Journal of Combinatorial Optimization (2023) 45:8
https://doi.org/10.1007/s10878-022-00941-3

Greedy guarantees for minimum submodular cost submodular/non-submodular cover problem

Majun Shi[1] · Zishen Yang[1] · Wei Wang[1]

Accepted: 20 October 2022 / Published online: 18 November 2022

Abstract

Minimum Submodular Cost Submodular Cover problem (MIN-SCSC) often occurs naturally in the areas of combinatorial optimization and particularly machine learning. It is well-known that the greedy algorithm proposed by Wan et al. yields a $\rho H(\delta)$-approximation for an integer-valued submodular function f, where ρ is the curvature of submodular cost function c, δ is the maximum value of f over all singletons and $H(\delta)$ is the δ-th harmonic number (Wan et al. in Comput Optim Appl 45(2):463–474). In this paper, we first extend MIN-SCSC to Minimum Submodular Cost Non-submodular Cover problem and analyze the performances of the widely used greedy algorithm for integer-valued and fraction-valued potential functions respectively. In addition, we also study MIN-SCSC with fraction-valued potential functions, with a new analysis of the performance ratio of the greedy algorithm, improving upon the result of Wan et al. (2010).

Keywords Greedy algorithm · Performance ratio · Submodular function · Submodular cover

1 Introduction

Let $f : 2^V \to \mathbb{R}_+$ be a normalized monotone set function defined on the ground set V. Let $c : 2^V \to \mathbb{R}_+$ be a non-negative cost function. Let $\Omega_f = \{A \subseteq V | f(A) = f(V)\}$ be the set of all feasible subsets, where f is called *potential* function. We consider the

Supported by the National Natural Science Foundation of China under Grant No. 11971376.

✉ Wei Wang
wang_weiw@xjtu.edu.cn

1 School of Mathematics and Statistics, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an 710049, People's Republic of China

following combinatorial optimization problem:

$$\begin{aligned} &\min \ c(A) \\ &\text{s.t. } A \in \Omega_f \end{aligned}$$

There exists a large body of literature studying the above problem and there is a beautiful line of research in this field. When c is a linear function and f is a submodular function, the above problem is known as Minimum Submodular Cover problem. The greedy algorithm proposed by Wolsey (1982) produces an $H(\delta)$-approximation solution for Submodular Set Covering problem, where δ is the maximum value of f over all singletons and $H(\delta)$ is the δ-th harmonic number. This captures many combinatorial optimization applications such as Minimum Set Cover, Minimum Hitting Set, Minimum Vertex Cover, Minimum Subset Interconnection Design and their corresponding weighted versions, to name a few; see e.g. (Du et al 2012; Feige 1996; Fujito 2000; Hongjie et al. 2011). When f is non-submodular, there exist several known results for real-world applications, including Network Steiner Tree (Du et al. 2008), Connected Dominating Set (Du et al 2012; Du et al. 2008), MOC-CDS problem in wireless sensor networks (Hongjie et al. 2011) and Connected Dominating Set with Labeling (Yang et al. 2020). In our prior work, defined some parameters (DR gap and DR ratio) to measure how far a set function is from being submodular, we propose a unified analysis framework for Minimum Non-submodular Cover problem (Shi et al. 2021).

However, there exist plenty of situations in which the cost function c is submodular and the potential function f is submodular or non-submodular. When c is submodular, inspired by Wan et al. (2010) established a similar result of $\rho H(\delta)$ for the minimum submodular cover problem with a submodular cost, where $\rho = \max_{S:\text{min}-\text{cost cover}} \frac{\sum_{e \in S} c(e)}{c(S)}$ is the curvature of the cost function; a general result on fraction-valued submodular cover was also presented, however, it depends on two additional hypothesis which make their method a little impractical for real-world problems. Iyer and Bilmes (2013) investigated two optimization problems — minimum submodular cover with a submodular cost (MIN-SCSC) and maximization of a submodular function with a submodular knapsack (MAX-SCSK) which were intimately connected and proved that any approximation algorithm for MAX-SCSK can be used to provide guarantees for MIN-SCSC. Nevertheless, as far as we know, there seems no study in the literature concerning minimum non-submodular cover problem with a submodular cost.

In this paper, we extend the potential function from submodular function to non-submodular one and first investigate Minimum Submodular Cost Non-submodular Cover problem (MIN-SCNSC). In addition, we also study MIN-SCSC with fraction-valued potential functions. Our main contributions are summarized as follows:

- We analyze the performance of the greedy algorithm for MIN-SCNSC with integer-valued and fraction-valued potential functions respectively, with a theoretical analysis (Theorem 1).

- We give a new analysis for MIN-SCSC with fraction-valued potential functions, expanding the scope of application of Wan et al. (2010) (Theorem 2).

The outline of the paper is as follows. In Sect. 2, we consider how MIN-SCSC and MIN-SCNSC occur in real-world problems and how they generalize a series of important optimization problems. In Sect. 3, we give the definitions of two parameters as well as two important lemmata. In Sect. 4, we propose a greedy algorithm for MIN-SCNSC, with a theoretical analysis for integer-valued and fraction-valued potential functions respectively. In Sect. 5, we show a new analysis for MIN-SCSC with fraction-valued potential functions. Sect. 6 contains the concluding remarks and future works.

