## 1.1. Minimization of total cost - Introduction.

the stationary discrete-time dynamic system:

* $x_{k+1} = f(x_k, u_k, w_k) \quad k = 0, 1, 2, \cdots$

  $x_k \in S$, $u_k \in C$, $w_k \in D$. $D$ is a countable set.

* $u_k \in U(x_k) \subseteq C$, $w_k \sim P(\cdot | x_k, u_k)$ ind of $k$.

* $p(w_k | x_k, u_k)$ is the probability $w_k$ occur, $w_k$ ind of $w_0, \cdots, w_{k-1}$

* Given $x_0$, we want to find a policy $\pi = \{\mu_0, \mu_1, \cdots\}$

  $\mu_k : S \to C$, $\mu_k(x_k) \in U(x_k)$, $\forall x_k \in S$, $k = 0, 1, 2, \cdots$

  that minimize the cost function.

$$J_\pi(x_0) = \lim_{N \to \infty} \underset{\substack{w_k \\ k=0,\cdots,N-1}}{E} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

$\alpha \in (0, 1)$ : discount factors.

* $\Pi$ : the set of all admissible policies $\pi$.
* Optimal cost function : $J^*(x) = \min_{\pi \in \Pi} J_\pi(x)$. $\forall x \in S$
* An optimal policy for a given state $x$ is one that attains the optimal cost $J^*(x)$. This policy depends on $x$, but sometimes ind. of $x$.
* Stationary : $\pi$ is stationary if $\pi = \{\mu, \mu, \cdots\}$ which is refered to as the stationary policy.
* $\mu$ is optimal if $J_\mu(x) = J^*(x)$.

### 1.1.1. The Finite - Horizon DP Algorithm.

Consider any admissible policy $\pi = \{\mu_0, \mu_1, \cdots\}$. any positive integer, any function $J : S \to \mathbb{R}$.

Suppose we accumulate the costs of the first $N$ stage, and we add to them some terminal cost of the form $\alpha^N J(x_N)$, where $J$ is some function for a total expected cost.

$$\underset{w_k}{E} \left\{ \alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

$k=0,\cdots N-1$

The minimum cost can be calculated by starting with $\alpha^N J(x)$ and by carrying out $N$ iterations of DP algorithm:

$$J_{N-k}(x) = \min_{u\in U(x)} E\{\alpha^{N-k} g(x,u,w) + J_{N-k+1}(f(x,u,w))\}.$$

with initial condition: $J_N(x) = \alpha^N J(x)$.

$J_{N-k}(x)$ : optimal cost of the last $k$ stages starting from state $x$.

$J_0(x)$ : the optimal $N$-stage cost.

More conveniently, we denote $V_k(x) \triangleq \frac{J_{N-k}(x)}{\alpha^{N-k}}$.

then $V_N(x) = J_0(x)$. rewrite DP Algorithm as.

$$V_{k+1}(x) = \min_{u\in U(x)} E\{g(x,u,w) + \alpha V_k(f(x,u,w))\}, \quad k=0,1,\cdots N-1.$$

$$V_0(x) = J(x).$$

## 1.1.2. Shorthand Notation and Monotonicity.

$\forall J: S\to \mathbb{R}$. consider DP mapping to $J$

$$(TJ)(x) = \min_{u\in U(x)} E_w\{g(x,u,w) + \alpha J(f(x,u,w))\}.$$

$TJ$ : optimal cost function for the one-stage problem that has stage cost $g$ and terminal cost $\alpha J$

$\forall J: S\to \mathbb{R}$. control $\mu: S\to C$

$$(T_\mu J)(x) = E\{g(x,\mu(x),w) + \alpha J(f(x,\mu(x),w))\}.$$

cost function associated with $\mu$.

$T^k$ : composition of the mapping $T$ with itself $k$ times.

$$(T^k J)(x) = (T(T^{k-1}J))(x)$$

$$(T^0 J)(x) = J(x). \quad x\in S.$$

Similarly : $(T_\mu^k J)(x) = (T_\mu(T_\mu^{k-1}J))(x)$

$(T_\mu^0 J)(x) = J(x)$.

$(T^k J)(x)$: optimal cost for the $k$-stage, $\alpha$-discounted problem with initial state $x$, cost per stage $g$ and terminal cost function $\alpha^k J$.

Similarly $(T_\mu^k J)(x)$: the cost of a policy $\{\mu, \mu, \cdots\}$ for the same problem.

