ECE276B: Planning & Learning in Robotics
Lecture 2: Markov Decision Processes

Instructor:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Zhichao Li: zhl355@eng.ucsd.edu
Ehsan Zobeidi: ezobeidi@eng.ucsd.edu
Ibrahim Akbar: iakbar@eng.ucsd.edu

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Notation and Terminology

$x \in \mathcal{X}$	Markov process state
$u \in \mathcal{U}(x)$	control/action avilable in state x
$p_f(x' \mid x, u)$	motion model, i.e., control-dependent transition pdf
$\ell(x, u)$	stage cost/reward for choosing control u in state x
$\mathfrak{q}(x)$	(optional) terminal cost/reward at state x
$\pi(x)$	control policy: mapping from state x to control $u \in \mathcal{U}(x)$
$V^\pi(x)$	value function: **cumulative cost/reward** for starting at state x and acting according to π thereafter
$\pi^*(x)$, $V^*(x)$	optimal control policy and corresponding value function

Problem Formulation

- **Motion model**: specifies how a dynamical system evolves

$$x_{t+1} = f(x_t, u_t, w_t) \sim p_f(\cdot \mid x_t, u_t), \quad t = 0, \ldots, T-1$$

  - discrete time $t \in \{0, \ldots, T\}$
  - state $x_t \in \mathcal{X}$
  - control $u_t \in \mathcal{U}(x_t)$ and $\mathcal{U} := \bigcup_{x \in \mathcal{X}} \mathcal{U}(x)$
  - motion noise w_t (random vector) with known probability density function (pdf) and assumed conditionally independent of other disturbances w_τ for $\tau \neq t$ for given x_t and u_t
  - the motion model is specified by the nonlinear function f or equivalently by the pdf p_f of x_{t+1} conditioned on x_t and u_t

- **Observation model**: the state x_t might not be observable but perceived through measurements:

$$z_t = h(x_t, v_t) \sim p_h(\cdot \mid x_t), \quad t = 0, \ldots, T$$

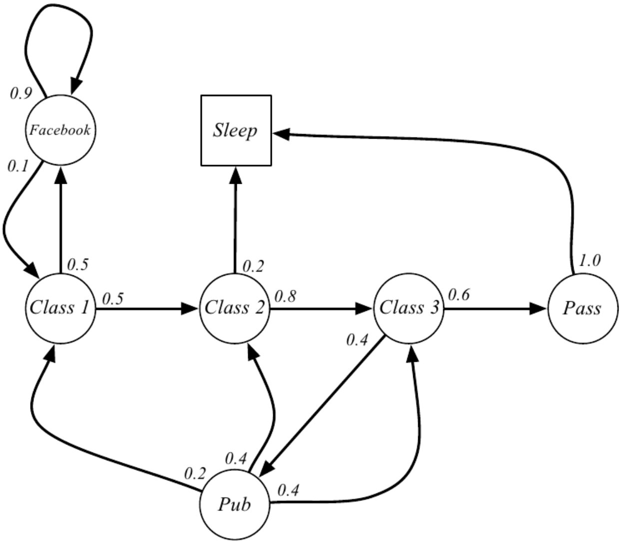
  - measurement noise v_t (random vector) with known pdf and conditionally independent of other disturbances v_τ for $\tau \neq t$ for given x_t and w_t for all t
  - the observation model is specified by the nonlinear function h or equivalently by the pdf p_h of z_t conditioned on x_t

Markov Chain

A **Markov Chain** is a stochastic process defined by a tuple $(\mathcal{X}, p_{0|0}, p_f)$:

- $\mathcal{X}$ is discrete/continuous set of states
- $p_{0|0}$ is a prior pmf/pdf defined on $\mathcal{X}$
- $p_f(\cdot \mid x_t)$ is a conditional pmf/pdf defined on $\mathcal{X}$ for given $x_t \in \mathcal{X}$ that specifies the stochastic process transitions. In the finite-dimensional case, the transition pmf is summarized by a matrix $P_{ij} := \mathbb{P}(x_{t+1} = j \mid x_t = i) = p_f(j \mid x_t = i)$

