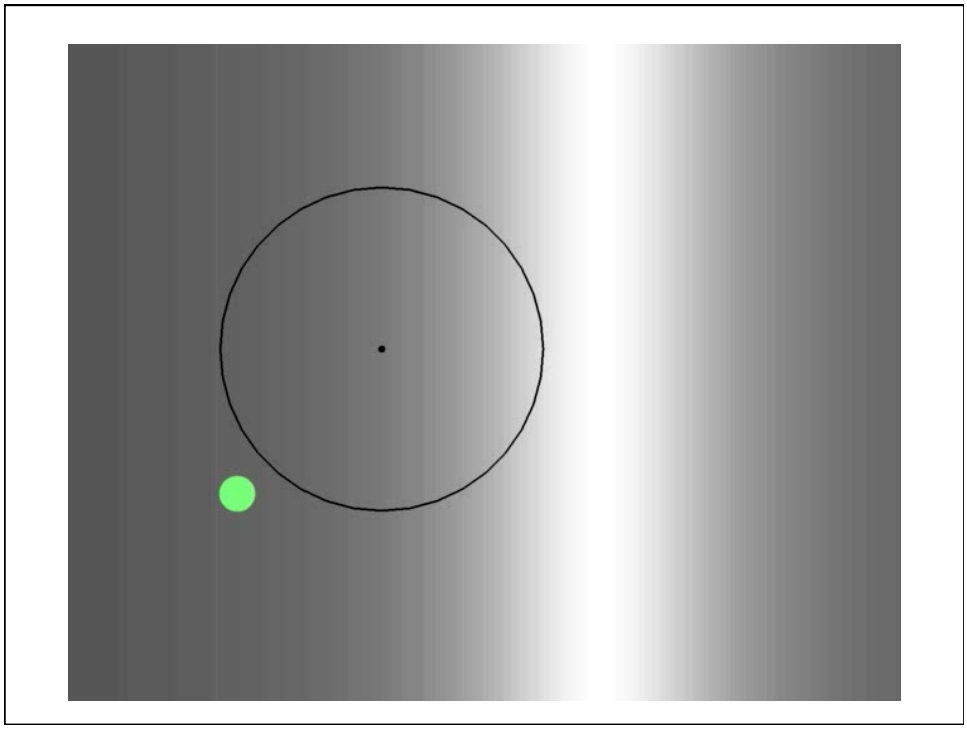
$$\begin{aligned}\mu_{t+1|0:t} &= A_t\mu_{t|0:t} + B_t u_t + w_t \\ \Sigma_{t+1|0:t} &= A_t\Sigma_{t|0:t}A_t^\top + Q_t\end{aligned}$$

