著者 | 韓 亜慧 岳 五
雑誌名 | 甲南大学紀要 工学部編
巻 | 2
号 | 1
ページ | 49-60
発行年 | 2003-07-31
URL | http://doi.org/10.14990/00000060
Properties of the Optimality Equation and Optimal Policies in Discrete Time Markov Decision Processes and Their Applications

†Qiying Hu and ‡Wuyi Yue

†College of International Business and Management, Shanghai University
Shanghai 201800, CHINA

‡Department of Information Science and Systems Engineering, Konan University
Kobe 658-8501, JAPAN

(Received April 15, 2003)

Abstract
This paper investigates properties of the optimality equation and optimal policies in discrete time Markov decision processes with expected discounted total rewards. Under conditions where the model is well defined and the optimality equation is true, it is shown that as a solution of the optimality equation, the solution called optimal value function is always the smallest one, and is also the unique one under another weak condition. Moreover, a structure of optimal policies is discussed. Finally, these properties are applied to state feedback control of discrete event systems with a numerical example.

keywords: Discrete time, Markov decision processes, Optimality equation, Optimal policies, Expected discounted total rewards, Control system.

1 Introduction

This paper will discuss some properties of the optimality equation and optimal policies in discrete time Markov decision processes (MDP) with expected discounted total rewards.

In the literature, the usual method is first to present a set of conditions, then to show that the model is well defined under these conditions, and finally to show that the optimality equation is true and some other properties [1]-[3]. One of the main properties discussed in the literature is that of optimal policies. For example, the authors of [4] and [5] discussed some properties of optimal policies respectively for discrete time MDP and discounted semi-Markov decision processes with unbounded rewards. In [6], it is shown that for a discrete time MDP with expected discounted total rewards, when the model is well defined, its state space $S$ can be partitioned into three sub-sets $S_{\infty}$, $S_{\infty}$, $S_0$, on which the
optimal value function equals $+\infty$, $-\infty$ or is finite, respectively. Furthermore, the validity of the optimality equation is shown when its right hand side is well defined, especially, when it is restricted to the subset $S_0$.

This paper is a sequent to [6]. Based on the results obtained in [6], this paper will discuss the properties of the optimality equation and that of optimal policies under the conditions where the model is well defined and the optimality equation is true. Moreover, these properties are applied to state feedback control of discrete event systems and a numerical example is given.

The remainder of the paper is organized as follows. In Section 2, the system model is given and in Section 3, the optimal value function is characterized as a solution of the optimality equation. In Section 4, the structure of optimal policies is discussed. Section 5 discusses an application to state feedback control of discrete event systems with a numerical example, while Section 6 is a concluding section.

2 The System Model

The model of discrete time MDP discussed here is $\{S, ((A(i), \mathcal{A}(i)), i \in S), p, r\}$, where the state space $S$ is countable; the action set $A(i)$ for $i \in S$ is nonempty and $(A(i), \mathcal{A}(i))$ is a measurable space. Each single point set of $A(i)$ is measurable. The system, when is in state $i$ and an action $a \in A(i)$ is taken at some period, will transit to $j$ at the next period with probability $p_{ij}(a)$, and incurs an extended real-valued reward $r(i, a)$. We assume that both $p_{ij}(a)$ and $r(i, a)$ are measurable in $a$ for any $i, j \in S$.

Let $\Gamma = \{(i, a)| i \in S, a \in A(i)\}$, $H_n = D^\Gamma_{n-1} \times S$ be the set of history up to $n$ for $n > 0$ and $H_0 = S$. We define a policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ as: for $n \geq 0$ and $h_n = (i_0, a_0, \ldots, i_n) \in H_n$, $\pi_{i_n}(\cdot|h_n)$ is a probability distribution on $A(i_n)$, and for each $A \in \mathcal{A}(i_n)$, $\pi_{i_n}(A|h_n)$ is measurable in $a_0, a_1, \ldots, a_{n-1}$. If $\pi_{i_n}(\cdot|h_n) = \pi_{i_n}(\cdot|i_n)$ for all $h_n$, then we call $\pi$ a Markov policy, the set of which is denoted by $\Pi_M$. The set of decision functions is $F = \times_i A(i)$, and $f \in F$ is also called a stationary policy.

