Maximizing Monotone DR-submodular Continuous Functions by Derivative-free Optimization

Yibo Zhang\(^1\) Chao Qian\(^1\) Ke Tang\(^2\)
\(^1\)Anhui Province Key Lab of Big Data Analysis and Application, USTC, China
\(^2\)Shenzhen Key Lab of Computational Intelligence, SUSTech, China
zyb233@mail.ustc.edu.cn chaoqian@ustc.edu.cn tangk3@sustc.edu.cn

Abstract

In this paper, we study the problem of monotone (weakly) DR-submodular continuous maximization. While previous methods require the gradient information of the objective function, we propose a derivative-free algorithm LDGM for the first time. We define \(\beta\) and \(\alpha\) to characterize how close a function is to continuous DR-submodular and submodular, respectively. Under a convex polytope constraint, we prove that LDGM can achieve a \((1 - e^{-\beta} - \epsilon)\)-approximation guarantee after \(O(1/\epsilon)\) iterations, which is the same as the best previous gradient-based algorithm. Moreover, in some special cases, a variant of LDGM can achieve a \(((\alpha/2)(1 - e^{-\alpha}) - \epsilon)\)-approximation guarantee for (weakly) submodular functions. We also compare LDGM with the gradient-based algorithm Frank-Wolfe under noise, and show that LDGM can be more robust. Empirical results on budget allocation verify the effectiveness of LDGM.

1 Introduction

Submodularity, which implies the diminishing return property, is usually defined on set functions. Submodular set function maximization arises in many applications, such as maximum coverage \([10]\), influence maximization \([14]\) and sensor placement \([16]\), to name a few. It is NP-hard in general, and has received a lot of attentions \([15]\). A well-known result is that for maximizing monotone submodular set functions with a size constraint, the greedy algorithm, which iteratively selects one element with the largest marginal gain, can achieve the optimal approximation guarantee of \((1 - 1/e)\) \([18, 19]\).

Meanwhile, many practical applications involve the objective functions defined over the integer lattice instead of subsets, e.g., budget allocation \([1]\) and welfare maximization \([13]\). Submodularity is thus also extended to functions over the integer lattice. Note that in this case, submodularity does not imply the diminishing return property, which is called DR-submodularity \([22]\). The latter is actually stronger, although they are equivalent for set functions. For monotone DR-submodular function maximization with a size constraint, the greedy algorithm can achieve a \((1 - 1/e)\)-approximation guarantee \([21]\); while for submodular functions, the generalized greedy algorithm, which can select multiple copies of the same element simultaneously in one iteration, achieves a \((1/2)(1 - 1/e)\)-approximation guarantee \([1]\).

Recently, submodularity has been further extended from discrete to continuous domains. Submodular continuous functions are a class of generally non-convex and non-concave functions, which also appear in many applications \([6]\). For maximizing monotone DR-submodular continuous functions with a convex polytope constraint, Chekuri et al. \([8]\) proposed a multiplicative weight update method which can achieve a \((1 - 1/e - \epsilon)\)-approximation guarantee after \(O(n^2/\epsilon^2)\) steps. Bian et al. \([6]\) considered a down-closed convex constraint, and proposed a Frank-Wolfe (FW) variant algorithm which can achieve a \((1 - 1/e - \epsilon)\)-approximation guarantee after \(O(1/\epsilon)\) iterations. Later, stochastic
We study the continuous functions \( f \) we assume that monotone functions are normalized, i.e.,wise minimum and maximum, respectively, that is, we prove that in some special situations, LDGM using the generalized greedy algorithm can achieve (1−
−\epsilon)-approximation guarantee after \( O(1/\epsilon^2) \) iterations. Note that there were also some works focusing on submodular continuous minimization \( 3, 4, 23 \) and non-monotone submodular continuous maximization \( 5, 6 \).

All the above mentioned algorithms require access to the gradients of the objective functions or their unbiased estimates. Thus, they cannot be directly applied to non-differentiable functions. In fact, there are many natural applications with non-differentiable objective functions \( 6 \). For the generalized maximum coverage problem, each subset \( C_i \) (\( 1 \leq i \leq n \)) has a confidence \( x_i \in [0, 1] \) and a monotone covering function \( p_i : \mathbb{R}^+ \to 2^{C_i} \), and the objective function is \( \bigcup_{i=1}^n p_i(x_i) \). For the extended text summarization problem, each sentence \( i \) has a confidence \( x_i \in [0, 1] \) and a monotone covering function \( p_i : \mathbb{R}^+ \to 2^{C} \) (where \( C \) denotes the set of concepts), and the objective function is \( \sum_{j \in \cup_i p_i(x_i)} c_j \) (where \( c_j \) denotes the credit of concept \( j \)). These two objective functions are obviously non-differentiable. Moreover, in noisy environments where only polluted objective values can be obtained, numerical differentiation is known to be ill-posed \( 9 \), and thus it can be difficult to acquire good gradient estimates.

In this paper, we propose the first derivative-free algorithm for the problem of maximizing monotone (weakly) DR-submodular continuous functions subject to a convex polytope constraint. The idea is to discretize the original continuous optimization problem into an optimization problem over the integer lattice by utilizing the frontier of the vertex set of the polytope, and then apply the greedy algorithm. This approach only requires the oracle access to the function value, and we call it Lattice Discretization Greedy Method (LDGM). We introduce the notion of the submodularity ratio \( \alpha \in [0, 1] \) and the DR-submodularity ratio \( \beta \in [0, 1] \) to characterize how close a general continuous function \( f \) is to submodularity and DR-submodularity, respectively. Our main theoretical results can be summarized as follows:

- For monotone (weakly) DR-submodular continuous maximization with a convex polytope constraint, we prove that LDGM can achieve a \((1−\epsilon^\beta−\epsilon)\)-approximation guarantee after \( O(1/\epsilon) \) iterations.
- For monotone (weakly) submodular continuous maximization with a convex polytope constraint, we prove that in some special situations, LDGM using the generalized greedy algorithm can achieve a \((\epsilon/2)(1−\epsilon^\beta)\)-approximation guarantee after \( O(1/\epsilon) \) iterations. Note that this is the first approximation guarantee for monotone submodular continuous maximization.
- We compare LDGM with FW \( 6 \) under noise, and show that LDGM can be more robust to noise.

