A PTAS for the chance-constrained knapsack problem with random item sizes

Vineet Goyal \(a,^*\), R. Ravi \(b\)

\(a\) Operations Research Center, Massachusetts Institute of Technology, United States
\(b\) Tepper School of Business, Carnegie Mellon University, United States

**ARTICLE INFO**

Article history:
Received 14 September 2009
Accepted 30 December 2009
Available online 18 January 2010

Keywords:
Stochastic knapsack
Chance constraint
Approximation schemes

**ABSTRACT**

We consider a stochastic knapsack problem where each item has a known profit but a random size that is normally distributed independent of other items. The goal is to select a profit maximizing set of items such that the probability of the total size exceeding the knapsack bound is at most a given threshold. We present a Polynomial Time Approximation Scheme (PTAS) for the problem via a parametric LP reformulation that efficiently computes a solution satisfying the chance constraint strictly and achieving near-optimal profit.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

We consider the following stochastic variant of the classical knapsack problem. We are given \(n\) items with profits \(p_1, p_2, \ldots, p_n\), a knapsack size \(B\) and a reliability level \(0.5 < \rho < 1\). Item \(i\) has a random size \(S_i\) distributed according to a known distribution and independent of the sizes of other items. The goal is to select a subset \(S\) of items such that,

\[
\Pr \left( \sum_{i \in S} S_i \leq B \right) \geq \rho, \tag{1.1}
\]

and the profit is maximized. We refer to (1.1) as the chance constraint and the problem as the chance-constrained knapsack problem. Our model finds applications in problems where there is only one stage of decision making under uncertainty. For instance, consider a typical capital investment problem where a central planner needs to select a set of projects to invest the available capital from a universe of projects where each project has an uncertain investment requirement and becomes known only after the project has been selected and started. The central planner would ideally like to invest in the set of projects that have a high profit or return while not exceeding the budget constraint. The solution might be highly conservative if we require that the budget is not exceeded in any possible realization of the investment requirements of selected projects. A chance-constrained model overcomes this drawback by allowing constraint violation for a small fraction of the unlikely realizations.

We present a PTAS for the chance-constrained knapsack problem when item sizes are normally distributed and independent of other items. It is known [1] that in the case of normally distributed item sizes the chance constraint can be formulated as a 0–1 conic program. However, we show that the integrality gap of the conic formulation is large. We reformulate the problem as a parametric LP and give a rounding algorithm that rounds any fractional solution to a \((1 + \epsilon)\)-approximate integral solution for any constant \(\epsilon > 0\) in time polynomial in the input size and \(\frac{1}{\epsilon}\).

Several stochastic variants of the classical knapsack problem have been studied in the literature. Henig [5] and Carraway et al. [2] consider the stochastic variant where item sizes are known but the profit of each item is distributed normally and independent of others and the goal is to maximize the probability that the profit is at least a given threshold. The authors present dynamic programming and branch and bound heuristics to solve this problem to optimality. Papastavrou et al. [9] and Kleywegt et al. [8] consider a variant called the stochastic and dynamic knapsack problem where items arrive online according to some stochastic process—the size and profit of each item is known only after the item arrives and you are required to decide whether to select the item or not when it arrives. Dean, Goemans and Vondrak [3] study the benefit of adaptivity in the online stochastic knapsack problem and give a polynomial time non-adaptive policy that is within a factor 4 of the optimal adaptive policy. Note that an adaptive policy depends on the remaining knapsack capacity while a non-adaptive policy does not.

Kleinberg et al. [7] and Goel and Indyk [4] consider a chance-constrained stochastic knapsack problem similar to the one considered in this paper. Kleinberg et al. [7] consider the case where item sizes have a Bernoulli-type distribution (with only two possible sizes for each item), and provide an \(O(\log \frac{1}{\epsilon})\)-approximation algorithm where \(\rho\) is the threshold probability.
Goel and Indyk \cite{4} provide a PTAS for the case when item sizes have Poisson or exponential distribution. However, the algorithm in \cite{4} violates the chance constraint by a factor of $(1+\epsilon)$. In contrast, we present a PTAS for the case when item sizes are normally distributed while satisfying the chance constraint strictly.

