SC42110
Dynamic Programming and Stochastic Control
Dynamic Programming Algorithm

Amin Sharifi Kolarijani

Delft Center for Systems and Control
Delft University of Technology

The Netherlands

2025

Markov decision process (MDP)

Markov chain (a stochastic process):

$$X_{t+1} = f(X_t, W_{t+1})$$

- $X_t \in \mathbb{X}$: state variable
- $W_{t+1} \in \mathbb{W}$: disturbance with distribution $p_{W_{t+1}|X_t} \in \Delta(\mathbb{W})$

Markov decision process (a controlled stochastic process):

$$X_{t+1} = f(X_t, U_t, W_{t+1})$$

- $X_t \in \mathbb{X}$: state variable
- $U_t \in \mathbb{U}$: control variable
- $W_{t+1} \in \mathbb{W}$: disturbance with distribution $p_{W_{t+1}|X_t,U_t} \in \Delta(\mathbb{W})$

Transition probability kernel of an MDP

Provides the distribution of X_{t+1} given X_t and U_t.

E.g., for a finite (state-action) MDP with $\mathbb{X} = [n]$ and $\mathbb{U} = [m]$ with $n, m \in \mathbf{N}$, we can define the family $\{P_k\}_{k \in \mathbb{U}}$ of transition probability matrices, where

$$P_k(i,j) = \mathbb{P}(X_{t+1} = j \mid X_t = i,\ U_t = k), \quad \forall (i,j,k) \in \mathbb{X} \times \mathbb{X} \times \mathbb{U}.$$

Control objective

Find "the best" control actions for steering the process in a "desired" fashion.

For a finite planning horizon $T \in \mathbf{N}$:

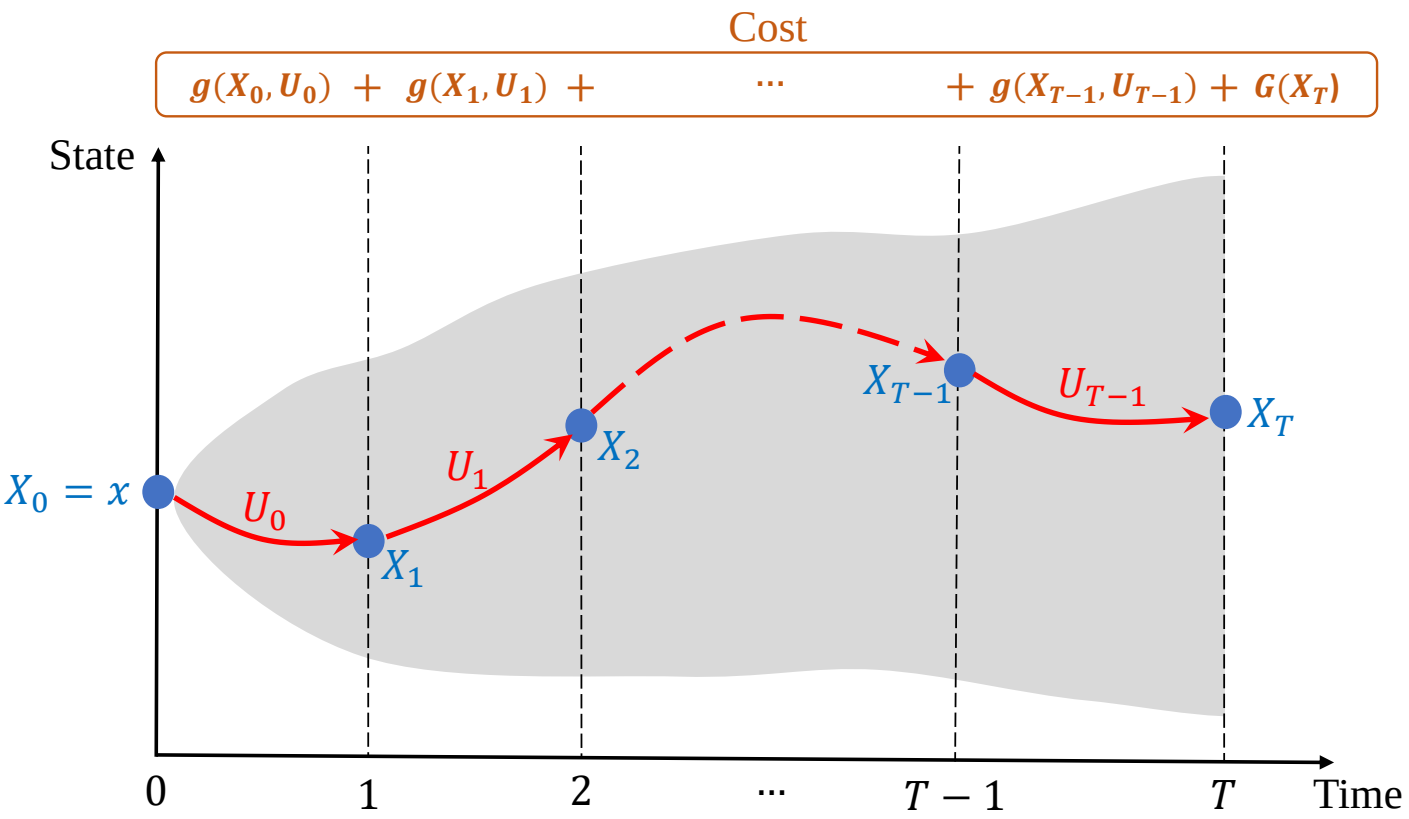
- Running (stage) cost $g(x, u)$: cost of action u in state x at time $t \in \{0, 1, \ldots, T-1\}$;
- Terminal cost $G(x)$: cost of being in state x at the end of planning horizon $t = T$.

