ECE276B: Planning & Learning in Robotics
Lecture 3: The Dynamic Programming Algorithm

Instructor:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Zhichao Li: zhl355@eng.ucsd.edu

Jinzhao Li: jil016@eng.ucsd.edu

UC San Diego

JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Dynamic Programming

- **Objective**: construct an optimal policy π^* (independent of $\mathbf{x}_0$):

$$\pi^* = \arg\min_{\pi \in \Pi} V_0^\pi(\mathbf{x}_0), \quad \forall \mathbf{x}_0 \in \mathcal{X}$$

- **Value function** $V_t^\pi(\mathbf{x})$: estimates how good (in terms of expected long-term cost/return) it is to be in state $\mathbf{x}$ at time t and follow policy π
- **Dynamic programming (DP)**: algorithm that can compute an optimal closed-loop policy given a known MDP model of the environment
  - Idea: use value functions to structure the search for good policies
  - Generality: can handle non-convex and non-linear problems
  - Complexity: **polynomial in the number of states and actions**
  - Efficiency: much more efficient than the brute-force approach of evaluating all possible strategies.

Principle of Optimality

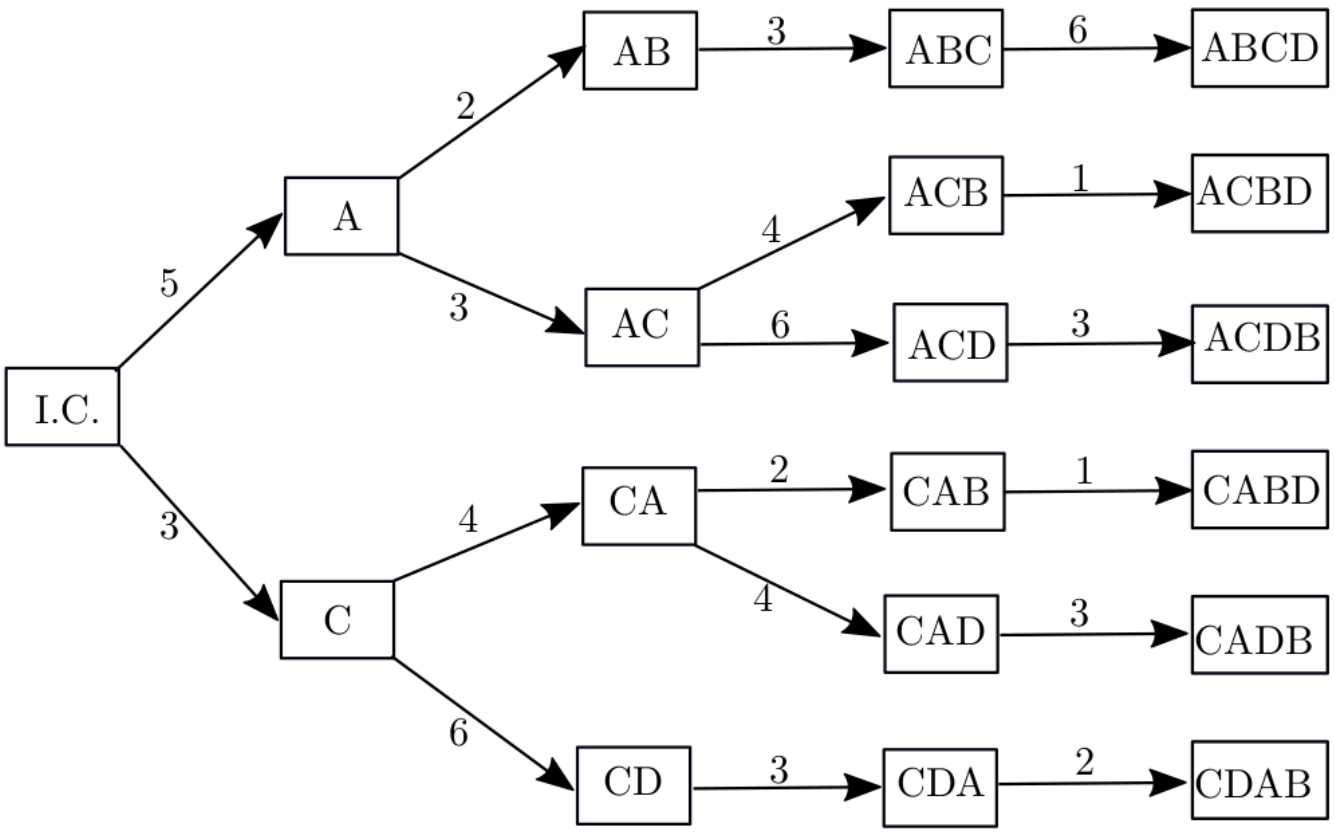
- Let $\pi^*_{0:T-1}$ be an optimal closed-loop policy
- Consider a **subproblem**, where the state is $\mathbf{x}_t$ at time t and we want to minimize:

