# Assignment 2, ECSE 526

Yulin Shi, 260629628

March 15, 2015

## 1 Description of the transient process

The dynamic Markov process of the "hospital problem" can be expressed as several variables and functions (Figure 1). In this problem, a time step $k$ is consists of $action_k$ and $state_k$.

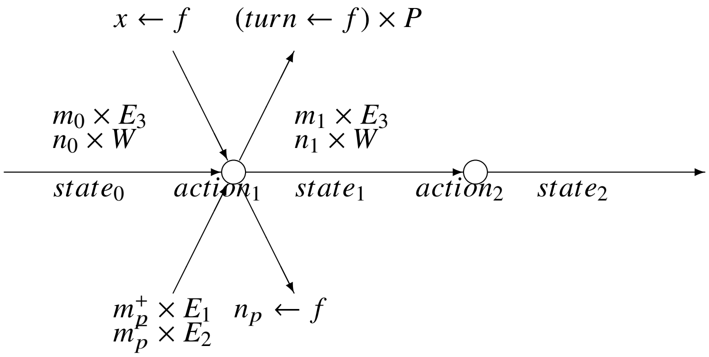


Figure 1: a set of 3 points in the plane that can be shattered by this hypothesis class

Variables in the figures are:

Table 1: variables table

| variables | description | cost | time | value |
|---|---|---|---|---|
| $m$ | open treatment room | $E_3$ | $state_k$ | $[0, M]$ |
| $n$ | patients in waiting | $W$ | $state_k$ | $[0, N]$ |
| $m_p^+$ | room to open | $E_1$ | $action_k$ | $[0, M - m]$ |
| $m_p^-$ | room to close | $E_2$ | $action_k$ | $[-m, 0]$ |
| $x$ | coming patients | - | $action_k$ | $\leftarrow f$ |
| $turn$ | patients turned back | $P$ | $action_k$ | $\min(0, x + n - m - N)$ |
| $n_p$ | increasement of waiting | - | $action_k$ | $x - turn - m$ |

symbol "$\leftarrow f$" means this variable is determined by probability $f$ (poisson or normal).

State transient functions are

- transient function of state $m_k$

$$m_{k+1} = m_k + m_p^+ - m_p^- \tag{1}$$

- transient function of state $n_k$

$$n_{k+1} = n_k + n_p = n_k + x - turn - m \tag{2}$$

According to transient equations (Equation 1, 2), in each step, the state $s(m, n)$ will transfer to new state $s'(m', n')$ wherein, $m'$ is can be determined by the certain policy, but $n'$ is a random value.

The cost function at time $k$ is the sum of action cost at $k$, and state cost at $k$.

$$cost_k = cost_{a_k} + cost_{s_k} = turn \times P + m_p^+ \times E_1 + m_p^- \times E_2 + m \times E_3 + n \times W \tag{3}$$

But, in this problem, the action of changing the opened treating room from $m_k$ to $m_{k+1}$ has no impact on $n_{k+1}$ which is influenced by the random variable $x$. (This is the second point where the "hospital problem" differs from the "4 × 3 world" problem in the textbook [1].) Therefore, before value iteration, we should define the reward function at time $k$ of this problem to be the expectation of the cost function defined in Equation 3.

$$Reward_k = \mathbb{E}[cost_k] = cost_{a_k} + \mathbb{E}[cost_{s_k}] = cost_{a_k} + \sum_{x_k} P(s_k|s_k, a_k)cost_{s_k} \tag{4}$$

Utility of a state sequence is defined as the sum of rewards along the sequence

$$Utility_N = \sum_{k=0}^{N} \gamma^k Reward_k \tag{5}$$

wherein, discount parameter $\gamma \in [0, 1]$ is chosen as 1 in this "hospital problem". Plus that there is no termination condition in this problem (This is the first point where the "hospital problem" differs from the "4 × 3 world" problem in the textbook [1].), we have

$$Utility \to \infty \text{ when } k \to \infty \tag{6}$$

Corresponding techniques about dealing with $\gamma = 1$ will be introduced in this article.

