Approximation Guarantees of Local Search Algorithms via Localizability of Set Functions

Kaito Fujii

Abstract

This paper proposes a new framework for providing approximation guarantees of local search algorithms. Local search is a basic algorithm design technique and is widely used for various combinatorial optimization problems. To analyze local search algorithms for set function maximization, we propose a new notion called localizability of set functions, which measures how effective local improvement is. Moreover, we provide approximation guarantees of standard local search algorithms under various combinatorial constraints in terms of localizability. The main application of our framework is sparse optimization, for which we show that restricted strong concavity and restricted smoothness of the objective function imply localizability, and further develop accelerated versions of local search algorithms. We conduct experiments in sparse regression and structure learning of graphical models to confirm the practical efficiency of the proposed local search algorithms.

1. Introduction

Local search is a widely used technique to design efficient algorithms for optimization problems. Roughly speaking, local search algorithms start with an initial solution and gradually increase the objective value by repeatedly moving the solution to a nearby point. While this approach leads to effective heuristics for many optimization problems in practice, it is not always easy to provide approximation guarantees on their performance.

In this paper, we propose a generic framework for providing approximation guarantees of local search algorithms for set function optimization. Set function optimization is a problem of finding an (approximately) optimal set from all feasible sets. Various machine learning tasks have been formulated as set function optimization problems, such as feature selection (Das & Kempe, 2011; Elenberg et al., 2018), summarization (Lin & Bilmes, 2011; Badanidiyuru et al., 2014), or active learning (Hoi et al., 2006; Golovin & Krause, 2011).

A promising approach to analyze local search algorithms for set function optimization is to utilize submodularity. Submodularity (Fujishige, 2005) is a property of set functions useful for designing efficient algorithms and has been extensively studied. Existing studies showed that for maximizing a submodular function subject to a certain constraint, local search procedures yield a constant-factor approximate solution to an optimal solution (Nemhauser et al., 1978; Fisher et al., 1978; Lee et al., 2010; Feldman et al., 2011). Local search algorithms have been applied to several machine learning tasks that have submodularity (Iyer et al., 2013; Balkanski et al., 2016), but there are many other practical set functions that deviate from submodularity.

To analyze local search algorithms for these problems, we propose a novel analysis framework that can be applied to local search algorithms for non-submodular functions. The key notion of our framework is a property of set functions, which we call localizability. Intuitively, localizability is a property that implies any local optimum is a good approximation to a global optimal solution. By utilizing this property, we show that for maximizing a set function with localizability under a certain constraint, simple local search algorithms achieve a good approximation.

The main application of our framework is sparse optimization. Sparse optimization is the problem of finding a sparse vector that optimizes a continuous objective function. It has various applications such as feature selection for sparse regression and structure learning of graphical models. An approach to sparse optimization is a reduction to a set function optimization problem, which is adopted by Jain et al. (2016) and Elenberg et al. (2017). We show that localizability of this set function is derived from restricted strong concavity and restricted smoothness of the original objective function, which implies approximation guarantees of local search algorithms. Furthermore, we devise accelerated
variants of our proposed local search algorithms by utilizing the structure of sparse optimization. An advantage of our approach over existing methods is its applicability to a broader class of combinatorial constraints.

**Our contribution.** In this paper, we propose a new property of set functions called localizability and provide a lower bound on the approximation ratio of local search algorithms for maximizing a set function with this property. Our contribution is summarized as follows.

- We define localizability of set functions and show that localizability of sparse optimization is derived from restricted strong concavity and restricted smoothness of the original objective function.

- Under the assumption of localizability, we provide lower bounds on the approximation ratio of a standard local search algorithm under a matroid constraint, p-matroid intersection constraint, or p-exchange system constraint.

- For sparse optimization, we propose two accelerated variants of local search algorithms, which we call semi-oblivious and non-oblivious local search algorithms.

- We conduct experiments on sparse regression and structure learning of graphical models to confirm the practical efficiency of our accelerated local search algorithms.

**1.1. Related Work**

**Local search for submodular maximization.** For monotone submodular maximization, several algorithms have been designed based on local search. Nemhauser et al. (1978) proposed a 1/2-approximation local search procedure for a cardinality constraint, which they call an interchange heuristic. Fisher et al. (1978) generalized this result to a single matroid constraint. For a p-matroid intersection constraint, Lee et al. (2010) proposed a (1/p - ε)-approximation local search algorithm. Feldman et al. (2011) proposed a novel class of constraints called p-exchange systems and devised a (1/p - ε)-approximation local search algorithm. Filmus & Ward (2014) devised a (1 - 1/e)-approximation local search algorithm for a matroid constraint. Also for non-monotone submodular maximization, constant-factor approximation local search algorithms have been devised (Feige et al., 2011; Lee et al., 2009). While local search algorithms for submodular maximization have been studied extensively, there are only a few results on those for non-submodular maximization.

**Approximation algorithms for non-submodular function maximization and sparse optimization.** For non-submodular function maximization, many existing studies have adopted an approach based on greedy algorithms. Das & Kempe (2011) analyzed greedy algorithms for sparse linear regression by introducing the notion of submodularity ratio. Elenberg et al. (2018) extended their results to sparse optimization problems with restricted strong concavity and restricted smoothness. Chen et al. (2018) showed that the random residual greedy algorithm achieves (γ/(1 + γ))^2-approximation for a set function maximization with submodularity ratio γ under a single matroid constraint. We compare our results with existing methods for sparse optimization in Table 1. Note that the results of Das & Kempe (2011) and Chen et al. (2018) hold for any monotone set function whose submodularity ratio is bounded, while we utilize a stronger property derived from restricted strong concavity and restricted smoothness. Bian et al. (2017) analyzed the greedy algorithm for maximizing a set function whose submodularity ratio and generalized curvature are both bounded. Sakae (2019) considered sparse optimization with a constraint expressed by a monotone set function with bounded superadditivity ratio and restricted inverse curvature, but their framework cannot deal with matroid constraints and p-exchange system constraints. Fujii & Soma (2018) developed a similar analysis to ours, but

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Local search</th>
<th>Greedy-based</th>
<th>Modular approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinality</td>
<td>m² / 2M²,2 exp⁡ (M₂,2 T)</td>
<td>1 - exp⁡ (−m₂,2 / Ms,1)</td>
<td>m₁ M₂ / M₁ M₂</td>
</tr>
<tr>
<td>Matroid</td>
<td>m² / 2M²,2 exp⁡ (−1/p)</td>
<td>1 / (1 + m²,2)</td>
<td>m₁ M₂ / M₁ M₂</td>
</tr>
<tr>
<td>p-Matroid intersection or p-Exchange systems</td>
<td>1 / (p−1+1/q)²m²,2 exp⁡ (−1/p+1/q)²Ms,2 T</td>
<td>N/A</td>
<td>1 / (p−1+1/q)²m₁ M₂ − ε</td>
</tr>
</tbody>
</table>

---

1They did not specify the subscripts of γ, but it is not larger than \(\min_{i=1, \ldots, s} \gamma_{i-1, s-i} \), where \(s\) is the rank of the matroid.
their focus lies in a different problem called dictionary selection. The Frank-Wolfe algorithm (Frank & Wolfe, 1956; Jaggi, 2013) is a continuous optimization method that is often applied to sparse optimization. A variant called the pairwise Frank-Wolfe algorithm (Lacoste-Julien & Jaggi, 2015) incorporates the technique of moving weight between two atoms at each iteration, which is similar to our local search procedures, but their guarantees are incomparable to ours.

Modular approximation for sparse optimization. For sparse optimization with structured constraints, there exists a trivial benchmark called modular approximation (Cevher & Krause, 2011). Modular approximation maximizes a linear function that approximates the original set function by ignoring all interactions between elements. We provide a detailed description and analysis of modular approximation in Appendix D.

Sparse recovery. Many existing studies on sparse optimization focus on sparse recovery guarantees, which cannot be compared directly with our guarantees on approximation ratios. In the context of compressed sensing, several algorithms similar to our non-oblivious local search algorithms have been developed (Needell & Tropp, 2010; Bahmani et al., 2013). Kyrillidis & Cevher (2012) and Baldassarre et al. (2016) developed frameworks that can be applied to a matroid constraint. For structure learning of graphical models, Jalali et al. (2011) provided sparse recovery guarantees for the forward-backward greedy algorithm by assuming restricted strong concavity and restricted smoothness. Recently, algorithms with recovery guarantees under weaker assumptions have been developed (Bresler, 2015; Klivans & Meka, 2017; Wu et al., 2019).

1.2. Organization

The rest of this paper is organized as follows. Section 2 specify the problem settings that we tackle in this paper. Section 3 introduces the notion of localizability and shows localization of sparse optimization. In Section 4, we propose local search algorithms for a matroid constraint, p-matroid intersection constraint, or p-exchange system constraint. In Section 5, we devise accelerated local search algorithms for sparse optimization. In Section 6, we describe applications of our problem settings: sparse regression and structure learning of graphical models. In Section 7, we empirically compare our proposed algorithms with existing methods. Due to space constraints, we defer all proofs to the appendix.

2. Problem Setting

In this section, we introduce the problem settings that we deal with in this paper.

Set function maximization. Let $N := [n]$ be the ground set and define a non-negative set function $f: 2^N \to \mathbb{R}_{\geq 0}$. Throughout the paper, we assume $f$ is monotone, i.e., $f(X) \leq f(Y)$ holds for any $X \subseteq Y \subseteq N$. We say $f$ is submodular when $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$ for any $S \subseteq T \subseteq N$ and $v \in N \setminus T$. Let $\mathcal{I} \subseteq 2^N$ be a set family that represents all feasible solutions. We assume $(N, \mathcal{I})$ is an independence system, that is, $\emptyset \in \mathcal{I}$ and $X \in \mathcal{I}$ for any $X \subseteq Y$ such that $Y \in \mathcal{I}$. A set function maximization problem can be written as

$$\text{Maximize } f(X) \text{ subject to } X \in \mathcal{I}. \quad (1)$$

In general, suppose we have access to a value oracle and independence oracle, which return the value of $f(X)$ and the Boolean value that represents whether $X \in \mathcal{I}$ or not for any input $X \in N$, respectively.

We consider three classes of independence systems: matroid constraints, p-matroid intersection, and p-exchange systems, which include structures that appear in applications. A standard setting of sparse optimization where $\mathcal{I} = \{X \subseteq N \mid |X| \leq s\}$ is a special case of matroid constraints.

Definition 1 (Matroids). An independence system $(N, \mathcal{I})$ is called a matroid if for any $S, T \in \mathcal{I}$ with $|S| < |T|$, there exists $v \in T \setminus S$ such that $S \cup \{v\} \in \mathcal{I}$.

