Lecture 14* (contd)

Infinite horizon discounted MDPs

(Ref: DPOC Vol. II, Chapter 1&2)

Goal:

$$J^*(i) = \min_{\pi \in \Pi} J_\pi(i), \quad \forall i, \quad \text{where}$$

($J^*(i)$: optimal discounted cost)

$$J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k g(x_k, \pi(x_k), x_{k+1}) \,\middle|\, x_0 = i \right)$$

(g: single-stage cost; α^k: discount, $0 < \alpha < 1$; π: policy; x_0: start state)

Let π^* denote the optimal policy i.e., $\arg\min_{\pi \in \Pi} J_\pi(i)$

Let $X = \{1, \ldots n\}$ denote the state space

(A1) We shall assume the single stage cost is bounded, i.e.,

$$|g(\cdot,\cdot,\cdot)| \leq M < \infty.$$

Note: We do not assume existence of a special termination state.

Bellman and another operator:

For $J = (J(1), \ldots J(n))$,

define Bellman operator T as follows:

$$(TJ)(i) = \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\left(g(i,a,j) + \alpha J(j)\right), \quad \forall i$$

for a stationary policy π,

$$(T_\pi J)(i) = \sum_{j=1}^{n} p_{ij}(\pi(i))\left(g(i,\pi(i),j) + \alpha J(j)\right), \quad \forall i$$

$$P_\pi = \begin{bmatrix} p_{11}(\pi(1)) & & \\ & \ddots & \\ & & p_{nn}(\pi(n)) \end{bmatrix}$$

$$g_\pi = \begin{pmatrix} \sum_{j=1}^{n} p_{1j}(\pi(1))\, g(1,\pi(1),j) \\ \vdots \\ \sum_{j=1}^{n} p_{nj}(\pi(n))\, g(n,\pi(n),j) \end{pmatrix} = \begin{pmatrix} g(1,\pi(1)) \\ \vdots \\ g(n,\pi(n)) \end{pmatrix}$$

if single stage cost is a function of the current state & current action, i.e., $g(i,a,j) = g(i,a)\ \forall j$.

So, $T_\pi J = g_\pi + \alpha P_\pi J$

Even in the discounted case, T, T_π are monotone.

Lemma 1: Let $J, J' \in \mathbb{R}^n$ & satisfy

$$J(i) \le J'(i), \quad \forall i$$

Then, for any $k = 1, 2, \ldots$

(i) $$(T^k J)(i) \le (T^k J')(i), \quad \text{and}$$

(ii) For any stationary policy π,

$$(T_\pi^k J)(i) \le (T_\pi^k J')(i)$$

Pf: H.W.

The constant shift lemma holds here as well.

Lemma 2:

Stationary π, $\delta \to$ positive scalar, $e \to$ vector of n ones.

Then, $\forall i = 1, \ldots n$, $\forall k = 1, 2, \ldots$, we have

(i) $$(T^k(J + \delta e))(i) = (T^k J)(i) + \alpha^k \delta$$

(ii) $$(T_\pi^k(J + \delta e))(i) = (T_\pi^k J)(i) + \alpha^k \delta$$

Note: It is an equality here, while the corresponding SSP claim had an inequality (owing to terminal state T).

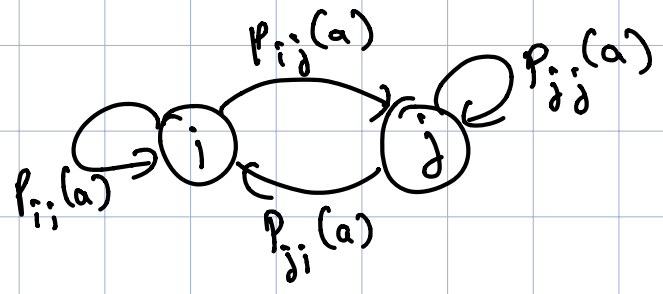
Pf: H.W.

"Every discounted problem has an equivalent SSP".

