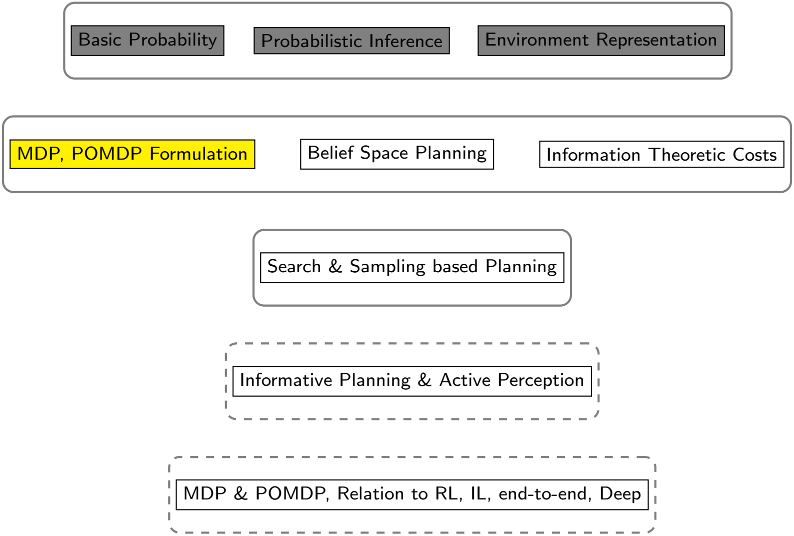
Lecture #2

MDP and POMDP Formulation

Course Diagram



Objectives of this Lecture

- Introduce formulation of decision making problems.
- Distinguish between Markov Decision Process (MDP) and Partially Observable MDP (POMDP) problems.

! Some of the material was adapted from David Silver (UCL, DeepMind), Mykel Kochenderfer (Stanford), Pieter Abbeel (Berkeley), Nikolay Atanasov (UCSD) and others.

Recall from Previous Lecture

State Transition and Observation models

- Motion model or (state transition model):

$$X_{k+1} = f(X_k, u_k, w_k) \sim \mathbb{P}_T(X_{k+1} \mid X_k, u_k)$$

- Observation model (Measurement likelihood):

$$z_k = h(X_k, v_k) \sim \mathbb{P}_Z(z_k \mid X_k)$$

- Discrete time domain
- For *given* functions $f(.)$ and $h(.)$, stochasticity is due to motion (process) and observation noise w_k and v_k.
- w_k and v_k are random variables with known/learned probability density functions (pdf). Common assumption - statistical independence: $\forall k$: $w_k \perp\!\!\!\perp v_k$; $\forall j \neq k$: $w_j \perp\!\!\!\perp w_k$
- More generally, the probabilistic models [e.g. $\mathbb{P}_T(X_{k+1} \mid X_k, u_k)$ and $\mathbb{P}_Z(z_k \mid X_k)$] could be learned from data.

Recall from Previous Lecture

Gaussian State Transition and Observation models

- For an additive Gaussian noise, with $w_k \sim \mathcal{N}(0, \Sigma_w)$ and $v_k \sim \mathcal{N}(0, \Sigma_v)$:

$$\begin{aligned} \mathbb{P}_T(X_{k+1} \mid X_k, u_k) &= \frac{1}{\sqrt{\det 2\pi\Sigma_w}} \exp\{-\frac{1}{2}\|X_{k+1} - f(X_k, u_k)\|^2_{\Sigma_w}\} \\ \mathbb{P}_Z(z_k \mid X_k) &= \frac{1}{\sqrt{\det 2\pi\Sigma_v}} \exp\{-\frac{1}{2}\|z_k - h(X_k)\|^2_{\Sigma_v}\} \end{aligned}$$

where $\|a\|^2_\Sigma \doteq a^T \Sigma^{-1} a$ is the squared Mahalanobis norm.

