九州大学学術情報リポジトリ
Kyushu University Institutional Repository

# ON EFFICIENCY OF INFERENCES OF TRANSITION PROBABILITIES TO MARKOVIAN DECISION PROCESSES

Kobayashi, Kingo
Department of Mathematics, Osaka Dental University

Fujikawa, H.
Faculty of Engineering Science, Osaka University

Kurano, Masami
Department of Mathematics, Faculty of Education, Chiba University

https://doi.org/10.5109/13077

出版情報：統計数理研究. 15 (3/4), pp.113-122, 1973-03. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

# ON EFFICIENCY OF INFERENCES OF TRANSITION PROBABILITIES TO MARKOVIAN DECISION PROCESSES

By

Y. KOBAYASHI*, H. FUJIKAWA** and M. KURANO***

(Received October 15, 1972)

## § 1. Introduction.

Markovian Decision Processes (M. D. P's) in case that transition probabilities are known to a decision maker have been investigated by many authors (for example [1], [2]).

However, it happens to a decision maker that transition probabilities, factors that M. D. P's are determined by, are unknown to him. Under such circumstances we must gather data useful in estimating the unknown transition probabilities.

We have an interest that if our efforts of gathering-information and estimation is good for the improvement of the circumstances for M. D. P's.

We shall define True Markovian Decision Processes (T. M. D. P's) and Estimated Markovian Decision Processes (E. M. D. P's), and we shall give the answer of the problem both in the discounted case and in the average case. Also, we shall give the example.

## § 2. Markovian Decision Processes.

In this section, we shall give the definitions and notations on a class of M. D. P's. We follow [1] and [3] closely as possible.

By a Borel set we mean a Borel subset of some complete separable metric space. The set of all probability measures on a non-empty Borel set $X$ is denoted by $P(X)$. For any non-empty Borel sets $X$, $Y$, a conditional probability on $Y$ given $X$ is a function $q(\cdot/\cdot)$ such that for each $x \in X$, $q(\cdot/x)$ is a probability measure on $Y$ and for each Borel set $B \subset Y$, $q(B/\cdot)$ is a Baire function on $X$.

The set of all conditional probabilities on $Y$ given $X$ is denoted by $Q(Y/X)$. The set of bounded Baire functions on $X$ is denoted by $M(X)$. The product space of $X$ and $Y$ will be denoted by $XY$.

For any $u \in M(XY)$ and any $q \in Q(Y/X)$, $qu$ denotes the element of $M(X)$ whose value of $x_0 \in X$ is

---

* Department of Mathematics, Osaka Dental University.

** Faculty of Engineering Science, Osaka University, Toyonaka, Osaka.

*** Department of Mathematics, Faculty of Education, Chiba University.

$$qu(x_0)=\int_Y u(x_0, y))dq(y/x_0)\,.$$

For any $p\in P(X)$ and any $u\in M(X)$, $pu$ is the integral of $u$ with respect to $p$. For any $p\in P(X)$, $q\in Q(Y/X)$, $pq$ is the probability measure on $XY$ such that, for every $u\in M(XY)$, $pq(u)=p(qu)$.

We extend the above notation in an obvious way to a finite or countable sequence of non-empty Borel sets $X_1, X_2, \cdots$. If $q_n\in Q(X_{n+1}/X_1\cdots X_n)$ for $n\geqq 1$ and $p\in P(X_1)$, $pq_1\cdots q_n$ is a probability measure on $X_1X_2\cdots X_{n+1}$, $pq_1q_2\cdots$ is a probability measure on the infinite product space $X_1X_2\cdots$, for any $u\in M(X_1X_2\cdots X_{n+1})$, $n\geqq 1$ and any $m$, $1\leqq m\leqq n$, $q_m\cdots q_n\ u\in M(X_1\cdots X_m)$, etc.