2. Conic integer formulation

We consider the case when each item $j$ has a normally distributed size with mean $\alpha_j$ and standard deviation $\sigma_j$ independent of the other items. Let $x_j$ denote whether item $j$ has been selected or not. Then the stochastic knapsack problem can be formulated as follows:

$$\max \left\{ \sum_{j=1}^{n} p_j x_j \mid \sum_{j} S_j x_j \leq B, \rho \right\}. \quad (2.1)$$

We simplify the probabilistic constraint as follows:

$$Pr \left( \sum_{j} S_j x_j \leq B \right) = Pr \left( \frac{\sum_{j} (S_j x_j - \alpha_j)}{\sqrt{\sum_j \sigma_j^2 x_j^2}} \leq \frac{B - \sum_j \alpha_j}{\sqrt{\sum_j \sigma_j^2 x_j^2}} \right) \leq \left( \phi^{-1}(\rho) \right)^2 \left( \sum_j \sigma_j^2 x_j^2 \right) \leq (B - \mu^*)^2 \quad (3.1)$$

where $Z = \frac{\sum_{j} (S_j x_j - \alpha_j)}{\sqrt{\sum_j \sigma_j^2 x_j^2}}$. Since the item sizes are normally distributed and independent of other items, $Z$ is a standard normal variable with mean 0 and standard deviation 1. Let $\phi$ denote the cumulative distribution function of the standard normal variate. Therefore, the probabilistic constraint can be rewritten as follows:

$$\Pr \left( Z \leq \frac{B - \sum_j \alpha_j}{\sqrt{\sum_j \sigma_j^2 x_j^2}} \right) \leq \rho \Rightarrow \frac{B - \sum_j \alpha_j}{\sqrt{\sum_j \sigma_j^2 x_j^2}} \geq \phi^{-1}(\rho).$$

Therefore, we can reformulate (2.1) as:

$$\max \left\{ \sum_{j=1}^{n} p_j x_j \mid \sum_{j} \phi^{-1}(\rho) \sqrt{\sum_j \sigma_j^2 x_j^2} \leq B, x \in \{0, 1\}^n \right\}. \quad (2.2)$$

If we relax the 0–1 constraints on $x_j$ to $0 \leq x_j \leq 1$ for all $j = 1, \ldots, n$, the formulation in (2.2) is a second order conic program and can be solved in polynomial time since $\phi^{-1}(\rho) \geq 0$ for $\rho \geq 0.5$. Note that we assume that $\rho > 0.5$ since for $\rho = 0.5$, $\phi^{-1}(\rho) = 0$ and the conic constraint in (2.2) reduces to a linear constraint which implies that the chance-constrained knapsack problem with normally distributed item sizes is equivalent to the deterministic knapsack problem where each item size is fixed at its expected size. Furthermore, we also require that $\rho$ is strictly less than 1 as $\phi^{-1}(1)$ is unbounded and the trivial solution $x = 0$ is the only feasible solution (assuming $\sigma_j > 0$ for all $j = 1, \ldots, n$). We show in the following example that the integrality gap of the conic relaxation is $\Omega(\sqrt{n})$.

Large integrality gap example. Consider the following instance: $p_j = \alpha_j = 1, \sigma_j = 1/\sqrt{n}$ for all $j = 1, \ldots, n$. Also, let the knapsack bound, $B = 3$, and the probability threshold, $\rho = 0.95$. Note that $\phi^{-1}(0.95) \approx 1.645$. It is easy to observe that any integral solution includes at most three items. Consider the conic constraint for any solution that contains four items.

$$4 - \frac{1}{\sqrt{n}} + \phi^{-1}(0.95) \cdot \sqrt{4} = 4 + 3.29 > 3,$$

which implies that any solution containing four items does not satisfy the conic constraint. Therefore, the integral optimal profit is at most 3. Now, consider the fractional solution $x_j = \frac{1}{\sqrt{n}}$. Then,

$$\sum_{j=1}^{n} a_j x_j + \phi^{-1}(\rho) \sqrt{\sum_{j=1}^{n} \sigma_j^2 x_j^2} = 1 + \phi^{-1}(\rho) < 3.$$

Therefore, the fractional solution is feasible and the optimal fractional profit is at least $\sqrt{n}$ which shows that the integrality gap of the conic formulation is $\Omega(\sqrt{n})$.