Let $\beta \in (-\infty, +\infty)$ be discount rate, $X_n$ and $\Delta_n$ denote the state and the action taken at period $n$. Then for $i \in S$, $\pi \in \Pi$, the expected discounted total reward is defined by

$$V_\beta(\pi, i) = \sum_{n=0}^{\infty} \beta^n E_{\pi_n}\{r(X_n, \Delta_n)\}. \tag{1}$$

The following two conditions are assumed to be true through this paper.

Condition 1: $V_\beta(\pi, i)$ is well defined for all $\pi \in \Pi$ and $i \in S$.

This condition means that the model is well defined, i.e., (1) for each policy $\pi$, state $i$ and $n \geq 0$, $E_{\pi_n}\{r(X_n, \Delta_n)\}$ is well defined (may be infinite); (2) as a series, $V_\beta(\pi, i)$ is convergent (may be also infinite).
Condition 2: For any policy \( \pi = (\pi_0, \pi_1, \ldots) \) \( \in \Pi \) and state \( i, j \in S \),

\[
V_\beta(\pi, i) = \int_{\Delta(i)} \pi_0(da | i) \left\{ r(i, a) + \beta \sum_j p_j(a) V_\beta(\pi^{i,a}, j) \right\}
\]

(2)

where \( \pi^{i,a} = (\sigma_0, \sigma_1, \ldots) \in \Pi \) with \( \sigma_n(\cdot | h_n) = \pi_{n+1}(\cdot | i, a, h_n) \) for \( n \geq 0 \).

This condition states that the total process under a policy \( \pi \) can be partitioned into two sub-processes: the first period and the remaining periods. So it is, in fact, the stage separation, which is essential for the optimality principle in dynamic programming and in MDP. Condition 2 implies also that the series and integration in Eq. (2) are convergent. This condition is shown usually in the literature under various conditions that \( r(i, a) \) is nonnegative, nonpositive, or satisfies boundedness conditions, etc. (see [2]).

Let the optimal value function \( V_\beta(i) = \sup_{\pi \in \Pi} \{ V_\beta(\pi, i) \} \) for \( i \in S \). For \( \epsilon \geq 0 \), \( \pi^* \in \Pi \), if \( V_\beta(\pi^*, i) \geq V_\beta(i) - \epsilon \) (if \( V_\beta(i) < +\infty \)) or \( \geq 1/\epsilon \) (if \( V_\beta(i) = +\infty \)), then \( \pi^* \) is called \( \epsilon \)-optimal. Here \( 1/0 = +\infty \) is assumed. An 0-optimal policy is simply called an optimal policy.

According to the results in [6], we further assume that the following condition is true.

Condition 3: The optimal value function \( V_\beta(i) \) is finite for each \( i \in S \) and satisfies the following optimality equation:

\[
V_\beta(i) = \sup_{\pi \in \Pi} \left\{ r(i, a) + \beta \sum_j p_j(a) V_\beta(j) \right\}.
\]

(3)

3 Properties of the Optimality Equation

In general, the solutions of Eq. (3) may be not unique. An example can be seen in [3]. This section will discuss which one among the solutions of the optimality equation is the optimal value function, and under which condition the solution of Eq. (3) is unique. First, we give the following lemma.

Lemma 1: For each policy \( \pi \) and state \( i \in S \) with that \( V_\beta(\pi, i) \) is finite,

\[
\lim_{n} \inf \beta^n E_{\pi_i}V_\beta(X_n) \geq 0.
\]

(4)

Moreover, if \( \pi = (f_0, f_1, \cdots) \in \Pi_\infty^d \), then \( \lim_{n} \sup \beta^n E_{\pi_i}V_\beta(X_n) \) is finite.

Proof: First, by Condition 2, we have

\[
V_\beta(\pi, i) = E_{\pi_i}r(X_0, \Delta_0) + \beta E_{\pi_i}V_\beta(\pi, X_1)
\]

= \cdots

= V_\beta(\pi, i) + \beta^n E_{\pi_i}V_\beta(\pi, X_n)\]
where $\beta^n E_{\pi}, V_\beta(\pi, X_n) = \sum_{r=1}^{\infty} \beta^r E_{\pi, r}(X_n, \Delta_n)$. So it follows the conditions 1 and 2 that
\[
\lim_{n \to \infty} \beta^n E_{\pi, i}, V_\beta(\pi, X_n) = 0 \quad \text{if} \ V_\beta(\pi, i) \text{ is finite. Thus}
\]
\[
\liminf_{n \to \infty} \beta^n E_{\pi, i}, V_\beta(X_n) \geq \liminf_{n \to \infty} \beta^n E_{\pi, i}, V_\beta(\pi, X_n) = 0.
\]

Second, it follows the optimality equation that
\[
V_\beta \geq r(f) + \beta P(f)V_\beta, \forall f \in F
\]
from which one can prove by the induction method that for $\pi = (f_{i_0}, f_i, \cdots) \in \Pi_m^d,$
\[
V_\beta \geq V_{\beta, n}(\pi) + \beta^n P(f_{i_0})P(f_i)\cdots P(f_n)V_\beta,
\]
which together with the finiteness of $V_\beta$ result in the lemma.