2 Motivation

In this section, we first give two real-world problems which can be formulated as Minimum Submodular Cost Submodular Cover problem, and then introduce two other examples which can be cast as Minimum Submodular Cost Non-submodular Cover problem.

2.1 Social influence spread (Kempe et al. 2003)

In Independent Cascade (IC) model, let directed graph $G = (V, A, p)$ denote a social network, where V is the set of users and A is the set of social relations between users. Any $uv \in A$ is assigned with a probability such that when u is active, v is activated by u with probability p_{uv}. Consider a submodular cost function $c : 2^V \to \mathbb{R}_+$. Then the objective is to find a set $S \subseteq V$ with minimum cost such that users in S influences all users in G. Let $I(S)$ be the set of active nodes obtained from the seed set S at the end of the diffusion process and its cardinality be $|I(S)|$. Define expected influence spread function $\sigma(S) = \mathbb{E}[I(S)]$ as the potential function and it's easy to verity that σ is submodular (Kempe et al. 2003). Then the problem can be cast as Minimum Submodular Cost Submodular Cover problem.

2.2 Sensor placement (Krause and Guestrin 2005; Krause et al. 2008)

Sensor placement problem is to choose sensor locations A from a given set of possible locations V. Denote by $f(A) = I(X_A; X_{V\setminus A})$ as the mutual information between the chosen locations A and the locations $V\setminus A$ which are not selected. Note that while f is non-monotone, it can be shown to approximately monotone. Alternatively, we can also define the mutual information between a set of chosen sensors X_A and a quantity of interest C as $f(A) = I(X_A; C)$ assuming that X_A are conditionally independent given C. Both these functions are submodular (Krause and Guestrin 2005). In many real-world settings, the cost of placing a sensor depends on the specific location. If the cost involved is submodular, then we can model this problem as Minimum Submodular Cost Submodular Cover problem; that is, find a set of sensors with minimal cooperative cost such that the sensors cover all the possible locations. In fact, submodularity of cost function is reasonable because there is basically a discount when purchasing sensors in bulk.

2.3 Minimum weight connected dominating set (Du et al 2012; Du et al. 2008)

The objective of Minimum Weight Connected Dominating Set problem is to find a minimum weight connected dominating set (CDS) of a given connected graph. This problem often occurs in wireless communication, playing a crucial important role. Specifically, given an input connected graph $G = (V, E)$. Consider a subset $A \subseteq V$. Let $\tau(A)$ be the set of edges incident to A. Denote by $\#(G, \tau(A))$ the number of connected components in the graph $(G, \tau(A))$ and $\#G[A]$ the number of connected components of the induced subgraph $G[A]$. Then the potential function is defined by $f(A) = |V| - \#(G, \tau(A)) - \#G[A], \forall A \subseteq V$ and it's easy to verity that f is a normalized monotone non-submodular function (Du et al. 2008). And A is a CDS of G iff $f(A) = |V| - 2 = f(V)$. If the non-negative weight function w is submodular, then the problem can be phrased as Minimum Submodular Cost Non-submodular Cover problem.

2.4 Subset selection (Das and Kempe 2011)

The subset selection problem is to select a subset of random variables with minimum total cost from a large set, in order to obtain the best prediction of another variable of interest. The problem has been widely studied, especially in feature selection, sparse approximation, compressed sensing in the areas of machine learning and signal processing. Define the squared multiple correlation R^2 as the potential function f, then f is non-submodular (Das and Kempe 2011). If the cost function c is submodular, then we can formulate this real-world problem as Minimum Submodular Cost Non-submodular Cover problem.

We refer the readers to Du et al (2012) for more examples in the field of combinatorial optimization. In the context of machine learning and data mining, there exist a number of applications occurring naturally, including machine translation, video summarization, probabilistic inference, recommendation systems, etc.

3 Preliminaries

Define $[n] := \{1, 2, \cdots, n\}$ for a positive integer $n \geq 1$. Let $V = [n]$ be a ground set, a set function $f : 2^V \to \mathbb{R}$ is *submodular* if for every $S, T \subseteq V$, $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. For any $S, T \subseteq V$, we use $\Delta_T f(S) = f(S \cup T) - f(S)$ to denote the marginal gain when add set T to S. For $i \in V$,we use the shorthand $\Delta_i f(S)$ for $\Delta_{\{i\}} f(S)$. Equivalently, f is submodular if it has diminishing returns (DR) property: $\Delta_i f(S) \geq \Delta_i f(T)$ for $S \subseteq T \subseteq V \backslash \{i\}$. f is *monotone* (*non-decreasing*) if $f(S) \leq f(T)$ for any $S \subseteq T \subseteq V$ or $\Delta_i f(S) \geq 0$ for any $S \subseteq V, i \in V \backslash S$. f is *normalized* if $f(\varnothing) = 0$. Every set function f can be normalized by setting $g(S) = f(S) - f(\varnothing)$. A normalized monotone submodular set function is called a *polymatroid* function. Throughout the paper, we assume that f is given via a value oracle; that is, given a set $S \subseteq V$, the oracle returns the function value of $f(S)$.

In the following, we define the total curvature for submodular function and the DR ratio for general set function. And we also provide two lemmta which will be used later in the analysis part.

