* Infinite Horizon Average Cost Problem

Recall that in this problem we will aim to minimize

$$J_\pi(i) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \mid x_0 = i\right]$$

We will first argue that for most problems of interest the average cost per stage of a policy and the optimal average cost per stage is independent of the initial state.

Consider a stationary policy μ and two states i and j such that under μ, the system will eventually reach j from i with prob 1.

Let $K_{ij}(\mu)$ be the first passage time from i to j under μ.

So, now

$$J_\mu(i) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{k=0}^{K_{ij}(\mu)-1} g(x_k, \mu(x_k))\right] + \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{k=K_{ij}(\mu)}^{N-1} g(x_k, \mu(x_k))\right]$$

If $\mathbb{E}[K_{ij}(\mu)] < \infty$, then by walds lemma

$$J_\mu(i) = 0 + J_\mu(j)$$

As $i, j \in S$ are arbitrary, $J_\mu(i) = J_\mu(j) \quad \forall i, j \in S$ with $\mathbb{E}[K_{ij}(\mu)] < \infty$

If there exists a stationary policy μ, under which j can be reached from i with probability 1 then it is not possible that

$$J^*(i) > J^*(j)$$

Since when starting from i we can play μ and switch to the optimal policy after reaching j

Since when starting from i we can play μ and switch to the optimal policy after reaching j.

So, $$J^*(i) = J^*(j) \quad \forall\, i,j \in S.$$

* Associated Stochastic Shortest Path Problem

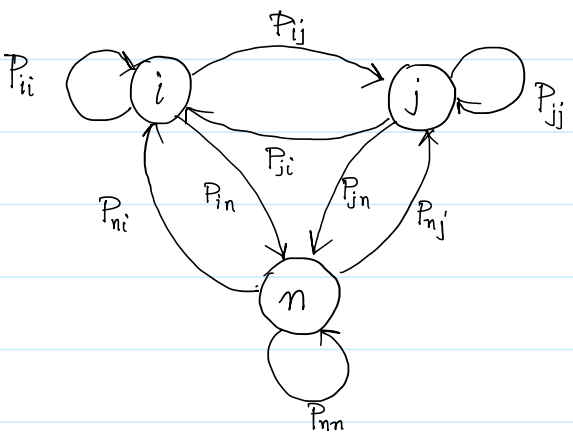
- We will associate a SSPP to the Average cost per stage problem and then we will use the results of SSPP.

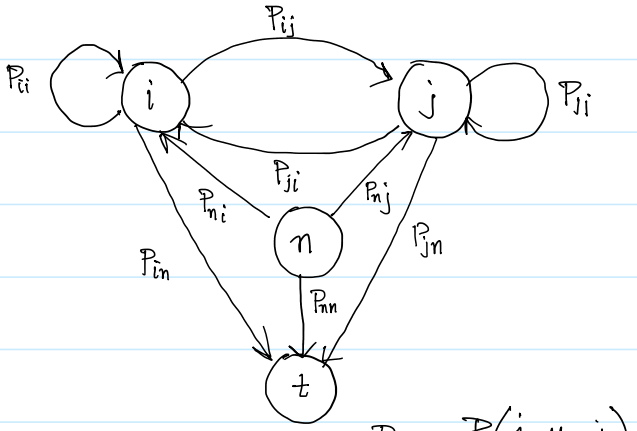
Assumption 1: One of the states, by convention state n, is such that for some integer $m > 0$, and for all initial state and all policies, n is visited with positive probability at least within the first m stages.

* Assumption 1 will make an important connection with SSPP.

- Consider a sequence of generated states. Divide the seq by successive visits to state n.
  - Cycle 1 is includes transition from initial state to first visit to state n.
  - Cycle $k \geq 2$ includes transitions from $(k-1)^{th}$ to the k^{th} visit to state n.

