# Robust recoverable and two-stage selection problems

Adam Kasperski*

*Department of Operations Research, Wrocław University of Technology,*
*Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, adam.kasperski@pwr.edu.pl*

Paweł Zieliński

*Department of Computer Science (W11/K2), Wrocław University of Technology,*
*Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, pawel.zielinski@pwr.edu.pl*

## Abstract

In this paper the following selection problem is discussed. A set of $n$ items is given and we wish to choose a subset of exactly $p$ items of the minimum total cost. This problem is a special case of 0-1 knapsack in which all the item weights are equal to 1. Its deterministic version has a trivial $O(n)$-time algorithm, which consists in choosing $p$ items of the smallest costs. In this paper it is assumed that the item costs are uncertain. Two robust models, namely two-stage and recoverable ones, under discrete and interval uncertainty representations, are discussed. Several positive and negative complexity results for both of them are provided.

**Keywords:** combinatorial optimization; robust optimization; computational complexity; approximation algorithms; selection problem

## 1 Introduction

In this paper we wish to investigate the following SELECTION problem. Let $E = \{e_1, \ldots, e_n\}$ be a set of items. Each item $e \in E$ has a nonnegative cost $c_e$ and we wish to choose a subset of exactly $p$ items, $p \in \{1, \ldots, n-1\}$, of the minimum total cost, $f(X) = \sum_{e \in X} c_e$. This problem has a trivial $O(n)$-time algorithm which works as follows. We first determine in $O(n)$-time the $p$th largest item cost, say $c_{[p]}$, and then choose $p$ items from $E$ whose costs are not greater than $c_{[p]}$. The SELECTION problem is a special, polynomially solvable version of the 0-1 knapsack problem, in which all the items have unit weights. It is also a special case of some other discrete optimization problems such as minimum assignment or a single machine scheduling problem with weighted number of late jobs criterion (see [14] for more details). It is also a matroidal problem, as the set of feasible solutions is composed of all bases of an uniform matroid [22] and it can be seen as a basic resource allocation problem [13].

Suppose that the item costs are uncertain and suppose that we are given a *scenario set* $\Gamma$ which contains all possible vectors of the item costs, called *scenarios*. We thus only know that one cost scenario $S \in \Gamma$ will occur but we do not know which one before a solution is computed. The cost of item $e \in E$ under scenario $S \in \Gamma$ is denoted as $c_e^S \geq 0$. No additional information for scenario set $\Gamma$, such as a probability distribution, is provided. Two methods

---

*Corresponding author

of defining scenario sets are popular in the existing literature (see, e.g, [18]). In the *discrete uncertainty representation*, $\Gamma^D = \{S_1, \ldots, S_K\}$ contains $K > 1$ explicitly listed scenarios. In the *interval uncertainty representation*, for each item $e \in E$ an interval $[\underline{c}_e, \overline{c}_e]$ of its possible costs is specified and $\Gamma^I = \Pi_{e \in E}[\underline{c}_e, \overline{c}_e]$ is the Cartesian product of these intervals. The cost of solution $X$ depends now on scenario $S \in \Gamma$, $\Gamma \in \{\Gamma^D, \Gamma^I\}$, and will be denoted as $f(X, S) = \sum_{e \in X} c_e^S$. In order to choose a solution two robust criteria, namely the *min-max* and *min-max regret* can be applied, which lead to the following two optimization problems:

$$\text{Min-Max Selection}: \min_{X \in \Phi} \max_{S \in \Gamma} f(X, S),$$

$$\text{Min-Max Regret Selection}: \min_{X \in \Phi} \max_{S \in \Gamma} f(X, S) - f^*(S),$$

where $\Phi = \{X \subseteq E : |X| = p\}$ is the set of all feasible solutions and $f^*(S)$ is the cost of an optimal solution under scenario $S$. The quantity $f(X, S) - f^*(S)$ is called a *regret* of solution $X$ under scenario $S$. Both robust versions of the Selection problem has attracted a considerable attention in the recent literature. It turns out that their complexity depends on the way in which scenario set $\Gamma$ is defined. It has been shown in [2] that under the discrete uncertainty representation Min-Max Regret Selection is NP-hard even for two scenarios. Min-Max Selection remains NP-hard even for two scenarios. A proof is similar to that for the min-max regret given in [2]. Both problems become strongly NP-hard when the number of scenarios is a part of input [15]. Furthermore, in this case Min-Max Selection is also hard to approximate within any constant factor [14]. Several approximation algorithms for Min-Max Selection has been recently proposed in [15, 14, 10]. The best known, designed in [10], has an approximation ratio of $O(\log K / \log \log K)$. For the interval uncertainty representation both robust problems are polynomially solvable. The min-max version is trivially reduced to a deterministic counterpart as it is enough to solve the deterministic problem for scenario $(\overline{c}_e)_{e \in E}$. On the other hand, Min-Max Regret Selection is more involving and some polynomial time algorithms for this problem have been constructed in [2, 7]. The best known algorithm with running time $O(n \cdot \min\{p, n - p\})$ has been shown in [7].

Many real world problems arising in operations research and optimization have a two-stage nature. Namely, a complete or partial solution is determined in the first stage and can be then modified or completed in the second stage. Typically, the costs in the first stage are known while the costs in the second stage are uncertain. This uncertainty is also modeled by providing a scenario set $\Gamma \in \{\Gamma^D, \Gamma^I\}$, which contains all possible vectors of the second stage costs. If no additional information with $\Gamma$ is provided, then the robust criteria can be applied to choose a solution. In this paper we investigate two well known concepts, namely *robust two-stage* and *robust recoverable* ones and apply them to the Selection problem. In the robust two-stage model, a partial solution is formed in the first stage and completed optimally when a true scenario reveals. In the robust recoverable model a complete solution must be formed in the first stage, but it can be modified to some extent after a true scenario occurs. Both models have been discussed in the existing literature for various problems. In particular, the robust two-stage versions of covering problems, the minimum assignment and the minimum spanning tree have been investigated in [8, 17, 16]. The robust recoverable approach to the linear programming, the weighted disjoint hitting set, the minimum perfect matching, the shortest path, the minimum spanning tree and the $0 - 1$ knapsack have been discussed in [4, 5, 6, 19, 21].

**Our results** In Section 3 we will investigate the robust recoverable model. We will show that it is NP-hard for two scenarios and becomes strongly NP-hard and not at all approximable when the number of scenarios is a part of input. A major part of Section 3 is devoted to constructing a polynomial $O((p-k+1)n^2)$ algorithm for the interval uncertainty representation, where $k$ is the recovery parameter. In Section 4 we will study the robust two-stage model. We will prove that it is NP-hard for three second-stage cost scenarios. Furthermore, when the number of scenarios is a part of input, then the problem becomes strongly NP-hard and hard to approximate within $(1-\epsilon)\log n$ for any $\epsilon > 0$, unless P=NP. For scenario set $\Gamma^D$, we will construct a randomized algorithm which returns an $O(\log K + \log n)$-approximate solution with high probability. If $K = poly(n)$, then the randomized algorithm gives the best approximation up to a constant factor. We will also show that for the interval uncertainty representation the robust two-stage model is solvable in $O(n)$ time.

## 2 Problems formulation

Before we show the formulations of the problems we recall some notations and introduce new ones. Let us fix $p \in \{1, \ldots, n-1\}$, and define

- $\Phi = \{X \subseteq E : |X| = p\}$,

- $\Phi_1 = \{X \subseteq E : |X| \le p\}$,

- $\Phi_X = \{Y \subseteq E \setminus X : |Y| = p - |X|\}$,

- $\Phi_X^k = \{Y \subseteq E : |Y| = p, |Y \setminus X| \le k\}$, $k \in \{0, \ldots, p\}$,

- $C_e$ is the deterministic, first-stage cost of item $e \in E$,

- $c_e^S$ is the second-state cost of item $e \in E$ under scenario $S \in \Gamma$, where $\Gamma \in \{\Gamma^D, \Gamma^I\}$.

- $f(X, S) = \sum_{e \in X} c_e^S$, for any subset $X \subseteq E$.

We now define the two-stage model as follows. In the first stage we choose a subset $X \in \Phi_1$ of the items, i.e. such that $|X| \le p$, and we add additional $p - |X|$ items to $X$ after observing which scenario in $\Gamma$ has occurred. The cost of $X$ under scenario $S$ is defined as

$$f_1(X, S) = \sum_{e \in X} C_e + \min_{Y \in \Phi_X} f(Y, S).$$

In the *robust two-stage selection* problem we seek $X \in \Phi_1$, which minimizes the maximum cost over all scenarios, i.e.

$$\text{Two-Stage Selection}: opt_1 = \min_{X \in \Phi_1} \max_{S \in \Gamma} f_1(X, S).$$

We now define the robust recoverable model. In the first stage we must choose a complete solution $X \in \Phi$, i.e. such that $|X| = p$. In the second stage additional costs occur for the selected items. However, a limited recovery action is allowed, which consists in replacing at most $k$ items in $X$ with some other items from $E \setminus X$, where $k \in \{0, \ldots, p\}$ is a given *recovery parameter*. The cost of $X$ under scenario $S$ is defined as follows:

$$f_2(X, S) = \sum_{e \in X} C_e + \min_{Y \in \Phi_X^k} f(Y, S).$$

In the *robust recoverable selection* problem we wish to find a solution $X \in \Phi$, which minimizes the maximum cost over all scenarios, i.e.

$$\text{RECOVERABLE SELECTION}: \ opt_2 = \min_{X \in \Phi} \max_{S \in \Gamma} f_2(X, S).$$

Observe that for the interval uncertainty representation, in both problems we can replace scenario set $\Gamma^I$ with one second-stage scenario $\overline{S} = (\overline{c}_e)_{e \in E}$, which follows from the fact that $\max_{S \in \Gamma^I} f_1(X, S) = f_1(X, \overline{S})$ and $\max_{S \in \Gamma^I} f_2(X, S) = f_2(X, \overline{S})$. For TWO-STAGE SELECTION, this observation leads to an easy polynomial time algorithm, discussed in Section 4. On the other hand, solving RECOVERABLE SELECTION for one second-stage scenario is a nontrivial problem and we will investigated it in Section 3.

