PROBABILISTIC ROBOTICS:
MARKOV DECISION PROCESSES

Pierre-Paul TACHER

2.

There are planning algorithms like Lifelong Planning A^* and Dynamic A^* Lite [1] or D^* lite [2] which adresse the context of time varying cost function and the problem of replanning reusing the results of previous searches.

3.

3.1. In this section the actions are deterministic and the state can be described using only the position, using for instance convention in figure 1. Note there is no need to use a discount factor ($\gamma = 1$) for future payoff in this setting. For the mathematically inclined reader, I give a proof of convergence of the value iteration algorithm in the stochastic shortest path setting (which encompass our simple example): cf. A.1. Algorithm is implemented in github repository. The quite obvious 2 equally optimal policies are represented in figure 2.

3.2. Optimal policies does not change by introducing a bit of stochasticity in action. See figure 3 for optimal expected payoff starting from each position.

3.3. We have to add a dimension to the state variable to handle the hidden state variable; there are 3 potential states regarding the knowledge of the position of the final reward. Thus we can

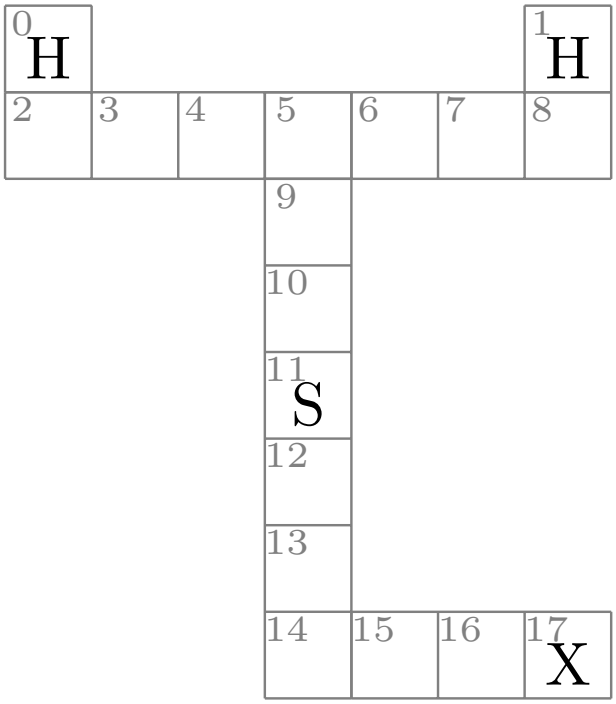

FIGURE 1. Problem setting

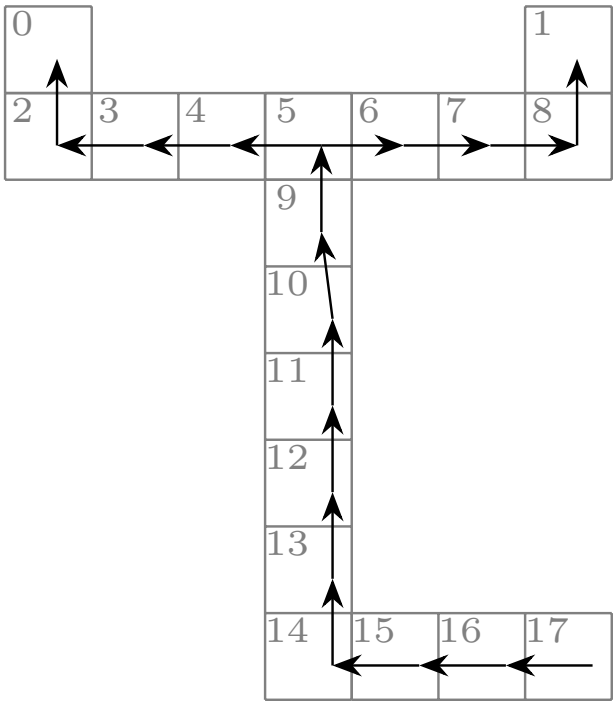

FIGURE 2. Optimal policies

modelize the state by

$$x = (x_1, x_2) \in \mathbb{R}^2$$

where

$$x_1 \in [\![0, 17]\!]$$

$$x_2 = \begin{cases} -1 & \text{if it is known the reward is on the left,} \\ 0 & \text{in absence of any a priori information,} \\ +1 & \text{if it is known the reward is on the right} \end{cases}$$

The knowledge x_2 can change only when reaching position X or ending game in either position H; the transition are represented in figure 4. The expected maximum payoff is represented in figure 5. The optimal policy is represented in figure 6.

