Finite horizon Markov decision processes (MDPs)
(Lecture-1)

Example 1: Machine replacement

A problem over N stages
(e.g. think of maintaining a bus, with each stage corresponding to a month).

The machine can be in one of the "n" states, i.e.,

$\{1, \dots, n\}$

1: perfect condition

n: worst condition

"Don't confuse state with stage"

Operating cost: $g(i)$ in state i

$$g(1) \le g(2) \le \dots \le g(n)$$

Actions: "Repair" or "Do nothing"

"Repair" → machine becomes new (i.e., goes to state 1)

"Do nothing" → "machine gets progressively worse"

"Stochastic system"

P_{ij} : probability that the machine transitions from state i to j "when you do nothing".

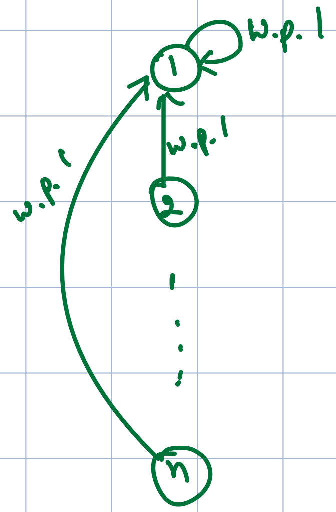
P_{ij} → transition probabilities

$$\sum_{j=1}^{n} P_{ij} = 1, \quad P_{ij} = 0 \text{ if } j < i$$

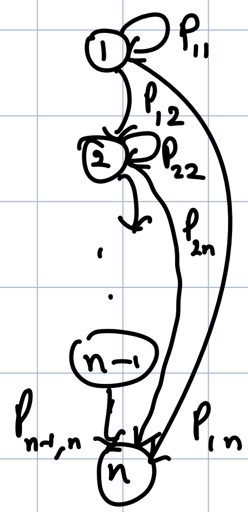
If you repair, the machine transition to state "1" & stays there for one stage

Transition diagram:

"Repair"



"Do nothing"

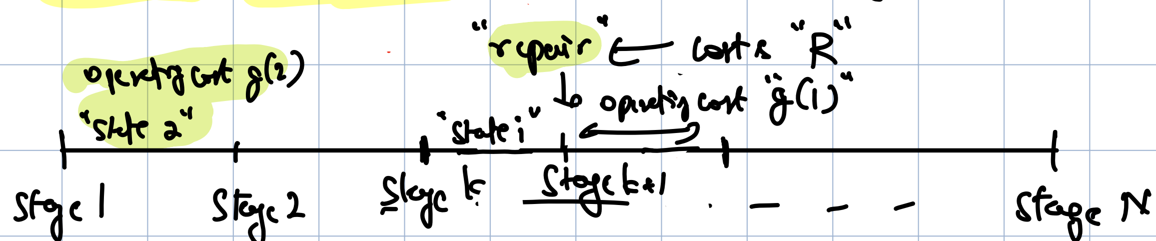


w.p. = with probability

"$P_{ij} = 0$ if $j < i$"

On "repair", machine goes to state "1", & remains there for one stage. Repair cost is "R".

Goal: choose actions so that the total cost (i.e., cumulative cost over N stages) is minimized.



Takeaway: MDP has "states", "actions", "transition probabilities", & "costs".

Another example:

"Chess match"

"Fix opponent"

You play 2 games →
- 1 pt win
- 0 loss
- $\frac{1}{2}$ draw

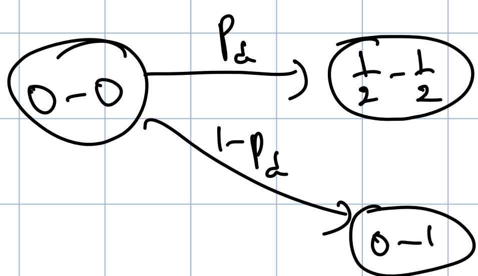
If there is a tie, then "sudden death" phase where you play games one after the other until a decisive result.