Empirical results on budget allocation show the superior performance of LDGM.

The rest of the paper is organized as follows. Section 2 introduces the studied problem and also gives some preliminaries. In Section 3, we propose the LDGM method and give its approximation guarantee. Section 4 presents the analysis under noise, and Section 5 gives the empirical studies. In Section 6, we conclude this paper. Appendix can be found in Section 7.

## 2 Monotone DR-submodular Continuous Maximization

**Notation.** Let \( \mathbb{R}, \mathbb{R}^+ \) and \( \mathbb{Z}^+ \) denote the set of reals, non-negative reals and non-negative integers, respectively. For two vectors \( x, y \in \mathbb{R}^n \), let \( x \wedge y \) and \( x \vee y \) denote the coordinate-wise minimum and maximum, respectively, that is, \( x \wedge y = \min \{ x_1, y_1 \}, \ldots, \min \{ x_n, y_n \} \) and \( x \vee y = \max \{ x_1, y_1 \}, \ldots, \max \{ x_n, y_n \} \). We use \( \| \cdot \| \) to denote the Euclidean norm of a vector.

The \( i \)-th unit vector is denoted by \( \chi_i \), that is, the \( i \)-th entry of \( \chi_i \) is 1 and others are 0; the all-zeros and all-ones vectors are denoted by \( \mathbf{0} \) and \( \mathbf{1} \), respectively. Let \( [n] \) denote the set \( \{ 1, 2, \ldots, n \} \). We denote \( conv(\cdot) \) as the convex hull of a set. For \( x, y \in \mathbb{R}^n \), we say \( x \leq y \) if \( x_i \leq y_i \), for every \( i \); \( x < y \) if \( x \leq y \) and \( x_i < y_i \) for some \( i \).

We study the continuous functions \( f : \mathcal{X} = \prod_{i=1}^n X_i \to \mathbb{R} \), where \( X_i \) is a compact subset of \( \mathbb{R}^+ \). A function \( f : \mathcal{X} \to \mathbb{R} \) is monotone if for any \( x \leq y \), \( f(x) \leq f(y) \). Without loss of generality, we assume that monotone functions are normalized, i.e., \( f(0) = 0 \). For a function \( f : \mathcal{X} \to \mathbb{R} \), submodularity (as presented in Definition \( 1 \)) does not imply the diminishing return property (called DR-submodularity as presented in Definition \( 8 \)). DR-submodularity is stronger than submodularity,
that is, a DR-submodular function is submodular, but not vice versa. In [6], it has been proved that submodularity is equivalent to a weak version of DR-submodularity, as presented in Definition 3.

**Definition 1 (Submodular [3]).** A function \( f : \mathcal{X} \to \mathbb{R} \) is submodular if for any \( x, y \in \mathcal{X} \),
\[
f(x) + f(y) \geq f(x \wedge y) + f(x \vee y).
\]

**Definition 2 (DR-Submodular [6]).** A function \( f : \mathcal{X} \to \mathbb{R} \) is DR-submodular if for any \( x \leq y, k \in \mathbb{R}^+ \) and \( i \in [n] \),
\[
f(x + kx_i) - f(x) \geq f(y + kx_i) - f(y).
\]

**Definition 3 (Weak DR-Submodular [6]).** A function \( f : \mathcal{X} \to \mathbb{R} \) is weak DR-submodular if for any \( x \leq y, k \in \mathbb{R}^+ \) and \( i \in [n] \) with \( x_i = y_i \),
\[
f(x + kx_i) - f(x) \geq f(y + kx_i) - f(y).
\]

According to the equivalence between submodularity and weak DR-submodularity, Definitions 2 and 3, we define the submodularity ratio \( \alpha \) as well as the DR-submodularity ratio \( \beta \), which measure to what extent a general continuous function \( f \) has submodular and DR-submodular properties, respectively. They are generalizations of that for functions over the integer lattice [20].

**Definition 4 (Submodularity Ratio).** The submodularity ratio of a continuous function \( f : \mathcal{X} \to \mathbb{R} \) is defined as
\[
\alpha = \inf_{x,y \in \mathcal{X} : x \leq y, k \in \mathbb{R}^+, i \in [n]} \frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)}.
\]

**Definition 5 (DR-Submodularity Ratio).** The DR-submodularity ratio of a continuous function \( f : \mathcal{X} \to \mathbb{R} \) is defined as
\[
\beta = \inf_{x,y \in \mathcal{X} : x \leq y, k \in \mathbb{R}^+, i \in [n]} \frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)}.
\]

Note that in [12], the notion of DR-submodularity ratio for differentiable functions was also defined as
\[
\gamma = \inf_{x,y \in \mathcal{X} : x \leq y, i \in [n]} \frac{[\nabla f(x)]_i}{[\nabla f(y)]_i},
\]
where \([\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}\) is the \( i \)-th component of the gradient. Our definition \( \beta \) does not require the differentiable property. Furthermore, it can be shown that \( \beta \) and \( \gamma \) are equivalent when \( f \) has continuous second derivatives in \( \mathcal{X} \). Let \( x, y, i \) correspond to the value of \( \gamma \). Then, \( \gamma = \frac{[\nabla f(x)]_i}{[\nabla f(y)]_i} = \lim_{m \to \infty} \frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)} \geq \beta \), where the last equality is because we do not need to care about \( 0 \) or \( \infty \), and the inequality is by the definition of \( \beta \). Next, we prove that \( \gamma \leq \beta \). For any \( x \leq y, i \in [n], k \in \mathbb{R}^+ \), we have
\[
\frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)} = \frac{\sum_{j=0}^{m-1} f(x + kx_i) - f(x + \frac{k}{m}x_i)}{\sum_{j=0}^{m-1} f(y + kx_i) - f(y + \frac{k}{m}x_i)} = \lim_{m \to \infty} \frac{\sum_{j=0}^{m-1} f(x + kx_i) - f(x + \frac{k}{m}x_i)}{\sum_{j=0}^{m-1} f(y + kx_i) - f(y + \frac{k}{m}x_i)} + R_{i,j},
\]
where \( R_{i,j} \) is the Lagrange remainder, i.e., \( R_{i,j} = k \frac{\partial^2 f(x)}{\partial x_j^2} |_{x=x_i} \). Since \( \frac{\partial^2 f(x)}{\partial x_j^2} \) is bounded by its maximum and minimum in \( \mathcal{X} \), we have \( \lim_{m \to \infty} \sum_{j=0}^{m-1} R_{i,j} = 0 \). Thus,
\[
\frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)} = \frac{\sum_{j=0}^{m-1} f(x + kx_i) - f(x)}{\sum_{j=0}^{m-1} f(y + kx_i) - f(y)} \geq \frac{\sum_{j=0}^{m-1} f(x + kx_i) - f(x)}{\sum_{j=0}^{m-1} f(y + kx_i) - f(y)},
\]
where the inequality is by the definition of \( \gamma \), hence \( \beta \geq \gamma \).