Policy $\pi \in \mathbb{R}^2 \to \mathbb{R}$ is defined as the mapping function

$$\pi : (m_k, n_k) \to m_{k+1} \tag{7}$$

# 2 Optimal policy

## 2.1 value iteration

For Markov decision process under a fixed policy (not necessary optimal policy), the the expected value of a state at time k can be solved with a one-step back propagation algorithm, also called value iteration [2].

$$U(s) = \sum_{s'} P_{\pi(s)}(s'|s, a)(R(s) + \gamma U(s'|s, a)) = R(s) + \gamma \sum_{s'} P_{\pi(s)}(s'|s, a)U(s'|s, a) \tag{8}$$

Since $\gamma = 1$ in the "hospital problem", and there is the value of expected utility after iteration can go to infinity. Then, how to compare different policies? Turn back to the physical meaning of this problem, we'll known that, once a policy is determined, the hospital will operate under that fixed policy forever.

Therefore, the contribution initial condition and rewards at the very initial several steps are trivial; on the contrary, what's real important is the converged utility at a infinity time point (if it converges) which can be solved by letting $U(s) = U(s'|s, a)$ in Equation 8.

$$R_\infty = \lim_{k\to\infty} \left\{ U_k - \gamma \sum_{s'} P_{\pi,k+1|k} U_{k+1} \right\} \tag{9}$$

In this problem, $R_\infty$ will usually converge to a nearly (99.9%) homogeneous field (when no source and no dismiss in a finite space) within 10 steps.

A very good property of $R_\infty$ is that it's irrelevant to initial state. It's determined only by the fixed policy.

$$R_\infty = R_\infty(\pi) \tag{10}$$

## 2.2 Policy Iteration

According to Equation 10, the optimal polity for a certain "hospital problem" with certain parameters ($M$, $N$, costs, distribution of $x$) can be defined as:

$$\pi^* = \arg\min_\pi R_\infty(\pi) \tag{11}$$

Since $\pi : (m, n) \to m'$ is a single value function, to find $\pi^*$, we design a update algorithm which change only one value of the policy function in the state space

**Algorithm 1** Complete Policy Iteration

```
Initialize: $\pi_{opt} \leftarrow \pi_0$
loop
    $\pi \leftarrow \pi_{opt}$
    for $m \leftarrow 0, 1, ..., M$ do
        for $n \leftarrow 0, 1, ..., N$ do
            $R_{\infty,min} \leftarrow \infty$
            for $m' \leftarrow 0, 1, ..., M$ do
                $\pi_{temp} \leftarrow \pi$
                $\pi_{temp}(m, n) \leftarrow m'$                                ▷ new policies
                if $R_\infty(\pi_{temp}) < R_{\infty,min}$ then
                    $R_{\infty,temp} \leftarrow R_\infty(\pi_{temp})$
                    $\pi_{opt}(m, n) \leftarrow m'$
                end if
            end for
        end for
    end for
    if $\pi_{opt} = \pi$ then
        BREAK
    end if
end loop
```

Complexity of the complete policy iteration algorithm is $O(n_1 \times M \times N \times M \times n_2)$ ($n_1$ is the number of loops of policy iteration, $n_2$ is the number of loops of value iteration). I limited the step of value iteration (Line 10 of Algorithm 1) to be 10, thus modified the algorithm to be a "modified policy algorithm" introduced in Reference [2], and reduced the complexity to be $\mathcal{O}(n_1 \times M \times N \times M)$

We might prefer reducing the step size of the policy updating, and avoiding oscillation. Therefore, I modify Line 7 of Algorithm 1 to be

$$m' \leftarrow \max(\pi(m,n) - 1, 0), ..., \min(\pi(m,n) + 1, M) \tag{12}$$

and permit $m'$ change only one unit per loop, and further reduced the complexity of the algorithm to be $\mathcal{O}(n_1 \times M \times N)$.

### 2.3 Bellman Equation

Bellman equation combines value iteration with policy iteration in one iteration step.