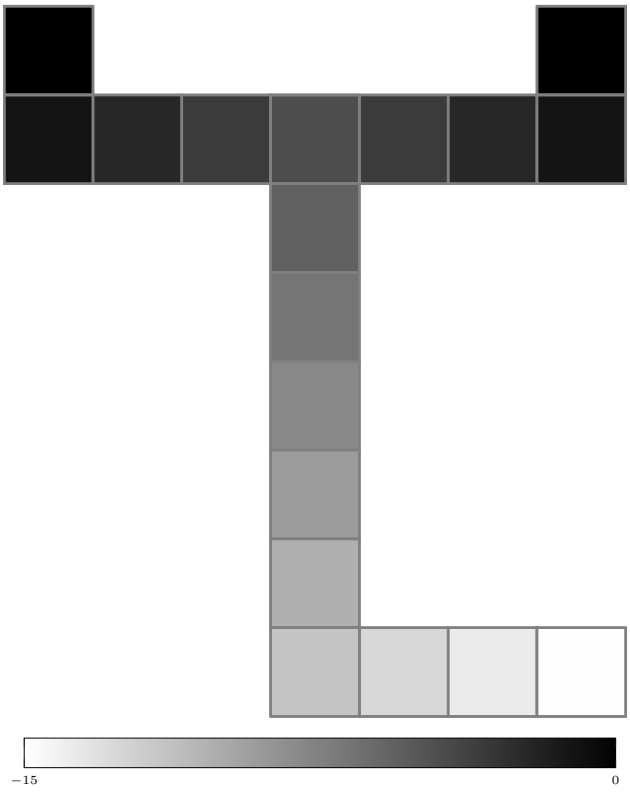

FIGURE 3. State values

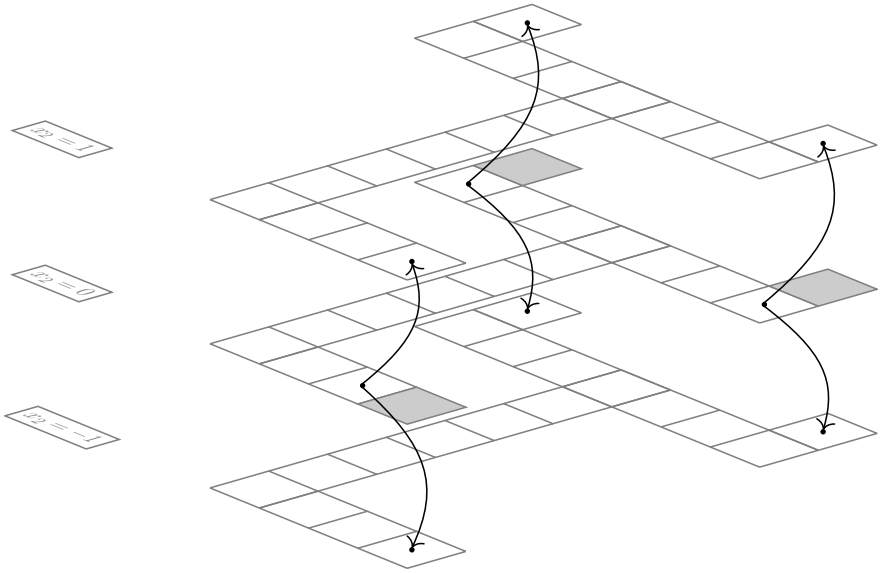

FIGURE 4. Transition between $x_2 = 0$ and $x_2 = \epsilon \in \{-1, 1\}$

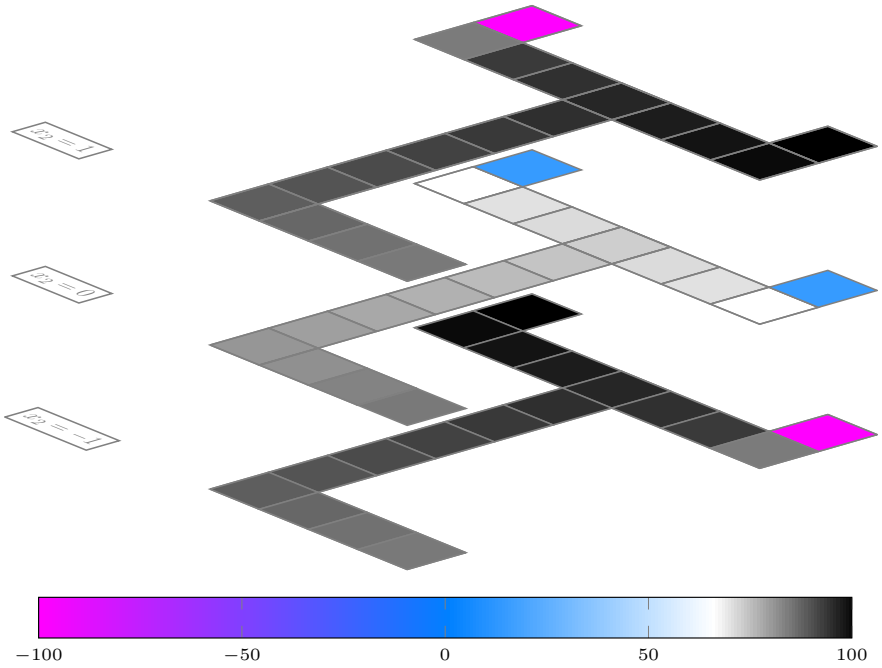

FIGURE 5. States values

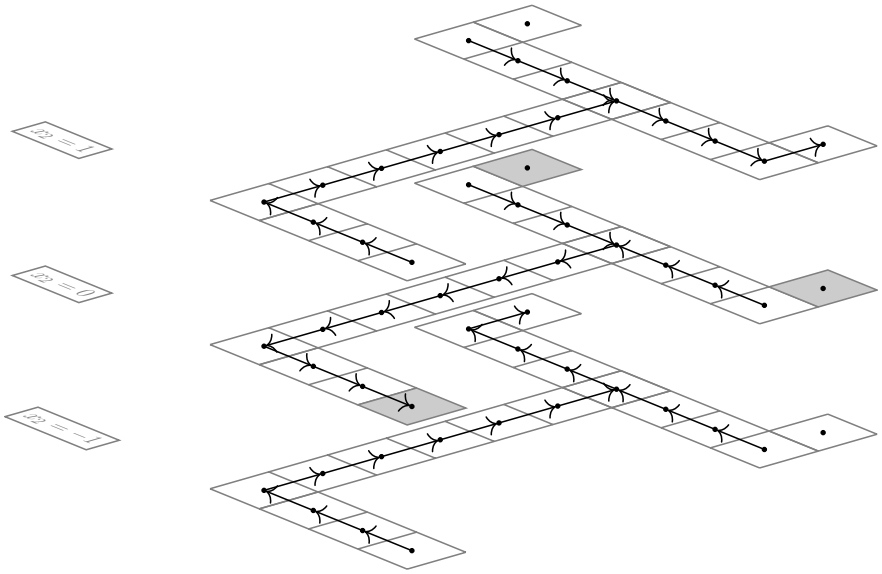

FIGURE 6. Optimal policy

Appendix A.

A.1.

Remark. We give most important convergence results from [3] relative to the stochastic shortest path problem. In what follows, one can replace all minimization by maximization, costs by rewards and apply to the problem of the book.

We consider n states $\{1,\dots,n\}$, and we suppose the existence of a terminal state t where all stops and costs ceases to accumulate. At each state i, there some actions available $u \in U(i)$, where the set $U(i)$ is finite. Any action $u \in U(i)$ started when in state i incurs a deterministic cost

$$g(i,u) \in \mathbb{R}^+$$

We consider a discrete time, infinite horizon setting. Let $J_k \in \mathbb{R}^n$ a vector whose component are minimum expected cost accumulated after $k \in \mathbb{N}$ time step starting from each state. A key point is to see that the minimum expected cost accumulated after $k+1$ time step will be

$$\forall i \in \{1,\dots,n\}, \quad J_{k+1}(i) = \min_{u \in (i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) \times J_k(j) \right]$$

where $p_{ij}(u)$ is the probability of transition from state i to state j when executing the action u. This motivates the need to introduce the following (continuous) operator

$$\begin{aligned} T : \mathbb{R}^n &\to \mathbb{R}^n \\ J &\mapsto T(J) \end{aligned}$$

where

$$\forall i \in \{1,\dots,n\}, \quad T(J)(i) = \min_{u \in (i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) \times J(j) \right]$$