Recall from Previous Lecture

Bayesian Inference

Apply Bayes rule, chain rule and use causality:

- Recursive formulation:

$$\mathbb{P}(X_k \mid H_k^-) = \int_{X_{k-1}} \mathbb{P}(X_k \mid X_{k-1}, a_{k-1}, H_{k-1}) \mathbb{P}(X_{k-1} \mid H_{k-1}) dX_{k-1}$$

$$b[X_k] = \frac{\mathbb{P}(z_k \mid X_k) \mathbb{P}(X_k \mid H_k^-)}{\mathbb{P}(z_k \mid H_k^-)}$$

- Smoothing formulation (e.g. $X_{0:k} \doteq \{X_0, \ldots, X_k\}$):

$$\mathbb{P}(X_{0:k} \mid H_k^-) = \mathbb{P}(X_k \mid X_{k-1}, a_{k-1}, H_{k-1}) \mathbb{P}(X_{0:k-1} \mid H_{k-1})$$

$$b[X_{0:k}] = \frac{\mathbb{P}(z_k \mid X_k) \mathbb{P}(X_{0:k} \mid H_k^-)}{\mathbb{P}(z_k \mid H_k^-)}$$

Alternative Notations (AI)

State Transition and Observation models

- Notations:
  - State: s or S (instead of X)
  - Observation: o (instead of z or y)
- Models:
  - State transition model: $T(s, a, s') = \mathbb{P}(s' \mid s, a)$
    Defines the probability of being in state s' after taking an action a in state s.
  - Observation model:
    - Probability of observing o given state s: $O(s, o) = \mathbb{P}(o \mid s)$
    - In some formulations, the observation can also depend on the action a: $O(s, a, o) = \mathbb{P}(o \mid s, a)$

Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

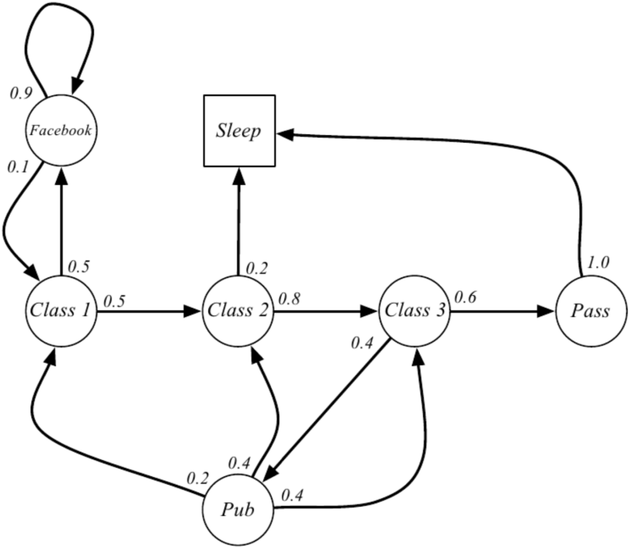
Markov Chain

A **Markov Chain** is a stochastic process defined by a tuple $(\mathcal{X}, \mathbb{P}_0, \mathbb{P}_T)$:

- $\mathcal{X}$ is a discrete/continuous/hybrid state space
- $\mathbb{P}_0$ is a prior pmf/pdf
- $\mathbb{P}_T(X' \mid X)$ is a conditional pmf/pdf representing the transition model

  In the (finite-dimensional) discrete case, the transition pmf can be summarized by a matrix $\mathbb{P}_{ij} \doteq \mathbb{P}_T(X_{t+1} = i \mid X_t = j)$

Example: Student Markov Chain ;)

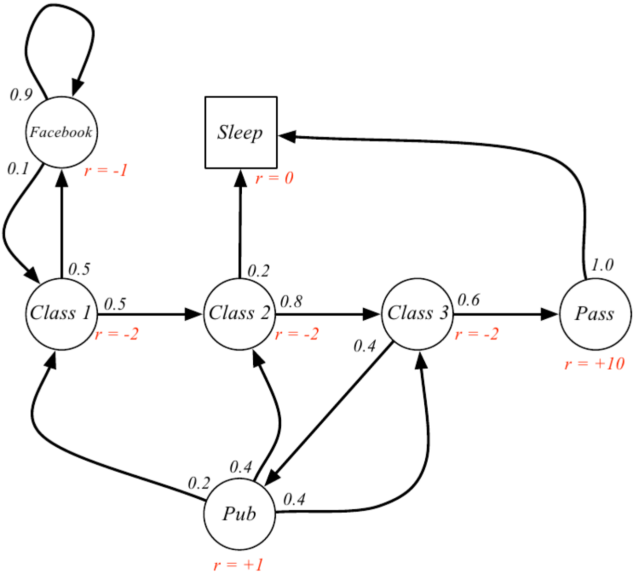


Markov Reward Process (MRP)