Definition 1 (total curvature, Conforti and Cornuéjols (1984)) Let $f : 2^V \to \mathbb{R}_+$ be a polymatroid function, the total curvature is defined as $\alpha = 1 - \min_{i\in V} \frac{\Delta_i f(V\backslash\{i\})}{f(\{i\})}$.

Definition 2 (DR ratio, Shi et al. (2021)) Let $f : 2^V \to \mathbb{R}_+$ be a monotone set function. The DR ratio of f is the largest scalar $\xi \in [0, 1]$ such that

$$\Delta_i f(S) \geq \xi \Delta_i f(T),\ \forall S \subseteq T \subseteq V\backslash\{i\}.$$

Remark 1 Note that $0 \leq \alpha \leq 1$. If $\alpha = 0$, then f is a modular function. If $\alpha = 1$, we say that f is fully curved. In this paper, we consider the submodular function with $\alpha \neq 1$. With regard to the DR ratio, f is submodular iff $\xi = 1$.

Lemma 3.1 *If $f : 2^V \to \mathbb{R}_+$ is a polymatroid function with the total curvature $\alpha \neq 1$, then for any $S \subseteq V$,*

$$f(S) \leq \sum_{i\in S} f(\{i\}) \quad \text{and} \quad \sum_{i\in S} f(\{i\}) \leq \frac{1}{1-\alpha} f(S).$$

Proof It holds obviously that $f(S) \leq \sum_{i\in S} f(\{i\})$ by the submodularity of f. Now we prove the second inequality. First, we claim that

$$f(S) \geq \sum_{i\in S} \Delta_i f(V\backslash\{i\}).$$

Let $S = \{i_1, \cdots, i_k\}$,

$$\begin{aligned} f(S) &= \Delta_{i_1} f(\emptyset) + \Delta_{i_2} f(\{i_1\}) + \cdots + \Delta_{i_k} f(\{i_1, \cdots, i_{k-1}\}) \\ &\geq \Delta_{i_1} f(V\backslash\{i_1\}) + \Delta_{i_2} f(V\backslash\{i_2\}) + \cdots + \Delta_{i_k} f(V\backslash\{i_k\}) \qquad \text{(submodularity of } f) \\ &= \sum_{i\in S} \Delta_i f(V\backslash\{i\}). \end{aligned}$$

By the definition of the total curvature of f, we have $\Delta_i f(V\backslash\{i\}) \geq (1-\alpha) f(\{i\})$. Therefore, $f(S) \geq (1-\alpha)\sum_{i\in S} f(\{i\})$. This completes the proof. □

Lemma 3.2 *Let $f : 2^V \to \mathbb{R}_+$ be a monotone set function with the DR ratio $\xi \neq 0$, then for any $S, T \subseteq V$,*

$$\Delta_S f(T) \leq \frac{1}{\xi} \sum_{i\in S} \Delta_i f(T).$$

Proof Let $S = \{i_1, \cdots, i_k\}$,

$$\begin{aligned}\Delta_S f(T) &= f(S \cup T) - f(T)\\ &= \Delta_{i_1} f(T) + \Delta_{i_2} f(T \cup \{i_1\}) + \cdots + \Delta_{i_k} f(T \cup \{i_1, \cdots, i_{k-1}\})\\ &\leq \Delta_{i_1} f(T) + \frac{1}{\xi}\Delta_{i_2} f(T) + \cdots + \frac{1}{\xi}\Delta_{i_k} f(T)\\ &\leq \frac{1}{\xi}\sum_{i \in S} \Delta_i f(T),\end{aligned}$$

where the first inequality follows from the definition of the DR ratio and the second one from $\xi \in (0, 1]$. This completes the proof. □

4 Minimum submodular cost non-submodular cover problem

In this section, we extend Minimum Submodular Cost Submodular Cover problem (MIN-SCSC) to Minimum Submodular Cost Non-submodular Cover problem (MIN-SCNSC) and propose a greedy algorithm for MIN-SCNSC, with an analysis of its theoretical guarantees.

The greedy algorithm is in fact the same as that for MIN-SCSC and we write here for the sake of completeness. It works as follows: starting with an empty set, at each step, an element with maximal value on $\Delta_x f(A)/c(x)$ is chosen and added to the current set A. Finally, if the marginal gain of any element is zero (or the potential function value of a set reaches the maximum value), the algorithm returns the greedy set. A more formal description is described in Algorithm 1. In the following, we indicate that Algorithm 1 is efficient and effective for solving MIN-SCNSC.

Algorithm 1 Greedy algorithm for MIN-SCNSC/MIN-SCSC

Input: A ground set V, a set function $f : 2^V \rightarrow \mathbb{R}_+$ and a polymatroid function $c : 2^V \rightarrow \mathbb{R}_+$.
Output: A greedy solution A_g.
1: $A \leftarrow \emptyset$.
2: **while** there exists $x \in V$ such that $\Delta_x f(A) > 0$ **do** % or $f(A) < f(V)$.
  select $x \in V$ that maximizes $\frac{\Delta_x f(A)}{c(x)}$;
  $A \leftarrow A \cup \{x\}$.
3: **end while**
4: **return** $A_g \leftarrow A$.