So, the cost of the trajectory $(X_0, U_0, X_1, U_1, \ldots, X_{T-1}, U_{T-1}, X_T)$ is

$$\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T).$$

Remark: In some problems, we may wish to maximize rewards, as opposed to minimizing costs.

Control objective



$$\min_{U_0, U_1, \ldots, U_{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T) \;\middle|\; X_0 = x \right)$$

Control policies

Sequential decision-making under uncertainty $\Longrightarrow$ closed-loop control laws

- For generic controlled stochastic processes, optimal policies may depend on the entire history, i.e.,

$$U_t = \mu_t(X_t, U_{t-1}, X_{t-1}, \ldots, U_0, X_0), \quad \forall t \in \{0, 1, \ldots, T-1\}.$$

- **Theorem**: MDPs have Markov (i.e., memory-less) optimal policies, i.e.

$$U_t = \mu_t(X_t), \quad \forall t \in \{0, 1, \ldots, T-1\}.$$

Remark: An MPD under a Markov policy becomes a Markov chain!

Optimal control of an MDP

Definition (Stochastic optimal control): Given an MDP with

- planning horizon $T \in \mathbf{N}$,
- state space $\mathbb{X}$ and action space $\mathbb{U}$,
- dynamics

$$X_{t+1} = f(X_t, U_t, W_{t+1}), \quad t = 0, 1, \ldots, T-1,$$

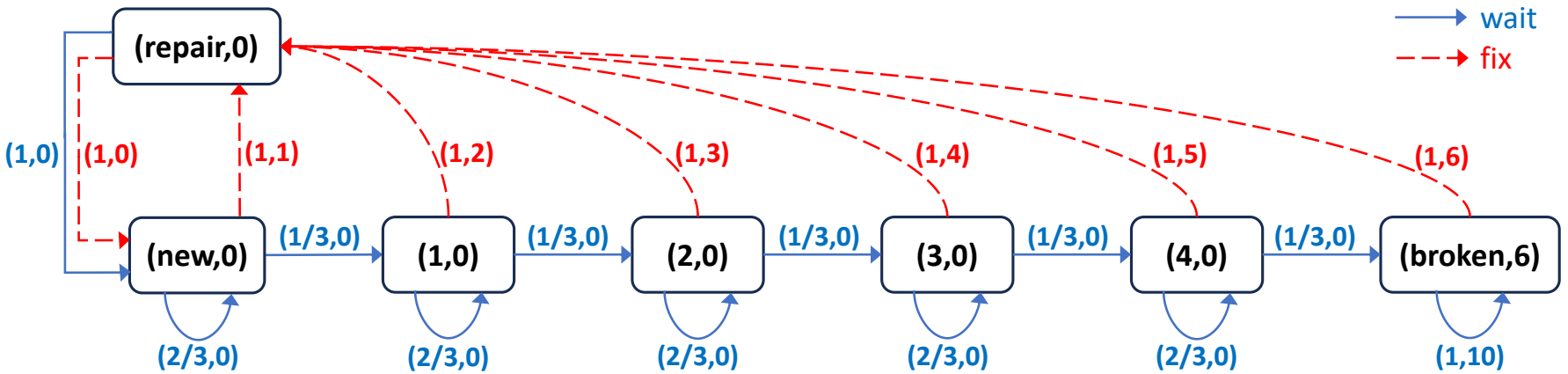
  where the disturbance $W_t \in \mathbb{W}$ has the distribution $p_{W_{t+1}|X_t,U_t} \in \Delta(\mathbb{W})$,
- running cost $g : \mathbb{X} \times \mathbb{U} \to \mathbf{R}$ and terminal cost $G : \mathbb{X} \to \mathbf{R}$,