Consider $k$-stage policy $\pi = \{\mu_0, \cdots, \mu_{k-1}\}$.

$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{k-1}} J)(x)$ is defined recursively.

$$(T_{\mu_i} T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J)(x) = (T_{\mu_i}(T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J))(x)$$

represents the cost under $\pi$ with initial state $x$, cost per stage $g$, terminal cost function $\alpha^k J$.

*Lemma 1.1.1 (Monotonicity Lemma).

For any functions $J: S \to \mathbb{R}$ and $J': S \to \mathbb{R}$

$$J(x) \le J'(x) \quad \forall x \in S.$$

for any stationary policy $\mu: S \to C$. we have

$$(T^k J)(x) \le (T^k J')(x). \quad \forall x \in S. \ k = 1, 2, \cdots.$$

$$(T_\mu^k J)(x) \le (T_\mu^k J')(x). \quad \forall x \in S. \ k = 1, 2, \cdots$$

Define unit function $e: S \to \mathbb{R}$. $e(x) = 1$. $\forall x \in S$.

$\forall$ scalar $r$, $(T(J + re))(x) = (TJ)(x) + \alpha r$

$$(T_\mu(J + re))(x) = (T_\mu J)(x) + \alpha r$$

* Lemma 1.1.2. $\forall k$. $J: S \to \mathbb{R}$. stationary policy $\mu$ and scalar $r$,

$$(T^k(J + re))(x) = (T^k J)(x) + \alpha^k r$$

$$(T_\mu^k(J + re))(x) = (T_\mu^k J)(x) + \alpha^k r$$

## 1.1.3. A Preview of Infinite Horizon Results.

a) convergence of the DP Algorithm.

$J_0$ denote zero function: $J_0(x) = 0 \quad \forall x \in S$.

Since the infinite horizon cost of a policy is the limit of its $k$-stage cost as $k \to \infty$. Hopefully,

$J^*(x) = \lim_{k\to\infty} (T^k J_0)(x) \quad x \in S$.

Also, $\alpha < 1$ and $J$ bounded. $\alpha^k J$ diminishes with $k$, then if $\alpha < 1$. $J^*(x) = \lim_{k\to\infty} (T^k J)(x)$. value iteration starts with any $J$.

b) Bellman's Equation.

Since by definition, we have for all $x \in S$.

$(T^{k+1} J_0)(x) = \min_{u \in U(x)} E_w \{ g(x,u,w) + \alpha (T^k J_0)(f(x,u,w)) \}$.

It is reasonable to speculate that if $\lim_{k\to\infty} T^k J_0 = J$, take limits on both sides of above equation:

$J^*(x) = \min_{u \in U(x)} E_w \{ g(x,u,w) + \alpha J^*(f(x,u,w)) \}$.

$\Rightarrow J^* = TJ^*$: $J^*$ is a fixed point of $T$.

↗ Bellman's Equation

c) Characterization of Optimal Stationary Police.

view Bellman's Equation $J^* = TJ^*$ as DP Algorithm

if $\mu(x)$ attains minimum on $TJ^*$ for all $x$, then the stationary policy is optimal.

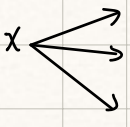
## 1.1.4. Randomized and history dependent policies.

At each time $k$. $\mu$ is applied to $x_k$. $\mu$ is Markov because

$\mu$ is independent of $x_0, \cdots, x_{k-1}$.

denote $h_k = \{x_0, \mu_0, x_1, \mu_1, \cdots, x_k\}$.

randomized policies :




| Prob |
|---|
| $p_1$ |
| $p_2$ |
| $p_3$. |

* Proposition 1.1.1 ( Adequacy of Markov Policies).

Assume Control space is countable. Initial state distribution takes value over a countable set

The probability distribution of each pair $(x_k, u_k)$ and the expected cost of each stage corresponding to a randomized – history-dependent policy can also be obtained with a randomized Markov policy.

## 1.2. Discounted problems with bounded cost per stage.

Assumption D. (not very restrictive)

The cost per stage $g$ satisfies $|g(x,u,w)| \le M$, $\forall (x,u,w) \in S \times C \times D$. $M$ is scalar. $0 < \alpha < 1$

If. S. C. D are finite, then $g$ is always bounded.

DP Algorithm convergences to the optimal cost function for an arbitrary bounded starting function J.