Lemma 4.1 *(Du et al (2012)) Let f be a monotone submodular function, then Ω_f can be rephrased as:*

$$\Omega_f = \{A \subseteq V | \Delta_x f(A) = 0, \forall x \in V\}.$$

This is an equivalent definition of Ω_f for polymatroid functions which explains why Algorithm 1 is reasonable and correct for solving MIN-SCSC. It means that

Ω_f contains the maximal sets A under f; that is, if $A \in \Omega_f$, then for any $C \subseteq V$, $f(C \cup A) = f(A)$. Denote by $\Lambda_f = \{A \subseteq V | \Delta_x f(A) = 0, \forall x \in V\}$. A natural question is what conditions a non-submodular function should satisfy, then $\Omega_f = \Lambda_f$ which enables Algorithm 1 efficient and effective for MIN-SCNSC. We give our discovery in the following lemma.

Lemma 4.2 *(Shi et al. (2021)) Let f be a monotone non-submodular function with $\xi \neq 0$, then*

$$\Omega_f = \Lambda_f,$$

where $\Omega_f = \{A \subseteq V | f(A) = f(V)\}$ and $\Lambda_f = \{A \subseteq V | \Delta_x f(A) = 0, \forall x \in V\}$.

Proof If $A \in \Omega_f$, then for any $x \in V$,

$$0 \leq \Delta_x f(A) = f(A \cup \{x\}) - f(A) \leq f(V) - f(A) = 0,$$

where the inequalities hold since f is monotone. Therefore, $\Delta_x f(A) = 0, \forall x \in V$, i.e., $A \in \Lambda_f$.

Conversely, if $A \in \Lambda_f$, then

$$0 \leq f(V) - f(A) = \Delta_{V \setminus A} f(A) \leq \frac{1}{\xi} \sum_{x \in V \setminus A} \Delta_x f(A) = 0,$$

where the first inequality follows from the monotonicity of f and the second one holds by Lemma 3.2. That is, $A \in \Omega_f$. This completes the proof. □

Lemma 4.2 indicates that Algorithm 1 can find a feasible solution for MIN-SCNSC with the potential functions of $\xi \neq 0$. And Algorithm 1 can obtain a competitive performance when it terminates, which is described in Theorem 1.

Theorem 1 *If f is an integer-valued normalized monotone non-submodular function with the DR ratio $\xi \neq 0$ and c is a polymatroid function with the total curvature $\alpha \neq 1$, then Algorithm 1 returns a solution whose objective function value never exceeds $\frac{1}{\xi}\frac{1}{1-\alpha}H(f(V))$ times the optimal value. If f is fraction-valued, then greedy value is at most $\frac{1}{\xi}\frac{1}{1-\alpha}(1 + \ln \frac{f(V)}{f(V)-f(A_{g-1})})$ times the optimal value.*

Proof Let $A_i = \{x_1, \cdots, x_i\}, i = 0, 1, \cdots, g$ be the successive sets returned by Algorithm 1 and $A_0 = \varnothing$. Let A^* be the optimal solution for MIN-SCNSC. For

$i \in [g]$, denote by $\theta_i = \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})}$.

$$
\begin{aligned}
\sum_{i=1}^{g} c(x_i) &= \sum_{i=1}^{g} \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})} \Delta_{x_i} f(A_{i-1}) \\
&= \sum_{i=1}^{g} \theta_i (f(A_i) - f(A_{i-1})) \\
&= \sum_{i=1}^{g} \theta_i [(f(V) - f(A_{i-1})) - (f(V) - f(A_i))].
\end{aligned}
$$

Let $l_i = f(V) - f(A_i)$ with $f(V) = l_0 \geq l_1 \geq \cdots \geq l_g = 0$, then we have

$$
\sum_{i=1}^{g} c(x_i) = \sum_{i=1}^{g} \theta_i (l_{i-1} - l_i) \leq \max_{i \in [g]} \{\theta_i l_{i-1}\} \sum_{i=1}^{g} (1 - \frac{l_i}{l_{i-1}}).
$$

For the quantity $\max_{i \in [g]}\{\theta_i l_{i-1}\}$, by the greedy rule, it's easy to see that $\theta_i = \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})} \leq \frac{c(y)}{\Delta_y f(A_{i-1})}$, $\forall y \in A^*$. Thus,

$$
\begin{aligned}
\max_{i \in [g]} \{\theta_i l_{i-1}\} &\leq \frac{c(y)}{\Delta_y f(A_{i-1})} (f(V) - f(A_{i-1})) \\
&= \frac{c(y)}{\Delta_y f(A_{i-1})} (f(A^*) - f(A_{i-1})) \\
&= \frac{c(y)}{\Delta_y f(A_{i-1})} (f(A^* \cup A_{i-1}) - f(A_{i-1})) \\
&\leq \frac{1}{\xi} \sum_{y \in A^*} \Delta_y f(A_{i-1}) \frac{c(y)}{\Delta_y f(A_{i-1})} = \frac{1}{\xi} \sum_{y \in A^*} c(y) \\
&\leq \frac{1}{\xi} \frac{1}{1-\alpha} c(A^*),
\end{aligned}
$$

where the second inequality follows from Lemma 3.2 and the last one from Lemma 3.1. Note that we do not consider whether $\Delta_y f(A_{i-1}) = 0 (i \in [g])$ or not since this quantity can be canceled out in line 4.

