RAIRO-Oper. Res. 51 (2017) 1119–1132
DOI: 10.1051/ro/2017015

RAIRO Operations Research
www.rairo-ro.org

OPTIMAL ASSIGNMENT OF SELLERS IN A STORE WITH A RANDOM NUMBER OF CLIENTS *VIA* THE ARMED BANDIT MODEL*

Víctor Hugo Vázquez-Guevara[1], Hugo Cruz–Suárez[1] and Fernando Velasco-Luna[1]

Abstract. The technique of Dynamic Programming for Armed Bandits is employed for solving the problem of maximizing the randomly depreciated gains of a store with unknown (finite random) number of clients with fixed (finite) number of sellers which skills are also random and will be represented as probability distributions which are themselves random. Hence, Armed Bandits's framework will be considered with horizon being a random variable with a finite support, that far as the authors know, it has not yet been discussed. In addition, numerical examples are detailed in order to illustrate the versatility and practical implementation of the approach presented in this paper in two general contexts, given by the number of available products: one product only, such situation coincides with that in which the number of sales needs to be maximized. And, more than one product, in this case, the amount of sales is not necessarily ruled by a Bernoulli distribution.

Mathematics Subject Classification. 49L20, 90C40, 93E20.

Received November 12, 2015. Accepted March 2, 2017.

1. Introduction

Maximizing gains is one of the most important objectives for every store, however, randomness may cause this labor harder since several levels of uncertainty may be considered:

(1) Known number of clients and amount of sales beforehand (non random),
(2) sellers with random sales with known distribution, and
(3) sellers with random sales with unknown distribution.

If such store can only provide service to one client at a time (this assumption remains valid in all the sequel) then the strategy is clear: you may have a lot of employees but you will always *"use"* the best one of them. But, what if we do not know which one is the best seller? In case 2, some appropriated stochastic order may be considered in determining the "best seller".

Case 3 is, in some how, more complicated since certain distribution is considered in the set of probability distributions of the total amount sold by every employee, we will assume that such distribution is known and

Keywords. Armed bandit model, dynamic programming, assignment of personal, random horizon, markov decision processes.

* *This work was partially supported by VIEP-BUAP, via the project: "Estimación de momentos de orden par del ruido en procesos ARX con ruido correlacionado".*

[1] Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, San Claudio y 18 sur. San Manuel, 72570, Puebla, Mexico. vvazquez@fcfm.buap.mx

Article published by EDP Sciences

that such knowledge comes from the product prices, an analysis of their *"résumés"*, the interview prior the engagement, the training period, etc. (this may happen is the sellers are new at the store).

We may consider another random factor: the number of clients. In case 1 this randomness is not considered, but, in situations 2 and 3 it may be incorporated, then it seems reasonable that a discrete random variable with a finite support can be used to model this random factor.

Finally, random discount factors may be taken into account in order to consider the depreciation of money. In economic and financial models, the discount factor is determined by interest rates (r): $(\frac{1}{1+r})$, which in turn are uncertain. This uncertainty in various situations is due to external random noises.

We will consider then, a store with k sellers for whom probability distribution of their sales are themselves random as well as the number of clients and the discount factors.

The store's manager needs to design an assignment *strategy* in order to distribute each of the customers that will arrive at the store among his personal for maximizing the randomly discounted sum of gains. After a certain number of clients have arrived, such strategy might be based on:

(1) The distribution of the probability distribution of sales of each seller,
(2) The previous selections, and
(3) The observed performances.

The approach that we will discuss is the one offered by the Armed Bandit model and the Dynamic Programming technique [1], since an Armed Bandit problem is related to sequential selections from k stochastic processes (known as arms, and in the context of this paper as sellers) characterized by parameters typically unknown. From the context of the assignment problem, we observe that both time and horizon (number of clients) may be considered finite, and hence the Dynamic Programming technique is a suitable tool.

To the best of our knowledge, the Armed Bandit model with random horizon has not been explored yet and we consider that the previous assignment problem is an intuitive way for its motivation. Of course that the considerations on the random components of the problem and the problem itself may be extrapolated to more complex situations.

The distribution of this paper is as follows: Section 2 is devoted to the Armed Bandit Model as well as its mathematical definition considering randomness in discount sequence and the horizon. Section 3 deals with an appropriate version of the Dynamic Programming technique that in Section 4 will be useful for solving some examples of the assignment problem. Finally, a conclusion and some references are provided.

2. The armed bandit model

2.1. Introduction

In this section, Armed Bandit model will be presented in detail. In addition, some ideas will be considered initially for the case in which the number of clients (horizon) is known beforehand (there exists an appointment system) and discount factors are deterministic, then it will be shown that if the number of customers is not random and the discount sequence is random non-observable (or observable as in the particular case presented in Sect. 2.3) there are no additional complications in its solution. Finally, a random number of clients will be considered and will be demonstrated that the initially explored ideas remain valid. Therefore, the most general case considered in this paper (random distributions of amount of sales, random number of clients and random discount factors) will be solved *via* a version of Dynamic Programming for Armed Bandits with finite deterministic horizon and deterministic discount factors after making some adjustments.

In a Bandit Problem, we must consider k stochastic processes from which one of them must be selected sequentially at some points of time in order to observe certain numerical characteristic. Even at each time of observation only one process is selected, we assume that the non observed processes produce also the numerical characteristic of interest and hence, we may consider a total of k sequences of random variables. In other words, an observation in any particular sequence is made when the corresponding process is selected, in addition the

classical objective is to maximize the expected value of the discounted payoff $\sum_{m=1}^{n} \alpha_m Z_m$ where Z_m is the observed variable at stage m, (α_m) is a sequence of non-negative random variables known as the discount factor sequence and n is the horizon of the bandit. At this point, we consider that horizon of the bandit is not random and finite.