3. Parametric LP reformulation

We reformulate the second order conic program as a parametric LP and obtain a PTAS for the chance-constrained knapsack problem. Suppose we know that the sum of mean sizes of the items selected in an optimal solution is $\mu^*$. Then, the conic constraint in (2.2) can be expressed as,

$$\sum_j \alpha_j x_j \leq \mu^* \leq \phi^{-1}(\rho) \sqrt{\sum_j \sigma_j^2 x_j^2} \leq (B - \mu^*)^2 \quad (3.1)$$

Since $x_j^2 = x_j$ for $x_j \in \{0, 1\}$, we can simplify (3.1) to:

$$(\phi^{-1}(\rho))^2 \left( \sum_j \sigma_j^2 x_j^2 \right) \leq (B - \mu^*)^2.$$

Therefore, we can formulate the chance-constrained knapsack problem as a parametric two-dimensional knapsack problem where $\mu$ is the parameter corresponding to the total mean size of the selected items. We consider powers of $(1+\epsilon)$, i.e., $(1+\epsilon)^{j}, j = 0, \ldots, \log(1+\epsilon)B$ for some constant $\epsilon > 0$, as different choices of the parameter $\mu$. Therefore, the number of different choices of $\mu$ is $O(\log B)$ which is polynomial in the input size.

We also give the idea of optimal profit $OPT$ by considering powers of $(1+\epsilon)$. Let $P = \sum_{j=1}^{n} p_j$; we consider $O(\log B)$ different choices of $OPT$. At most $\frac{1}{\epsilon}$ items can have profit greater than $\epsilon$ $OPT$. Therefore, for each guess of $OPT = (1+\epsilon)$ we consider all subsets of cardinality at most $\frac{1}{\epsilon}$ of the items that have profit more than $\epsilon$ $OPT$ to include in the solution. For each guess $OPT$ of $OPT$ and each choice of subset of items of individual profit more than $\epsilon$ $OPT$, we solve a subproblem $\Pi(S_1, S_2, \emptyset)$ where $S_2$ is the set of items whose profit is more than $\epsilon$ and are included in the final solution and we are required to choose a subset of items from $S_1 \subset \{n\}$ that maximizes the total profit. Let $\Pi(S_1, S_2, \emptyset, \mu)$ denote the problem where the total mean size of all items selected from $S_1$ is at most $\mu$. Therefore, we can formulate $\Pi(S_1, S_2, \emptyset, \mu)$ as the following two-dimensional knapsack problem:

$$\max \sum_{j \in S_1} p_j x_j + \sum_{j \in S_2} \phi^{-1}(\rho) \sqrt{\sum_{j \in S_1} \sigma_j^2 x_j^2} \leq B - \mu \quad (3.2)$$

$$\sum_{j \in S_2} a_j x_j \leq \mu$$

$$\leq \left( B - \mu - \sum_{j \in S_2} a_j \right) - \phi^{-1}(\rho) \sqrt{\sum_{j \in S_1} \sigma_j^2} \quad (3.3)$$

$x_j \in \{0, 1\}$.
Algorithm $A$ for Chance-constrained Knapsack Problem.

**Input:** Given $n$ items where item $j$ has profit $p_j$ and a normally distributed size with mean $a_j$ and standard deviation $\sigma_j$, knapsack size $B$, reliability level $0 < \rho < 1$ and a constant $\epsilon > 0$. Let $p_m = \min_{j \in [n]} p_j$, $P = \sum_{j \in [n]} p_j$.