It is easy to see that \( \beta \leq \alpha \). Note that \( \frac{f(x + kx_i) - f(x)}{f(y + kx_i) - f(y)} \) reaches 1 by letting \( x = y \), so \( \beta \leq \alpha \leq 1 \). For a monotone continuous function \( f \), we make the following observations:

**Remark 1.** For a monotone continuous function \( f : \mathcal{X} \to \mathbb{R} \), it holds that 1) \( 0 \leq \beta \leq \alpha \leq 1 \); 2) \( f \) is submodular iff \( \alpha = 1 \); 3) \( f \) is DR-submodular iff \( \beta = 1 \).
Our studied problem as presented in Definition \ref{def:general_problem} is to maximize a monotone continuous function \( f \) in a convex polytope \( P \). A convex polytope in \( \mathbb{R}^n \) has two equivalent definitions, i.e., \( V \)-type polytope defined by the convex hull of a finite number of points in \( \mathbb{R}^n \), and \( H \)-type polytope defined by the intersection of a finite number of half-spaces. There exist algorithms to convert from one to the other, e.g., the reverse search method \cite{conf/stoc/Borgwardt08} can find all vertices of a \( H \)-type polytope in time linear to the product of the number of vertices, the number of half-spaces and the dimension.

**Definition 6 (The General Problem).** Given a convex polytope \( P = \text{conv}(E) \subseteq \mathbb{R}_+^n \), where \( E \) is a set of points located in the positive space, it is to maximize a monotone function \( f : \mathcal{X} \rightarrow \mathbb{R} \) in \( P \), i.e.,

\[
\arg \max_{x \in P} f(x)
\]  

(1)

### 3 The Proposed Approach

In this section, we propose the derivative-free algorithm LDGM for maximizing monotone (weakly) DR-submodular continuous functions with a convex polytope constraint. The procedure of LDGM is presented in Algorithm \ref{alg:ldgm}.

Our algorithm iteratively adds one point from \( E \) with the largest marginal gain until \( l \) points are selected (i.e., lines 3-7). Note that \( l \) is a parameter that controls step size, and the size of iteration step \( (v \text{ at line 5}) \) is not necessarily the same as the size of lookahead step \( (\gamma \text{ at line 4}) \). The following analysis in this section assumes \( \gamma = 1 \). Setting \( \gamma > 1 \) would help mitigate impacts from noise, which is shown in section \ref{sec:approx}."
Assumption 1. The function $f$ is Lipschitz continuous with constant $L$ in $\text{Frontier}(\mathcal{P})$, i.e., $\forall x, y \in \text{Frontier}(\mathcal{P}), \ |f(x) - f(y)| \leq L \cdot \|x - y\|$. 

Next, we make an assumption on the radius of $\mathcal{P}$.

Assumption 2. The convex body $\mathcal{P}$ is bounded, otherwise the maximization is meaningless given that the objective function $f$ is monotone. Thus, there must exist $D \in \mathbb{R}_+$ such that $\forall x \in \mathcal{P} : \|x\| \leq D$.

Let $OPT$ denote the optimal function value. We show in Theorem 1 that LDGM can achieve the approximation guarantee of $(1 - e^{-\beta} - \epsilon)$ after $O(1/\epsilon)$ iterations. Note that we set the lookahead step size $\gamma = 1$, as we do not consider noise in this section.

Theorem 1. For maximizing monotone (weakly) DR-submodular continuous functions with a convex polytope constraint $\mathcal{P} = \text{conv}(E)$, LDGM with $l$ iterations and $\gamma = 1$ can find a solution $x \in \mathcal{P}$ having $f(x) \geq (1 - e^{-\beta}) \cdot OPT - \frac{(1-e^{-\beta})mDL}{1/l}$, where $m \leq |E|$.

Before giving the proof of Theorem 1 we introduce some lemmas. Lemma 1 extends the definition of DR-submodularity ratio to show diminishing return property exists for arbitrary vector increment. Lemma 2 shows that for the optimization over the integer lattice $\mathbb{Z}_c^e$, there always exists a point from $E$, the inclusion of which can bring an improvement on $f$ proportional to the current distance to the optimum on the lattice.

Lemma 1. For a monotone continuous function $f : \mathcal{X} \to \mathbb{R}$, any $x, y \in \mathcal{X}$ with $x \leq y$, and any vector $v \in \mathbb{R}^n_+$, we have

$$
\frac{f(x+v) - f(x)}{f(y+v) - f(y)} \geq \beta
$$

Proof. Let $v(i) = v_i \chi_i$. Then, we have

$$
\frac{f(x+v) - f(x)}{f(y+v) - f(y)} = \frac{\sum_{j=1}^{n} f(x + \sum_{i=1}^{j} v(i)) - f(x + \sum_{i=1}^{j-1} v(i))}{\sum_{j=1}^{n} f(y + \sum_{i=1}^{j} v(i)) - f(y + \sum_{i=1}^{j-1} v(i))} \\
\geq \frac{\beta \cdot \sum_{j=1}^{n} f(y + \sum_{i=1}^{j} v(i)) - f(y + \sum_{i=1}^{j-1} v(i))}{\sum_{j=1}^{n} f(y + \sum_{i=1}^{j} v(i)) - f(y + \sum_{i=1}^{j-1} v(i))} = \beta,
$$

where the inequality is derived from the definition of DR-submodularity ratio, i.e., Definition 5.