The following lemma is well-known and can be proved easily.

**Lemma 2:** For any policy $\pi = (\pi_0, \pi_1, \cdots) \in \pi$ and state $i, j \in S$, we define a Markov policy $\sigma = (\sigma_0, \sigma_1, \cdots) \in \Pi_m$ by
\[
\sigma_n(a | j) = P_{\pi}[\Delta_n = a | X_0 = i, X_n = j]
\]
where $(j, a) \in \Gamma, n \geq 0$. Then for any $n \geq 0$ and $(j, a) \in \Gamma$ we have
\[
P_{\sigma}[X_n = j, \Delta_n = a | X_0 = i] = P_{\pi}[X_n = j, \Delta_n = a | X_0 = i].
\]

This lemma says that the policies can be limited in region of Markov policies, $\Pi_m$, and
\[
V_\beta(i) = \sup_{\pi \in \Pi_m} V_\beta(\pi, i), \ i, j \in S.
\]  \hspace{1cm} (5)

We introduce a vector $TV$ for any vector $V$ by
\[
TV(i) = \sup_{a \in A(i)} \{r(i, a) + \beta \sum_{j} P_{\pi}(a) V(j)\}, \ i \in S.
\]
Then the optimality equation can be rewritten by $V = TV$. Certainly, $TV$ may be not defined, but $T0 (V = 0)$ and $TV_\beta (V = V_\beta)$ are well defined.

**Lemma 3:** If $V$ is a finite solution of the optimality equation, then

1. $V \leq V_\beta$ if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,
\[
\lim_{n \to \infty} \beta^n E_{\pi, i}, \{V(X_n)\} \leq 0, \ \forall i.
\]  \hspace{1cm} (6)

2. $V \geq V_\beta$ if and only if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,
\[
\limsup_{n \to \infty} \beta^n E_{\pi, i}, \{V(X_n)\} \geq 0, \ \forall i.
\]  \hspace{1cm} (7)

3. $V = V_\beta$ if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,
\[
\liminf_{n \to \infty} \beta^n E_{\pi, i}, \{V(X_n)\} = 0, \ \forall i.
\]  \hspace{1cm} (8)

So, if the optimality equation has a solution $V$ satisfying the above Eq. (8), then $V_\beta$
also satisfies Eq. (8).

Proof:

(1) Suppose that \( \{ \varepsilon_n \geq 0 \} \) with \( \sum_{n=0}^{\infty} \beta^n \varepsilon_n < \infty \) and \( \{ f_n \in F \} \) satisfy \( T_{s_n} V \geq TV - \varepsilon_n \) for \( n \geq 0 \).

Let a policy \( \pi = (f_0, f_1, \ldots) \), then we can prove by the induction method that for \( N \geq 0 \),

\[
\sum_{n=0}^{N} \beta^n E_{\pi_n} r(X_n, \Delta_n) + \beta^{N+1} E_{\pi_{N+1}} \{ V(X_{N+1}) \} = T_{s_0} T_{s_1} \cdots T_{s_N} V(i) \geq V(i) - \sum_{n=0}^{N} \beta^n \varepsilon_n.
\]

Taking \( \lim \inf \pi_n \rightarrow \infty \), we get that \( V_\beta \geq V_\beta(\pi) \geq V - \sum_{n=0}^{\infty} \beta^n \varepsilon_n \), which implies that \( V_\beta \geq V \) for the arbitrariness of \( \varepsilon_n \).