$\lim_{k\to\infty} E\{\sum_{k=N}^{k} \alpha^k g(x_k, \mu_k(x_k), w_k)\} \to 0$ as $N \to \infty$.

* Proposition 1.2.1 : (Convergence of DP Algorithm).

For any bounded function $J: S \to \mathbb{R}$. the optimal cost function satisfies :

$$J^*(x) = \lim_{N\to\infty} (T^N J)(x). \qquad \forall x \in S.$$

proof. $\forall$ positive integer $k$, $x_0 \in S$ ( initial states). policy $\pi = \{\mu_0, \mu_1, \cdots\}$. we break down the cost $J_\pi(x_0)$ into

the portions incurred over the first k stage and over the remaining stage.

$$J_\pi(x_0) = \lim_{N\to\infty} E\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k)\}$$
$$= \lim_{N\to\infty} E\{\sum_{k=0}^{K-1} \alpha^k g(x_k, \mu_k(x_k), w_k)\} + \lim_{N\to\infty} E\{\sum_{k=K}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k)\}.$$

By Assumption D, $|g(x_k, \mu_k(x_k), w_k)| \le M$.

$$\left|\lim_{N\to\infty} E\{\sum_{k=K}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k)\}\right| \le M \sum_{k=K}^{\infty} \alpha^k = \frac{\alpha^K M}{1-\alpha}$$

$$J_\pi(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \max_{x\in S} |J(x)|$$
$$\le E\{\alpha^K J(x_K) + \sum_{k=0}^{K-1} \alpha^k g(x_k, \mu_k(x_k), w_k)\}$$
$$\le J_\pi(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \max_{x\in S} |J(x)|$$

Take minimum over $\pi$, we obtain for all $x_0$, $K$.

$$J^*(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \max_{x\in S} |J(x)| \le (T^K J)(x_0)$$
$$\le J^*(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \max_{x\in S} |J(x)|.$$

let $K\to\infty \Rightarrow J^*(x_0) = (T^K J)(x_0)$.

Gives us a way to compute $J^*$ approximately.

For a given stationary policy $\mu$, we can compute its approximation

* Corollary 1.2.1.1. For every stationary policy $\mu$, the associated cost function satisfies

$$J_\mu(x) = \lim_{N\to\infty} (T_\mu^N J)(x). \quad \forall\, x\in S.$$

$J^*$ is the unique solution to the Bellman's equation:

* Proposition 1.2.2. the optimal cost Function $J^*$ satisfies

$$J^*(x) = \min_{u\in U(x)} E_w\{g(x,u,w) + \alpha J^*(f(x,u,w))\}. \text{ or } J^* = TJ^*$$

Furthermore, $J^*$ is the unique solution of this equation within the class of bounded functions.

proof. $\forall\, x\in S$, $N$

$$J^*(x) - \frac{\alpha^N M}{1-\alpha} \le (T^N J_0)(x) \le J^*(x) + \frac{\alpha^N M}{1-\alpha}$$

$J_0$ is the zero function,

apply $T$ to this relation and using the Monotonicity Lemma 1.1.1 and 1.1.2, we obtain $\forall x \in S$, $N$.

$$(TJ^*)(x) - \frac{\alpha^{N+1} M}{1-\alpha} \le (T^{N+1} J_0)(x) \le (TJ^*)(x) + \frac{\alpha^{N+1} M}{1-\alpha}$$

Take $N \to \infty$, and $\lim_{N\to\infty} (T^{N+1} J_0)(x) = J^*(x)$

$\Rightarrow (TJ^*)(x) \le J^*(x) \le (TJ^*)(x) \Rightarrow J^* = TJ^*$

To show uniqueness, if $J$ is bounded and $J = TJ$, then

$J = \lim_{N\to\infty} T^N J$. by Prop 1.2.1 we have $J = J^*$.

* Corollary 1.2.2.1. For every stationary policy $\mu$.

$$J_\mu(x) = E_w\{g(x, \mu(x), w) + \alpha J_\mu(f(x, \mu(x), w))\}$$

or $J_\mu(x) = T_\mu J_\mu$.

Furthermore, $J_\mu$ is the unique solution of this equation within the class of bounded functions.

* Proposition 1.2.3. (Necessary and Sufficient Condition for Optimality).