Definition 2 (p-Matroid intersection). An independence system $(N, \mathcal{I})$ is a p-matroid intersection if there exist $p$ matroids $(N, \mathcal{I}_1), \ldots, (N, \mathcal{I}_p)$ such that $\mathcal{I} = \bigcap_{i=1}^p \mathcal{I}_i$.

Definition 3 (p-Exchange systems (Feldman et al., 2011)). An independence system $(N, \mathcal{I})$ is a p-exchange system if for any $S, T \in \mathcal{I}$, there exists a map $\phi: (T \setminus S) \to 2^{S \setminus T}$ such that (a) for any $v \in T \setminus S$, it holds that $|\phi(v)| \leq p$, (b) each $v \in S \setminus T$ appears in $(\phi(v))_{v \in T \setminus S}$ at most $p$ times, and (c) for any $X \subseteq T \setminus S$, it holds that $(S \setminus \bigcup_{v \in X} \phi(v)) \cup X \in \mathcal{I}$.

Sparse optimization. Sparse optimization is the problem of finding a sparse solution that maximizes a continuously differentiable function $u: \mathbb{R}^n \to \mathbb{R}$. Assume we have an access to a zeroth and first-order oracle that returns the value of $u(w)$ and gradient $\nabla u(w)$ given $w \in \mathbb{R}^n$. To define the approximation ratio properly, we need to assume $u(0) \geq 0$, but we can normalize any function $u': \mathbb{R}^n \to \mathbb{R}$ by setting $u'(w) := u'(w) - u'(0)$. Let $N = [n]$ be the set of all variables and $\mathcal{I} \subseteq 2^N$ a family of feasible supports. We can write a sparse optimization problem with structured constraints as

$$\text{Maximize } u(w) \text{ subject to } \text{supp}(w) \in \mathcal{I}, \quad (2)$$

where $\text{supp}(w)$ represents the set of non-zero elements of $w$, that is, $\text{supp}(w) = \{i \in N \mid w_i \neq 0\}$. We define $|w|_0 = |\text{supp}(w)|$. 

2. Problem Setting
We assume restricted strong concavity and restricted smoothness of the objective function $u$, which are defined as follows.

**Definition 4** (Restricted strong concavity and restricted smoothness (Negahban et al., 2012; Jain et al., 2014)). Let $\Omega$ be a subset of $\mathbb{R}^d \times \mathbb{R}^d$ and $u: \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. We say that $u$ is **restricted strongly concave** with parameter $m_\Omega$ and **restricted smooth** with parameter $M_\Omega$ on domain $\Omega$ if

$$-\frac{m_\Omega}{2}\|y - x\|^2 \geq u(y) - u(x) - \langle \nabla u(x), y - x \rangle \geq -\frac{M_\Omega}{2}\|y - x\|^2$$

for all $(x, y) \in \Omega$.

Let $\Omega_s = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid \|x\|_0 \leq s, \|y\|_0 \leq s, \|x - y\|_0 \leq s\}$ and $\Omega_{s,t} = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid \|x\|_0 \leq s, \|y\|_0 \leq s, \|x - y\|_0 \leq t\}$. Let $m_s$ be restricted strong concavity parameter on $\Omega_s$, and $M_{s,t}$ the restricted smoothness parameter on $\Omega_{s,t}$ for arbitrary positive integers $s, t \in \mathbb{Z}_{\geq 0}$. Due to the restricted strong concavity of $u$, $\arg \max_{\text{supp}(w) \subseteq X} u(w)$ is uniquely determined. We denote this maximizer by $w(X)$.

By introducing a set function $f: 2^N \to \mathbb{R}_{\geq 0}$ defined as

$$f(X) = \max_{\text{supp}(w) \subseteq X} u(w),$$

we can regard the sparse optimization problem as a set function optimization problem (1).

**Notations.** Vectors are denoted by bold lower-case letters (e.g. $x$ and $y$) and matrices are denoted by bold upper-case letters (e.g. $A$ and $B$). Sets are denoted by upper-case letters (e.g. $X$ and $Y$). For $X \subseteq N$ and $a \in N$, we define $X + a := X \cup \{a\}$ and $X - a := X \setminus \{a\}$. We define the symmetric difference by $X \Delta Y := (X \setminus Y) \cup (Y \setminus X)$.

### 3. Localizability of Set Functions

In this section, we define a property of set functions, which we call **localizability**. Since the general version of the definition of localizability is complicated, we first introduce a simplified definition as a warm-up, and then state the general definition.

Intuitively, localizability measures how much small modifications of a solution increase the objective value. Localizability is defined as a property that the sum of the increases yielded by small modifications is no less than the increase yielded by a large modification.

**Definition 5** (Localizability (simplified version)). Let $f: 2^N \to \mathbb{R}_{\geq 0}$ be a non-negative monotone set function. For some $\alpha, \beta \in \mathbb{R}_{\geq 0}$, we say $f$ is $(\alpha, \beta)$-**localizable with size $s$** if for arbitrary subsets $X, X^* \subseteq N$ of size $s$ and bijection $\phi: X \setminus X^* \to X^* \setminus X$, we have

$$\sum_{x \in X \setminus X^*} \{f(X - x + \phi(x)) - f(X)\} \geq \alpha f(X^*) - \beta f(X).$$

This property is sufficient to provide an approximation guarantee of local search algorithms for matroid constraints. However, we need a generalized version of localizability that considers exchanges of multiple elements to deal with more complicated constraints.

**Definition 6** (Localizability). Let $f: 2^N \to \mathbb{R}_{\geq 0}$ be a non-negative monotone set function. For some $\alpha, \beta_1, \beta_2 \in \mathbb{R}_{\geq 0}$, we say $f$ is $(\alpha, \beta_1, \beta_2)$-**localizable with size $s$ and exchange size $t$** if for arbitrary subsets $X, X^* \subseteq N$ of size $s$ and any collection $\mathcal{P}$ of subsets of $X \Delta X^*$ such that $|P| \leq t$ for each $P \in \mathcal{P}$, we have

$$\sum_{P \in \mathcal{P}} \{f(X \Delta P) - f(X)\} \geq \alpha k f(X^*) - (\beta_1 \ell + \beta_2) f(X),$$

where $k$ and $\ell$ are positive integers such that each element in $X^* \setminus X$ appears at least $k$ times in $\mathcal{P}$ and each element in $X \setminus X^*$ appears at most $\ell$ times in $\mathcal{P}$.

If we consider the case when $k = 1$ and $\ell = 1$, this definition coincides with the simplified version with $\beta = \beta_1 + \beta_2$. In existing studies on submodular maximization, Lee et al. (2010) and Feldman et al. (2011) utilized this property of linear functions and non-negative monotone submodular functions to prove the approximation bounds for local search algorithms.

**Proposition 7** (Proved in the proof of Lemma 3.1 of Lee et al. (2010)). Any linear function is $(1, 1, 0)$-localizable and any non-negative monotone submodular function is $(1, 1, 1)$-localizable with any size and any exchange size.

In the following proposition, we show that the set function derived from sparse optimization satisfies localizability under the restricted strong concavity and restricted smoothness assumption.

**Proposition 8.** Suppose $u: 2^N \to \mathbb{R}$ is a continuously differentiable function with $u(0) \geq 0$. Let $s, t \in \mathbb{Z}_{\geq 0}$ be arbitrary integers. Assume $u$ is restricted strong concave on $\Omega_{2s}$ and restricted smooth on $\Omega_{s,t}$. If $f: 2^N \to \mathbb{R}$ is a set function defined as $f(X) = \max_{\text{supp}(w) \subseteq X} u(w)$, then $f$ is $\left(\frac{m_{2s}}{M_{s,t}}, \frac{M_{s,t} - M_{s,t}}{m_{2s}}\right)$-localizable with size $s$ and exchange size $t$.

### 4. Local Search Algorithms

In this section, we describe our proposed local search algorithms for a matroid constraint, a $p$-matroid intersection constraint, and a $p$-exchange system constraint, and provide approximation ratio bounds in terms of localizability.
We can provide an approximation guarantee in terms of $O(1)$ while keeping the feasibility, which requires local search algorithms for a matroid constraint.

Algorithm 1: Local search algorithms for a matroid constraint
1: Let $X \leftarrow \emptyset$.
2: Add arbitrary elements to $X$ until $X$ is maximal in $I$.
3: for $i = 1, \cdots, T$ do
4: Find the pair of $x \in X$ and $x' \in N \setminus X$ such that $(x, x') \in \arg\max\{f(X - x + x') \mid X - x + x' \in I\}$
5: if $f(X - x + x') - f(X) > 0$ then
6: Update the solution $X \leftarrow X - x + x'$.
7: else
8: return $X$.
9: end if
10: end for
11: return $X$.

4.1. Algorithms for a Matroid Constraint

Here, we describe our proposed algorithm for a matroid constraint. The algorithm starts with an initial solution, which is any base of the given matroid. The main procedure of the algorithm is to repeatedly improve the solution by replacing an element in the solution with another element. At each iteration, the algorithm seeks a pair of an element $x \in X$ and another element $x' \in N \setminus X$ that maximizes $f(X - x + x')$ while keeping the feasibility, which requires $O(sn)$ oracle calls. The detailed description of the algorithm is given in Algorithm 1.

We can provide an approximation guarantee in terms of localizability of the objective function as follows.

Theorem 9. Suppose $I$ is the independence set family of a matroid and $s = \max\{|X| \mid X \in I\}$. Assume the objective function $f$ is non-negative, monotone, and $(\alpha, \beta_1, \beta_2)$-localizable with size $s$ and exchange size 2. If $X$ is the solution obtained by executing $T$ iterations of Algorithm 1 and $X^*$ is an optimal solution, then we have

$$f(X) \geq \frac{\alpha}{\beta_1 + \beta_2} \left(1 - \exp\left(-\frac{(\beta_1 + \beta_2)T}{s}\right)\right) f(X^*).$$

If $X$ is the output returned by Algorithm 1 when it stops by finding no pair to improve the solution, then we have

$$f(X) \geq \frac{\alpha}{\beta_1 + \beta_2} f(X^*).$$

4.2. Algorithms for $p$-Matroid Intersection and $p$-Exchange System Constraints

In this section, we consider two more general constraints, $p$-matroid intersection and $p$-exchange system constraints with $p \geq 2$. The proposed algorithms for these two constraints can be described as almost the same procedure by using the different definitions of $q$-reachability as follows.