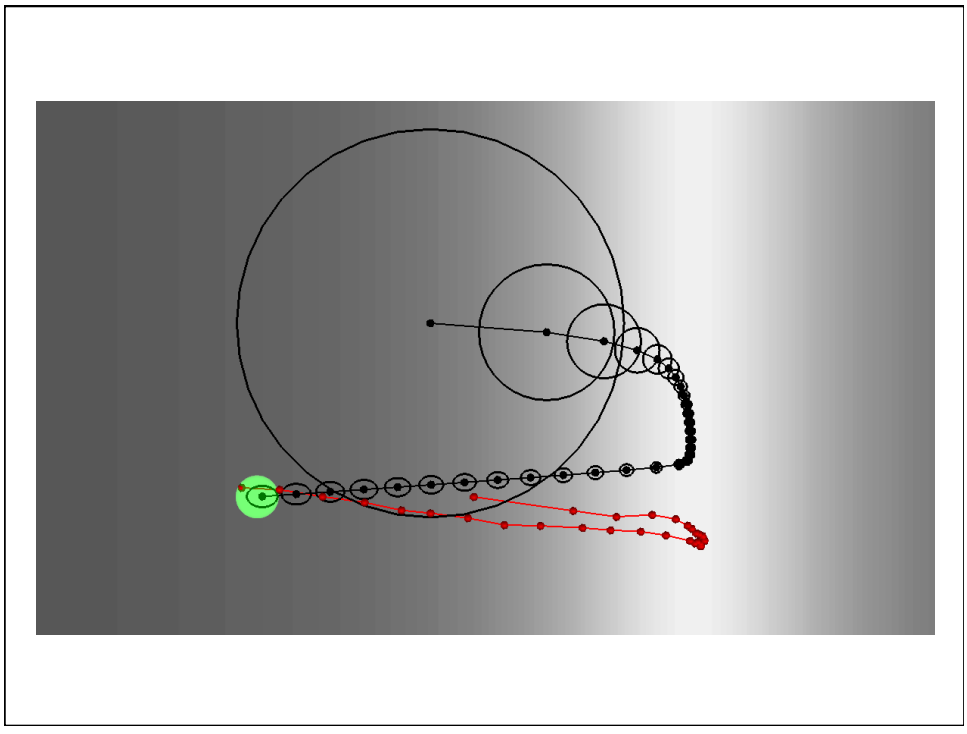
$$\begin{aligned}K_{t+1} &= \Sigma_{t+1|0:t}C_{t+1}^\top(C_{t+1}\Sigma_{t+1|0:t}C_{t+1}^\top + R_{t+1})^{-1} \\ \mu_{t+1|0:t+1} &= \mu_{t+1|0:t} + K_{t+1}(z_{t+1} - (C_{t+1}\mu_{t+1|0:t} + d)) \\ \Sigma_{t+1|0:t+1} &= (I - K_{t+1}C_{t+1})\Sigma_{t+1|0:t}\end{aligned}$$

- Why will we not end up with certainty equivalent solution?

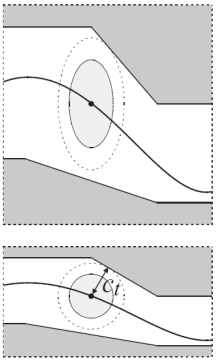
$$A_t(\mu_{t|0:t}), B_t(\mu_{t|0:t}), C_t(\mu_{t|0:t}), Q_t(\mu_{t|0:t}), R_t(\mu_{t|0:t})$$

→ Uncertainty is influenced by trajectory! (and linearization captures this)



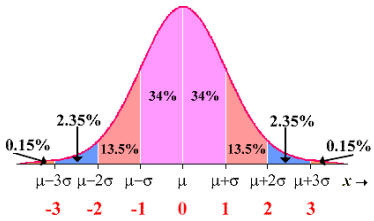


Obstacles



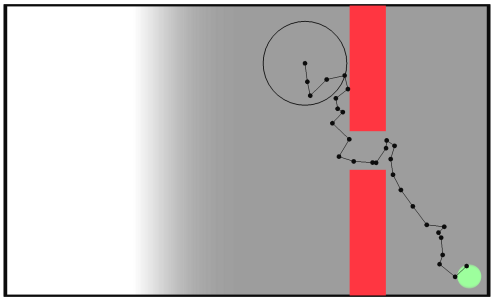
- Account for probability of collision
- Collision checker: number of standard deviations before obstacle is hit $\sigma[\mathbf{b}_t]$
- Probability of collision-free $\gamma[n/2, \sigma[\mathbf{b}_t]^2/2],$
- Cost:

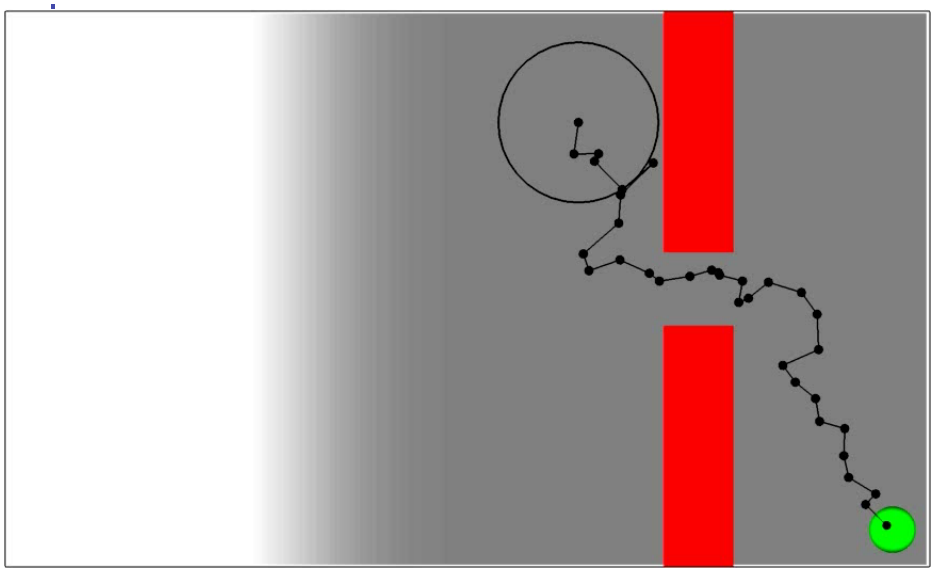
$$f[\sigma[\mathbf{b}]] = -\log \bar{\gamma}[n/2, \sigma[\bar{\mathbf{b}}]^2/2]$$



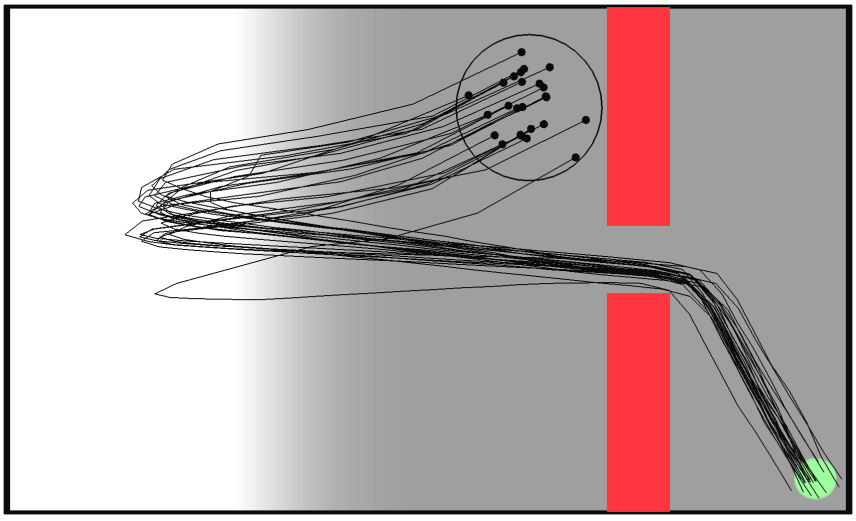
Experimental Results: Light/dark

- Dynamics: $\mathbf{x}_{t+1} = \mathbf{f}[\mathbf{x}_t, \mathbf{u}_t, \mathbf{m}_t] = \mathbf{x}_t + \mathbf{u}_t + M[\mathbf{u}_t] \cdot \mathbf{m}_t,$
- Sensing: $\mathbf{z}_t = \mathbf{h}[\mathbf{x}_t, \mathbf{n}_t] = \mathbf{x}_t + N[\mathbf{x}_t] \cdot \mathbf{n}_t,$
- Cost: $c_\ell[\mathbf{b}_\ell] = \hat{\mathbf{x}}_\ell^T Q_\ell \hat{\mathbf{x}}_\ell + \mathrm{tr}[Q_\ell \Sigma_\ell], \quad c_t[\mathbf{b}_t, \mathbf{u}_t] = \mathbf{u}_t^T R_t \mathbf{u}_t + \mathrm{tr}[Q_t \Sigma_t] + f[\sigma[\mathbf{b}_t]],$

