Time-Constrained Restless Bandits and the Knapsack Problem for Perishable Items (Extended Abstract) [1]

Peter Jacko [2] and José Niño-Mora [3]

Department of Statistics
Universidad Carlos III de Madrid
Leganés (Madrid), Spain

Abstract

Motivated by a food promotion problem, we introduce the *Knapsack Problem for Perishable Items* (KPPI) to address a dynamic problem of optimally filling a knapsack with items that disappear randomly. The KPPI naturally bridges the gap and elucidates the relation between the PSPACE-hard restless bandit problem and the NP-hard knapsack problem. Our main result is a problem decomposition method resulting in an approximate transformation of the KPPI into an associated 0-1 knapsack problem. The approach is based on calculating explicitly the marginal productivity indices in a generic finite-horizon restless bandit subproblem.

Keywords: knapsack problem, perishable items, restless bandits, finite horizon, index policy heuristic

[1] This research has been supported in part by the Spanish Ministry of Education and Science under grant MTM2004-02334 and an associated Postgraduate Research Fellowship awarded to the first author, by the Autonomous Community of Madrid-UC3M through grant UC3M-MTM-05-075, and by the European Union's Network of Excellence Euro-NGI.
[2] Email: peter.jacko@uc3m.es
[3] Email: jose.nino@uc3m.es

1 Introduction

1.1 Motivating Problem: Food Promotion

The strong quality control in the food merchandizing industry requires that perishable products cannot be sold after their "best before" date. In order to sell the products before they perish, merchandisers can decide to reallocate them to a promotion space (e.g., close to the cash desk), where they are likely to attract more customers. We address the problem of filling the promotion space so that the expected lost profit is minimal.

1.2 Our Model and Knapsack Problem Literature

In this brief communication we deal with the *Knapsack Problem for Perishable Items* (KPPI) when the items are mutually independent. If referring to the food promotion problem, this is the case with one unit of each product where all the products have independent demands. The model is the simplest case of the problem allowing to present the main ideas as clearly as possible, and it itself is of practical interest. The analysis of a more general model (allowing for multiple items with a common demand process) will be published elsewhere.

A *perishable item* is an item with an associated *deadline* (e.g., the "best before" date), when the item perishes and a cost is incurred. A *favorable event* can happen before the deadline (e.g., selling the product), causing that the item disappears immediately, and consequently, the deadline cost is avoided. The probability of this favorable event only depends on whether the item is in the knapsack (e.g., being promoted), or not. By being the items mutually independent we mean their favorable events being mutually independent.

The question is then to select regularly a subset of items for the knapsack (e.g., promotion space) so that the total expected cost is minimized. A formal statement of the KPPI model is given in the next section. In general, the KPPI defines a stochastic variant of the knapsack problem with multiple units of items. As time evolves, some items disappear randomly and some items perish deterministically at their deadlines.

We have not found any literature that would deal with such problem. There exists work on the so-called *Dynamic and Stochastic Knapsack Problem* (DSKP), which is, however, different in nature [see, e.g. 3]. The DSKP is a problem of finding an online rule to immediately reject or accept arriving items with a random value and/or random weight (which never disappear). Thus, the DSKP is an admission control problem, whereas the KPPI deals with items whose values change over time (become worthless when disappear).

A natural mathematical setting for the KPPI problem is the multi-armed bandit framework [cf. 2], in which one wants to dynamically choose between various *bandits* (reward-yielding processes) one in an optimal fashion. That model captures the fundamental trade-off between *exploitation* of current rewards and *exploration* of possible future rewards.

The bandits are now items, with a single negative reward at their deadline. In our model, however, there are four complications: the bandits are *restless*, because they can disappear regardless of being in the knapsack or not, the time horizon is *finite*, and we are to select *more than one* item for the knapsack, which is allowed to be *filled partially*, due to the heterogeneity of the items. We thus mix up two models: the bandit problem and the knapsack problem.