# 3 Robust recoverable selection

In this section we deal with the RECOVERABLE SELECTION problem. Consider first the discrete uncertainty representation, i.e. the problem with scenario set $\Gamma^D$. It is easy to observe that when all the first stage costs are equal to zero and recovery parameter $k = 0$, then RECOVERABLE SELECTION is equivalent to MIN-MAX SELECTION with scenario set $\Gamma^D$. It follows from the fact that the solution formed in the first stage cannot be changed in the second stage and thus $f_2(X, S) = f(X, S)$. Hence, we have an immediate consequence of the results obtained in [2, 14], namely under scenario set $\Gamma^D$, the RECOVERABLE SELECTION problem is weakly NP-hard when $K = 2$. Furthermore, it becomes strongly NP-hard and hard to approximate within any constant factor when $K$ is a part of input. The following two theorems extend these results to case when recovery parameter $k$ is positive.

**Theorem 1.** *The* RECOVERABLE SELECTION *problem with scenario set $\Gamma^D$ is weakly NP-hard, even for two scenarios and any constant $k > 0$.*

*Proof.* See Appendix A. □

**Theorem 2.** *If $K$ is a part of input and $k > 0$, then* RECOVERABLE SELECTION *with scenario set $\Gamma^D$ is strongly NP-hard and not at all approximable unless P=NP.*

*Proof.* See Appendix A. □

In the remaining part of this section we will provide a polynomial algorithm for the interval uncertainty representation. The RECOVERABLE SELECTION problem with scenario set $\Gamma^I$ can be rewritten as follows:

$$\min_{X \in \Phi} \left\{ \sum_{e \in X} C_e + \min_{Y \in \Phi_X^k} \max_{S \in \Gamma^I} \sum_{e \in Y} c_e^S \right\} = \min_{X \in \Phi} \left\{ \sum_{e \in X} C_e + \min_{Y \in \Phi_X^k} \sum_{e \in Y} \overline{c}_e \right\}. \tag{1}$$

Consequently, in order to compute a solution under the interval uncertainty representation, it suffices to provide an algorithm for RECOVERABLE SELECTION with only one cost scenario in the second stage. We now construct such a polynomial algorithm for this problem. For notation purposes we will use $S = (\overline{c}_e)_{e \in E}$ to denote the second-stage cost scenario. We will

also use $C_i$ to denote $C_{e_i}$ and $\bar{c}_i$ to denote $\bar{c}_{e_i}$ for $i \in [n]$. In problem (1) we need to find a pair of solutions $X \in \Phi$ and $Y \in \Phi_X^k$. Since $|X| = |Y| = p$, the problem (1) is equivalent to:

$$
\begin{aligned}
\min \quad & \textstyle\sum_{e \in X} C_e + \sum_{e \in Y} \bar{c}_e \\
\text{s.t.} \quad & |X \cap Y| \geq p - k \\
& X, Y \in \Phi
\end{aligned}
\tag{2}
$$

Let us introduce 0-1 variables $x_i$, $y_i$ and $z_i$, $i \in [n]$, that indicate the chosen parts of $X$, $Y$ and $X \cap Y$, respectively. Problem (2) can be then represented as the following MIP model:

$$
\begin{aligned}
opt_2 = \min \quad & \textstyle\sum_{i=1}^{n} C_i x_i + \sum_{i=1}^{n} \bar{c}_i y_i + \sum_{i=1}^{n}(C_i + \bar{c}_i) z_i \\
\text{s.t.} \quad & x_1 + \cdots + x_n + z_1 + \cdots + z_n = p \\
& y_1 + \cdots + y_n + z_1 + \cdots + z_n = p \\
& z_1 + \cdots + z_n \geq p - k \\
& x_i + z_i \leq 1 & i \in [n] \\
& y_i + z_i \leq 1 & i \in [n] \\
& x_i, y_i, z_i \in \{0, 1\} & i \in [n]
\end{aligned}
\tag{3}
$$

Let $x_i, y_i, z_i$, $i \in [n]$ be an optimal solution to (3). Then $e_i \in X$ if $x_i = 1$ or $z_i = 1$, and $e_i \in Y$ if $y_i = 1$ or $z_i = 1$. The following theorem shows the unimodularity property of the constraint matrix of (3).

**Theorem 3.** *The constraint matrix of (3) is totally unimodular.*

*Proof.* See Appendix A. $\qquad\square$

From Theorem 3 we immediately get that every extreme solution of (3), after removing the integrality constraints, is integral and in consequence RECOVERABLE SELECTION for one second-stage scenario is polynomially solvable. Our goal is now to construct an efficient combinatorial algorithm for this problem. In order to do this, we first apply Lagrangian relaxation (see, e.g., [1]) to (3). Relaxing the cardinality constraint $\sum_{i \in [n]} z_i \geq p - k$ with a nonnegative multiplier $\theta$, we obtain the following linear programming problem:

$$
\begin{aligned}
\phi(\theta) = \min \quad & \textstyle\sum_{i=1}^{n} C_i x_i + \sum_{i=1}^{n} \bar{c}_i y_i + \sum_{i=1}^{n}(C_i + \bar{c}_i - \theta) z_i + (p - k)\theta \\
\text{s.t.} \quad & x_1 + \cdots + x_n + z_1 + \cdots + z_n = p \\
& y_1 + \cdots + y_n + z_1 + \cdots + z_n = p \\
& x_i + z_i \leq 1 & i \in [n] \\
& y_i + z_i \leq 1 & i \in [n] \\
& x_i, y_i, z_i \geq 0 & i \in [n]
\end{aligned}
\tag{4}
$$

The Lagrangian function $\phi(\theta)$ for any $\theta \geq 0$ is a lower bound on the optimum value $opt_2$. It is well-known that $\phi(\theta)$ is concave and piecewise linear function. We now find a nonnegative multiplier $\theta$ together with an optimal solution $x_i^*, y_i^*, z_i^*$, $i \in [n]$, to (4) which is also feasible in (3) and satisfies the complementary slackness condition, i.e. $\theta((p - k) - \sum_{i \in [n]} z_i) = 0$. By the optimality test, such a solution is optimal to the original problem (3). We will do this by iteratively increasing the value of $\theta$, starting with $\theta = 0$. For $\theta = 0$ the following lemma holds:

**Lemma 1.** *The value of $\phi(0)$ can be computed in $O(n)$ time.*

*Proof.* Let $X$ be a set of $p$ items of the smallest values of $C_i$ and let $Y$ be the set of $p$ items of the smallest $\overline{c}_i$, $e_i \in E$. Clearly, $\hat{c} = \sum_{e_i \in X} C_i + \sum_{e_i \in Y} \overline{c}_i$ is a lower bound on $\phi(0)$. A feasible solution of the cost $\hat{c}$ can be obtained by setting $z_i = 1$ for $e_i \in X \cap Y$, $x_i = 1$ for $e_i \in X \setminus Y$ and $y_i = 1$ for $e_i \in Y \setminus X$. The sets $X$ and $Y$ can be found in $O(n)$ time and the lemma follows. $\square$

Given a $0-1$ optimal solution to (4) for a fixed $\theta \geq 0$, let $E_X = \{e_i \in E : x_i = 1\}$, $E_Y = \{e_i \in E : y_i = 1\}$, $E_Z = \{e_i \in E : z_i = 1\}$. For $\theta > 0$ the sets $E_X$, $E_Y$ and $E_Z$ are pairwise disjoint and thus form a partition of the set $X \cup Y$ into $X \setminus Y$, $Y \setminus X$ and $X \cap Y$, respectively, $X, Y \in \Phi$. Indeed, $E_X \cap E_Z = \emptyset$ and $E_Y \cap E_Z = \emptyset$ by the constraints of (4). If $e_i \in E_X \cap E_Y$, then we can find a better solution by setting $z_i = 1$, $x_i = 0$ and $y_i = 0$. The same property holds for the optimal solution when $\theta = 0$ (see the construction in the proof of Lemma 1). Let us state the above reasoning as the following property:

**Property 1.** *For each $\theta \geq 0$ there is an optimal solution $x_i, y_i, z_i$, $i \in [n]$, to (4) such that $E_X$, $E_Y$ and $E_Z$ form a partition of the set $X \cup Y$ into $X \setminus Y$, $Y \setminus X$ and $X \cap Y$, respectively.*

From now on, we will represent an optimal solution to (4) for any $\theta \geq 0$ as a triple $(E_X, E_Y, E_Z)$ that has Property 1. It is easily seen that (4) is equivalent to the minimum cost flow problem in the network $G_\theta$ shown in Figure 1a. All the arcs in $G_\theta$ have capacities equal to 1. The costs of arcs $(s, x_i)$ and $(y_i, t)$ for $i \in [n]$ are equal to 0. The costs of the arcs $(x_i, y_i)$ for $i \in [n]$ are $C_i + \overline{c}_i - \theta$, and the costs of the arcs $(x_i, y_j)$, $i \neq j$, are $C_i + \overline{c}_j$. The supply at the source $t$ is of $p$ and the demand at the sink $t$ is of $-p$.
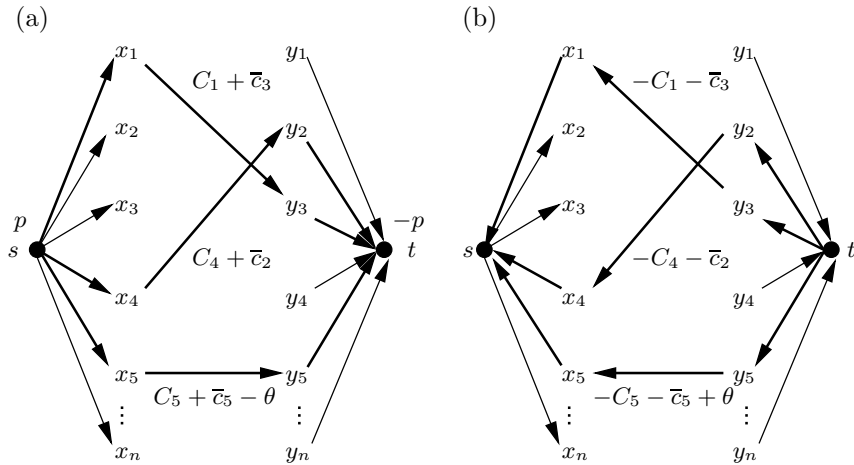


Figure 1: (a) Network $G_\theta$ (all the arc capacities are equal to 1, not all arcs between $x_i$ and $y_j$ are shown). (b) Residual network $G_\theta^r$ for $E_X = \{e_1, e_4\}, E_Y = \{e_2, e_3\}, E_Z = \{e_5\}$ (the arcs leading from $x_i$ to $y_j$ are not shown).