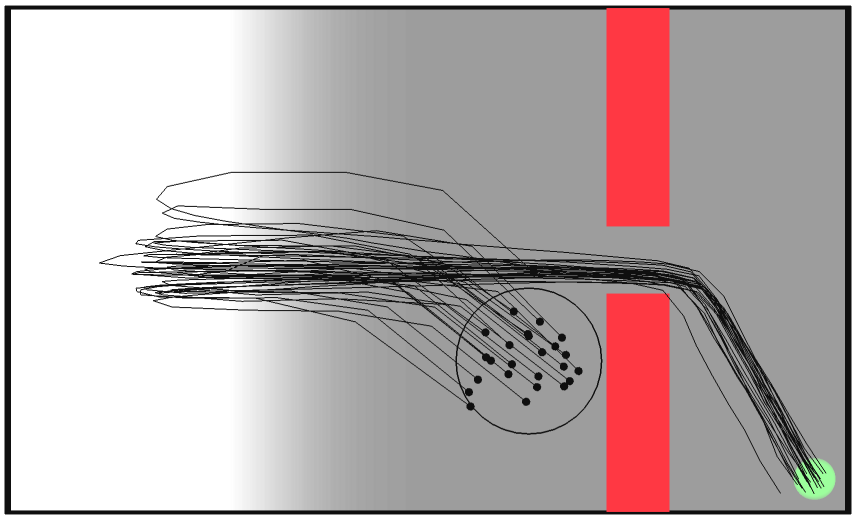




Convergence: 2.75s



Traces of 25 simulations with synthetic noise



Different initial belief, same control policy

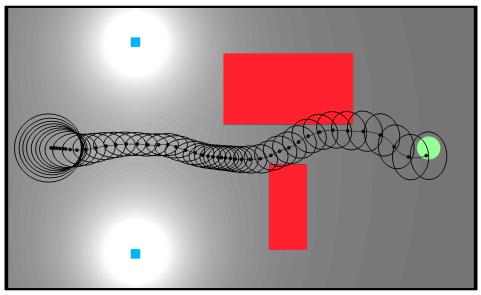
Experimental Results: Car robot

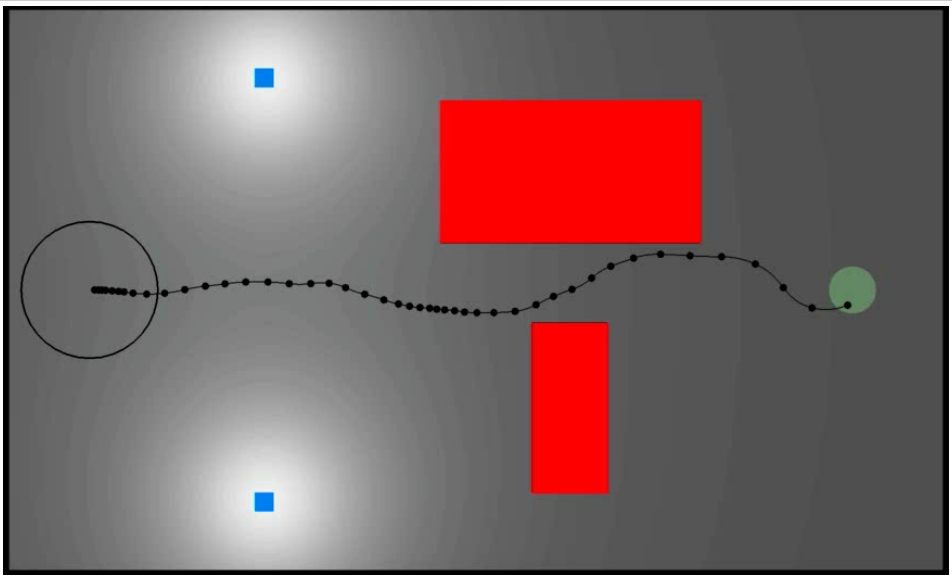
- Dynamics:

$$\mathbf{x}_{t+1} = \mathbf{f}[\mathbf{x}_t, \mathbf{u}_t, \mathbf{m}_t] = \begin{bmatrix} x_t + \tau v_t \cos\theta_t \\ y_t + \tau v_t \sin\theta_t \\ \theta_t + v_t \tan(\phi)/d \\ v_t + \tau a \end{bmatrix} + M\mathbf{m}_t,$$