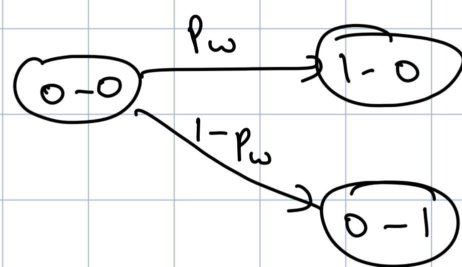
Actions:
- timid: → draw w.p. p_d; → lose w.p. $1-p_d$
- bold: → win w.p. p_w; → lose w.p. $1-p_w$

"$p_d > p_w$"

"Timid play" "Bold play"

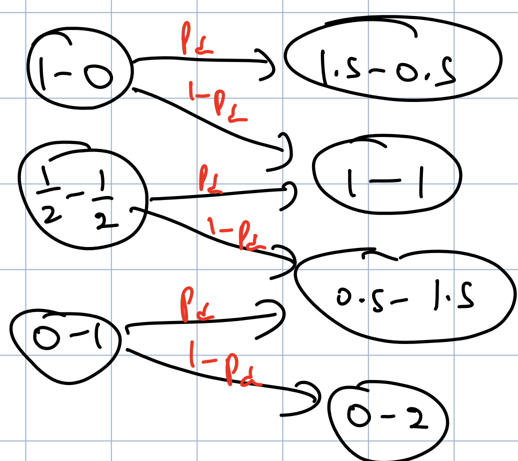
Game 1:





"Timid" "Bold"

Game 2:



H.W.

Sudden death: Play bold.

MDP framework:

Let $\mathcal{X}$ denote the state space,
$\mathcal{A}$ denote the action space.

$x_k \rightarrow$ state in stage k, $x_k \in \mathcal{X}$

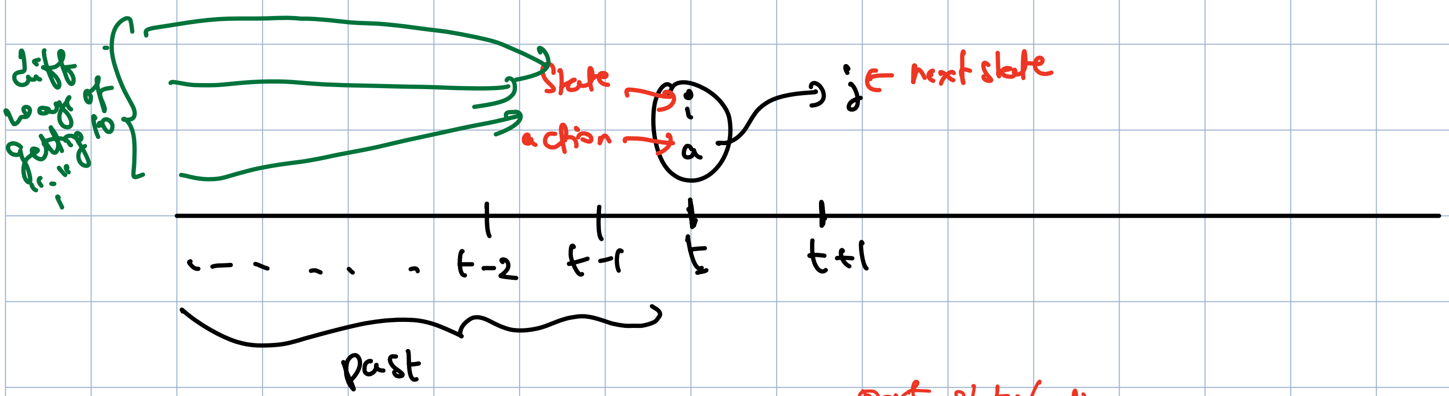
$a_k \rightarrow$ action that can be taken in state x_k, $a_k \in \mathcal{A}(x_k) \subset \mathcal{A}$.

set of possible actions depend on the current state

Transition probabilities:

$$P_{ij}(a) = P(x_{k+1} = j \mid x_k = i, a_k = a)$$

State evolution satisfies "Controlled Markov property".



$$P(\overbrace{x_{t+1} = j}^{\text{future}} \mid \overbrace{x_t = i, a_t = a}^{\text{current state \& action}}, \overbrace{x_{t-1}, a_{t-1}, \dots x_0}^{\text{past states/actions}})$$

$$= P_{ij}(a)$$

"Single-stage cost" $g_k(i, a, j)$

In stage k, if you are in state i & chose to take action a to transition to state j, the cost incurred is $g_k(i, a, j)$

$g_N(i) \to$ Terminal cost

final stage if state is i, then cost is $g_N(i)$

Policy $\{\mu_0, -----, \mu_{N-1}\}$ $\qquad$ $\text{state} \to \boxed{\mu_k} \to \text{action}$

$\mu_k : \mathcal{X} \to \mathcal{A}$ is a mapping from states to actions for stage k.