  So each cycle can be viewed as trajectory of SSPP with termination state being n.
- Precisely, the problem is obtained by two changes:
  1. Change in transition probabilities:
     i. $p(i,u,j)$ are unchanged $\forall\, j \neq n$
     ii. $p(i,u,n) = 0$
     iii. Introduce new artificial termination state t and $p(i,u,t) = p(i,u,n)$





$p_{ij} := P(i, u, j)$

2. Change in cost per stage : Now we will argue that if we fix the expected stage cost incurred at state i to be $g(i,u) - \lambda^*$

where λ^* be the optimal average cost per stage, the associated SSPP becomes equivalent to the original average cost per stage problem.

It can be shown that under a stationary policy μ, average cost per stage

$$\lambda_\mu = \frac{C_{nn}(\mu)}{N_{nn}(\mu)}$$

where for fixed μ

$C_{nn}(\mu)$: expected cost starting from n up to first return to n

$N_{nn}(\mu)$: expected number of stages to return to n starting from n.

Since $\lambda^* \leq \lambda_\mu$, $C_{nn}(\mu) - N_{nn}(\mu)\lambda^* \geq 0$.
with equality if μ is optimal.

* $C_{nn}(\mu) - N_{nn}(\mu)\lambda^*$ is the total expected cost of the assoc. SSPP with stage cost $g(i,u) - \lambda^*$.

Let $h^*(i)$ be the optimal cost of this SSPP when starting from state $i \in [n]$. Then the Bellman' equation is

$$\overset{n-1}{\longleftarrow}$$

from state $i \in [n]$. Then the Bellman' equation is

$$h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) - \lambda^* + \sum_{j=1}^{n-1} p(i,u,j)\, h^*(j) \right\}, \quad i \in [n]$$

If μ^* is an optimal policy then

$$C_{nn}(\mu^*) - N_{nn}(\mu^*)\, \lambda^* = 0$$

and $$h^*(n) = C_{nn}(\mu^*) - N_{nn}(\mu^*)\, \lambda^* = 0$$

By this equation we can write Bellmans' equation

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j) \cdot h^*(j) \right\}, \quad i \in [n]$$

with $h^*(n) = 0$.

* Proposition 1 : Under assumption 1 the following hold for the average cost per stage problem:

a. The optimal average cost λ^* is the same for all initial states and together with some vector $h^* = [h^*(1), \ldots, h^*(n)]$ satisfies Bellmans equation

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h^*(j) \right\}, \quad i \in [n] \qquad (1)$$

Furthermore, if $\mu(i)$ attains the minimum in the above equation for all i, the stationary policy μ is optimal. In addition, out of all vectors h^* satisfying (1), there is a unique vector with $h^*(n) = 0$.

b. If a scalar λ and a vector $h = [h(1) \ldots h(n)]^T$ satisfy (1), then λ is the average optimal cost per stage for each initial state.

c. Given a stationary policy μ with corresponding average cost per stage λ_μ, there is a unique vector $h_\mu = [h_\mu(1), \ldots, h_\mu(n)]^T$ such that $h_\mu(n) = 0$ and

$$\lambda_\mu + h_\mu(i) = g(i, \mu(i)) + \sum^{n} p(i, \mu(i), i)\, h_\mu(i) \qquad i \in [n]$$

$$\lambda_\mu + h_\mu(i) = g(i,\mu(i)) + \sum_{j=1}^{n} P(i,\mu(i),j)\, h_\mu(j), \quad j \in [n] \qquad -(2)$$

Lecture 15

* **Proposition 1**: Under assumption 1 the following hold for the average cost per stage problem:

a. The optimal average cost λ^* is the same for all initial states and together with some vector $h^* = [h^*(1), \dots, h^*(n)]$ satisfies Bellmans equation

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h^*(j) \right\}, \quad i \in [n] \qquad -(1)$$

Furthermore, if $\mu(i)$ attains the minimum in the above equation for all i, the stationary policy μ is optimal. In addition, out of all vectors h^* satisfying (1), there is a unique vector with $h^*(n) = 0$.

b. If a scalar λ and a vector $h = [h(1) \dots h(n)]^T$ satisfy (1), then λ is the average optimal cost per stage for each initial state.