(2) Suppose that Eq. (7) is true. It follows \( V = TV \) that

\[
V(i) \geq r(i, a) + \beta \sum_{j} p_{ij}(a)V(j), \quad (i, a) \in \Gamma.
\]

Then for each policy \( \pi \) we have that

\[
V(i) \geq \sum_{n=0}^{\infty} \beta^n E_{\pi_n} r(X_n, \Delta_n) + \beta^{N+1} E_{\pi_{N+1}} V(X_{N+1}).
\]

Again by taking \( \lim \inf \pi_n \rightarrow \infty \), we know that \( V \geq V_\beta (\pi) \). So \( V \geq V_\beta \).

On the other hand, Eq. (7) can be obtained by \( V \geq V_\beta \) and Lemma 1.

(3) It follows (1) and (2).

By Lemma 1 we know that the fact of \( V_\beta \) satisfying Eq. (6) is equivalent to the fact of \( V_\beta \) satisfying Eq. (8). By this and Lemma 3, we have the following theorem about the uniqueness of solutions of the optimality equation.

**Theorem 1:**

(1) \( V_\beta \) is the smallest solution of the optimality equation satisfying Eq. (7).

(2) \( V_\beta \) is the unique solution of the optimality equation satisfying Eq. (8) (or Eq. (6) equivalently) if and only if the optimality equation has a solution \( V \) that satisfies Eq. (8).

In the literature (see [7]), it was said only that \( V_\beta \) is the maximal solution of the optimality equation satisfying Eq. (6), or is the smallest solution of the optimality equation satisfying Eq. (7), while here, \( V_\beta \) is the smallest solution satisfying Eq. (7) with no conditions, and is the unique one if there is a solution of the optimality equation satisfying Eq. (8).
4 Properties of Optimal Policies

Some properties of optimal policies will now be discussed. The following theorem, about the existence of \( \epsilon \)-optimal policies, follows immediately the proofs of the lemmas 1 and 2.

**Theorem 2:**

(1) Suppose that \( \{ \epsilon_n \geq 0 \} \) satisfies \( \sum_{n=0}^{\infty} \beta^n \epsilon_n < \infty \) and for each \( n, f_n \in F \) attains the \( \epsilon_n \)-supremum of the optimality equation. Then the policy \( \pi = (f_0, f_1, \ldots) \) is \( \sum_{n=0}^{\infty} \beta^n \epsilon_n \)-optimal if \( \pi \) satisfies the condition of Eq. (8) with \( V = V_\beta \).

(2) For any policy \( \pi, \pi \) is optimal if and only if \( V_\beta(\pi) \) is a solution of the optimality equation and satisfies the condition of Eq. (7) with \( V = V_\beta \) \( (V_\beta(\pi) \geq 0, \text{ or } r \geq 0 \) especially).

The first conclusion of Theorem 2 discusses the optimality of policies attaining the \( \epsilon \)-supremum of the optimality equation, while the second conclusion characterizes the optimality of a policy \( \pi \) by its objective value \( V_\beta(\pi) \).

The following corollary on \( \epsilon \)-optimal policies follows immediately (1) of Theorem 2.

**Corollary 1:** If \( \beta \in (0, 1), f \) attains the \( \epsilon(\geq 0) \)-supremum of the optimality equation and satisfies the condition of Eq. (6) with \( V = V_\beta \), then \( f \) is \( (1-\beta)^{-1} \epsilon \)-optimal. Moreover, when \( \epsilon = 0, f \) is optimal, irrespectively of the value of \( \beta \).

The following theorem shows the dominance of stationary policies.

**Theorem 3:**

(1) If \( V_\beta \) satisfies Eq. (8) for each \( \pi \in \Pi_{\alpha^d, (especially V_\beta \leq 0 or r \leq 0)} \), then \( V_\beta = \sup_{\pi \in \Pi_{\alpha^d}} V_\beta(\pi) \). Moreover, \( V_\beta = \sup_{f \in F} V_\beta(f) \) if \( \beta \in (0, 1) \) or there is \( f \in F \) attaining the supremum of the optimality equation.

(2) If there is a policy \( \pi \) such that \( V_\beta(\pi) \geq 0 \) \( (r \geq 0 \text{ or } V_\beta \geq 0 \) especially), then \( V_\beta = \sup_{f \in F} V_\beta(f) \).

**Proof:** (1) follows (1) of Theorem 2 and Corollary 1. About (2), since \( V_\beta \geq V_\beta(\pi) \geq 0 \), so for any \( i \in S \) with \( V_\beta(i) = 0, V_\beta(\pi, i) = V_\beta(i) = 0 \), which imply the result together with Theorem 1.1 in [8].