Lecture-2*

"**Admissible policy**": $\mu_k(i) \in \mathcal{A}(i) \leftarrow$ set of feasible action in state i

We say a policy $\pi = \{\mu_0, --- \mu_{N-1}\}$ is admissible if $\mu_k(i) \in \mathcal{A}_k(i) \quad \forall k = 0, -- N-1 \;\&\; \forall i \in \mathcal{X}$.

Total cost: Initial state $x_0 \in \mathcal{X}$

$$J_\pi(x_0) = \mathop{E}_{x_1, --- x_N}\left[\underbrace{g_N(x_N)}_{\text{terminal cost}} + \sum_{k=0}^{N-1} \underbrace{g_k(\underset{\text{current state}}{x_k}, \underset{\text{action chosen}}{\mu_k(x_k)}, \underset{\text{next state}}{x_{k+1}})}_{\text{cost in stage } k}\right]$$

$\pi = \{\mu_0, -- \mu_{N-1}\}$

state transition here

Stage k $\qquad$ Stage $k+1$

"Optimization objective"

$$\underbrace{J_{\pi^*}(x_0)}_{\text{optimal total cost}} = \min_{\pi \in \underbrace{\Pi}_{\text{set of "admissible" policies}}} \underbrace{J_\pi(x_0)}_{\text{total cost of policy } \pi}, \qquad \forall x_0 \in \mathcal{X}$$

π^*: optimal policy

$$\pi^* = \operatorname*{argmin}_{\pi \in \Pi} J_\pi(x_0)$$

Open vs. closed loop policies:

Closed loop policy: $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$

μ_i: decision based on the state

$\mu_i(x_i)$, $x_i \in \mathcal{X}$.

Open loop policy: Fix the sequence of actions beforehand

Chess match (revisited):

Possible open loop policies:

	Game 1	Game 2	Prob(win)?
Policy π_1	Timid	Timid	$p_d^2 p_w$
Policy π_2	Bold	Bold	$p_w^2 + 2p_w^2(1-p_w)$
Policy π_3	Bold	Timid	$p_w p_d + p_w^2(1-p_d)$
Policy π_4	Timid	Bold	$p_w p_d + p_w^2(1-p_d)$

Lets ignore π_1 (If $3p_w > p_d$, then π_2 is better than π_1)

Which among π_2, π_3, π_4 is the best?

$$\max\left(p_w^2(3-2p_w),\ p_w p_d + p_w^2(1-p_d) \right)$$

$$= \max\left(p_w^2 + 2p_w^2(1-p_w),\ p_w^2 + p_w p_d(1-p_w) \right)$$

$$= p_w^2 + p_w(1-p_w)\max(2p_w, p_d)$$

If $2p_w < p_d$, then π_3 / π_4 are better

Else, π_2 is better

Set $p_w = 0.45$, $p_d = 0.9$, Then $\underbrace{\text{Prob(winning match)} = 0.425}$

< 50% chance of match win

Closed loop policy: Play timid if leading, else play bold.
π_c

Prob(match win with π_c)

$$= p_w p_d + p_w \left(p_w (1-p_d) + (1-p_w) p_w \right)$$

$$= p_w^2 (2-p_w) + p_w (1-p_w) p_d$$

If $p_w = 0.45$, $p_d = 0.9$, then Prob(winning match with π_c)

$= 0.53 \rightarrow$ > 50% chance of winning match

Optimality principle

Let $\pi^* = \{ \mu_0^*, \mu_1^*, \cdots, \mu_{N-1}^* \}$ denote the optimal policy.

