*The Annals of Probability*
1978, Vol. 6, No. 3, 491–498

# ON THE MANY-ARMED BANDIT PROBLEM

By Leiba Rodman

*Tel-Aviv University*

Two models for the "many-armed bandit" problem with two distributions are considered. Feldman's result is extended to these models.

**1. Introduction.** We consider two models for the "many-armed bandit" problem involving two distributions $P$ and $Q$.

Model (A): We have $n$ experiments. A trial on experiment $i$ yields an observation from distribution $P_i$, where all $P_i$ except one are equal to $Q$ and the exceptional one is equal to $P$. We do not know which $P_i$ is $P$ but assign an a priori probability $s_i$ to this event. If on some trial experiment $i$ was performed, we lose (unknowingly) 1 on that trial if $P_i = Q$, and 0 if $P_i = P$. For a fixed number of trials (possibly $\infty$), the goal is to find a rule for the choice of experiments at each trial that will minimize the (discounted or undiscounted) expected total loss.

Model (B): The same as model (A), the sole difference being that all $P_i$ except one are equal to $P$ and the exceptional one is equal to $Q$. The framework for these problems is negative dynamic programming (see Strauch (1966)). They generalize Feldman's "two-armed bandit" problem (1962), which coincides with both models (A) and (B) which we have introduced, when $n = 2$. Feldman proved in the case of $n = 2$ for the undiscounted problem that the optimal rule for one trial, if applied again and again, will provide an optimal rule for any number of trials. This rule yields a finite expected total loss in the undiscounted case involving an infinite number of trials (unless, of course, $P = Q$).

The purpose of this work is to extend Feldman's result to models (A) and (B). Simple examples show that Feldman's result can not be extended to the general three-armed bandit problem involving 3 distributions, or to the model involving 2 distributions, but where more than one experiment yields an observation from each one of them.

**2. Two-armed bandit case.** In this section we deal with Feldman's two-armed bandit model (1962). We establish a preliminary lemma and formulate for convenience the result Feldman has obtained.

We represent Feldman's two-armed bandit model as a negative dynamic programming problem (Strauch (1966)). The set of states $S$ is the standard one-dimensional simplex:

$$S = \{(s_1, s_2)/s_1 + s_2 = 1; s_1, s_2 \geqq 0\}.$$

Received April 15, 1976; revised April 14, 1977.
*AMS* 1970 *subject classification.* Primary 62C10.
*Key words and phrases.* Many-armed bandit problem, dynamic programming, optimal policy.

The state $(s_1, s_2)$ means that we assign a priori probability $s_1$ to the event $P_1 = P$ and $P_2 = Q$, and probability $s_2$ to the event $P_1 = Q$ and $P_2 = P$. The set of acts is $U = \{A_1, A_2\}$, where the act $A_1(A_2)$ means to choose experiment 1(2) for the next trial. If we have been in state $(s_1, s_2)$ and applied the act $A_1$, then the new state will be

$$(s_1 p(x)/(s_1 p(x) + s_2 q(x)),\ s_2 q(x)/(s_1 p(x) + s_2 q(x)))\,,$$

where $p(x)$ and $q(x)$ are densities of $P$ and $Q$ respectively, with respect to some measure $\mu$. If the act $A_2$ is chosen, then the new state is

$$(s_1 q(x)/(s_1 q(x) + s_2 p(x)),\ s_2 p(x)/(s_1 q(x) + s_2 p(x)))\,.$$

The loss function is defined as follows:

$$r((s_1, s_2), A_1) = s_2\,; \qquad r((s_1, s_2), A_2) = s_1\,.$$

Consider the discounted problem with the discount factor $\beta$; $0 \leqq \beta \leqq 1$. Let $\pi$ be a stationary policy which is optimal for one trial (i.e., optimal in the game consisting of one trial only). Let $I_k(s_1, s_2)$ be the expected discounted loss from $\pi$ in $k$ trials, when the initial state is $s = (s_1, s_2)$. Define functions $J_k(t_1, t_2)$, $k = 1, 2, \cdots$ for every pair $(t_1, t_2)$ of numbers, such that $t_1, t_2 \geqq 0$:

$$\begin{aligned} J_k(t_1, t_2) &= (t_1 + t_2) I_k(t_1/(t_1 + t_2),\ t_2/(t_1 + t_2)) && \text{if} \quad t_1 + t_2 > 0\,; \\ &= 0 && \text{otherwise.} \end{aligned}$$

Define $J_0(t_1, t_2) \equiv 0$. We use the following abbreviations: instead of $p(x)$, $q(x)$, $\mu(x)$, $p(y)$, $q(y)$, $\mu(y)$ we shall write $p$, $q$, $\mu$, $p_y$, $q_y$, $\mu_y$ respectively.