Let $(E_X, E_Y, E_Z)$ be an optimal solution to (4). The corresponding integer optimal flow $f_\theta$ in $G_\theta$ is constructed as follows. We send 1 unit of flow along arcs $(s, x_i)$, $(x_i, y_i)$ and $(y_i, t)$ if $e_i \in E_Z$; we then pair the items from $e_i \in E_X$ and $e_j \in E_Y$ in any fashion and send 1 unit of flow along the arcs $(s, x_i)$, $(x_i, y_j)$ and $(y_j, t)$ for each such a pair $(e_i, e_j)$. An example for $E_Z = \{e_5\}$, $E_X = \{e_1, e_4\}$ and $E_Y = \{e_2, e_3\}$ is shown in Figure 1a, where $X = \{e_1, e_4, e_5\}$, $Y = \{e_2, e_3, e_5\}$, $p = 3$. Assume that $f_\theta$ is an optimal flow in $G_\theta$. We can assume that

this flow is integer by the integrality property of optimal solutions to the minimum cost flow problem (see, e.g., [1]). Let $A \subseteq [n]$ be the set of indices of all nodes $x_i$ which receive 1 unit of flow from $s$ and let $B \subseteq [n]$ be the set of indices of nodes $y_j$ which send 1 unit of flow to $t$ in $f_\theta$. Clearly $|A| = |B| = p$. Let $E_Z = \{e_i \in E : i \in A \cap B\}$, $E_X = \{e_i \in E : i \in A \setminus B\}$, $E_Y = \{e_i \in E : i \in B \setminus A\}$. It is easy to see that the cost of the resulting feasible solution $(E_X, E_Y, E_Z)$ to (4) is the same as the cost of $f_\theta$.

Consider now an optimal solution $(E_X, E_Y, E_Z)$ to (4) for some fixed $\theta \geq 0$. Without loss of generality we can assume that this optimal solution has Property 1. Let $f_\theta$ be the corresponding optimal flow in $G_\theta$. The residual network $G_\theta^r$ with respect to $f_\theta$ is depicted in Figure 1b. By the negative cycle optimality condition (see, e.g., [1]), $G_\theta^r$ does not contain any negative cost directed cycle. Suppose we increase $\theta$ in $G_\theta^r$. Then, a negative cost directed cycle may appear in $G_\theta^r$, which means that the flow $f_\theta$ becomes not optimal in $G_\theta$. We now investigate the structure of such negative cycles. This will allow us to find the largest value of $\theta$ for which the flow $f_\theta$ (and thus the corresponding solution to (4)) remains optimal.

Denote by $\mathcal{F}$ the set of arcs of the form $(x_i, y_i)$ in $G_\theta^r$ (the *forward* arcs) and by $\mathcal{B}$ the set of arcs of the form $(y_i, x_i)$ in $G_\theta^r$ (the *backward* arcs). These arcs will play a crucial role as only their costs depend on $\theta$ in $G_\theta^r$. Clearly, the costs of the arcs in $\mathcal{F}$ decrease and the costs of the arcs in $\mathcal{B}$ increase when the value of $\theta$ increases. In the example shown in Figure 1b the arc $(y_5, x_5)$ is a backward arc and all the remaining arcs $(x_i, y_i)$ are forward arcs.
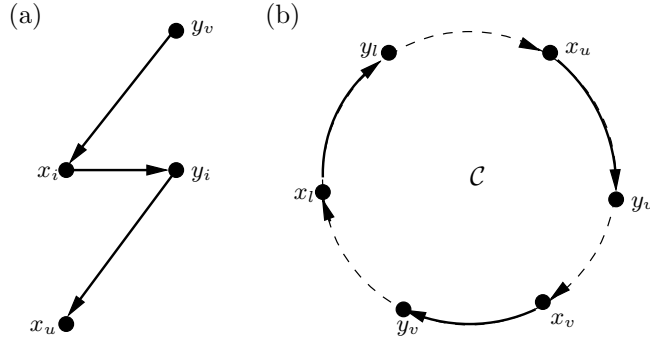


Figure 2: (a) A path which cannot appear in $G_\theta^r$. (b) A cycle $\mathcal{C}$ in $G_\theta^r$ which contains three arcs from $\mathcal{F}$ (the dashed arcs represent paths between nodes that may traverse $s$ or $t$).

We start by establishing the following lemma:

**Lemma 2.** *The residual network $G_\theta^r$ does not contain a path composed of arcs $(y_v, x_i)$, $(x_i, y_i)$ and $(y_i, x_u)$, where $i, u, v \in [n]$ and $(x_i, y_i) \in \mathcal{F}$.*

*Proof.* If such a path exists (see Figure 2a), then the flow $f_\theta$ corresponds to solution $(E_X, E_Y, E_Z)$ in which $x_i = 1$ and $y_i = 1$. In consequence $E_X \cap E_Y \neq \emptyset$ which violates Property 1. □

The following lemma is crucial in investigating the structure of cycles in $G_\theta^r$.

**Lemma 3.** *Each simple cycle in $G_\theta^r$ contains at most two arcs from $\mathcal{F}$.*

*Proof.* Assume, contrary to our claim, that there exists a cycle $\mathcal{C}$ in $G_\theta^r$ that contains at least three arcs from $\mathcal{F}$ (see Figure 2b). This cycle may (or may not) contain nodes $s$ and $t$. However, in all possible four cases the cycle $\mathcal{C}$ must violate Lemma 2, i.e. it must contain a path of the form shown in Figure 2a. □

7

Suppose we increase $\theta$ to some value $\theta' > \theta$ in $G^r_\theta$ and denote the resulting residual network as $G^r_{\theta'}$. Assume that a negative cycle $\mathcal{C}$ appears in $G^r_{\theta'}$. It is obvious that $\mathcal{C}$ must contain at least one arc from $\mathcal{F}$, since only the costs of these arcs decrease when $\theta$ increases. By Lemma 3, the cycle $\mathcal{C}$ contains either one or two arcs from $\mathcal{F}$.

Consider first the case when $\mathcal{C}$ contains exactly one arc from $\mathcal{F}$, say $(x_u, y_u)$. Then $\mathcal{C}$ cannot contain any arc from $\mathcal{B}$. Otherwise, after computing the cost of $\mathcal{C}$, the value of $\theta'$ would be canceled and $\mathcal{C}$ is a negative cycle in $G^r_\theta$, a contradiction. All the possible cases are shown in Figure 3. Notice that in the case (b), by Lemma 2, the arc $(y_u, t)$ must belong to $\mathcal{C}$. The dashed arcs represent paths between nodes, say $y_a$ and $x_b$, that do not use any arcs from $\mathcal{F}$ and $\mathcal{B}$ and none of the nodes $s$ or $t$. An easy computation shows that the cost of such paths equals $-C_a - \overline{c}_b$ as the remaining terms are canceled while traversing this path.
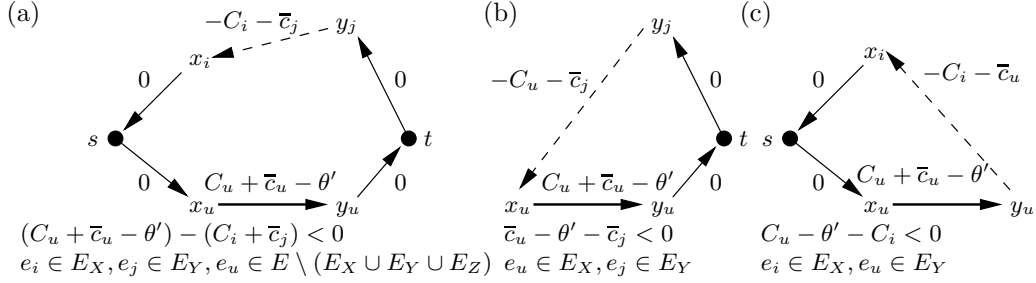


Figure 3: All possible situations when $\mathcal{C}$ contains only one arc from $\mathcal{F}$.