Lemma 2. Let $v^*$ be the best solution one can achieve using $l$ vectors in $E$, denoted as $v^* = \sum_{i=1}^{l} e_i$, where $e_i \in E$. Then for any $x \in \mathcal{X}$, there exists a vector $e^* \in E$ such that

$$
f(x + e^*) - f(x) \geq \frac{\beta}{l} (f(v^*) - f(x)).
$$

Proof. Let $e^* \in \arg\max_{e \in E} f(x+e)$. Then, we have

$$
f(v^*) - f(x) \leq f(x + v^*) - f(x) = f \left( x + \sum_{i=1}^{l} e_i \right) - f(x) \\
= \sum_{k=1}^{l} f \left( x + \sum_{i=1}^{k} e_i \right) - f \left( x + \sum_{i=1}^{k-1} e_i \right) \\
\leq \frac{1}{\beta} \sum_{k=1}^{l} f(x + e_k) - f(x) \leq \frac{l}{\beta} \cdot (f(x + e^*) - f(x)),
$$

where the first inequality is by the monotonicity of $f$, and the second is by Lemma 1.

Based on Lemma 2 we can prove an approximation guarantee w.r.t. the best solution on the lattice $\mathbb{Z}_c^e$. To further bound the difference between the best solution on the lattice and the true optimal solution $x^*$, we give Lemma 4, which shows that there exists a point on the lattice which is close enough to $x^*$. Note that $x^* \in \text{Frontier}(\mathcal{P})$ due to the monotonicity. Before introducing Lemma 4
we prove a geometry result regarding \( \text{Frontier}(\mathcal{P}) \). It is later used in Lemma 4 to show that there must exist a solution \( \psi' \) in \( \text{Frontier}(\mathcal{P}) \) and on the lattice, close to the global optimal solution \( x^* \) in Euclidean distance sense. This suggests that the Lipschitz Assumption is only required in \( \text{Frontier}(\mathcal{P}) \).

**Lemma 3.** Let \( X = \{x_1, \ldots, x_m\} \), where \( \forall i \in [m], x_i \in \mathcal{P} \). If there exists \( \theta_1, \ldots, \theta_m > 0 \) such that \( \sum_{i=1}^m \theta_i = 1 \) and \( x' = \sum_{i=1}^m \theta_i x_i \in \text{Frontier}(\mathcal{P}) \), then we have \( \text{conv}(X) \subseteq \text{Frontier}(\mathcal{P}) \).

**Proof.** When \( m = 1 \), obviously the lemma holds. When \( m = 2 \), we assume that there exists \( x' = \theta_1 x_1 + \theta_2 x_2 \in \text{Frontier}(\mathcal{P}) \) where \( \theta_1, \theta_2 > 0 \) and \( \theta_1 + \theta_2 = 1 \), but the lemma does not hold, i.e., \( \exists t \in \text{conv}(X) \) such that \( t \notin \text{Frontier}(\mathcal{P}) \). We then show this makes a contradiction. Let \( t = \eta_1 x_1 + \eta_2 x_2 \), where \( \eta_1, \eta_2 \geq 0 \) and \( \eta_1 + \eta_2 = 1 \). Since \( t \notin \text{Frontier}(\mathcal{P}) \), there exists \( t' \in \mathcal{P} \) such that \( t' > t \). Let \( \Delta t = t' - t \) and \( x'' = x' + \xi t \). Since \( \theta_1, \theta_2 > 0 \) and \( \eta_1, \eta_2 \geq 0 \), there exists a convex combination of \( x_1, x_2, t' \). As \( x_1, x_2, t' \in \mathcal{P} \), we have \( x'' \in \mathcal{P} \). Furthermore, \( x'' > x' \). Thus, we have \( x'' \notin \text{Frontier}(\mathcal{P}) \), which makes a contradiction. Thus, the lemma holds when \( m = 2 \). Next we prove it for any \( m \in \mathbb{Z}_+ \) by induction.

Let \( X = \{x_1, \ldots, x_m\} \) and \( x' = \sum_{i=1}^m \theta_i x_i \in \text{Frontier}(\mathcal{P}) \). We are to prove that for any \( x = \sum_{i=1}^m \eta_i x_i \in \text{conv}(X) \), \( x \in \text{Frontier}(\mathcal{P}) \). Let \( y = x' - \xi (x - x') \), where \( \xi > 0 \). Then, we have \( y = (\sum_{i=1}^m \theta_i x_i) - \xi (\sum_{i=1}^m \eta_i x_i) = \sum_{i=1}^m ((1 + \xi) \theta_i - \xi \eta_i) x_i \). Note that \( \theta_i > 0 \) and \( \eta_i \geq 0 \), so we can always find a sufficient small \( \xi > 0 \), such that \( \forall i \in [m] : (1 + \xi) \theta_i - \xi \eta_i > 0 \), i.e., \( y \in \text{conv}(X) \). From the definition of \( y \), we get \( x' = \frac{1}{1+\xi} y + \frac{\xi}{1+\xi} x \). Note that \( x, y \in \text{conv}(X) \subseteq \mathcal{P}, x' \in \text{Frontier}(\mathcal{P}) \) and \( \xi > 0 \). Since the lemma with \( m = 2 \) has been proved, we have \( \text{conv} \{x, y\} \subseteq \text{Frontier}(\mathcal{P}) \), which implies that \( x \in \text{Frontier}(\mathcal{P}) \). As \( x \) is chosen arbitrarily in \( \text{conv}(X) \), we can conclude that \( \text{conv}(X) \subseteq \text{Frontier}(\mathcal{P}) \).