From the definitions it may be easily seen that the functions $J_k$ are positive homogeneous, symmetric, and satisfy the following equation: if $t_1 \geqq t_2$, then

$$J_k(t_1, t_2) = t_2 + \beta \int J_{k-1}(t_1 p,\ t_2 q)\, d\mu \qquad k = 1, 2, \cdots\,. \tag{1}$$

Define functions $D_k(t_1, t_2)$ as follows:

$$D_k(t_1, t_2) = t_1 - t_2 + \beta \int [J_k(t_1 q,\ t_2 p) - J_k(t_1 p,\ t_2 q)]\, d\mu \qquad k = 0, 1, \cdots\,.$$

From this definition we obtain immediately that $D_k(t_1, t_2) = -D_k(t_2, t_1)$.

LEMMA 1. *$D_k(t_1, t_2)$ is an increasing function of $t_1$, when $t_2$ is kept fixed.*

PROOF. By induction. Suppose Lemma 1 is proven for $k - 1$. We have:

$$\begin{aligned} D_k(t_1, t_2) &= t_1 - t_2 + \beta \int [J_k(t_1 q,\ t_2 p) - J_k(t_1 p,\ t_2 q)]\, d\mu \\ &= t_1 - t_2 + \beta \int_{\{t_1 q \leqq t_2 p\}} t_1 q\, d\mu + \beta^2 \int\int_{\{t_1 q \leqq t_2 p\}} J_{k-1}(t_1 q q_y,\ t_2 p p_y)\, d\mu_y\, d\mu \\ &\quad + \beta \int_{\{t_1 q > t_2 p\}} t_2 p\, d\mu + \beta^2 \int\int_{\{t_1 q > t_2 p\}} J_{k-1}(t_1 q p_y,\ t_2 p q_y)\, d\mu_y\, d\mu \\ &\quad - \beta \int_{\{t_1 p \leqq t_2 q\}} t_1 p\, d\mu - \beta^2 \int\int_{\{t_1 p \leqq t_2 q\}} J_{k-1}(t_1 p q_y,\ t_2 q p_y)\, d\mu_y\, d\mu \\ &\quad - \beta \int_{\{t_1 p > t_2 q\}} t_2 q\, d\mu - \beta^2 \int\int_{\{t_1 p > t_2 q\}} J_{k-1}(t_1 p p_y,\ t_2 p p_y)\, d\mu_y\, d\mu \end{aligned}$$

and after rearranging terms we get

$$\begin{aligned} D_k(t_1, t_2) = {} & \beta \int_{\{t_1 q \leqq t_2 p\}} D_{k-1}(t_1 q,\ t_2 p)\, d\mu + \beta \int_{\{t_1 p > t_2 q\}} D_{k-1}(t_1 p,\ t_2 q)\, d\mu \\ & + (1 - \beta)(t_1 - t_2)\,. \end{aligned} \tag{2}$$

If we take $t_1^* > t_1$ and calculate $D_k(t_1^*, t_2) - D_k(t_1, t_2)$ by using (2) and the induction assumption, we get Lemma 1.

Feldman (1962) proved the following theorem under $\beta = 1$, for the two-armed bandit problem.

THEOREM 1. *Let $\pi$ be a stationary policy, which is optimal for one trial. Then $\pi$ is optimal for every number of trials (including $\infty$) and for every discount factor $\beta \in [0, 1]$. The expected total loss under $\pi$ for the undiscounted problem involving infinitely many trials is finite (unless $P = Q$).*

REMARK. One can easily obtain the optimality of $\pi$ from Lemma 1 (because the optimality of $\pi$ means exactly $D_k(t_1, t_2) \geqq 0$ if $t_1 \geqq t_2$, and $D_k(t_1, t_2) \leqq 0$ if $t_1 \leqq t_2$).