For the quantity $\sum_{i=1}^{g}(1-\frac{l_i}{l_{i-1}})$, if f is integral,

$$\begin{aligned}\sum_{i=1}^{g}(1-\frac{l_i}{l_{i-1}}) &= 1+\sum_{i=1}^{g-1}(1-\frac{l_i}{l_{i-1}})\\ &\le 1+\sum_{i=1}^{g-1}\sum_{j=l_i+1}^{l_{i-1}}\frac{1}{j}\\ &= 1+\sum_{i=1}^{g-1}(\sum_{j=1}^{l_{i-1}}\frac{1}{j}-\sum_{j=1}^{l_i}\frac{1}{j})\\ &= 1+\sum_{i=1}^{g-1}(H(l_{i-1})-H(l_i))\\ &= 1+H(l_0)-H(l_{g-1})\\ &\le H(l_0)=H(f(V)).\end{aligned}$$

Therefore,

$$c(A_g)\le\sum_{i=1}^{g}c(x_i)\le\frac{1}{\xi}\frac{1}{1-\alpha}H(f(V))c(A^*).$$

If f is fractional,

$$\begin{aligned}\sum_{i=1}^{g}(1-\frac{l_i}{l_{i-1}}) &= 1+\sum_{i=1}^{g-1}(1-\frac{l_i}{l_{i-1}})\\ &\le 1+\sum_{i=1}^{g-1}\int_{l_i}^{l_{i-1}}\frac{1}{s}ds\\ &= 1+\sum_{i=1}^{g-1}\ln\frac{l_{i-1}}{l_i}\\ &= 1+\ln\frac{l_0}{l_{g-1}}\\ &= 1+\ln\frac{f(V)-f(A_0)}{f(V)-f(A_{g-1})}=1+\ln\frac{f(V)}{f(V)-f(A_{g-1})}.\end{aligned}$$

Therefore,

$$c(A_g)\le\sum_{i=1}^{g}c(x_i)\le\frac{1}{\xi}\frac{1}{1-\alpha}\left(1+\ln\frac{f(V)}{f(V)-f(A_{g-1})}\right)c(A^*).$$

This completes the proof. □

5 New analysis for minimum submodular cost submodular cover problem with fraction-valued potential functions

In this section, we present the approximation guarantee of the greedy algorithm for MIN-SCSC with fraction-valued polymatroid functions, with a new analysis.

Let $A_i = \{x_1, \cdots, x_i\}, i = 0, 1, \cdots, g$ be the successive sets returned by Algorithm 1 and $A_0 = \varnothing$. Let A^* be the optimal solution for MIN-SCSC. For $i = 0, 1, \cdots, g$, denote by $\theta_i = \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})}$ and $\theta_0 = 0$. It' obvious to see that

$$0 = \theta_0 < \theta_1 \le \theta_2 \le \cdots \le \theta_g.$$

Because for any $i \in [g], \theta_i > 0$ as otherwise Algorithm 1 terminates, and for $i = 2, \cdots, g$,

$$\theta_{i-1} = \frac{c(x_{i-1})}{\Delta_{x_{i-1}} f(A_{i-2})} \le \frac{c(x_i)}{\Delta_{x_i} f(A_{i-2})} \le \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})} = \theta_i,$$

where the first inequality holds by the greedy rule and the second one follows from the submodularity of f.

Let $m \le g$ be the first index i such that $\Delta_y f(A_i) = 0, y \in A^*$; that is, $\Delta_y f(A_{i-1}) > 0, \forall i \in [m]$ and $\Delta_y f(A_m) = \Delta_y f(A_{m+1}) = \cdots = \Delta_y f(A_g) = 0$.

Theorem 2 *If f is a fraction-valued polymatroid function and c is a polymatroid function with the total curvature $\alpha \neq 1$, then Algorithm 1 returns a solution whose objective value is at most $\frac{1}{1-\alpha}(1 + \ln \lambda)$ times the optimal value, where $\lambda = \min\{\lambda_1, \lambda_2, \lambda_3\}$ is one of three possible problem parameters and*

$$\lambda_1 = \frac{\theta_g}{\theta_1}, \lambda_2 = \frac{f(\{y\})}{\Delta_y f(A_{m-1})}, \lambda_3 = \frac{f(V)}{f(V) - f(A_{g-1})}.$$

Before proving Theorem 2, we need the following two lemmata.

Lemma 5.1 $\sum_{i=1}^{g} c(x_i) \le \sum_{y \in A^*} \varphi(y)$, *where* $\varphi(y) = \sum_{i=1}^{g} \Delta_y f(A_{i-1})(\theta_i - \theta_{i-1})$.