A $p\in P(X)$ is degenerate if it is concentrated at some one point $x\in X$; a $q\in Q(Y/X)$ is degenerate if each $q(\cdot/x)$ is degenerate. The degenerate $q$ are exactly those for which there is a Baire function $f$ mapping $X$ into $Y$ for which $q(\{f(x)\}/x)=1$ for all $x\in X$. Any such $f$ will also denote its associated degenerate $q$, so that, for any $u\in M(XY)$, $fu(x)=u(x, f(x))$ for all $x\in X$. Markovian Decision Processes are controlled dynamic systems defined by $S$, $A$, $Q$, $r$, where $S$, $A$ are any non-empty Borel sets, $Q\in Q(S/SA)$, $r\in M(SAS)$.

A policy $\pi$ is a sequence $(\pi_1, \pi_2, \cdots)$, where $\pi_n\in Q(A/H_n)$ and $H_n=SA\cdots AS$ ($2n-1$ factors) is the set of possible histories of the system when the $n$-th act must be chosen. $\pi=(\pi_1, \pi_2, \cdots)$ is a non-randomized stationary policy if each $\pi_n$ is a degenerate element of $Q(A/S)$, and if there is a Baire function $f$ mapping $S$ into $A$ such that $\pi_n=f$ for all $n$. The non-randomized stationary policy defined by $f$ is denoted by $f^{(\infty)}$.

We interpret $S$ as the set of states of some system, and $A$ as the set of actions available at each state. When the system is in state $s$ and we take action $a$, we move to a new state $s'$ selected according to $Q(\cdot/s, a)$ and we recieve a return $r(s, a, s')$. The process is then repeated from the new state $s'$, and we wish to maximize the total expected return over the infinite future.

Any policy $\pi$, together with the law of motion $Q$ of the system, defines for each intial state $s$ a conditional distribution on the set $\Omega=ASAS\cdots$ of futures of the system, i.e. it defines an element of $Q(\Omega/S)$, namely, $e_\pi=\pi_1q\pi_2q\cdots$. Denote the coordinate function on $S\Omega$ by $s_1, a_1, s_2, a_2, \cdots$, and our return on the $n$-th day is $r(s_n, a_n, s_{n+1})$.

The total expected return starting from $s$ and using $\pi$ may well be infinite. There are, however, two case in which the problem is well defined, which may be described as follows:

(1) The discounted case.

We discount our future return with a discount factor $\alpha$, $0\leqq\alpha<1$, so that a return of one unit $n$ stages in the future is worth $\alpha^n$ now. The expected total discounted return from $\pi$, as a function of the initial state, is

$$\phi(s_1, \pi)=e_\pi u=\sum_{n=1}^{\infty}\alpha^{n-1}\pi_1Q\cdots\pi_nQr$$

where $u=\sum_{n=1}^{\infty}\alpha^{n-1}r(s_n, a_n, s_{n+1})\,.$

(2) The average case.

We give the expected average return per unit day over the infinite future. Let

$$\Psi^{(T)}(s_1, \pi) = \frac{1}{T} \sum_{n=1}^{T} \pi_1 Q \cdots \pi_n Q r ,$$

thus, the expected average return per unit day over the infinite future from $\pi$, as a function of the initial state, is

$$\Psi(s_1, \pi) = \lim_{T \to \infty} \inf \Psi^{(T)}(s_1, \pi) .$$

## § 3. Estimated and True Markovian Decision Processes.

In this section we shall formulate our problem investigated in this paper. The Markovian Decision Problem is determined by four objects, $S, A, Q$ and $r$. We assume that $S, A$ and $r$ are known to a decision maker but that $Q$ is unknown to him.

Let $Q_0$ be the true value of the unknown transition probability and let $Q_n$ ($n = 1, 2, \cdots$) be the estimated value of the unknown transition probabilities.

Markovian Decision Processes determined by $S, A, r, Q_0$, will be called True Markovian Decision Processes and denoted by T. M. D. P ($S, A, r, Q_0$).

