11-2018

Optimal Policy for Sequential Stochastic Resource Allocation

Kalyanam Krishnamoorthy
InfoSciTex Corporation

Meir Pachter
Air Force Institute of Technology

David W. Casbeer
Air Force Research Laboratory

Recommended Citation

Krishnamoorthy, K., Pachter, M., & Casbeer, D. W. (2016). Optimal Policy for Sequential Stochastic Resource Allocation. Procedia Computer Science, 95, 483–488. https://doi.org/10.1016/j.procs.2016.09.325

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 95 (2016) 483 – 488

Procedia Computer Science

Complex Adaptive Systems, Publication 6
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2016 - Los Angeles, CA

Optimal Policy for Sequential Stochastic Resource Allocation

K. Krishnamoorthy[a,*], M. Pachter[b], D. Casbeer[c]

[a] *InfoSciTex Corporation, a DCS Company, Wright-Patterson A.F.B., OH 45433, USA*
[b] *Air Force Institute of Technology, Wright-Patterson A.F.B., OH 45433, USA*
[c] *Air Force Research Laboratory, Wright-Patterson A.F.B., OH 45433, USA*

Abstract

A gambler in possession of R chips/coins is allowed $N(> R)$ pulls/trials at a slot machine. Upon pulling the arm, the slot machine realizes a random state $i \in \{1, \dots, M\}$ with probability $p(i)$ and the corresponding positive monetary reward $g(i)$ is presented to the gambler. The gambler can accept the reward by inserting a coin in the machine. However, the dilemma facing the gambler is whether to spend the coin or keep it in reserve hoping to pick up a greater reward in the future. We assume that the gambler has full knowledge of the reward distribution function. We are interested in the optimal gambling strategy that results in the maximal cumulative reward. The problem is naturally posed as a Stochastic Dynamic Program whose solution yields the optimal policy and expected cumulative reward. We show that the optimal strategy is a threshold policy, wherein a coin is spent if and only if the number of coins r exceeds a state and stage/trial dependent threshold value. We illustrate the utility of the result on a military operational scenario.

Keywords: Resource Allocation; Stochastic Optimization; Threshold Policy

1. Introduction

We are interested in the optimal sequential allocation of R resources to a system over N stages, where $R < N$.

* Corresponding author. Tel.: +1-937-713-7017.
E-mail address: krishnak@ucla.edu

doi:10.1016/j.procs.2016.09.325

At each stage, no more than 1 resource can be allocated to the system. The system state, $s \in S = \{1, \dots, M\}$ evolves randomly and at each stage, $p(s) > 0$ is the probability that the system state will be s. If the system is at state s and a resource is allocated to the system, then an immediate reward, $g(s) > 0$, is gained. We wish to compute the optimal allocation that results in the maximal cumulative reward.

The problem considered herein is a special case of the Sequential Stochastic Assignment Problem (SSAP)[1]. The SSAP deals with the assignment of N *differently abled* men to N jobs that arrive sequentially. The fitness of the i^{th} man is given by $m_i; 0 \le m_i \le 1$. Associated with job $j \in \{1, \dots, N\}$ is a random variable X_j that takes on the value x_j. The value/ reward associated with the assignment of the i^{th} man to job j is given by the product $m_i x_j$. The $X_j; j = 1, \dots, N$ are i.i.d. random variables with a known distribution. The goal is to maximize the total expected reward. In our simplified setting, the $R(< N)$ men are identical. The solution in Ref. 1 can therefore be applied by assigning $m_i = 1,\ i = 1, \dots, R$ and $m_i = 0,\ i = (R+1), \dots, N$. Moreover, in the resource allocation setting we consider, the continuous valued random variable X_j is replaced by a discrete valued random variable (with known distribution) that takes values from the finite set: $\{g(1), \dots, g(M)\}$. Optimal and asymptotically optimal decision rules for general resource allocation problem and its connection to the SSAP are discussed in Ref. 2. Finitely valued random rewards are also considered in Ref. 3; but the time between successive pulls is modelled as a renewal process and the performance metric is a (exponentially) discounted sum of rewards. In our work, we consider a simpler model with no discounting; thereby rendering the time between successive pulls irrelevant. In doing so, we uncover a structurally elegant solution. A related work[4] considers the problem of optimal donor-recipient assignment in live-organ transplants. Optimal sequential inspection polices that deal with allocation of a continuous valued decision variable (fuel/ time) is considered in Ref 5-6; therein a threshold policy is shown to be optimal as well. For a military operational scenario that involves optimal inspection of sequential targets, see Ref. 7.