A **Markov Reward Process** (MRP) is a Markov Chain with state costs (rewards) defined by a tuple $(\mathcal{X}, \mathbb{P}_0, \mathbb{P}_T, r, \gamma)$:

- $\mathcal{X}, \mathbb{P}_0$ and $\mathbb{P}_T$ are defined as in Markov Chain
- $r(X)$ is a function specifying the reward of state $X \in \mathcal{X}$
  - Alternatively: cost function $c(X)$
- $\gamma \in [0, 1]$ is a discount factor

Example: Student Markov Reward Process ;)



Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

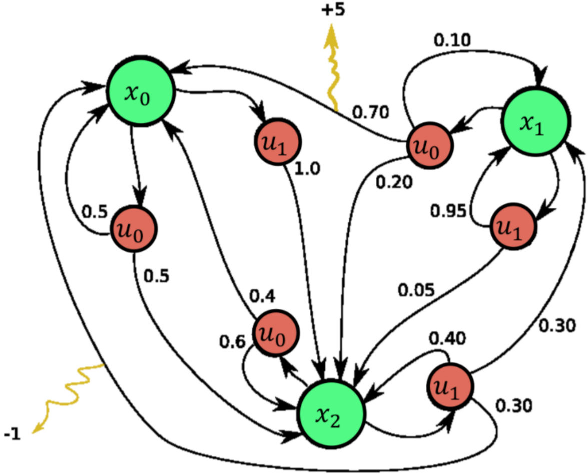
Problem Variations & Illustrative Examples

Markov Decision Process (MDP)

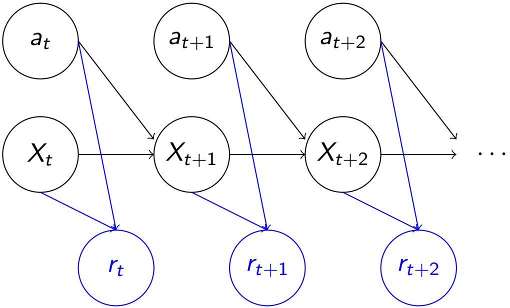
A **Markov Decision Process** (MDP) is a Markov Reward Process with controlled transitions defined by a tuple $(\mathcal{X}, \mathcal{A}, \mathbb{P}_T, r, \gamma)$:

- $\mathcal{X}$ is a discrete/continuous state space
- $\mathcal{A}$ is a discrete/continuous action set
- $\mathbb{P}_T(X' \mid X, a)$ is the transition (motion) model
- $r(X, a)$ is a function specifying the reward of applying action $a \in \mathcal{A}$ in state $X \in \mathcal{X}$
  - Alternatively: cost function $c(X, a)$
  - We want to **minimize** cost $c(X, a)$, or **maximize** reward $r(X, a)$
  - In this course, we shall use both settings
- $\gamma \in [0, 1]$ is a discount factor

Example: Markov Decision Process



Graphical View of MDP



Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

Objective Function

- Consider planning session at time $t \doteq 0$
- Denote the set of possible actions at time i by $\mathcal{A}_i$
- Consider a sequence of actions of length T,

$$a_{0:T-1} \doteq \{a_0, \dots, a_{T-1}\}, \quad \text{with} \quad a_i \in \mathcal{A}_i$$

Objective Function

- Consider planning session at time $t \doteq 0$
- Denote the set of possible actions at time i by $\mathcal{A}_i$
- Consider a sequence of actions of length T,

$$a_{0:T-1} \doteq \{a_0, \ldots, a_{T-1}\}, \quad \text{with} \quad a_i \in \mathcal{A}_i$$

- Objective function - expected cumulative reward (cost) starting from state $X_0 \in \mathcal{X}$ for an action sequence $a_{0:T-1}$:

$$J(X_0, a_{0:T-1}) \doteq \mathbb{E}\{\sum_{t=0}^{T-1} r(X_t, a_t) + r_T(X_T)\}$$

  - r_T is the terminal reward (cost)
  - **Q:**Expectation over what?