Definition 10. Let $I \subseteq 2^N$ be a $p$-matroid intersection. A feasible solution $T \in I$ is $q$-reachable from $S \in I$ if $|T \setminus S| \leq q$ and $|S \setminus T| \leq pq - q + 1$.

Definition 11. $q$-Reachability for $p$-exchange systems (Feldman et al., 2011). Let $I \subseteq 2^N$ be a $p$-exchange system. A feasible solution $T \in I$ is $q$-reachable from $S \in I$ if $|T \setminus S| \leq q$ and $|S \setminus T| \leq pq - q + 1$.

We denote by $F_q(X)$ the set of all $q$-reachable sets from $X$ that is determined by each definition of $q$-reachability for $p$-matroid intersection or $p$-exchange systems.

First, we must decide parameter $q \in \mathbb{Z}_{\geq 1}$ that determines the neighborhood to be searched at each iteration. When we select larger $q$, we search larger solution space for improvement at each iteration; thus, we can obtain a better bound on its approximation ratio, while the number of oracle calls $n^{O(q)}$ becomes larger as well. The initial solution of the proposed algorithms is any feasible solution. Then the algorithms repeatedly replace the solution with a $q$-reachable solution that increases the objective value the most. The detailed description of this local search algorithm is given in Algorithm 2.

We can provide an approximation ratio bound under the assumption of localizability of the objective function as follows.

Theorem 12. Suppose $I$ is the independence set family of a $p$-matroid intersection or $p$-exchange system and $s = \max\{|X| \mid X \in I\}$. Let $t = 2p(q + 1)$ for the $p$-matroid intersection case and $t = pq + 1$ for the $p$-exchange system case. Assume the objective function $f$ is non-negative, monotone, and $(\alpha, \beta_1, \beta_2)$-localizable with size $s$ and exchange size $t$. If $X$ is the output obtained by executing $T$ iterations of Algorithm 2 with parameter $q$ and $X^*$ is an optimal solution, then the approximation ratio is lower-bounded by

$$\frac{\alpha \left(1 - \exp\left(\frac{\beta_1(p - 1 + 1/q) + \beta_2 T}{s}\right)\right)}{\beta_1(p - 1 + 1/q) + \beta_2}.$$
finding no better \( q \)-reachable solution, then we have
\[
f(X) \geq \frac{\alpha}{\beta_1(p - 1 + 1/q) + \beta_2} f(X^*).
\]

5. Acceleration for Sparse Optimization

We consider two accelerated variants of the proposed local search algorithms in the case of sparse optimization. To distinguish the original one from the accelerated variants, we call Algorithm 1 and Algorithm 2 oblivious local search algorithms.

5.1. Acceleration for a Matroid Constraint

The oblivious version computes the value of \( f(X - x + x') \) for \( O(sn) \) pairs of \((x, x')\) at each iteration. We can reduce the computational cost by utilizing the structure of sparse optimization.

The first variant is the semi-oblivious local search algorithm. For each element \( x' \in N \setminus X \) to be added, it computes the value of \( f(X - x + x') \) only for \( x \in X \) with the smallest \( (w(X))_2^2 \) among those satisfying \( X - x + x' \in \mathcal{I} \). Thus, we can reduce the number of times we compute the value of \( f(X - x + x') \) from \( O(sn) \) to \( O(n) \).

The second variant is the non-oblivious local search algorithm. It uses the value of
\[
\frac{1}{2M_{s,t}^2} \left( \nabla u(w(X)) \right)_x^2 - \frac{M_{s,2}}{2} \left( w(X) \right)_x^2
\]
in place of the increase of the objective function \( f(X - x + x') - f(X) \). We need to evaluate \( \nabla u(w(X)) \) and \( w(X) \) at the beginning of each iteration, but it is not necessary to compute the value of \( f(X - x + x') \).

The detailed description of these algorithms are given in Algorithm 3 in Appendix A.

**Theorem 13.** Suppose \( f(X) = \max_{\supp(w) \subseteq X} u(w) \) and \( \mathcal{I} \) is the independence set family of a matroid. If \( X \) is the solution obtained by executing \( T \) iterations of the semi-oblivious or non-oblivious local search algorithms and \( X^* \) is an optimal solution, then we have
\[
f(X) \geq \frac{m_{2s}^2}{M_{s,2}^2} \left( 1 - \exp \left( \frac{M_{s,2}T}{sm_{2s}} \right) \right) f(X^*),
\]
where \( s = \max\{|X| : X \in \mathcal{I}\} \). If \( X \) is the output returned when the algorithm stops by finding no pair to improve the solution, then we have
\[
f(X) \geq \frac{m_{2s}^2}{M_{s,2}^2} f(X^*).
\]

5.2. Acceleration for \( p \)-Matroid Intersection and \( p \)-Exchange System Constraints

Similarly to the case of matroid constraints, we can develop the semi-oblivious and non-oblivious local search algorithms for \( p \)-matroid intersection and \( p \)-exchange system constraints. The semi-oblivious variant only checks \( X' \in F_q(X) \) that minimizes \( \| (w(X))_X \|_2^2 \) among \( X'' \in F_q(X) \) such that \( X'' \setminus X = X' \setminus X \). The non-oblivious version selects the solution \( X' \in F_q(X) \) that maximizes
\[
\frac{1}{2M_{s,t}^2} \left\| \nabla u(w(X)) \right\|_{X \setminus X}^2 - \frac{M_{s,2}}{2} \left\| w(X) \right\|_{X \setminus X}^2.
\]

The detailed description of these algorithms are given in Algorithm 4 in Appendix A. While the oblivious local search algorithm requires \( O(n^p) \) times of the evaluation of \( f \) for finding the most suitable exchange at each iteration in general, the non-oblivious local search reduces it to a linear function maximization problem. In several cases such as a partition matroid constraint or a matching constraint, we can find the most suitable exchange in time polynomial in \( n \) and \( q \) by using standard techniques of combinatorial optimization. We can provide the same approximation guarantees for these accelerated variants as the oblivious variant.

**Theorem 14.** Suppose \( f(X) = \max_{\supp(w) \subseteq X} u(w) \) and \( \mathcal{I} \) is the independence set family of a \( p \)-matroid intersection or \( p \)-exchange system. Let \( t = 2(pq + 1) \) for the \( p \)-matroid intersection case and \( t = pq + 1 \) for the \( p \)-exchange system case. If \( X \) is the output obtained by executing \( T \) iterations of the semi-oblivious or non-oblivious local search algorithms with parameter \( q \) and \( X^* \) is an optimal solution, then its approximation ratio is lower-bounded by
\[
\frac{1}{p - 1 + 1/q} \left( 1 - \exp \left( \frac{(p - 1 + 1/q)M_{s,t}T}{sm_{2s}} \right) \right),
\]
where \( s = \max\{|X| : X \in \mathcal{I}\} \). If \( X \) is the output returned when the algorithm stops by finding no better \( q \)-reachable solution, then we have
\[
f(X) \geq \frac{1}{p - 1 + 1/q} \frac{m_{2s}^2}{M_{s,t}^2} f(X^*).
\]

**Remark 15.** We also develop another version of our local search algorithms that increases the objective value at a predetermined rate, which is described in Appendix C.

**Remark 16.** The parameter \( M_{s,t} \) used in the non-oblivious variant can be replaced with an upper bound on \( M_{s,t} \), which leads to the approximation ratio bounds whose \( M_{s,t} \) is also replaced with the upper bound.

6. Applications

In this section, we provide two applications of our framework: feature selection for sparse regression and structure
learning of graphical models.

6.1. Feature Selection for Sparse Regression

In sparse regression, given a design matrix $A \in \mathbb{R}^{n \times d}$ and a response vector $y$, we aim to find a sparse vector $w \in \mathbb{R}^n$ that optimizes some criterion. We can formulate this problem as a sparse optimization problem of maximizing $u(w)$ subject to $|\text{supp}(w)| \leq s$, where $u : \mathbb{R}^n \to \mathbb{R}$ is the criterion determined by $A$ and $y$. Das & Kempe (2011) devised approximation algorithms in the case where $u$ is the squared multiple correlation $R^2$, i.e., $u(w) := 1 - \|y - Aw\|^2_2/\|y\|^2_2$, and Elenberg et al. (2018) extended their results to general objectives with restricted strong concavity and restricted smoothness.

Here we consider sparse regression with structured constraints. In practical scenarios, we often have prior knowledge of relationships among features and can improve the quality of the estimation by incorporating it into structured constraints (Barnabik et al., 2010; Huang et al., 2009). We formulate sparse regression with structured constraints as the problem of maximizing $u(w)$ subject to $\text{supp}(w) \in \mathcal{I}$, where $\mathcal{I}$ is the set family of feasible supports.

An advantage of our local search framework is its applicability to a broad class of structured constraints, including matroid constraints. For example, the following constraint is a special case of matroid constraints. Suppose the set of all conditionally dependent pairs. To obtain a more interpretable graphical model, we often impose a sparsity constraint on the set of edges. This task can be formulated as a sparse optimization problem.

While most existing methods solve the neighborhood estimation problem separately for each vertex under a sparsity constraint (Jalali et al., 2011; Klivans & Meka, 2017), our framework provides an optimization method that handles the sparsity constraints for all vertices simultaneously. Suppose we aim to maximize some likelihood function (e.g., pseudo-log-likelihood (Besag, 1975)) under the sparsity constraint on each vertex, i.e., the degree of each vertex is at most $b$, where $b \in \mathbb{Z}_{\geq 0}$ is the maximum degree. This degree constraint is called a $b$-matching constraint, which is a special case of 2-exchange system. Hence, we can apply our local search algorithms to this problem.

7. Experiments

In this section, we conduct experiments on two applications: sparse regression and structure learning of graphical models. All the algorithms are implemented in Python 3.6. We conduct the experiments in a machine with Intel Xeon E3-1225 V2 (3.20 GHz and 4 cores) and 16 GB RAM.

7.1. Experiments on Sparse Regression

Datasets. We generate synthetic datasets with a partition matroid constraint. First, we determine the design matrix $A \in \mathbb{R}^{n \times d}$ by generating each of its entries according to the uniform distribution on $[0, 1]$. Then we normalize each of columns to ensure that the mean will be 0 and the standard deviation will be 1. Suppose the set of all features are partitioned into $n_c$ equal-size categories. We randomly select a sparse subset $S^*$ by selecting $n_p$ parameters from each category. The response vector is determined by $y = A_{S^*} w + \epsilon$, where $w$ is a random vector generated from the standard normal distribution $\mathcal{N}(0, 1)$ and $\epsilon$ is a noise vector whose each element is generated from $\mathcal{N}(0, 0.2)$. We consider two settings with different parameters. We set $(n, d, n_e, n_p) = (200, 50, 5, 5)$ in one setting and $(n, d, n_e, n_p) = (1000, 100, 10, 10)$ in the other setting. We use $R^2$ as the objective function to be maximized. For each parameter, we conduct 10 trials and plot the average.