Let $V(k, r, s)$ indicate the maximal cumulative reward (“payoff to go") at stage k, when the system state is s with $r(> 0)$ resources in hand. It stands to reason that $V(k, r, s)$ satisfies the Bellman recursion:

$$V(k, r, s) = \max_{u=0,1} \ \left\{\overline{V}(k+1, r), g(s) + \overline{V}(k+1, r-1)\right\},\ s \in S, 1 \le k < N, \qquad (1)$$

where the average return: $\overline{V}(k, r) = \sum_{x=1}^{M} p\,(x) V(k, r, x)$. The decision variable $u = 0{,}1$ indicates the number of resources allocated to the system at stage k. The optimal decision is therefore given by:

$$u(k, r, s) \quad = \begin{cases} 1, \text{if } g(s) \ge \overline{\Delta}(k+1, r), \\ 0, \text{otherwise}, \end{cases} \quad s \in S, r > 0{,}1 \le k < N.$$

where the marginal expected reward obtained by allocating an additional resource over and above $r - 1$ resources to the downstream stages $k + 1$ to N is given by:

$$\overline{\Delta}(k+1, r) \quad = \overline{V}(k+1, r) - \overline{V}(k, r-1).$$

The boundary condition for the recursion (1) is given by:

$$V(N, r, s) \quad = \begin{cases} 0, \ r = 0, \\ g(s), \ r \ge 1. \end{cases}, s = 1, \dots, M.$$

$$\Rightarrow \overline{V}(N, r) \quad = \begin{cases} 0, \ r = 0, \\ \overline{g}, \ r \ge 1, \end{cases}$$

where the average reward, $\overline{g} = \sum_{x=1}^{M} p\,(x) g(x)$.

2. Monotonic marginal reward

Lemma 1. For $k = 1, \dots, (N-1)$, we have:

$$0 = \overline{\Delta}(k+1, N-k+1) < \dots < \overline{\Delta}(k+1{,}1).$$

Proof. We show the result by backward induction on k. By definition,

$$\overline{\Delta}(N, r) \quad = \begin{cases} \overline{g}, \ r = 1, \\ 0, \ r = 2, \end{cases}$$

and so, $0 = \overline{\Delta}(N, 2) < \overline{\Delta}(N, 1) = \overline{g}$.

Let us assume that for some $k = 2, \dots, (N-2)$:

$$0 = \overline{\Delta}(k+1, N-k+1) < \dots < \overline{\Delta}(k+1{,}1). \qquad (2)$$

In other words, the marginal reward, $\overline{\Delta}(k+1, r)$, is a monotonic decreasing function of r with finite support. Given the monotonicity property, let the threshold $\gamma(k,s)$ be the smallest positive integer j such that $g(s) \geq \overline{\Delta}(k+1,j)$. Recall that the optimal policy is given by:

$$u(k,r,s) \quad = \begin{cases} 1, \text{if } g(s) \geq \overline{\Delta}(k+1,r), \\ 0, \text{otherwise}, \end{cases} \quad s \in S, r > 0, 1 \leq k < N.$$

It follows that:

$$u(k,r,s) \quad = \begin{cases} 1, \text{if } r \geq \gamma(k,s) \\ 0, \text{otherwise}, \end{cases}, s = 1, \dots, M.$$

Accordingly, the maximal reward satisfies:

$$V(k,r,s) = \quad \begin{cases} g(s) + \overline{V}(k+1,r-1),\ r \geq \gamma(k,s), \\ \overline{V}(k+1,r), r < \gamma(k,s), \end{cases} \quad s = 1, \dots, M.$$

Let $\Delta(k,r,s) = V(k,r,s) - V(k,r-1,s)$. It follows that:

$$\Delta(k,r,s) \quad = \begin{cases} \overline{\Delta}(k+1,r),\ r < \gamma(k,s), \\ g(s),\ r = \gamma(k,s), \\ \overline{\Delta}(k+1,r-1),\ r > \gamma(k,s). \end{cases} \tag{3}$$

From the definition of the threshold value $\gamma(k,s)$, we have:

$$\overline{\Delta}(k+1, \gamma(k,s) - 1) > g(s) \geq \overline{\Delta}(k+1, \gamma(k,s)).$$

Also, from (3), we have:

$$\Delta(k, N-k+2, s) = \overline{\Delta}(k+1, N-k+1) = 0. \tag{4}$$

So, combining (2), (3) and (4), we have:

$$0 = \Delta(k, N-k+2, s) < \cdots < \Delta(k,1,s),\ s = 1, \dots, M.$$

Since $\overline{\Delta}(k+1,r) = \sum_{x=1}^{M} p\,(x)\Delta(k+1,r,x)$, and probability $p(x) \geq 0$, it follows that:

$$0 = \overline{\Delta}(k, N-k+2) < \cdots < \overline{\Delta}(k,1).$$

The above result shows that the optimal policy is structured and is in fact a *control limit* policy. The state and stage dependent threshold is given by $\gamma(k,s)$. Structured policies are appealing to decision makers in that they are easy to implement and often enable efficient computation - for details, see Sec 4.7.1 of Ref. 8. Applying Lemma 1 to the most and least profitable states, we get the following result.