Now we are ready to prove Lemma 4, which is to bound the difference between the closest solution \( \psi' \) on lattice and the global optimal solution \( x^* \).

**Lemma 4.** Let \( x^* \in \text{Frontier}(\mathcal{P}) \) denote a global optimal solution and \( |\mathcal{E}| = m \). There exists \( e_1', \ldots, e_l' \in \mathcal{E} \) such that \( \psi' = \sum_{i=1}^l e_i' \in \text{Frontier}(\mathcal{P}) \) and \( ||x^* - \psi'|| \leq mD/l \).

**Proof.** Let \( l \cdot \mathcal{E} = \{r_i \mid r_i = l \cdot e_i, e_i \in \mathcal{E}\} \). Suppose \( x^* \notin \text{conv}(l \cdot \mathcal{E}) \). Since \( x^* \in \mathcal{P} = \text{conv}(E) \), there exist \( y_1 \ldots y_t \in E \) such that \( x^* = \sum_{i=1}^t \theta_i y_i \), where \( \theta_1, \ldots, \theta_t > 0 \) and \( \sum_{i=1}^t \theta_i = 1 \). As \( x^* \notin \text{conv}(l \cdot \mathcal{E}) \), there exists \( i \in \lfloor l \rfloor : y_i \notin l \cdot \mathcal{E} \). By the construction of \( \mathcal{E} \) (i.e., line 1 of Algorithm 1), we have \( y_i \notin \text{Frontier}(\mathcal{P}) \). So we can find \( y_i' > y_i \) in \( \mathcal{P} \). Let \( x' = \sum_{j=1}^{t-1} \theta_j y_j + \theta_t y_i' + \sum_{j=t+1}^m \theta_j y_j > x^* \). As \( x' \) is a convex combination of points in \( \mathcal{P} \), we know that \( x' \) is still in \( \mathcal{P} \). That is, \( x' \in \mathcal{P} \) and \( x' > x^* \), which makes a contradiction with \( x^* \in \text{Frontier}(\mathcal{P}) \). Thus, \( x^* \in \text{conv}(l \cdot \mathcal{E}) \).

Due to \( x^* \in \text{conv}(l \cdot \mathcal{E}) \), there exist \( \theta_i > 0 \) \( (i = 1, 2, \ldots, m) \), \( m' \leq m \) and \( \sum_{i=1}^{m'} \theta_i = 1 \) such that \( x^* = \sum_{i=1}^{m'} \theta_i r_i \). Note that we have made an assumption that those \( \theta_i = 0 \) correspond to the last \( m - m' \) terms. We then show that there exists \( x' = \sum_{i=1}^{m'} \theta'_i r_i \) with \( \sum_{i=1}^{m'} \theta'_i = 1 \) and \( \theta'_i = k_i \cdot \frac{1}{l} \in \mathbb{Z}_+ \) such that \( \forall i \in [m'], |\theta_i - \theta'_i| \leq 1/l \). We first construct an initial \( x' \) by randomly assigning \( k_i \in \mathbb{Z}_+ \) subject to \( \sum_{i=1}^{m'} \theta'_i = 1 \). If currently \( x' \) does not satisfy \( \forall i \in [m'], |\theta_i - \theta'_i| \leq 1/l \), suppose there are in total \( t \) positions violating the constraint. Let \( j \) denote one position violating the constraint, then \( \theta_j - \theta'_j > 1/l \) or \( \theta'_j - \theta_j > 1/l \). We consider \( \theta_j - \theta'_j > 1/l \) and the other one can be similarly analyzed. Meanwhile, there must exist another position \( k \) with \( \theta_k - \theta'_k < 0 \); otherwise, \( \sum_{i=1}^{m'} \theta_i > 1/l + \sum_{i=1}^{m'} \theta'_i = 1 + 1/l \), which makes a contradiction. Then, we make such a change: \( \theta'_j := \theta'_j + 1/l \) and \( \theta'_k := \theta'_k - 1/l \). Thus, \( \theta'_j \) becomes closer to \( \theta_j \). For \( \theta'_k \), we have \( \theta_k - \theta'_k \leq 1/l \), which implies that the change at position \( k \) will not increase the number \( t \) of violations. By repeating this procedure, we can decrease \( t \) to 0, i.e., \( x' = \sum_{i=1}^{m'} \theta'_i r_i \) satisfies that \( \sum_{i=1}^{m'} \theta'_i = 1 \) and \( \forall i \in [m'], |\theta_i - \theta'_i| \leq 1/l \). Note that \( \forall i \in [m'] : \theta_i > 0 \) and \( \theta'_i \) can only be \( 1/l \).
times an integer, thus \( \forall i \in [m'] : \theta_i' \geq 0 \). Let \( v' = x' \). Then, we have

\[
\|x^* - v'\| \leq \sum_{i=1}^{m'} |\theta_i - \theta_i'| \cdot \|r_i\| \leq \frac{1}{\ell} \sum_{i=1}^{m'} \|r_i\| \leq \frac{m'D}{\ell} \leq \frac{mD}{\ell}.
\]

Finally, we only need to show that \( v' \in \text{Frontier}(\mathcal{P}) \). Note that \( \forall i \in [m'] : r_i \in \mathcal{P} \), and \( x^* = \sum_{i=1}^{m'} \theta_i r_i \in \text{Frontier}(\mathcal{P}) \) is a convex combination of them with all \( \theta_i > 0 \). According to Lemma 3, \( \text{conv}\{r_1, \ldots, r_{m'}\} \subseteq \text{Frontier}(\mathcal{P}) \). As \( x' \in \text{conv}\{r_1, \ldots, r_{m'}\} \), we have \( x' \in \text{Frontier}(\mathcal{P}) \). Thus, the lemma holds.

Then, we can prove Theorem 1 by using Lemmas 2 and 4.