Methods. We implement the oblivious, semi-oblivious, and non-oblivious local search algorithms. As benchmarks, we implement the random residual greedy algorithm (Chen et al., 2018) and modular approximation. We apply these methods to a partition matroid constraint with capacity $n_p'$ for each $n_p' \in \{1, \ldots, 10\}$.

Results. First, we compare the proposed methods in the case of $n = 200$ (Figure 1(a) and Figure 1(b)). We can observe that the non-oblivious variant is approximately 10 times faster than the oblivious variant, while achieving an objective value comparable to that of the oblivious variant. In comparison to the random residual greedy algorithm, the non-oblivious variant achieves a higher objective value in a similar running time. Modular approximation is consid-
The experimental results on sparse linear regression under a partition matroid constraint (1(a), 1(b), and 1(c)) and structure learning of graphical models under a $b$-matching constraint (1(d), 1(e), and 1(f)). 1(a) shows the running time in the case where $n = 100$. 1(b) and 1(c) show the objective value ($R^2$) in the case where $n = 200$ and $n = 1000$, respectively. 1(d) shows the running time in the case where $|V| = 10$. 1(e) and 1(f) show the ratio between the objective achieved by the algorithms and the optimal objective value in the case where $|V| = 10$ and $|V| = 20$, respectively.

7.2. Experiments on Structure Learning of Graphical Models

Datasets. We consider Ising models $G = (V, E)$ with parameter $(w_{uv})_{u, v \in V}$. First we generate the true graphical model randomly from the configuration model with degree $d$ for all vertices. For each edge $(u, v) \in E$, we set the parameter $w_{uv} = +0.5$ or $w_{uv} = -0.5$ uniformly at random. We synthetically generate 100 samples by Gibbs sampling from this Ising model. We consider two settings with different parameters. We set $(|V|, d) = (10, 5)$ in one setting and $(|V|, d) = (20, 7)$ in the other setting. We consider the problem of maximizing the pseudo-log-likelihood. For each setting, we conduct 10 trials and plot the average.

Methods. We implement the oblivious, semi-oblivious, and non-oblivious local search algorithms with parameter $q = 1$. Since calculating $M_{s,3}$ requires much computational cost, we use an upper bound $4 \sum_{i=1}^{N} \|x_i\|^3$ instead of $M_{s,3}$ in the non-oblivious variant. As a benchmark, we implement modular approximation, in which to maximize a linear function over a $b$-matching constraint, we use the reduction from a max-weight $b$-matching problem to a max-weight matching problem (Schrijver, 2003, Theorem 32.4) and the max-weight matching problem solver in NetwerkX library. We also implement random selection, which randomly samples a subgraph whose degree is $d$ at all vertices. In all methods, we use the L-BFGS-G solver in scipy.optimize library for evaluating the value of $f$. We apply these methods to pseudo-log-likelihood maximization under a $b$-matching constraint for each $b \in \{1, \cdots, d\}$.

Results. First, we compare the proposed methods in the case of $|V| = 10$ (Figure 1(d) and Figure 1(e)). We can observe that our acceleration techniques work well in practice. The running time of the non-oblivious variant is competitive with that of modular approximation and its solution quality is higher than that of modular approximation for larger solution size. Next, we conduct experiments on larger graphs with $|V| = 20$ (Figure 1(f)). Since the oblivious and semi-oblivious local search algorithms are too slow to be
applied to this setting, we omit them. Also, in this setting, we can observe that the non-oblivious variant outperforms the benchmarks particularly in cases of larger solution size.

Acknowledgements

The author would like to thank Andreas Krause for providing insightful comments in the early stages of this study. The author is thankful to Takeru Matsuda and Kazuki Matoya for inspiring discussions. This study was supported by JSPS KAKENHI Grant Number JP 18J12405.

References


A. Omitted pseudocodes

Algorithm 3 Local search algorithms for a matroid constraint
1: Let $X \leftarrow \emptyset$.
2: Add arbitrary elements to $X$ until $X$ is maximal in $\mathcal{I}$.
3: for $i = 1, \cdots, T$ do
4:   Determine the pair of $x \in X$ and $x' \in N \setminus X$ by the following rules:
5:     \[
      x' \in \text{argmax}\{ f(X - x + x') \mid X - x + x' \in \mathcal{I} \} \quad \text{(oblivious)}
     \]
6:     Let $x' \in \text{argmax}\{ f(X - \phi_X(x') + x') - f(X) \}$ and $x = \phi_X(x')$, where $\phi_X : N \setminus X \rightarrow X$ is a map that satisfies
7:     \[
      \phi_X(x') \in \argmin_{x \in X} : X - x + x' \in \mathcal{I} \left( \mathbf{w}^{(X)} \right)_{x}^{2} \quad \text{(semi-oblivious)}
     \]
8:     $\{ (x, x') \in \text{argmax}_{(x, x')} : X - x + x' \} \begin{cases} \frac{1}{2M_{x', 2}} \left( \nabla u(x') \right)_{x'}^{2} - \frac{M_{x', 2}}{2} \left( w(x') \right)_{x'}^{2} > 0 \quad \text{(non-oblivious)} \\ \end{cases}$
9:   if $f(X - x + x') - f(X) > 0$ (oblivious or semi-oblivious) then
10:     Update the solution $X \leftarrow X - x + x'$.
11:   else
12:     return $X$.
13: end if
14: end for
15: return $X$.

Algorithm 4 Local search algorithms for a $p$-matroid intersection or $p$-exchange system ($p \geq 2$)
1: Let $t = \begin{cases} 2p(q + 1) & \text{in the case of $p$-matroid intersection constraints} \\ pq + 1 & \text{in the case of $p$-exchange system constraints}. \end{cases}$
2: Let $X \leftarrow \emptyset$.
3: Add arbitrary elements to $X$ until $X$ is maximal in $\mathcal{I}$.
4: for $i = 1, \cdots, T$ do
5:   Determine $X'$ that is $q$-reachable from $X$ such that:
6:     \[
     X' \in \text{argmax}_{X' \in \mathcal{F}_{q}(X)} f(X') \quad \text{(oblivious)}
     \]
7:     Let $X' \in \text{argmax}_{X' \in \mathcal{F}_{q}(X)} : \exists S, X' = (X \cup S) \setminus \phi_{X}(S) f(X')$, where $\phi_{X} : 2^{N} \rightarrow 2^{N}$ is a map that satisfies
8:     \[
     \phi_{X}(S) \in \text{argmin}_{T : (X \cup S) \setminus \{ T \in \mathcal{F}_{q}(X) \}} \left( \nabla u(x') \right)_{T}^{2} \quad \text{(semi-oblivious)}
     \]
9:     $\{ X' \in \text{argmax}_{X' \in \mathcal{F}_{q}(X)} \begin{cases} \frac{1}{2M_{x', t}} \left( \nabla u(w(x')) \right)_{X' \setminus X}^{2} - \frac{M_{x', t}}{2} \left( w(x') \right)_{X' \setminus X}^{2} > 0 \quad \text{(non-oblivious)} \\ \end{cases}$
10:   if $f(X) - f(X') > 0$ (oblivious or semi-oblivious) then
11:     Update the solution $X \leftarrow X'$.
12:   else
13:     return $X$.
14: end if
15: end for
16: return $X$.

B. Omitted Proofs

B.1. Properties of Matroids, $p$-Matroid Intersection, and $p$-Exchange Systems

Here we provide important lemmas used in the proofs.
The following lemma is the exchange property of matroids.

**Lemma 17** (Corollary 39.12a in Schrijver (2003)). Let \(\mathcal{M} = (\mathcal{N}, \mathcal{I})\) be a matroid and \(I, J \in \mathcal{I}\) with \(|I| = |J|\). There exists a bijection \(\varphi : I \setminus J \to J \setminus I\) such that \(I - v + \varphi(v) \in \mathcal{I}\) for all \(v \in I \setminus J\).

The following lemma is on the exchange property of \(p\)-matroid intersection, which was first used for analyzing local search algorithms for submodular maximization.

**Lemma 18** (Lee et al., 2010). Suppose \(\mathcal{I}\) is a \(p\)-matroid intersection. Let \(q \in \mathbb{Z}\) be any positive integer. For any \(S, T \in \mathcal{I}\), there exists a multiset \(\mathcal{P} \subseteq 2^\mathcal{N}\) and an integer \(\eta\) (depending on \(p\) and \(q\)) that satisfies the following conditions.

1. For all \(P \in \mathcal{P}\), the symmetric difference is feasible, i.e., \(S \triangle P \in \mathcal{I}\), and \(S \triangle P\) is \(q\)-reachable (Definition 10) from \(S\).
2. Each element \(v \in T \setminus S\) appears in exactly \(\eta q\) sets in \(\mathcal{P}\).
3. Each element \(v \in S \setminus T\) appears in at most \((pq - q + 1)\eta\) sets in \(\mathcal{P}\).

A property similar to that for \(p\)-matroid intersection was known for \(p\)-exchange systems as follows.

**Lemma 19** (Feldman et al., 2011). The full proof can be found in Feldman (2013). Suppose \(\mathcal{I}\) is a \(p\)-exchange system. Let \(q \in \mathbb{Z}\) be any positive integer. For any \(S, T \in \mathcal{I}\), there exists a multiset \(\mathcal{P} \subseteq 2^\mathcal{N}\) and an integer \(\eta\) (depending on \(p\) and \(q\)) that satisfies the following conditions.

1. For all \(P \in \mathcal{P}\), the symmetric difference is feasible, i.e., \(S \triangle P \in \mathcal{I}\), and \(S \triangle P\) is \(q\)-reachable (Definition 11) from \(S\).
2. Each element \(v \in T \setminus S\) appears in at most \(\eta q\) sets in \(\mathcal{P}\).
3. Each element \(v \in S \setminus T\) appears in at most \((pq - q + 1)\eta\) sets in \(\mathcal{P}\).

### B.2. Proof of the Localizability of Sparse Optimization

To prove the localizability of sparse optimization, we use the following lemmas.