The restless bandit problem (with infinite horizon) has been proven to be PSPACE-hard even in its deterministic version [8]. For the analysis we use the framework and methodology proposed for restless bandits by Niño-Mora [5, 6], modeling the problem as a Markov decision process. That work provided a sufficient condition for a restless bandit to be *indexable* together with an adaptive-greedy algorithm which in $\mathcal{O}(n^3)$ computes corresponding *marginal productivity indices* (MPIs) that extend earlier indices of Gittins [classical bandits, 1] and Whittle [restless bandits, 10].

The indexability property of a single item modeled as a restless bandit means that the optimal solution is to select the item for the knapsack whenever its index is higher than the cost of space occupation. When coupling the bandits back into a multi-armed bandit problem, the indices define an optimal policy: At every decision moment choose a bandit with highest index (*index policy*). Index policy is in general not optimal for the restless case, in which it becomes a well-grounded, efficient and practical heuristic.

Regarding the bandit problems with finite horizon, interesting results of index nature appear very sporadically, because of the intractability of such model, and therefore other methods (such as dynamic programming) are usually used. Even then, the problem is computationally intractable. Nevertheless, there is a tractable instance, the so-called *deteriorating case*, first presented for an infinite-horizon bandit problem by Gittins [1], which was also successfully applied in a problem with finite-horizon objective [4]. In that setting, the bandits, however, were not restless. The same is the case for the index policies for the finite-horizon multi-armed bandit problem: Niño-Mora [7] showed that such problem is indexable and provided an $\mathcal{O}(Tn^2)$ algorithm (where T is the time horizon) to compute the corresponding index (MPI).

1.4 Paper Overview

We develop a method of solving the KPPI approximately by solving a 0-1 knapsack problem. Thought being NP-hard, the last decade provided a number of techniques which have led to the development of fast algorithms capable to find an optimal solution of large knapsack problems in a few seconds [cf. 9].

The problem we analyze has a variety of applications, since the items considered in knapsack problems are often perishable, either naturally or due to some restrictions. Other examples of product promotion include fashion and seasonal goods or hotel rooms, which need to be sold by a specific time. In warfare operations, the KPPI captures the situation when a number of attacking units must be destroyed before they arrive to a particular area and damage it. Another application arises in surgery, when only a limited number of patients (e.g., waiting for a transplantation) may be chosen to undertake an alternative treatment. Further, the task management problem where tasks have associated deadlines falls to the general KPPI setting.

Our contributions include (1) a new problem decomposition method, which results in an approximate transformation of the KPPI into a deterministic 0-1 knapsack problem, (2) a heuristic with a very good performance based on that decomposition, (3) the first results for the restless bandit problem with finite horizon, (4) a closed-form formula for the marginal productivity index.

2 Knapsack Problem for Perishable Items

In this section we describe formally the simplest case of the KPPI. Suppose we have I mutually independent items ($I \geq 2$). The items are assumed to be perishable, that is, if item i has not disappeared within T_i periods, a cost c_i at the deadline T_i is incurred ($T_i \geq 1, c_i > 0$). Item i occupies space w_i (positive integer) and can be either *rested* (left where it is) or *replaced* to a knapsack (common for all the items). The knapsack has a restricted integer capacity $W > 0$. To avoid trivial cases we assume that $w_i \leq W$ for all $i = 1, 2, \ldots, I$ and $\sum_{i=1}^{I} w_i > W$. Decisions are made at time epochs $0, 1, \ldots, T-1$, where $T = \max\{T_1, T_2, \ldots, T_I\}$ is the problem's time horizon. The costs are discounted by a one-period discount factor β ($0 < \beta \leq 1$).

Item i disappears with probability $1 - q_i$ before the next period if it is rested, or with probability $1 - p_i$ if it is in the knapsack. We assume that selecting an item for the knapsack increases the probability of the favorable event (disappearing of the item), i.e. $q_i > p_i$. Otherwise, it is always optimal to leave the item rested.