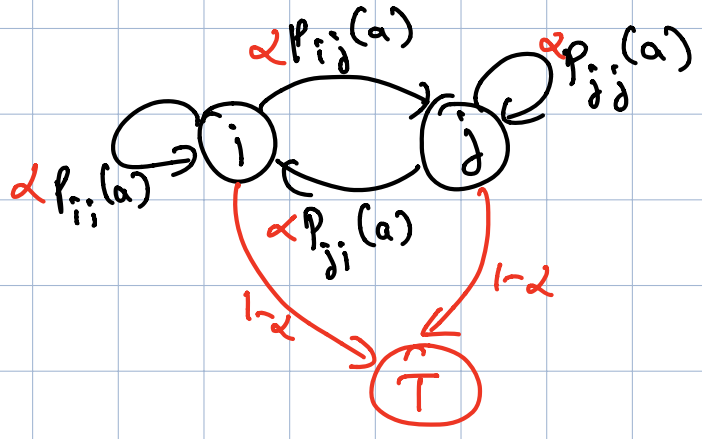
Given discounted MDP on states $\{1, \ldots n\}$,
form an SSP on states $\{1, \ldots n\} \cup \{T\}$

↑ add this extra state & make it cost-free & absorbing.

Idea: In the SSP, w.p. α pick a next state according to transition probabilities of discounted MDP
& w.p. $(1-\alpha)$ move to "T".



discounted - MDP



Equivalent SSP

Why are these two MDPs equivalent?

① Note that in the SSP, all policies are proper.

② In the discounted MDP, the expected kth stage cost is $E\left(\alpha^k g(i,a,j)\right) = \alpha^k \sum_j p_{ij}(a)\, g(i,a,j)$

③ In the SSP, the expected kth stage cost is

$$\left[\sum_j p_{ij}(a)\, g(i,a,j)\right] \times \alpha^k$$

If the terminal state is not hit upto kth stage, then the underlying probabilities will have a α^k multiplier.

Lecture-15

Prop 1: (VI converges)

Assume (A1).

For any finite J, the optimal cost satisfies

$$J^*(i) = \lim_{N\to\infty} (T^N J)(i), \quad \forall i$$

Corollary: For a stationary policy π, we have

$$J_\pi(i) = \lim_{N\to\infty} (T_\pi^N J)(i), \quad \forall i \text{ for any finite } J.$$

Pf) Given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ and a state $i \in X$,

$$J_\pi(i) = \lim_{N\to\infty} E\left(\sum_{l=0}^{N-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

→ For convenience, drop the dependence on i on RHS.

$$= E\left(\sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$+ \lim_{N\to\infty} E\left(\sum_{l=L}^{N-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

term B — ①

Since $|g(\cdot,\cdot,\cdot)| \le M$ by assumption,

$$|\text{term B}| \le M \sum_{l=L}^{\infty} \alpha^l = \frac{M\alpha^L}{1-\alpha} \quad -②$$

$$E\left(\sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$= J_\pi(i) - \text{term B}$$

$$E\left(\alpha^L J(x_L) + \sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$= J_\pi(i) - \text{term B} + E(\alpha^L J(x_L))$$

Using ②,

$$J_\pi(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j\in X} |J(j)|$$

$$\le E\left(\alpha^L J(x_L) + \sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

apply T_π L times on J

$$\le J_\pi(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j\in X} |J(j)| \quad -③$$

Taking minimum over π on all sides of (3), we obtain $\forall i \in X$ and any $L > 0$ that

$$J^*(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j \in X} |J(j)|$$

$$\leq (T^L J)(i) \longrightarrow \text{because } \min_{\pi} T_\pi^L = T^L$$

$$\leq J^*(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j \in X} |J(j)| \quad -(4)$$

Taking $L \to \infty$ on all sides of (4), we obtain (Note: $\alpha \in (0,1)$)

$$J^*(i) \leq \lim_{L \to \infty} (T^L J)(i) \leq J^*(i)$$

& the claim follows. ■

Corollary: The claim follows by considering an MDP where the only feasible action in a state i is $\pi(i)$, $\forall i$, and invoking Prop 1. (Note $T = T_\pi$ for this MDP).