We write $\mathcal{D}$ for representing the space of probability distributions on $\mathbb{R}$ together with the topology of convergence in distribution. This space is: separable, locally compact, metrizable and complete [6].

The space $\mathcal{D}^k$ of order k-tuples of members of $\mathcal{D}$ will be considered with the corresponding product topology. The component Q_i of $(Q_1, Q_2, \ldots, Q_k) \in \mathcal{D}^k$ governs observations in arm i. An element $G \in \mathcal{D}(\mathcal{D}^k)$ represents the prior knowledge on the k arms.

It can be shown [3] that there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\{X_{im} : 1 \leq i \leq k, m = 1, 2, \ldots, n\} \tag{2.1}$$

is an independent family of random variables, where X_{im} represent result of arm i at stage m.

2.2. Discount sequences

Let us consider the space $\mathcal{A}$ of all the deterministic discount sequences and a probability distribution H on $\mathcal{A}$, such distribution represents the prior knowledge of the decision maker about the discount sequence. In addition, we assume that

$$\mathbb{E}\left[\sum_{m=1}^{n} \alpha_m \middle| H\right] < \infty,$$

and that arms are independent of the discount sequence.

In this paper, we will notice three different approaches on the discount sequence:

(1) Random Non observable,
(2) Random Observable, and
(3) Deterministic.

On the one hand, in the non observable case, the discount factors are not observed and, hence, arm selected at any time do not depend on them. On the other hand, in the observable case arms selected may depend on previous discount factors as well as on previous observations.

The deterministic sequences correspond to a degenerated case in which probability distribution $H = \delta_A$ with $A \in \mathcal{A}$; *i.e.*, the discount sequence is known and the selections may be considered as non dependent on the discount sequence.

2.3. Strategies

In this section we will consider a non observable discount sequence and, at its final part we will discuss the observable case in a particular context. A strategy τ assigns to each partial history of observations an integer between 1 and k, which indicates the arm that will be selected at next stage. Thus, $\tau(\phi)$ indicates the arm selected initially when strategy τ is followed, $\tau(z_1)$ indicates the arm that will be chosen at stage two, given that z_1 was observed initially, $\tau(z_1, z_2)$ is the arm selected at stage three given that at stage one was observed z_1 while at stage two z_2 was observed, etc. We have then, that the sequence of observed random variables is recursively defined by

$$Z_1 = X_{\tau(\phi),1}, \tag{2.2}$$

$$Z_m = X_{\tau(Z_1,\ldots,Z_{m-1}),m}, \quad 1 < m \leq n. \tag{2.3}$$

In addition, we will assume that each component Q_i of $(Q_1, Q_2, \ldots, Q_k) \in \mathcal{D}^k$ has a finite first absolute moment with G-probability one and that this moment at the same time has finite G-expectation.

We have then that Z_m is integrable since

$$|Z_m| \leq \vee_{i=1}^{k} |X_i| \leq \sum_{i=1}^{k} |X_i| \tag{2.4}$$

and for each strategy τ

$$\left| \mathbb{E}_\tau \left[\sum_{m=1}^{n} \alpha_m Z_m \middle| G, H \right] \right| \leq \mathbb{E}_\tau \left[\sum_{i=1}^{k} |X_i| \middle| G \right] \mathbb{E} \left[\sum_{m=1}^{n} \alpha_m \middle| H \right] < \infty, \tag{2.5}$$

where the subscript τ indicates the dependence of the expectation on the strategy τ.

The number $\mathbb{E}_\tau \left[\sum_{m=1}^{n} \alpha_m Z_m | G, H \right]$ is called the *worth* of strategy τ and will be denoted by $W(G, \mathbf{H}, \tau)$. Hence, we have the following definition.

Definition 2.1. *The value of the (G,H)-bandit* is given by

$$V(G, \mathbf{H}) = \sup_\tau W(G, \mathbf{H}, \tau) \tag{2.6}$$

$$= \sup_\tau \mathbb{E}_\tau \left[\sum_{m=1}^{n} \alpha_m Z_m \middle| G, H \right]. \tag{2.7}$$

A strategy τ for which $V(G, \mathbf{H})$ is attained is called *optimal*.

Since arms are independent of the discount sequence, we have for each τ that:

$$\begin{aligned} W(G, \mathbf{H}, \tau) &= \mathbb{E}_\tau \left[\sum_{m=1}^{n} \alpha_m Z_m \middle| G, H \right] \\ &= \sum_{m=1}^{n} \mathbb{E} \left[\alpha_m | H \right] \mathbb{E}_\tau \left[Z_m | G \right] \\ &= \mathbb{E}_\tau \left[\sum_{m=1}^{n} \mathbb{E} \left[\alpha_m | H \right] Z_m \middle| G \right]. \end{aligned} \tag{2.8}$$

If for simplicity $\beta_m := \mathbb{E} \left[\alpha_m | H \right]$ then

$$W(G, \mathbf{H}, \tau) = \mathbb{E}_\tau \left[\sum_{m=1}^{n} \beta_m Z_m \middle| G \right]. \tag{2.9}$$

In addition, if we consider the non random discount sequence $B = (\beta_1, \beta_2, \ldots, \beta_n) \in \mathcal{A}$ then, we observe that

$$W(G, \delta_B, \tau) = \mathbb{E}_\tau \left[\sum_{m=1}^{n} \beta_m Z_m \middle| G \right] \tag{2.10}$$

or in the simpler way

$$W(G, B, \tau) = \mathbb{E}_\tau \left[\sum_{m=1}^{n} \beta_m Z_m \middle| G \right]. \tag{2.11}$$