One might formulate the KPPI as a dynamic program, discovering that the complexity of the resulting system of equations grows exponentially with the number of items, and thus becomes intractable quickly. Therefore, we apply a different approach, as presented next.

2.1 Restless Bandit Formulation

We will show how the KPPI can be formulated as a multi-armed restless bandit problem, using the framework by Niño-Mora [5, 6]. Each item i is defined as a Markov decision chain as follows. The state space $\mathcal{X}_i$ contains $T_i + 2$ states, T_i of which are *controllable* ($\mathcal{T}_i = \{-T_i, \ldots, -2_i, -1_i\}$), and the remaining two, 0_i and u_i, are *uncontrollable*.

A controllable state $-t_i$ refers to the situation t periods before the deadline, when the item is still available (not disappeared). Thus, there are two possible actions: *resting* the item or *selecting* the item for the knapsack, and no *immediate reward* is associated to either action ($r_i(-t_i) = 0$). The transition probability corresponds to the probability of disappearing, so the transition is defined as follows: if the item disappears, the state changes to the absorbing terminal state u_i; not disappearing refers to the state change to $-(t-1)_i$.

The uncontrollable state 0_i represents the act of perishing of the item, when an immediate reward $r_i(0_i) = -c_i$ is incurred, and the state moves to the terminal state u_i, for which $r_i(u_i) = 0$. For uncontrollable states only one action (resting) is available, that is, no decision is to be made.

In the restless bandit framework [cf. 5], the *immediate work* is assumed to be 1 for the active action and 0 for being passive. In our model the active action (selecting) does not require a uniform utilization of the knapsack, so we need to reflect this feature in the model. Therefore, we define the active immediate work $w_i^1(-t_i)$ of an item by its volume w_i (the immediate work for resting an item remains zero). The restless bandit formulation follows,

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{s=0}^{\infty} \beta^s \sum_{i \in \mathcal{I}} r_i(x_i(s)) \right]$$
$$\text{subject to} \quad \sum_{i \in \mathcal{I}^1(s)} w_i^1(x_i(s)) \leq W \text{ at each time } s = 0, 1, \ldots, \infty \qquad \text{(RB)}$$

where $\mathcal{I}^1(s)$ is the set of all items that are selected in time epoch s. Further, $x_i(s)$ denotes the state of item i at time epoch s, starting at state $x_i(0) = -T_i$.

The standard (approximate) solution method for restless bandit problems is to make the so-called *Whittle's relaxation*, or Lagrangian relaxation, and

calculate the marginal productivity indices for each item separately. Whittle [10] proposed the following heuristic for the standard framework with a uniform resource utilization ($w_i = 1$): "select the W items with the highest indices". Since the marginal productivity index measures the value of selecting the item instead of resting it, we propose the following heuristic for the KPPI: "select the items that are given by an optimal solution to a knapsack problem with MPIs as the objective function value coefficients and w_i's as the knapsack constraint weights". In particular, we define the value of item i as $v_i = \nu^*_{-T_i}$ (see the next section) and solve the following 0-1 knapsack problem

$$\begin{aligned} \max_{\boldsymbol{x}} & \sum_{i \in \mathcal{I}} v_i x_i \\ \text{subject to} & \sum_{i \in \mathcal{I}} w_i x_i \leq W \\ & x_i \in \{0,1\} \text{ for all } i \in \mathcal{I} \end{aligned} \tag{KP}$$

where $\boldsymbol{x} = (x_i : i \in \mathcal{I})$ is the vector of binary decision variables denoting whether the item i is selected for the knapsack or not. Note that the heuristic proposed by Whittle is indeed an optimal solution of the knapsack problem (KP), when all the weights w_i are uniform.