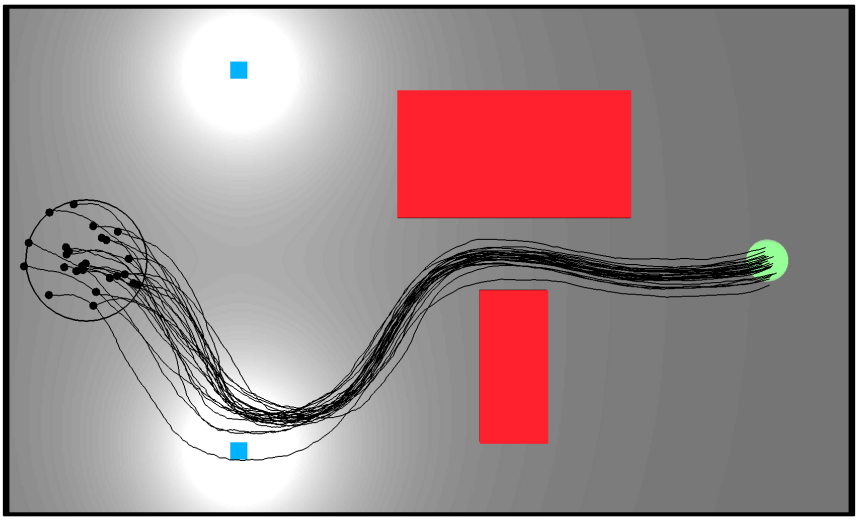
- Sensing:

$$\mathbf{z}_t = \mathbf{h}[\mathbf{x}_t, \mathbf{n}_t] = \begin{bmatrix} 1/((x_t - \check{x}_1)^2 + (y_t - \check{y}_1)^2 + 1) \\ 1/((x_t - \check{x}_2)^2 + (y_t - \check{y}_2)^2 + 1) \\ v_t \end{bmatrix} + N\mathbf{n}_t,$$