We shall call the Markovian Decision Processes determined by $S, A, r, Q_n$ by the Estimated Markovian Decision Processes and denote it by E. M. D. P. ($S, A, r, Q_n$).

A specific question then arises: How do we construct the good policy for T. M. D. P ($S, A, r, Q_0$) from E. M. D. P ($S, A, r, Q_n$).

In section 4, we shall give an answer of the above question both in the discounted case and in the average case. In section 5, we shall give an example (Automobile Replacement Problem).

## § 4. Efficiency of E. M. D. P ($S, A, r, Q_n$) to T. M. D. P ($S, A, r, Q_0$).

Our main results will be taken in this section. A condition that we shall need to assume is that

(A) In T. M. D. P ($S, A, r, Q_0$), there is a non-randomized stationary optimal policy, denoted by $f_0^{(\infty)} = (f_0, f_0, \cdots)$, respectively, in the discounted case and in the average case.

The expected total discounted return from $\pi$, as a functional of the initial state, in T. M. D. P ($S, A, r, Q_0$) will be denoted by $\phi_0(\cdot, \pi)$, and the expected average return per unit day over the infieite future in T. M. D. P ($S, A, r, Q_0$) will be denoted by $\Psi_0(\cdot, \pi)$.

Suppose

(B) In each of E. M. D. P ($S, A, r, Q_n$) ($n = 1, 2, \cdots$), there is a non-randomized stationary optimal policy, denoted by $f_n^{(\infty)} = (f_n, f_n, \cdots)$, respectively in the discounted case and in the average case.

The expected total discounted return with a discounted factor $\alpha$, as a function of the initial state, in E. M. D. P ($S, A, r, Q_n$) will be denoted $\phi_n(\cdot, \pi)$, and the expected average return per unit day in E. M. D. P $(S, A, r, Q_n)$ will be denoted by $\Psi_n(\cdot, \pi)$.

REMARK. When action space $A$ is finite, condition (A) and (B) are sstisfied in the discounted case, and, when both action space $A$ and state space $S$ are finite, condition (A) and (B) are satisfied in the average case. (see [1], [2], [4]).

Let $q_n(s'/s, a)$ be the probability density function of $Q_n(\cdot/s, a)$ with respect to the Lebesque-measure $\mu$ for $n=0, 1, 2, \cdots$. Let

$$C_n = \sup_{\substack{s \in S \\ a \in A}} \int |q_n(s'/s, a) - q_0(s'/s, a)| d\mu(s') .$$

We shall make use of the following conditions.

(D) (Regularity Condition)

$$C_n \longrightarrow 0 \qquad (\text{as } n \to +\infty) .$$

### 4.1. Discounted Case $\alpha \in (0, 1)$.

In this sub-section, we shall treat the discounted case. Let $f_0^{(\infty)}$ and $f_n^{(\infty)}$, respectively, be optimal policies of T. M. D. P $(S, A, r, Q_0)$ and E. M. D. P $(S, A, r, Q_n)$ in the discounted case.

Our main results is stated in the following theorem.

THEOREM 1. *Under conditions* (A), (B) *and* (C), *as* $n \to \infty$, $\phi_0(s_1, f_n^{(\infty)})$ *converges to* $\phi_0(s_1, f_0^{(\infty)})$ *in uniform for* $s_1 \in S$.

The above theorem says that if we use the policy $f_n^{(\infty)}$ taken by the E. M. D. P $(S, A, r, Q_n)$ in the place of the $f_0^{(\infty)}$, the effort of the inference concerning the unknown transition probabilities is rewarded.

Three lemmas will be given to prove the theorem 1.