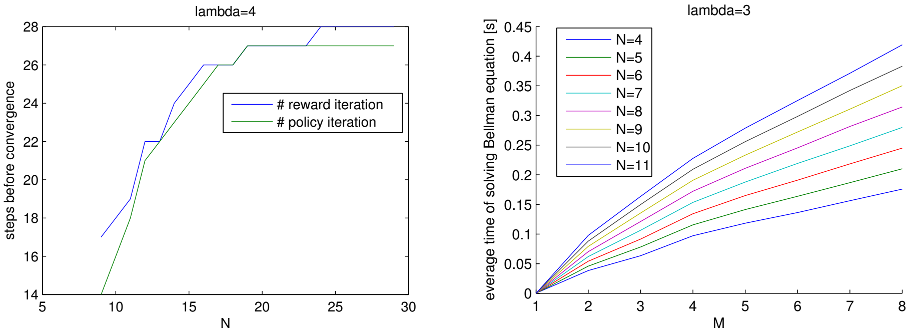
$$U(m,n) \leftarrow R(m,n,\pi) + \max_{m' \in A(m,n,\pi)} \sum_{n'} P(n')U(m,n) \tag{13}$$

wherein, we search suboptimal actions from a small set

$$A(m,n,\pi) = \max(a(m,n,\pi) - 1, 0), ..., \max(a(m,n,\pi) + 1, M) \tag{14}$$

It's corresponding algorithm for the "hospital problem" is Algorithm 2.

Complexity of my Algorithm 2 is $\mathcal{O}(n_1 \times M \times N)$, wherein, variable $n_1$ is the estimate of Policy iteration steps whose maximum value is $M^2N$ in worst case. It's function w.r.t. variables $M, N$ can be estimated by test. In the test of Poisson distribution (Picture 2a) wherein $\lambda = 4$, policy iteration steps appears to be always irrelevant to $M$, irrelevant to $N$ when $N > 25$ for $\lambda = 4$. Therefore, combining the result in Picture 2a who shows average time cost of each step of policy iteration are almost proportional to $M \times N$, the estimated complexity according to experiment is $\mathcal{O}(MN)$.



(a) steps of policy iteration

(b) time cost w.r.t. $M, N$

Figure 2: complexity w.r.t $M, N$

The convergence performance of the Bellman iteration is demonstrated by an example ($M = 3, N = 7, \lambda = 2$). The log of the experiment is archived in file "logDemoBellmanIteration.txt". The convergence is indicated by the decreasing of norm errors in Figure 3.

**Algorithm 2** Bellman Equation

```
Initialize: $\pi_{opt} \leftarrow \pi_0, \quad U \leftarrow 0$
loop
    $\pi \leftarrow \pi_{opt}$
    $U' \leftarrow U$
    for $m \leftarrow 0, 1, ..., M$ do
        for $n \leftarrow 0, 1, ..., N$ do
            $U_{expect,min} \leftarrow \infty$
            for $m' \leftarrow \max(\pi(m, n) - 1, 0), ..., \min(\pi(m, n) + 1, M)$ do
                $\pi_{temp} \leftarrow \pi$
                $\pi_{temp}(m, n) \leftarrow m'$                          ▷ new policies
                $U_{expect} \leftarrow 0$
                for $x \leftarrow \arg_x P(x)$ do                         ▷ solve the expectation
                    $U_{expect} \leftarrow U_{expect} + P(x)U'(m', n')$
                end for
                if $U_{expect} < U_{expect,min}$ then     ▷ find the minimal expectation and its
    corresponding updated policy
                    $U_{expect,min} \leftarrow U_{expect}$
                    $\pi_{opt}(m, n) \leftarrow m'$
                end if
                $U(m, n) \leftarrow R(m, n) + U_{expect,min}$
            end for                                                        ▷ end of loop over m’
        end for                                                            ▷ end of loop over n
    end for                                                                ▷ end of loop over m
    if $U - U'$ convergences then
        BREAK
    end if
end loop
```