Example: Student Markov Chain

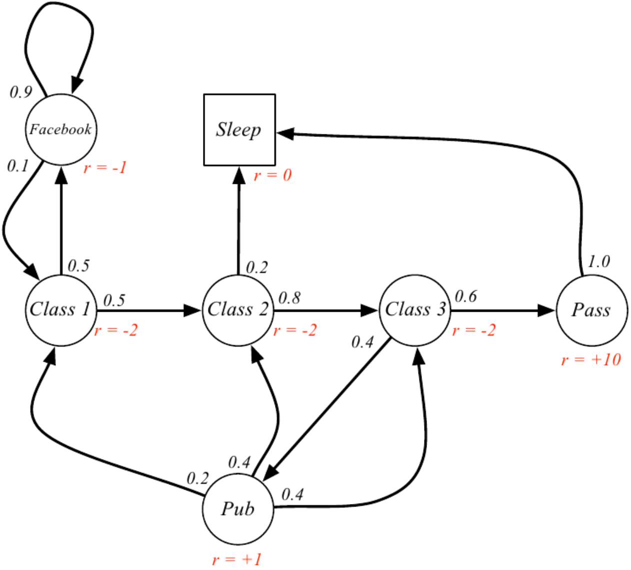


Markov Reward Process

A Markov Reward Process (MRP) is a Markov chain with state costs (rewards) defined by a tuple $(\mathcal{X}, p_{0|0}, p_f, \ell, \gamma)$

- $\mathcal{X}$ is a discrete/continuous set of states
- $p_{0|0}$ is a prior pmf/pdf defined on $\mathcal{X}$
- $p_f(\cdot \mid x_t)$ is a conditional pmf/pdf defined on $\mathcal{X}$ for given $x_t \in \mathcal{X}$ and summarized by a matrix $P_{ij} := p_f(j \mid x_t = i)$ in the finite-dimensional case.
- $\ell(x)$ is a function specifying the cost/reward of state $x \in \mathcal{X}$
- $\gamma \in [0, 1]$ is a discount factor