Corollary 1. If $g(\overline{s}) = \max_{s \in S} g(s)$ and $g(\underline{s}) = \min_{s \in S} g(s)$, then $\gamma(k, \overline{s}) = 1$ and $\gamma(k, \underline{s}) = N - k + 1$.
In other words, for the state with the highest reward, it is always optimal to assign a resource (if available). On the other hand, for the least profitable state, it is optimal to assign a resource if and only if the number of resources is greater than the number of stages/trials left i.e., if $r > N - k$. So, for the simple case of 2 states, i.e., $M = 2$, the resulting optimal policy is trivial and requires no computation whatsoever. This simple result will be applied to the practical scenario considered later. For $M > 2$, we wish to establish a direct recursion equation to compute the threshold values. In doing so, we circumvent solving for the value function and somewhat alleviate the curse of dimensionality associated with Dynamic Programming.

2.1. Direct recursion for generating the partitions

For $r = 1, \dots, (N-k+2)$, we have the marginal expected reward given by:

$$\begin{aligned} \overline{\Delta}(k,r) \quad &= \textstyle\sum_{s=1}^{M} \Delta\,(k,r,s)p(s) \\ &= \bar{\Delta}(k+1,r) \textstyle\sum_{x \in S_3^k} p(x) + \bar{\Delta}(k+1,r-1) \textstyle\sum_{x \in S_2^k} p(x) + \textstyle\sum_{x \in S_1^k} g(x)p(x), \end{aligned} \tag{5}$$

where the sub-sets:

$$S_1^k = \{s : \overline{\Delta}(k+1,r-1) > g(s) \geq \overline{\Delta}(k+1,r)\}, S_2^k = \{s : g(s) \geq \overline{\Delta}(k+1,r-1)\}, S_3^k = \{s : g(s) < \overline{\Delta}(k+1,r)\}.$$

Note that we arrived at the recursion (5) by substituting for $\Delta(k,r,s)$ from (3). So, we have established a direct recursion from $\overline{\Delta}(k+1,r)$ to $\overline{\Delta}(k,r)$ with the boundary condition given by:

$$\overline{\Delta}(N,r) \quad = \begin{cases} \overline{g}, & r = 1, \\ 0, & r = 2. \end{cases}$$

The optimal *threshold* policy is given by:

$$u(k,r,s) \quad = \begin{cases} 1, \text{if } r \geq \gamma(k,s), \\ 0, \text{otherwise}, \end{cases}, \ r = 1, \dots, (N-k+2).$$

As before, $\gamma(k,s)$ is the smallest positive integer j such that $g(s) \geq \overline{\Delta}(k+1,j)$.

2.2. Single coin case

Suppose the casino provides a coin for "free" and charges the gambler c_N for the N trials purchased. This would be the special case where $R = 1$. Indeed, we can drop the dependence on r and let v_k indicate the maximal expected cumulative reward with k trials to go. So, $v_1 = \overline{g}$ and

$$v_k = \quad v_{k-1}(1 - P_{k-1}) + \sum\nolimits_{x \in I_{k-1}} g(x)p(x), k > 1,$$

where the set I_{k-1} and probability P_{k-1} are given by:

$$I_{k-1} = \{x | g(x) \geq v_{k-1}\} \text{and } P_{k-1} = \sum\nolimits_{x \in I_{k-1}} p(x) .$$

The casino should charge $c_N > v_N$ for it to remain profitable. With k trials to go, let T_k be the average number of pulls/ trials expended before the coin/resource is spent. It follows that:

$$T_k \quad = P_{k-1} + (1 - P_{k-1})(1 + T_{k-1}).$$

In other words, with k trials available, the coin is either spent now with probability P_{k-1} or after $1 + T_{k-1}$ trials with probability $1 - P_{k-1}$. The boundary condition is given by: $T_1 = 1$. The gambler can take into consideration three factors before purchasing N trials: 1) the expected return, v_N, 2) cost, c_N and time spent in completing the T_N trials.