Proof

$$\begin{aligned}\sum_{i=1}^{g} c(x_i) &= \sum_{i=1}^{g} \theta_i \Delta_{x_i} f(A_{i-1}) = \sum_{i=1}^{g} \theta_i (f(A_i) - f(A_{i-1})) \\ &= \sum_{i=1}^{g} \theta_i [(f(A^*) - f(A_{i-1})) - (f(A^*) - f(A_i))] \\ &= \sum_{i=1}^{g} \theta_i (f(A^*) - f(A_{i-1})) - \sum_{i=1}^{g} \theta_i (f(A^*) - f(A_i)).\end{aligned}$$

Since

$$\sum_{i=1}^{g}\theta_i(f(A^*)-f(A_i))=\sum_{i=1}^{g-1}\theta_i(f(A^*)-f(A_i))$$
$$=\sum_{i=2}^{g}\theta_{i-1}(f(A^*)-f(A_{i-1}))=\sum_{i=1}^{g}\theta_{i-1}(f(A^*)-f(A_{i-1})),$$

where the first equality follows from $f(A^*)=f(A_g)$ and the last one from $\theta_0=0$. We then have

$$\begin{aligned}\sum_{i=1}^{g}c(x_i)&=\sum_{i=1}^{g}\theta_i(f(A^*)-f(A_{i-1}))-\sum_{i=1}^{g}\theta_{i-1}(f(A^*)-f(A_{i-1}))\\&=\sum_{i=1}^{g}(\theta_i-\theta_{i-1})(f(A^*)-f(A_{i-1}))\\&=\sum_{i=1}^{g}(\theta_i-\theta_{i-1})(f(A^*\cup A_{i-1})-f(A_{i-1}))\\&\le\sum_{i=1}^{g}(\theta_i-\theta_{i-1})\sum_{y\in A^*}\Delta_y f(A_{i-1})\\&=\sum_{y\in A^*}\sum_{i=1}^{g}\Delta_y f(A_{i-1})(\theta_i-\theta_{i-1})\end{aligned}$$

where the inequality follows from the submodularity of f. Let

$$\varphi(y)=\sum_{i=1}^{g}\Delta_y f(A_{i-1})(\theta_i-\theta_{i-1}),$$

it means that $\sum_{i=1}^{g}c(x_i)\le\sum_{y\in A^*}\varphi(y)$. This completes the proof. □

Lemma 5.2 $\sum_{i=1}^{g}c(x_i)\le\sum_{y\in A^*}\psi(y)$, *where* $\psi(y)=\sum_{i=1}^{g}\theta_i(\Delta_y f(A_{i-1})-\Delta_y f(A_i))$.

Proof According to the proof of Lemma 5.1, we have

$$\begin{aligned}\sum_{i=1}^{g}c(x_i)&\le\sum_{y\in A^*}\sum_{i=1}^{g}\Delta_y f(A_{i-1})(\theta_i-\theta_{i-1})\\&=\sum_{y\in A^*}\left[\sum_{i=1}^{g}\theta_i\Delta_y f(A_{i-1})-\sum_{i=1}^{g}\theta_{i-1}\Delta_y f(A_{i-1})\right]\end{aligned}$$

Since

$$\sum_{i=1}^{g}\theta_{i-1}\Delta_y f(A_{i-1})=\sum_{i=2}^{g}\theta_{i-1}\Delta_y f(A_{i-1})=\sum_{i=1}^{g-1}\theta_i\Delta_y f(A_i)=\sum_{i=1}^{g}\theta_i\Delta_y f(A_i),$$

where the first equality follows from $\theta_0=0$ and the last one from $\Delta_y f(A_g)=0$. We then have

$$\begin{aligned}\sum_{i=1}^{g}c(x_i)&\leq\sum_{y\in A^*}[\sum_{i=1}^{g}\theta_i\Delta_y f(A_{i-1})-\sum_{i=1}^{g}\theta_i\Delta_y f(A_i)]\\&=\sum_{y\in A^*}\sum_{i=1}^{g}\theta_i(\Delta_y f(A_{i-1})-\Delta_y f(A_i)).\end{aligned}$$

Let

$$\psi(y)=\sum_{i=1}^{g}\theta_i(\Delta_y f(A_{i-1})-\Delta_y f(A_i)),$$

it means that $\sum_{i=1}^{g}c(x_i)\leq\sum_{y\in A^*}\psi(y)$. This completes the proof. □

We then give the proof of Theorem 2 in the following.

Proof of Theorem 2 **First**, according to Lemma 5.1, for any $y\in A^*$, we have

$$\varphi(y)=\sum_{i=1}^{g}\Delta_y f(A_{i-1})(\theta_i-\theta_{i-1}).$$

By the greedy rule,

$$\frac{\Delta_y f(A_{i-1})}{c(y)}\leq\frac{\Delta_{x_i} f(A_{i-1})}{c(x_i)}=\frac{1}{\theta_i},\forall i\in[g].$$

Thus,

$$\begin{aligned}\varphi(y)&\leq c(y)\sum_{i=1}^{g}\left(1-\frac{\theta_{i-1}}{\theta_i}\right)=c(y)\left[1+\sum_{i=2}^{g}\left(1-\frac{\theta_{i-1}}{\theta_i}\right)\right]\\&\leq c(y)(1+\sum_{i=2}^{g}\int_{\theta_{i-1}}^{\theta_i}\frac{1}{s}ds)=c(y)(1+\sum_{i=2}^{g}\ln\frac{\theta_i}{\theta_{i-1}})=c(y)(1+\ln\frac{\theta_g}{\theta_1}).\end{aligned}$$