**Lemma 20.** Suppose \(u : 2^\mathcal{N} \to \mathbb{R}\) is a continuously differentiable function with \(u(\emptyset) \geq 0\). Assume \(u\) is restricted strong concave on \(\Omega_{2s}\) and restricted smooth on \(\Omega_s\). If \(f : 2^\mathcal{N} \to \mathbb{R}\) is a set function defined as \(f(X) = \max_{\supp(w) \subseteq X} u(w)\), then for any \(X, X' \subseteq \mathcal{N}\) with \(s = \max\{\|X\|, \|X'\|\}\) and \(t = |X \triangle X'|\), we have

\[
f(X') - f(X) \geq \frac{1}{2M_{s,t}} \left\| \nabla u(w(X)) \right\|_{X \setminus X'}^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{X \setminus X'} \right\|^2.
\]

**Proof.** From the restricted smoothness of \(u\), for any \(z \in \mathbb{R}^n\) with \(\supp(z) \subseteq X' \setminus X\), we have

\[
f(X') - f(X) = u(w(X')) - u(w(X))
\geq u((w(X))_{X \cap X'} + z) - u(w(X))
\geq \left\langle \nabla u(w(X)), z - (w(X))_{X \setminus X'} \right\rangle - \frac{M_{s,t}}{2} \left\| z - (w(X))_{X \setminus X'} \right\|^2.
\]

Since this inequality holds for every \(z\) with \(\supp(z) \subseteq X' \setminus X\), by optimizing it for \(z\), we obtain

\[
f(X') - f(X) \geq \frac{1}{2M_{s,t}} \left\| \nabla u(w(X)) \right\|_{X \setminus X'}^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{X \setminus X'} \right\|^2.
\]

**Lemma 21.** Suppose \(u : 2^\mathcal{N} \to \mathbb{R}\) is a continuously differentiable function with \(u(\emptyset) \geq 0\). Assume \(u\) is restricted strong concave on \(\Omega_{2s}\) and restricted smooth on \(\Omega_s\). If \(f : 2^\mathcal{N} \to \mathbb{R}\) is a set function defined as \(f(X) = \max_{\supp(w) \subseteq X} u(w)\), then for any \(X, X' \subseteq \mathcal{N}\) with \(s = \max\{\|X\|, \|X'\|\}\), we have

\[
f(X^*) - f(X) \leq \frac{1}{2m_{2s}} \left\| \nabla u(w(X)) \right\|_{X^* \setminus X}^2 - \frac{m_{2s}}{2} \left\| (w(X))_{X^* \setminus X} \right\|^2.
\]
Proof. From the restricted strong concavity of $u$, we obtain
\[
\begin{align*}
f(X^*) - f(X) &= u(w(X^*)) - u(w(X)) \\
&\leq \left\langle \nabla u(w(X^*)), w(X^*) - w(X) \right\rangle - \frac{m_{2s}}{2} \left\| w(X^*) - w(X) \right\|^2 \\
&\leq \max_{z : \text{supp}(z) \subseteq X^*} \left\{ \left\langle \nabla u(w(X)), z - w(X) \right\rangle - \frac{m_{2s}}{2} \left\| z - w(X) \right\|^2 \right\} \\
&= \frac{1}{2m_{2s}} \left\| \nabla u(w(X)) \right\|^2_{X^* \setminus X} - \frac{m_{2s}}{2} \left\| w(X) \right\|_{X \setminus X^*}^2.
\end{align*}
\]

Now we prove Proposition 8 from the above two lemmas.

Proof of Proposition 8. From Lemma 20, we have
\[
f(X \Delta P) - f(X) \geq \frac{1}{2M_{s,t}} \left\| \nabla u(w(X)) \right\|^2_{P \setminus X} - \frac{M_{s,t}}{2} \left\| w(X) \right\|^2_{P \cap X}
\]
for all $P \in \mathcal{P}$. By adding this inequality for each $P \in \mathcal{P}$, we obtain
\[
\sum_{P \in \mathcal{P}} \{f(X \Delta P) - f(X)\} \geq k \frac{1}{2M_{s,t}} \left\| \nabla u(w(X)) \right\|^2_{X^* \setminus X} - \ell \frac{M_{s,t}}{2} \left\| w(X) \right\|^2_{X \setminus X^*},
\]
where we used the fact that each element in $v \in X^* \setminus X$ appears in at least $k$ sets in $\mathcal{P}$ and each element in $X \setminus X^*$ appears in at most $\ell$ sets in $\mathcal{P}$. From the strong concavity of $u$, by applying Lemma 21, we obtain
\[
f(X^*) - f(X) \leq \frac{1}{2m_{2s}} \left\| \nabla u(w(X)) \right\|^2_{X^* \setminus X} - \frac{m_{2s}}{2} \left\| w(X) \right\|^2_{X \setminus X^*}
\]
and
\[
-f(X) \leq f(\emptyset) - f(X) \\
\leq -\frac{m_{2s}}{2} \left\| w(X) \right\|^2_{X^*} \\
\leq -\frac{m_{2s}}{2} \left\| w(X) \right\|^2_{X \setminus X^*}.
\]

By adding the former inequality multiplied by $km_{2s}/M_{s,t}$ and the latter inequality multiplied by $\ell M_{s,t}/m_{2s} - km_{2s}/M_{s,t}$, we obtain
\[
\sum_{P \in \mathcal{P}} \{f(X \Delta P) - f(X)\} \geq \frac{m_{2s}}{M_{s,t}} kf(X^*) - \frac{M_{s,t}}{m_{2s}} \ell f(X).
\]

B.3. Proof of Theorem 9

Proof of Theorem 9. Let $X$ be the output of the algorithm and $X^*$ an optimal solution. From Lemma 17, we have a bijection $\phi : X^* \setminus X \to X \setminus X^*$ such that $X - \phi(x^*) + x^* \in \mathcal{I}$ for all $x^* \in X^* \setminus X$.

Suppose at some iteration the solution is updated from $X$ to $X - x + x'$. Since $(x, x') \in \text{argmax}\{f(X - x + x') \mid X - x + x' \in \mathcal{I}\}$ and $f(X - x + x') - f(X) > 0$, we have
\[
f(X - x + x') - f(X) = \max_{(x, x') : X - x + x' \in \mathcal{I}} f(X - x + x') - f(X)
\]
Approximation Guarantees of Local Search Algorithms via Localizability of Set Functions

\[ \geq \frac{1}{s} \sum_{x^* \in X^* \setminus X} \{f(X - \phi(x^*) + x^*) - f(X)\}. \]

By setting \( P = \{\{x, \phi(x)\} \mid x \in X \setminus X^*\} \), each element in \( X^* \setminus X \) and \( X \setminus X^* \) appears exactly once in \( P \). Since \( f \) is \((\alpha, \beta_1, \beta_2)\)-localizable with size \( s \) and exchange size 2, we obtain

\[ \sum_{x^* \in X^* \setminus X} \{f(X - \phi(x^*) + x^*) - f(X)\} \geq \alpha f(X^*) - (\beta_1 + \beta_2)f(X). \]

By combining these inequalities, we have

\[ f(X - x + x') - f(X) \geq \frac{1}{s} \left\{ \alpha f(X^*) - (\beta_1 + \beta_2)f(X) \right\} \]
\[ = \frac{\beta_1 + \beta_2}{s} \left\{ \frac{\alpha}{\beta_1 + \beta_2} f(X^*) - f(X) \right\}, \]

which implies that the distance from the current solution to \( \alpha/(\beta_1 + \beta_2) \) times the optimal value is decreased by the rate \( 1 - (\beta_1 + \beta_2)/s \) at each iteration. Hence, the approximation ratio after \( T \) iterations can be bounded as

\[ f(X) \geq \frac{\alpha}{\beta_1 + \beta_2} \left( 1 - \left( 1 - \frac{\beta_1 + \beta_2}{s} \right)^T \right) f(X^*) \]
\[ \geq \frac{\alpha}{\beta_1 + \beta_2} \left( 1 - \exp \left( - \frac{(\beta_1 + \beta_2)T}{s} \right) \right) f(X^*), \]

which proves the first statement of the theorem.

Next, we consider the case where the algorithm stops by finding no pair to improve the objective value. When the oblivious variant stops, we have \( f(X) \geq f(X - x + x') \) for all \( x \in X \) and \( x' \in X \setminus X \) such that \( X - x + x' \in \mathcal{I} \). In the same manner as the above analysis, we obtain

\[ 0 \geq \sum_{x^* \in X^* \setminus X} \{f(X - \phi(x^*) + x^*) - f(X)\} \]
\[ \geq \alpha f(X^*) - (\beta_1 + \beta_2)f(X), \]

which implies

\[ f(X) \geq \frac{\alpha}{\beta_1 + \beta_2} f(X^*). \]

\[ \square \]

B.4. Proof of Theorem 12

Proof of Theorem 12. Let \( X \) be the output of the algorithm and \( X^* \) an optimal solution. From Lemma 18 and Lemma 19 for each case, respectively, we can observe that there exists a multiset \( P \subseteq 2^N \) and an integer \( \eta \) that satisfy the following conditions.

1. For all \( P \in \mathcal{P} \), the symmetric difference is \( q \)-reachable from \( X \), i.e., \( X \triangle P \in \mathcal{F}_q(X) \).
2. Each element \( v \in X^* \setminus X \) appears in exactly \( q\eta \) sets in \( \mathcal{P} \).
3. Each element \( v \in X \setminus X^* \) appears in at most \((pq - q + 1)\eta \) sets in \( \mathcal{P} \).