We now turn to the case when $\mathcal{C}$ contains two arcs from $\mathcal{F}$, say $(x_l, y_l)$ and $(x_m, y_m)$. It is easy to check that in this case $\mathcal{C}$ may also contain an arc from $\mathcal{B}$, but at most one such arc. The cycle $\mathcal{C}$ must traverse $s$ and $t$. Accordingly, it is of the form presented in Figure 4a, which is a consequence of Lemma 2 and Lemma 3. Assume that a path from $t$ to $s$ in $\mathcal{C}$ uses arc $(y_v, x_u)$ such that $u \neq v$ (see Figure 4a). Let us add to $\mathcal{C}$ two arcs $(x_u, y_j)$ and $(x_l, y_v)$ whose costs are $C_u + \overline{c}_j$ and $C_l + \overline{c}_v$, respectively (see Figure 4a). A trivial verification shows that the cost of the cycle $x_u \to y_j \to x_l \to y_v \to x_u$ is 0. Hence the cost of $\mathcal{C}$ is the sum of the costs of two disjoint cycles $\mathcal{C}_1$ and $\mathcal{C}_2$ (see Figures 4b and 4c). Since the cost of $\mathcal{C}$ is negative, the cost of at least one of $\mathcal{C}_1$ and $\mathcal{C}_2$ must be negative. If no arc from $\mathcal{B}$ belongs to $\mathcal{C}$, then the cycle $\mathcal{C}_1$ is of the form depicted in Figure 3c and $\mathcal{C}_2$ is of the form given in Figure 3b. Now suppose that $\mathcal{C}$ contains one arc from $\mathcal{B}$. If that arc is in path from $x_i$ to $y_j$ or from $x_u$ to $s$, then it belongs to cycle $\mathcal{C}_1$. In this case $\mathcal{C}_2$ is negative and is of the form shown in Figure 3b. If that arc is in path from $t$ to $y_v$, then we get a symmetric case. Consequently, the case $u \neq t$, can be reduced to the cases presented in Figure 3. The last case of cycle $\mathcal{C}$, which
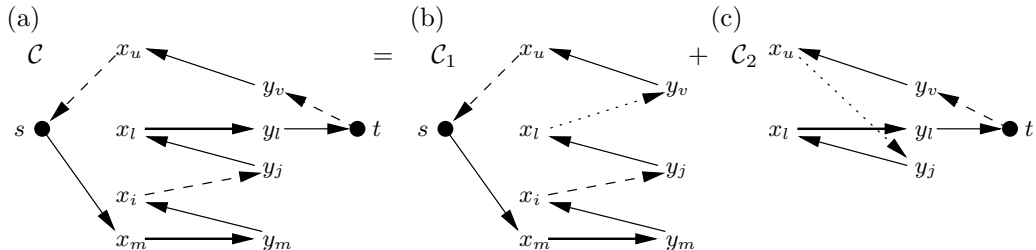


Figure 4: A situation when $\mathcal{C}$ contains two arcs from $\mathcal{F}$ and $u \neq v$.

$$(C_m + \overline{c}_l) - (C_r + \overline{c}_r) - \theta' < 0$$
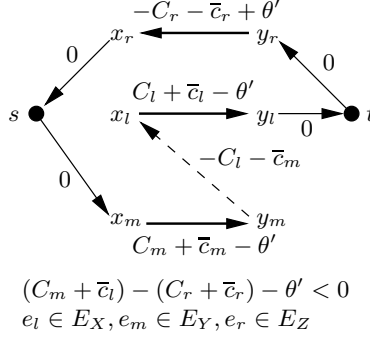$$e_l \in E_X, e_m \in E_Y, e_r \in E_Z$$

Figure 5: A situation when $\mathcal{C}$ contains two arcs from $\mathcal{F}$ and one arc from $\mathcal{B}$.

cannot be reduced to any cases previously discussed, and thus must be treated separately, is shown in Figure 5. Observe that in this case contains two arcs from $\mathcal{F}$ and one arc from $\mathcal{B}$.

We are thus led to the following optimality conditions.

**Theorem 4.** *Let $(E_X, E_Y, E_Z)$ be an optimal solution to (4) for $\theta \geq 0$. Then $(E_X, E_Y, E_Z)$ remains optimal for all $\theta' > \theta$ which satisfy the following inequalities:*

$$C_u + \overline{c}_u - C_i - \overline{c}_j \geq \theta' \qquad \text{for } e_i \in E_X, e_j \in E_Y, e_u \in E \setminus (E_X \cup E_Y \cup E_Z), \quad (5)$$
$$\overline{c}_u - \overline{c}_j \geq \theta' \qquad \text{for } e_u \in E_X, e_j \in E_Y, \quad (6)$$
$$C_u - C_i \geq \theta' \qquad \text{for } e_i \in E_X, e_u \in E_Y, \quad (7)$$
$$C_m + \overline{c}_l - (C_r + \overline{c}_r) \geq \theta' \qquad \text{for } e_l \in E_X, e_m \in E_Y, e_r \in E_Z. \quad (8)$$

*Proof.* Consider the optimal flow $f_\theta$ associated with $(E_X, E_Y, E_Z)$ and the corresponding residual network $G_\theta^r$. Fix $\theta' > \theta$ so that the conditions (5)-(8) are satisfied and suppose that $f_\theta$ is not optimal in $G_{\theta'}$. Then a negative cost directed cycle must appear in $G_{\theta'}^r$ with respect to $f_\theta$. This cycle must be of the form depicted in Figures 3 and 5. In consequence, at least of the conditions (5)-(8) must be violated. $\square$

Assume that $(E_X, E_Y, E_Z)$ is an optimal solution to (4) for some $\theta \geq 0$. This solution must satisfy the optimality conditions (5)-(8) (when $\theta' = \theta$). Suppose that some of the inequalities (5)-(8) is binding, i.e. it is satisfied as equality. In this case we can construct a new solution $(E_X', E_Y', E_Z')$ whose cost is the same as $(E_X, E_Y, E_Z)$ by using the following transformations:

1. If (5) is binding, then $E_Z' = E_Z \cup \{e_u\}$, $E_X' = E_X \setminus \{e_i\}$, $E_Y' = E_Y \setminus \{e_j\}$,

2. If (6) is binding, then $E_Z' = E_Z \cup \{e_u\}$, $E_X' = E_X \setminus \{e_u\}$, $E_Y' = E_Y \setminus \{e_j\}$,

3. if (7) is binding, then $E_Z' = E_Z \cup \{e_u\}$, $E_X' = E_X \setminus \{e_i\}$, $E_Y' = E_Y \setminus \{e_u\}$,

4. if (8) is binding, then $E_Z' = E_Z \cup \{e_l, e_m\} \setminus \{e_r\}$, $E_Y' = E_Y \setminus \{e_m\}$, $E_X' = E_X \setminus \{e_l\}$.

It is worth pointing out that the above transformations correspond to augmenting a flow around the corresponding cycle in $G_\theta^r$ whose cost is 0. In consequence the cost of $(E_X', E_Y', E_Z')$ is the same as the cost of $(E_X, E_Y, E_Z)$. Hence $(E_X', E_Y', E_Z')$ is also optimal to (4) for $\theta$. Furthermore it must satisfy conditions (5)-(8) (otherwise we could decrease the cost of $(E_X', E_Y', E_Z')$ by augmenting a flow around some negative cycle). Observe that $|E_Z'| = |E_Z| + 1$

9

and $(E'_X, E'_Y, E'_Z)$ also has Property 1. If none of the inequalities (5)-(8) is binding, then we increase $\theta$ until at least one inequality among (5)-(8) becomes binding, preserving the optimality of $(E_X, E_Y, E_Z)$. We start with $\theta = 0$ and we repeat the procedure described until we find a optimal solution $(E_X, E_Y, E_Z)$ to (4) for $\theta$, which is feasible in (3) and satisfies the complementary slackness condition: $\theta((p - k) - |E_Z|) = 0$. Notice that $|E_Z| = p - k$, when $\theta > 0$. Such a solution is an optimal one to the original problem (3). Indeed,

$$opt_2 \geq \phi(\theta) = \sum_{e_i \in E_X} C_i + \sum_{e_i \in E_Y} \overline{c}_i + \sum_{e_i \in E_Z} (C_i + \overline{c}_i - \theta) + (p - k)\theta \tag{9}$$

$$= \sum_{e_i \in E_X} C_i + \sum_{e_i \in E_Y} \overline{c}_i + \sum_{e_i \in E_Z} (C_i + \overline{c}_i). \tag{10}$$

The inequality (9) follows from the Lagrangian bounding principle (see, e.g., [1]). Applying the complementary slackness condition yields the equality (10). Moreover $(E_X, E_Y, E_Z)$ is feasible in (3), and so it is optimal solution to (3). We have thus arrived to the following theorem.

**Theorem 5.** *The* Recoverable Selection *problem for one second-stage scenario is solvable in $O((p - k + 1)n^2)$ time.*

*Proof.* By Lemma 1 the first solution for $\theta = 0$ can be computed in $O(n)$ time. Given an optimal solution for some $\theta \geq 0$, the next optimal solution can be found in $O(n^2)$ times, since we need to analyze $O(n^2)$ inequalities (5)-(8). The cardinality of $|E_Z|$ increases by 1 at each step until $|E_Z| = p-k$. Hence, the overall running time of the algorithm is $O((p-k+1)n^2)$. □

From Theorem 5 and (1) we immediately get the following corollary.

**Corollary 1.** *For scenario set $\Gamma^I$, the* Recoverable Selection *problem is solvable in $O((p - k + 1)n^2)$ time.*

## 4 Robust two stage selection

In this section we explore the complexity of Two-Stage Selection. We start by establishing the following easy observation:

**Observation 1.** *Under scenario set $\Gamma^I$,* Two-Stage Selection *is solvable in $O(n)$ time.*

*Proof.* Define $\hat{c}_e = \min\{C_e, \overline{c}_e\}$, $e \in E$. Let $Z$ be the set of $p$ items of the smallest values of $\hat{c}_e$. Clearly, $\hat{c} = \sum_{e \in Z} \hat{c}_e$ is a lower bound on $opt_1$. A solution $X$ with the cost $\hat{c}$ can be now constructed as follows. For each $e \in Z$, if $C_e \leq \overline{c}_e$, then we add $e$ to $X$; otherwise we add $e$ to $Y$. It holds $f_1(X, S) = \sum_{e \in X} C_e + \sum_{e \in Y} \overline{c}_e = \hat{c}$ and $X$ must be optimal. It is easy to see that $X$ can be computed in $O(n)$ time. □

The problem is much harder for scenario set $\Gamma^D$. The following theorems hold.