Control Policy & Value Function

Admissible control policy: a sequence $\pi_{0:T-1}$ of functions π_t that map **any** state $X_t \in \mathcal{X}$ to a feasible action/control $a_t \in \mathcal{A}(X_t)$, i.e.

$$\pi_t : X_t \mapsto a_t \quad \forall X_t \in \mathcal{X}$$

Value function: The expected cumulative reward, of a policy π applied to an MDP $(\mathcal{X}, \mathcal{A}, \mathbb{P}_T, r, \gamma)$ starting from state $X \in \mathcal{X}$ at time $t = 0$:

- Finite horizon:

$$\mathtt{V}_0^{\pi}(X) \doteq \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t, a_t) + r_T(X_T) \mid X_0 = X, a_t = \pi_t(X_t)\right]$$

- Discounted infinite-horizon:

$$\mathtt{V}_0^{\pi}(X) \doteq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, a_t) \mid X_0 = X, a_t = \pi(X_t)\right]$$

- Stochastic policy: change to $a_t \sim \pi(x_t)$

As $T \to \infty$, optimal policies become stationary, i.e. $\pi \equiv \pi_0 = \pi_1 = \ldots$.

Discount Factor

The discount factor γ specifies the present value of future costs:

- γ close to 0 leads to myopic (greedy) evaluation
- γ close to 1 leads to non-myopic (long horizon) evaluation
- Mathematically convenient, as it avoids infinite cumulative rewards/costs as $T \to \infty$

Optimal Policy and Action Sequence

Recall:

$$
\begin{aligned}
\mathtt{V}_0^{\pi}(X) &\doteq \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t, a_t) + r_T(X_T) \mid X_0 = X, a_t = \pi_t(X_t)\right] \\
J(X_0, a_{0:T-1}) &\doteq \mathbb{E}\{\sum_{t=0}^{T-1} r(X_t, a_t) + r_T(X_T)\}
\end{aligned}
$$

Optimal policy $\pi^\star$ is the one that maximizes the expected cumulative reward (or minimizes the expected cumulative cost):

$$\pi^\star = \arg\max_{\pi} \mathtt{V}_0^{\pi}(x)$$

Optimal action sequence $a^\star_{0:T-1}$**:**

$$a^\star_{0:T-1} = \arg\max_{a_{0:T-1}} J(X_0, a_{0:T-1})$$

Comparison of Markov Models

	Observed	Partially Observed
Uncontrolled ("Passive")	Markov Chain / MRP	HMM
Controlled ("Active")	MDP	POMDP

- Hidden Markov Model (HMM) = Markov Chain & Partial Observability
- Markov Decision Process (MDP) = Markov Chain & Control
- Partially Observable Markov Decision Process (POMDP) =
  - = Markov Chain & Partial Observability & Control
  - = HMM & Control
  - = MDP & Partial Observabilty

Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

Partially Observable Markov Decision Process (POMDP)

A **Partially Observable Markov Decision Process** (POMDP) is a Markov Decision Process with hidden states.

A POMDP is a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathbb{P}_0, \mathbb{P}_T, \mathbb{P}_Z, r, \gamma)$

- $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{Z}$ are state, action and observation spaces
  Can be discrete, continuous, or hybrid
- $\mathbb{P}_0$ is a priori pmf/pdf
- $\mathbb{P}_T(X_{k+1} \mid X_k, a_k)$ and $\mathbb{P}_Z(z \mid X)$ are state transition and observation models
- $r(X, a)$ or $r(b, a)$ are functions specifying the reward (cost) of applying action/control $a \in \mathcal{A}$ in state $X \in \mathcal{X}$
- $\gamma \in [0, 1]$ is the discount factor