Prop 2: (Bellman equation)

The optimal discounted cost J^* satisfies

$$J^* = T J^*$$

, i.e.,

$$J^*(i) = \min_a \sum_j p_{ij}(a) \left(g(i,a,j) + \alpha J^*(j)\right)$$

Also, J^* is the unique fixed point of T.

Pf) Eq ④ in the proof above is

$$J^*(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j \in X} |J(j)|$$

$$\leq (T^L J)(i)$$

$$\leq J^*(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j \in X} |J(j)| \quad — ⑤$$

Applying operator T on all sides,

$$TJ^*(i) - \frac{M\alpha^{L+1}}{1-\alpha} - \alpha^{L+1} \max_{j \in X} |J(j)| \leq (T^{L+1} J)(i)$$

$$\leq TJ^*(i) + \frac{M\alpha^{L+1}}{1-\alpha} + \alpha^{L+1} \max_{j \in X} |J(j)|$$

→ we used constant shift lemma.

Taking $L \to \infty$ on all sides of the equation above,

$$T J^*(i) = J^*(i), \quad \forall i.$$

Uniqueness: Let J' be another fixed point of T.

$$J' = T J' = T^2 J' = \cdots = \lim_{L \to \infty} T^L J' = J^*$$

■

Corollary: For a stationary policy π, the associated cost J_π satisfies

$$J_\pi = T_\pi J_\pi \quad \text{(or)}$$

$$J_\pi(i) = \sum_j p_{ij}(\pi(i)) \left(g(i, \pi(i), j) + \alpha J_\pi(j) \right), \quad \forall i$$

Also, J_π is the unique fixed point of T_π.

Pf) Follows from Prop 2.

■

Lecture-16*

Necessary & sufficient condition for optimal policy:

Prop 3: A stationary policy π is optimal if and only if $\pi(i)$ attains the minimum in the Bellman equation, $\forall i \in X$. Or, equivalently,

$$TJ^* = T_\pi J^*$$

Pf> Assume $TJ^* = T_\pi J^*$

We know $J^* = TJ^*$

So, $J^* = T_\pi J^*$

$\Rightarrow J^* = J_\pi \Rightarrow \pi$ is optimal.

Converse: π is optimal

$\Rightarrow J^* = J_\pi \Rightarrow J^* = T_\pi J^*$

Also, Bellman equation $\Rightarrow J^* = TJ^*$

So, $T_\pi J^* = TJ^*$

■

Contraction property of T and T_π:

Max-norm: $\|J\|_\infty = \max_{i=1 \sim n} |J(i)|$

we will show that T, T_π are α-contractions in $\|\cdot\|_\infty$.

Prop 4: For any two bounded J, J', and $\forall k \geq 1$

$$\| T^k J - T^k J' \|_\infty \leq \alpha^k \| J - J' \|_\infty \quad \text{— (*)}$$

Pf: Let $c = \max_{i=1 \sim n} |J(i) - J'(i)|$

$$J(i) - c \leq J'(i) \leq J(i) + c \quad \text{— ①}$$

Apply T^k on all sides of ① & use constant shift lemma to obtain

$$(T^k J)(i) - \alpha^k c \leq (T^k J')(i) \leq (T^k J)(i) + \alpha^k c$$

$$\Rightarrow \quad |(T^k J)(i) - (T^k J')(i)| \leq \alpha^k c$$ ← holds for any i

Taking max over i implies (*). ■

Corollary: For any stationary π & bounded J, J', and $\forall k \geq 1$

$$\| T_\pi^k J - T_\pi^k J' \|_\infty \leq \alpha^k \| J - J' \|_\infty$$

Value Iteration: Start with J_0 & repeatedly apply T.

Error bound for VI:

$$\| T^k J_0 - J^* \|_\infty \leq \alpha^k \| J_0 - J^* \|_\infty$$

Why? Set $J' = J^*$ in (*) & note $T^k J^* = J^*$.