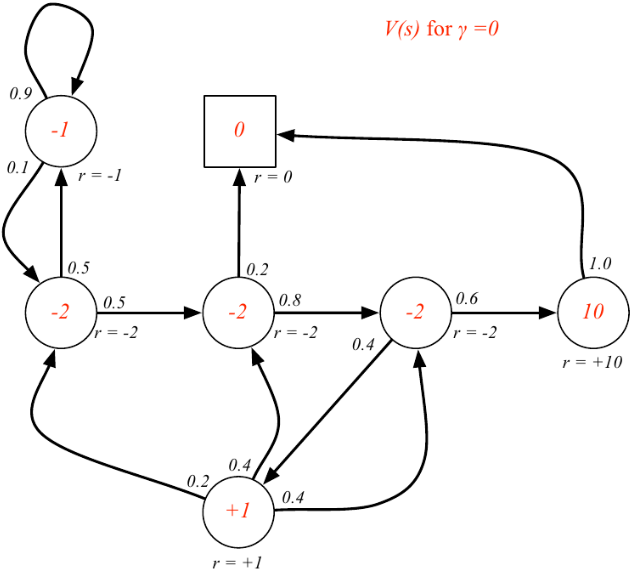
Example: Student Markov Reward Process



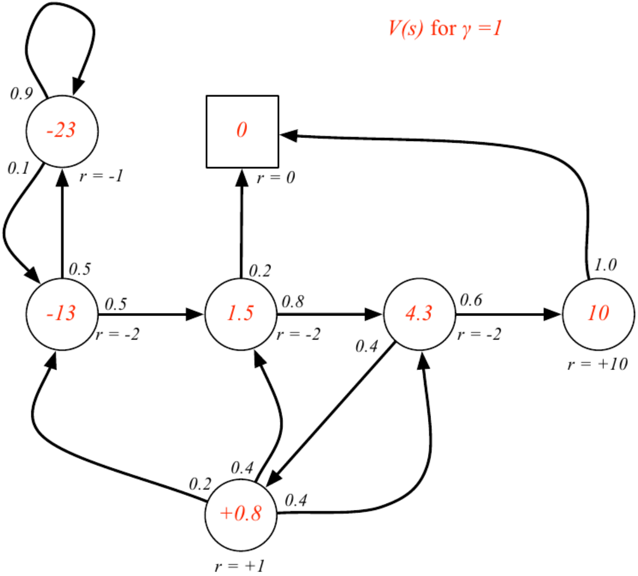
Cumulative Cost

- **Value function**: The cumulative cost/reward of an MRP $(\mathcal{X}, p_f, \ell, \gamma)$ starting from state $x \in \mathcal{X}$ at time 0:
  - **Finite-horizon**: $V_0(x) := \mathbb{E}\left[\underbrace{\mathfrak{q}(x_T)}_{\text{terminal cost}} + \sum_{t=0}^{T-1} \ell(x_t) \mid x_0 = x\right]$
  - **Discounted Infinite-horizon**: $V(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t) \mid x_0 = x\right]$
  - **Average-reward**: $V(x) := \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \ell(x_t) \mid x_0 = x\right]$
- 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 nonmyopic/far-sighted evaluation
  - Mathematically convenient since it avoids infinite costs as $T \to \infty$
  - The long-term future may be hard to model anyways
  - Animal/human behavior shows preference for immediate reward
  - It is possible to use an undiscounted MRP if all sequences terminate (**first-exit** formulation). The finite-horizon formulation is a special case of the first-exit formulation.

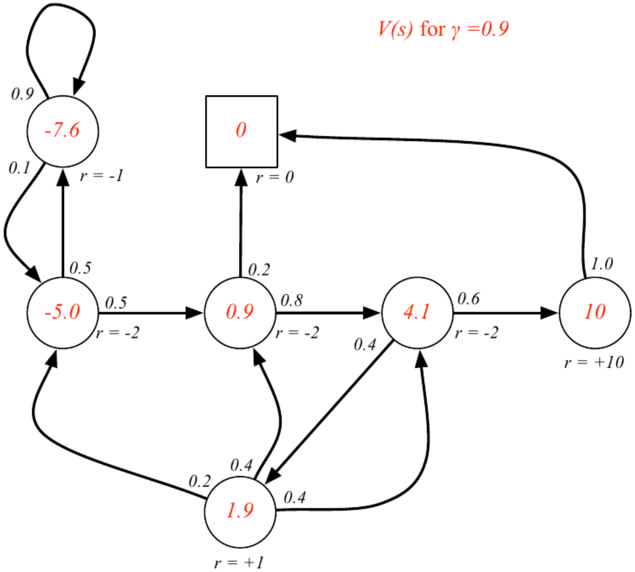
Example: Cumulative Reward of the Student MRP



Example: Cumulative Reward of the Student MRP



Example: Cumulative Reward of the Student MRP



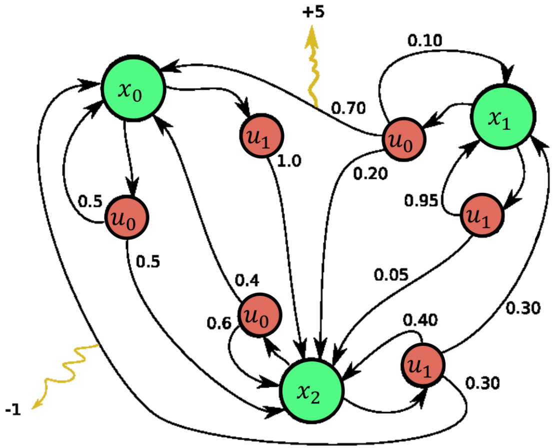
Markov Decision Process

A Markov Decision Process (MDP) is a Markov Reward Process with controlled transitions defined by a tuple $(\mathcal{X}, \mathcal{U}, p_{0|0}, p_f, \ell, \gamma)$