A stationary policy $\mu$ is optimal if and only if $\mu(x)$ attains the minimum in Bellman's Equation ($J^* = TJ^*$) for each $x \in S$

i.e. $TJ^* = T_\mu J^*$

proof: $\Rightarrow$ If $TJ^* = T_\mu J^*$. By Bellman's Equation

$J^* = TJ^* = T_\mu J^*$, which is unique solution of Bellman's equation : $J^* = J_\mu$. and $\mu$ is optimal

$\Leftarrow$ If stationary policy $\mu$ is optimal, $J^* = J_\mu$. By corollary 1.2.2.1,

$J^* = T_\mu J^*$

Combining this with Bellman's Equation $J^* = TJ^*$

$\Rightarrow TJ^* = T_\mu J^*$

* Proposition 1.2.4. For any two bounded functions

$$J: S \to \mathbb{R}, \quad J': S \to \mathbb{R} \qquad \forall k = 0, 1, \cdots$$

$$\max_{x \in S} |(T^k J)(x) - (T^k J')(x)| \leq \alpha^k \max_{x \in S} |J(x) - J'(x)|$$

proof. Denote $c = \max_{x \in S} |J(x) - J'(x)|$

$$\Rightarrow \quad J(x) - c \leq J'(x) \leq J(x) + c$$

Apply $T^k$ in this relation and use Monotonicity lemma 1.1.1, 1.1.2

$$(T^k J)(x) - \alpha^k c \leq (T^k J')(x) \leq (T^k J)(x) + \alpha^k c \qquad x \in S$$

$$\Rightarrow \max_{x \in S} |(T^k J)(x) - (T^k J')(x)| \leq \alpha^k c \qquad x \in S$$

* Corollary 1.2.4.1. $J, J': S \to \mathbb{R}$. $\forall$ stationary policy $\mu$.

$$\max_{x \in S} |(T_\mu^k J)(x) - (T_\mu^k J')(x)| \leq \alpha^k \max_{x \in S} |J(x) - J'(x)|, \quad k = 0, 1, \cdots$$

## 1.3. Finite-State System — Computational Methods

the state, control and disturbance spaces are finite sets

DP ⇒ control of a finite-state Markov chain

Notations:

State space $S = \{1, 2, \cdots, n\}$

transition probability: $p_{ij}(u) = P(x_{k+1} = j \mid x_k = i, u_k = u) \qquad i, j \in S, \ u \in U(i)$

update function: $x_{k+1} = f(x_k, u_k, w_k)$

disturbance probability: $P(\cdot \mid x, u)$.

$$p_{ij}(u) = P(w_{ij}(u) \mid i, u)$$

where $w_{ij}(u) \triangleq \{w \in D \mid f(i, u, w) = j\}$.

To simplify, assume cost per stage does not depend on $w$

$$(g(i, u, w) = g(i, u))$$

define $g(i, u) = \sum_{j=1}^{n} p_{ij}(u)\, \tilde{g}(i, u, j)$

$\tilde{g}(i, u, j)$: the cost of starting from $i$, use control $u$, transite to $j$.

$g(i, u)$: expected cost at state $i$, use control $u$.

$$TJ(i) = \min_{u \in U(i)} \left[ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right]$$

$$T_\mu J(i) = g(i, \mu(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j)$$

$TJ$, $T_\mu J$ can be represented by the $n$-dimensional vectors.

$$J = \begin{bmatrix} J(1) \\ \vdots \\ J(n) \end{bmatrix} \qquad TJ = \begin{bmatrix} (TJ)(1) \\ \vdots \\ (TJ)(n) \end{bmatrix} \qquad T_\mu J = \begin{bmatrix} (T_\mu J)(1) \\ \vdots \\ (T_\mu J)(n) \end{bmatrix}$$

For a stationary policy $\mu$, we denote $P_\mu$ as transition probability matrix

$$P_\mu = \begin{bmatrix} p_{11}(\mu(1)) & \cdots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \cdots & p_{nn}(\mu(n)) \end{bmatrix}$$

$g_\mu$: the cost vector $g_\mu = \begin{bmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{bmatrix}$

$$\Rightarrow J_\mu = T_\mu J = g_\mu + \alpha P_\mu J_\mu$$

$\forall$ stationary policy $\mu$, $J_\mu$ can be computed

$$(I - \alpha P_\mu) J_\mu = g_\mu \Rightarrow J_\mu = (I - \alpha P_\mu)^{-1} g_\mu$$

$(I - \alpha P_\mu)^{-1}$ exists : $\alpha < 1$. $P_\mu$'s eigenvalue $\leq 1 \Rightarrow \alpha P_\mu$'s eigenvalue $< 1$