Consider the tail sub-problem

$$J_i^*(x_i) = \min_{\pi^i = (\mu_i, \cdots, \mu_{N-1})} E \left(g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}) \right) \quad (*)$$

Optimality principle

The policy $\{ \mu_i^*, \cdots, \mu_{N-1}^* \}$ is optimal for the $(N-i)$ stage problem in $(*)$

Proof: see Sec 1.5 of Bertsekas DPOC. Vol. I

DP algorithm:

$$\boxed{\begin{aligned} &\text{Set} \quad J_N(x_N) = g_N(x_N), \ \forall x_N \in \mathcal{X} \\ &\text{For} \quad k = N-1, \ldots\ldots, 0 \\ &\{ \\ &\quad J_k(x_k) = \min_{a_k \in A(x_k)} E_{x_{k+1}} \Big(\underbrace{g_k(x_k, a_k, x_{k+1}) + J_{k+1}(x_{k+1})}_{\text{sol}^n \text{ of tail problem}} \Big) \\ &\qquad\qquad\qquad\qquad\qquad \forall x_k \in \mathcal{X}. \\ &\} \end{aligned}}$$

Idea: Going backward, J_0 from DP algorithm is the optimal cost i.e., $J_{\pi^*} \equiv J^*$

Applying DP algorithm to "machine replacement" example:

$$J_N(i) = 0 \quad (\text{no terminal cost})$$

For $i = 1 \ldots n$, $$J_k(i) = \min\Big(\underbrace{R + g(1) + J_{k+1}(1)}_{\text{repair}}, \ \underbrace{g(i) + \sum_{j=i}^{n} P_{ij} J_{k+1}(j)}_{\text{do nothing}} \Big)$$

Lecture - 3

"DP algorithm finds the best policy"

Claim: $\forall x_0 \in \mathcal{X}$, the function $J_0(x_0)$ obtained at the end of the DP algorithm coincides with the optimal cost $J^*(x_0) (= J_{\pi^*}(x_0))$.

Proof: For any admissible policy $\pi = \{\mu_0, \ldots, \mu_{N-1}\}$, let $\pi^k = \{\mu_k, \ldots, \mu_{N-1}\}$

$J_k^*(x_k)$ be the optimal cost of the tail sub-problem beginning in stage k, in state x_k.

$$J_k^*(x_k) = \min_{\pi^k = \{\mu_k, \dots, \mu_{N-1}\}} E\left(g_N(x_N) + \sum_{i=k}^{N-1} g_i(x_i, \mu_i(x_i), x_{i+1}) \right)$$

Claim: $J_k^*(x_k) = \underbrace{J_k(x_k)}_{\text{obtained by DP algorithm}}$, $\forall k$

<Pf written Pf>

Base Case: $J_N^*(x_N) = g_N(x_N) = J_N(x_N)$, $\forall x_N \in X$

Induction hypothesis: Assume the claim for "$k+1$", i.e,

$$J_{k+1}^*(x_{k+1}) = J_{k+1}(x_{k+1}), \ \forall x_{k+1}$$

$$J_k^*(x_k) = \min_{(\mu_k, \pi^{k+1})} E_{x_{k+1}, \dots, x_N} \Big(g_N(x_N) + g_k(x_k, \mu_k(x_k), x_{k+1}) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu_i(x_i), x_{i+1}) \Big)$$

$$= \min_{\mu_k} E_{x_{k+1}} \Big(g_k(x_k, \mu_k(x_k), x_{k+1}) + \underbrace{\min_{\pi^{k+1}} E_{x_{k+2}, \dots, x_N} \Big[g_N(x_N) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu_i(x_i), x_{i+1}) \Big| x_{k+1} \Big]}_{\text{splitting of mins} \to \text{optimality principle}} \Big)$$

splitting of mins → optimality principle (for rigorous proof, check Sec 1.5 of DPOC Vol. I)

$$J_k^*(x_k) = \min_{\mu_k} E_{x_{k+1}} \Big(g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}^*(x_{k+1}) \Big)$$

Induction hypothesis

$$= \min_{\mu_k} E\left(g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}(x_{k+1}) \right)$$

$$= \min_{a_k \in A(x_k)} E\left(g_k(x_k, a_k, x_{k+1}) + J_{k+1}(x_{k+1}) \right)$$