**3. Model A.** In this section we prove Theorem 1 for model (A) described in the introduction. Represent model (A) as a negative dynamic programming problem. The set of states $S$ is the $(n-1)$-dimensional standard simplex:

$$S = \{(s_1, \cdots, s_n) \mid \textstyle\sum_{i=1}^n s_i = 1; s_i \geqq 0\}.$$

The state $(s_1, s_2, \cdots, s_n)$ means that we assign a priori probability $s_i$ to the event that $P_i$ is equal to $P$ and all other $P_j$ are equal to $Q$. The set of acts is $U = \{A_1, \cdots, A_n\}$, where the act $A_i$ means to choose experiment $i$ for the next trial. If act $A_i$ is applied, when the state is $(s_1, \cdots, s_n)$, an observation $x$ is obtained from $P_i$, then the new state is

$$(s_1 q(x)/\Sigma, \cdots, s_i p(x)/\Sigma, \cdots, s_n q(x)/\Sigma),$$

where $\Sigma = s_i p(x) + (1 - s_i)q(x)$.

The loss function is defined as follows:

$$r((s_1, \cdots, s_n), A_i) = 1 - s_i; \qquad i = 1, \cdots, n.$$

Let $\pi$ be a stationary policy which is optimal for one trial. Let $I_k(s_1, \cdots, s_n)$ be the discounted expected loss in the first $k$ trials under $\pi$, when the initial state is $s = (s_1, \cdots, s_n)$. Define functions $J_k(t_1, \cdots, t_n)$; $k = 1, 2, \cdots$ for $t_1, t_2, \cdots, t_n \geqq 0$:

$$\begin{aligned} &J_k(t_1, \cdots, t_n) \\ &\quad = (\textstyle\sum_{i=1}^n t_i) I_k(t_1/\sum_{i=1}^n t_i, t_2/\sum_{i=1}^n t_i, \cdots, t_n/\sum_{i=1}^n t_i) && \text{if } \textstyle\sum_{i=1}^n t_i > 0; \\ &\quad = 0 && \text{otherwise.} \end{aligned}$$

Define $J_0 \equiv 0$.

From the definitions it may easily be seen that $J_k(t_1, \cdots, t_n)$ is positive homogeneous, symmetric, and satisfies the following equation: if $t_1 \geqq \cdots \geqq t_n$, then

$$J_k(t_1, \cdots, t_n) = t_2 + t_3 + \cdots + t_n + \beta \int J_{k-1}(t_1 p, t_2 q, \cdots, t_n q)\, d\mu \qquad k = 1, 2, \cdots.$$

In the sequel we use the following notations:

$$t \quad \text{will denote a vector} \quad (t_1, \cdots, t_n), \quad \text{where} \quad t_i \geqq 0;$$
$$j_k^{(i)}(t, x) \equiv J_k(t_1 q(x), \cdots, t_{i-1} q(x), t_i p(x), t_{i+1} q(x), \cdots, t_n q(x));$$
$$T = T(t) \equiv \sum_{i=1}^{n} t_i;$$
$$\Lambda_i(t, x) \equiv (t_1 q(x), \cdots, t_{i-1} q(x), t_i p(x), t_{i+1} q(x), \cdots, t_n q(x)).$$

LEMMA 2. *Let $t_1 \geqq \cdots \geqq t_n \geqq 0$ and fix integer $i$ with $1 < i < n$. Then*

$$\int [j_k^{(1)}(t, x) - j_k^{(i)}(t, x)]\, d\mu \tag{3}$$

*does not depend on $t_n$.*

PROOF. We shall prove that

$$\int [j_k^{(i-1)}(t, x) - j_k^{(i)}(t, x)]\, d\mu \tag{4}$$

does not depend on $t_n$. Use induction on $k$. The case $k = 1$ being trivial, we proceed to the induction step. Suppose at first $i = 2$. We have:

$$\begin{aligned}
&\int [j_k^{(1)}(t, x) - j_k^{(2)}(t, x)]\, d\mu \\
&\quad = \int_{\{t_1 p > t_2 q\}} j_k^{(1)}(t, x)\, d\mu + \int_{\{t_1 p \leqq t_2 q\}} j_k^{(1)}(t, x)\, d\mu \\
&\qquad - \int_{\{t_2 p > t_1 q\}} j_k^{(2)}(t, x)\, d\mu - \int_{\{t_2 p \leqq t_1 q\}} j_k^{(2)}(t, x)\, d\mu \\
&\quad = \int_{\{t_1 p > t_2 q\}} [T(\Lambda_1(t, x)) - t_1 p + \beta \int j_{k-1}^{(1)}(\Lambda_1(t, x), y)\, d\mu_y]\, d\mu \\
&\qquad + \int_{\{t_1 p \leqq t_2 q\}} [T(\Lambda_1(t, x)) - t_2 q + \beta \int j_{k-1}^{(2)}(\Lambda_1(t, x), y)\, d\mu_y]\, d\mu \\
&\qquad - \int_{\{t_2 p > t_1 q\}} [T(\Lambda_2(t, x)) - t_2 p + \beta \int j_{k-1}^{(2)}(\Lambda_2(t, x), y)\, d\mu_y]\, d\mu \\
&\qquad - \int_{\{t_2 p \leqq t_1 q\}} [T(\Lambda_2(t, x)) - t_1 q + \beta \int j_{k-1}^{(1)}(\Lambda_2(t, x), y)\, d\mu_y]\, d\mu.
\end{aligned}$$

It is sufficient to check that the sum of the double integrals does not depend on $t_n$. But this sum is equal to:

$$\begin{aligned}
&\iint_{\{t_1 p > t_2 q\}} [j_{k-1}^{(1)}(\Lambda_1(t, x), y) - j_{k-1}^{(2)}(\Lambda_1(t, x), y)]\, d\mu_y\, d\mu \\
&\qquad + \iint j_{k-1}^{(2)}(\Lambda_1(t, x), y)\, d\mu_y\, d\mu \\
&\qquad\quad + \iint_{\{t_2 p > t_1 q\}} [j_{k-1}^{(1)}(\Lambda_2(t, x), y) - j_{k-1}^{(2)}(\Lambda_2(t, x), y)]\, d\mu_y\, d\mu \\
&\qquad\quad\quad - \iint j_{k-1}^{(1)}(\Lambda_2(t, x), y)\, d\mu_y\, d\mu.
\end{aligned}$$

The second and fourth integrals are equal. (After changing the order of integration, and the names of variables, the second integral is seen to be equal to the fourth integral.) From the assumption of induction we learn that (4) does not depend on $t_n$.

Suppose now that $i > 2$. Then, in the same way, we discover that independence of $t_n$ in the expression

$$\iint [j_{k-1}^{(1)}(\Lambda_{i-1}(t, x), y) - j_{k-1}^{(1)}(\Lambda_i(t, x), y)]\, d\mu_y\, d\mu \tag{5}$$

is sufficient for proving independence of $t_n$ in (4). But by examining two cases, $t_1 p_y \geqq t_n q_y$ and $t_1 p_y < t_n q_y$, and by induction on $k$ and $n$, we obtain that (5) does not depend on $t_n$.

LEMMA 3. *Let* $1 \leqq i \leqq n-1$; *let* $t_1 \geqq t_2 \geqq \cdots \geqq t_n \geqq 0$; $t_i \geqq t_n' \geqq t_n$. *Denote* $t = (t_1, \cdots, t_n)$; $t' = (t_1, \cdots, t_{n-1}, t_n')$. *Then*

$$\beta \int [j_k^{(i)}(t', x) - j_k^{(n)}(t', x) - j_k^{(i)}(t, x) + j_k^{(n)}(t, x)]\, d\mu \geqq t_n - t_n' . \tag{6}$$

PROOF. By induction on $k$ and $n$. If $n = 2$, then (6) is a consequence of Lemma 1. For $k = 1$ one can easily obtain (6) by straightforward calculation. Suppose now that $n \geqq 3$, and we have proven (6) for $k-1$ in the model with $n$ experiments, and for every $k$ in the model with $n-1$ experiments. We shall prove (6) for $k$ in the model with $n$ experiments.

Suppose at first $i \neq 1$. Computing as in the proof of Lemma 2, express the left-hand side of (6) as follows:

$$\begin{aligned}
&\beta[t_n' - t_n + \textstyle\int_{\{t_1 q < t_n' p\}} t_1 q\, d\mu - \int_{\{t_1 q \geqq t_n' p\}} t_n' p\, d\mu + \int_{\{t_1 q < t_n p\}} t_1 q\, d\mu + \int_{\{t_1 q \geqq t_n p\}} t_n p\, d\mu] \\
&\quad + \beta^2[\textstyle\int\int_{\{t_i p > t_1 q\}} [j_{k-1}^{(i)}(\Lambda_i(t', x), y) - j_{k-1}^{(1)}(\Lambda_i(t', x), y)]\, d\mu_y\, d\mu \\
&\quad + \textstyle\int\int j_{k-1}^{(1)}(\Lambda_i(t', x), y)\, d\mu_y\, d\mu \\
&\quad - \textstyle\int\int_{\{t_n' p > t_1 q\}} [j_{k-1}^{(n)}(\Lambda_n(t', x), y) - j_{k-1}^{(1)}(\Lambda_n(t', x), y)]\, d\mu_y\, d\mu \\
&\quad - \textstyle\int\int j_{k-1}^{(1)}(\Lambda_n(t', x), y)\, d\mu_y\, d\mu \\
&\quad - \textstyle\int\int_{\{t_i p > t_1 q\}} [j_{k-1}^{(i)}(\Lambda_i(t, x), y) - j_{k-1}^{(1)}(\Lambda_i(t, x), y)]\, d\mu_y\, d\mu \\
&\quad - \textstyle\int\int j_{k-1}^{(1)}(\Lambda_i(t, x), y)\, d\mu_y\, d\mu \\
&\quad + \textstyle\int\int_{\{t_n p > t_1 q\}} [j_{k-1}^{(n)}(\Lambda_n(t, x), y) - j_{k-1}^{(1)}(\Lambda_n(t, x), y)]\, d\mu_y\, d\mu \\
&\quad + \textstyle\int\int j_{k-1}^{(1)}(\Lambda_n(t, x), y)\, d\mu_y\, d\mu] \\
&\qquad = (\text{by force of Lemma 2}) \quad \beta[t_n' - t_n - \textstyle\int_{\{t_1 q < t_n' p\}} t_1 q\, d\mu \\
&\qquad\quad - \textstyle\int_{\{t_1 q \geqq t_n' p\}} t_n' p\, d\mu + \int_{\{t_1 q < t_n p\}} t_1 q\, d\mu + \int_{\{t_1 q \geqq t_n p\}} t_n p\, d\mu] \\
&\qquad\quad + \beta^2 \textstyle\int\int_{\{t_n p > t_1 q\}} [J_{k-1}(t_1 q q_y, t_n p p_y) - J_{k-1}(t_1 q p_y, t_n p q_y)]\, d\mu_y\, d\mu \\
&\qquad\quad - \beta^2 \textstyle\int\int_{\{t_n' p > t_1 q\}} [J_{k-1}(t_1 q q_y, t_n' p p_y) - J_{k-1}(t_1 q p_y, t_n' p q_y)]\, d\mu_y\, d\mu \\
&\qquad\quad + \beta^2 \textstyle\int\int [j_{k-1}^{(i)}(\Lambda_1(t', x), y) - j_{k-1}^{(n)}(\Lambda_1(t', x), y) \\
&\qquad\quad - j_{k-1}^{(i)}(\Lambda_1(t, x), y) + j_{k-1}^{(n)}(\Lambda_1(t, x), y)]\, d\mu_y\, d\mu .
\end{aligned}$$

The last integral is greater than or equal to $\beta(t_n - t_n')$ by the induction assumption. All that precedes it is nonnegative, as can be checked using Lemma 1.

Suppose now $i = 1$. Then we can restrict ourselves to the case $t_n \leqq t_{n-1}'$. But then by the same arguments as in the case $i \neq 1$, we obtain (6).

We start now with the proof of Theorem 1 for model (A). We must show that

$$J_k(t_1, \cdots, t_n) = \min_i (T(t) - t_i + \beta \int j_{k-1}^{(i)}(t, x)\, d\mu) . \tag{7}$$

Use induction on $n$ and on $k$. Without loss of generality, we can assume $t_1 \geqq t_2 \geqq \cdots \geqq t_n$. Then

$$J_k(t_1, \cdots, t_n) = T(t) - t_1 + \beta \int j_{k-1}^{(1)}(t, x)\, d\mu$$

and the inequality

$$T(t) - t_1 + \beta \int j_{k-1}^{(1)}(t, x)\, d\mu \leqq T(t) - t_i + \beta \int j_{k-1}^{(i)}(t, x)\, d\mu$$

for $i \neq n$ follows from Lemma 2 and from the assumption of induction. Now (7) is a consequence of the inequality

$$T(t) - t_{n-1} + \beta \int j_{k-1}^{(n-1)}(t, x)\, d\mu \leqq T(t) - t_n + \beta \int j_{k-1}^{(n)}(t, x)\, d\mu ,$$

which follows immediately from Lemma 3. Theorem 1 is proven for every finite $k$. From the results of negative dynamic programming (see Strauch (1966)) Theorem 1 follows also for $k = \infty$.