We may conclude that in the non observable case, discount sequence (α_m) may be replaced by the deterministic one (β_m), to be more specific:

$$W(G, H, \tau) = W \left(G, \mathbb{E} \left[A | H \right], \tau \right) \tag{2.12}$$

and

$$V(G,H)=V\left(G,\mathbb{E}\left[A|H\right]\right). \tag{2.13}$$

If we consider a sequence of independent random variables (U_i) with corresponding distributions $H_1,H_2,\ldots$ which may themselves be random with distributions $\eta_1,\eta_2,\ldots$ then we have that

$$\mathbb{E}\left[U_i\right]=\int_{\mathcal{D}}\mathbb{E}\left[U_i|H_i\right]\eta_i(\mathrm{d}H_i) \tag{2.14}$$

hence, if $\alpha_m=\Pi_{i=1}^m U_i$ then $\mathbb{E}\left[\alpha_m\right]=\Pi_{i=1}^m\mathbb{E}\left[U_i\right]$. In addition, if the random variables $U_1,U_2,\ldots$ have the same mean μ (for example, if all of them have the same distribution) then $\mathbb{E}\left[\alpha_m\right]=\mu^m$.

The observable case is simplified in the sense that the random discount sequence can be replaced by the deterministic sequence $(\mathbb{E}\left[A|H\right])$ if the latter context is held [2,3].

Hence, we may consider the random non observable and the observable cases in the multiplicative context exposed above as equivalent to a bandit with non random discount sequence. Henceforth, we will assume that discount sequences are non random by virtue of the last discussion; *i.e.* in an implicit way, we will work with random non observable, deterministic or random observable (only in the special case previously presented) discount sequences. Then, we will refer to the (G,A)-bandit, the worth of strategy τ will be denoted by $W(G,A,\tau)$, the value of the (G,A)-bandit will be given by $V(G,A)$ and in the conditional expectations the dependence on H will be removed.

Hence, if we consider that the horizon is a finite discrete random variable T [5] with probability mass function $\rho_k=\mathbb{P}[T=k]$ for $k=1,2,\ldots,N$ then, we must find

$$W(G,A,\tau)=\mathbb{E}_\tau\left[\sum_{m=1}^{T}\alpha_m Z_m\middle|\ G\right], \tag{2.15}$$

in addition, if we assume independence between T and the observed sequence $Z_1,Z_2,\ldots,Z_N$ then [4]

$$\begin{aligned}
W(G,A,\tau)&=\mathbb{E}_\tau\left[\sum_{m=1}^{T}\alpha_m Z_m\middle|\ G\right]\\
&=\mathbb{E}\left[\mathbb{E}_\tau\left[\sum_{m=1}^{N}\alpha_m Z_m\middle|\ G,T\right]\right]\\
&=\sum_{n=1}^{N}\mathbb{E}_\tau\left[\sum_{m=1}^{n}\alpha_m Z_m\middle|\ G\right]\rho_n\\
&=\sum_{m=1}^{N}\sum_{n=m}^{N}\mathbb{E}_\tau\left[\alpha_m Z_m\middle|\ G\right]\rho_n\\
&=\mathbb{E}_\tau\left[\sum_{m=1}^{N}\alpha_m P_m Z_m\middle|\ G\right],
\end{aligned}$$

where $P_m=\mathbb{P}(T\geq m)$ for $m=1,2,\ldots,N$. With this discussion, we have seen that the assignment problem with random horizon may be transformed into another one with finite (non random) horizon with discount sequence $\gamma_k=\alpha_k P_k$ for $k=1,2,\ldots,N$ and $\gamma_k=0$ for $k\geq N+1$.

In summary, we may work with a bandit problem with a random non-observable (or observable in the case discussed earlier) discount sequence and a finite random horizon and reduce it into a problem with deterministic finite horizon and a deterministic discount sequence. This is the main reason for which the following section

deals with the technique of Dynamic Programming for bandits with finite horizon and deterministic discount sequence.

Optimal strategy in a random horizon situation is ready for making N assignments, however, given the random nature of the horizon (number of clients), it is possible that not all of them have to be made.

For a given discount sequence $A = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, we denote by $A^{(m)}$ to the sequence $(\alpha_{m+1}, \alpha_{m+2}, \ldots, \alpha_n)$. At the second stage, the decision maker will be faced with a bandit characterized by the discount sequence $A^{(1)}$ and the posteriori distribution of $(Q_1, \ldots, Q_k)$. For indicating the dependence of this posteriori distribution on the observation at stage one, the symbol $(x)_i G$ will be used for denoting the conditional distribution of $(Q_1, \ldots, Q_k)$ given observation of x in arm i.

In the case that $G = F_1 \times \ldots \times F_k$ we have that:

$$(x)_i G = (x)_i (F_1 \times \ldots \times F_k) = (F_1 \times \ldots (x)F_i \times \ldots \times F_k),$$

where $(x)F_i$ is the conditional distribution of Q_i given the observation of x in arm i.

3. Dynamic programming for finite horizon

The following theorem contains the Dynamical Programming technique for armed bandits which will be used in order to solve our assignment problem.

Theorem 3.1. *There exists a strategy $\tau_{G,A}$ for every $A \in \mathcal{A}$ with finite horizon and every $G \in \mathcal{D}(\mathcal{D}^k)$ such that*

$$\left\{(G, A, z_1, \ldots, z_{m-1}) : \tau_{(G,A)}(z_1, \ldots, z_{m-1}) = i\right\}$$

is a measurable set of $\mathcal{D}(\mathcal{D}^k) \times \mathcal{A} \times \mathbb{R}^{m-1}$ for every $i = 1, 2, \ldots, k$ and $m = 1, 2, \ldots$ and that is optimal for the (G, A)-bandit. In addition, the function

$$(G, A) \to V(G, A) \tag{3.1}$$

is measurable and satisfies

$$V(G, A) = \vee_{i=1}^{k} \mathbb{E}\left[\alpha_1 X_{i1} + V((X_{i1})_i G, A^{(1)} | G)\right]. \tag{3.2}$$

The proof is made *via* mathematical induction on the horizon of the bandit and may be found in [3].