c. Given a stationary policy μ with corresponding average cost per stage λ_μ, there is a unique vector $h_\mu = [h_\mu(1), \dots, h_\mu(n)]^T$ such that $h_\mu(n) = 0$ and

$$\lambda_\mu + h_\mu(i) = g(i,\mu(i)) + \sum_{j=1}^{n} p(i,\mu(i),j)\, h_\mu(j), \quad j \in [n]. \qquad -(2)$$

Proof:

a. Let us denote

$$\bar{\lambda} := \min_{\mu} \frac{C_{nn}(\mu)}{N_{nn}(\mu)} \qquad -(i)$$

Note that $C_{nn}(\mu)$ and $N_{nn}(\mu)$ are finite $\forall\, \mu$ by assumption 1. Also we have that

$$C_{nn}(\mu) - N_{nn}(\mu)\bar{\lambda} \geq 0 \qquad -(ii)$$

with equality if μ attains minimum in (i). By a previous proposition, the costs $h^*(1), \dots, h^*(n)$ solve uniquely corresponding

with equality if μ attains minimum in (i). By a previous proposition, the costs $h^*(1), \dots, h^*(n)$ solve uniquely corresponding Bellmans equation

$$h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) - \bar{\lambda} + \sum_{j=1}^{n-1} P(i,u,j)\, h^*(j) \right\}, \quad \text{—(iii)}$$

Since the transition probability $P(i,u,n) = 0 \ \forall i,u$ in the assoc. SSPP. An optimal stationary policy should minimize

$$C_{nn}(\mu) - N_{nn}(\mu)\bar{\lambda}$$

to zero, so we have that $h^*(n) = 0$.

Thus,

$$h^*(i) + \bar{\lambda} = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} P(i,u,j) \cdot h^*(j) \right\}, \quad i \in [n] \quad \text{—(iv)}$$

We will show that this relation implies $\bar{\lambda} = \lambda^*$.

Let $\pi = \{\mu_0, \mu_1, \dots\}$ be any admissible policy, let N be a positive integer, and for all $k = 0, \dots, N-1$, define $J_k(i)$ using the following recursion

$$J_0(i) = h^*(i), \quad i \in [n]$$

$$J_{k+1}(i) = g\left(i, \mu_{N-k-1}(i)\right) + \sum_{j=1}^{n} P\left(i, \mu_{N-k-1}(i), j\right) \cdot J_k(j), \quad i \in [n]$$

Note that $J_N(i)$ is the N-stage cost of π when the starting state is i and the terminal cost function is h^*. By (iv)

$$\bar{\lambda} + J_0(i) \leq J_1(i), \quad i \in [n]$$

Using this relation, we have

$$g\left(i, \mu_{N-2}(i)\right) + \bar{\lambda} + \sum_{j=1}^{n} P\left(i, \mu_{N-2}(i), j\right) J_0(i)$$

$$\leq g\left(i, \mu_{N-2}(i)\right) + \sum_{j=1}^{n} P\left(i, \mu_{N-2}(i), j\right) J_1(j)$$

and as $J_0(i) = h^*(i)$

$$2\bar{\lambda} + h^*(i) \leq J_2(i)$$

Repeating this argument we obtain

$$k\bar{\lambda} + h^*(i) \leq J_k(i), \quad k = 0, \dots, N, \quad i \in [n]$$

'Repeating this argument we obtain

$$k\bar{\lambda} + h^*(i) \leq J_k(i) \quad , \quad k = 0, \ldots, N, \quad i \in [n]$$

and in particular for $k = N$,

$$\bar{\lambda} + \frac{h^*(i)}{N} \leq \frac{J_N(i)}{N}$$

Taking limit as $N \to \infty$,

$$\bar{\lambda} \leq J_\pi(i)$$

for all admissible policy π, with equality if π is a stationary policy μ such that $\mu(i)$ attains minimum in (iv) $\forall\, i, k$.