The set of the measureable function mapping $S$ into $A$ is denoted by $F(S: A)$. Associated with each $f \in F(S: A)$ and each $n$ $(n=0, 1, 2, \cdots)$ are two corresponding operators $T_{(f,n)}$ and $r_{(f,n)}$, mapping $M(S)$ into $M(S)$, denoted as following. For $u \in M(S)$,

$$T_{(f,n)}u(s) = \alpha \int_S u(v) q_n(v/s, f(s)) d\mu(v) ,$$

$$r_{(f,n)}(s) = \int_S r(s, f(s), v) q_n(v/s, f(s)) d\mu(v) .$$

LEMMA 1. (i) *For any* $f \in F(S: A)$ *and any* $n$, $T_{(f,n)}$ *is a linear operator on* $M(S)$ *and* $\|T_{(f,n)}\| \leq \alpha < 1$.

(ii) *Under condition* (C), *for any* $f \in F(S: A)$,

$$T_{(f,n)} \longrightarrow T_{(f,0)} \qquad (as\ n \to +\infty) ,$$

*and*

$$r_{(f,n)} \longrightarrow r_{(f,0)} \qquad (as\ n \to +\infty) .$$

PROOF. (i) is immediate.

For (ii),

$$|T_{(f,n)}u(s)-T_{(f,0)}u(s)|\leqq\|u\|C_n$$

and

$$|r_{(f,n)}u(s)-r_{(f,0)}u(s)|\leqq\|r\|C_n \qquad \text{for any} \quad u\in M(S)\,,$$

where $\|u\|=\sup_{s\in S}|u(s)|$ for any $u\in M(S)$.

This completes the proof.

LEMMA 2. *Under condition* (C), *as* $n\to\infty$, $\phi_n(s,f^{(\infty)})$ *converges to* $\phi_0(s,f^{(\infty)})$ *in uniform for any* $s\in S$ *and any* $f\in F(S:A)$.

PROOF. For any $n$ $(n=0,1,2,\cdots)$,

$$\phi_n(s,f^{(\infty)})=\sum_{k=0}^{\infty}T^{(k)}_{(f,n)}r_{(f,n)}(s)\,.$$

By Lemma 1,

$$\phi_n(s,f^{(\infty)})=(I-T_{(f,n)})^{-1}r_{(f,n)}(s)$$

where $I$ is an identity operator.

Also, by Lemma 1,

$$\begin{aligned}\|(I-T_{(f,n)})^{-1}u-(I-T_{(f,0)})^{-1}u\|&\leqq\|T_{(f,n)}u-T_{(f,0)}u\|\cdot\|(I-T_{(f,n)})^{-1}\|\cdot\|(I-T_{(f,0)})^{-1}\|\\&\leqq\frac{1}{(1-\alpha)^2}\|T_{(f,n)}-T_{(f,0)}\|\\&\leqq\frac{\|u\|}{(1-\alpha)^2}C_n \qquad \text{for any} \quad u\in M(S)\,.\end{aligned}$$

Therefore, since $\|r_{(f,n)}-r_{(f,0)}\|\leqq\|r\|C_n$,

$$\begin{aligned}\|\phi_n(s,f^{(\infty)})-\phi_0(s,f^{(\infty)})\|&=\|(I-T_{(f,n)})^{-1}r_{(f,n)}(s)-(I-T_{(f,n)})^{-1}r_{(f,0)}(s)\\&\quad+(I-T_{(f,n)})^{-1}r_{(f,0)}(s)-(I-T_{(f,0)})^{-1}r_{(f,0)}(s)\|\\&\leqq\frac{1}{1-\alpha}\|r_{(f,n)}-r_{(f,0)}\|+\|r\|\cdot\|(I-T_{(f,n)})^{-1}-(I-T_{(f,0)})^{-1}\|\\&\leqq\|r\|\Big(\frac{1}{(1-\alpha)}+\frac{1}{(1-\alpha)^2}\Big)C_n\,.\end{aligned}$$

This completes the proof.

LEMMA 3. *Under condition* (C), *as* $n\to+\infty$, $\phi_n(s,f_n^{(\infty)})$ *converges to* $\phi_0(s,f_0^{(\infty)})$ *in uniform for* $s\in S$.