the goal is to find an optimal policy $\mu^\star = (\mu_t^\star)_{t=0}^{T-1}$ by solving the optimization problem

$$J_0^\star(x) \coloneqq \min_{(\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} g\big(X_t, \mu_t(X_t)\big) + G(X_T) \;\middle|\; X_0 = x \right), \quad \forall x \in \mathbb{X}.$$

Example: Machine replacement

A machine has states 'repair', 'new', 1,2,3,4 and 'broken' (ordered by deteriorating operating conditions). At the beginning of each period, we can either 'fix' the machine at some cost or 'wait'. In the broken state, waiting incurs a large penalty.



- Rectangles (x, G): state x and corresponding terminal cost G
- Arcs (p, g): transition prob. p and corresponding running cost g

When should we repair the machine to minimize the expected costs if the machine is initially at state '1' and used for 10 periods?

Example: Machine replacement

The standard formulation of the problem:

Example: Machine replacement

Example: Machine replacement

The problem to be solved is then

Summary

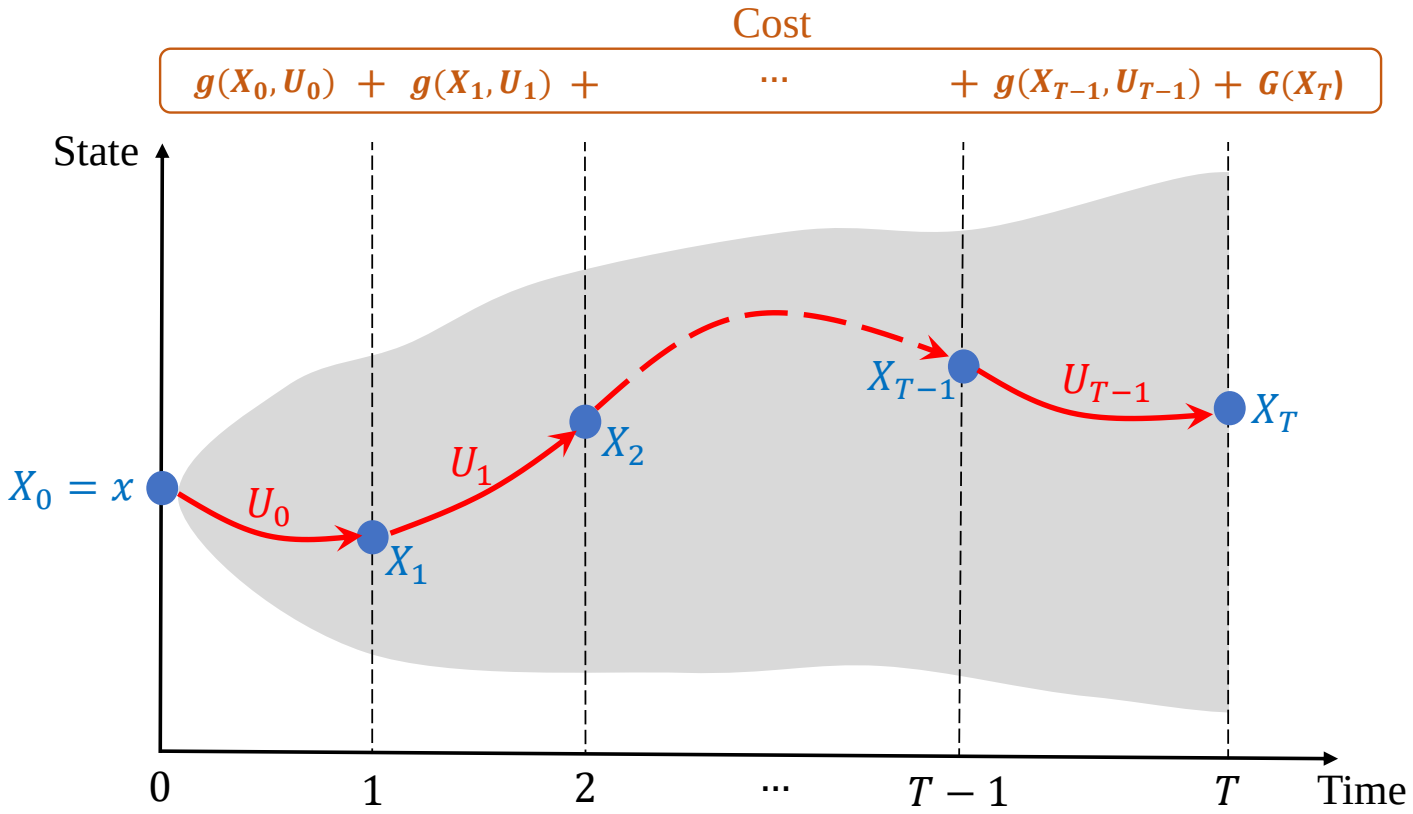
The first step of solving a stochastic optimal control problem is its standard formulation by identifying:

- the planning horizon,
- the state and action spaces,
- the dynamics (either recursion or transition probability kernel), and,
- the running and terminal costs.

Modeling, i.e., formulating the optimal control problem, is one of the most difficult parts of real applications (and exams).

Among different elements of the problem formulation, proper identification of the state variable is of utmost importance: It should contain all the necessary and sufficient information required for the process to be Markovian.

Stochastic optimal control



$$J_0^\star(x) \coloneqq \min_{(\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} g\big(X_t, \mu_t(X_t)\big) + G(X_T) \;\middle|\; X_0 = x \right), \qquad \forall x \in \mathbb{X}.$$

Difficulty of the problem

Consider a finite MDP with $\mathbb{X} = [n]$ and $\mathbb{U} = [m]$ with $n, m \in \mathbf{N}$.

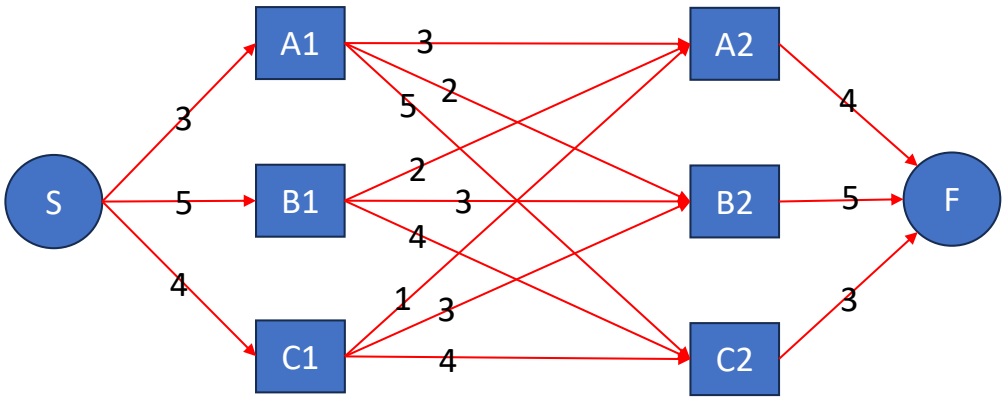
Question: What is the total number of open-loop policies $(u_t)_{t=0}^{T-1} \in \mathbb{U}^T$?

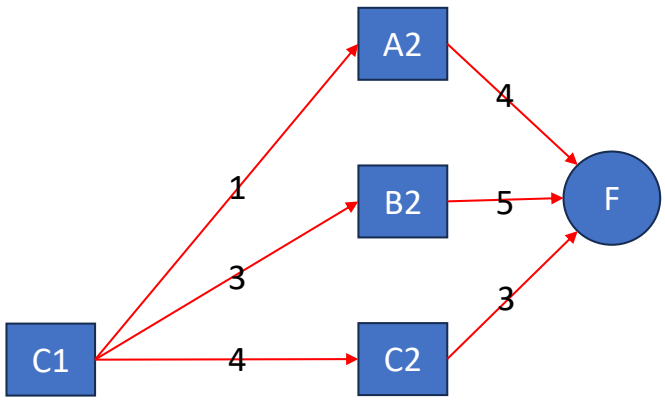
Question: What is the total number of closed-loop policies $(\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{T-1}$?