$$V_t^{\pi}(\mathbf{x}_t) = \mathbb{E}_{\mathbf{x}_{t+1:T}}\left[\gamma^{T-t}\mathfrak{q}(\mathbf{x}_T) + \sum_{\tau=t}^{T-1}\gamma^{\tau-t}\ell_\tau(\mathbf{x}_\tau, \pi_\tau(\mathbf{x}_\tau)) \,\middle|\, \mathbf{x}_t\right]$$

- **Principle of optimality**: the truncated policy $\pi^*_{t:T-1}$ is optimal for the subproblem starting at time t
- **Intuition**: Suppose $\pi^*_{t:T-1}$ were not optimal for the subproblem. Then, there would exist a policy yielding a lower cost on at least some portion of the state space.

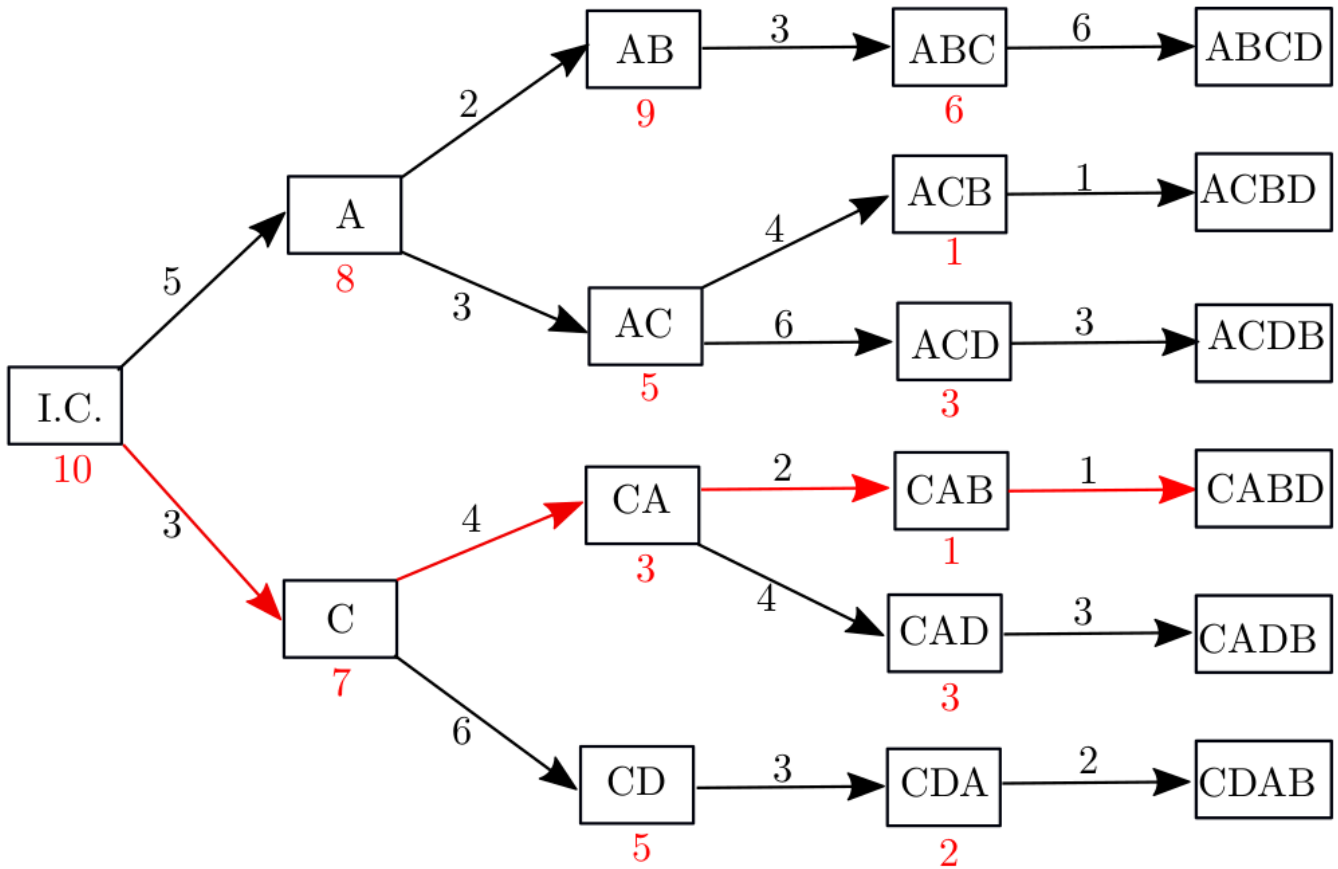
Example: Deterministic Scheduling Problem