$\Rightarrow I - \alpha P_\mu$ is invertible

### 1.3.1. Value Iteration and Error Bounds.

$\forall$ $n$-dimensional vector $J$, compute $J, TJ, \cdots$ successively. By prop 1.2.1.

$$\lim_{k \to \infty} (T^k J)(i) = J^*(i) \qquad \forall i$$

Furthermore, by Prop 1.2.4. error sequence $|(T^k J)(i) - J^*(i)|$ is bounded by $\alpha^k C$. $C$ is a constant.

this method is called. value iteration.

$$J_\mu(i) = g(i, \mu(i)) + \sum_{k=1}^{\infty} \alpha^k E\{ g(x_k, \mu(x_k)) \mid x_0 = i \}.$$

$$\Rightarrow g_\mu + \left(\frac{\alpha \underline{\beta}}{1-\alpha}\right) e \leq J_\mu \leq g_\mu + \left(\frac{\alpha \bar{\beta}}{1-\alpha}\right) e.$$

where $e = (1, 1, \cdots, 1)'$, $\underline{\beta} = \min_i g(i, \mu(i))$, $\bar{\beta} = \max_i g(i, \mu(i))$

$$\Rightarrow \left(\frac{\underline{\alpha}}{1-\alpha}\right)e \le g_\mu + \left(\frac{\alpha\underline{\alpha}}{1-\alpha}\right)e \le J_\mu \le g_\mu + \left(\frac{\alpha\bar{\alpha}}{1-\alpha}\right)e \le \left(\frac{\bar{\alpha}}{1-\alpha}\right)e$$

suppose we have a vector $J$, compute

$$T_\mu J = g_\mu + \alpha P_\mu J \quad (1)$$

subtracting this equation from the relation

$$J_\mu = g_\mu + \alpha P_\mu J_\mu \quad (2)$$

$(1),(2) \Rightarrow \quad J_\mu - J = T_\mu J - J + \alpha P_\mu (J_\mu - J).$

variational form of the equation $J_\mu = T_\mu J_\mu$.

replace $J_\mu$ with $J_\mu - J$, $g_\mu$ with $T_\mu J - J$, we have

$$\left(\frac{\gamma}{1-\alpha}\right)e \le T_\mu J - J + \left(\frac{\alpha\gamma}{1-\alpha}\right)e$$

$$\le J_\mu - J \le T_\mu J - J + \left(\frac{\alpha\bar{\gamma}}{1-\alpha}\right)e \le \left(\frac{\bar{\gamma}}{1-\alpha}\right)e$$

where $\gamma = \min_i [(T_\mu J)(i) - J(i)]$, $\bar{\gamma} = \max_i [(T_\mu J)(i) - J(i)]$

$$\Leftrightarrow \quad J + \frac{\underline{c}}{\alpha}e \le T_\mu J + \underline{c}e \le J_\mu \le T_\mu J + \bar{c}e \le J + \frac{\bar{c}}{\alpha}e$$

where $\underline{c} = \frac{\alpha\gamma}{1-\alpha}$, $\bar{c} = \frac{\alpha\bar{\gamma}}{1-\alpha}$

* Proposition 1.3.1. $\forall J$, state $i$ and $k$, we have.

$$(T^k J)(i) + \underline{c}_k \le (T^{k+1}J)(i) + \underline{c}_{k+1}$$

$$\le J^*(i)$$

$$\le (T^{k+1}J)(i) + \bar{c}_{k+1} \le (T^k J)(i) + \bar{c}_k$$

where $\underline{c}_k = \frac{\alpha}{1-\alpha} \min_{i=1,\dots,n} [(T^k J)(i) - (T^{k-1}J)(i)]$

$\bar{c}_k = \frac{\alpha}{1-\alpha} \max_{i=1,\dots,n} [(T^k J)(i) - (T^{k-1}J)(i)]$

* Termination Issue — Optimality of the obtained Policy.