PROOF. By Lemma 2, for any $\varepsilon>0$, there is a $N=N(\varepsilon)$ such that

$$\|\phi_n(\cdot,f^{(\infty)})-\phi_0(\cdot,f^{(\infty)})\|\leqq\varepsilon \qquad \text{for any} \quad n\geqq N,$$

in uniform for $f\in F(S:A)$.

Since the definition $f_n^{(\infty)}$ and the above result,

$$\begin{aligned}\phi_0(s,f_0^{(\infty)})-\varepsilon&\leqq\phi_n(s,f_0^{(\infty)})\leqq\phi_n(s,f_n^{(\infty)})\\&\leqq\phi_0(s,f_n^{(\infty)})+\varepsilon\leqq\phi_0(s,f_0^{(\infty)})+\varepsilon\,.\end{aligned}$$

Therefore,

$$\|\phi_n(s,f_n^{(\infty)})-\phi_0(s,f_0^{(\infty)})\|\leqq\varepsilon \qquad \text{for any} \quad n\geqq N.$$

As $n\to\infty$ and $\varepsilon\to0$, Lemma 3 is proved.

PROOF OF THEOREM.

$$\|\phi_0(\cdot, f_n^{(\infty)})-\phi_0(\cdot, f_0^{(\infty)}\| \leqq \|\phi_0(\cdot, f_n^{(\infty)})-\phi_n(\cdot, f_n^{(\infty)})\|+\|\phi_n(\cdot, f_n^{(\infty)})-\phi_0(\cdot, f_0^{(\infty)})\| .$$

As $n \to +\infty$, the first term on the right converges to 0 by Lemma 2 and the second converges to 0 by Lemma 3. This completes the proof of the theorem.

### 4.2. Average Case.

In this sub-section, we shall treat the average case. Let both $S$ and $A$ be finite after this from now.

We shall make use of the following conditions.

(D) The Markov chain induced by any non-randomized stationary policy $f^{(\infty)}$ is completely ergodic. Let

$$C_n' = \underset{\substack{s', s \in S \\ a \in A}}{\text{Max}} |q_n(s'/s, a)-q_0(s'/s, a)| .$$

(E) (Regularity condition)

$$C_n' \longrightarrow 0 \qquad (\text{as } n \to \infty) .$$

Let $f_0^{(\infty)}$ and $f_n^{(\infty)}$, respectively, be optimal policies of T.M.D.P $(S, A, Q_0, r)$ and E.M.D.P $(S, A, Q_n, r)$.

We can state the following theorem.

THEOREM 2. *Under conditions* (D) *and* (E), *as* $n \to \infty$, $\phi_0(s, f_n^{(\infty)})$ *converges to* $\Psi_0(s, f_0^{(\infty)})$.

Two Lemmas will be given to prove the theorem. We associate with each $f \in F(S: A)$.

(1) a $N \times 1$ column vector $r_n(f)$ whose $s$-th element is $r_{(f,n)}(s)$.

(2) a $N \times N$ stochastic matrix $Q_n(f)$ whose $(s, s')$ element is $q_n(s'/s, f(s))$.

Then, for each $n$ $(n=0, 1, 2, \cdots)$,

$$\Psi_n(s, f) = X_n^t(f) r_n(f)$$

where $X_n$ is a $N \times 1$ column vector and a stationary absolute probability, and $X_n$ is uniquely determined as a solution of

$$X_n(f) = Q_n^t(f) X_n(f) \qquad \text{under condition (E)} .$$

For $N \times 1$ vector sequence $\{u_n\}$ and $N \times 1$ vector $u$, that $u_n$ converges to $u$ means that each element of $u_n$ converges to the same element of $u$.

LEMMA 4. *Under condition* (D) *and* (E), *as* $n \to \infty$,

(i) *for any* $f \in F(S: A)$, $X_n(f) \to X_0(f)$,

(ii) *for any* $f \in F(S: A)$, $\Psi_n(s, f^{(\infty)}) \to \Psi_0(s, f^{(\infty)})$.