Suppose at some iteration the solution is updated from \( X \) to \( X' \). Since \( X' \in \arg\max_{X' \in \mathcal{F}_q(X)} f(X') \) and \( f(X') - f(X) > 0 \), we have

\[ f(X') - f(X) = \max_{X' \in \mathcal{F}_q(X)} f(X') - f(X) \]
\[ \geq \frac{1}{| \mathcal{P} |} \sum_{P \in \mathcal{P}} \{ f(X \triangle P) - f(X) \}. \]

Since \( f \) is \((\alpha, \beta_1, \beta_2)\)-localizable with size \( s \) and exchange size \( t \), we have
\[ \sum_{P \in \mathcal{P}} \{ f(X \triangle P) - f(X) \} \geq \alpha q f(X^*) - (\beta_1(pq - q + 1) + \beta_2 q) f(X). \]

By combining these inequalities, we have
\[ f(X') - f(X) \geq \frac{1}{| \mathcal{P} |} \{ \alpha q f(X^*) - (\beta_1(pq - q + 1) + \beta_2 q) f(X) \}. \]

Since each element in \( T \setminus S \) appears in \( q \eta \) sets in \( \mathcal{P} \) and \(| T \setminus S | \leq s \), it holds that \(| \mathcal{P} | \leq sq\eta \). Hence, we obtain
\[ f(X') - f(X) \geq \frac{1}{s} \{ \alpha f(X^*) - (\beta_1(p - 1 + 1/q) + \beta_2) f(X) \} \]
\[ = \frac{\beta_1(p - 1 + 1/q) + \beta_2}{s} \left\{ \frac{\alpha}{\beta_1(p - 1 + 1/q) + \beta_2} f(X^*) - f(X) \right\}, \]

which implies that the distance from the current solution to \( \frac{\alpha}{\beta_1(p - 1 + 1/q) + \beta_2} \) times the optimal value decreases by the rate \( 1 - \frac{\beta_1(p - 1 + 1/q) + \beta_2}{\beta_1(p - 1 + 1/q) + \beta_2} \) at each iteration. Therefore, the solution \( X \) after \( T \) iterations satisfies
\[ f(X) \geq \frac{\alpha}{\beta_1(p - 1 + 1/q) + \beta_2} \left( 1 - \exp \left( \frac{(\beta_1(p - 1 + 1/q) + \beta_2)T}{s} \right) \right) f(X^*). \]

Next, we consider the case where the algorithm stops by finding no pair to improve the objective value. In this case, we have \( f(X) \geq f(X') \) for all \( X' \in \mathcal{F}_q(X) \). In the same manner as the above analysis, we obtain
\[ 0 \geq \sum_{P \in \mathcal{P}} \{ f(X \triangle P) - f(X) \} \]
\[ \geq \frac{1}{s} \{ \alpha f(X^*) - (\beta_1(p - 1 + 1/q) + \beta_2) f(X) \}, \]

which implies
\[ f(X) \geq \frac{\alpha}{\beta_1(p - 1 + 1/q) + \beta_2} f(X^*). \]

\[ \square \]

**B.5. Proof of Theorem 13**

**Proof of Theorem 13.** Let \( X \) be the output of the algorithm and \( X^* \) an optimal solution. Suppose at some iteration the solution is updated from \( X \) to \( X - x + x' \). From Lemma 17, we have a bijection \( \phi : X^* \setminus X \rightarrow X \setminus X^* \) such that \( X - \phi(x^*) + x^* \in \mathcal{I} \) for all \( x^* \in X^* \setminus X \). Here we show that
\[ f(X - x + x') - f(X) \geq \frac{1}{n} \frac{M_{s,2}}{m_{2s}} \left\{ \frac{m_{2s}}{M_{s,2}^2} f(X^*) - f(X) \right\} \]
holds at each iteration of the semi-oblivious and non-oblivious variants.

When using the semi-oblivious variant, due to the property of the algorithm, we have
\[ (\mathbf{w}(X))^2 \geq (\mathbf{w}(X))^2 \]
for any \( \tilde{x} \in X \) such that \( X - \tilde{x} + x' \in \mathcal{I} \). If \( \phi_X : N \setminus X \rightarrow X \) is a map defined as \( \phi_X(x') \in \arg\min_{x \in X} \{(\mathbf{w}(X))^2 \mid X - x + x' \in \mathcal{I} \} \), then
\[ f(X - x + x') - f(X) \]
When using the non-oblivious variant, we have

\[= \max_{x' \in \mathbb{N}\setminus X} f(X - \phi_X(x') + x') - f(X)\]

\[\geq \frac{1}{s} \sum_{x'' \in \mathbb{N}\setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x''} - \frac{M_{s,2}}{2} \(w(X)\)^2_{\phi(x'')} \right\} \]

\[\geq \frac{1}{s} \sum_{x'' \in \mathbb{N}\setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x''} - \frac{M_{s,2}}{2} \(w(X)\)^2_{\phi(x'')} \right\} \quad \text{(From Lemma 20)}\]

\[\geq \frac{1}{s} \sum_{x'' \in \mathbb{N}\setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x''} - \frac{M_{s,2}}{2} \(w(X)\)^2_{\phi(x'')} \right\} \quad \text{(since } (w(X))^{2}_{\phi(x'')} \geq (w(X))^{2}_{\phi(x''')}\)}

\[= \frac{1}{s} \left\{ \frac{1}{2M_{s,2}} \left\| \nabla u(w(X)) \right\|_{X''}^2 - \frac{M_{s,2}}{2} \left\| w(X) \right\|_{X'\setminus X''}^2 \right\} \]

\[\geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,2}} f(X') - \frac{M_{s,2}}{m_{2s}} f(X) \right\} \quad \text{(From Lemma 21)}\]

where we used Lemma 20 and Lemma 21 as in the oblivious case.

When using the non-oblivious variant, we have

\[f(X - x + x') - f(X)\]

\[\geq \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x'} - \frac{M_{s,2}}{2} \(w(X)\)^2_{x'} \quad \text{(From Lemma 20)}\]

\[= \max_{(x', x'')} x' + x'' \in X \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x'} - \frac{M_{s,2}}{2} \(w(X)\)^2_{x'} \right\} \]

\[\geq \frac{1}{s} \sum_{x'' \in \mathbb{N}\setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)^2_{x''} - \frac{M_{s,2}}{2} \(w(X)\)^2_{\phi(x'')} \right\} \]

\[= \frac{1}{s} \left\{ \frac{1}{2M_{s,2}} \left\| \nabla u(w(X)) \right\|_{X''}^2 - \frac{M_{s,2}}{2} \left\| w(X) \right\|_{X'\setminus X''}^2 \right\} \]

\[\geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,2}} f(X') - \frac{M_{s,2}}{m_{2s}} f(X) \right\} \quad \text{(From Lemma 21)}\]

Therefore, in semi-oblivious and non-oblivious variants, we have

\[f(X - x + x') - f(X) \geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,2}} f(X') - \frac{M_{s,2}}{m_{2s}} f(X) \right\} \]

\[= \frac{1}{s} \frac{M_{s,2}}{m_{2s}} \left\{ \frac{m_{2s}}{M_{s,2}} f(X') - f(X) \right\} \]

which implies that the distance from the current solution to \(m_{2s}/M_{s,2}\) times the optimal value decreases by the rate \(1 - M_{s,2}/(sm_{2s})\) at each iteration. Hence, the approximation ratio after \(T\) iterations can be bounded as

\[f(X) \geq \frac{m_{2s}}{M_{s,2}} \left( 1 - \left( 1 - \frac{M_{s,2}}{sm_{2s}} \right)^T \right) f(X') \]

\[\geq \frac{m_{2s}}{M_{s,2}} \left( 1 - \exp \left( - \frac{M_{s,2}T}{sm_{2s}} \right) \right) f(X') \]

which proves the first statement of the theorem.

Next, we consider the case where the algorithm stops by finding no pair to improve the objective value. For the semi-oblivious and non-oblivious variants, we show that

\[0 \geq \frac{m_{2s}}{M_{s,2}} f(X') - \frac{M_{s,2}}{m_{2s}} f(X) \]
holds when the algorithm stops, from which the second statement of the theorem follows. When the semi-oblivious variant stops, we have \( f(X) \geq f(X - \phi_X(x') + x') \) for all \( x' \in N \setminus X \), where \( \phi_X(x') \) is defined in the algorithm. Hence, in the same manner as the above analysis, we obtain

\[
0 \geq \sum_{x^* \in X^* \setminus X} \{ f(X - \phi_X(x^*) + x^*) - f(X) \} \\
\geq \frac{m_{2s}}{M_{s,2}} f(X^*) - \frac{M_{s,2}}{m_{2s}} f(X).
\]

When the non-oblivious variant stops, we have

\[
0 \geq \frac{1}{2M_{s,2}} \left( \nabla u(w^{(X)}) \right)^2_{x'} - \frac{M_{s,2}}{2} \left( w^{(X)} \right)^2_{x'}
\]

for all \( x \in X \) and \( x' \in N \setminus X \) such that \( X - x + x' \in T \). Therefore, we have

\[
0 \geq \sum_{x^* \in X^* \setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w^{(X)}) \right)^2_{x'} - \frac{M_{s,2}}{2} \left( w^{(X)} \right)^2_{x'} \right\} \\
\geq \frac{m_{2s}}{M_{s,2}} f(X^*) - \frac{M_{s,2}}{m_{2s}} f(X)
\]

by using the above analysis for the first statement.

\[\square\]

B.6. Proof of Theorem 14

*Proof of Theorem 14.* Let \( X \) be the output of the algorithm and \( X^* \) an optimal solution. Suppose at some iteration the solution is updated from \( X \) to \( X' \). From Lemma 18 and Lemma 19 for each case, respectively, we can observe that there exist a multiset \( P \subseteq 2^N \) and an integer \( \eta \) that satisfy the following conditions.

1. For all \( P \in \mathcal{P} \), the symmetric difference is \( q \)-reachable from \( X \), i.e., \( X \triangle P \in \mathcal{F}_q(X) \).
2. Each element \( v \in X^* \setminus X \) appears in exactly \( \eta \) sets in \( \mathcal{P} \).
3. Each element \( v \in X \setminus X^* \) appears in at most \((pq - q + 1)\eta\) sets in \( \mathcal{P} \).

Here we show that

\[
f(X') - f(X) \geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,t}} f(X^*) - (p - 1 + 1/q) \frac{M_{s,t}}{m_{2s}} f(X) \right\}
\]

holds at each iteration of the semi-oblivious and non-oblivious variants.

When using the semi-oblivious variant, due to the property of the algorithm, we have

\[
\| (w^{(X)})_T \|^2 \geq \| (w^{(X)})_{X \setminus X'} \|^2
\]

for any \( T \subseteq X \) such that \( (X \cup X') \setminus T \in \mathcal{F}_q(X) \). If \( \phi_X : 2^N \rightarrow 2^N \) is a map defined as \( \phi_X(S) \in \text{argmin}_{T} : (X \cup S) \setminus T \in \mathcal{F}_q(X) \) \( \| (w^{(X)})_T \|^2 \), then

\[
f(X') - f(X) = \max_{X' \in \mathcal{F}_q(X)} : \exists S, X' = (X \cup S) \setminus \phi_X(S) f(X') - f(X) \\
\geq \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \{ f((X \cup P) \setminus \phi_X(P \setminus X)) - f(X) \} \\
\geq \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \left\{ \frac{1}{2M_{s,t}} \left\| \nabla u(w^{(X)}) \right\|_{P \setminus X}^2 - \frac{M_{s,t}}{2} \left\| w^{(X)} \right\|_{\phi_X(P \setminus X)}^2 \right\} \quad \text{(From Lemma 20)}
\]
When using the non-oblivious variant, we have

\[ \frac{1}{|P|} \sum_{p \in P} \left\{ \frac{1}{2M_{s,t}} \left\| (\nabla u(w(X)))_{p \setminus X} \right\|^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{p \setminus X} \right\|^2 \right\} \]

(since \( \left\| (w(X))_{p \setminus X} \right\|^2 \geq \left\| (w(X))_{\phi X(p \setminus X)} \right\|^2 \))

\[ \geq \frac{1}{|P|} \left\{ \frac{q \eta}{2M_{s,t}} \left\| (\nabla u(w(X)))_{X^* \setminus X} \right\|^2 - (pq - q + 1) \frac{M_{s,t}}{2} \left\| (w(X))_{X^* \setminus X} \right\|^2 \right\} \]

\[ \geq \frac{1}{|P|} \left\{ \frac{q \eta M_{s,t}}{2M_{s,t}} f(X^*) - (pq - q + 1) \eta \frac{M_{s,t}}{m_{2s}} f(X) \right\} \]

\[ \geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,t}} f(X^*) - (p - 1 + 1/q) \frac{M_{s,t}}{m_{2s}} f(X) \right\}, \]

where we used Lemma 20 and Lemma 21 as in the oblivious case.