Example: Machine replacement

Recall n-states 1 .. n

Operating cost $g(i)$

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

$p_{ij} \to$ transition probabilities

actions: do nothing & repair (Repair cost R)

Goal: minimize infinite horizon discounted cost
($\alpha \leftarrow$ discount factor)

Bellman equation:

$$J^*(i) = \min \Big\{ \underbrace{R + g(1) + \alpha J^*(1)}_{\text{repair}},$$

$$\underbrace{g(i) + \alpha \sum_{j=1}^{n} p_{ij} J^*(j)}_{\text{do nothing}} \Big\}$$

So, optimal action is to repair if

$$R + g(1) + \alpha J^*(1) < g(i) + \alpha \sum_{j=1}^{n} p_{ij} J^*(j)$$

& "do nothing" otherwise.

Assume · ① $P_{ij} = 0$ if $j < i$ ← machine can't get better if you do nothing

② $P_{ij} \leq P_{(i+1)j}$ if $i < j$

Suppose J is monotone non-decreasing, i.e.,

$J(1) \leq J(2) \cdots \leq J(n)$

Then,

$$\sum_{j=1}^{n} P_{ij} J(j) \leq \sum_{j=1}^{n} P_{(i+1)j} J(j), \quad i = 1 \cdots n-1$$

Since $g(i)$ is non-decreasing, we have

$(TJ)(i)$ is non-decreasing in i, if J is non-decreasing.

$\Rightarrow (T^k J)(i)$ is non-decreasing in i, $\forall k$

So, $J^*(i) = \lim_{k \to \infty} (T^k J)(i)$ is non-decreasing in i

So, the function $\left(g(i) + \alpha \sum_j P_{ij} J^*(j) \right)$

is non-decreasing in i.

Set of states $S_R = \left\{ i \;\middle|\; R + g(i) + \alpha J^*(1) \leq g(i) + \alpha \sum_j P_{ij} J^*(j) \right\}$

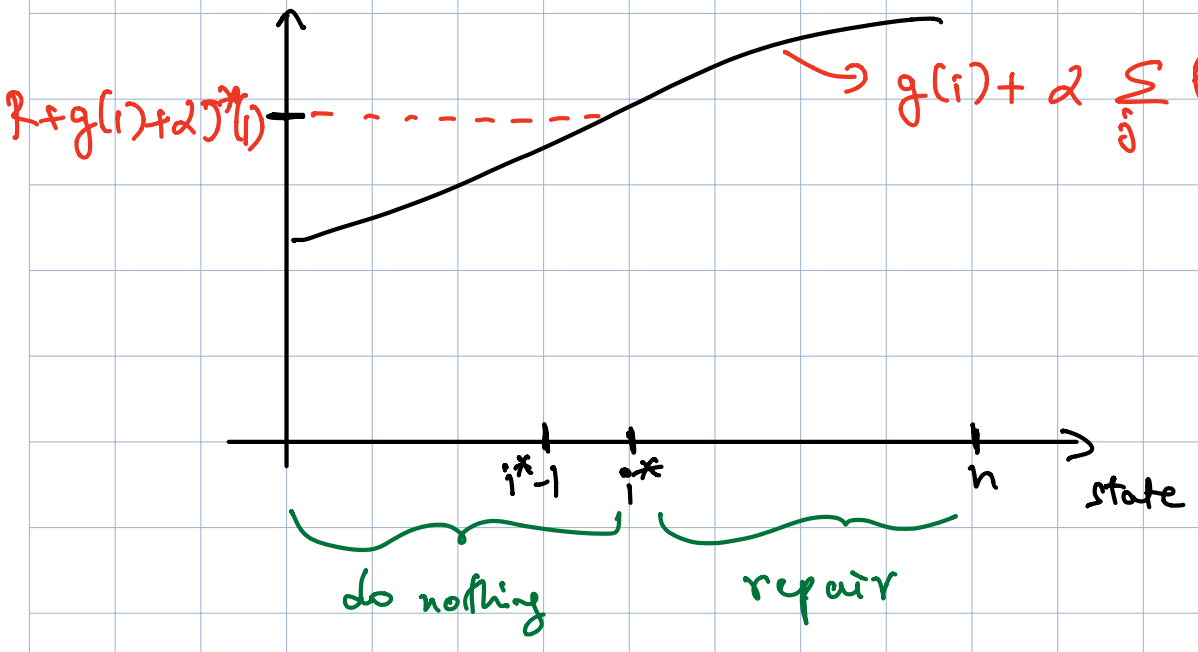
S_R = set of states where it is optimal to repair