PROOF. For (i), suppose that $X^*$ is a limit point of $\{X_n(f)\}$, and there exists a subsequence $\{X_{n_j}(f)\}$ such that $X_{n_j}$ converges to $X^*$. Since $X_{n_j} = Q_{n_j}^t(f) X_{n_j}$, $X^* = Q_0^t(f) X^*$. Because of the uniqueness of a solution, $X^* = X_0(f)$.

For (ii), since $\Psi_n(s, f) = X_n^t(f) r_n(f)$, $\Psi_0(s, f) = X_0^t(f) r_0(f)$ and $r_n(f)$ converges to $r_0(f)$, $\Psi_n(s, f)$ converges to $\Psi_0(s, f)$.

This completes the proof.

LEMMA 5. *Under conditions* (D) *and* (E), *as* $n \to \infty$, $\Psi_n(s, f_n^{(\infty)})$ *converges to* $\Psi_0(s, f_0^{(\infty)})$.

PROOF. There exists a sub-sequence $\{f_{n_j}\}$ and $f$ such $f_{n_j} = f$ for all $j$. Since $\Psi_n(s, f_{n_j}^{(\infty)}) = \Psi_n(s, f^{(\infty)}) \geqq \Psi_n(s, f_0^{(\infty)})$, by Lemma 4, as $n \to \infty$ $\Psi_0(s, f^{(\infty)}) \geqq \Psi_0(s, f^{(\infty)})$.

On the other hand, $\Psi_0(s, f^{(\infty)}) \leqq \Psi_0(s, f_0^{(\infty)})$. Therefore, $\Psi_0(s, f^{(\infty)}) = \Psi_0(s, f_0^{(\infty)})$. This completes the proof.

PROOF OF THEOREM 2.

$$|\Psi_0(s, f_n^{(\infty)}) - \Psi_0(s, f_0^{(\infty)})| \leqq |\Psi_n(s, f_n^{(\infty)}) - \Psi_0(s, f_0^{(\infty)})| + |\Psi_n(s, f_n^{(\infty)}) - \Psi_0(s, f_n^{(\infty)})| .$$

The first term on the right converges to 0 by Lemma 5, and the second term converges to 0 by Lemma 4. This completes the proof.

## § 5. **Example** (Automobile Replacement Problem).

Let us consider the problem of Automobile Replacement over a time interval of six years. We agree to review our current situation every six months and to make a decision whether to keep our present car or to trade it in at that time. The state of the system, $s$, is decribed by the age of the car in six month periods; $s$ may run from 0 to 12 and the State Space is $S = \{0, 1, \cdots, 12\}$. The immediate return function $r$ is Table 1. We denote by $X = X(s, s')$ the transition probability that the car is in age $s'$ after six months, given that the car is in age $s$ at that time. Then,

$$q(s'/s, a) = X(a, s') \quad \text{for } s', \qquad s \in S \text{ and } a \in A .$$

We denote by $q(X)$ the transition probability determined by $X$. We will change $X$ step by step for the purpose of illustrating out results. The first step $X^{(1)}$ is Table 2.

The $n$-th step $X^{(n)}$, for $n = 2, 3, \cdots$, is determined iteratively by the following way;

$$\text{(i)} \qquad X^{(n)}(s, s') = X^{(n-1)}(s, s') \qquad \text{for} \quad s = 10, 11, 12, \quad s' = 0, 1, \cdots, 12 .$$

$$\text{(ii)} \qquad X^{(n)}(s, s') = \frac{(12 - s' + 1) X^{(n-1)}(s, s')}{\sum_{s' > s} (12 - s' + 1) X^{(n-1)}(s, s')} , \qquad \text{if} \quad s' > s ,$$

$$X^{(n)}(s, s') = 0 , \qquad \text{if} \quad s' \leqq s \qquad \text{for} \quad s = 0, 1, \cdots, 9 .$$

Then, $X^{(2)}$ and $X^{(30)}$ are, respectively, Table 3 and Table 4.