Shortest path problem

Question: Find the shortest path

- from 'S' to 'F'
- from 'C1' to 'F'

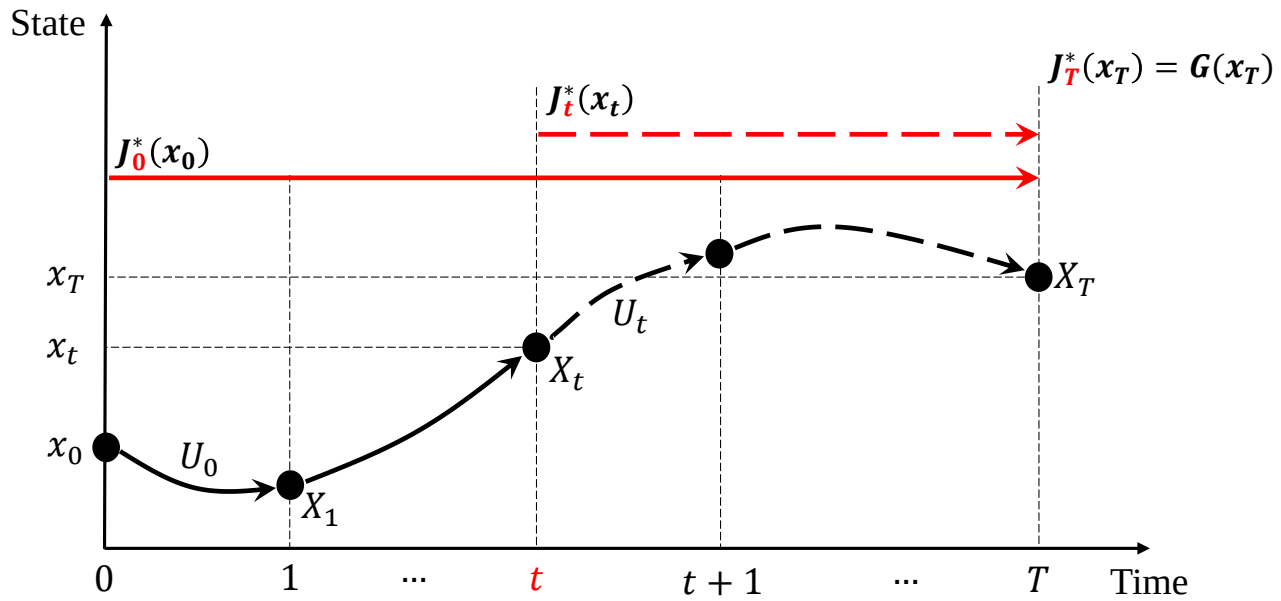




Principle of optimality

Consider the tail problem $\mathcal{P}_t$ from time t to terminal time T:

$$J_t^\star(x) \coloneqq \min_{(\mu_k:\mathbb{X}\to\mathbb{U})_{k=t}^{T-1}} \mathbb{E}\left(\sum_{k=t}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_t = x \right), \qquad \forall x \in \mathbb{X}$$