**Theorem 6.** *Under scenario set $\Gamma^D$,* Two-Stage Selection *is NP-hard when the number of scenarios equals three.*

*Proof.* See Appendix A. □

**Theorem 7.** *When $K$ is a part of input, then* TWO-STAGE SELECTION *under scenario set $\Gamma^D$ is strongly NP-hard. Furthermore, for every $\epsilon > 0$, it is NP-hard to approximate within $(1 - \epsilon) \ln n$.*

*Proof.* see Appendix A. $\qquad\square$

We now present a positive result for TWO-STAGE SELECTION under scenario set $\Gamma^D$. Namely, we construct an LP-based randomized approximation algorithm for this problem, which returns an $O(\log K + \log n)$-approximate solution with high probability. Consider the following linear program:

$$\mathcal{LP}(L) : \sum_{e \in E} C_e x_e + c_e^S y_e^S \leq L, \qquad\qquad S \in \Gamma^D, \qquad (11)$$

$$\sum_{e \in E} x_e + y_e^S = p, \qquad\qquad S \in \Gamma^D, \qquad (12)$$

$$x_e + y_e^S \leq 1, \qquad\qquad e \in E, S \in \Gamma^D, \qquad (13)$$

$$x_e = 0, \qquad\qquad e \notin E^1(L), \qquad (14)$$

$$y_e^S = 0, \qquad\qquad e \notin E^S(L), S \in \Gamma^D, \qquad (15)$$

$$x_e \geq 0, \qquad\qquad e \in E^1(L), \qquad (16)$$

$$y_e^S \geq 0, \qquad\qquad e \in E^S(L), S \in \Gamma^D, \qquad (17)$$

where $E^1(L) = \{e \in E : C_e \leq L\}$ and $E^S(L) = \{e \in E : c_e^S \leq L\}$. Minimizing $L$ subject to (11)-(17), we obtain an LP relaxation of TWO-STAGE SELECTION. Let $L^*$ denote the smallest value of the parameter $L$ for which $\mathcal{LP}(L)$ is feasible. Obviously, $L^*$ is a lower bound on $opt_1$, and can be determined in polynomial time by using binary search. We lose nothing by assuming, from now on, that $L^* = 1$, and all the item costs are such that $C_e, c_e^S \in [0, 1]$, $e \in E, S \in \Gamma^D$. One can easily meet this assumption by dividing all the item costs by $L^*$. Notice that we can assume that $L^* > 0$. Otherwise, when $L^* = 0$, there exists an optimal integral solution of the zero total cost. Such a solution can be constructed by picking all the items with zero first and second stage costs under all scenarios.

---

**Algorithm 1** A randomized algorithm for TWO-STAGE SELECTION.

---

1: $c_{\max} := \max_{e \in E}\{C_e, \max_{S \in \Gamma} c_e^S\}$
2: Use binary search in $[0, (n-1)c_{\max}]$ to find the minimal value of $L^*$ for which there exists a feasible solution $(\hat{\boldsymbol{x}}, (\hat{\boldsymbol{y}}^S)_{S \in \Gamma^D})$ to $\mathcal{LP}(L^*)$
   **Randomized rounding**
3: $\hat{t} := \lceil 32 \ln n + 8 \ln(2K) \rceil$
4: $X := \emptyset$, $Y^S := \emptyset$ for each $S \in \Gamma^D$
5: For each $e \in E$, flip an $\hat{x}_e$-coin $\hat{t}$ times, if it comes up heads at least once add $e$ to $X$
6: For each $e \in E$ and each $S \in \Gamma^D$, flip an $\hat{y}_e^S$-coin $\hat{t}$ times, if it comes up heads at least once add item $e$ to $Y^S$
   **End of randomized rounding**
7: Add to $X$ at most 4 arbitrary items, which have been not selected in steps 5 and 6
8: **If** $|X \cup Y^S| \geq p$ for each $S \in \Gamma^D$ **then return** $X, Y^S, S \in \Gamma^D$ **else fail**

---

Now our aim is to convert a feasible solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = (\hat{\boldsymbol{x}}, (\hat{\boldsymbol{y}}^S)_{S \in \Gamma}) \in [0, 1]^{n+nK}$ to $\mathcal{LP}(L^*)$ into a feasible solution of TWO-STAGE SELECTION. Let *x-coin* be a coin which comes up

head with probability $x \in [0, 1]$. We use such a device to construct a randomized algorithm (see Algorithm 1) for the problem. If Algorithm 1 outputs a solution such that $|X \cup Y^S| \geq p$ for each $S \in \Gamma^D$, then the sets $X$ and $Y^S$ can be converted into a feasible solution in the following way. For each scenario $S \in \Gamma^D$, If $e \in X \cap Y^S$, then we remove $e$ from $Y^S$. Next, if $|X \cup Y^S| > p$, then we remove arbitrary items, first from $Y^S$ and then from $X$ so that $|X \cup Y^S| = p$. Notice that this operation does not increase the total cost of the selected items under any scenario. The algorithm fails if $|X \cup Y^S| < p$ for at least one scenario $S$. We will show, however, that this bad event occurs with a small probability. Let us first analyze the cost of the obtained solution.

**Lemma 4.** *Fix scenario $S \in \Gamma^D$. The probability that the total cost of the items selected in the first and the second stage under $S$, after the randomized rounding, is at least $(\hat{t} + (\mathrm{e} - 1)\sqrt{\hat{t}\ln(2Kn^2)})opt_1$ is at most $\frac{1}{2Kn^2}$.*

*Proof.* Fix scenario $S \in \Gamma^D$. Let $\mathrm{X}_e$ be a random variable such that $\mathrm{X}_e = 1$ if item $e$ is included in $X$; and $\mathrm{X}_e = 0$ otherwise, and let $\mathrm{Y}_e^S$ be a random variable such that $\mathrm{Y}_e^S = 1$ if item $e$ is included in $Y^S$; and $\mathrm{Y}_e^S = 0$ otherwise. Obviously, $\Pr[\mathrm{X}_e = 1] = 1 - (1 - \hat{x}_e)^{\hat{t}}$ and $\Pr[\mathrm{Y}_e^S = 1] = 1 - (1 - \hat{y}_e^S)^{\hat{t}}$. Because $\hat{t} \geq 1$ and $\hat{x}_e, \hat{y}_e^S \in [0, 1]$, an easy computation shows that $\Pr[\mathrm{X}_e = 1] \leq \hat{x}_e \hat{t}$ and $\Pr[\mathrm{Y}_e^S = 1] \leq \hat{y}_e^S \hat{t}$. Define $\mathrm{C}^S = \sum_{e \in E} C_e \mathrm{X}_e + c_e^S \mathrm{Y}_e^S$ and let $\psi^S$ be the event that $\mathrm{C}^S > (\hat{t} + (\mathrm{e} - 1)\sqrt{\hat{t}\ln(2Kn^2)})opt_1$. It holds

$$\mathrm{E}[\mathrm{C}^S] = \sum_{e \in E} C_e \Pr[\mathrm{X}_e = 1] + c_e^S \Pr[\mathrm{Y}_e^S = 1] = \sum_{e \in E} C_e(1 - (1 - \hat{x}_e)^{\hat{t}}) + c_e^S(1 - (1 - \hat{y}_e^S)^{\hat{t}})$$

$$\leq \hat{t} \sum_{e \in E} C_e \hat{x}_e + c_e^S \hat{y}_e^S \leq \hat{t}L^* = \hat{t}. \tag{18}$$

Recall that the item costs are such that $C_e, c_e^S \in [0, 1]$, $e \in E$ and $L^* = 1 \leq opt_1$. Using (18) and applying Chernoff-Hoeffding bound given in [23, Theorem 1 and inequality (1.13) for $D(\hat{t}, 1/2Kn^2)$], we obtain

$$\Pr\left[\psi^S\right] \leq \Pr\left[\mathrm{C}^S > \hat{t} + (\mathrm{e} - 1)\sqrt{\hat{t}\ln(2Kn^2)}\right] < \frac{1}{2Kn^2}, \tag{19}$$

which completes the proof. □

We now analyze the feasibility of the obtained solution. In order to do this, it is convenient to see steps 5 and 6 of the algorithm in the following equivalent way. In a *round* we flip an $\hat{x}$-coin for each $e \in E$ and add $e$ to $X$ when it comes up head; we then flip an $\hat{y}^S$-coin for each $e \in E$ and $S \in \Gamma^D$ and add $e$ to $Y^S$ if it comes up head. Clearly, steps 5 and 6 can be seen as performing $\hat{t}$ such rounds independently. Let us fix scenario $S \in \Gamma^D$. Let $X_t$ and $Y_t^S$ be the sets of items selected in the first and second stage under $S$ (i.e. added to $X$ and $Y^S$), respectively, after $t$ rounds. Define $E_t^S = E \setminus (X_t \cup Y_t^S)$ and $|E_t^S| = N_t^S$. Initially, $E_0^S = E$, $N_0^S = n$. Let $P_t^S$ denote the number of items remaining for selection out of the set $E_t^S$ under scenario $S$ after the $t$th round. Initially $P_0^S = p$. We say that a round $t$ is "successful" if either $P_{t-1}^S < 5$ (at most 4 items is to be selected) or $P_t^S < 0.88P_{t-1}^S$; otherwise, it is "failure".