It follows that

$$\bar{\lambda} = \min_\pi J_\pi(i) = \lambda^* \quad , \quad i \in [n]$$

Equation (iv) with $h^*(n) = 0$ is equivalent to equation (iii). Since the solution of (iii) is unique, the same is true for (iv) with $h^*(n) = 0$.

* b and c are home work! "$\blacksquare$"

* Value Iteration :

The natural one : Initialize iterate J_0 arbitrarily. For $k = 0, 1, 2, \ldots$ Compute J_{k+1} iteratively as follows :

$$J_{k+1}(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} P(i,u,j)\, J_k(j) \right\} , \quad i \in [n]$$

It is natural to expect that $\frac{J_k(i)}{k}$ should converge to λ^* as $k \to \infty$, i.e.,

$$\lim_{k \to \infty} \frac{J_k(i)}{k} = \lambda^* \quad \forall\, i.$$

To show this, let us define the recursion

$$J^*_{k+1}(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} P(i,u,j)\, J^*_k(j) \right\} , \quad i \in [n]$$

with $J^*_0(i) = h^*(i)$, $\forall\, i \in [n]$.

By induction it is easy to see that

$$J^*_k(i) = k\lambda^* + h^*(i) \quad , \quad \forall\, i \in [n]$$

It can also be shown that

$$|J_k(i) - J_k^*(i)| \leq \max_{j\in[n]} |J_0(j) - h^*(i)|, \quad i\in[n].$$

So, $|J_k(i) - k\lambda^*| \leq \max_{j\in[n]} |J_0(j) + h^*(j)| + \max_{j\in[n]} |h^*(j)|, \quad i\in[n]$

So, $\frac{J_k(i)}{k}$ converges to λ^*

* But see that λ^* is finite hence as $k\to\infty$, $J_k(i)\to \underset{-\infty}{\infty}$ $\forall\, i\in[n]$. So there is a computational problem!
* Also we have no information about the differential cost vector h^*.

* Relative Value Iteration :

- To overcome the problems in previous section we subtract a constant (cleverly) to get h^* as well as λ^*.

Consider the algorithm :

$$h_k(i) = J_k(i) - J_k(s), \quad i = 1,\dots,n, \quad s\in[n] \text{ be fixed.}$$

Then

$$\begin{aligned} h_{k+1}(i) &= J_{k+1}(i) - J_{k+1}(s) \\ &= \min_{u\in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, J_k(j) \right\} \\ &\quad - \min_{u\in U(s)} \left\{ g(s,u) + \sum_{j=1}^{n} p(s,u,j)\, J_k(j) \right\} \\ &= \min_{u\in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j) \left(J_k(j) - J_k(s)\right) \right\} \\ &\quad - \min_{u\in U(s)} \left\{ g(s,u) + \sum_{j=1}^{n} p(s,u,j) \left(J_k(j) - J_k(s)\right) \right\} \\ &= \min_{u\in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h_k(j) \right\} \\ &\quad - \min_{u\in U(s)} \left\{ g(s,u) + \sum_{j=1}^{n} p(s,u,j)\, h_k(j) \right\} \end{aligned}$$

It can be seen that if relative value iteration converges to some vector h, then we have

to some vector h, then we have

$$\lambda + h(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h(j) \right\}$$

with $h(s) = 0$ and

$$\lambda = \min_{u \in U(s)} \left\{ g(s,u) + \sum_{j=1}^{n} p(s,u,j) \cdot h(j) \right\}$$

Lecture 16

* Policy Iteration :

Algorithm

Step 1 : (Initialization) Guess an initial policy μ^0

Step 2 : (Policy evaluation) Given the stationary policy μ^k, Compute corresponding average and differential costs λ^k and $h^k(i)$ satisfying

$$\lambda^k + h^k(i) = g(i, \mu^k(i)) + \sum_{j=1}^{n} p(i, \mu^k(i), j)\, h^k(j) \quad \forall\, i \in [n]$$

$$h^k(n) = 0$$

Step 3 : (Policy improvement) Obtain a new stationary policy μ^{k+1} satisfying, where $\forall\, i$, $\mu^{k+1}(i)$ is such that

$$g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p(i, \mu^{k+1}(i), j)\, h^k(j) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h^k(j) \right\}$$

If $\lambda^{k+1} = \lambda^k$ and $h^{k+1}(i) = h^k(i)\ \forall\, i$, then stop else return to step 2 and repeat.