2.3. Heterogeneous coins case

Suppose we have N different coins ordered such that the immediate reward upon using coin i at state $s \in S$ yields the reward $m_i g(s)$, where $m_1 < m_2 < \cdots < m_N$. We wish to determine the optimal assignment of coins with N pulls/trials to go such that the expected cumulative reward is a maximum. As mentioned earlier, the scenario considered herein is a variation of the SSAP[1]. So, the results therein apply here. In particular, we state below the relevant result i.e., Theorem 1 in Ref. 1, as it applies to our discrete valued problem.

Theorem 1. There exist numbers:

$$0 = a_{0,N} < a_{1,N} < \cdots < a_{N,N} = \infty,$$

such that when there are N stages to go, the optimal choice in the 1^{st} stage is to use the i^{th} coin if the 1^{st} stage reward, $g(s_1) \in [a_{i-1,N}, a_{i,N})$. The $a_{i,N}$ depend on the probabilities, $p(x)$, but are independent of m_i's. Furthermore, the $a_{i,n}; i = 1, \dots, N$ are computed via the recursion below:

$$\begin{aligned} a_{i,n+1} \quad &= \textstyle\sum_{x \in I_{i,n}} g(x)p(x) + a_{i-1,n} Prob\{g(x) < a_{i-1,n}\} \\ &+ a_{i,n} Prob\{g(x) \geq a_{i,n}\}, \end{aligned} \qquad (6)$$

where, $I_{i,n} = \{x | a_{i-1,n} \leq g(x) < a_{i,n}\}$.

With the association: $k \to N - n + 1$ and $r \to n - i + 1$, it is easy to show that:

$$a_{i-1,n} \quad = \overline{\Delta}(k+1, r), \ i = 1, \dots, n.$$

Therefore, the recursive equations (5) and (6) are equivalent.

3. Military Application

A bomber travels along a designated route/ path and sequentially encounters enemy target sites numbered 1 to N on the ground. Upon reaching a target site, the bomber is provided feedback information on the nature of the enemy site. This could come from an Automatic Target Recognition (ATR) module on-board the vehicle or a human operator looking at the target site via an on-board camera. We assume that the feedback sensor/ classifier is error-prone and a and b respectively indicate the probabilities that a True and False Target are correctly identified. The bomber equipped with $R(< N)$ homogenous weapons can either deploy a weapon at the current location or keep it in reserve for future sites. We stipulate that the bomber gains a reward of 1 if it destroys a True Target and 0 otherwise. We are interested in the optimal weapon allocation (feedback) strategy that maximizes the expected cumulative reward.

3.1. Error-prone classifier

The *imperfect* classifier in the feedback path identifies the target site to be either a True or a False Target. Let the random variable $x \in X = \{T, F\}$ specify whether a target site contains a True Target, T or False Target, F. Let the classifier decision, $y \in X$ specify whether the target site is identified to be a True or False Target. Consider an environment where the true target density, i.e., a priori probability that a target site is a True Target, $P\{x = T\} = \alpha$, where $0 < \alpha < 1$. The conditional probabilities, which specify whether the classifier correctly identified True and False Targets, are given by:

$$a := P\{y = T | x = T\} \text{ and } b := P\{y = F | x = F\}.$$

Together, a and b determine the entries of the binary *confusion matrix* (see Table 1) of the classifier.

Table 1. Classifier confusion matrix.

	Target site	
Classifier decision	Target	False target
Target	a	$1 - b$
False target	$1 - a$	b

Suppose the classifier decision is T. From Bayes' rule, the a posteriori probability that the target site is a True Target is given by:

$$g(T) := P\{x = T | y = T\} \quad = \quad \frac{\alpha a}{p(T)},$$

where $p(T) = \alpha a + (1 - \alpha)(1 - b)$ is the probability that the classifier's decision is T. On the other hand, if the classifier decision is F, the a posteriori probability that the target site is a True Target is given by:

$$g(F) := P\{x = T | y = F\} \quad = \quad \frac{\alpha(1 - a)}{p(F)},$$

where $p(F) = \alpha(1 - a) + (1 - \alpha)b$ is the probability that the classifier's decision is F.

We make the following standard assumption regarding the Type I and II error rates.

Assumption 1. $a > 1 - b$.

The above assumption implies that the classifier is more likely to correctly classify a True Target than misclassify a False Target. Also, when $\alpha = 0.5$, the probability of correct classification, $a\alpha + b(1 - \alpha) > 0.5$ i.e., the outcome is better than a random guess, which is intuitively appealing. We shall show that, under this assumption, the optimal decision takes a remarkably simple form, i.e., bomb a site if and only if the classifier identifies it to be a True Target. Thereafter, we shall also highlight how the optimal solution changes, when this assumption is violated.