We find the optimal policy in Average Case, $f_n$, of E. M. D. P $(S, A, r, q(X^{(n)}))$ by using Howard's Policy Iteration Algorithm. The optimal policies, $\{f_n, n = 1, 2, \cdots, 30\}$, are Table 5.

Table 5 shows that $f_n = f_{30}$ for $n = 2, 3, \cdots, 29$. In other words, this example is dependent upon a tendency of the distribution of Estimated transition Probability for the determination of the optimal policy of T. M. D. P $(S, A, r, q(X^{(30)}))$.

We conjecture that the determination of the optimal policy of Markovian Decision Processes is generally more dependent on a return function than on a transition probability.

Table 1. (Immediate Return Function $r(s, a)$).

| $s \backslash a$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -31.0 | 10.5 | 46.0 | 52.0 | 56.5 | 62.5 | 66.0 | 74.0 | 72.0 | 93.0 | 94.0 | 8.0 | -200.0 |
| 1 | -62.0 | -20.5 | 15.0 | 21.0 | 25.5 | 31.5 | 35.0 | 43.0 | 41.0 | 62.0 | 63.0 | -23.0 | -200.0 |
| 2 | -82.5 | -41.0 | -5.5 | .5 | 5.0 | 11.0 | 14.5 | 22.5 | 20.5 | 41.5 | 42.5 | -43.5 | -200.0 |
| 3 | -88.0 | -46.5 | -11.0 | -5.0 | -.5 | 5.5 | 9.0 | 17.0 | 15.0 | 36.0 | 37.0 | -49.0 | -200.0 |
| 4 | -93.0 | -51.5 | -16.0 | -10.0 | -5.5 | .5 | 4.0 | 12.0 | 10.0 | 31.0 | 32.0 | -54.0 | -200.0 |
| 5 | -98.5 | -57.0 | -21.5 | -15.5 | -11.0 | -5.0 | -1.5 | 6.5 | 4.5 | 25.5 | 26.5 | -59.5 | -200.0 |
| 6 | -103.5 | -62.0 | -26.5 | -20.5 | -16.0 | -10.0 | -6.5 | 1.5 | -.5 | 20.5 | 21.5 | -64.5 | -200.0 |
| 7 | -110.0 | -68.5 | -33.0 | -27.0 | -22.5 | -16.5 | -13.0 | -5.0 | -7.0 | 14.0 | 15.0 | -71.0 | -200.0 |
| 8 | -115.0 | -73.5 | -38.0 | -32.0 | -27.5 | -21.5 | -18.0 | -10.0 | -12.0 | 9.0 | 10.0 | -76.0 | -200.0 |
| 9 | -127.0 | -85.5 | -50.0 | -44.0 | -39.5 | -33.5 | -30.0 | -22.0 | -24.0 | -3.0 | -2.0 | -88.0 | -200.0 |
| 10 | -130.0 | -88.5 | -53.0 | -47.0 | -42.5 | -36.5 | -33.0 | -25.0 | -27.0 | -6.0 | -5.0 | -91.0 | -200.0 |
| 11 | -135.0 | -93.5 | -58.0 | -52.0 | -47.5 | -41.5 | -38.0 | -30.0 | -32.0 | -11.0 | -10.0 | -96.0 | -200.0 |
| 12 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 | -200.0 |

Table 2. (Transition Probability $X^{(1)}(s, s')$ where $q(s'/s, a) = X^{(1)}(a, s')$).