* The algorithm is based only the following proposition :

Proposition 2 : Under assumption 1, in the policy iteration algo, for each k we either have

$$\lambda^{k+1} < \lambda^k$$

or else we have

$$\lambda^{k+1} = \lambda^k, \quad h^{k+1}(i) \leq h^k(i), \quad i \in [n]$$

Furthermore, the algorithm terminates and the policies μ^k and μ^{k+1} obtained upon termination are optimal.

Proof: To simplify notation, denote $\mu^k = \mu$, $\mu^{k+1} = \bar{\mu}$, $\lambda^k = \lambda$, $\lambda^{k+1} = \bar{\lambda}$, $h^k = h$, $h^{k+1} = \bar{h}$. Define for $N = 1, 2, \ldots$

$$h_N(i) = g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_{N-1}(j), \quad i \in [n]$$

where $h_0(i) = h(i)$

Note that $h_N(i)$ is the N-stage cost of policy $\bar{\mu}$ starting from i when the termination cost function is h. Thus we have

$$\bar{\lambda} = J_{\bar{\mu}}(i) = \lim_{N\to\infty} \frac{1}{N} h_N(i) \quad \forall\, i \in [n]$$

Since the contribution of the terminal cost function vanishes as $N \to \infty$. By definition of $\bar{\mu}$ and by Prop 1(c), we have $\forall\, i$

$$\begin{aligned} h_1(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_0(j) \\ &\leq g(i, \mu(i)) + \sum_{j=1}^{n} P(i, \mu(i), j)\, h_0(j) \\ &= \lambda + h_0(i) \end{aligned}$$

From the above equation we also obtain

$$\begin{aligned} h_2(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\, h_1(j) \\ &\leq g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} P(i, \bar{\mu}(i), j)\,(\lambda + h_0(j)) \\ &\leq \lambda + g(i, \mu(i)) + \sum_{j=1}^{n} P(i, \mu(i), j)\, h_0(i) \\ &= 2\lambda + h_0(i) \end{aligned}$$

and proceeding in this way, $\forall\, i$ and N we have

$$\frac{1}{N} h_N(i) \leq \lambda + \frac{1}{N} h_0(i)$$

and by taking limit as $N \to \infty$, we get $\bar{\lambda} \leq \lambda$.

If $\bar{\lambda} = \lambda$ then μ^{k+1} is a policy improvement step for assoc. SSPP with cost per stage

$$g(i, u) - \lambda$$

Furthermore, $h(i)$ and $\bar{h}(i)$ are the optimal costs starting from

Furthermore, $h(i)$ and $\bar{h}(i)$ are the optimal costs starting from i and corresponding to μ and $\bar{\mu}$, respectively, is associated SSPP. Thus by a previous proposition

$$\bar{h}(i) \leq h(i) \quad \forall\, i.$$

Let us now show that when the algorithm terminates with $\bar{\lambda} = \lambda$ and $\bar{h}(i) = h(i)$ $\forall\, i \in [n]$, μ and $\bar{\mu}$ are optimal.

We have then $\forall\, i$

$$\begin{aligned}
\lambda + h(i) = \bar{\lambda} + \bar{h}(i) &= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} p(i, \bar{\mu}(i), j)\, \bar{h}(j) \\
&= g(i, \bar{\mu}(i)) + \sum_{j=1}^{n} p(i, \bar{\mu}(i), j)\, h(j) \\
&= \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p(i,u,j)\, h(j) \right\}
\end{aligned}$$

Thus λ and h satisfy Bellman's equation. By prop 1, λ must be equal to optimal average cost.

Furthermore, $\bar{\mu}(i)$ attains minimum of rhs. of the Bellman's equation $\forall\, i$. So $\bar{\mu}$ is optimal. ▨