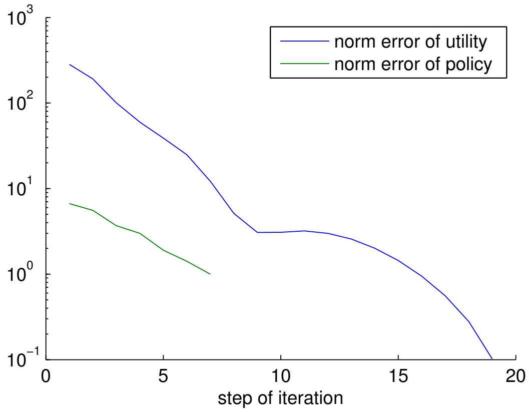


Figure 3: a demo of Bellman iteraton ($M = 3, N = 7, x \leftarrow$ "poisson" , $\lambda = 2$)

# 3 Attachment

## 3.1 value iteration in vector form

Recall from value iteration equation that, the reward and utility are only determined by states. For a fixed policy, actions are only determined by states. Then, transit probability, which is the function of states and action, are only determined by states now. Therefore, given policy, we eliminate action variable in value iteration equation, map every variable to states, and rewrite Equation 8 to be

$$u_i^{(n-1)}(s_i) = \gamma_1 r_i(s_i) + \gamma_2 \sum_j p_{ij}(s_j, s_i) u_j^{(n)}(s_j) \tag{15}$$

In vector form, Equation 15 is

$$\boldsymbol{u}^{(n-1)} = \gamma_1 \boldsymbol{r} + \gamma_2 \boldsymbol{P}\boldsymbol{u}^{(n)} \tag{16}$$

wherein, function space included are all defined on state space as:

$$\begin{aligned} &\boldsymbol{s} \in \mathbb{R}^{MN} \\ &\boldsymbol{u}^{(n)} \in \mathbb{R}^{MN}, \quad \boldsymbol{u}^{(n)} \leftarrow \boldsymbol{s} \\ &\boldsymbol{r} \in \mathbb{R}^{MN}, \quad \boldsymbol{r} \leftarrow \boldsymbol{s} \\ &\boldsymbol{P} \in \mathbb{R}^{MN} \times \mathbb{R}^{MN}, \quad \boldsymbol{P} \leftarrow \boldsymbol{s} \times \boldsymbol{s} \end{aligned} \tag{17}$$

wherein, superscript $MN$ is the dimension of state space in the "hospital problem".

Now, given value iteration equation in vector form (Equation 16) and a initial utility $\boldsymbol{u}^{(n)}$, $\boldsymbol{u}^{(k)}, k < n$, can be solved recursively

$$\begin{aligned} \boldsymbol{u}^{(n-2)} &= \gamma_1 \boldsymbol{r} + \gamma_2 \boldsymbol{P}\boldsymbol{u}^{(n-1)} \\ &= \gamma_1 \boldsymbol{r} + \gamma_2 \boldsymbol{P}(\gamma_1 \boldsymbol{r} + \gamma_2 \boldsymbol{P}\boldsymbol{u}^{(n)}) \\ &= \gamma_1 (\boldsymbol{I} + \gamma_2 \boldsymbol{P})\boldsymbol{r} + \gamma_2^2 \boldsymbol{P}^2 \boldsymbol{u}^{(n)} \end{aligned} \tag{18}$$

$$\boldsymbol{u}^{(0)} = \gamma_1 (\boldsymbol{I} + \gamma_2 \boldsymbol{P} + \cdots + \gamma_2^{n-1} \boldsymbol{P}^{n-1})\boldsymbol{r} + \gamma_2^n \boldsymbol{P}^n \boldsymbol{u}^{(n)} \tag{19}$$

Recall the definition of the probability matrix which gives the following properties

$$\sum_j p_{ij} = 1, \quad \max |\lambda_i(\boldsymbol{P})| = 1, \quad 0 < \|\boldsymbol{P}^\infty\| < \infty \tag{20}$$

$$\begin{aligned} &\forall \boldsymbol{u} \text{ s.t. } 0 < \|\boldsymbol{u}\|_p < \infty, \quad \exists \alpha \in (0, 1], \quad \text{ s.t. } \\ &\alpha \|\boldsymbol{u}\|_\infty \le \|\boldsymbol{P}^\infty \boldsymbol{u}\|_\infty \le \|\boldsymbol{u}\|_\infty < \infty, \quad \alpha \|\boldsymbol{u}\|_p \le \|\boldsymbol{P}^\infty \boldsymbol{u}\|_p \le \|\boldsymbol{u}\|_p < \infty \end{aligned} \tag{21}$$