| $s \backslash s'$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .0000 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 | .0833 |
| 1 | .0000 | .0000 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 | .0909 |
| 2 | .0000 | .0000 | .0000 | .1000 | .1000 | .1000 | .1000 | .1000 | .1000 | .1000 | .1000 | .1000 | .1000 |
| 3 | .0000 | .0000 | .0000 | .0000 | .1111 | .1111 | .1111 | .1111 | .1111 | .1111 | .1111 | .1111 | .1111 |
| 4 | .0000 | .0000 | .0000 | .0000 | .0000 | .1250 | .1250 | .1250 | .1250 | .1250 | .1250 | .1250 | .1250 |
| 5 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .1429 | .1429 | .1492 | .1429 | .1429 | .1429 | .1429 |
| 6 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .1667 | .1667 | .1667 | .1667 | .1667 | .1667 |
| 7 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .2000 | .2000 | .2000 | .2000 | .2000 |
| 8 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .2500 | .2500 | .2500 | .2500 |
| 9 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .3333 | .3333 | .3333 |
| 10 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .5000 | .5000 |
| 11 | .5000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .5000 |
| 12 | 1.0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |

Table 3. (Transition Probability $X^{(2)}(s, s')$ where $q(s'/s, a)=X^{(2)}(a, s/))$.

| $s$ \ $s'$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .0000 | .1538 | .1410 | .1282 | .1154 | .1026 | .0897 | .0769 | .0641 | .0513 | .0385 | .0256 | .0128 |
| 1 | .0000 | .0000 | .1667 | .1515 | .1364 | .1212 | .1061 | .0909 | .0758 | .0606 | .0455 | .0303 | .0152 |
| 2 | .0000 | .0000 | .0000 | .1818 | .1636 | .1455 | .1273 | .1091 | .0909 | .0727 | .0545 | .0364 | .0182 |
| 3 | .0000 | .0000 | .0000 | .0000 | .2000 | .1778 | .1556 | .1333 | .1111 | .0889 | .0667 | .0444 | .0222 |
| 4 | .0000 | .0000 | .0000 | .0000 | .0000 | .2222 | .1944 | .1667 | .1389 | .1111 | .0833 | .0556 | .0278 |
| 5 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .2500 | .2143 | .1786 | .1429 | .1071 | .0714 | .0357 |
| 6 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .2857 | .2381 | .1905 | .1429 | .0952 | .0476 |
| 7 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .3333 | .2667 | .2000 | .1333 | .0667 |
| 8 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .4000 | .3000 | .2000 | .1000 |
| 9 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .5000 | .3333 | .1667 |
| 10 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .5000 | .5000 |
| 11 | .5000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .5000 |
| 12 | 1.0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |

Table 4. (Transition Probability $X^{(30)}(s, s')$ where $q(s'/s, a)=X^{(30)}(a, s'))$.

| $s$ \ $s'$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0000 | 0.9212 | 0.0739 | 0.0047 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1 | 0.0000 | 0.0000 | 0.9380 | 0.0591 | 0.0028 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 0.0000 | 0.0000 | 0.0000 | 0.9536 | 0.0449 | 0.0015 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9675 | 0.0318 | 0.0007 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9794 | 0.0204 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 5 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9886 | 0.0113 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9950 | 0.0050 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 7 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9985 | 0.0015 | 0.0000 | 0.0000 | 0.0000 |
| 8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9998 | 0.0002 | 0.0000 | 0.0000 |
| 9 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| 10 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5000 | 0.5000 |
| 11 | 0.5000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5000 |
| 12 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Table 5. (Optimal policy).

| $s$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 12 |
| $f_2$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 12 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $f_k$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 12 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $f_{29}$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 12 |
| $f_{30}$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 12 |

## References

[1] David Blackmell, *Discounted dynamic programming,* Ann. Math. Statist. **36** (1965), 226-235.

[2] Howard, Ronald A, *Dynamic Programming and Markov Process,* M. I. T. Press (1960).

[3] R. E. Strauch, *Negative dynamic programming,* Ann. Math. Statist. **37** (1966), 871-890.

[4] Derman, Cyrus, *On sequential decisions and Markov chains,* Managiment Sci., (1967).