Now, a policy is a way to choose an action from the one available at each states, at every time step. Formally, a policy is a sequence of application $(\mu_k)_{k \in \mathbb{N}}$, with

$$\mu_k : \{1,\dots,n\} \to \prod_{i=1}^{n} U(i)$$

$\mu_k(i)$ is the action chosen at time step k, when in state i. Of particular interest are *stationnary* policies, where the same action is chosen at any point in time

$$\forall i \in \{1,\dots,n\}, \quad \forall k \in \mathbb{N}, \quad \mu_k(i) = \mu_0(i) = \mu(i)$$

If J_k is the vector of the expected costs accumulated after k time steps, starting from each state, when following the policy μ, the cost accumulated under the same policy during $k+1$ time steps will be

$$\forall i \in \{1,\dots,n\}, \quad J_{k+1}(i) = g(i,\mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) \times J_k(j)$$

We note:

$$\begin{aligned} T_\mu : \mathbb{R}^n &\to \mathbb{R}^n \\ J &\mapsto T_\mu(J) \end{aligned}$$

$$\forall i \in \{1,\dots,n\}, \quad T_\mu(J)(i) = g(i,\mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) \times J(j)$$

if we note

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

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

we can write T_μ as a matrix affine transformation:

$$T_\mu(J) = g_\mu + P_\mu J$$

By iterating previous induction, the total cost of executing a policy μ during N time steps will be

$$\underbrace{T_\mu \circ T_\mu \cdots \circ T_\mu}_{N\text{times}}(J_0) = T_\mu^N(J_0)$$

where J_0 is the zero vector; to define the infinite horizon costs of the policy J_μ, we use (some components might be infinite a priori):

$$\begin{aligned} J_\mu &= \lim_{N\to+\infty} \sup_{k\geqslant N} T_\mu^k(J_0) \\ &= \limsup_{N\to+\infty} T_\mu^N(J_0) \end{aligned}$$

In what follows, $\leqslant$ between two vectors in $\mathbb{R}^n$ is meant to be element wise.To prove the convergence result we need to make two assumptions:

Assumption. There exists a stationnary policy such that, starting for any state, the terminal state is reached after n steps with a positive probability, i.e.

$$\forall i \in \{1,\dots,n\}, \quad P(x_n = t \mid x_0 = i, \mu) > 0$$

Such a policy is called *proper*.

Assumption. If μ is improper, there is some component of the sum $\sum_{m=0}^{k-1} P_\mu^m g_\mu$ which diverges to $+\infty$ as $k \to +\infty$: for some state i, following the policy μ gives rise to infinite cost.

Now to the theorem:

Lemma. Properties of proper policies.

(i) for a proper policy μ, the associated cost vector J_μ satisfies

$$\forall J \in \mathbb{R}^n, \quad \forall i \in \{1,\dots,n\}, \quad \lim_{k\to+\infty} (T_\mu^k J)(i) = J_\mu(i)$$

and J_μ is the unique solution of equation

$$J = T_\mu J$$

(ii) A stationnary policy is proper if and only if

$$\exists J \in \mathbb{R}^n, \quad \forall i \in \{1,\dots,n\}, \quad J(i) \geqslant (T_\mu J)(i)$$

Proof. (i) Starting from

$$T_\mu(J) = g_\mu + P_\mu J$$

and using induction we have

$$\forall k \in \mathbb{N}^*, \quad T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu \qquad ①$$

Let

$$\rho_\mu = \max_{i \in \{1,\dots,n\}} P(x_n \neq t \mid x_0 = i, \mu)$$

By assumption, $\rho_\mu < 1$.