**Proof of Theorem 1.** Let \( x^* \in \text{Frontier}(\mathcal{P}) \) denote a global optimal solution, i.e., \( f(x^*) = \text{OPT} \).
Let \( v' \in \text{Frontier}(\mathcal{P}) \) be the point suggested by Lemma 4. Let \( v^* \) be the best solution one can achieve using the sum of \( l \) vectors in \( \mathcal{E} \). According to Lemma 4 and Assumption 1 we get

\[
f(x^*) - f(v^*) \leq f(x^*) - f(v') \leq L \cdot \|x^* - v'\| \leq mDL/l,
\]

which implies that \( f(v^*) \geq f(x^*) - mDL/l = \text{OPT} - mDL/l \).

According to the algorithm procedure and Lemma 3 at the \( t \)-th iteration where \( t = 1, 2, \ldots, l - 1 \),

\[
f(x_{t+1}) - f(x_t) \geq \frac{\beta}{l} (f(v^*) - f(x_t)) \geq \frac{\beta}{l} (\text{OPT} - mDL/l - f(x_t)).
\]

By a simple transformation, we can equivalently get

\[
\text{OPT} - mDL/l - f(x_{t+1}) \leq (1 - \beta/l) \cdot (\text{OPT} - mDL/l - f(x_t)).
\]

Thus, \( \text{OPT} - mDL/l - f(x_t) \leq (1 - \beta/l) \cdot (\text{OPT} - mDL/l - f(x_0)) \leq e^{-\beta} (\text{OPT} - mDL/l) \), which leads to \( f(x_t) \geq (1 - e^{-\beta}) \cdot \text{OPT} - (1 - e^{-\beta}) mDL/l \).

If the set \( \mathcal{E} \) produced by line 1 of Algorithm 1 is an orthogonal set, i.e., the inner product of any two vectors from \( \mathcal{E} \) equals zero, the LDGM variant can be able to handle the weakly submodular case. In LDGM, for the optimization over the integer lattice \( \mathbb{Z}^n_+ \), we will apply the generalized greedy algorithm instead of the greedy algorithm. That is, in each iteration, it selects a combination \((e, j)\) (where \( e \in \mathcal{E} \)) such that the average marginal gain by adding \( j \) copies of \( e \) is maximized. The detailed algorithm procedure and the proof of Theorem 2 are provided in the appendix.

**Theorem 2.** For maximizing monotone (weakly) submodular continuous functions with a convex polytope constraint \( \mathcal{P} = \text{conv}(\mathcal{E}) \), if \( \mathcal{E} = \{1/2 x \mid x \in \text{ Frontier(\mathcal{E})}\} \) is an orthogonal set, LDGM with the generalized greedy algorithm can find a solution \( x \) having \( f(x) \geq \frac{\alpha}{\ell} (1 - \epsilon^{-\alpha}) \text{OPT} - \frac{\alpha(1 - \epsilon^{-\alpha}) mDL}{2\ell^2} \) with \( l \) iterations and \( \gamma = 1 \), where \( m \leq |\mathcal{E}| \).

3.2 Discussion

LDGM works well when the constraint is given by \( \mathcal{V} \)-type polytope and some classes of \( \mathcal{H} \)-type polytope. Although in general the number of vectors (i.e., \( |\mathcal{E}| \) ) can be exponential given a \( \mathcal{H} \)-type polytope, many common classes of \( \mathcal{H} \)-type polytopes have limited vertices. One trivial example having \( n + 1 \) vertices is \( \mathcal{P} = \{ x \mid |x|_1 \leq b, x \geq 0 \} \). In fact for any \( \alpha > 0 \), \( \mathcal{P} = \{ x \mid a^T x \leq b, x \geq 0 \} \) has \( n + 1 \) vertices; for any \( A \in \mathbb{R}^{r \times n} \) and \( r = O(1) \), \( \mathcal{P} = \{ x \mid A^T x \leq b, x \geq 0 \} \) has \( O(n^r) \) (i.e., polynomial) vertices.

Moreover, although the common convex polytope constraint \( \mathcal{P} = \{ x \mid a^T x \leq b, 0 \leq x \leq c \} \) (where \( \alpha > 0 \) ) can have exponential number of vertices, LDGM with a slight modification can still obtain the (1 - \( 1/\epsilon \) - approximation guarantee after \( O(1/\epsilon) \) iterations.

**Modification.** Let \( \mathcal{P} = \mathcal{Q} \cap \mathcal{C} \), where \( \mathcal{Q} = \{ x \mid a^T x \leq b, x \geq 0 \} \) and \( \mathcal{C} = \{ x \mid 0 \leq x \leq c \} \). We run LDGM on \( \mathcal{Q} \) instead of \( \mathcal{P} \) (so the number of vertices is \(|\mathcal{E}| = n + 1\) and \(|\mathcal{E}| = |\text{ Frontier}(\mathcal{E})| = n\)), and change line 4 of LDGM from \( \text{”} v := \arg\max_{e \in \mathcal{E}} f(x_t + e) - f(x_t) \text{”} \) to \( \text{”} v := \arg\max_{e \in \mathcal{E}} f(x_t + e) - f(x_t) \text{”} \) (so the output solution must belong to \( \mathcal{Q} \cap \mathcal{C} = \mathcal{P} \)). With a stronger Assumption 1, i.e., \( f(\cdot) \) is Lipschitz continuous with constant \( L \) in \( \mathcal{P} \) instead of only in \( \text{ Frontier}(\mathcal{P}) \), we can similarly prove the same approximation guarantee for this case. As Lemma 1 and Lemma 3 have nothing to do with this adaption, we only need to care about Lemma 2 and Lemma 4. We first derive Lemma 5 adapted from Lemma 4. Note that we still set \( \gamma = 1 \) here.
Lemma 5. Given the above modification, let \( v^* \) be the best solution one can achieve using \( l \) vectors in \( E \), denoted as \( v^* = \sum_{i=1}^l e_i \), where \( e_i \in \mathcal{E} \) and \( v^* \in Q \cap C \). Note that in this case each \( e_i \) has only one entry that is not zero but a positive value. Then for step \( t < l \) in the algorithm, let \( e^* \in \arg \max_{e \in \mathcal{E}} f(x_t + e) \) s.t. \( x_t + e \in C \), we have

\[
f(x_t + e^*) - f(x_t) \geq \frac{\beta}{l} (f(v^*) - f(x_t)).
\]