Initialize $N_1 = [\log_{1+\epsilon} p_m], N_2 = [\log_{1+\epsilon} P], x_A = 0, P_A \leftarrow 0$.

1. For $t = N_1, \ldots, N_2$,
   (a) Let $O = (1 + \epsilon)^t$ and let $S_t = \{ j \in [n] | p_j \geq \epsilon \cdot O \}$.
   (b) For each set $S \subset S_t$ such that $|S| < \frac{1}{2}$,
      i. Solve $\Pi([n] \setminus S_t, S, O)$ and let $x_S$ denote the integral solution returned by $A(\Pi)$.
      ii. If $P_A < p^T x_S$, then $x_A \leftarrow x_S$ and $P_A \leftarrow p^T x_S$.

2. Return the solution $x_A$.

**Fig. 1.** Algorithm $A$ for chance-constrained Knapsack problem.

Algorithm $A(\Pi)$ for $\Pi(S_1, S_2, O)$.

Let $\mu_{\min} = \min_{j \in [n]} a_j$. Initialize $N_1 = [\log_{1+\epsilon} \mu_{\min}], N_2 = [\log_{1+\epsilon} B], x_s = 0, P_s \leftarrow 0$.

1. For $t = N_1, \ldots, N_2$,
   (a) Let $\mu = (1 + \epsilon)^t$ and let $x(\mu)$ be a basic optimal solution for $\Pi(S_1, S_2, O, \mu)$.
   (b) Using Lemma 3.1 find an integral solution $\tilde{x}(\mu)$ such that
      \[
      \sum_{j=1}^{n} p_j \cdot \tilde{x}(\mu)_{j} \geq \sum_{j=1}^{n} p_j \cdot x(\mu)_{j} - 2\epsilon \cdot O
      \]
   (c) If $P_s < \sum_{j \in S_1} p_j \tilde{x}(\mu)_{j} + \sum_{j \in S_2} p_j$, then
      \[
      x_s \leftarrow \tilde{x}(\mu),
      P_s \leftarrow \sum_{j \in S_1} p_j \tilde{x}(\mu)_{j} + \sum_{j \in S_2} p_j
      \]

2. Return the solution $x_s$.

**Fig. 2.** Algorithm $A(\Pi)$ for $(S_1, S_2, O)$.

Algorithm $A$ for the Chance-constrained Knapsack Problem is described in Fig. 1 and Algorithm $A(\Pi)$ for $\Pi(S_1, S_2, O, \mu)$ is described in Fig. 2.

In the following lemma, we show that we can find a good integral solution to the problem $\Pi(S_1, S_2, O, \mu)$. 

**Lemma 3.1.** Consider the problem $\Pi(S_1, S_2, O, \mu)$ such that $p_j \leq \epsilon \cdot O$ for all $j \in S_1$. If $P^*$ is the optimal profit for $\Pi(S_1, S_2, O, \mu)$, then there is a polynomial time algorithm to find a feasible set of items whose profit is at least $(P^* - 2\epsilon \cdot O)$.

**Proof.** Consider the two-dimensional knapsack formulation of $\Pi(S_1, S_2, O, \mu)$ and consider a basic optimal solution $\tilde{x}$ of the LP relaxation. Since there are only two constraints other than the bound constraints, at least $|S_1| - 2$ bound constraints must be tight for $\tilde{x}$. Therefore, at least $|S_1| - 2$ variables out of $|S_1|$ variables are integral in any basic optimal solution. Let $j_1, j_2 \in S$ such that $\tilde{x}_{j_1}, \tilde{x}_{j_2}$ are fractional. We know that $p_j \leq \epsilon \cdot O$ for all $j \in S_1$. Consider the following solution, $\hat{x}$ where for any $j \in S, \hat{x}_j = 0$ if $j = j_1, j_2$ and $\hat{x}_j = \tilde{x}_j$ otherwise. Clearly, $\hat{x}$ is an integral solution and $\sum_{j \in S_1} p_j \hat{x}_j \geq \sum_{j \in S_1} p_j \tilde{x}_j - 2\epsilon \cdot O$. Therefore, we obtain an integral solution such that $\sum_{j \in S_1} p_j \hat{x}_j + \sum_{j \in S_2} p_j \geq P^* - 2\epsilon \cdot O$.