- Consider a deterministic scheduling problem where 4 operations A, B, C, D are used to produce a product
- Rules: Operation A must occur before B, and C before D
- Cost: there is a transition cost between each two operations:



Example: Deterministic Scheduling Problem

- The DP algorithm is applied backwards in time. First, construct an optimal solution at the last stage and then work backwards.
- The optimal cost-to-go at each state of the scheduling problem is denoted with red text below the state:



The Dynamic Programming Algorithm

Algorithm 1 Dynamic Programming

Input: MDP $(\mathcal{X}, \mathcal{U}, p_0, p_f, T, \ell, \mathfrak{q}, \gamma)$

$V_T(\mathbf{x}) = \mathfrak{q}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}$

for $t = (T-1) \dots 0$ **do**

$Q_t(\mathbf{x}, \mathbf{u}) = \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\mathbf{u})} \left[V_{t+1}(\mathbf{x}') \right], \quad \forall \mathbf{x} \in \mathcal{X}, \mathbf{u} \in \mathcal{U}(\mathbf{x})$

$V_t(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} Q_t(\mathbf{x}, \mathbf{u}), \quad \forall \mathbf{x} \in \mathcal{X}$

$\pi_t(\mathbf{x}) = \arg\min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} Q_t(\mathbf{x}, \mathbf{u}), \quad \forall \mathbf{x} \in \mathcal{X}$

return policy $\pi_{0:T-1}$ and value function V_0

Theorem: Optimality of the DP Algorithm

The policy $\pi_{0:T-1}$ and value function V_0 returned by the DP algorithm are optimal for the finite-horizon optimal control problem.

The Dynamic Programming Algorithm

- At each recursion step, the optimization needs to be performed over all possible values of $\mathbf{x} \in \mathcal{X}$ because we do not know a priori which states will be visited
- This point-wise optimization for each $\mathbf{x} \in \mathcal{X}$ is what gives us a policy π_t, i.e., a function specifying the optimal control for **every** state $\mathbf{x} \in \mathcal{X}$
- 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 $N_u^{N_x(T-1)+1} = 10^{31}$ different closed-loop strategies
  - For each stage t and each state x_t, the DP algorithm goes through the N_u control inputs to determine the optimal input. In total, there are $N_u N_x(T-1) + N_u = 310$ such operations.