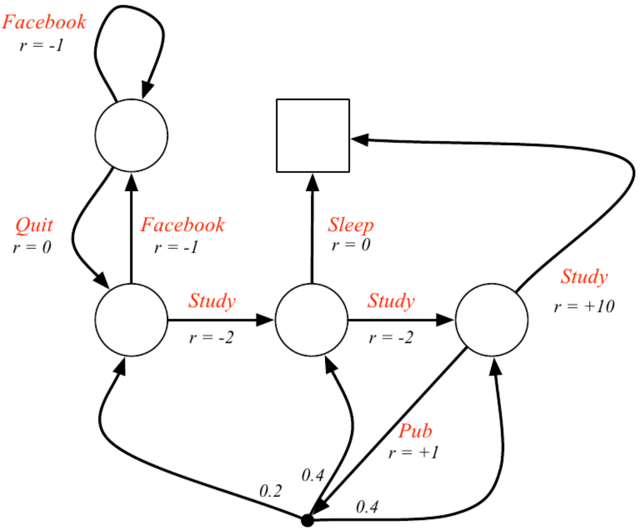
- $\mathcal{X}$ is a discrete/continuous set of states
- $\mathcal{U}$ is a discrete/continuous set of controls
- $p_{0|0}$ is a prior pmf/pdf defined on $\mathcal{X}$
- $p_f(\cdot \mid x_t, u_t)$ is a conditional pmf/pdf defined on $\mathcal{X}$ for given $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ and summarized by a matrix $P^u_{ij} := p_f(j \mid x_t = i, u_t = u)$ in the finite-dimensional case.
- $\ell(x, u)$ is a function specifying the cost/reward of applying control $u \in \mathcal{U}$ in state $x \in \mathcal{X}$
- $\gamma \in [0, 1]$ is a discount factor

Example: Markov Decision Process

- An action $u_t \in \mathcal{U}(x_t)$ applied in state $x_t \in \mathcal{X}$ determines the next state x_{t+1} and the obtained cost/reward $\ell(x_t, u_t)$



Example: Student Markov Decision Process



Control Policy and Cumulative Cost

- **Admissible control policy**: a sequence $\pi_{0:T-1}$ of functions π_t that map a state $x_t \in \mathcal{X}$ to a feasible control input $u_t \in \mathcal{U}(x_t)$

- **Value function**: the cumulative cost/reward of a policy π applied to an MDP $(\mathcal{X}, \mathcal{U}, p_f, \ell, \gamma)$ with initial state $x \in \mathcal{X}$ at time $t = 0$:

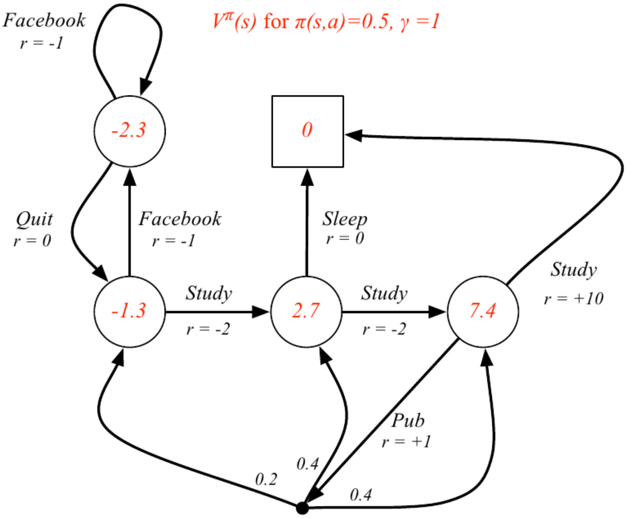
  - **Finite-horizon**: $V_0^\pi(x) := \mathbb{E}\left[\underbrace{\mathfrak{q}(x_T)}_{\text{terminal cost}} + \sum_{t=0}^{T-1} \ell(x_t, \pi_t(x_t)) \mid x_0 = x\right]$

  - **Discounted Infinite-horizon**: $V^\pi(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) \mid x_0 = x\right]$

  - **Average-reward**: $V^\pi(x) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \ell(x_t, \pi(x_t)) \mid x_0 = x\right]$

- **Note**: we will show that as $T \to \infty$, optimal policies become stationary, i.e., $\pi := \pi_0 \equiv \pi_1 \equiv \cdots$, and independent of x_0

Example: Value Function of Student MDP



Alternative Cost Formulations

- **Noise-dependent costs**: a more general model allows the stage costs ℓ' to depend on the motion noise w_t:

$$V_0^\pi(x) := \mathbb{E}_{w_{0:T}, x_{1:T}} \left[\mathfrak{q}(x_T) + \sum_{t=0}^{T-1} \ell'(x_t, \pi_t(x_t), w_t) \mid x_0 = x \right]$$