In the following, we discuss the structure of optimal policies under the restriction that all the action sets are countable. This restriction is not essential and is just to avoid measure theory. First we give the following two concepts.

**Definition 1:** A history \( h_n = (i_0, a_0, i_1, \cdots, a_{n-1}, i) \in H_n \) is called a realized history under policy \( \pi \) if the probability of \( h_n \) under policy \( \pi \) is positive, namely,

\[
\pi_0(a_0 | i_0) p_{a_0, i_1}(a_0) \cdots p_{a_{n-1}, (a_{n-1}) | i_0, a_0, i_1, \cdots, i_{n-1}} p_{a_{n-1}, (a_{n-1}) | i_0, a_0, i_1, \cdots, i_{n-1}} > 0.
\]
And \(k_{n-1} = (i_0, a_0, i_1, \ldots, a_{n-1})\) in \(h_n\) is also called a realized history if its probability under policy \(\pi\) is positive. For \(n = 0\), we always call \(h_0 = (i)\) to be realized.

**Definition 2:** For any policy \(\pi = (\pi_0, \pi_1, \ldots)\) and a history \(k_n\), we define a new policy \(\pi^{k_n} = (\pi'_0, \pi'_1, \ldots)\) by

\[
\pi'_m(\bullet | h_m) = \pi_{m+n+1}(\bullet | k_n, h_m), \quad m \geq 0, \quad h_m \in H_m.
\]

Obviously, \(\pi^{k_n}\) is exactly the remaining part of policy \(\pi\) when the history \(k_n\) has happened. For convenience, we denote for any policy \(\pi = (\pi_0, \pi_1, \ldots)\), state \(i\) and a function \(V\) on \(\Gamma\) that

\[
\pi_0 p_{\pi_1} p \cdots p_{\pi_n} V(i) = \sum_{a_0 \in A(i)} \pi_0(a_0 \mid i) \sum_{a_1 \in A(i)} \pi_1(a_1 \mid i, a_0) \cdots \sum_{a_{n-1} \in A(i)} \pi_{n-1}(a_{n-1} \mid i, a_0, a_1, \ldots, a_{n-2}) \pi_n(a_{n-1} \mid i, a_0, a_1, \ldots, a_{n-1}) \sum_{a_n \in A(i)} \pi_n(a_n \mid i, a_0, a_1, \ldots, a_{n-1}, a_{n-1}) V(i_{n+1})
\]

and similarly for \(\pi_0 p_{\pi_1} p \cdots p_{\pi_n-1} p V\) when \(V\) is a function on \(S\).

**Lemma 4:** \(\pi = (\pi_0, \pi_1, \ldots)\) is an optimal policy if and only if \(V_\beta(i) = V_\beta(\pi^{k_{n-1}}, i)\) for any realized history \(h_n = (k_{n-1}, i) = (i_0, a_0, i_1, \ldots, i_{n-1}, a_{n-1}, i)\) under policy \(\pi\).

**Proof:** The sufficiency is obvious only by taking \(n = 0\) and \(h_0 = (i)\). We now prove the necessity. Suppose that \(\pi\) is optimal, i.e., \(V_\beta(i) = V_\beta(\pi, i) = V_\beta(\pi^{k_{n-1}}, i)\). So the result for \(n = 0\) is true.

For \(n \geq 1\), we have

\[
V_\beta = V_\beta(\pi) = \sum_{r=0}^{\infty} \beta^r \pi_0 p_{\pi_1} p \cdots p_{\pi_{r-1}} p \pi_r r
\]

\[
= \sum_{r=0}^{\infty} \beta^r \pi_0 p \cdots p_{\pi_r} r + \beta^0 \sum_{r=n}^{\infty} \beta^{n-r} \pi_0 p \cdots p_{\pi_n} p_{\pi_{n-1}} p \cdots p_{\pi_r} r
\]

\[
= \sum_{r=0}^{\infty} \beta^r \pi_0 p \cdots p_{\pi_r} r + \beta^0 \pi_0 p_{\pi_{n-1}} p V_\beta(\pi^{k_{n-1}})
\]

\[
= \sum_{r=0}^{\infty} \beta^r \pi_0 p \cdots p_{\pi_r} r + \beta^0 \pi_0 p_{\pi_{n-1}} p V_\beta(\pi)
\]

\[
= V_\beta(\pi') \leq V_\beta
\]

where the first inequality follows the fact that \(\pi\) is optimal, while \(\pi' = (\pi_0, \pi_1, \ldots, \pi_{n-1}, \pi_0, \pi_1, \pi_2, \ldots)\) is a policy. So all the above inequalities should be equal and the first term results in that should be

\[
\pi_0 p_{\pi_{n-1}} p [V_\beta - V_\beta(\pi^{k_{n-1}})] = 0.
\]

Since \(\pi\) is optimal, the first term implies that \(V_\beta(i) = V_\beta(\pi^{k_{n-1}}, i)\) if \((k_{n-1}, i)\) is realized under
\( \pi \) for \( V_\beta \geq V_\beta(\pi^{k_{i-1}}) \).