For computational purposes we write (3.2) as follows:

$$V(G, A) = \vee_{i=1}^{k} \int_{\mathcal{D}^k} \int_{\mathbb{R}} \left[\alpha_1 x + V((x)_i G, A)\right] Q_i(\mathrm{d}x) G\left(d(Q_1, \ldots, Q_k)\right) \tag{3.3}$$

$$= \vee_{i=1}^{k} \int_{\mathcal{D}} \int_{\mathbb{R}} \left[\alpha_1 x + V((x)_i G, A)\right] Q_i(\mathrm{d}x) F_i(\mathrm{d}Q_i). \tag{3.4}$$

In order to simplify expressions, we will write $G^{(m)}$ for representing conditional distributions in $\mathcal{D}^k$ after stage m.

If the (G, A)-bandit has horizon one then, since horizon of $A^{(1)}$ is zero, we have that for any distribution $G^{(1)}$

$$V(G^{(1)}, A^{(1)}) = 0, \tag{3.5}$$

hence (3.3) simplifies to:

$$V(G, (\alpha_1, 0, \ldots)) = \alpha_1 \vee_{i=1}^{k} \int_{\mathcal{D}^k} \int_{\mathbb{R}} x Q_i(\mathrm{d}x) G\left(d(Q_1, \ldots, Q_k)\right) \tag{3.6}$$

$$= \alpha_1 \vee_{i=1}^{k} \int_{\mathcal{D}} \int_{\mathbb{R}} x Q_i(\mathrm{d}x) F_i(\mathrm{d}Q_i). \tag{3.7}$$

An optimal selection is any arm for which the maximum is attained.

If the (G, A)-bandit has horizon 2, in order to use (3.3) we need the values of the bandits having $A^{(1)}$ as discount sequence. We observe that we only need to consider distributions of the form $(x)_i G$ for some possible observation x in arm i, the values $V((x)_i G, A^{(1)})$ are obtained with (3.6). When (3.3) is used for finding $V(G, A)$, the optimal initial selections are those indicated by the i's for which the maximum in (3.3) is attained.

In general, when we are in the presence of a (G, A)-bandit with horizon $n < \infty$ the first step is to calculate all the possible conditional distributions $G^{(m)}$ of $(Q_1, \ldots, Q_k)$ given the corresponding observations. Since $A^{(n-1)}$ has horizon one, every $V(G^{(n-1)}, A^{(n-1)})$ is obtained *via* (3.6). Next, every $V(G^{(n-2)}, A^{(n-2)})$ is calculated by (3.3). This process may be repeated until $V(G, A)$ is found.

4. Dynamic programming and the seller's assignment problem

In this section we will apply the Dynamic Programming technique to our problem of maximizing the randomly discounted gains of a store with a random number of clients. For this, we will consider two cases:

(1) There is only one product available at the store.
(2) There are many products for selling.

4.1. One product only

When there is only one product in the store, but sales are not sure (seller must convince to customer of buying the product), the price of such product (P) is not so important if we are only interested in optimal selections and not in the optimal gain. In this case the problem of making the biggest gain coincides with the problem of maximizing the number of sales.

We have that the corresponding element $G \in \mathcal{D}(\mathcal{D}^k)$ is supported by a subset of

$$(\text{Bernoulli distributions}) \times \ldots \times (\text{Bernoulli distributions}),$$

i.e., we may consider that each Q_i is a Bernoulli distribution (since every sale may or may not be made) with unknown associated parameter θ_i. Hence, for simplicity we may say that G is a distribution on $[0, 1]^k$.

Because of the arm's nature, we will write $(1)_i$ for representing a success in arm i (seller i succeeded in its sale) and $(0)_i$ for a failure in arm i (seller i did not make the sale).

Next, we present some examples of particular choices of G for illustrative purposes:

1) **Non independent arms.** Let us assume that after an evaluation of the "résumés" of two sellers who used to work together in another store, the interview prior the engagement and the training period, the manager concludes that both are excellent or very bad sellers and suppose also that the manager thinks that is much more probable the second possibility. Explicitly

$$G = \frac{4}{5}\delta_{(1/10,15/100)} + \frac{1}{5}\delta_{(9/10,8/10)}. \tag{4.1}$$

If we assume that only two customers will arrive at the store and that distribution H assigns probability $\frac{1}{2}$ to the discount sequences $(1, 1)$ and $(.9, .81)$ then $\mathbb{E}[A|H] = (\frac{19}{20}, \frac{181}{200})$.

First of all, we find posteriori distributions $G^{(2)}$ through the likelihood function and the prior distribution G [7]. Let us make only the details for finding $(1)_1 G$. We must find $a, b \in \mathbb{R}$ such that:

$$(1)_1 G = a\delta_{(1/10,15/100)} + b\delta_{(9/10,8/10)}. \tag{4.2}$$

We have from Bayes' law that

$$a = \mathbb{P}\left[\theta_1 = \frac{1}{10}, \theta_2 = \frac{15}{100} | (1)_1\right] = \frac{\frac{1}{10}\frac{4}{5}}{\frac{1}{10}\frac{4}{5} + \frac{9}{10}\frac{1}{5}}$$
$$= \frac{4}{13},$$

hence

$$(1)_1 G = \frac{4}{13}\delta_{(1/10,15/100)} + \frac{9}{13}\delta_{(9/10,8/10)}. \tag{4.3}$$