Proof of Dynamic Programming Optimality

- **Claim**: The policy $\pi_{0:T-1}$ and value function V_0 returned by the DP algorithm are optimal
- Let $J_t^*(\mathbf{x})$ be the optimal cost for the $(T-t)$-stage problem that starts at time t in state $\mathbf{x}$.
- Proceed by induction
- **Base-case**: $J_T^*(\mathbf{x}) = \mathfrak{q}(\mathbf{x}) = V_T(\mathbf{x})$
- **Hypothesis**: Assume that for $t+1$, $J_{t+1}^*(\mathbf{x}) = V_{t+1}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$
- **Induction**: Show that $J_t^*(\mathbf{x}_t) = V_t(\mathbf{x}_t)$ for all $\mathbf{x}_t \in \mathcal{X}$

Proof of Dynamic Programming Optimality

$$
\begin{aligned}
J_t^*(\mathbf{x}_t) &= \min_{\pi_{t:T-1}} \mathbb{E}_{\mathbf{x}_{t+1:T}|\mathbf{x}_t} \left[\ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) + \gamma^{T-t} \mathfrak{q}(\mathbf{x}_T) + \sum_{\tau=t+1}^{T-1} \gamma^{\tau-t} \ell_\tau(\mathbf{x}_\tau, \pi_\tau(\mathbf{x}_\tau)) \right] \\
&\overset{(1)}{=\!=} \min_{\pi_{t:T-1}} \ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) + \mathbb{E}_{\mathbf{x}_{t+1:T}|\mathbf{x}_t} \left[\gamma^{T-t} \mathfrak{q}(\mathbf{x}_T) + \sum_{\tau=t+1}^{T-1} \gamma^{\tau-t} \ell_\tau(\mathbf{x}_\tau, \pi_\tau(\mathbf{x}_\tau)) \right] \\
&\overset{(2)}{=\!=} \min_{\pi_{t:T-1}} \ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) + \gamma \mathbb{E}_{\mathbf{x}_{t+1}|\mathbf{x}_t} \left[\mathbb{E}_{\mathbf{x}_{t+2:T}|\mathbf{x}_{t+1}} \left[\gamma^{T-t-1} \mathfrak{q}(\mathbf{x}_T) + \sum_{\tau=t+1}^{T-1} \gamma^{\tau-t-1} \ell_\tau(\mathbf{x}_\tau, \pi_\tau(\mathbf{x}_\tau)) \right] \right] \\
&\overset{(3)}{=\!=} \min_{\pi_t} \left\{ \ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) + \gamma \mathbb{E}_{\mathbf{x}_{t+1}|\mathbf{x}_t} \left[\min_{\pi_{t+1:T-1}} \mathbb{E}_{\mathbf{x}_{t+2:T}|\mathbf{x}_{t+1}} \left[\gamma^{T-t-1} \mathfrak{q}(\mathbf{x}_T) + \sum_{\tau=t+1}^{T-1} \gamma^{\tau-t-1} \ell_\tau(\mathbf{x}_\tau, \pi_\tau(\mathbf{x}_\tau)) \right] \right] \right\} \\
&\overset{(4)}{=\!=} \min_{\pi_t} \left\{ \ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) + \gamma \mathbb{E}_{\mathbf{x}_{t+1} \sim p_f(\cdot|\mathbf{x}_t, \pi_t(\mathbf{x}_t))} \left[J_{t+1}^*(\mathbf{x}_{t+1}) \right] \right\} \\
&\overset{(5)}{=\!=} \min_{\mathbf{u}_t \in \mathcal{U}(\mathbf{x}_t)} \left\{ \ell_t(\mathbf{x}_t, \mathbf{u}_t) + \gamma \mathbb{E}_{\mathbf{x}_{t+1} \sim p_f(\cdot|\mathbf{x}_t, \mathbf{u}_t)} \left[V_{t+1}(\mathbf{x}_{t+1}) \right] \right\} \\
&= V_t(\mathbf{x}_t), \quad \forall \mathbf{x}_t \in \mathcal{X}
\end{aligned}
$$

Proof of Dynamic Programming Optimality

(1) Since $\ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t))$ is not a function of $\mathbf{x}_{t+1:T}$

(2) Using conditional probability
$p(\mathbf{x}_{t+1:T}|\mathbf{x}_t) = p(\mathbf{x}_{t+2:T}|\mathbf{x}_{t+1}, \mathbf{x}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ and the Markov assumption

(3) The minimization can be split since the term $\ell_t(\mathbf{x}_t, \pi_t(\mathbf{x}_t))$ does not depend on $\pi_{t+1:T-1}$. The expectation $\mathbb{E}_{\mathbf{x}_{t+1}|\mathbf{x}_t}$ and $\min_{\pi_{t+1:T}}$ can be exchanged since the functions $\pi_{t+1:T-1}$ make the cost small for all initial conditions., i.e., independently of $\mathbf{x}_{t+1}$.