$$= J_k(x_k)$$

To infer this, we used

$$\min_{\mu \in B} F(x, \mu(x)) = \min_{a \in A(x)} F(x, a).$$

$B = \{\mu \mid \mu(x) \in A(x)\}$

So, $\quad J_k^*(x_k) = J_k(x_k)$

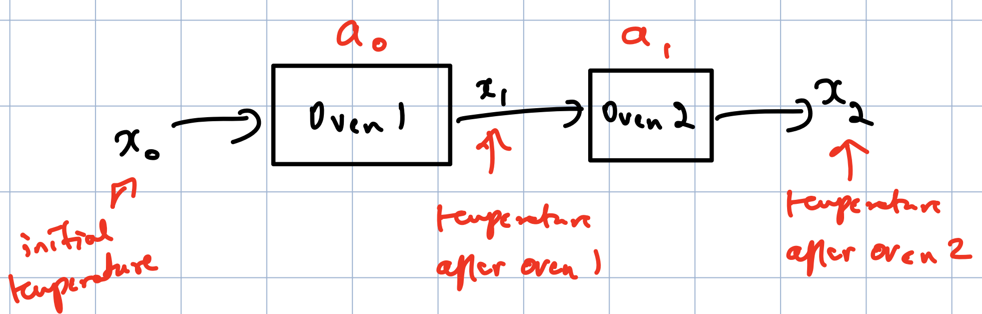
$\Rightarrow \quad J_0(x_0) = J^*(x_0)$

(End of pf within Pf) ■

Lecture-4*

Examples:

① Linear system with quadratic cost



Goal: to get x_2 as close as possible to a target temperature T

State Evolution: $\quad \underbrace{x_{k+1} = (1-\alpha)x_k + \alpha a_k}_{\text{linear state evolution}}, \quad k = 0, 1 \qquad (*)$

$\alpha \in (0,1)$ fixed

Total cost : $a_0^2 + a_1^2 + (x_2 - T)^2$
(to be minimized)

Apply DP algorithm:

Final stage : $J_2(x_2) = (x_2 - T)^2$ } Terminal cost

Going back one stage :

$$J_1(x_1) = \min_{a_1} \left(a_1^2 + J_2(x_2) \right)$$

$$= \min_{a_1} \left(a_1^2 + J_2\big((1-\alpha)x_1 + \alpha a_1\big) \right)$$

$$= \min_{a_1} \left(a_1^2 + \big((1-\alpha)x_1 + \alpha a_1 - T\big)^2 \right)$$

Linear in x_1 →

$$\boxed{\mu_1^*(x_1) = \frac{\alpha\left(T - (1-\alpha)x_1\right)}{1+\alpha^2}}$$

$$\boxed{J_1^*(x_1) = \frac{\big((1-\alpha)x_1 - T\big)^2}{1+\alpha^2}}$$

→ quadratic in x_1

Check:

$$\mu_0^*(x_0) = \frac{(1-\alpha)\alpha\left(T - (1-\alpha)^2 x_0\right)}{1+\alpha^2\left(1+(1-\alpha)^2\right)}$$

← linear in x_0

$$J_0^*(x_0) = \frac{\big((1-\alpha)^2 x_0 - T\big)^2}{1+\alpha^2\left(1+(1-\alpha)^2\right)}$$

← quadratic in x_0

Adding randomness to ovens:

Stochastic "state" evolution

$$x_{k+1} = (1-\alpha)x_k + \alpha a_k + \omega_k, \quad k=0,1$$

zero-mean r.v. iid with bounded variance (e.g. $N(0,\sigma^2)$) (indep of $\{x_k\}$)

Applying DP algorithm:

$$J_1(x_1) = \min_{a_1} E_{\omega_1}\left(a_1^2 + \left((1-\alpha)x_1 + \alpha a_1 + \omega_1 - T\right)^2\right)$$

$$= \min_{a_1} \Big(a_1^2 + \left((1-\alpha)x_1 + \alpha a_1 - T\right)^2$$

$$+ 2 \cancel{E\omega_1}^{0} \left((1-\alpha)x_1 + \alpha a_1 - T\right)$$

$$+ E\omega_1^2 \Big)$$

$$= \min_{a_1} \left(a_1^2 + \left((1-\alpha)x_1 + \alpha a_1 - T\right)^2 + E\omega_1^2 \right)$$

Minimizing RHS above leads to the same action as in the deterministic setting, i.e.,

$$\mu_1^*(x_1) = \frac{\alpha\left(T - (1-\alpha)x_1\right)}{1+\alpha^2}$$

Chess match - revisited (last time)

Consider an extension to N games

timid $\to p_d$, bold $\to p_w$

$p_d > p_w$

Players play N games & enter sudden death if the score is tied.