Proposition 2.1 (Reduction of KPPI to KP) *If the deadline $T_i = 1$ for all $i \in \mathcal{I}$, then any optimal solution $\boldsymbol{x}^*$ of the knapsack problem (KP) is an optimal solution of the KPPI.*

2.2 Calculation of Marginal Productivity Index

In this section we analyze a perishable item in isolation, and drop the subscript i from the previous notation. We now examine the "economics" of selecting an item, or, in the vocabulary of the food promotion example, the efficiency of promoting an item if one must pay for the promotion. We will identify circumstances in which it is not worth to replace the item.

In addition to the previous problem parameters, suppose that we have to pay a replacement cost $\nu > 0$ in all time epochs when the selecting action is applied (there is no additional cost for letting the item rested). We apply the standard restless bandit analysis, as described in Niño-Mora [5, 6].

We will prove that a perishable item defined as a restless bandit is *indexable*, that is, the optimal decisions are prescribed by an *index policy*, using *marginal productivity indices* (MPIs) assigned to controllable states. Niño-Mora [5] gave a sufficient condition for indexability, which holds in our case.

Proposition 2.2 *A perishable item is indexable, and the marginal productivity index for a controllable state $-t$ is*

$$\nu^*_{-t} = \frac{c\beta(q-p)(\beta p)^{t-1}}{1-\beta(q-p)\frac{1-(\beta p)^{t-1}}{1-\beta p}}. \tag{1}$$

A list of the most important properties of the MPI, which define priorities for selecting if various items are considered, follows.

Corollary 2.3 *An item with lower probability of disappearing when rested gets higher priority for being selected.*

Corollary 2.4 *The marginal productivity index given by (1) is proportional to cost c and positive.*

Corollary 2.5 *An item with closer deadline gets higher priority for being selected.*

To prove Proposition 2.2, a more detailed description of the restless bandit framework would be needed, so we omit it here. We only point out that the result follows from emulation of the adaptive-greedy algorithm [5].

3 Experimental Results

For fixed $2 \leq I \leq 8$ and $2 \leq T \leq 40$, we randomly generated 10'000 instances of the KPPI problems and analyzed the performance of the MPI heuristic with two naïve heuristics. Let RND be the following policy: select the items for the knapsack in a greedy manner following a random order. Let further EDF be the *Earlier-Deadline-First* strategy, where the items are selected greedily by their deadlines. Finally, let PAS denote the passive strategy: "select no items for the knapsack", giving the worst-case performance.

In our experiments, the average *relative suboptimality gap* rsg(MPI) is within 5% for all (I, T) pairs, while the average rsg(RND) ranges between 30% and 100%, and the average rsg(EDF) gives even worse performance (50% – 150%). Both RND and EDF perform on average increasingly worse as the time horizon increases, on the other hand, the MPI attains a peak around $T = 10$ and then performs increasingly better. The MPI almost never leads to a higher total expected cost than its alternatives. In addition, the MPI heuristic dramatically outperforms both RND and EDF in the worst-case analysis.

References

[1] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177.

[2] Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. J. Wiley & Sons, New York.

[3] Kleywegt, A. J. and Papastavrou, J. D. (1998). The dynamic and stochastic knapsack problem. *Operations Research*, 46(1):17–35.

[4] Manor, G. and Kress, M. (1997). Optimality of the greedy shooting strategy in the presence of incomplete damage information. *Naval Research Logistics*, 44:613–622.

[5] Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98.

[6] Niño-Mora, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Mathematical Programming, Series A*, 93(3):361–413.

[7] Niño-Mora, J. (2005). Marginal productivity index policies for the finite-horizon multiarmed bandit problem. In *Proceedings of the 44th IEEE CDC and ECC '05*, pages 1718–1722.

[8] Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queueing network. *Mathematics of Operations Research*, 24(2):293–305.

[9] Pisinger, D. (2005). Where are the hard knapsack problems? *Computers & Operations Research*, 32:2271–2284.

[10] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *A Celebration of Applied Probability, J. Gani (Ed.), Journal of Applied Probability*, 25A:287–298.