Proof. From monotonocity we have

\[
f(v^*) - f(x_t) \leq f(x_t \vee v^*) - f(x_t)
\]

denote \( v = x_t \vee v^* - x_t = \sum_{i=1}^{l'} e'_i \). As \( x_t \vee v^* \in C \), we have for every \( i \) that: \( x_t + e'_i \in C \). Then

\[
f(x_t \vee v^*) - f(x_t) = f \left( x_t + \sum_{i=1}^{l'} e'_i \right) - f(x_t)
\]

\[
= \sum_{k=1}^{l'} f \left( x_t + \sum_{i=1}^{k} e'_i \right) - f \left( x_t + \sum_{i=1}^{k-1} e'_i \right)
\]

\[
\leq \frac{1}{\beta} \sum_{k=1}^{l'} f(x_t + e_k) - f(x_t) \leq \frac{1}{\beta} \cdot (f(x_t + e^*) - f(x_t)),
\]

where the first inequality is by Lemma 1 and the second is due \( l' \leq l \) as well as the definition of \( e^* \).

Similarly to Lemma 4, we derive the following lemma to bound the global optimal solution and its closest feasible solution on lattice.

Lemma 6. Using the slightly modified algorithm above, let \( \mathcal{P} = Q \cap C \), where \( Q = \{x | \alpha^T x \leq b, x \geq 0\} \) and \( C = \{x | 0 \leq x \leq c\} \). Let \( x^* \in \text{Frontier}(\mathcal{P}) \) denote a global optimal solution and \( |E'| = m \). There exists \( e'_1, \ldots, e'_l \in E \) such that \( v' = \sum_{i=1}^{l'} e'_i \in \mathcal{P} \) and \( \|x^* - v'\| \leq mD/l \), where \( l' \leq l \) is an integer.

Proof. The Lemma 4 present upper bounds for the distance between true optimum \( x^* \) and its closest point on the lattice \( v' \). In this particular case, we first use the same process as in Lemma 4 to construct \( x' = \sum_{i=1}^{m'} \theta'_i r_i \) satisfying \( \sum_{i=1}^{m'} \theta'_i = 1 \) and \( \forall i \in [m'], |\theta'_i - \theta_i| \leq 1/l \). Note that \( x^* = \sum_{i=1}^{m} \theta_i r_i \), and as we run the algorithm under constraint \( Q \), \( \{r_i\} \) are in fact vertices of \( Q \). Now the tricky thing is we want every \( \theta'_i r_i \in C \). Due to \( C \) is down-closed and \( x^* \in C \), we have \( \forall i : \theta_i r_i \in C \). Therefore we have \( \theta'_i > \theta_i \) for every \( \theta'_i : \theta_i r_i \notin C \). Given that we know \( |\theta'_i - \theta_i| \leq 1/l \), we reassign \( \theta'_i = \theta'_i - 1/l \) for every \( \theta'_i \) where \( \theta_i r_i \notin C \), so that \( \theta'_i \leq \theta_i \) while the property \( |\theta'_i - \theta_i| \leq 1/l \) still holds. Note that now \( x' \in C \).

Now we have a solution \( x' = \sum_{i=1}^{m'} \theta'_i r_i \) satisfying \( \sum_{i=1}^{m'} \theta'_i = 1 \) and \( \forall i \in [m'], |\theta'_i - \theta_i| \leq 1/l \). As \( 0 \in Q \), we can write \( x' = (1 - \sum_{i=1}^{m'} \theta'_i)0 + \sum_{i=1}^{m'} \theta'_i r_i \), showing that \( x' \) is a convex combination of points in the convex body \( Q \), which is saying \( x' \in Q \). Thus \( x' \in Q \cap C = \mathcal{P} \). Let \( v' = x' \), we have

\[
\|x^* - v'\| \leq \sum_{i=1}^{m'} |\theta_i - \theta'_i| \cdot \|r_i\| \leq \frac{1}{l} \cdot \sum_{i=1}^{m'} \|r_i\| \leq \frac{m'D}{l} \leq \frac{mD}{l}.
\]

Thus, the lemma holds.

Now it is basically the same to derive the guarantee (as Theorem 1) for this modified version of LDGM. The only difference is that \( v' \) in Lemma 6 is not necessarily in \( \text{Frontier}(\mathcal{P}) \), which requires a stronger Assumption 1 i.e., \( f \) is Lipschitz continuous with constant \( L \) in \( \mathcal{P} \) instead of only in \( \text{Frontier}(\mathcal{P}) \), to derive \( f(x^*) - f(v') \leq mD/L \). Since \( f(v^*) \geq f(v') \), we have \( f(x_t) \geq (1 - 1/e) f(x^*) - (1 - 1/e) mD/L \). This gives the same guarantee for the common convex polytope constraint \( \mathcal{P} = \{x \mid \alpha^T x \leq b, 0 \leq x \leq c\} \).

In addition, we can see that LDGM requires a convex polytope constraint, which cannot handle a general convex constraint. In these situations, we can use a polytope with limited number of vertices to approximate the original convex body \( \mathcal{P} \), and then apply LDGM.
4 Analysis under Noise

In this section, we compare LDGM with the gradient-based algorithm FW \cite{FW} under noise. We consider the problem of maximizing monotone DR-submodular functions with a convex polytope constraint. Let $x^*$ be an optimal solution, and $x_t$ be the solution after $t$ iterations of the FW algorithm.

Let $v_t^* = x^* \lor x_t - x_t$. Assume that in the $t$-th iteration of FW, a gradient call at $x_t$ introduces a noise term $\epsilon_t$. That is, instead of using $\nabla f(x_t)$ in its update policy, it uses $\nabla f(x_t) + \epsilon_t$. Denote $v_t$ as the vector chosen in the $t$-th iteration of FW, i.e., $v_t = \arg \max_{v \in P} \langle v, \nabla f(x_t) + \epsilon_t \rangle$. Suppose FW performs $l$ iterations and uses a constant step size $1/l$. Then, we can show the approximation guarantee of FW under noise as follows. For the sake of readability, we assume $f(\cdot)$ is DR-submodular in this section, i.e., $\beta = 1$, and it is similar when $\beta$ is introduced. The proof is inspired from that of Theorem 1 in \cite{FW}.