Via similar arguments we find that

$$(0)_1 G = \frac{36}{37}\delta_{(1/10,15/100)} + \frac{1}{37}\delta_{(9/10,8/10)} \tag{4.4}$$

$$(1)_2 G = \frac{3}{7}\delta_{(1/10,15/100)} + \frac{4}{7}\delta_{(9/10,8/10)} \tag{4.5}$$

$$(0)_2 G = \frac{17}{18}\delta_{(1/10,15/100)} + \frac{1}{18}\delta_{(9/10,8/10)}. \tag{4.6}$$

From (3.6), it can be seen that

$$\begin{aligned}
\frac{200}{181}V((1)_1 G, A^{(1)}) &= \left(\frac{4}{13}\frac{1}{10} + \frac{9}{13}\frac{9}{10}\right) \vee \left(\frac{4}{13}\frac{15}{100} + \frac{9}{13}\frac{8}{10}\right) = \frac{85}{130} \vee \frac{132}{130} \\
\frac{200}{181}V((0)_1 G, A^{(1)}) &= \left(\frac{36}{37}\frac{1}{10} + \frac{1}{37}\frac{9}{10}\right) \vee \left(\frac{36}{37}\frac{15}{100} + \frac{1}{37}\frac{8}{10}\right) = \frac{45}{370} \vee \frac{62}{370} \\
\frac{200}{181}V((1)_2 G, A^{(1)}) &= \left(\frac{3}{7}\frac{1}{10} + \frac{4}{7}\frac{9}{10}\right) \vee \left(\frac{3}{7}\frac{15}{100} + \frac{4}{7}\frac{8}{10}\right) = \frac{39}{70} \vee \frac{365}{700} \\
\frac{200}{181}V((0)_2 G, A^{(1)}) &= \left(\frac{17}{18}\frac{1}{10} + \frac{1}{18}\frac{9}{10}\right) \vee \left(\frac{17}{18}\frac{15}{100} + \frac{1}{18}\frac{8}{10}\right) = \frac{26}{180} \vee \frac{335}{1800}.
\end{aligned}$$

Through (3.3), we find that

$$V(G, A) = \frac{579}{968} \vee \frac{401}{869} = \frac{579}{968}.$$

Then, the optimal strategy is as follows

$$\tau(\phi) = 1 \tag{4.7}$$

and

$$\tau(z_1) = \begin{cases} 2\, z_1 = (1)_1 \\ 2\, z_1 = (0)_1. \end{cases}$$

In other words, seller number one must be chosen initially and no matter what his performance was, the second client will be assigned to seller number two even if he makes a sale. Then, in some sense in this example the optimal strategy does not depend on the observed performances.

2) **Arms with uniform distributions**. Let us consider the following example considered in [3]: now assume that there are three sellers at the store. Suppose in addition that third seller is known in the sense that we know his skills; *i.e.* θ_3 is known, for example $\theta_3 = 8/15$. Prior distribution of other two sellers is uniform in $[0,1]^2$, this may be seen as the situation in which no prior information about the skills of such sellers is available. Then, we have that $G = F_1 \times F_2 \times F_3$ where F_1 and F_2 are uniform distributions on $[0,1]$ and $F_3 = \delta_{8/15}$. Finally, suppose that the number of customers which arrive at the store is three. In order to keep this example as simple as possible, we assume that the discount sequence is non observable or observable in the multiplicative case explained in Section 2.3 and that $\mathbb{E}[A|H] = (1,1,1)$.

For being able to find the distributions $G^{(2)}$, we note that both θ_1 and θ_2 have Beta posterior distributions [7]. Let us illustrate the calculations for one of the values of $V(G^{(2)}, A^{(2)})$; for example, in order to find

	$(0)_1$	$(1)_1$	$(0)_1\,(0)_1$	$(0)_1\,(1)_1$	$(1)_1\,(1)_1$
	(1/2,1/2,8/15)	(1/3,1/2,8/15)	(2/3,1/2,8/15)	(1/4,1/2,8/15)	(1/2,1/2,8/15)	(3/4,1/2,8/15)
$(0)_2$	(1/2,1/3,8/15)	(1/3,1/3,8/15)	(2/3,1/3,8/15)			
$(1)_2$	(1/2,2/3,8/15)	(1/3,2/3,8/15)	(2/3,2/3,8/15)			
$(0)_2\,(0)_2$	(1/2,1/4,8/15)					
$(0)_2\,(1)_2$	(1/2,1/2,8/15)					
$(1)_2\,(1)_2$	(1/2,3/4,8/15)					

FIGURE 1. $m = 2$.

	$(0)_1$	$(1)_1$
	(198/180,198/180,192/180)	(156/180,198/180,192/180)	(242/180,210/180,216/180)
$(0)_2$	(198/180,156/180,192/180)		
$(1)_2$	(210/180,242/180,216/180)		

FIGURE 2. $m = 1$.

$V\left((1)_1(1)_1G, A^{(2)}\right)$ we must calculate and compare the following three integrals:

$$\begin{aligned} V\left((1)_1(1)_1G, A^{(2)}\right) &= \frac{\Gamma(4)\Gamma(2)}{\Gamma(3)\Gamma(1)^3}\int_0^1\int_0^1 \theta_1^3 \mathrm{d}\theta_1\mathrm{d}\theta_2 \\ &\vee \frac{\Gamma(4)\Gamma(2)}{\Gamma(3)\Gamma(1)^3}\int_0^1\int_0^1 \theta_1^2\theta_2 \mathrm{d}\theta_1\mathrm{d}\theta_2 \\ &\vee \int_{\mathbb{R}} x d\delta_{8/15} \\ &= \frac{3}{4}\vee\frac{1}{2}\vee\frac{8}{15} \end{aligned}$$

In addition, we observe the following commutativity:

$$V\left((a)_i(b)_jG, A^{(2)}\right) = V\left((b)_j(a)_iG, A^{(2)}\right),$$

for $a, b = 0, 1$ and $i, j = 1, 2, 3$. In addition, if first two clients are assigned to seller three then $V\left((a)_3(b)_3G, A^{(2)}\right) = \frac{1}{2}\vee\frac{1}{2}\vee\frac{8}{15} = \frac{8}{15}$, this can be seen in the blue cell in Figure 1. Green cells at the same figure help us to find $V\left((a)_3(b)_iG, A^{(2)}\right)$, for $a, b = 0, 1$ and $i = 1, 2$.