**Lemma 5.** *Fix scenario $S \in \Gamma^D$. The conditional probability that round $t$ is "successful", given any set of items $E_{t-1}^S$ and number $P_{t-1}^S$, is at least $1/2$.*

*Proof.* If $P_{t-1}^S < 5$, then we are done. Assume that $P_{t-1}^S \geq 5$ and consider the set of items $E_{t-1}^S$, $|E_{t-1}^S| = N_{t-1}^S$ and the number of items $P_{t-1}^S$, remaining for selection in round $t$ (i.e. after round $t-1$). Let $\mathrm{I}_e$ be a random variable such that $\mathrm{I}_e = 1$ if item $e$ is picked from $E_{t-1}^S$; and $\mathrm{I}_e = 0$, otherwise. It is easily seen that $\Pr[\mathrm{I}_e = 1] = 1 - (1 - \hat{x}_e)(1 - \hat{y}_e^S)$. The expected number of items selected out of $E_{t-1}^S$ in round $t$ is

$$\mathrm{E}\left[\sum_{e \in E_{t-1}^S} \mathrm{I}_e\right] = \sum_{e \in E_{t-1}^S} \Pr[\mathrm{I}_e = 1] = N_{t-1}^S - \sum_{e \in E_{t-1}^S} (1 - \hat{x}_e)(1 - \hat{y}_e^S)$$

$$\geq N_{t-1}^S - \sum_{e \in E_{t-1}^S} \left(1 - \frac{\hat{x}_e + \hat{y}_e^S}{2}\right)^2 \geq N_{t-1}^S - \sum_{e \in E_{t-1}^S} \left(1 - \frac{\hat{x}_e + \hat{y}_e^S}{2}\right) \geq \frac{P_{t-1}^S}{2}.$$

The first inequality follows from the fact that $ab \leq (a + b)/2$ for any $a, b \in [0, 1]$ (indeed, $0 \leq (a - b)^2 = a^2 - 2ab + b^2 \leq a - 2ab + b$). The last inequality follows from the fact that the feasible solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ satisfies constraints (12). Using Chernoff bound (see, e.g.,[20], Theorem 4.2 and inequality (4.6) for $\delta = \sqrt{4 \ln 2 / P_{t-1}^S}$), we get

$$\Pr\left[\sum_{e \in E_{t-1}^S} \mathrm{I}_e < \frac{P_{t-1}^S}{2} - \sqrt{P_{t-1}^S \ln 2}\right] \leq \frac{1}{2}.$$

Thus, with probability at least $1/2$, the number of selected items in round $t$ is at least $P_{t-1}^S/2 - \sqrt{P_{t-1}^S \ln 2}$. Hence, with probability at least $1/2$ it holds

$$P_t^S \leq P_{t-1}^S - P_{t-1}^S/2 + \sqrt{P_{t-1}^S \ln 2} = P_{t-1}^S(1/2 + \sqrt{\ln 2 / P_{t-1}^S}).$$

Consequently, when $P_{t-1}^S \geq 5$ we get

$$P_t^S < 0.88 P_{t-1}^S$$

with probability at least $1/2$. $\qquad\square$

**Lemma 6.** *Fix scenario $S \in \Gamma^D$. The probability of the event that $P_{\hat{t}}^S \geq 5$ is at most $1/(2Kn^2)$.*

*Proof.* Let $\xi^S$ be the event that $P_{\hat{t}}^S \geq 5$, i.e. that the number of items remaining for selection after $\hat{t} = \lceil 32 \ln n + 8 \ln 2K \rceil$ rounds is at last 5. We now estimate the number $\ell$ of successful rounds which are enough to achieve $P_\ell^S < 5$. It is easy to see that $\ell$ satisfies $(0.88)^\ell p \leq 4$. In particular, this inequality holds when $\ell \geq 8 \ln p$. Let $\mathrm{Z}$ be a random variable denoting the number of "successful" rounds among $\hat{t}$ performed rounds. We estimate $\Pr[\mathrm{Z} < 8 \ln p]$ from above by $\Pr[\mathrm{B}(\hat{t}, 1/2) < 8 \ln p]$, where $\mathrm{B}(\hat{t}, 1/2)$ is a binomial random variable. This can be done, since we have a lower bound on success of given any history. Applying Chernoff bound (see, e.g., [20, Theorem 4.2 and inequality (4.6) for $\delta = \ln(2Kn^2)/\ln(2Kn^4)$] and $\hat{t} = \lceil 32 \ln n + 8 \ln(2K) \rceil$) and the fact that $p < n$, we obtain the following upper bound:

$$\Pr[\xi^S] \leq \Pr[\mathrm{Z} < 8 \ln p] \leq \Pr[\mathrm{B}(\hat{t}, 1/2) < 8 \ln p] \leq \mathrm{e}^{-(8 \ln n + 4 \ln(2K))^2/2(16 \ln n + 4 \ln(2K))}$$

$$< \mathrm{e}^{-(8 \ln n + 4 \ln(2K))/4} = \frac{1}{2Kn^2}, \tag{20}$$

and the lemma follows. $\qquad\square$

Lemma 4 and Lemma 6 (see inequalities (19) and (20)) and the union bound imply that $\Pr[\xi^{S_1} \cup \cdots \cup \xi^{S_K} \cup \psi^{S_1} \cdots \cup \psi^{S_K}] < 1/n^2$. The addition of 4 elements to $X$ in step 7 can increase the cost of the computed solution by at most $4 \cdot opt_1$. As all the costs are nonnegative, repairing $X \cup Y^S$, $S \in \Gamma^D$, to obtain a feasible solution cannot increase the cost of the computed solution. We thus get the following result:

**Theorem 8.** *There is a polynomial time randomized algorithm for* TWO-STAGE SELECTION *that returns, with probability at least $1 - \frac{1}{n^2}$, a solution whose cost is $O(\log K + \log n)opt_1$.*

It is worth pointing out that if $K = poly(n)$, then our randomized algorithm gives the best approximation ratio up to a constant factor (see Theorem 7).

## 5   Conclusions and open problems

In this paper we have discussed two robust versions of the SELECTION problem, which have a two-stage nature. In the first problem, a partial solution is formed in the first stage and completed optimally when a true state of the world reveals. In the second problem a complete solution must be formed in the first stage, but it can be modified to some extent after a true state of the world becomes known. Such two-stage problems often appear in practical applications. In this paper we have presented some positive and negative complexity results for two types of uncertainty representations, namely the discrete and interval ones. In particular, we have shown that both problems are polynomially solvable for one second-stage scenario, which implies the they are also polynomially solvable for the interval uncertainty representation. We believe that a similar method might be applied to other combinatorial optimization problems, in particular for those possessing a matroidal structure. When the number of scenarios is a part of input, then the recoverable model is not at all approximable and the two-stage model is hard to approximate within $(1 - \epsilon) \log n$ for any $\epsilon > 0$. We have shown that the latter one admits a randomized $O(\log n + \log K)$-approximation algorithm.

There are still some open questions concerning the considered problems. The complexity of TWO-STAGE SELECTION for two scenarios is open. Also, a deterministic $\log n$-approximation algorithm for this problem may exists and it can be a subject of further research. When $K$ is constant, then both robust problems are only proven to be weakly NP-hard. So, they might be solved in pseudopolynomial time and even admit an FPTAS. The interval uncertainty representation can be generalized by adopting the scenario set proposed in [3]. The complexity of both robust problems under this scenario set is open.

## References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: theory, algorithms, and applications.* Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] I. Averbakh. On the complexity of a class of combinatorial optimization problems with uncertainty. *Mathematical Programming*, 90:263–272, 2001.

[3] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming*, 98:49–71, 2003.

[4] C. Büsing. *Recoverable Robustness in Combinatorial Optimization*. PhD thesis, Technical University of Berlin, Berlin, 2011.

[5] C. Büsing. Recoverable robust shortest path problems. *Networks*, 59:181–189, 2012.

[6] C. Büsing, A. M. C. A. Koster, and M. Kutschka. Recoverable robust knapsacks: the discrete scenario case. *Optimization Letters*, 5:379–392, 2011.

[7] E. Conde. An improved algorithm for selecting p items with uncertain returns according to the minmax regret criterion. *Mathematical Programming*, 100:345–353, 2004.

[8] K. Dhamdhere, V. Goya, R. Ravi, and M. Singh. How to Pay, Come What May: Approximation Algorithms for Demand-Robust Covering Problems. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 367–378, 2005.

[9] I. Dinur and D. Steurer. Analytical approach to parallel repetition. In *Symposium on Theory of Computing, STOC*, pages 624–633, 2014.

[10] B. Doerr. Improved approximation algorithms for the Min-Max selecting Items problem. *Information Processing Letters*, 113:747–749, 2013.

[11] M. R. Garey and D. S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[12] A. Ghouila-Houri. Caractérisation des matrices totalement unimodulaires. *C. R. Acad. Sci. Paris*, 254:1192–1194, 1962.

[13] T. Ibaraki and N. Katoh. *Resource Allocation Problems*. The MIT Press, 1988.

[14] A. Kasperski, A. Kurpisz, and P. Zieliński. Approximating the min-max (regret) selecting items problem. *Information Processing Letters*, 113:23–29, 2013.

[15] A. Kasperski and P. Zieliński. On the approximability of minmax (regret) network optimization problems. *Information Processing Letters*, 109:262–266, 2009.

[16] A. Kasperski and P. Zieliński. On the approximability of robust spanning problems. *Theoretical Computer Science*, 412:365–374, 2011.

[17] I. Katriel, C. Kenyon-Mathieu, and E. Upfal. Commitment under uncertainty: Two-stage matching problems. *Theoretical Computer Science*, 408:213–223, 2008.

[18] P. Kouvelis and G. Yu. *Robust Discrete Optimization and its applications*. Kluwer Academic Publishers, 1997.

[19] C. Liebchen, M. E. Lübbecke, R. H. Möhring, and S. Stiller. The Concept of Recoverable Robustness, Linear Programming Recovery, and Railway Applications. In *Robust and Online Large-Scale Optimization*, volume 5868 of *Lecture Notes in Computer Science*, pages 1–27. Springer-Verlag, 2009.

[20] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[21] E. Nasrabadi and J. B. Orlin. Robust optimization with incremental recourse. *CoRR*, abs/1312.4075, 2013.

[22] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover Publications Inc., 1998.

[23] P. Raghavan. Probabilistic Construction of Deterministic Algorithms: Approximating Packing Integer Programs. *Journal of Computer and System Sciences*, 37:130–143, 1988.

# A    Some proofs

*Proof.* (Theorem 1) Consider an instance of the NP-complete BALANCED PARTITION problem, where $\mathcal{A} = \{a_1, \ldots, a_{2n}\}$, and $\sum_{i \in [2n]} a_i = 2b$. Recall that we ask if there is a subset $I \subset [2n]$ such that $|I| = n$ and $\sum_{i \in I} a_i = \sum_{i \in [2n] \setminus I} a_i$. We construct the corresponding instance of RECOVERABLE SELECTION as follows. The set of items $E = \{e_1, \ldots, e_{2n}, f_1, \ldots, f_k, r_1, \ldots, r_k\}$, where $k > 0$ and $r_1, \ldots, r_k$ are $k$ *recoverable items*. The first-stage costs and two second-stage cost scenarios are depicted in Table 1, where $M_1$ and $M_2$ are large constants such that $M_1 \ll M_2$, say $M_1 = 4nb$ and $M_2 = 2nM_1$. We also fix $p = n + k$.