State : net score (e.g. (0-1) state = -1)

Apply DP algorithm:

(*) $J_k(x_k) = \max\Big(\underbrace{p_d J_{k+1}(x_k) + (1-p_d) J_{k+1}(x_k - 1)}_{\text{timid play}},$

we are playing for rewards (change min to max in DP algo)

$\underbrace{p_w J_{k+1}(x_k+1) + (1-p_w) J_{k+1}(x_k-1)}_{\text{bold play}}\Big)$

$$J_N(x_N) = \begin{cases} 1 & \text{if } x_N > 0 \\ p_w & \text{if } x_N = 0 \\ 0 & \text{if } x_N < 0 \end{cases}$$

It is better to play bold when

$$\frac{p_w}{p_d} > \frac{J_{k+1}(x_k) - J_{k+1}(x_k - 1)}{J_{k+1}(x_k+1) - J_{k+1}(x_k - 1)}$$

→ inferred from (*)

Given that we have J_N specified, we can go back & calculate J_{N-1}.

x_{N-1}	J_{N-1}	Best play
>1	1	does not matter
1	$\max(\underbrace{p_d + (1-p_d)p_w}_{\text{Timid}}, \underbrace{p_w + (1-p_w)p_w}_{\text{Bold}})$ $= p_d + (1-p_d)p_w$	Timid
0	p_w	Bold
-1	p_w^2	Bold
<-1	0	does not matter

For the 2-game match → we can figure the optimal strategy by knowing $J_{N-2}(0)$

$$J_{N-2}(0) = \max\Big(\underbrace{p_d p_w + (1-p_d)p_w^2}_{\text{Timid}},\ \underbrace{p_w\big(p_d + (1-p_d)p_w + (1-p_w)p_w^2\big)}_{\text{bold play}}\Big)$$

$$= \max\Big\{p_w\big(p_d + (1-p_d)p_w\big),\ p_w\big[p_d + (1-p_d)p_w + (1-p_w)p_w\big]\Big)$$

$$= p_w\big(p_d + (1-p_d)p_w + (1-p_w)p_w\big]\Big) \Rightarrow \text{play bold}$$

As noted before, one could choose $p_w < 0.5$ & still get a better than 50-50 chance of winning the match if $p_w\big(p_d + (1-p_d)p_w + (1-p_w)p_w\big]\Big) > 0.5$

"Skipped in class"

Another example: (Job scheduling)

N jobs to schedule

$T_i \rightarrow$ time taken for ith job to complete

T_i is a r.v. $\{T_i, i=1 \dots N\}$ independent

Each job "i" has a reward R_i associated with it.

So, if job "i" finishes at time "t", then the reward is $\alpha^t R_i$, α = discount factor $0<\alpha<1$

Cumulative reward = sum of each job's reward.

Goal: schedule jobs to maximize cumulative reward.

"Interchange argument" to figure optimal schedule

$$L = \{ i_0, i_1, \dots, i_{k-1}, i, j, i_{k+2}, \dots, i_{N-1} \}$$

$$L' = \{ i_0, i_1, \dots, i_{k-1}, j, i, i_{k+2}, \dots, i_{N-1} \}$$

$$J_L = E\left[\alpha^{t_0} R_{i_0} + \dots + \alpha^{t_{k-1}} R_{i_{k-1}} + \alpha^{t_{k-1}+T_i} R_i + \alpha^{t_{k-1}+T_i+T_j} R_j + \dots + \alpha^{t_{N-1}} R_{i_{N-1}} \right]$$

$$J_{L'} = E\left[\alpha^{t_0} R_{i_0} + \dots + \alpha^{t_{k-1}} R_{i_{k-1}} + \alpha^{t_{k-1}+T_j} R_j + \alpha^{t_{k-1}+T_i+T_j} R_i + \dots + \alpha^{t_{N-1}} R_{i_{N-1}} \right]$$

Schedule L is better than L' if

$$E\left[\alpha^{t_{k-1}+T_i} R_i + \alpha^{t_{k-1}+T_i+T_j} R_j \right] \geq E\left[\alpha^{t_{k-1}+T_j} R_j + \alpha^{t_{k-1}+T_i+T_j} R_i \right]$$

Using t_{k-1}, T_i, T_j are independent,

$$\frac{E(\alpha^{T_i})R_i}{1-E(\alpha^{T_i})} \geq \frac{E(\alpha^{T_j})R_j}{1-E(\alpha^{T_j})} \quad \text{--- } (*)$$

From $(*)$, the optimal schedule works out as follows:

Assign $\mu_i = \dfrac{E(\alpha^{T_i})R_i}{1-E(\alpha^{T_i})}$ as the index for job i, $i=1 \cdots N$