$$i^* = \begin{cases} \text{smallest state in } S_R & \text{if } S_R \neq \emptyset \\ n+1 & \text{else} \end{cases}$$

Optimal policy

$$= \begin{cases} \text{repair} & \text{if } i \geq i^* \\ \text{do nothing} & \text{else} \end{cases}$$

} A threshold-based optimal policy



H.W. Think about policy iteration for this problem. In particular, if we start with a threshold-based policy & do policy improvement, then does it lead to another threshold policy?

If yes, then PI converges to optimal policy in at most n iterations.

Illustrative example for VI:

MDP $\quad S = \{1, 2\} \quad A = \{a, b\}$

$$P(a) = \begin{bmatrix} P_{11}(a) & P_{12}(a) \\ P_{21}(a) & P_{22}(a) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}$$

$$P(b) = \begin{bmatrix} P_{11}(b) & P_{12}(b) \\ P_{21}(b) & P_{22}(b) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Costs: $g(1,a) = 2, \quad g(1,b) = 0.5$

$g(2,a) = 1, \quad g(2,b) = 3$

Discount $\alpha = 0.9$

$J_0 = (0, 0)$

$$(TJ)(i) = \min\left\{ g(i,a) + \alpha \sum_{j=1}^{2} P_{ij}(a)\, J(j), \right.$$
$$\left. g(i,b) + \alpha \sum_{j=1}^{2} P_{ij}(b)\, J(j) \right\}$$

$J_1 = TJ_0 = (0.5, 1)$

$J_2 = (1.28, 1.56)$

& so on.

PI algorithm:

Step 1: Start with a policy π_0

Step 2: (Policy Evaluation) Evaluate π_k, i.e., compute J_π by solving $J = T_{\pi_k} J$

$$\Leftrightarrow J(i) = \sum_j p_{ij}(\pi_k(i))\left(g(i,\pi_k(i),j) + \alpha J(j)\right), \quad \forall i \quad (*)$$

(here $J(1) \ldots J(n)$ are the unknowns & solving $(*)$ given J_{π_k})

Step 3: Policy improvement

Find a new policy π_{k+1} by

$$T_{\pi_{k+1}} J_{\pi_k} = T J_{\pi_k}$$

$$\Leftrightarrow \pi_{k+1}(i) = \arg\min_{a \in A(i)} \sum_j p_{ij}(a)\left(g(i,a,j) + \alpha J_{\pi_k}(j)\right)$$

If $J_{\pi_{k+1}}(i) < J_{\pi_k}(i)$ for at least one state i,

then go to step 2 & repeat.

Remark: Policy improvement claim holds even in the discounted setting.

Policy improvement claim:

Let π, π' be two proper policies s.t.

$$T_{\pi'} J_\pi = T J_\pi$$

Then, $J_{\pi'}(i) \leq J_\pi(i) \quad \forall i$

with strict inequality for at least one of the states if π is not optimal.

Pf: Follows by a parallel argument to the proof in SSP case.