  This is equivalent to our formulation since the pdf $p_w(\cdot \mid x_t, u_t)$ of w_t is known and we can always compute:

$$\ell(x_t, u_t) := \mathbb{E}_{w_t \mid x_t, u_t} \left[\ell'(x_t, u_t, w_t) \right] = \int \ell(x_t, u_t, w_t) p_w(w_t \mid x_t, u_t) dw_t$$

- **Joint cost-state pdf**: a more general model allows random costs ℓ' by specifying the joint pdf $p(x', \ell' \mid x, u)$. This is equivalent to our formulation as follows:

$$p_f(x' \mid x, u) := \int p(x', \ell' \mid x, u) d\ell'$$

$$\ell(x, u) := \mathbb{E}\left[\ell' \mid x, u \right] = \int \int \ell' p(x', \ell' \mid, x, u) dx' d\ell'$$

Comparison of Markov Models

	observed	partially observed
uncontrolled	**Markov Chain/MRP**	**HMM**
controlled	**MDP**	**POMDP**

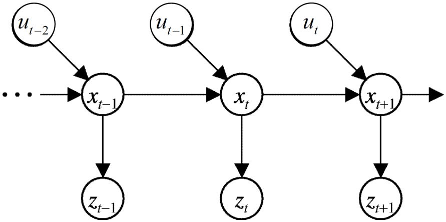
- Markov Chain + Partial Observability = HMM
- Markov Chain + Control = MDP
- Markov Chain + Partial Observability + Control = HMM + Control = MDP + Partial Observability = POMDP

Partially Observable Markov Decision Process

A Partially Observable Markov Decision Process (POMDP) is a Markov Decision Process with partially observable states defined by a tuple $(\mathcal{X}, \mathcal{U}, \mathcal{Z}, p_{0|0}, p_f, p_h, g, \gamma)$

- $\mathcal{X}$ is a discrete/continuous set of states
- $\mathcal{U}$ is a discrete/continuous set of controls
- $\mathcal{Z}$ is a discrete/continuous set of observations
- $p_{0|0}$ is a prior pmf/pdf defined on $\mathcal{X}$
- $p_f(\cdot \mid x_t, u_t)$ is a conditional pmf/pdf defined on $\mathcal{X}$ for given $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ and summarized by a matrix $P^u_{ij} := p_f(j \mid x_t = i, u_t = u)$ in the finite-dimensional case.
- $p_h(\cdot \mid x_t)$ is a conditional pmf/pdf defined on $\mathcal{Z}$ for given $x_t \in \mathcal{X}$ and summarized by a matrix $O_{ij} := p_h(j \mid x_t = i)$ in the finite-dim case.
- $\ell(x, u)$ is a function specifying the cost/reward of applying control $u \in \mathcal{U}$ in state $x \in \mathcal{X}$
- $\gamma \in [0, 1]$ is a discount factor

Bayes Filter



- **Motion model**:
  $x_{t+1} = f(x_t, u_t, w_t) \sim p_f(\cdot \mid x_t, u_t)$
- **Observation model**:
  $z_t = h(x_t, v_t) \sim p_h(\cdot \mid x_t)$
- **Filtering**: keeps track of

$$p_{t|t}(x_t) := p(x_t \mid z_{0:t}, u_{0:t-1})$$
$$p_{t+1|t}(x_{t+1}) := p(x_{t+1} \mid z_{0:t}, u_{0:t})$$

- **Bayes filter**:

$$p_{t+1|t+1}(x_{t+1}) = \underbrace{\overbrace{\frac{1}{p(z_{t+1}|z_{0:t}, u_{0:t})}}^{\frac{1}{\eta_{t+1}}} p_h(z_{t+1} \mid x_{t+1}) \overbrace{\int p_f(x_{t+1} \mid x_t, u_t) p_{t|t}(x_t) dx_t}^{\textbf{Predict: } p_{t+1|t}(x_{t+1})}}_{\textbf{Update}}$$