Table 1: The instance of RECOVERABLE SELECTION.

|        | $C_e$ | $S_1$         | $S_2$             |
|--------|-------|---------------|-------------------|
| $e_1$  | $M_1$ | $b + a_1$     | $b + 2b/n - a_1$  |
| $e_2$  | $M_1$ | $b + a_2$     | $b + 2b/n - a_2$  |
| $\vdots$ | $\vdots$ | $\vdots$   | $\vdots$          |
| $e_{2n}$ | $M_1$ | $b + a_{2n}$ | $b + 2b/n - a_{2n}$ |
| $f_1$  | $0$   | $M_2$         | $M_2$             |
| $\vdots$ | $\vdots$ | $\vdots$   | $\vdots$          |
| $f_k$  | $0$   | $M_2$         | $M_2$             |
| $r_1$  | $M_2$ | $0$           | $0$               |
| $\vdots$ | $\vdots$ | $\vdots$   | $\vdots$          |
| $r_k$  | $M_2$ | $0$           | $0$               |

We now show that the answer to BALANCED PARTITION is yes if and only if $opt_2 \leq nM_1 + (n+1)b$. Suppose that answer to PARTITION is yes. Thus there exists a subset $I \subset [2n]$ such that $|I| = n$ and $\sum_{i \in I} a_i = b$. Let $X$ be a solutions composed of $n + k$ items $e_i$, $i \in I$ and $f_i, i \in [k]$. Solution $X$ does not use any recovery item. Under two scenarios $S_1$ and $S_2$ we can decrease its cost to $nM_1 + nb + b$ by replacing items $f_1, \ldots, f_k$ with the recovery items $r_1, \ldots, r_k$. In consequence, $opt_2 \leq nM_1 + (n+1)b$.

Assume that $opt_2 \leq nM_1 + (n+1)b$. Let $X$ be an optimal solution, $|X| = n + k$. Let $I = \{i : e_i \in X\}$. We claim that $|I| = n$. If $|I| < n$, then $X$ must contain at least one item $r_i, i \in [k]$, with the first-stage cost of $M_2$, since the number of recovery items equals $k$. This gives $\max\{f_2(X, S_1), f_2(X, S_1)\} \geq M_2 > nM_1 + (n+1)b$, a contradiction. If $|I| > n$, then $\sum_{e \in X} C_e \geq (n+1)M_1$. Since $M_1 > (n+1)b$, $\max\{f_2(X, S_1), f_2(X, S_1)\} >$

Table 2: The scenarios set for $x_1, \ldots, x_3$ and the clauses $(x_1 \vee \overline{x}_2 \vee \overline{x}_3)$, $(\overline{x}_1 \vee x_2 \vee x_3)$, $(x_1 \vee x_2 \vee x_3)$, $p = 3$ and $k = 1$.

| $E$ | $C_e$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $e_1^1$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $e_2^1$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| $e_3^1$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| $e_1^2$ | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| $e_2^2$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $e_3^2$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| $e_1^3$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| $e_2^3$ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| $e_3^3$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| $r$ | $n$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

$nM_1 + (n+1)b$, a contradiction. Hence $|I| = n$ and $X$ must also contain all the items $f_1, \ldots, f_k$. Since the costs of these items under $S_1$ and $S_2$ are very large, they must be replaced by $r_1, \ldots, r_k$ under both scenarios. In consequence, the cost of $X$ is as follows $\max\{nM_1 + nb + \sum_{i \in I} a_i, nM_1 + nb + 2b - \sum_{i \in I} a_i\}$ and it is equal to $nM_1 + (n+1)b$ if $\sum_{i \in I} a_i = b$, i.e. when the answer to BALANCED PARTITION is yes. $\square$

*Proof.* (Theorem 2) Consider an instance of the strongly NP-complete 3-SAT problem [11], in which we are given a set of boolean variables $x_1, \ldots, x_n$ and a collection of clauses $A_1, \ldots A_m$, where each clause is disjunction of exactly three literals (variables or their negations). We ask if there is a truth assignment to the variables which satisfies all the clauses. We now construct the corresponding instance of the RECOVERABLE SELECTION problem as follows. We associate with each clause $A_i = (l_1^i \vee l_2^i \vee l_3^i)$ three items $e_1^i$, $e_2^i$, $e_3^i$ corresponding to three literals in $A_i$. We also create one recovery item $r$. This gives us the item set $E$, such that $|E| = 3m+1$. The first-stage cost of the recovery item $r$ is set to $n$ the first-stage costs of the remaining items are set to zero. The scenario set $\Gamma^D$ is formed as follows. For each pair of items $e_u^s$ and $e_w^t$, that corresponds to contradictory literals $l_u^s$ and $l_w^t$, i.e. $l_u^s = \overline{l}_w^t$, we create scenario $S$ such that under this scenario the costs of $e_u^s$ and $e_w^t$ are set to 1 and the costs of all the remaining items are set to 0. For each clause $A_i$, $i \in [m]$, we form scenario in which the costs of $e_1^i$, $e_2^i$, $e_3^i$ are set to 1 and the rest of items have zero costs. We complete the reduction by setting $p = m$ and $k = 1$. An example of the reduction is shown in Table 2.

It is easy to check that if the answer to 3-SAT is yes, then there is a selection $X$ out of $E$ such that $|X| = p$, containing items that do not correspond to contradictory literals and literals from the same clauses ($X$ corresponds to exactly one literal from each clause). Thus the costs of $X$ under each scenario $S \in \Gamma^D$ are at most 1. Hence we can decrease them to zero by replacing an item from $X$ with the cost of 1 under any scenario with recovery item $r$ and $\max_{S \in \Gamma^D} f_2(X, S) = 0$. On the other hand, if the answer is no, then all selections $X$, $|X| = p$, must contain at least two items corresponding to contradictory literals or literals from the same clause. Note that the recovery action under each scenario is limited to one. So, $\max_{S \in \Gamma^D} f_2(X, S) \geq 1$. Accordingly RECOVERABLE SELECTION with scenario set $\Gamma^D$ is strongly NP-hard and not at all approximable unless P=NP. $\square$

*Proof.* (Theorem 3) In order to show that the constraint matrix of (3) is totally unimodular we

Table 3: The constraint matrix of (3).

| | $x_1$ | $x_2$ | $x_3$ | $\ldots$ | $x_n$ | $y_1$ | $y_2$ | $y_3$ | $\ldots$ | $y_n$ | $z_1$ | $z_2$ | $z_3$ | $\ldots$ | $z_n$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$: | 1 | 1 | 1 | $\ldots$ | 1 | 0 | 0 | 0 | $\ldots$ | 0 | 1 | 1 | 1 | $\ldots$ | 1 |
| $a_2$: | 0 | 0 | 0 | $\ldots$ | 0 | 1 | 1 | 1 | $\ldots$ | 1 | 1 | 1 | 1 | $\ldots$ | 1 |
| $a_3$: | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 0 | 1 | 1 | 1 | $\ldots$ | 1 |
| $b_1$: | 1 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 0 | 1 | 0 | 0 | $\ldots$ | 0 |
| $b_2$: | 0 | 1 | 0 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 1 | 0 | $\ldots$ | 0 |
| $b_3$: | 0 | 0 | 1 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 1 | $\ldots$ | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| $b_n$: | 0 | 0 | 0 | $\ldots$ | 1 | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 1 |
| $c_1$: | 0 | 0 | 0 | $\ldots$ | 0 | 1 | 0 | 0 | $\ldots$ | 0 | 1 | 0 | 0 | $\ldots$ | 0 |
| $c_2$: | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 1 | 0 | $\ldots$ | 0 | 0 | 1 | 0 | $\ldots$ | 0 |
| $c_3$: | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 1 | $\ldots$ | 0 | 0 | 0 | 1 | $\ldots$ | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| $c_n$: | 0 | 0 | 0 | $\ldots$ | 0 | 0 | 0 | 0 | $\ldots$ | 1 | 0 | 0 | 0 | $\ldots$ | 1 |

will use the following Ghouira-Houri's characterization of totally unimodular matrices [12]. An $m \times n$ integral matrix is totally unimodular if and only if each set $R \subseteq [m]$ can be partitioned into two disjoint sets $R_1$ and $R_2$ such that

$$\sum_{i \in R_1} a_{ij} - \sum_{i \in R_2} a_{ij} \in \{-1, 0, 1\}, \quad j \in [n]. \tag{21}$$

This criterion can alternatively be stated as follows. An $m \times n$ integral matrix is totally unimodular if and only if for any subset of rows $R = \{r_1, \ldots, r_l\} \subseteq [m]$ there exists a coloring of rows of $R$, with 1 or -1, i.e. $l(r_i) \in \{-1, 1\}$, $r_i \in R$, such that the weighted sum of every column (while restricting the sum to rows in $R$) is $-1$, 0 or 1. The constraint matrix of (3) is shown in Table 3. Consider a subset of rows $R = A \cup B \cup C$, where $A \subseteq \{a_1, a_2, a_3\}$, $B \subseteq \{b_1, \ldots, b_n\}$, $C \subseteq \{c_1, \ldots, c_n\}$. We examine the following cases and for each of them we show a valid coloring.

1. $A = \emptyset$ . Then $l(b_i) = 1$ for $b_i \in B$ and $l(c_i) = -1$ for $c_i \in C$.

2. $A = \{a_1\}$. Then $l(a_1) = 1$, $l(b_i) = -1$ for $b_i \in B$ and $l(c_i) = -1$ if $c_i \in C$.

3. $A = \{a_2\}$. Symmetric to Case 2.

4. $A = \{a_3\}$. Then $l(a_3) = 1$, $l(b_i) = -1$ for $b_i \in B$ and $l(c_i) = -1$ if $c_i \in C$.