$\forall J$, if we compute $TJ$ and policy $\mu$ attaining the minimum in the calculation of $TJ$ ($T_\mu J = TJ$), then we can obtain the bound on the suboptimality of $\mu$.

$$\max_i [J_\mu(i) - J^*(i)] \le \frac{\alpha}{1-\alpha} (\max_i [(TJ)(i) - J(i)] - \min_i [(TJ)(i) - J(i)]) \quad (3)$$

By Prop 1.3.1, $k=1$ $\quad \underline{c}_1 \le J^*(i) - (TJ)(i) \le \bar{c}_1 \quad (4)$

replacing $T$ with $T_\mu$:

$$\underline{c}_1 \le J_\mu(i) - (T_\mu J)(i) = J_\mu(i) - TJ(i) \le \bar{c}_1 \quad (5).$$

subtracting the above two equations : (4) − (5) ⇒ (3)

In practice, we terminate as $\bar{c}_k - \underline{c}_k$ is sufficiently small.

One can take the final estimate of $J^*$ the median

$$\tilde{J}_k = T^k J + \left(\frac{\bar{c}_k - \underline{c}_k}{2}\right) e$$

or the average :

$$\hat{J}_k = T^k J + \frac{\alpha}{n(1-\alpha)} \sum_{i=1}^{n} \left( (T^k J)(i) - (T^{k-1} J)(i) \right) e$$

Because for eq(3). $\exists\ \bar{\varepsilon} > 0$, s.t. if $\mu$ satisfy

$$\max_i \left[ J_\mu(i) - J^*(i) \right] < \bar{\varepsilon},$$

then $\mu$ is optimal.

let $\bar{k}$ be such that for all $k > \bar{k}$, we have

$$\frac{\alpha}{1-\alpha} \left( \max_i \left[ (T^k J)(i) - (T^{k-1} J)(i) \right] - \min_i \left[ (T^k J)(i) - (T^{k-1} J)(i) \right] \right) < \bar{\varepsilon}$$

then eq(3) implies that $\forall\ k \geq \bar{k}$, the stationary policy that attains the minimum in the $k$th value iteration is optimal.

### 1.3.3. Policy Iteration.

The policy iteration generates a sequence of stationary polices, each with improved cost over the preceding one. Given stationary policy $\mu$ and cost function $J_\mu$, an improved $\{\bar{\mu}, \bar{\mu}, \cdots\}$ is computed by minimization in the DP equation corresponding to $J_\mu$ : $T_{\bar{\mu}} J_\mu = T J_\mu$, and the process is repeated.

* Proposition 1.3.4. Let $\mu$ and $\bar{\mu}$ be stationary policies such that

$$T_{\bar{\mu}} J_\mu = T J_\mu \quad , \text{ or equivalently}$$

$$g(i, \bar{\mu}(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\bar{\mu}(i))\, J_\mu(j) = \min_{u \in U(i)} \left[ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u)\, J_\mu(j) \right]$$

Then we have.

$$J_{\bar{\mu}}(i) \leq J_\mu(i)$$

Furthermore, if $\mu$ is not optimal, strict inequality holds in the above

equation for at least one state $i$

proof $J_\mu = T_\mu J_\mu$ and $T_{\bar{\mu}} J_\mu = T J_\mu$, $\forall i$:

$$J_\mu(i) = g(i, \mu(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\mu(i)) J_\mu(j)$$

$$\geq g(i, \bar{\mu}(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\bar{\mu}(i)) J_\mu(j)$$

$$= (T_{\bar{\mu}} J_\mu)(i)$$

Apply $T_{\bar{\mu}}$ on both sides of this inequality and use the monotonicity of $T_{\bar{\mu}}$ and Corollary 1.2.1.1. we have.

$$J_\mu \geq T_{\bar{\mu}} J_\mu \geq \cdots \geq T_{\bar{\mu}}^k J_\mu \geq \cdots \geq \lim_{N \to \infty} T_{\bar{\mu}}^N J_\mu = J_{\bar{\mu}}$$

If $J_\mu = J_{\bar{\mu}}$, then $J_\mu = T_{\bar{\mu}} J_\mu$.

By hypothesis $\Rightarrow$ $T_{\bar{\mu}} J_\mu = T J_\mu \Rightarrow J_\mu = T J_\mu$.

* Policy Iteration Algorithm

Step 1 (Initialization) Guess an initial stationary policy $\mu^0$

Step 2. (Policy Evaluation) Given the stationary policy $\mu^k$, compute the corresponding cost function $J_{\mu^k}$ from the linear system of equations: $(I - \alpha P_{\mu^k}) J_{\mu^k} = g_{\mu^k}$

Step 3. (Policy Improvement) Obtain a new stationary policy $\mu^{k+1}$ satisfying

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$$

If $J_{\mu^k} = T J_{\mu^k}$, stop. else, return to step 2 and repeat.