In the following lemma we show that for an appropriately chosen value of $O$ and $\mu$ and subsets $S_1, S_2 \subset [n]$, the problem $\Pi(S_1, S_2, O, \mu)$ has optimal profit at least $OPT/(1+\epsilon)$.

**Lemma 3.2.** Let $S^*$ be the set of items selected by an optimal solution and let $OPT = \sum_{j \in S^*} p_j$. Consider $l$, such that $(1 + \epsilon)^{-1} \leq OPT < (1 + \epsilon)^l$. Let $O = (1 + \epsilon)^l$ and let $S_\epsilon = \{ i \in [n] | p_i \geq \epsilon \cdot O \}$, $S_1 = [n] \setminus S_\epsilon$, and $S_2 = S_\epsilon \cap S^*$. Then the optimal profit for the problem $\Pi(S_1, S_2, O)$ is at least $OPT/(1+\epsilon)$.

**Proof.** Let $\mu^* = \sum_{j \in S_1} a_j, \mu_1 = \sum_{j \in S_1 \setminus S^*} a_j, \mu_2 = \sum_{j \in S_2} a_j$ and let $k$ be such that $(1+\epsilon)^{k-1} \leq \mu_1 < (1+\epsilon)^k$. Let $\beta = (1+\epsilon)^{k-1}$ and we consider the problem $\Pi(S_1, S_2, O, \beta)$. Consider the following fractional solution, $\tilde{x}$ for $\Pi(S_1, S_2, O, \beta)$ where $\tilde{x}_j = \frac{1}{1+\epsilon}$ for all $j \in S_1 \cap S^*$ and 0 otherwise. We first show that $\tilde{x}$ is a feasible fractional solution for $\Pi(S_1, S_2, O, (1+\epsilon)^{k-1})$. It is easy to observe that $\tilde{x}$ satisfies (3.2) as $\sum_{j \in S} p_j \tilde{x}_j = \sum_{j \in S \cap S^*} a_j \cdot \frac{1}{1+\epsilon} \leq \beta$. Also, 

\[
(\phi^{-1}(\rho))^2 \cdot \left( \sum_{j \in S_1} \sigma_j^2 \tilde{x}_j \right)^2 \cdot \left( \sum_{j \notin S_1 \cap S^*} \sigma_j^2 \cdot \frac{1}{1+\epsilon} \right)^2 
\]

\[
(1+\epsilon)^{2k-2} \leq (\phi^{-1}(\rho))^2 \cdot \left( \sum_{j \in S_1} \sigma_j^2 \right)^2 \cdot \left( \sum_{j \notin S_1 \cap S^*} \sigma_j^2 \cdot \frac{1}{1+\epsilon} \right)^2 
\]

\[
(B - \mu_1)^2 - (\phi^{-1}(\rho))^2 \cdot \left( \sum_{j \in S_2} \sigma_j^2 \right)^2 
\]

\[
\leq \frac{1+\epsilon}{1+\epsilon} 
\]

\[
(B - \mu_1)^2 - (\phi^{-1}(\rho))^2 \cdot \left( \sum_{j \in S_2} \sigma_j^2 \right)^2 
\]

\[
\leq \frac{1+\epsilon}{1+\epsilon} 
\]

(3.4)
where the last equality follows because \( S^* = (S_1 \cap S^*)_U S_2 \).

**Theorem 3.3.** Given \( \varepsilon > 0 \), there is a polynomial time algorithm that gives a \((1 - 3\varepsilon)\)-approximation for the chance-constrained knapsack problem with reliability, \((0.5 + \delta) \leq \rho \leq (1 - \delta)\) for some fixed \( \delta > 0 \). Furthermore, the running time of \( \mathcal{A} \) is

\[
O \left( \frac{\log (B/\mu_m) \cdot \log (P/p_m) \cdot n^2}{\varepsilon^2} \cdot n^{1.5} \cdot L \right),
\]

where \( P = \sum_{i=1}^m p_i \), \( p_m = \min\{p_i \mid j = 1, \ldots, n\} \), \( \mu_m = \min\{a_i \mid j = 1, \ldots, n\} \) and \( L \) is the input size of the problem.