We denote for \((i, a) \in \Gamma\) that
\[
G(i, a) = V_\beta(i) - \{ r(i, a) + \beta \sum_j p_{ij}(a)V_\beta(j) \} \geq 0.
\]
It represents the value deviated from the optimal value in state \( i \) if action \( a \in A(i) \) is chosen. The optimal action set \( A'(i) \) at state \( i \) is defined by
\[
A'(i) = \{ a \mid a \in A(i), G(i, a) = 0 \}, i \in S.
\]
That action \( a \in A(i) \) is optimal at state \( i \) means that no deviation from the optimal value occurs when choosing \( a \), or equivalently, optimal actions at state \( i \) are the actions that attain the supremum of the optimality equation.

**Lemma 5:** If there is an optimal policy \( \pi^* = (\pi_0^*, \pi_1^*, \cdots) \), then all the optimal action sets \( A'(i) \) are nonempty and \( \pi_0^*(A'(i) \mid i) = 1 \) for each \( i \in S \).

**Proof:** It has been shown already in the proof of Lemma 4 that \( V_\beta = \pi_0^* r + \beta \pi_0^* pV_\beta \). That is for \( i \in S \),
\[
\pi_0^* \{ V_\beta - (r + \beta pV_\beta) \}(i) = \sum_{a \in A(i)} \pi_0^*(a \mid i) G(i, a) = 0,
\]
which implies that \( \pi_0^*(A'(i) \mid i) = 1 \) for each \( i \in S \) since \( G(i, a) \geq 0 \) and \( \sum_{a \in A(i)} \pi_0^*(a \mid i) = 1 \). Certainly, all the optimal action sets \( A'(i) \) are nonempty.

Based on the above two lemmas 4 and 5, we can now give the structure of optimal policies, which characterizes the necessary and sufficient condition of the optimality of a policy.

**Theorem 4:** There is an optimal policy if and only if the optimal action set \( A'(i) \) is nonempty for each \( i \in S \). Under this condition, \( \pi^* = (\pi_0^*, \pi_1^*, \cdots) \) is optimal if and only if the following two conditions hold:

1. \( \pi_0^*(A'(i) \mid h_n) = 1 \) for any \( n \geq 0 \) and realized history \( h_n = (i_0, a_0, \cdots, a_{n-1}, i) \) under policy \( \pi^* \);
2. \( (\pi^*, V_\beta) \) satisfies condition of Eq. (8).

**Proof:** Necessity. Suppose that \( h_n = (k_{i-1}, i) = (i_0, a_0, \cdots, a_{n-1}, i) \) is a realized history under policy \( \pi^* \). Then it follows Lemma 4 that \( V_\beta(\pi^{k_{i-1}}, i) = V_\beta(i) \). Define a new policy \( \pi^{'k_{i-1}} \) by using the policy \( \pi^{k_{i-1}} \) if the initial state is \( i \), and using the policy \( \pi \) otherwise. Then
\[
V_\beta(\pi^{'k_{i-1}}, i) = \begin{cases} 
V_\beta(\pi, j) = V_\beta(j), & \text{if } j \neq i \\
V_\beta(\pi^{k_{i-1}}, i) = V_\beta(i), & \text{if } j = i.
\end{cases}
\]
So \( \pi^{'k_{i-1}} \) is also optimal, which together with Lemma 5 that implies that
\[ \pi_n(A'(i) \mid h_n) = \pi_0^{h_n+1}(A'(i) \mid i) = \pi_0(A'(i) \mid i) = 1. \]

Thus (1) of Theorem 4 is true, while (2) of Theorem 4 follows (2) of Theorem 2 and \( V_\beta = V_\beta(\pi) \).