- **Joint distribution**:

$$p(x_{0:T}, z_{0:T}, u_{0:T-1}) = \underbrace{p_{0|0}(x_0)}_{\text{prior}} \prod_{t=0}^{T} \underbrace{p_h(z_t \mid x_t)}_{\text{observation model}} \prod_{t=0}^{T} \underbrace{p_f(x_t \mid x_{t-1}, u_{t-1})}_{\text{motion model}}$$

Information Space and Sufficient Statistics

- The information available to the robot at time t to choose the control input u_t is $i_t := (z_{0:t}, u_{0:t-1}) \in \mathcal{I}$
- The **information space** $\mathcal{I}$ is the space of sequences of observations and controls
- A **statistic** $y_t = s(i_t)$ is a function of the information available at time t to estimate x_t
- The statistic $y_t = s(i_t)$ is **sufficient** for x_t if the conditional distribution of x_t given the statistic y_t does not depend on the information i_t
- Under the Markov and measurement and motion noise independence (over time, from the state, and from each other) assumptions, the distribution of the state x_t conditioned on the information state i_t is a sufficient statistic for x_t. In other words, $p_{t|t}(x_t) := p(x_t \mid i_t)$ is a compact representation of i_t.

Equivalence of POMDPs and MDPs

- The **Bayes filter** ψ tracks precisely the needed sufficient statistic:

$$p(x_t \mid i_t) = \boxed{p_{t|t}(x_t) = \psi(p_{t-1|t-1}, u_{t-1}, z_t)}$$
$$= \frac{1}{\eta_t} p_h(z_t \mid x_t) \int p_f(x_t \mid x_{t-1}, u_{t-1}) p_{t-1|t-1}(x_{t-1}) dx_{t-1}$$

- Because $p_{t|t}$ is a sufficient statistic for x_t, we can convert a POMDP $(\mathcal{X}, \mathcal{U}, \mathcal{Z}, p_f, p_h, \ell, \gamma)$ into an equivalent MDP $(\mathcal{B}, \mathcal{U}, p_\psi, \rho, \gamma)$ where:
  - The state space $\mathcal{B} := \mathcal{P}(\mathcal{X})$ is the continuous space of pdfs/pmfs over $\mathcal{X}$, e.g., if $|\mathcal{X}| = N$, then $\mathcal{B} = \{b \in [0,1]^N \mid \mathbf{1}^T b = 1\}$
  - The transformed motion model is the Bayes filter $b_{t+1} = \psi(b_t, u_t, z_t)$, where z_t plays the role of noise or in probabilistic terms:

$$p_\psi(b_{t+1} \mid b_t, u_t) := \int \mathbb{1}\{b_{t+1} = \psi(b_t, u_t, z)\} \eta(z \mid b_t, u_t) dz$$
$$\eta(z \mid b_t, u_t) := \int \int p_h(z \mid x_{t+1}) p_f(x_{t+1} \mid x_t, u_t) b_t(x_t) dx_t dx_{t+1}$$

  - The transformed stage cost/reward function $\rho(b, u) = \int \ell(x, u) b(x) dx$ is the expected stage cost/reward

The Problem of Acting Optimally in a POMDP

- An infinite-dimensional dynamic optimization problem defined for a POMDP $(\mathcal{X}, \mathcal{U}, \mathcal{Z}, p_f, p_h, \ell, \gamma)$ as follows:

$$\begin{aligned}
\min_{\pi_{0:T-1}} \quad & \mathbb{E}\left[\gamma^T \mathfrak{q}(x_T) + \sum_{t=0}^{T-1} \gamma^t \ell_t(x_t, u_t)\right] \\
\text{s.t.} \quad & x_{t+1} \sim p_f(\cdot \mid x_t, u_t), \quad t = 0, \dots, T-1 \\
& z_{t+1} \sim p_h(\cdot \mid x_t), \quad t = 0, \dots, T-1 \\
& u_t \sim \pi_t(\cdot \mid i_t), \quad t = 0, \dots, T-1 \\
& x_0 \sim b_0(\cdot) \equiv \text{prior pdf over the hidden state } x_0
\end{aligned}$$