$$\begin{aligned}
\forall k \geqslant n, \quad P(x_k \neq t \mid x_0 = i, \mu) &\leqslant P(x_{\lfloor \frac{k}{n} \rfloor \times n} \neq t \mid x_0 = i, \mu) \\
&= P(x_{\lfloor \frac{k}{n} \rfloor \times n} \neq t \mid x_{(\lfloor \frac{k}{n} \rfloor - 1) \times n} \neq t, x_0 = i, \mu) \times P(x_{(\lfloor \frac{k}{n} \rfloor - 1) \times n} \neq t \mid x_0 = i, \mu) \\
&\leqslant \rho_\mu \times P(x_{(\lfloor \frac{k}{n} \rfloor - 1) \times n} \neq t \mid x_0 = i, \mu) \\
&\leqslant \rho_\mu^{\lfloor \frac{k}{n} \rfloor}
\end{aligned}$$

where we have grouped the timesteps by batch of n, used bayes and straightforward induction. This shows that

$$\lim_{k \to +\infty} P(x_k \neq t \mid x_0 = i, \mu) = 0$$

so

$$\forall j \in \{1, \dots, n\}, \quad \lim_{k \to +\infty} P(x_k = j \mid x_0 = i, \mu) = 0$$

This is equivalent to say

$$\lim_{k \to +\infty} P_\mu^k = 0 \qquad ②$$

On the other hand,

$$\left\| P_\mu^m g_\mu \right\| \leqslant \rho_\mu^{\lfloor \frac{m}{n} \rfloor} \max_{i \in \{1,\dots,n\}} |g(i, \mu(i))|$$

The latter being the term of converging series, this shows that the cost series $(\sum_{m=0}^{k-1} P_\mu^m g_\mu)_{k \in \mathbb{N}}$ converges and

$$\begin{aligned}
J_\mu &= \limsup_{k \to +\infty} T^k J_0 \\
&= \limsup_{k \to +\infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu \\
&= \lim_{k \to +\infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu \qquad ③
\end{aligned}$$

where J_0 is the null vector.so,

$$①, ②, ③ \Rightarrow \lim_{k \to +\infty} T_\mu^k J = J_\mu$$

Also,

$$\begin{aligned}
T_\mu^{k+1} J &= T_\mu(T_\mu^k J) \\
&= g_\mu + P_\mu T_\mu^k J
\end{aligned}$$

Letting k to $+\infty$, and using continuity of linear application T_μ

$$J_\mu = g_\mu + P_\mu J_\mu$$

Besides, if for some J, we have $J = T_\mu J$, by induction:

$$\begin{aligned} &\forall k \in \mathbb{N}^*, \quad J = T_\mu^k J \\ \Rightarrow \qquad & J \qquad = \lim_{k\to+\infty} T_\mu^k J \\ & \qquad = J_\mu \end{aligned}$$

(ii) If μ is proper, then $J_\mu \geqslant T_\mu(J_\mu)$. Conversely, if $J \geqslant T_\mu J$ for some $J \in \mathbb{R}^n$, by iterating and using the non decreasing property of T_μ we have

$$\begin{aligned} \forall k \in \mathbb{N}^*, \quad J &\geqslant T_\mu^k J \\ &= P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu \end{aligned}$$

If μ were not proper, some component in the sum in the right hand side would diverge to $+\infty$ by our second assumption, which is a contradiction.

□

Proposition. Bellman's equation.

(i) The optimal cost vector J^* satisfies the Bellman's equation

$$J^* = TJ^*$$

and it is the only solution of this equation.

(ii) The value iteration algorithm converges for any starting cost vector J, i.e.

$$\forall J \in \mathbb{R}^n, \quad \lim_{k\to+\infty} T^k(J) = J^*$$

Proof. There exists a fixed point of mapping T: Let $\mu = \mu_0$ a proper policy. Let us define policy μ_1 by

$$T_{\mu_1} J_{\mu_0} = TJ_{\mu_0}$$

Thus

$$\begin{aligned} T_{\mu_1} J_{\mu_0} &\leqslant T_{\mu_0} J_{\mu_0} \\ &= J_{\mu_0} \end{aligned}$$

and by lemma (ii) the policy μ_1 is proper. Applying T_{μ_1} repeatedly to previous inequality, we have

$$\forall k \in \mathbb{N}^*, \quad T_{\mu_1}^k J_{\mu_0} \leqslant TJ_{\mu_0} \leqslant J_{\mu_0}$$

and letting $k \to +\infty$,

$$J_{\mu_1} \leqslant TJ_{\mu_0} \leqslant J_{\mu_0}$$

Reiterating we see that we can construct a sequence of proper policy $(\mu_k)_{k\in\mathbb{N}}$ satisfying:

$$\forall k \in \mathbb{N}^*, \quad J_{\mu_{k+1}} \leqslant TJ_{\mu_k} \leqslant J_{\mu_k} \qquad ①$$