 Sufficiency. From the proof of Lemma 4 we have

\[ V_\beta = \sum_{r=0}^{\infty} \beta^r p \pi_0 p \cdots \pi_r r + \beta^n p \pi_0 p \cdots \pi_{n-1} p V_\beta, \]

in which by letting \( n \to \infty \) we can know that \( V_\beta = V_\beta(\pi) \). So \( \pi \) is optimal.

 The condition (1) of Theorem 4 is equivalent to say that \( \pi \) chooses only the optimal actions in the realized history, or equivalently, all actions in any realized history belong to the optimal action set \( a_\pi \in A'(i_\pi) \). From Theorem 4, we know that the set of all optimal stationary policies is exactly \( \times_{i} A'(i) \). We can say that the conditions (1) and (3) of Theorem 4 characterize the structure of the set of all optimal policies.

 From Theorem 4, all properties of optimal policies, e.g., the convex combination of any optimal policy, discussed in the literature (see [5]), can be obtained easily. The details are omitted here.

 Finally, the following corollary follows Theorem 4.

**Corollary 2:** If there exists an optimal policy, then there must exist optimal stationary policies and \( V_\beta(i) = \sup_f V_\beta(f, i), i \in S \).

### 5 State Feedback Control of DES

State feedback control is a branch in Supervisory control of discrete event systems (DES for short), which was first presented in [9] and [10].

A discrete event system based on automaton is \( \{ Q, \Sigma, \delta, q_0 \} \), where \( Q \) is a countable state space, \( \Sigma \) is a finite event set with \( N = |\Sigma| \), \( \delta \) is a partial function from \( \Sigma \times Q \) to \( Q \), while \( q_0 \in Q \) is the initial state. We denote by \( \delta(\sigma, q) \) if it is well defined, and \( \Sigma(q) = \{ \sigma \mid \delta(\sigma, q) \} \).

The event set \( \Sigma \) is further divided into a uncontrolled event set \( \Sigma_u \) and a controlled event set \( \Sigma_c \). A control input is a \( \gamma \) satisfying \( \Sigma_u \subset \gamma \subset \Sigma \), the set of which is denoted by \( \Gamma \). Now a state feedback control is defined as a map \( f : Q \to \Gamma \). For each \( f \), the controlled system \( f/G \) is defined as a DES \( \{ Q, \Sigma, \delta, q_0 \} \) where \( \delta(\sigma, q) = \delta(\sigma, q) \) is well defined only if \( \sigma \in f(q) \) and \( \delta(\sigma, q) \). Let \( R(f/G) \) be the reachable states set under \( f/G \).

We call subsets of \( Q \) predicates. A predicate \( P \) is said to be control invariant if \( \delta(\sigma, q) \in P \) for each \( q \in P \) and \( \sigma \in \Sigma_u \) with \( \delta(\sigma, q) \). This is equivalently to say that there is a state feedback control \( f \) such that \( \delta(\sigma, q) \in P \) for each \( q \in P \) and \( \sigma \in \Sigma_u \) with \( \delta(\sigma, q) \) and \( \sigma \in f(q) \). Such an \( f \) is said to be permissive for \( P \), the set of which is denoted by \( F(P) \).
Moreover, $P$ is said to be controllable if there is $f$ such that $P = R(f/G)$.

A basic problem in the state feedback control for DES is that for a given predicate $P$ to find an $f$ such that

$$\max_i R(f/G), \quad \text{s.t.} \ R(f/G) \subseteq P.$$  \hspace{1cm} (9)

We will use Markov decision processes to solve this problem. First, we define a Markov decision process model with the state space $Q$ and action set $\Gamma$, while the reward function $r(q) = \chi^\rho(q) - 1$ is irrespective of the action, and the state transition probability is $p_{\rho \gamma}(q) = 1/N$ if $q' = \delta(\sigma, q)$ for some $\sigma \in \gamma$ and $p_{\rho \gamma}(q) = 0$ otherwise.

Surely, a state feedback control is exactly a stationary policy in Markov decision processes. Then its optimality equation Eq. (3) is

$$V_\beta(q) = \chi^\rho(q) - 1 + \beta \max_\gamma \frac{1}{N} \sum_{\sigma \in \nu(q)} V_\beta(\delta(\sigma, q)), \quad q \in Q$$  \hspace{1cm} (10)

where $\gamma(q) = \gamma \cap \Sigma(q)$. Now we let $P' = \{q \mid V_\beta(q) = 0\}$, $A'(q)$ the optimal action sets, and $F' = \times_q A'(q)$. Then we have the following result.