- Equivalently, using the information-space MDP $(\mathcal{B}, \mathcal{U}, p_\psi, \rho, \gamma)$ with sufficient statistic b_t:

$$\begin{aligned}
\min_{\pi_{0:T-1}} \quad & V_0^\pi(b_0) = \mathbb{E}\left[\gamma^T \rho_T(b_T) + \sum_{t=0}^{T-1} \gamma^t \rho_t(b_t, u_t)\right] \\
\text{s.t.} \quad & b_{t+1} = \psi(b_t, u_t, z_{t+1}), \quad t = 0, \dots, T-1 \\
& z_{t+1} \sim \eta(\cdot \mid b_t, u_t), \quad t = 0, \dots, T-1 \\
& u_t \sim \pi_t(\cdot \mid b_t), \quad t = 0, \dots, T-1
\end{aligned}$$

Final Problem Formulation

- Due to the equivalence between POMDPs and (information-space) MDPs, we will focus exclusively on MDPs

- First, we will consider the **finite-horizon** formulation

$$
\begin{aligned}
\min_{\pi} \; & V_0^{\pi}(x_0) := \mathbb{E}_{x_{1:T}}\left[\mathfrak{q}(x_T) + \sum_{t=0}^{T-1} \ell_t(x_t, \pi_t(x_t)) \,\middle|\, x_0\right] \\
\text{s.t.} \; & x_{t+1} \sim p_f(\cdot \mid x_t, \pi_t(x_t)), \qquad t = 0, \ldots, T-1 \\
& x_t \in \mathcal{X}, \;\; \pi_t(x_t) \in \mathcal{U}(x_t)
\end{aligned}
$$

- Then, we will consider the discounted **infinite-horizon** formulation:

$$
\begin{aligned}
\min_{\pi} \; & V^{\pi}(x_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) \,\middle|\, x_0\right] \\
\text{s.t.} \; & x_{t+1} \sim p_f(\cdot \mid x_t, \pi_t(x_t)), \\
& x_t \in \mathcal{X}, \;\; \pi_t(x_t) \in \mathcal{U}(x_t)
\end{aligned}
$$

Open Loop vs Closed Loop Control

- There are two different control methodologies:
  - **Open loop**: control inputs $u_{0:T-1}$ are determined at once at time 0 as a function of x_0 (fully observable case) or $p_{0|0}$ (partially observable case)
  - **Closed loop**: control inputs are determined "just-in-time" as a function of the state x_t (fully observable case) or measurement history $z_{0:t}$, $u_{0:t-1}$ (partially observable case)
- A special case of closed loop control is to simply disregard state/measurement information (open loop control). Thus, open loop control can never give better performance than closed loop control.
- In the absence of disturbances (or in the special linear quadratic Gaussian case), the two give theoretically the same performance.
- When good models are available, open-loop control is a viable strategy for short time horizons

Open Loop vs Closed Loop Control

- Open loop control is typically much less demanding than closed loop control
- Consider a discrete-space example with $N_x = 10$ states, $N_u = 10$ control inputs, planning horizon $T = 4$, and given x_0:
  - There are $N_u^T = 10^4$ different open-loop strategies
  - There are $N_u(N_u^{N_x})^{T-1} = N_u^{N_x(T-1)+1} = 10^{31}$ different closed-loop strategies (10 orders of magnitude larger than the number of stars in the observable universe!)

Example: Chess Strategy Optimization

- **Objective**: come up with a strategy that maximizes the chances of winning a 2 game chess match.
- Possible outcomes:
  - Win/Lose: 1 point for the winner, 0 for the loser
  - Draw: 0.5 points for each player
  - If the score is equal after 2 games, the players continue playing until one wins (sudden death)
- Playing styles:
  - **Timid**: draw with probability p_d and lose with probability $(1 - p_d)$
  - **Bold**: win with probability p_w and lose with probability $(1 - p_w)$
  - **Assumption**: $p_d > p_w$

Finite-state Model of the Chess Match

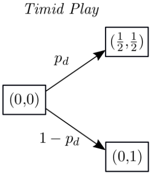
- The **state** x_t is a 2-D vector with our and the opponent's score after the t-th game
- The **control** u_t is the play style: timid or bold
- The **noise** w_t is the score of the next game
- Since timid play does not make sense during the sudden death stage, the planning horizon is $T = 2$
- We can construct a **time-dependent motion model** P^u_{ijt} for $t \in \{0, 1\}$ (shown on the next slide)
- **Cost**: minimize loss probability: $-P_{win} = \mathbb{E}_{x_{1:2}}\left[\ell_2(x_2) + \sum_{t=0}^{1} \ell_t(x_t, u_t)\right]$