Suppose now that $P \neq Q$. We shall prove that the optimal expected total loss in an infinite number of trials for the undiscounted problem is finite. It is sufficient to find a real-valued function $u(s)$, $s \in S$, for which

$$T_\pi u(s) \leqq u(s) \tag{8}$$

for all $s$, where $T_\pi$ is the operator associated with the stationary policy $\pi$ (see Strauch (1966)). Define

$$u(s_1, \cdots, s_n) = C(\prod_{i=1}^n (1 - s_i))^{1/n} .$$

Now (8) follows (when the constant $C$ is large enough) from the inequality

$$\int (x_1, \cdots, x_n)^{1/n}\, d\mu \leqq (\prod_{i-1}^n \int x_i\, d\mu)^{1/n} , \tag{9}$$

where $x_i$ are measurable nonnegative functions, and from the following condition for a strict inequality in (9): if $\int x_i\, d\mu > 0$ for every $i$, and not all the expressions $x_i / \int x_i\, d\mu$ are equal a.s., then the strict inequality holds in (9).

**4. Model (B).** In this section we prove Theorem 1 for model (B). The proof is analogous to the proof of Theorem 1 for model (A). Therefore, we shall mention only the main steps of the proof.

The set of states $S$ is as in model (A), but now $s_i$ is the a priori probability of the event that $P_i$ is equal to $Q$, and all the other $P_j$ are equal to $P$. The set of acts $A$ is as in model (A). If act $A_i$ is applied when the state is $(s_1, \cdots, s_n)$, an observation $x$ is obtained from $P_i$, then the new state is

$$(s_1 p(x)/\Sigma, s_2 p(x)/\Sigma, \cdots, s_i q(x)/\Sigma, \cdots, s_n p(x)/\Sigma) ,$$

where $\Sigma = s_i q(x) + (1 - s_i)p(x)$. The loss function is defined as follows:

$$r((s_1, \cdots, s_n), A_i) = s_i .$$

Let $\pi$ be a stationary policy which is optimal for one trial. Define $I_k^{(n)}$ and $J_k^{(n)}$ as for model (A) (the index $(n)$ stresses the fact that the values $I_k^{(n)}$ and $J_k^{(n)}$ are associated with the $n$-armed bandit problem). $J_k^{(n)}$ is symmetric, positive homogeneous and satisfies the equation: if $t_1 \geqq t_2 \geqq \cdots \geqq t_n$, then

$$J_k^{(n)}(t_1, \cdots, t_n) = t_n + \beta \int J_{k-1}^{(n)}(t_1 p, t_2 p, \cdots, t_n q)\, d\mu ; \qquad k = 1, 2, \cdots .$$

We use the following notations (in a way analogous to the preceding section):

$$t = (t_1, \cdots, t_n) ;$$
$$j_{k,n}^{(i)}(t, x) = J_k^{(n)}(t_1 p(x), \cdots, t_{i-1} p(x), t_i q(x), t_{i+1} p(x), \cdots, t_n p(x)) ;$$
$$\Lambda_i(t, x) = (t_1 p(x), \cdots, t_i q(x), \cdots, t_n p(x)) .$$

LEMMA 4. *Let* $n \geqq 3$; $t_1 \geqq \cdots \geqq t_n \geqq 0$; $1 < i < n$. *Denote* $t^1 = (t_2, \cdots, t_n)$. *Then*

$$\int [j_{k,n}^{(i)}(t, x) - j_{k,n}^{(n)}(t, x)]\, d\mu = \int [j_{k,n-1}^{(i-1)}(t^1, x) - j_{k,n-1}^{(n-1)}(t^1, x)]\, d\mu\,.$$

LEMMA 5. *Let* $2 \leqq i \leqq n$; *let* $t_1 \geqq t_2 \geqq \cdots \geqq t_n \geqq 0$, *and let* $t_1'$ *satisfy* $t_i \leqq t_1' \leqq t_1$. *Denote* $t' = (t_1', t_2, \cdots, t_n)$. *Then*

$$\beta \int [j_{k,n}^{(i)}(t, x) - j_{k,n}^{(1)}(t, x) - j_{k,n}^{(i)}(t', x) + j_{k,n}^{(1)}(t', x)]\, d\mu \leqq t_1 - t_1'\,.$$

The proofs of Lemma 4 and Lemma 5 are analogous to the proofs of Lemma 2 and Lemma 3 respectively.