Order $\{\mu_1, \cdots, \mu_N\}$, say $\mu_{[1]} \geq \mu_{[2]} \cdots \geq \mu_{[N]}$

$$\text{Optimal schedule} = \underbrace{\{[1], [2], \cdots [N]\}}_{\text{index-based optimal policy}}$$

Further reading: Check out Gittins index
Sec 1.3 of PPOC Vol. II

Yet-another example: <Optimal stopping>

"Asset-selling"

A technical note before asset-selling:

Discrete-time MDPs can be formulated as

$$x_{k+1} = f(x_k, a_k, \underbrace{w_k}_{\text{disturbance}})$$

$\{w_k\}$ could be i.i.d or could depend on x_k, a_k
$x_k \in$ infinite set.

For the case when $x_k \in \{1, \ldots n\}$, it is enough to know $P_{ij}^a = P(x_{k+1} = j \mid x_k = i, a_k = a)$

Now to asset-selling!

Want to sell an asset.

You get offers $w_0, w_1, \ldots, w_{N-1}$

Assume: $\{w_k\}$ iid with some distribution function that has finite mean

Action → sell the asset (by accepting the offer) a^1
↘ don't sell & wait for more offers a^2

Add a special state "T" to denote that the asset is sold.

Also, $x_0 = 0$

$x_{k+1} = f(x_k, a_k, w_k)$, where

$$f(x_k, a_k, w_k) = \begin{cases} T & \text{if } (x_k \neq T, a_k = a^1) \text{ or } (x_k = T) \\ w_k & \text{else.} \end{cases}$$

Work with rewards.

Goal: maximize $E\left[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, a_k, w_k) \right]$

expectation over $w_0, w_1, \ldots w_{N-1}$

$$g_N(x_N) = \begin{cases} x_N & \text{if } x_N \neq T \\ 0 & \text{else} \end{cases}$$

$$g_k(x_k, a_k, w_k) = \begin{cases} \underbrace{(1+r)^{N-k}}_{\text{"interest added"}} x_k & x_k \neq T,\ a_k = a' \\ 0 & \text{else} \end{cases}$$

asset not sold / sell

$1 > r > 0$ is the interest rate.

Apply DP algorithm:

$$J_N(x_N) = \begin{cases} x_N & \text{if } x_N \neq T \\ 0 & \text{else} \end{cases}$$

$$J_k(x_k) = \begin{cases} \max\Big(\underbrace{(1+r)^{N-k} x_k}_{\text{sell}},\ \underbrace{E_{w_k}\big(J_{k+1}(w_k)\big)}_{\text{don't sell}}\Big) & \text{if } x_k \neq T \\ 0 & \text{if } x_k = T \end{cases}$$

Then is x_{k+1} if one doesn't sell

Let $\alpha_k = \dfrac{E\big(J_{k+1}(w_k)\big)}{(1+r)^{N-k}} = \dfrac{E\big(J_{k+1}(w)\big)}{(1+r)^{N-k}}$ (Since w_k iid)

Optimal policy: <threshold-based policy> "non-stationary thresholds α_k"

sell if $x_k > \alpha_k$

don't sell if $x_k < \alpha_k$

if $x_k = \alpha_k$ both actions are fine

Understanding the optimal policy:

Lecture 5*

Suppose $\alpha_k \geq \alpha_{k+1} \quad \forall k$

For notational convenience, let $V_k(x_k) = \dfrac{J_k(x_k)}{(1+r)^{N-k}}$,

for $x_k \neq T$

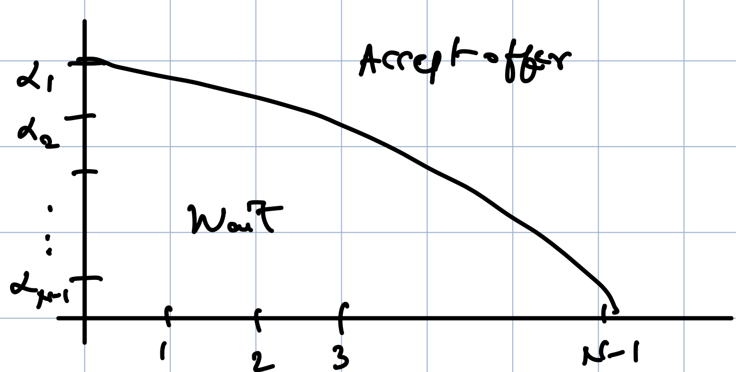
DP algo in "V" notation is

$$V_N(x_N) = x_N$$

$$V_k(x_k) = \max\left(x_k, \frac{E\left(V_{k+1}(w_k)\right)}{(1+r)} \right)$$

Optimal policy $\mu_k^*(x_k) = \begin{cases} a^1 & \text{if } x_k \geq \alpha_k \\ a^2 & \text{else} \end{cases}$