When using the non-oblivious variant, we have

\[ f(X') - f(X) \]

\[ \geq \frac{1}{2M_{s,t}} \left\| (\nabla u(w(X)))_{X^* \setminus X} \right\|^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{X^* \setminus X} \right\|^2 \]

\[ = \max_{X' \in \mathcal{F}_{q}(X)} \left\{ \frac{1}{2M_{s,t}} \left\| (\nabla u(w(X)))_{X^* \setminus X} \right\|^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{X^* \setminus X} \right\|^2 \right\} \]

\[ \geq \frac{1}{|P|} \sum_{p \in P} \left\{ \frac{1}{2M_{s,t}} \left\| (\nabla u(w(X)))_{p \setminus X} \right\|^2 - \frac{M_{s,t}}{2} \left\| (w(X))_{p \setminus X} \right\|^2 \right\} \]

\[ \geq \frac{1}{|P|} \left\{ \frac{q \eta}{2M_{s,t}} \left\| (\nabla u(w(X)))_{X^* \setminus X} \right\|^2 - (pq - q + 1) \frac{M_{s,t}}{2} \left\| (w(X))_{X^* \setminus X} \right\|^2 \right\} \]

\[ \geq \frac{1}{s} \left\{ \frac{m_{2s}}{M_{s,t}} f(X^*) - (p - 1 + 1/q) \frac{M_{s,t}}{m_{2s}} f(X) \right\}, \]

where we used \( |P| \leq sq \eta \) in the last inequality. Therefore, in semi-oblivious and non-oblivious variants, we have

\[ f(X') - f(X) \geq (p - 1 + 1/q) \frac{M_{s,t}}{sm_{2s}} \left\{ \frac{1}{p - 1 + 1/q} \frac{m_{2s}}{M_{s,t}} f(X^*) - f(X) \right\}. \]

which implies that the distance from the current solution to \( \frac{1}{p - 1 + 1/q} \frac{m_{2s}}{M_{s,t}} \) times the optimal value decreases by the rate \( 1 - (p - 1 + 1/q)m_{2s}/(sM_{s,2}) \) at each iteration. Hence, the approximation ratio after \( T \) iterations can be bounded as

\[ f(X) \geq \frac{1}{p - 1 + 1/q} \frac{m_{2s}}{M_{s,t}^2} \left( 1 - \left( 1 - \frac{(p - 1 + 1/q)M_{s,t}}{sm_{2s}} \right)^T \right) f(X^*) \]

\[ \geq \frac{1}{p - 1 + 1/q} \frac{m_{2s}}{M_{s,t}^2} \left( 1 - \exp \left( - \frac{(p - 1 + 1/q)M_{s,t}T}{sm_{2s}} \right) \right) f(X^*), \]

which proves the first statement of the theorem.

Next, we consider the case where the algorithm stops by finding no pair to improve the objective value. For the semi-oblivious and non-oblivious variants, we show that

\[ 0 \geq q \eta \frac{m_{2s}}{M_{s,t}} f(X^*) - (pq - q + 1) \frac{M_{s,t}}{m_{2s}} f(X) \]
holds when the algorithm stops, from which the second statement of the theorem follows. When the semi-oblivious variant stops, we have \( f(X) \geq f(X - \phi_X(x')) \) for all \( x' \in N \setminus X \), where \( \phi_X(x') \) is defined in the algorithm. Hence, in the same manner as the above analysis, we obtain

\[
0 \geq \sum_{P \in P} \left\{ \frac{1}{2M_{s,t}} \left( \left( \nabla u(w(X)) \right)_{X \setminus X'} \right)^2 - \frac{M_{s,t}}{2} \right\} - \sum_{P \in P} \left\{ \frac{1}{2M_{s,t}} \left( \left( \nabla u(w(X)) \right)_{P \setminus X} \right)^2 - \frac{M_{s,t}}{2} \right\} \geq q \eta \frac{m_{2s}}{M_{s,t}} f(X^*) - \eta \frac{M_{s,t}}{m_{2s}} f(X).
\]

When the non-oblivious variant stops, we have

\[
0 \geq \sum_{P \in P} \left\{ \frac{1}{2M_{s,t}} \left( \left( \nabla u(w(X)) \right)_{X \setminus X} \right)^2 - \frac{M_{s,t}}{2} \right\} \geq q \eta \frac{m_{2s}}{M_{s,t}} f(X^*) - \eta \frac{M_{s,t}}{m_{2s}} f(X).
\]

by using the above analysis for the first statement.

\[\square\]

### C. Variants with Geometric Improvement

In this section, we analyze variants of our proposed local search algorithms for sparse optimization under a single matroid constraint. These variants increase the objective value by at least \( 1 + \epsilon \) times at each iteration, where \( \epsilon > 0 \) is a prescribed rate.

To analyze the singleton with the largest objective, which is used as an initial solution for the variants in this section, we use the following fact.

**Lemma 22.** Suppose \( u : 2^N \to \mathbb{R} \) is a continuously differentiable function with \( u(\emptyset) \geq 0 \). Assume \( u \) is restricted strong concave on \( \Omega_s \) and restricted smooth on \( \Omega_t \). Define \( f : 2^N \to \mathbb{R} \) by \( f(X) = \max_{\text{supp}(w) \subseteq X} u(w) \). Let \( x^* \in \arg\max \{ f(x) \mid x \in N \} \). We have \( f(\{ x^* \}) \geq \frac{2}{\lambda_{\max}} f(X) \) for any \( X \in I \), where \( s = \max \{|X| \mid X \in I \} \).

**Proof.** From the restricted smoothness of \( u \), we have

\[
u(c_x e_x) \geq u(\emptyset) + \langle \nabla u(\emptyset), c_x e_x \rangle - \frac{M_1}{2} c_x^2
\]

for any \( x \in N \) and \( c_x \in \mathbb{R} \). From Lemma 21, we have

\[
f(X) - f(\emptyset) \leq \frac{1}{2m_s} \left\| (\nabla u(\emptyset))_X \right\|^2
\]

for any \( X \in I \). By utilizing these inequalities, for any \( X \in I \), we have

\[
\max_{x \in N} f(\{ x \}) \geq f(\emptyset) + \frac{1}{s} \sum_{x \in X} f(x \setminus \emptyset)
\]

\[
= f(\emptyset) + \frac{1}{s} \sum_{x \in X} \max_{c_x \in \mathbb{R}} \{ u(c_x e_x) - u(\emptyset) \}
\]

\[
\geq f(\emptyset) + \frac{1}{s} \sum_{x \in X} \max_{c_x \in \mathbb{R}} \left\{ \langle \nabla u(\emptyset), c_x e_x \rangle - \frac{M_1}{2} c_x^2 \right\}
\]

\[
= f(\emptyset) + \frac{1}{s} \left( \left\| (\nabla u(\emptyset))_X \right\|^2 \right) \frac{1}{2M_1}
\]

holds.
We can provide bounds on the approximation ratio of these variants as follows.

**Algorithm 5** Local search algorithms for a matroid constraint with geometric improvement

1. Let $\delta \leftarrow \varepsilon/n$.
2. Let $X \leftarrow \text{argmax}\{f(v) \mid v \in N\}$.
3. Add arbitrary elements to $X$ until $X$ is maximal in $I$.
4. **loop**
   5. Search for a pair of $x \in X$ and $x' \in N \setminus X$ such that $X - x + x' \in I$ and
      
      \[
      f(X - x + x') \geq (1 + \delta)f(X)
      \]
      (oblivious)
      \[
      f(X - \phi_X(x') + x') \geq (1 + \delta)f(X) \text{ and } x = \phi_X(x'),
      \]
      where $\phi_X : N \setminus X \to X$ is a map that satisfies
      \[
      \phi_X(x') \in \arg\min_{x \in X : x-x'+x' \in I} \left(w^X\right)^2_{x'}
      \]
      (semi-oblivious)
      \[
      \frac{1}{2M_{s,2}} \left(\nabla u(w^X)\right)^2_{x'} - \frac{M_{s,2}}{2} \left(w^X\right)^2_x \geq \delta f(X)
      \]
      (non-oblivious)
   6. **if** $\exists(x, x')$ satisfying the above condition **then**
     7. Let $X \leftarrow X - x + x'$.
   8. **else**
   9. **return** $X$.
10. **end if**
11. **end loop**

\[
\geq f(\emptyset) + \frac{m_s}{sM_1}(f(X) - f(\emptyset))
\]
\[
\geq \frac{m_s}{sM_1}f(X),
\]
which concludes the statement. 

Here we introduce other variants of local search algorithms that use a different type of criteria for finding a pair $(x, x')$ to improve the solution. These new variants use any pair that increases some function by the rate $(1 + \delta)$, while the previously introduced variants find the pair that yields the largest improvement of some function. We consider three variants, the oblivious, semi-oblivious, and non-oblivious, similarly to the previous ones. The oblivious variant searches for any pair $(x, x')$ that increases the objective function by the rate $(1 + \delta)$, that is, $f(X - x + x') \geq (1 + \delta)f(X)$. The semi-oblivious variant constructs a map $\phi_X : N \setminus X \to X$ that satisfies $\phi_X(x') = \arg\min_{x \in X : x-x'+x' \in I} \left(w^X\right)^2_{x'}$ and searches for $x' \in N \setminus X$ with $f(x - \phi_X(x') + x') \geq (1 + \delta)f(X)$. The non-oblivious variant searches for any $(x, x')$ that satisfies
\[
\frac{1}{2M_{s,2}} \left(\nabla u(w^X)\right)^2_{x'} - \frac{M_{s,2}}{2} \left(w^X\right)^2_x \geq \delta f(X).
\]

All variants stop when they do not find any solution that satisfies the criteria. The detailed description of these algorithms is given in Algorithm 5.