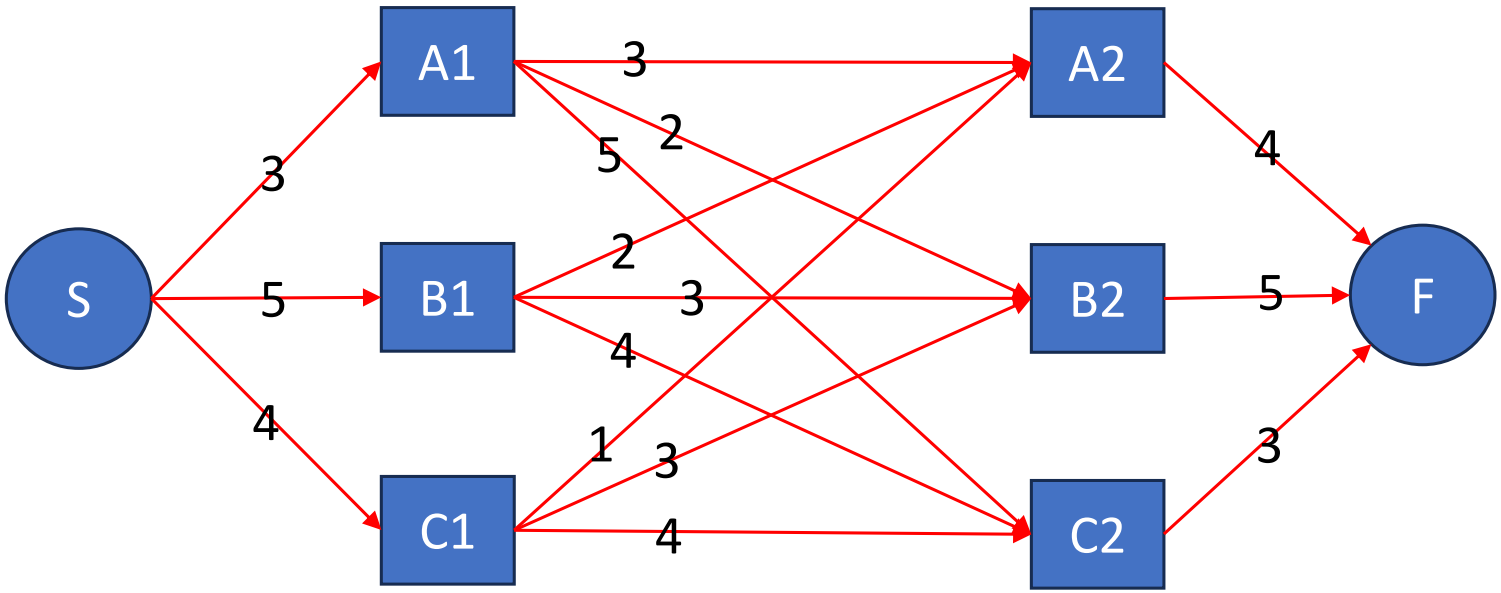


Principle of optimality: If the policy $(\mu_k^\star)_{k=0}^{T-1}$ is optimal in the problem $\mathcal{P}_0$, then the tail policy $(\mu_k^\star)_{k=t}^{T-1}$ is optimal in the tail problem $\mathcal{P}_t$ for all t.

Shortest path problem (cont'd)

Question: Use the principle of optimality to find the shortest path 'S' → 'F'.

Hint: Solve the problem backward from 'F' to 'S'!



Dynamic programming algorithm (DPA)

Use the principle of optimality and solve the tail problems one by one backward in time!

Algorithm (DPA):

(1) Initialization at $t = T$: Set $J_T(x) = G(x), \ \forall x \in \mathbb{X}$.

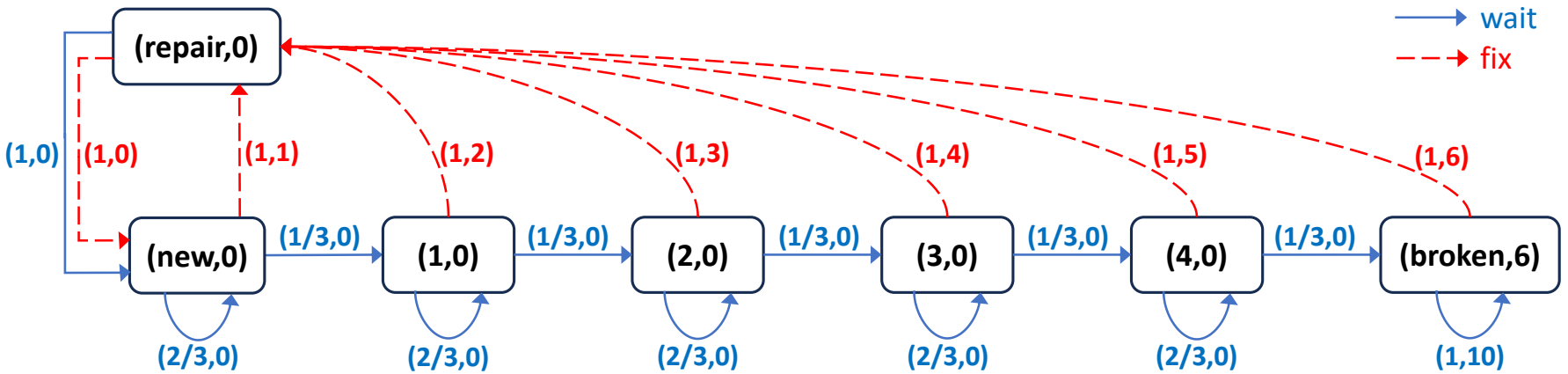
(2) Backward iteration for $t = T-1, \ldots, 0$: Set[1]

$$
\begin{aligned}
J_t(x) &= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \,\middle|\, X_t = x, U_t = u \right) \right\} \\
&= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}\big(f(x,u,W_{t+1})\big) \,\middle|\, X_t = x, U_t = u \right) \right\} \\
&= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \sum_{x_+ \in \mathbb{X}} P_u(x,x_+) J_{t+1}(x_+) \right\}, \quad \forall x \in \mathbb{X}, \\
\mu_t(x) &= \operatorname*{argmin}_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \,\middle|\, X_t = x, U_t = u \right) \right\}, \quad \forall x \in \mathbb{X}.
\end{aligned}
$$

Lemma (DPA): DPA outputs the optimal costs-to-go and an optimal policy.