**Theorem 5:**

(1) $P'$ is the maximal control invariant sub-predicate of $P$, $F' = F(P')$.

(2) For each $q \in P'$, $A'(q)$ has the unique maximal element and

$$f'(q) = \{\sigma \mid \delta(\sigma, q) \text{ is undefined, or } \sigma \in P'\}, \quad q \in P'$$

and $R(f'/G)$ is the maximal controllable sub-predicate of $P$. So $f'$ is the optimal solution of problem Eq. (9).

**Proof:**

(1) Since $V_\beta \leq 0$, we know from Eq. (10) that for $q \in P'$, $r(q) \geq V_\beta(q) = 0$, which implies that $r(q) = 0$ and so $q \in P$. Thus $P' \subseteq P$.

Now, $q \in P'$ iff there is $\gamma$ such that $V_\beta(\delta(\sigma, q)) = 0$ for all $\sigma \in \gamma(q)$, which is equivalent to that $\delta(\sigma, q) \in P'$ for all $\sigma \in \gamma(q)$. Thus for any $f \in F'$, if $q \in P'$ and $\sigma \in f'(q)$ with $\delta(\sigma, q) \neq \gamma$, then $\delta(\sigma, q) \in P'$. So $P'$ is control invariant.

Suppose that $P' \subset P$ is also control invariant, then from Eq. (10) we have that for $q \in P'$, $V_\beta(q) \geq (\beta/N) \sum_{\sigma \in \gamma(q)} V_\beta(\delta(\sigma, q))$. These together with the induction method can show that $V_\beta(q) \geq 0$ for $q \in P'$. So $P' \subseteq P'$, that is, $P'$ is the maximal control invariant sub-predicate of $P$.

(2) It is apparent that $f'$ is the maximal one, while the maximum of $R(f'/G)$ follows (1) and the definition of $f'$.

Now we consider a numerical example, which was first presented for partially observable DES in [11]. Here we consider it for complete observable case and will use
Theorem 6 to solve it. The state variable is $x = (x_1, x_2, x_3)$ and the event set is $\Sigma = \Sigma_e = \{\alpha, \beta, \theta\}$, so it has no uncontrollable events. While the state transition function is described by

$$
\delta(\alpha, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (-1, 1, 0), \text{ if } x_1 \geq 1,
$$

$$
\delta(\beta, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (-1, 1, 1), \text{ if } x_1 \geq 1,
$$

$$
\delta(\theta, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (0, -1, 1), \text{ if } x_2 \geq 1,
$$

with the initial state $x^0 = (1, 0, 0)$.

We consider the given predicate $P = \{x \mid x_1 \leq 1\}$. Since for each $x$, either $V_\beta(x) = 0$ or $V_\beta(x) \leq 0$, and there is no uncontrolled events, the maximum term in the optimality equation Eq. (10) becomes

$$
\max_{\gamma} \sum_{\sigma \gamma(x)} V_\beta(\delta(\sigma, x)) = \sum_{\sigma \Sigma, V_\beta(\delta(\sigma, x)) = 0} V_\beta(\delta(\sigma, x)) = 0.
$$

Thus

$$
V_\beta(x) = \chi_P(x) - 1, \quad A^*(x) = \{\sigma \in \Sigma \mid \delta(\sigma, x) \in P\}, \quad x \in Q.
$$

(11)

So from Theorem 6 we know that the maximal control invariant sub-predicate of $P$ is itself, i.e., $P^* = P$, and the maximal permissive state feedback control of $P$ is $f^*(x) = A^*(x)$, while the maximal controllable sub-predicate of $P$ is $R(f^*/G)$. In fact,

$$
R(f^*/G) = \{(1, 0, 0), (0, 1, 0), (0, 1, 1), (0, 0, 1)\},
$$

$$
f^*(1, 0, 0) = f^*(0, 1, 0) = f^*(0, 0, 1) = \{\alpha, \beta, \theta\},
$$

$$
f^*(0, 1, 1) = \{\alpha, \beta\}.
$$

6 Conclusions

In this paper, we have shown that as a solution of the optimality equation, the solution called optimal value function is always the smallest one and is the unique one motions under weak conditions. Furthermore, the structure of optimal policies is given. Finally, these properties are applied to the state feedback control of discrete event systems with a numerical example.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 69904008) and in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 13650440), Japan.
References