We may find now $V(G^{(1)}, A^{(1)})$. For example, for $V\left((1)_1G, A^{(1)}\right)$ we have that

$$\begin{aligned} V\left((1)_1G, A^{(1)}\right) &= \left(2\int_0^1 \left[0 + V\left((0)_1(1)_1G, A^{(2)}\right)\right](1-\theta_1) + \left[1 + V\left((1)_1(1)_1G, A^{(2)}\right)\right]\theta_1\mathrm{d}\theta_1\right) \\ &\vee \left(\int_0^1 \left[0 + V\left((0)_2(1)_1G, A^{(2)}\right)\right](1-\theta_1) + \left[1 + V\left((1)_2(1)_1G, A^{(2)}\right)\right]\theta_1\mathrm{d}\theta_1\right) \\ &\vee \left(\frac{7}{15}\left[0 + V\left((0)_3(1)_1G, A^{(2)}\right)\right] + \frac{8}{15}\left[1 + V\left((1)_3(1)_1G, A^{(2)}\right)\right]\right) \\ &= \frac{242}{180}\vee\frac{210}{180}\vee\frac{216}{180} = \frac{242}{180}. \end{aligned}$$

In Figure 2 we observe every value of $V(G^{(1)}, A^{(1)})$ where again, blue cell correspond to $V((1)_3G, A^{(1)})$ and $V((0)_3G, A^{(1)})$.

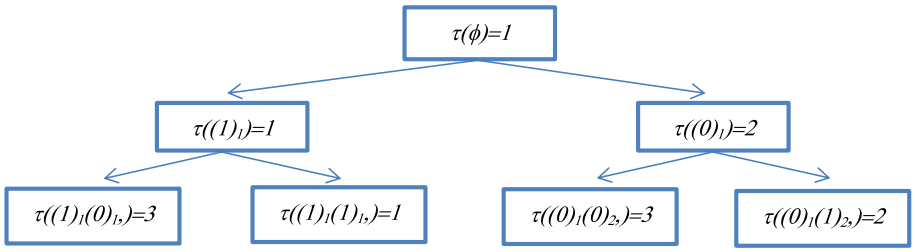

FIGURE 3. Optimal strategies with $\tau(\phi) = 1$.

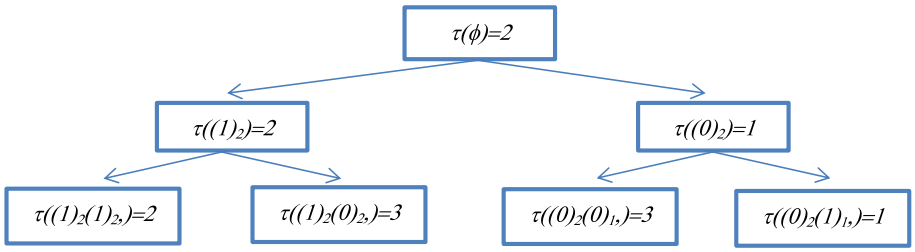

FIGURE 4. Optimal strategies with $\tau(\phi) = 2$.

Finally:

$$\begin{aligned} V(G,A) &= \int_0^1 \left[\left(0 + V((0)_1 G, A^{(1)})\right)(1-\theta_1) + \left(1 + V((1)_1 G, A^{(1)})\right)\theta_1\right] \mathrm{d}\theta_1 \\ &\vee \int_0^1 \left[\left(0 + V((0)_2 G, A^{(1)})\right)(1-\theta_2) + \left(1 + V((1)_2 G, A^{(1)})\right)\theta_2\right] \mathrm{d}\theta_2 \\ &\vee \frac{7}{15}\left(0 + V((0)_3 G, A^{(1)})\right) + \frac{8}{15}\left(1 + V((1)_3 G, A^{(1)})\right) \\ &= \frac{310}{180} \vee \frac{310}{180} \vee \frac{294}{180}. \end{aligned}$$

We have then, two optimal initial selections: $\tau(\phi) = 1$ and $\tau(\phi) = 2$. In Figures 3 and 4 we observe the resulting optimal strategies that, unlike the previous example, they depend on the performances of sellers.

For example, if seller number one is chosen for attending to first client and he makes a sale, then the second one will be assigned to the same seller but, if he does not make a second sale the third customer will be assigned to seller three.

4.2. More than one product

Let us present now some examples in which there is more than one product in stock; in other words, distributions Q_i are not necessarily of Bernoulli type.

1) Consider that we have two employees and that seller number one, always make a sale with the known probability mass function

$$Q_1 = \frac{3}{4}\delta_3 + \frac{1}{4}\delta_5 \tag{4.8}$$

and consequently $F_1 = \delta_{Q_1}$. For seller number two, let us consider that his sales follow a Binomial distribution with parameters 4 and θ, where θ is a uniform random variable over $(0, 1)$ (no information about his skills is available). Finally, we assume that $G = F_1 \times F_2$ and that $n = 3$.