**Proof.** Let \( OPT \) denote the profit value of an optimal solution and let \( S^* \) be the set of items selected in \( OPT \). Consider \( I \), such that, \((1 + \varepsilon)^{k-1} \leq OPT \leq (1 + \varepsilon)^{k} \) and let \( O = (1 + \varepsilon)^{k} \). Let \( S_1 = \{ i \in [n] \mid p_i \geq O \} \), \( S_1 = [n] \setminus S_2 \) and \( S_2 = S \cap S^* \). Note that Algorithm \( \mathcal{A} \) considers the guess \( O \) for the optimal value. Also, since \( |S_2| < \frac{1}{2} \) the subproblem \( \Pi(S_1, S_2, O) \) is considered as one of the subproblems in the algorithm \( \mathcal{A} \). Let \( \mu(1) = \sum_{j \in S_1 \cap S^*} a_j \) and \( k \) such that \((1 + \varepsilon)^{k-1} \leq \mu(1) \leq (1 + \varepsilon)^{k} \) and let \( \beta = (1 + \varepsilon)^{k-1} \). Clearly, the subproblem \( \Pi(S_1, S_2, O, \beta) \) is considered in the algorithm \( \mathcal{A}(\Pi) \) while solving \( \Pi(S_1, S_2, O) \). From Lemma 3.2, we know that the optimal profit for the subproblem \( \Pi(S_1, S_2, O, \beta) \) is at least \( OPT \). Furthermore, using Lemma 3.1 we can find a set of items \( S \) for the problem \( \Pi(S_1, S_2, O, \beta) \) such that \( \sum_{j \in S} p_j \geq \frac{OPT}{1 + \varepsilon} - 2\varepsilon \cdot O \geq (1 - 3\varepsilon) \cdot OPT \). Therefore, Algorithm \( \mathcal{A} \) finds an integral solution that has profit at least \((1 - 3\varepsilon) \cdot OPT \). To bound the running time of \( \mathcal{A} \), note that we consider \( O \left( \log (P/p_m) / \varepsilon \right) \) different choices of the optimal profit value \( O \). Also, we consider \( O \left( n^{1.5} \right) \) choices of the set of items \( S \) for the subproblem \( \Pi \) for each choice of \( O \). Furthermore, in the subroutine \( \mathcal{A}(\Pi) \), we solve \( O \left( \log (B/\mu_m) / \varepsilon \right) \) different subproblems for solving \( \Pi(S_1, S_2, O) \) for given subsets \( S_1, S_2 \subseteq [n] \) and a choice for optimal profit \( O \). Therefore, Algorithm \( \mathcal{A} \) solves and rounds

\[
O \left( \frac{\log (B/\mu_m) \cdot \log (P/p_m) \cdot n^2}{\varepsilon^2} \cdot n^{1.5} \cdot L \right),
\]

linear programs. Each LP has at most \( n \) variables and it is easy to observe that the input size of each LP is \( O(L) \) since \( \delta \leq \phi^{-1}(\rho) \leq \phi^{-1}(1-\delta) \) which implies that \( \phi^{-1}(\rho) \) is bounded between constants for a fixed \( \delta > 0 \). Therefore, each LP can be solved in \( O(n^{1.5}L) \) time \[6\] and can be rounded in \( O(n) \) time by scanning each variable and setting the fractional ones to zero (following the proof of Lemma 3.2), which completes the proof of the running time of the algorithm. \( \square \)

**Acknowledgements**

The first author’s research is partially supported by NSF Grant CCF-0430751. The second author’s research is partially supported by NSF grants CCF-0430751 and CCF-0728841.

**References**