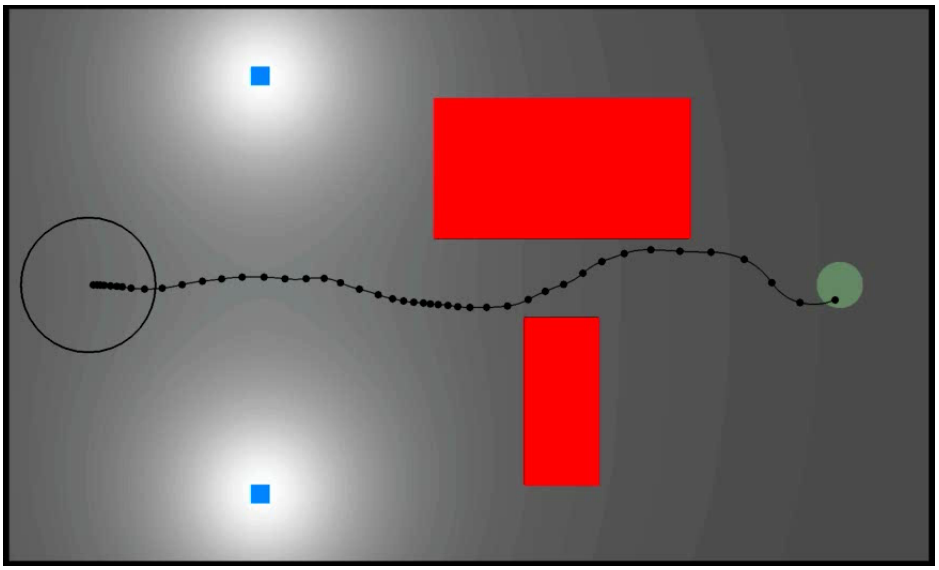




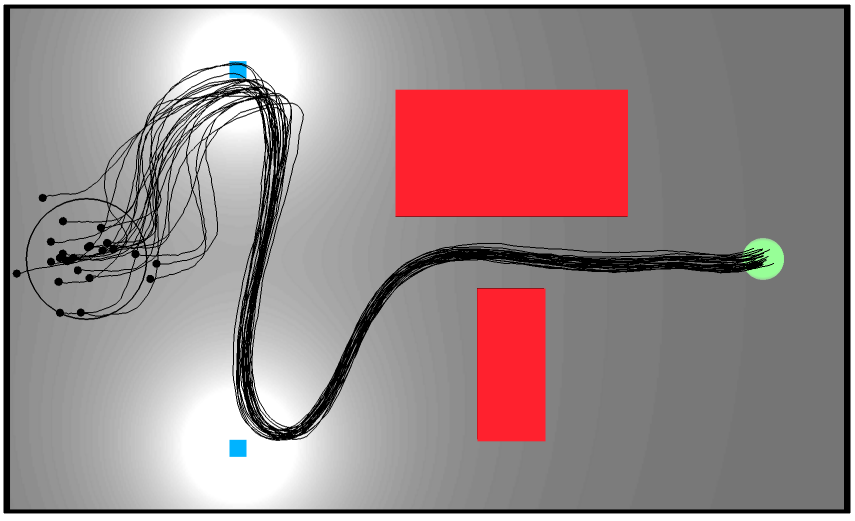
Convergence: 15.3s



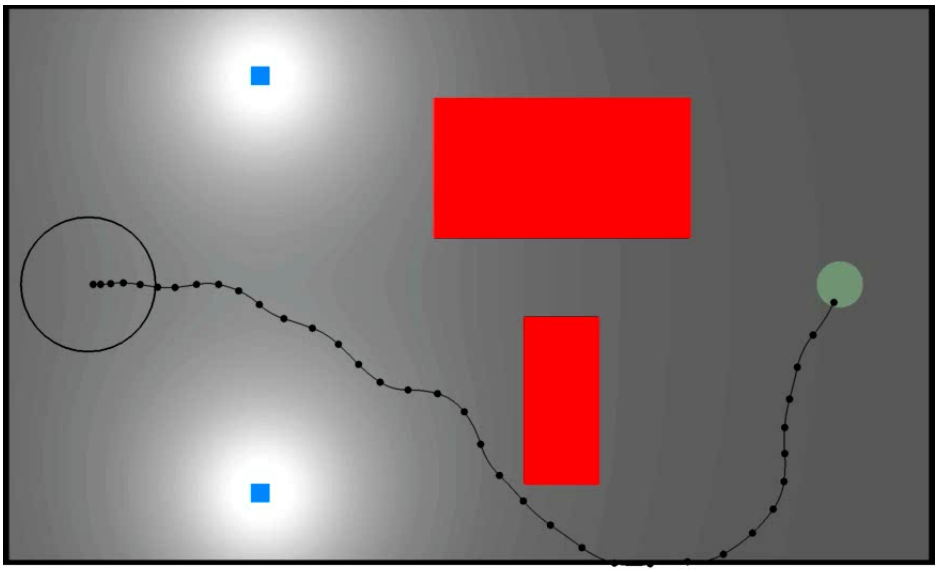
Traces of 25 simulations with synthetic noise



Uncertainty 10 times more heavily penalized



Traces of 25 simulations with synthetic noise



Initial trajectory in different homotopy class

Conclusion, Future work

- General approach for motion planning under uncertainty
- Continuous POMDP
- Locally optimal solution in polynomial time
- Current research directions
  - Nonlinear optimization rather than iterative LQR
  - Better handling of collision probability
  - Better handling of field of view of sensors
  - Distinction between maximum likelihood observation assumption and incorporating noise in the optimization