[1] The third equality is under the assumption that the state space $\mathbb{X}$ is finite.

Example: Machine replacement (cont'd)



$J_t(x)/\mu_t(x)$

$t\backslash x$	repair	new	1	2	3	4	broken
10	$*$	$*$	$*$	$*$	$*$	$*$	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	$*$	$*$	$*$	$*$	$*$	$*$	$*$

Example: Machine replacement (cont'd)

Dynamic programming algorithm:

Initialization at $t = T = 10$:

$$J_{10}(x) = G(x) = \begin{cases} 6 & \text{if } x = \text{broken}, \\ 0 & \text{otherwise.} \end{cases}$$

$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	6.00/-

Backward iteration for $t = T - 1 = 9$ to $t = 0$:

$$\begin{aligned} J_t(x) &= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u \right) \right\} \\ &= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \sum_{x_+ \in \mathbb{X}} P_u(x, x_+) J_{t+1}(x_+) \right\}, \quad \forall x \in \mathbb{X}. \end{aligned}$$

Example: Machine replacement (cont'd)

Iteration $t = 9$:

$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	6.00/-
9	0.00/w	0.00/w	0.00/w	0.00/w	0.00/w	2.00/w	6.00/**f**

Example: Machine replacement (cont'd)

Iteration $t = 8$:

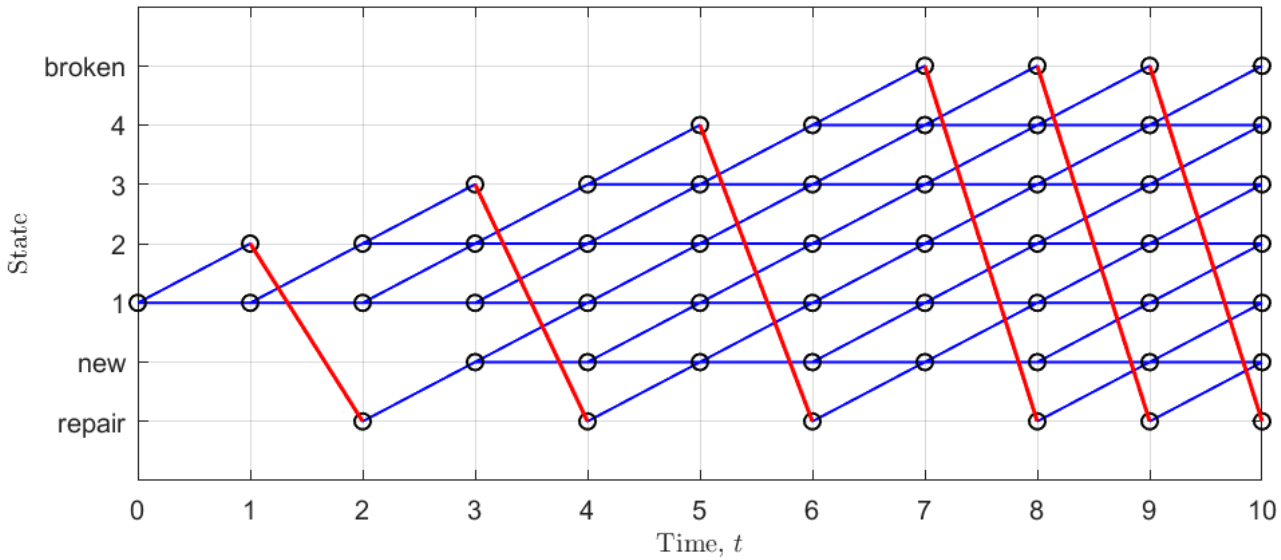
$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	6.00/-
9	0.00/w	0.00/w	0.00/w	0.00/w	0.00/w	2.00/w	6.00/**f**
8	0.00/w	0.00/w	0.00/w	0.00/w	0.67/w	3.33/w	6.00/**f**