Now Theorem 1 for model (B) follows from Lemma 4 and Lemma 5 in the same way as in the preceding section.

For establishing the finiteness of the optimal expected total loss in an infinite number of trials for the undiscounted problem, choose

$$u(s_1, \cdots, s_n) = C(s_1 \cdots s_n)^{1/n}$$

and use the inequality

$$\int p^{(n-1)/n} q^{1/n}\, d\mu < 1\,,$$

if $p$ and $q$ are different.

**5. Examples.** We present here two examples we have mentioned in the introduction.

EXAMPLE 1. Let $P_1$, $P_2$, $P_3$ be mutually singular distributions. We have 3 experiments; a trial from experiment $i$ yields an observation from distribution $P_{\sigma(i)}$, where $\sigma$ is a permutation of the set $\{1, 2, 3\}$; $i = 1, 2, 3$. If experiment $i$ was performed we lose (unknowingly) 1 if $\sigma(i) = 2$ or $\sigma(i) = 3$, and 0 if $\sigma(i) = 1$. Denote by $\sigma_1$, $\sigma_2$, $\sigma_3$, $\sigma_4$, $\sigma_5$, $\sigma_6$ the permutations of the set $\{1, 2, 3\}$ which move $\{1, 2, 3\}$ to $\{1, 2, 3\}$; $\{1, 3, 2\}$; $\{2, 1, 3\}$; $\{3, 1, 2\}$; $\{2, 3, 1\}$; $\{3, 2, 1\}$ respectively. We assign a priori probability $s_j$ to the permutation $\sigma_j$; $j = 1, 2, 3, 4, 5, 6$. Then for the state $s_1 = 0.01$; $s_2 = 0.47$; $s_3 = 0.44$; $s_4 = 0.03$; $s_5 = 0.01$; $s_6 = 0.04$ the optimal policy for one trial is optimal for no number of trials greater than 1.

EXAMPLE 2. Let $P$ and $Q$ be singular distributions. We have 4 experiments; a trial from experiment $i$ yields an observation from distribution $P_i$; $i = 1, 2, 3, 4$. Two $P_i$ are equal to $P$, and two other $P_i$ are equal to $Q$. If experiment $i$ was performed we lose (unknowingly) 1 if $P_i = Q$, and 0 if $P_i = P$. We assign a priori probabilities $s_1, s_2, s_3, s_4, s_5, s_6$ to the events that $\{P_1 = P_2 = P; P_3 = P_4 = Q\}$; $\{P_1 = P_3 = P; P_2 = P_4 = Q\}$; $\{P_1 = P_4 = P; P_2 = P_3 = Q\}$; $\{P_2 = P_3 = P; P_1 = P_4 = Q\}$; $\{P_2 = P_4 = P; P_1 = P_3 = Q\}$; $\{P_3 = P_4 = P; P_1 = P_2 = Q\}$ respectively. Then for the state $s_1 = \frac{7}{24}$; $s_2 = \frac{8}{24}$; $s_3 = 0$; $s_4 = \frac{4}{24}$; $s_5 = \frac{3}{24}$; $s_6 = \frac{2}{24}$ the optimal policy for one trial is optimal for no number of trials greater than 1.

It is possible to construct Examples 1 and 2 with mutually absolutely continuous distributions. Such examples should satisfy a strict inequality between integrals. We can change a little the mutually singular distributions of Examples

1 and 2, making them mutually absolutely continuous and maintaining the inequality.

**Acknowledgment.** This work is based on a thesis submitted to the Department of Statistics, Tel-Aviv University as partial fulfillment of the requirements for the M.A. degree. The work was done under the supervision of Dr. I. Meilijson, whom I wish to thank for his most valuable help.

## REFERENCES

[1] FELDMAN, DORIAN (1962). Contributions to the two-armed bandit problem. *Ann. Math. Statist.* **33** 847–856.

[2] STRAUCH, RALPH E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.

DEPARTMENT OF MATHEMATICS
TEL-AVIV UNIVERSITY
TEL-AVIV
ISRAEL