According to property Equation 20, eigenvalue matrix $\boldsymbol{D} = \boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^{-1}$ has the following property

$$\max |\lambda_i(\boldsymbol{D})| = 1, \quad \boldsymbol{D}^{k+1} \le \boldsymbol{D}^k \tag{22}$$

1. If $0 < \gamma_2 < 1$,

$$\begin{aligned} \boldsymbol{u}^{(0)} &= \gamma_1 \boldsymbol{V}^{-1} (\boldsymbol{I} + \gamma_2 \boldsymbol{D} + \cdots + \gamma_2^{n-1} \boldsymbol{D}^{n-1}) \boldsymbol{V}\boldsymbol{r} + \gamma_2^n \boldsymbol{P}^n \boldsymbol{u}^{(n)} \\ &= \gamma_1 \boldsymbol{V}^{-1} (\boldsymbol{I} - \gamma_2 \boldsymbol{D})^{-1} (\boldsymbol{I} - \gamma_2^n \boldsymbol{D}^n) \boldsymbol{V}\boldsymbol{r} + \gamma_2^n \boldsymbol{P}^n \boldsymbol{u}^{(n)} \end{aligned} \tag{23}$$

$$\lim_{n\to\infty} \boldsymbol{u}^{(0)} = \gamma_1 \boldsymbol{V}^{-1} (\boldsymbol{I} - \gamma_2 \boldsymbol{D})^{-1} \boldsymbol{V}\boldsymbol{r} \tag{24}$$

2. If $\gamma_2 = 1$, Since $\max|\lambda_i(\boldsymbol{D})| = 1$, the Geometric matrix series in Equation 19 diverge

$$\lim_{n\to\infty} \boldsymbol{u}^{(0)} \to \infty + \lim_{n\to\infty} \gamma_2^n \boldsymbol{P}^n \boldsymbol{u}^{(n)} \to \infty \tag{25}$$

On the contrary, the newly defined utility function called average utility converges

$$\bar{\boldsymbol{u}}^{(0)} = \frac{1}{n}\boldsymbol{u}^{(0)} \quad = \frac{1}{n}\gamma_1(\boldsymbol{I} + \boldsymbol{P} + \cdots + \gamma_2^{n-1}\boldsymbol{P}^{n-1})\boldsymbol{r} + \frac{1}{n}\boldsymbol{P}^n\boldsymbol{u}^{(n)} \tag{26}$$

$$\begin{aligned} \|\bar{\boldsymbol{u}}^{(0)}\| &\le \frac{1}{n}\gamma_1(\|\boldsymbol{r}\| + \|\boldsymbol{P}\boldsymbol{r}\| + \cdots + \|\boldsymbol{P}^{n-1}\boldsymbol{r}\|) + \frac{1}{n}\|\boldsymbol{P}^n\boldsymbol{u}^{(n)}\| \\ &\le \frac{1}{n}\gamma_1(\|\boldsymbol{r}\| + \|\boldsymbol{r}\| + \cdots + \|\boldsymbol{r}\|) + \frac{1}{n}\|\boldsymbol{P}^n\boldsymbol{u}^{(n)}\| \\ &= \gamma_1\|\boldsymbol{r}\| + \frac{1}{n}\|\boldsymbol{P}^n\boldsymbol{u}^{(n)}\| \end{aligned} \tag{27}$$

$$\lim_{n\to\infty} \|\bar{\boldsymbol{u}}^{(0)}\| \le \gamma_1\|\boldsymbol{r}\| + \lim_{n\to\infty} \frac{1}{n}\|\boldsymbol{P}^n\boldsymbol{u}^{(n)}\| = \gamma_1\|\boldsymbol{r}\| \tag{28}$$

# References

[1] Russell, Stuart, Peter Norvig, *Artificial Intelligence, A Modern Approach*. Prentice-Hall, Egnlewood Cliffs 25, 1995.

[2] "Markov Decision Process." *Wikipedia: The Free Encyclopedia* . Wikimedia Foundation, Inc. 21 Feb, 2015.