Posterior Belief and Bayesian Inference

- Posterior belief at time instant k:

$$b[X_k] \doteq \mathbb{P}(X_k \mid a_{0:k-1}, z_{1:k}) \equiv \mathbb{P}(X_k \mid H_k)$$

where history H_k and propagated history H_k^- are defined as

$$H_k \doteq \{a_{0:k-1}, z_{1:k}\} = H_k^- \cup \{z_k\} \quad , \quad H_k^- \doteq \{a_{0:k-1}, z_{1:k-1}\}$$

- From previous lecture - Bayesian inference:
  - Joint distribution:

$$\mathbb{P}(X_{0:k}, a_{0:k-1}, z_{1:k}) = \mathbb{P}(X_0) \prod_{i=1}^{k} \mathbb{P}_T(X_i \mid X_{i-1}, a_{i-1}) \mathbb{P}_Z(z_i \mid X_i)$$

  - Bayes filter:

$$\mathbb{P}(X_k \mid H_k) = \frac{1}{\mathbb{P}(z_k \mid H_k^-)} \mathbb{P}_Z(z_k \mid X_k) \int_{X_{k-1}} \mathbb{P}_T(X_k \mid X_{k-1}, a_{k-1}) \mathbb{P}(X_{k-1} \mid H_{k-1}) dX_{k-1}$$

Sufficient Statistics

- The posterior belief $b[X_k] \doteq \mathbb{P}(X_k \mid H_k)$ is a *sufficient statistics* for X_k, under the undertaken assumptions (Markov, measurement and process noise statistical independence).

- Sufficient statistics:
  - The data/information available to the robot at time k to determine its action/control a_k is $H_k \doteq \{a_{0:k-1}, z_{1:k}\}$.
  - A statistic $\zeta_k = s(H_k)$ is a function of the information available at time k to infer the state X_k.
  - The statistic $\zeta_k = s(H_k)$ is **sufficient** for X_k if the conditional distribution of X_k given the statistic ζ_k does not depend on H_k.

- In other words: $b[X_k]$ is a compact representation of H_k.

- **Example:** Two first moments for a Gaussian distribution.

Value Function and Policy

- Recall POMDP is a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathbb{P}_0, \mathbb{P}_T, \mathbb{P}_Z, r, \gamma)$, where prior distribution/belief is over state X at planning time $t \doteq 0$, i.e. $b_0 \doteq \mathbb{P}_0(X)$.

- Policy: $\pi : b \mapsto a$ for all possible beliefs

- Value function (e.g. discounted infinite horizon):

$$\mathtt{V}_0^{\pi}(b_0) \doteq \mathbb{E}\{\sum_t \gamma^t r(b_t, a_t) \mid a_t = \pi(b_t)\}$$

  A particular case is (more soon):

$$\mathtt{V}_0^{\pi}(b_0) \doteq \mathbb{E}\{\sum_t \gamma^t r(X_t, a_t) \mid X_0 \sim b_0, a_t = \pi(b_t)\}$$

- As previously, policy could be also stochastic $(a_t \sim \pi(b_t))$

Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

Belief MDP

- The Bayes filter tracks and updates sufficient statistics (the belief). In general: $b_k = \psi(b_{k-1}, a_{k-1}, z_k)$
  - E.g. in a recursive formulation: $b[X_k] = \eta \mathbb{P}_Z(z_k \mid X_k) \int_{X_{k-1}} \mathbb{P}_T(X_k \mid X_{k-1}, a_{k-1}) b[X_{k-1}] dX_{k-1}$

- Because the posterior belief is a sufficient statistic for the state, we can convert a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathbb{P}_0, \mathbb{P}_T, \mathbb{P}_Z, r, \gamma)$ into an equivalent **belief MDP**, $(\mathcal{B}, \mathcal{A}, \mathbb{P}_\psi, r, \gamma)$, where
  - $\mathcal{B}$ represents the **belief space**, a *continuous* space of pdfs/pmfs over $\mathcal{X}$, i.e. space of distributions
  - $\mathbb{P}_\psi(b_{k+1} \mid b_k, a_k)$ is a transformed transition model (next slide)
  - $r(b, a)$ is either (as earlier):
    - the transformed reward (cost): $r(b, a) = \int_X r(X, a) b[X] dX$
    - an information-theoretic reward (to be discussed later)