- (1)-(3) is the *principle of optimality*

(4) By definition of $J^*_{t+1}(\cdot)$ and the motion model $\mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \mathbf{u}_t)$

(5) By the induction hypothesis

Example: Chess Strategy Optimization

- State: $x_t \in \mathcal{X} := \{-2, -1, 0, 1, 2\}$ – the difference between our and the opponent's score at the end of game t
- Input: $u_t \in \mathcal{U} := \{timid, bold\}$
- Dynamics: with $p_d > p_w$:

$$\begin{aligned} p_f(x_{t+1} = x_t \mid u_t = timid, x_t) &= p_d \\ p_f(x_{t+1} = x_t - 1 \mid u_t = timid, x_t) &= 1 - p_d \\ p_f(x_{t+1} = x_t + 1 \mid u_t = bold, x_t) &= p_w \\ p_f(x_{t+1} = x_t - 1 \mid u_t = bold, x_t) &= 1 - p_w \end{aligned}$$

- Cost: $V_t^*(x_t) = \mathbb{E}\left[\mathfrak{q}(x_2) + \sum_{t=\tau}^{1} \underbrace{\ell_\tau(x_\tau, u_\tau)}_{=0}\right]$ with

$$\mathfrak{q}(x) = \begin{cases} -1 & \text{if } x > 0 \\ -p_w & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Dynamic Programming Applied to the Chess Problem

- Initialize: $V_2(x_2) = \begin{cases} -1 & \text{if } x_2 > 0 \\ -p_w & \text{if } x_2 = 0 \\ 0 & \text{if } x_2 < 0 \end{cases}$
- Recursion: for all $x_t \in \mathcal{X}$ and $t = 1, 0$:

$$V_t(x_t) = \min_{u_t \in \mathcal{U}} \left\{ \ell_t(x_t, u_t) + \mathbb{E}_{x_{t+1}|x_t,u_t} \left[V_{t+1}(x_{t+1}) \right] \right\}$$

$$= \min \left\{ \underbrace{p_d V_{t+1}(x_t) + (1 - p_d) V_{t+1}(x_t - 1)}_{\text{timid}}, \underbrace{p_w V_{t+1}(x_t + 1) + (1 - p_w) V_{t+1}(x_t - 1)}_{\text{bold}} \right\}$$

DP Applied to the Chess Problem ($t = 1$)

- $x_1 = 1$:

$$V_1(1) = -\max\{p_d + (1-p_d)p_w, p_w + (1-p_w)p_w\} \underset{p_d > p_w}{\overset{\text{since}}{=\!=}}$$
$$= -p_d - (1-p_d)p_w$$
$$\pi_1^*(1) = \textit{timid}$$

- $x_1 = 0$:

$$V_1(0) = -\max\{p_d p_w + (1-p_d)0, p_w + (1-p_w)0\} = -p_w$$
$$\pi_1^*(0) = \textit{bold}$$

- $x_1 = -1$:

$$V_1(-1) = -\max\{p_d 0 + (1-p_d)0, p_w p_w + (1-p_w)0\} = -p_w^2$$
$$\pi_1^*(-1) = \textit{bold}$$

DP Applied to the Chess Problem ($t = 0$)

- $x_0 = 0$:

$$
\begin{aligned}
V_0(0) &= -\max\{p_d V_1^*(0) + (1-p_d)V_1^*(-1), p_w V_1^*(1) + (1-p_w)V_1^*(-1)\} \\
&= -\max\left\{p_d p_w + (1-p_d)p_w^2, p_w(p_d + (1-p_d)p_w) + (1-p_w)p_w^2\right\} \\
&= -p_d p_w - (1-p_d)p_w^2 - (1-p_w)p_w^2
\end{aligned}
$$

$\pi_0^*(0) = \textit{bold}$

- Thus, as we saw before, the optimal strategy is to play timid iff ahead in the score