To reconcile the application scenario with the model considered earlier, we note that there are only two states, i.e., $y \in S = \{T, F\}$. The probabilities that $y = T, F$ are given by $p(T)$ and $p(F)$ respectively and the reward associated with the two states are given by $g(T)$ and $g(F)$ respectively. Under Assumption 1, we show that the reward function satisfies the following property.

Lemma 2.

$$0 < g(F) < \alpha < g(T).$$

Proof. From Assumption 1, we have:

$$\begin{aligned}
\beta &= a + b - 1 > 0, \\
\Rightarrow \alpha\beta &> \alpha^2\beta, \text{since } \alpha < 1, \\
\Rightarrow \alpha\beta + \alpha(1-b) &> \alpha^2\beta + \alpha(1-b), \\
\Rightarrow g(T) = \frac{\alpha\beta + \alpha(1-b)}{\alpha\beta + (1-b)} &> \alpha.
\end{aligned}$$

A similar argument shows that $g(F) < \alpha$ and by definition, $g(F) > 0$.

Lemma 2 implies that the classifier is reliable in that its output nudges the a posteriori probability in the right direction.

3.2. Optimal bombing strategy

Suppose the bomber is at the k^{th} (out of N) target site. Since $g(T) > g(F)$, Corollary 1 tells us that the corresponding threshold values, $\gamma(k,T) = 1$ and $\gamma(k,F) = N - k + 1$. In other words, it is optimal to bomb a target site k only if either:

1. The site is identified to be a True Target or
2. The number of weapons in hand is greater than the number of target sites/stages left to visit.

In light of the above policy, the expected maximal cumulative reward is given by:

$$V = \sum_{k=0}^{R} \binom{N}{k} p(T)^k p(F)^{N-k} (kg(T) + (R-k)g(F)) + Rg(T) \sum_{k=R+1}^{N} \binom{N}{k} p(T)^k p(F)^{N-k}.$$

The above calculation is based on the optimal strategy which yields a reward of $kg(T) + (R-k)g(F)$ when k out of the N trials yield in a positive (True Target) identification. We sum over all possible k wherein the cumulative reward associated with each k is multiplied by the probability of occurrence of k True Target identifications out of N sites.

Suppose Assumption 1 is not true and $a > 1 - b$. It is trivial to show that $g(F) > g(T)$ and so, the optimal strategy is reversed in that it is optimal to bomb a site only if it is identified to be a False Target. This seemingly strange result is due to the classifier being a counter indicator or a reliable liar! Finally, if $a = 1 - b$, the classifier is useless since $g(F) = g(T) = \alpha$. So, any policy is optimal and will result in the expected cumulative reward, $R\alpha$.

4. Conclusion

We consider a variant of the Sequential Stochastic Assignment Problem (SSAP), wherein the rewards for incoming jobs are drawn from a discrete (finitely valued) distribution and the men assigned to do the job are identical. We show that an available resource (man) is assigned to an incoming job if and only if the number of resources left is no less than a state and stage dependent threshold value. In doing so, we uncover an interesting structure in the optimal policy. For the special case where the incoming jobs are of two types only, the policy becomes trivial in that an available resource is only assigned to the more profitable state except when there are more resources available than jobs left to process. This result is applied to an operational military example; where the optimal policy is to bomb a true target (site) so long as a reliable classifier is used to identity the site.

References

1. Derman, C., Lieberman, G.J., Ross, S.M. A sequential stochastic assignment problem. *Management Science* 1972;**18**(7):349–357.
2. Pronzato, L. Optimal and asymptotically optimal decision rules for sequential screening and resource allocation. *IEEE Transactions on Automatic Control* 2001;**46**(5):687–697.
3. David, I., Levi, O. A new algorithm for the multi-item exponentially discounted optimal selection problem. *European Journal of Operational Research* 2004;**153**:782–789.
4. David, I., Yechiali, U. One-attribute sequential assignment match processes in discrete time. *Operations Research* 1995;**43**(5):879–884.
5. Pachter, M., Chandler, P., Darbha, S. Optimal sequential inspection. In: *IEEE Conference on Decision and Control*. San Diego, CA; 2006, p. 5930–5934.
6. Pachter, M., Chandler, P., Darbha, S. Optimal MAV operations in an uncertain environment. *International Journal of Robust and Nonlinear Control* 2008;**18**(2):248–262.
7. Kalyanam, K., Pachter, M., Patzek, M., Rothwell, C., Darbha, S. Optimal human-machine teaming for a sequential inspection operation. *IEEE Transactions on Human-Machine Systems* 2016;URL: http://dx.doi.org/10.1109/THMS.2016.2519603.
8. Puterman, M.L. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience; 1994.