Belief MDP

- The transformed transition/motion model $\mathbb{P}_\psi(b_{k+1} \mid b_k, a_k)$ is

$$\mathbb{P}_\psi(b_{k+1} \mid b_k, a_k) = \underset{z_{k+1} \sim \mathbb{P}(.|b_k, a_k)}{\mathbb{E}} \left[\mathbb{P}(b_{k+1} \mid b_k, a_k, z_{k+1}) \right]$$

$$= \int_{z_{k+1}} \mathbb{P}(z_{k+1} \mid b_k, a_k) \mathbb{1} \left[b_{k+1} = \psi(b_k, a_k, z_{k+1}) \right] dz_{k+1}$$

Outline

Markov Chain and Markov Reward Process (MRP)

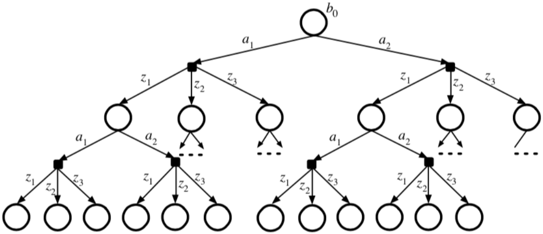
Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

Computational Complexity

- "Curse of dimensionality" & "curse of history": the complexity of planning grows **exponentially** with the size of the state space and the planning horizon (see e.g. [Papadimitriou and Tsitsiklis, 1987])



Belief tree. Figure from [Ye et al., 2017]

Outline

Markov Chain and Markov Reward Process (MRP)

Markov Decision Process (MDP)
- MDP Definition
- Value Function and (Optimal) Policy

Partially Observable Markov Decision Process (POMDP)
- POMDP Definition
- Belief MDP
- Computational Complexity
- Open Loop vs. Closed Loop Control

Problem Variations & Illustrative Examples

Open Loop vs. Closed Loop Control

- **Open loop**: actions/controls $a_{0:T-1}$ are determined at once at time 0 as a function of the initial state X_0 (fully observable case) or initial belief b_0 (partially observable case)
- **Closed loop** (policy): actions/controls are determined "just-in-time" as a function of the state X_t (fully observable case) or history $H_k \doteq \{a_{0:t-1}, z_{0:T}\}$ (partially observable case)

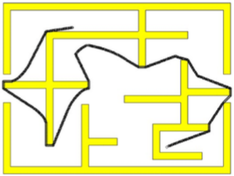
A special case of the closed control methodology is to disregard current state/history information, which yields an open loop setting.

Problem Variations

- Fully observable vs partially observable (MDP vs POMDP)
- Stationary vs. nonstationary (time (in-)dependent models)
- Finite vs. continuous state space $\mathcal{X}$ and action/control space $\mathcal{A}$
  - Represent probabilistic models with a tabular approach vs. function approximation (e.g. neural networks)
- Parametric vs. non-parametric probabilistic models
- Discrete vs. continuous (planning) time:
  - Finite-horizon vs. infinite-horizon discrete time
  - Continuous time: Hamilton-Jacobi-Bellman (HJB) Partial Differential Equation (PDE) [outside scope]
- MDP or POMDP models are unknown - Reinforcement Learning (RL) and Imitation Learning (IL)
  - Model-based approaches: explicitly learn/approximate models from experience and use optimal control algorithms
  - Model-free approaches: directly learn a control policy, without explicitly learning/approximating motion/observation models
- Offline vs. online methods

Example: Grid World Navigation

- Navigate to a goal w/o crashing into obstacles, given map
- Formalization:
  - State space: robot pose (2D or 3D)
  - Actions: allowable robot movement (can be discrete or continuous)
    examples: $\{\uparrow, \leftarrow, \rightarrow, \downarrow\}$, control angle w/ constant velocity
  - Reward: 1 until the goal is reached, $-\infty$ for colliding with an obstacle
  - Can be deterministic or stochastic, fully or partially observable



References I

Papadimitriou, C. and Tsitsiklis, J. (1987).
The complexity of markov decision processes.
Mathematics of operations research, 12(3):441–450.

Ye, N., Somani, A., Hsu, D., and Lee, W. S. (2017).
Despot: Online pomdp planning with regularization.
JAIR, 58:231–266.