Combining with Lemma 5.1, we have

$$\sum_{i=1}^{g}c(x_i)\leq\sum_{y\in A^*}\varphi(y)\leq(1+\ln\frac{\theta_g}{\theta_1})\sum_{y\in A^*}c(y).\tag{1}$$

Next, according to Lemma 5.2, for any $y \in A^*$, we have

$$\psi(y) = \sum_{i=1}^{g} \theta_i (\Delta_y f(A_{i-1}) - \Delta_y f(A_i)).$$

Because $m \leq g$ is the first index i such that $\Delta_y f(A_i) = 0$, then by the greedy rule,

$$\theta_i = \frac{c(x_i)}{\Delta_{x_i} f(A_{i-1})} \leq \frac{c(y)}{\Delta_y f(A_{i-1})}, \forall i \in [m].$$

Thus,

$$\begin{aligned}
\psi(y) &= \sum_{i=1}^{m} \theta_i (\Delta_y f(A_{i-1}) - \Delta_y f(A_i)) \\
&\leq c(y) \sum_{i=1}^{m} \left(1 - \frac{\Delta_y f(A_i)}{\Delta_y f(A_{i-1})}\right) \\
&= c(y) \left[1 + \sum_{i=1}^{m-1} \left(1 - \frac{\Delta_y f(A_i)}{\Delta_y f(A_{i-1})}\right)\right] \\
&\leq c(y)(1 + \sum_{i=1}^{m-1} \int_{\Delta_y f(A_i)}^{\Delta_y f(A_{i-1})} \frac{1}{s} ds) \\
&= c(y) \left(1 + \sum_{i=1}^{m-1} \ln \frac{\Delta_y f(A_{i-1})}{\Delta_y f(A_i)}\right) \\
&= c(y) \left(1 + \ln \frac{\Delta_y f(A_0)}{\Delta_y f(A_{m-1})}\right) = c(y) \left(1 + \ln \frac{f(\{y\})}{\Delta_y f(A_{m-1})}\right).
\end{aligned}$$

Combining with Lemma 5.2, we have

$$\sum_{i=1}^{g} c(x_i) \leq \sum_{y \in A^*} \psi(y) \leq (1 + \ln \frac{f(\{y\})}{\Delta_y f(A_{m-1})}) \sum_{y \in A^*} c(y). \tag{2}$$

Finally, analogous to the proof in Theorem 1, let $l_i = f(V) - f(A_i), i = 0, 1, \cdots, g$, we have

$$
\begin{aligned}
\sum_{i=1}^{g} c(x_i) &= \sum_{i=1}^{g} \theta_i (l_{i-1} - l_i) \\
&\leq \max_{i \in [g]} \{\theta_i l_{i-1}\} \sum_{i=1}^{g} (1 - \frac{l_i}{l_{i-1}}) \\
&\leq \max_{i \in [g]} \{\theta_i l_{i-1}\} (1 + \ln \frac{f(V)}{f(V) - f(A_{g-1})})
\end{aligned}
$$

It's easy to see that

$$
\begin{aligned}
\max_{i \in [g]} \{\theta_i l_{i-1}\} &\leq \frac{c(y)}{\Delta_y f(A_{i-1})} (f(V) - f(A_{i-1})) \\
&= \frac{c(y)}{\Delta_y f(A_{i-1})} (f(A^*) - f(A_{i-1})) \\
&\leq \sum_{y \in A^*} \Delta_y f(A_{i-1}) \frac{c(y)}{\Delta_y f(A_{i-1})} = \sum_{y \in A^*} c(y)
\end{aligned}
$$

Therefore,

$$
\sum_{i=1}^{g} c(x_i) \leq \left(1 + \ln \frac{f(V)}{f(V) - f(A_{g-1})}\right) \sum_{y \in A^*} c(y). \tag{3}
$$

Combining inequalities (1), (2) and (3), let $\lambda_1 = \frac{\theta_g}{\theta_1}, \lambda_2 = \frac{f(\{y\})}{\Delta_y f(A_{m-1})}, \lambda_3 = \frac{f(V)}{f(V) - f(A_{g-1})}$, and denote by $\lambda = \min\{\lambda_1, \lambda_2, \lambda_3\}$, we have

$$
c(A_g) \leq \sum_{i=1}^{g} c(x_i) \leq (1 + \ln \lambda) \sum_{y \in A^*} c(y) \leq \frac{1}{1 - \alpha} (1 + \ln \lambda) c(A^*),
$$

where the last inequality follows from Lemma 3.1. This completes the proof. □

Remark 2 Wan et al. (2010) presented a general result of $\rho H(\delta)$ on integer-valued submodular cover with a submodular cost; in fact, the curvature ρ they defined for the submodular cost function can be expressed as $\rho = \frac{1}{1-\alpha}$, which can be computed efficiently in linear time. They also gave an approximation performance of $1 + \rho \ln \frac{f(V)}{c(A^*)}$ for MIN-SCSC with fraction-valued potential functions with an assumption[1] which make it impractical for real-world problems, because one cannot guarantee

[1] In fact, one of assumptions $f(V) \geq opt$ is unnecessary, since we can get it by $\Delta_x f(A)/c(x) \geq 1$. That is, $f(V) = f(A_g) = \sum_{i=1}^{g} \Delta_{x_i} f(A_{i-1}) \geq \sum_{i=1}^{g} c(x_i) \geq c(A_g) \geq c(A^*) = opt$.