In this case, we consider the random observable discount sequence

$$(U_1, U_1U_2, U_1U_2U_3)\,,$$

with U_1, U_2 and U_3 independent random variables with corresponding distributions H_1, H_2 and H_3 where H_1 is the Binomial$(3, 1/4)$ distribution (η_1 has a point of mass one), H_2 is a Bernoulli distribution uniformly distributed on the set of all Bernoulli distributions (η_2 may be seen as uniform on $(0,1)$) and H_3 has the following distribution:

$$\eta_3 = \frac{1}{2}\delta^1_{H_1} + \frac{1}{2}\delta^2_{H_1},$$

with

$$H^1_1 = \frac{1}{3}\delta_1 + \frac{2}{3}\delta_2 \text{ and } H^2_1 = \delta_{\frac{1}{3}}.$$

After some calculations, it may be found that $\mathbb{E}[A|H] = (\frac{3}{4}, \frac{1}{2}, \frac{19}{12})$.

In order to find the values of $V((k)_i(r)_jG, A^{(2)})$, we may consider the following cases:

Case 1. Both outcomes come from seller number two; *i.e.*, we must find $V((k)_2(r)_2G, A^{(2)})$, for $k, r = 0, 1, 2, 3, 4$. In this case, we may see that $(k)_2(r)_2F_2$ is Beta with parameters $k + r + 1$ and $9 - k - r$ [7]. Hence we must contrast the following integrals:

$$\frac{19}{12}\left[3 \times \frac{3}{4} + 5 \times \frac{1}{4} \vee \frac{1}{B(k + r + 1, 9 - k - r)} \int_0^1 8\theta_2\theta_2^{k+r}(1 - \theta_2)^{8-k-r}\mathrm{d}\theta_2\right] = \frac{133}{24} \vee \frac{19}{15}(k + r + 1),$$

where $B(\cdot, \cdot)$ stands for the Beta function. Hence, $V((k)_2(r)_2G, A^{(2)}) = \frac{133}{24} \vee \frac{19}{15}(k+r+1)$. We may easily conclude that if the sum of sales made by seller number two is at least \$4 then the maximum in (3.6) corresponds to him/her.

Case 2. If there is one assignation to each seller, then we must find

$$V((k)_1(r)_2G, A^{(2)})$$

for $r = 0, 1, 2, 3, 4$ and $k = 3, 5$. In this case:

$$V((k)_1(r)_2G, A^{(2)}) = \frac{133}{24} \vee \frac{19}{12} \times \frac{1}{B(r + 1, 5 - r)} \int_0^1 4\theta_2\theta_2^r(1 - \theta_2)^{4-r}\mathrm{d}\theta_2 = \frac{133}{24} \vee \frac{19}{18}(r + 1),$$

we conclude then, that the maximum in (3.6) always is attained by arm one.

Case 3. If both observed results come from arm one, we have to find

$$V((k)_1(r)_1G, A^{(2)})$$

for $k, r = 3, 5$. We observe that

$$V((k)_1(r)_1G, A^{(2)}) = \frac{133}{24} \vee \frac{19}{12} \times \int_0^1 4\theta_2\mathrm{d}\theta_2 = \frac{133}{24}.$$

In other words, if only arm one is observed, then the maximum in (3.6) is again reached by the arm one.

The above discussion may be summarized in Figures 5 and 6, where we omit the comparison between arms since the value associated with arm one is always the same (133/24). In addition, green cells represent those situations in which the maximum is reached by the arm one, and brown cells to those in which the value associated with arm two is bigger.

	$(0)_2$	$(1)_2$	$(2)_2$	$(3)_2$	$(4)_2$	$(0)_2(0)_2$	$(0)_2(1)_2$	$(0)_2(2)_2$	$(0)_2(3)_2$	$(0)_2(4)_2$
$(3)_1$	19/18	38/18	57/18	76/18	59/18	19/15	38/15	57/15	76/15	95/15
$(5)_1$	19/18	38/18	57/18	76/18	59/18					
$(3)_1(3)_1$	19/6									
$(5)_1(5)_1$	19/6									
$(3)_1(5)_1$	19/6									

FIGURE 5. $m = 2$.

$(1)_2(1)_2$	$(1)_2(2)_2$	$(1)_2(3)_2$	$(1)_2(4)_2$	$(2)_2(2)_2$	$(2)_2(3)_2$	$(2)_2(4)_2$	$(3)_2(3)_2$	$(3)_2(4)_2$	$(4)_2(4)_2$
57/15	76/15	95/15	114/15	95/15	114/15	133/15	133/15	152/15	171/15

FIGURE 6. $m = 2$.

In order to find $V((k)_iG, A^{(1)})$ for $i = 1, 2$ *via* (3.3), we find after some calculations that:

$$V((k)_1G, A^{(1)}) = \frac{175}{24} \vee \int_0^1 \sum_{j=0}^{4} \left[j/2 + V\left((j)_2(k)_1G, A^{(2)}\right)\right] \binom{4}{j}$$
$$\theta_2^j(1-\theta_2)^{4-j}\mathrm{d}\theta_2 = \frac{175}{24} \vee \frac{154}{24} = \frac{175}{24}.$$

In addition, we also have that

$$V((k)_2G, A^{(1)}) = \frac{175}{24} \vee \int_0^1 \sum_{j=0}^{4} \left[j/2 + V\left((j)_2(k)_2G, A^{(2)}\right)\right] \binom{4}{j} \theta_2^{k+j}(1-\theta_2)^{8-k-j}\mathrm{d}\theta_2$$
$$= \begin{cases} \dfrac{175}{24} & \text{if } k = 0, 1, 2 \\ 9.7966 & \text{if } k = 3 \\ 12.2037 & \text{if } k = 4. \end{cases}$$

We may deduce that if, seller number one is observed at stage one then, posterior customer has to be assigned to the same seller, but if we observed seller 2, we will continue with this arm only if a sale of \$3 or \$4 is made.