The space of admissible policies being finite,

$$\exists i, j, \quad i \neq j, \quad \mu_i = \mu_j$$

but then,

$$① \Rightarrow \exists i, \quad J_{\mu_i} = TJ_{\mu_i}$$

The fixed point of mapping T is unique: Let J, J' such that $J = TJ$ and $J' = TJ'$. The actions available being in finite number, the minimum in T exists and we can find policies μ, μ' such that $J = T_\mu J$ and $J' = T_\mu J'$. By lemma (ii), μ and μ' are proper and by lemma (i), J and J' are the respective cost of each policy: $J = J_\mu$ and $J' = J_{\mu'}$. We have $J = T^k J \leqslant T^k_{\mu'} J$ so $J \leqslant \lim_{k\to+\infty} T^k_{\mu'} J = J'$ by lemma (i). Similarly, $J' \leqslant J$ so that finally $J = J'$.

Now let consider the equation

$$J = g_\mu + \delta e + P_\mu J$$

where μ is the proper fixed point of T, $\delta > 0$ and

$$e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

By lemma (i), it has one unique solution $\hat{J}$ (only difference is the g costs have all been augmented by δ). It is clear that $J_\mu \leqslant \hat{J}$ (the expected cost cannot decrease when g is higher) implying

$$J_\mu = TJ_\mu \leqslant T\hat{J} \leqslant T_\mu \hat{J} = \hat{J} - \delta e \leqslant \hat{J}$$

By induction,

$$\forall k \in \mathbb{N}^*, \quad J_\mu = T^k_\mu J_\mu \leqslant T^k \hat{J} \leqslant T^{k-1}\hat{J} \leqslant \hat{J}$$

This shows that the sequence $(T^k \hat{J})_{k\in\mathbb{N}}$ is non increasing with lower bound, so converges to some $\tilde{J}$. Furthermore,

$$\begin{aligned} T\tilde{J} &= T(\lim_{k\to+\infty} T^k \hat{J}) \\ &= \lim_{k\to+\infty} T(T^k \hat{J}) \\ &= \lim_{k\to+\infty} T^{k+1}\hat{J} \\ &= \tilde{J} \end{aligned}$$

by the continuity of T, unicity of fixed point proves that $\tilde{J} = J_\mu$. Now we construct a second sequence whose limit is J_μ, with non decreasing values:

$$J_\mu - \delta e = TJ_\mu - \delta e \leqslant T(J_\mu - \delta e) \leqslant TJ_\mu = J_\mu$$

By iterating in similar way as before, we show that the sequence $(T^k(J_\mu - \delta e))_{k\in\mathbb{N}}$ converges to J_μ. Now for any $J \in \mathbb{R}^n$, for some fixed $\delta > 0$, we can bound J by the two sequences:

$$\begin{aligned} J_\mu - \delta e \quad &\leqslant \quad J \quad \leqslant \hat{J} \\ \Rightarrow \forall k \in \mathbb{N}^*, \quad T^k(J_\mu - \delta e) &\leqslant T^k J \leqslant T^k \hat{J} \end{aligned}$$

and letting $k \to +\infty$ shows

$$\forall J \in \mathbb{R}^n, \quad \lim_{k\to+\infty} T^k J = J_\mu$$

To conclude the proof, let a (non necessarily stationary) policy $\pi = (\mu_k)_{k\in\mathbb{N}}$, its k step cost is bounded below by

$$T_{\mu_0} T_{\mu_1} \dots T_{\mu_{k-1}} J_0 \geqslant T^k J_0$$

Taling the limit superior,

$$J_\pi \geqslant J_\mu$$

showing μ is an optimal stationary policy and $J_\mu = J^*$ is the minimum expected cost. □

References

[1] Likhachev, Maxim; Koenig, Sven: *Lifelong Planning A* and Dynamic A* Lite: The proofs*, (2001)
[2] Likhachev, Maxim; Koenig, Sven: *D* lite*, American Association for Artificial Intelligence (2002)
[3] Bertsekas, Dimitri P.: *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming*, Athena Scientific; 4th edition (2012)