$\Delta_x f(A)/c(x) \geq 1$ in each step in the course of the greedy algorithm; however, we obtain a data-dependent approximation ratio without any additional hypothesis.

In the following, we show the approximation qualities for MIN-SCSC with fraction-valued potential functions between Wan et al. and ours. If Wan et al. do not inflict any hypothesis upon this problem, then they only have the inequality:

$$c(x_i) \leq \rho c(A^*) \frac{l_{i-1} - l_i}{l_{i-1}}, i \in [g],$$

where $l_i = f(V) - f(A_i)$. According to the proof of Theorem 2, their result is reduced to $\rho(1 + \ln \frac{f(V)}{f(V)-f(A_{g-1})})$ which is inferior to ours since we choose the minimum among three approximation ratios of $\rho(1 + \ln \frac{\theta_g}{\theta_1})$, $\rho(1 + \ln \frac{f(\{y\})}{\Delta_y f(A_{m-1})})$ and $\rho(1 + \ln \frac{f(V)}{f(V)-f(A_{g-1})})$, where $\rho = \frac{1}{1-\alpha}$. On the other hand, under the same hypothesis, we take two situations into consideration. The first is that the cost function c is linear (i.e. $\rho = 1$), then our result beats that of Wan et al. This is because

$$1 + \ln \frac{f(V)}{f(V) - f(A_{g-1})} \leq 1 + \ln \frac{f(V)}{c(A_g)} \leq 1 + \ln \frac{f(V)}{c(A^*)},$$

where the first inequality follows from the hypothesis of $\frac{\Delta_x f(A)}{c(x)} \geq 1$, thus, $f(V) - f(A_{g-1}) = f(A_g) - f(A_{g-1}) \geq c(A_g)$ and $c(A_g) \geq c(A^*)$ is based on the optimality of A^*. The other is that c is submodular, then the approximation performances between Wan et al. and ours depends on the real-world problems. Because $c(A_g) \geq c(A^*)$, then

$$\rho \ln \frac{f(V)}{f(V) - f(A_{g-1})} \leq \rho \ln \frac{f(V)}{c(A^*)};$$

however, $\rho > 1$. Thus, it is difficult to judge which one is better and it depends on the concrete problems.

6 Conclusions and future works

In this paper, we first study Minimum Submodular Cost Non-submodular Cover problem, with a theoretical analysis for integer-valued and fraction-valued potential functions respectively. In addition, we give a new analysis for Minimum Submodular Cost Submodular Cover problem with fraction-valued potential functions, improving upon the result in Wan et al. (2010). As a future work, it would be interesting to study some natural problems for which the DR ratio and the total curvature can be estimated directly, and hence the performance ratio of the greedy algorithm can be given explicitly which are no longer data-dependable.

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Competing interest There is no conflict of interest.

References

Conforti M, Cornuéjols G (1984) Submodular functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem. Discret Appl Math 7(3):251–274

Das A, Kempe D (2011) Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In: Proceedings of the 28th international conference on machine learning, pp 1057–1064

Du D, Ko K-I, Hu X (2012) Design and analysis of approximation algorithms. Springer, Berlin

Du D, Graham RL, Pardalos PM, Wan P, Wu W, Zhao W (2008) Analysis of greedy approximations with nonsubmodular potential functions. In: Proceedings of the 19th annual ACM-SIAM symposium on discrete algorithms, pp 167–175

Feige U (1996) A threshold of $\ln n$ for approximating set cover. In: Proceedings of the 28th annual ACM symposium on theory of computing, pp 314–318

Fujishige S (2005) Submodular functions and optimization. Elsevier, Amsterdam

Fujito T (2000) Approximation algorithms for submodular set cover with applications. IEICE Trans Inf Syst E83-D(3):488–495

Hongjie Du, Weili Wu, Lee Wonjun, Liu Qinghai, Zhang Zhao, Dingzhu Du (2011) On minimum submodular cover with submodular cost. J Glob Optim 50(2):229–234

Iyer R, Bilmes J (2013) Submodular optimization with submodular cover and submodular knapsack constraints. In: Proceedings of the 8th conference and workshop on neural information processing systems, pp 2436–2444

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, pp 137–146

Krause A, Singh A, Guestrin C (2008) Near-optimal sensor placements in gaussian processes: theory, efficient algorithms and empirical studies. J Mach Learn Res 9(3):235–284

Krause A, Guestrin C (2005) Near-optimal nonmyopic value of information in graphical models. In: Proceedings of the 21st conference on uncertainty in artificial intelligence, pp 324–331

Shi Majun, Yang Zishen, Wang Wei (2021) Minimum non-submodular cover problem with applications. Appl Math Comput 410(4):126442

Wan P, Du DP, Pardalos M, Weili W (2010) Greedy approximations for minimum submodular cover with submodular cost. Comput Optim Appl 45(2):463–474

Wolsey LA (1982) An analysis of the greedy algorithm for submodular set covering problem. Combinatorica 2(4):385–393

Yang Zishen, Shi Majun, Wang Wei (2020) Greedy approximation for the minimum connected dominating set with labeling. Optim Lett 15(2):685–700