Finally, we find $V(G, A)$ from (3.3) as follows:

$$V(G, A) = \left[3\frac{3}{4} + V((3)_1G, A^{(1)})\right]\left(\frac{3}{4}\right) + \left[5\frac{3}{4} + V((5)_1G, A^{(1)})\right]\left(\frac{1}{4}\right)$$
$$\vee \int_0^1 \sum_{j=0}^{4} \left[j\frac{3}{4} + V((j)_2G, A^{(1)})\right] \binom{4}{j} \theta_2^j(1-\theta_2)^{4-j}\mathrm{d}\theta_2$$
$$= 10.2751 \vee 9.9167 = 10.2751.$$

From previous discussion, we find that the only optimal strategy is to choose arm one at every stage.

2) Let us now consider a store with three employees and assume that the number of customers that may come into the store is $1, 2, 3$ or 4, all of them with probability $1/4$ of occurrence. In addition, the discount sequence is deterministic with

$$A = \left(\frac{9}{10}, \frac{4}{3}\left(\frac{9}{10}\right)^2, 2\left(\frac{9}{10}\right)^3, 4\left(\frac{9}{10}\right)^4\right)$$

hence, the modified discount sequence is $(\frac{9}{10}, (\frac{9}{10})^2, (\frac{9}{10})^3, (\frac{9}{10})^4)$.

We may assume that sale's distribution is known for sellers one and two. On the one hand, seller number one makes no sale with probability 1/2, and with identical probability he makes a sale of \$3; *i.e.*

$$Q_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_3$$

and, on the other hand seller number two always earns \$2 for the store: $Q_2 = \delta_2$.

Sales distribution of employee three is similar to the seller's one in the sense that they are either \$3 or \$0. But, the probability of making no sale (p) follows the distribution

$$\frac{2}{3}\delta_{1/2} + \frac{1}{3}\delta_{4/5},$$

in other words,

$$F_3 = \frac{2}{3}\delta_{Q_3^1} + \frac{1}{3}\delta_{Q_3^2},$$

with

$$Q_3^1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_3 \quad \text{and} \quad Q_3^2 = \frac{4}{5}\delta_0 + \frac{1}{5}\delta_3.$$

Intuitively, there is a direct competition between sellers one and three that is won by the first one, since his probability of making a sale of \$3 is greater or equal than both options for seller number one. Then, there will be no surprises if optimal selections do not include seller number one. Hence, the real competition is between sellers one and two because there is not (at least at first sight) a clear winner.

Via Theorem 1 and a short recursive code (written, for example, in R) we find that at any stage, no matter what the observed results are and where they come from, the optimal strategy is always to choose seller number two. In particular, the optimal strategy selects all the to time to employee two.

However, if we set:

$$Q_2 = \delta_1$$

then all decisions are in favor of seller number one. But, what if we keep with this choice of Q_2 but the distribution of p is now

$$\frac{2}{3}\delta_{1/2} + \frac{1}{3}\delta_{1/5},$$

we observe that the probability of making a sale of \$3 is now 4/5 with probability 1/3. However, even that there are cases in which optimal selection is seller number one, the optimal strategy always selects to seller number two.

Finally, if we reconsider the distribution $Q_2 = \delta_2$ for seller number two and assume that parameter p is distributed according to

$$\frac{1}{3}\delta_{1/2} + \frac{2}{3}\delta_{1/5},$$

then the only optimal initial selection is seller three. In Figure 7, we present the optimal strategy if his initial sale is \$3.

If initially, seller three can not make a sale, then optimal strategy is: choose seller number two for attending next two customers and then assign the last one's attention to seller number three.

5. Conclusions

Machinery offered by the Dynamic Programming technique in the context of Armed Bandits has been exploited, in order to give a solution for an assignment problem concerning with sellers in a store, when all we know is the distribution of the distributions of sales associated with each seller. The goal is to maximize the randomly discounted gains with a random number of clients. Some examples are detailed in order to illustrate its utility.

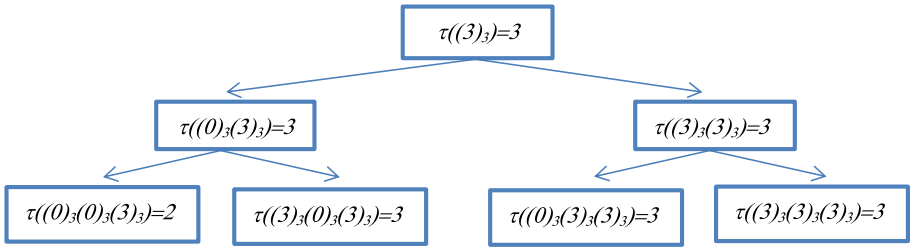

FIGURE 7. Optimal strategy when initial sale is \$3.

Acknowledgements. Authors would like to thank to anonymous reviewers for their constructive comments which helped to improve this paper substantially.

REFERENCES

[1] R. Bellman, On the Theory of Dynamic Programming. *Proc. of the National Academy of Sciences* (1952).
[2] D.A. Berry, Bandit Problems with random discounting, Mathematical learning. Models-Theory and Algorithms. Springer Verlag (1983).
[3] D.A. Berry and B. Fristedt, Bandit Problems. Chapman and Hall (1985).
[4] H. Cruz−Suárez, R. Ilhuicatzi−Roldán and R. Montes-de-Oca, Markov Decision Processes on Borel Spaces with Total Cost and Random Horizon. *J. Optimiz. Theory Appl.* **162** (2014) 329–346.
[5] D. Levhari and L.J. Mirman, Savings and consumption with uncertain horizon. *J.Political Econ.* **85** (1977) 265–281.
[6] K.R. Parthasarathy, Probability Measures on Metric Spaces. Academic Press (1967).
[7] J. Wakefield, Bayesian and Frequentist Regression Methods. Springer Verlag (2013).