Lecture-17

PI example:

$S = \{1, 2\}, \quad A = \{a, b\}$

$$P(a) = \begin{bmatrix} P_{11}(a) & P_{12}(a) \\ P_{21}(a) & P_{22}(a) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}$$

$$P(b) = \begin{bmatrix} P_{11}(b) & P_{12}(b) \\ P_{21}(b) & P_{22}(b) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Costs: $g(1,a) = 2, \quad g(1,b) = 0.5$

$g(2,a) = 1, \quad g(2,b) = 3$

Discount $\alpha = 0.9$

Initialization: $\pi_0(1) = a, \quad \pi_0(2) = b$

Policy evaluation:

Finding J_{π_0}:

$$J_{\pi_0}(1) = g(1,a) + \alpha P_{11}(a) J_{\pi_0}(1) + \alpha P_{12}(a) J_{\pi_0}(2)$$

$$J_{\pi_0}(2) = g(2,b) + \alpha P_{21}(b) J_{\pi_0}(1) + \alpha P_{22}(b) J_{\pi_0}(2)$$

Using MDP data, we have

$$J_{\pi_0}(1) = 2 + 0.9 \times \frac{3}{4} \times J_{\pi_0}(1) + 0.9 \times \frac{1}{4} \times J_{\pi_0}(2)$$

$$J_{\pi_0}(2) = 3 + 0.9 \times \frac{1}{4} \times J_{\pi_0}(1) + 0.9 \times \frac{3}{4} \times J_{\pi_0}(2)$$

Solving, $J_{\pi_0}(1) = 24.12$, $J_{\pi_0}(2) = 25.96$

Policy improvement:

$$T_{\pi_1} J_{\pi_0} = T J_{\pi_0}$$

$$(T J_{\pi_0})(1) = \min\{ \underbrace{2 + 0.9\left(\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96\right)}_{\text{action } a},$$

$$\underbrace{0.5 + 0.9\left(\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96\right)}_{\text{action } b} \}$$

$$= \min\{24.12, 23.45\} = 23.45 \leftarrow \text{for action } b$$

$$(T J_{\pi_0})(2) = \min\{ \underbrace{1 + 0.9\left(\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96\right)}_{\text{action a}},$$

$$\underbrace{3 + 0.9\left(\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96\right)}_{\text{action b}} \}$$

$$= \min\{23.12, 25.95\} = 23.12 \quad \leftarrow \text{for action a}$$

$\pi_1(1) = b, \quad \pi_1(2) = a$

Policy evaluation: J_{π_1} ?

$J_{\pi_1}(1) = 7.33, \quad J_{\pi_1}(2) = 7.67$

Policy improvement: $T_{\pi_2} J_{\pi_1} = T J_{\pi_1}$

$\pi_2(1) = b, \quad \pi_2(2) = a$

So, stop & output π_2

Linear programming

Want to solve $J^* = T J^*$

Idea: is to form a linear optimization problem whose solution is J^*

How? We know $\lim_{N\to\infty} T^N J = J^*$ for any J.

Suppose $J \le TJ$

Then, $J \le T^2 J$

$\vdots$

$J \le T^k J$

$\Rightarrow$ $J \le J^* = TJ^*$

$\downarrow$

J^* is the largest solution that satisfies $J \le TJ$

Optimization problem:

constraints: $J \le TJ$

objective: max J

More precisely, $J \le TJ$

$\Updownarrow$

$$J(i) \le g(i,a) + \alpha \sum_{j=1}^{n} P_{ij}(a) J(j)$$

→ for simplicity, assume single stage cost doesn't depend on next state.

So, the LP formulation is:

Variables: $\lambda_1, \ldots, \lambda_n$

Objective: $\max \sum_i \lambda_i$

subject to

$$\lambda_i \le g(i,a) + \alpha \sum_j p_{ij}(a)\lambda_j, \quad \text{for } i=1,\ldots,n, \text{ and } a \in \mathcal{A}(i) \qquad (*)$$

Remark: Assuming $\mathcal{A}(i) = \mathcal{A}$ for $|\mathcal{A}| = q$, we have $n \times q$ constraints in the LP $(*)$

variables $= n$ ← Cardinality of the state space

On problems with a large state space, LP is not practical.

Remark: Can we LP approach for solving SSPs as well.

H.W.: Write down the LP for the 2-state 2-action example used for VI/PI above.