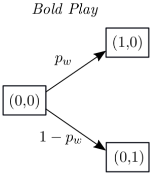
where $\ell_t(x_t, u_t) = 0$ for $t \in \{0, 1\}$ and

$$\ell_2(x_2) = \begin{cases} -1 & \text{if } x_2 = \left(\frac{3}{2}, \frac{1}{2}\right) \text{ or } (2, 0) \\ -p_w & \text{if } x_2 = (1, 1) \\ 0 & \text{if } x_2 = \left(\frac{1}{2}, \frac{3}{2}\right) \text{ or } (0, 2) \end{cases}$$

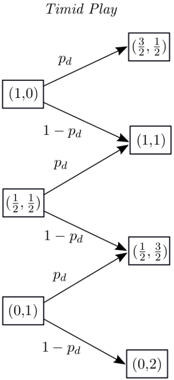
Chess Transition Probabilities

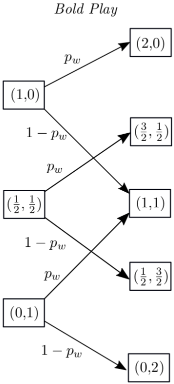
Game 1:





Game 2:





Open Loop Chess Strategy

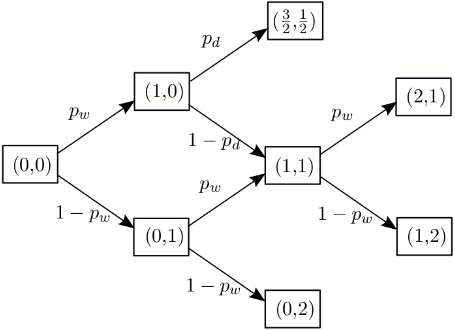
- There are 4 admissible open-loop policies:
  1. timid-timid: $P_{win} = p_d^2 p_w$
  2. bold-bold: $P_{win} = p_w^2 + p_w(1 - p_w)p_w + (1 - p_w)p_w p_w = p_w^2(3 - 2p_w)$
  3. bold-timid: $P_{win} = p_w p_d + p_w(1 - p_d)p_w$
  4. timid-bold: $P_{win} = p_d p_w + (1 - p_d)p_w^2$
- Since $p_d^2 p_w \leq p_d p_w \leq p_d p_w + (1 - p_d)p_w^2$, timid-timid is not optimal
- The best achievable winning probability is:

$$
\begin{aligned}
P_{win}^* &= \max\{\overbrace{p_w^2(3 - 2p_w)}^{\text{bold-bold}}, \overbrace{p_d p_w + (1 - p_d)p_w^2}^{\text{3. or 4.}}\} \\
&= p_w^2 + p_w(1 - p_w)\max\{2p_w, p_d\}
\end{aligned}
$$

- In the open-loop case, if $p_w \leq 0.5$, then $P_{win}^* \leq 0.5$
  - For $p_w = 0.45$ and $p_d = 0.9$, $P_{win}^* = 0.43$
  - For $p_w = 0.5$ and $p_d = 1.0$, $P_{win}^* = 0.5$
- If $p_d > 2p_w$, bold-timid and timid-bold are optimal open-loop policies; otherwise bold-bold is optimal

Closed Loop Chess Strategy

- There are 16 admissible policies
- Consider one option: play timid if and only if ahead (it will turn out that this is optimal)



- The probability of winning is:
  $P_{win} = p_d p_w + p_w((1-p_d)p_w + p_w(1-p_w)) = p_w^2(2-p_w) + p_w(1-p_w)p_d$
- Note that in the closed-loop case we can achieve P_{win} larger than 0.5 even when p_w is less than 0.5:
  - For $p_w = 0.45$ and $p_d = 0.9$, $P_{win} = 0.5$
  - For $p_w = 0.5$ and $p_d = 1.0$, $P_{win} = 0.625$