5. $A = \{a_1, a_2\}$. Then $l(a_1) = 1$, $l(a_2) = -1$, $l(b_i) = -1$ for $b_i \in B$ and $l(c_i) = 1$ for $c_i \in C$.

6. $A = \{a_1, a_3\}$. Then $l(a_1) = 1$, $l(a_3) = -1$, $l(b_i) = -1$ for $b_i \in B$ and $l(c_i) = 1$ for $c_i \in C$.

7. $A = \{a_2, a_3\}$. Symmetric to Case 6.

8. $A = \{a_1, a_2, a_3\}$. Then $l(a_1) = 1$, $l(a_2) = 1$, $l(a_3) = -1$, $l(b_i) = -1$ for $b_i \in B$, $l(c_i) = -1$ for $c_i \in C$.

□

*Proof.* (Theorem 6) Consider the following NP-complete BALANCED PARTITION problem [11]. We are given a collection of positive integers $\mathcal{A} = (a_1, a_2, \ldots, a_n)$, where $a_i > 0$ for all $i \in [n]$, and $n$ is even. We ask if there is a subset $I \subseteq [n]$, $|I| = n/2$, such that $\sum_{i \in I} a_i = \frac{1}{2} \sum_{i \in [n]} a_i$. Given an instance of BALANCED PARTITION, we form the corresponding instance of TWO-STAGE SELECTION as follows. Fix $b = \frac{1}{2} \sum_{i \in [n]} a_i$. The set of items is $E = \{e_1, \ldots, e_n, f_1, \ldots, f_{n/2}\}$. The first stage costs and three second-stage cost scenarios $S_1$, $S_2$, and $S_3$, are shown in Table 4, where $M$ is a sufficiently large constant, say $M = 4nb$. Finally, we set $p = n$. We will show that the answer to BALANCED PARTITION is yes if and only if $opt_1 \leq (\frac{3}{2}n + 3)b$ in the resulting instance of TWO-STAGE SELECTION.

Table 4: The instance of TWO-STAGE SELECTION corresponding to BALANCED PARTITION.

|           | $C_e$       | $S_1$   | $S_2$         | $S_3$ |
|-----------|-------------|---------|---------------|-------|
| $e_1$     | $3b + a_1$  | $2a_1$  | $6b/n - a_1$  | $M$   |
| $e_2$     | $3b + a_2$  | $2a_2$  | $6b/n - a_2$  | $M$   |
| $\vdots$  | $\vdots$    | $\vdots$| $\vdots$      |       |
| $e_n$     | $3b + a_n$  | $2a_n$  | $6b/n - a_n$  | $M$   |
| $f_1$     | $M$         | $M$     | $M$           | $0$   |
| $\vdots$  | $\vdots$    | $\vdots$| $\vdots$      |       |
| $f_{n/2}$ | $M$         | $M$     | $M$           | $0$   |

Assume that the answer to BALANCED PARTITION is yes. Let $I \subseteq [n]$, $|I| = n/2$, be such that $\sum_{i \in I} a_i = b$. Let us form solution $X$ by choosing all items $e_i$, $i \in I$. Clearly $|X| = n/2 \leq p$ and thus $X$ is feasible. Under scenarios $S_1$ and $S_2$ we add to $X$ all the items $e_i$ for $i \in [n] \setminus X$. Easy computations show that $f_1(X, S_1) = f(X, S_2) = (\frac{3}{2}n + 3)b$. Furthermore, $f(X, S_3) = (\frac{3}{2}n + 1)b \leq (\frac{3}{2}n + 3)b$, since we can complete $X$ by $f_1, \ldots, f_{n/2}$ under $S_3$. We thus get $opt_1 \leq (\frac{3}{2}n + 3)b$.

Assume that $opt_1 \leq (\frac{3}{2}n + 3)b$. We first note that each optimal solution $X$ must be such that $|X| = n/2$. Indeed, suppose that $|X| > n/2$. Then, since all $a_i$ are positive, the first stage cost of $X$ is greater than $\frac{3}{2}nb + 3b = (\frac{3}{2}n + 3)b \geq opt_1$. On the other hand, if $|X| < n/2$, then we have to add more than $n/2$ items to $X$ under $S_3$ and $opt_1 \geq M$. Let $X$ be an optimal solution. Since $|X| = n/2$, the value of $f(X, S_3)$ is not greater than $f(X, S_1)$. Hence

$$opt_1 = \max\{f(X, S_1), f(X, S_2)\}.$$

Let $\sum_{e \in X} C_e = \frac{3}{2}nb + b_1$, where $b_1 = \sum_{e \in X} a_i$. Then $f(X, S_1) = \frac{3}{2}nb + b_1 + 2(2b - b_1) = \frac{3}{2}nb + 4b - b_1$ and $f(X, S_2) = \frac{3}{2}nb + b_1 + 3b - (2b - b_1) = \frac{3}{2}nb + b + 2b_1$. In consequence

$$opt_1 = \frac{3}{2}nb + \max\{4b - b_1, b + 2b_1\}.$$

It holds $\max\{4b - b_1, b + 2b_1\} \leq 3b$ if and only if $b_1 = b$. Hence $opt_1 \leq (\frac{3}{2}n + 3)b$ if and only if $b = b_1$, i.e. when the set $I = \{i : e_i \in X\}$ proves that the answer to BALANCED PARTITION is yes. □

*Proof.* (Theorem 7) Consider the following MIN-SET COVER problem. We are given a finite set $\mathcal{U}$, called the *universe*, and a collection $\mathcal{A} = \{A_1, \ldots, A_m\}$ of subsets of $\mathcal{U}$. A subset

$\mathcal{D} \subseteq \mathcal{A}$ covers $\mathcal{U}$ (it is called a *cover*) if for each element $i \in \mathcal{U}$, there exists $A_j \in \mathcal{D}$ such that $i \in A_j$. We seek a cover $\mathcal{D}$ of the smallest size $|\mathcal{D}|$. The MIN-SET COVER problem is known to be strongly NP-hard and for every $\epsilon > 0$, it is also NP-hard to approximate within $(1 - \epsilon)\ln n$, where $n$ is the size of the instance [9]. We now show a cost preserving reduction from MIN-SET COVER to TWO-STAGE SELECTION.

Table 5: The instance of TWO-STAGE SELECTION for $\mathcal{U} = \{1,\ldots,7\}, \mathcal{C} = \{\{2,4,3\},\{1\},\{3,5,7\},\{1,4,6,7\},\{2,5,6\},\{1,6\}\}$

|  | $C_e$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|---|
| $e_{\{2,4,3\}}$ | 1 | 0 | $M$ | $M$ | $M$ | 0 | 0 | 0 |
| $e_{\{1\}}$ | 1 | $M$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $e_{\{3,7\}}$ | 1 | 0 | 0 | $M$ | 0 | 0 | 0 | $M$ |
| $e_{\{1,4,6,7\}}$ | 1 | $M$ | 0 | 0 | $M$ | 0 | $M$ | $M$ |
| $e_{\{2,5,6\}}$ | 1 | 0 | $M$ | 0 | 0 | $M$ | $M$ | 0 |
| $e_{\{1,6\}}$ | 1 | $M$ | 0 | 0 | 0 | 0 | $M$ | 0 |
| $u_1$ | $M$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $u_2$ | $M$ | 0 | 0 | 0 | 0 | $M$ | 0 | 0 |
| $u_3$ | $M$ | 0 | $M$ | $M$ | $M$ | $M$ | 0 | $M$ |
| $u_4$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ |
| $u_5$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ |
| $u_6$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ |

Given an instance $(\mathcal{U},\mathcal{C})$ of MIN-SET COVER we construct the corresponding instance of TWO-STAGE SELECTION as follows. For each set $A \in \mathcal{A}$ we create an item $e_A$ with first-stage cost equal to 1. We also create additional $m$ items labeled as $u_1,\ldots,u_m$ with the first-stage costs equal to a sufficiently large constant $M$. The scenario set $\Gamma^D$ is constructed in the following way. For each element $i \in \mathcal{U}$ we form scenario $S_i$ under which the cost of $e_A = M$ if $i \in A$ and 0 otherwise. Let $r$ be the number of 0's created under scenario $S_i$ for the items $e_A$, $A \in \mathcal{A}$. Hence $r$ is the number of sets in $\mathcal{A}$ which do not contain $i$. We set the costs of $u_1,\ldots u_r$ equal to 0 and the costs of $u_{r+1},\ldots,u_m$ equal to $M$. Notice that under each scenario $S_i$ exactly $m$ items have costs equal to 0. We fix $p = m + 1$.

We now show that there is a cover $\mathcal{D}$ such that $|\mathcal{D}| \leq k$ if and only if there is a solution $X$ such that $f_1(X) \leq k$. Let $\mathcal{D}$ be a cover of a size at most $k$. In the first stage we choose the items $e_A$ for each $A \in \mathcal{D}$. Consider any scenario $S_i$. Let $E' = E \setminus X$. Since the element $i$ is covered, there must exist at least $m - |X| + 1$ elements in $E'$ with 0 costs under $S_i$. We complete $X$ by using these elements, which gives $f_1(X,S) = k$ and, consequently, $opt_1 \leq k$. Assume now that $opt_1 \leq k$. By the construction, there must exist a solution $X$ such that $|X| = k_1 \leq k$, i.e. we choose exactly $k_1$ elements in the first stage. Let $\mathcal{D} = \{A : e_A \in X\}$, $|\mathcal{D}| = k_1$. Consider scenario $S_i$. It must be that $f_1(X,S_i) = k_1$, so w must be able to complete $X$ with $m - k_1 + 1$ elements of 0 cost under $S_i$. It is possible only when the cost of some $e_A \in X$ under $S_i$ is $M$, i.e. when the element $i$ is covered by $\mathcal{D}$. In consequence, each element $i$ is covered by $\mathcal{D}$ and the size of the minimum cover is not greater than $k$. It is clear that the presented reduction is cost preserving and the theorem follows. $\square$