We can provide bounds on the approximation ratio of these variants as follows.

**Theorem 23.** Suppose $f(X) = \max_{u \in \text{supp}(w) \subseteq X} u(w) \text{ and } I \text{ is the independence set family of a matroid}. \text{ Then any of the oblivious, semi-oblivious, and non-oblivious variants of Algorithm 5 stops after at most } O\left(\frac{n}{\varepsilon} \ln\left(M_{s^2}\right)\right) \text{ iterations and returns an output } X \text{ that satisfies}

\[
f(X) \geq \left(\frac{m^2_s}{M_{s,2}^2} - \varepsilon\right) f(X^*),
\]

where $X^*$ is an optimal solution and $s = \max\{|X| : X \in I\}$ is the rank of the matroid.

**Proof.** Let $X$ be the output of the algorithm. Let $X^*$ be an optimal solution. From Lemma 17, we have a bijection $\phi : X^* \setminus X \to X \setminus X^*$ such that $X - \phi(x^*) + x^* \in I$ for all $x^* \in X^* \setminus X$. For each of three variants, we prove
\[
0 \geq \sum_{x^* \in X^* \setminus X} \left\{ \frac{1}{2M_{s,2}} \left(\nabla u(w^X)\right)^2_{x^*} - \frac{M_{s,2}}{2} \left(w^X\right)^2_{\phi(x^*)} - \delta f(X) \right\},
\]
\[
\frac{1}{2M_{s,2}} \left(\nabla u(w^X)\right)^2_{x^*} - \frac{M_{s,2}}{2} \left(w^X\right)^2_{\phi(x^*)} \geq \frac{m^2_s}{M_{s,2}^2} f(X^*) - \varepsilon f(X),
\]
which concludes the proof.
when we use the non-oblivious variant, since we have
\[ \frac{m_{2s}}{M_{s,2}} f(X) - \left( \frac{M_{s,2}}{m_{2s}} + \delta n \right) f(X), \]
which implies
\[
0 \geq \frac{1}{2M_{s,2}} \left\| \left( \nabla u(w(X)) \right)_{X^* \setminus X} \right\|^2 - \frac{M_{s,2}}{2} \left\| \left( w(X) \right)_{X^* \setminus X} \right\|^2 - \delta n f(X)
\]
\[
\geq \frac{m_{2s}}{M_{s,2}} f(X^*) - \left( \frac{M_{s,2}}{m_{2s}} + \delta n \right) f(X),
\]
where the second inequality is due to Lemma 21. Since we set \( \delta = \epsilon/n \), we obtain
\[
f(X) \geq \left( \frac{m_{2s}^2}{M_{s,2}^2} - \epsilon \right) f(X).
\]
In the case of the oblivious variant, since \( f(X - x + x') \leq (1 + \delta) f(X) \) for all \( x \in X \) and \( x' \in N \setminus X \), we have
\[
0 \geq \sum_{x^* \in X^* \setminus X} \{ f(X - \phi(x^*) + x^*) - (1 + \delta) f(X) \}
\]
\[
\geq \sum_{x^* \in X^* \setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)_{x^*}^2 - \frac{M_{s,2}}{2} \left( w(X) \right)_{x^*}^2 - \delta f(X) \right\}
\]
similarly to the proof of Theorem 9. In the case of the semi-oblivious variant, since \( f(X - \phi_X(x^*) + x') \) for any \( x' \in N \setminus X \), we have
\[
0 \geq \sum_{x^* \in X^* \setminus X} \{ f(X - \phi_X(x^*) + x^*) - (1 + \delta) f(X) \}
\]
\[
\geq \sum_{x^* \in X^* \setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)_{x^*}^2 - \frac{M_{s,2}}{2} \left( w(X) \right)_{x^*}^2 - \delta f(X) \right\}
\]
When we use the non-oblivious variant, since
\[
0 \geq \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)_{x^*}^2 - \frac{M_{s,2}}{2} \left( w(X) \right)_{x^*}^2 - \delta f(X)
\]
for all \( x^* \in X^* \setminus X \), we obtain
\[
0 \geq \sum_{x^* \in X^* \setminus X} \left\{ \frac{1}{2M_{s,2}} \left( \nabla u(w(X)) \right)_{x^*}^2 - \frac{M_{s,2}}{2} \left( w(X) \right)_{x^*}^2 - \delta f(X) \right\}
\]
Finally, we bound the number of iterations. At each iteration, the objective value is improved at least at a rate of \( (1 + \delta) \). From Lemma 22, the initial solution is \( \frac{m_{2s}}{sM_1} \)-approximation. Therefore, the number of iterations is at most
\[
\log_{1+\delta} \left( \frac{m_{2s}}{sM_1} \right) = O \left( \frac{n}{\epsilon} \ln \left( \frac{sM_1}{m_{2s}} \right) \right).
\]
To obtain the same bound by using Theorem 13 for Algorithm 1, the number of iterations \( T \) is required to be larger than
\[
T = \frac{sM_{s,2}}{m_{2s}} \log \left( \frac{m_{2s}^2}{\epsilon M_{s,2}^2} \right),
\]
which can be larger than Algorithm 5 in some cases and smaller in other cases.

**Remark 24.** In the same manner, we can devise local search algorithms with geometric improvement for \( p \)-matroid intersection and \( p \)-exchange system constraints. Since they are a straightforward combination of techniques in Theorem 12 and this section, we omit the description.
Algorithm 6 Modular approximation

1. Apply the $\alpha$-approximation algorithm to

\[
\hat{f}(X) = f(\emptyset) + \sum_{x \in X} f(x|\emptyset)
\]

subject to $X \in \mathcal{I}$ and obtain the output $X$.

2. return $X$.

D. Modular Approximation for Sparse Optimization

Modular approximation is a generic method for nonlinear optimization. This method maximizes a linear function that approximates the original objective function instead of maximizing the original function. If we can exactly or approximately solve linear function optimization under the sparsity constraint, we can bound the approximation ratio by the restricted strong concavity and restricted smoothness constants. While we can provide approximation ratio bounds for modular approximation, the empirical performance of modular approximation is mostly poor since it completely ignores correlations between variables in the objective function.

**Proposition 25.** Suppose $f(X) = \max_{\text{supp}(w) \subseteq X} u(w)$. Assume we use an $\alpha$-approximation algorithm for maximizing a linear function under constraint $\mathcal{I}$ as a subroutine. Modular approximation is $\alpha m_1 m_s M_1$-approximation for sparse optimization with constraint $X \in \mathcal{I}$, where $s = \max\{|X|: X \in \mathcal{I}\}$.

**Proof.** We consider a set function $\tilde{f}: 2^N \to \mathbb{R}$ defined by

\[
\tilde{f}(X) = f(\emptyset) + \sum_{x \in X} f(x|\emptyset)
\]

for each $X \subseteq N$ to be a modular approximation of $f$. From the restricted strong concavity and restricted smoothness of $u$, we have

\[
\langle \nabla u(0), c_i e_i \rangle - \frac{M_1}{2} c_i^2 \leq u(c_i e_i) - u(0) \leq \langle \nabla u(0), c_i e_i \rangle - \frac{m_1}{2} c_i^2
\]

(3)

for any $i \in N$ and $c_i \in \mathbb{R}$. From Lemma 20 and Lemma 21, we have

\[
\frac{1}{2M_s} \|\nabla u(0)\|_X^2 \leq f(X) - f(\emptyset) \leq \frac{1}{2m_s} \|\nabla u(0)\|_X^2
\]

(4)

for any $X \in \mathcal{I}$. By using (3) and (4), for any $X \in \mathcal{I}$, we obtain

\[
\hat{f}(X) = u(0) + \sum_{i \in X} \max_{c_i \in \mathbb{R}} \{u(c_i e_i) - u(0)\}
\]

\[
\leq u(0) + \sum_{i \in X} \max_{c_i \in \mathbb{R}} \left\{\langle \nabla u(0), c_i e_i \rangle - \frac{m_1}{2} c_i^2\right\}
\]

(3)

\[
= u(0) + \sum_{i \in X} \frac{(\nabla u(0))^2}{2m_1}
\]

\[
\leq \frac{M_s}{m_1} f(X),
\]

(4)

where we use $u(0) \geq 0$ for the last inequality. Similarly, for any $X \in \mathcal{I}$, we obtain

\[
\hat{f}(X) = u(0) + \sum_{i \in X} \max_{c_i \in \mathbb{R}} \{u(c_i e_i) - u(0)\}
\]

\[
\geq u(0) + \sum_{i \in X} \max_{c_i \in \mathbb{R}} \left\{\langle \nabla u(0), c_i e_i \rangle - \frac{M_1}{2} c_i^2\right\}
\]

(3)
\[
= u(0) + \sum_{i \in X} \left(\frac{(\nabla u(0))^2_i}{2M_1}\right)
\geq \frac{m_s}{M_1}f(X), \quad \text{(from (4))}
\]

where we use \(u(0) \geq 0\) for the last inequality. Let \(X_{\text{MA}}\) be the output of the \(\alpha\)-approximation algorithm applied to maximizing \(\tilde{f}(X)\) subject to \(X \in \mathcal{I}\). Then we have

\[
\tilde{f}(X_{\text{MA}}) \geq \alpha \tilde{f}(X^*),
\]

where \(X^* \in \text{argmax}_{X \in \mathcal{I}} f(X)\). Finally, we have

\[
f(X_{\text{MA}}) \geq \frac{m_1}{M_s} \tilde{f}(X_{\text{MA}}) \geq \alpha \frac{m_1}{M_s} \tilde{f}(X^*) \geq \alpha \frac{m_1 m_s}{M_1 M_s} f(X^*).
\]

Since there exists an exact greedy algorithm for maximizing a linear function over a matroid constraint and \((1/(p - 1 + 1/q) - \epsilon)\)-approximation local search algorithms for a \(p\)-matroid intersection constraint (Lee et al., 2010) or \(p\)-exchange system constraint (Feldman et al., 2011), we obtain the following approximation ratio bounds.

**Corollary 26.** Modular approximation with the greedy algorithm is \(\frac{m_1 m_s}{M_1 M_s}\)-approximation for a matroid constraint.

**Corollary 27.** Modular approximation with local search algorithms is \((\frac{1}{p - 1 + 1/q} \frac{m_1 m_s}{M_1 M_s} - \epsilon)\)-approximation for a \(p\)-matroid intersection or \(p\)-exchange system constraint.