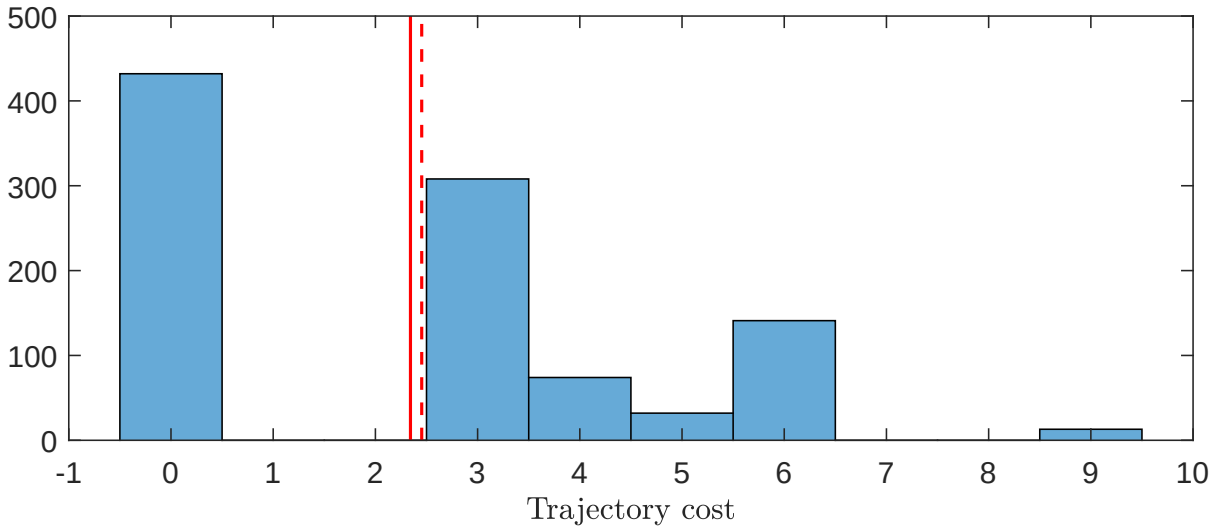
Example: Machine replacement (cont'd)

$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	6.00/-
9	0.00/w	0.00/w	0.00/w	0.00/w	0.00/w	2.00/w	6.00/f
8	0.00/w	0.00/w	0.00/w	0.00/w	0.67/w	3.33/w	6.00/f
7	0.00/w	0.00/w	0.00/w	0.22/w	1.56/w	4.22/w	6.00/f
6	0.00/w	0.00/w	0.07/w	0.67/w	2.44/w	4.81/w	6.00/f
5	0.00/w	0.02/w	0.27/w	1.26/w	3.23/w	5.00/f	6.00/f
4	0.02/w	0.11/w	0.60/w	1.92/w	3.82/w	5.00/f	6.00/f
3	0.11/w	0.27/w	1.04/w	2.55/w	4.02/f	5.02/f	6.02/f
2	0.27/w	0.53/w	1.54/w	3.04/w	4.11/f	5.11/f	6.11/f
1	0.53/w	0.87/w	2.04/w	3.27/f	4.27/f	5.27/f	6.27/f
0	0.87/w	1.26/w	2.45/w	3.53/f	4.53/f	5.53/f	6.53/f

Example: Machine replacement (cont'd)

1000 simulations of the system under the optimal policy:





Difficulty of the problem – revisited

Consider a finite MDP with $\mathbb{X} = [n]$ and $\mathbb{U} = [m]$ with $n, m \in \mathbf{N}$.

Question: What is the total number of closed-loop policies $(\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{T-1}$ while using DPA?

Summary: DPA and its extension

DPA can be extended in a straightforward manner to handle

- time-varying running cost g_t and dynamics f_t,
- stochastic running cost $g(x, u, w)$,
- state-dependent input constraints $u \in \mathbb{U}(x)$ for each $x \in \mathbb{X}$.

DPA:

(1) Initialize at $t = T$: For each $x \in \mathbb{X}$, set $J_T(x) = G(x)$.

(2) Iterate for $t = T-1, \ldots, 0$: For each $x \in \mathbb{X}$, set

$$J_t(x) = \min_{u \in \mathbb{U}(x)} \mathbb{E}\left(g_t(x, u, W_{t+1}) + J_{t+1}\big(f_t(x, u, W_{t+1})\big) \,\middle|\, X_t = x, U_t = u\right),$$

$$\mu_t(x) = \operatorname*{argmin}_{u \in \mathbb{U}(x)} \mathbb{E}\left(g_t(x, u, W_{t+1}) + J_{t+1}\big(f_t(x, u, W_{t+1})\big) \,\middle|\, X_t = x, U_t = u\right).$$