Claim: $\alpha_k \geq \alpha_{k+1}$



$$\alpha_k = \frac{E\left(V_{k+1}(w)\right)}{1+r}$$

To show $\alpha_k \geq \alpha_{k+1}$, it is enough if we establish that $V_k(x) \geq V_{k+1}(x) \quad \forall x$

For $k = N-1$,

$$V_{N-1}(x) = \max\left(x, \frac{E(V_N(\omega))}{1+r}\right)$$

$$= \max\left(x, \frac{E(\omega)}{1+r}\right) \geq x = V_N(x)$$

For $k = N-2$,

$$V_{N-2}(x) = \max\left(x, \frac{E(V_{N-1}(\omega))}{1+r}\right)$$

$$\geq \max\left(x, \frac{E(V_N(\omega))}{1+r}\right)$$

$$= V_{N-1}(x)$$

Proceeding similarly, we get $V_k(x) \geq V_{k+1}(x), \forall x, \forall k$

Understanding the asset selling problem for large N:

Suppose "ω" is a continuous, positive-valued r.v. with distribution F_ω & density h.

$$V_{k+1}(\omega) = \begin{cases} \omega & \text{if } \alpha_{k+1} \leq \omega \\ \alpha_{k+1} & \text{else} \end{cases}$$

rewriting the max in terms of V_{k+1}

$$\alpha_k = \frac{E\left(V_{k+1}(\omega)\right)}{1+r}$$

$$= \frac{1}{1+r}\int_0^\infty V_{k+1}(\omega)\, h(\omega)\, d\omega$$

$$= \frac{1}{1+r}\int_0^{\alpha_{k+1}} \alpha_{k+1}\, h(\omega) d(\omega) + \frac{1}{1+r}\int_{\alpha_{k+1}}^\infty \omega\, h(\omega)\, d\omega$$

$$\alpha_k = \frac{\alpha_{k+1}}{1+r} F_\omega(\alpha_{k+1}) + \frac{1}{1+r}\int_{\alpha_{k+1}}^\infty \omega\, h(\omega)\, d\omega$$

$$\alpha_k \le \frac{\alpha_{k+1}}{1+r} + \frac{E(\omega)}{1+r}$$

$$\boxed{\text{used } \int_{\alpha_{k+1}}^\infty \omega\, h(\omega)\, d\omega \le E(\omega)}$$

$\alpha_k \geq \alpha_{k+1}$

$$\alpha_k \le \frac{\alpha_k}{1+r} + \frac{E(\omega)}{1+r}.$$

From the above, we can conclude

$$0 \le \alpha_k \le \frac{E\omega}{r} \quad \text{——} \quad (*)$$

The sequence $\{\alpha_k\}$ is non-increasing and bounded because of $(*)$.

Since $\alpha_{k+1} \le \alpha_k$, the sequence $\{\alpha_k\}$ converges as $k \to -\infty$, say to some

$$\bar{\alpha} \in \left[0, \frac{E\omega}{r}\right]$$

$$\alpha_k = \frac{\alpha_{k+1}}{1+r} F_w(\alpha_{k+1}) + \frac{1}{1+r} \int_{\alpha_{k+1}}^{\infty} w\, h(w)\, dw$$

Taking $k \to -\infty$ in eqn above, we obtain

$$\underbrace{\bar{\alpha} = \frac{\bar{\alpha}\, F_w(\bar{\alpha})}{1+r} + \frac{1}{1+r} \int_{\bar{\alpha}}^{\infty} w\, h(w)\, dw}$$

Given the distribution F_w of w, we can calculate $\bar{\alpha}$

For very large N, an approximately optimal policy is

sell at x_k if $x_k > \bar{\alpha}$
don't sell at x_k if $x_k \le \bar{\alpha}$

Simpler "threshold" based policy because the threshold $\bar{\alpha}$ is the same for all stages.

<End of Finite horizon MDPs>