**Theorem 3.** The FW algorithm under noise finds a solution $x_t \in P$ with

$$f(x_t) \geq (1 - 1/e) \cdot \text{OPT} - \frac{L}{2l} - \frac{1}{l} \sum_{t=0}^{l-1} (1 - 1/l)i_{1-t}(v_t - v^*_t, \epsilon_t).$$

**Proof.** In each iteration of FW, we have

$$f(x^*) - f(x_t) = \langle v^*_t, \nabla f(x_t) \rangle = \langle v_t, \nabla f(x_t) + \epsilon_t \rangle - \langle v_t, \nabla f(x_t) \rangle + \langle v_t - v^*_t, \epsilon_t \rangle.$$

where the first inequality is due to the monotonicity and coordinate-wise concavity of $f$, and the second one is by the update policy, i.e., $v_t = \arg \max_{v \in P} \langle v, \nabla f(x_t) + \epsilon_t \rangle$. As the FW algorithm assumes $L$-Lipschitz continuous gradient, we have, for any $x, v$,

$$f(x) + \frac{1}{l} \langle v, \nabla f(x) \rangle - f(x + v/l) \leq \frac{L}{2l^2}.$$

By combining the above two formulas and using $x_{t+1} = x_t + \frac{1}{l}v_t$, we get

$$f(x_{t+1}) - f(x^*) \geq (1 - 1/l)(f(x_t) - f(x^*)) - \frac{L}{2l^2} - \frac{1}{l} \langle v_t - v^*_t, \epsilon_t \rangle.$$

By induction, we can derive that $f(x_t) - f(x^*) \geq (1 - 1/l)^t[f(x^*)] - \frac{L}{2l^2} - \frac{1}{l} \sum_{t=0}^{l-1} (1 - 1/l)^{t-1} \langle v_t - v^*_t, \epsilon_t \rangle$. Thus, the theorem holds. \hfill \Box

For LDGM, we set $\gamma > 1$ to mitigate impacts from noise. The intuition here is that we want the differences among all options ($e \in \mathcal{E}$) be sufficient large to avoid being interfered by noises. Let $x_t$ denote the solution after $t$ iterations. Let $\mathcal{E} = \{e_1, \ldots, e_m\}$. In each iteration of LDGM, it will make calls to obtain $f(x_t + e_i)$ for each $e_i \in \mathcal{E}$. To analyze LDGM under noise, we assume that in the $(t + 1)$-th iteration of LDGM, it introduces a noise of $\epsilon_{t,i}$ to $f(x_t + e_i)$, i.e., the actual value returned is $f(x_t + e_i) + \epsilon_{t,i}$.

Similarly to Lemma\cite{LDGM}, Lemma\cite{LDGM} shows that even under noise, the improvement on $f$ in each step is proportional to the current distance to the optimum on the lattice. The difference, aside from noise, is the best solution on lattice $v^*$ is substituted by the best solution on a 'sparser' lattice, i.e. $v^*_t$ in Lemma\cite{LDGM}.

To simplify the proof, we assume $\gamma$ is an integer and $l$ is divisible by $\gamma$.

**Lemma 7.** Let $v^*_t$ be the best solution one can achieve using $1/\gamma$ vectors in $\mathcal{E}$, denoted as $v^* = \sum_{j=1}^{l/\gamma} \gamma e_{ij}$, where $e_{ij} \in \mathcal{E}$. Then for any $x_t$, line 4 in Algorithm\cite{LDGM} finds a vector $e_{i}^* \in \mathcal{E}$ such that

$$f(x_t + e_i^*) - f(x_t) \geq \frac{1}{l} \left( f(v^*_t) - f(x_t) - \sum_{k=1}^{l/\gamma} (\epsilon_{t,i}^* - \epsilon_{t,ik}^*) \right).$$

**Proof.** The proof is similar to that of Lemma\cite{LDGM}. Let $e_{i}^* \in \arg \max_{e \in \mathcal{E}} f(x_t + \gamma e_i) + \epsilon_{t,i}$. By the monotonicity and DR-submodularity of $f$, we can get

$$f(v^*_t) - f(x_t) \leq f(x_t + v^*_t) - f(x_t) = f \left( x_t + \sum_{j=1}^{l/\gamma} \gamma e_{ij} \right) - f(x_t)$$

9
We can see that as

\[ \sum_{k=1}^{t/\gamma} f\left(x_t + \sum_{j=1}^{k} \gamma e_{i_j} \right) - f\left(x_t + \sum_{j=1}^{k-1} \gamma e_{i_j} \right) \leq \sum_{k=1}^{t/\gamma} f(x_t + \gamma e_{i_k}) - f(x_t) \]

\[ = \sum_{k=1}^{t/\gamma} f(x_t + \gamma e_{i_k}) + \epsilon_{t,i_k} - \epsilon_{t,i_k} - f(x_t) \leq \frac{1}{\gamma} \cdot (f(x_t + \gamma e_{i_k}) - f(x_t)) + \sum_{k=1}^{t/\gamma} (\epsilon_{t,i_k} - \epsilon_{t,i_k}). \]

(1)

Where the last inequality is due to the selection procedure, i.e. line 4 in Algorithm 1.

Note that we assume \( \gamma \) is an integer that greater than 1, thus we have

\[ \frac{(f(x_t + \gamma e_{i_k}^*) - f(x_t))}{\gamma} = \sum_{j=1}^{\gamma} f(x_t + j e_{i_k}) - f(x_t + (j-1) e_{i_k}) \]

\[ \leq \sum_{j=1}^{\gamma} f(x_t + e_{i_k}^*) - f(x_t) = f(x_t + e_{i_k}^*) - f(x_t) \]

(2)

Where the inequality is due to DR-submodularity. By combining (1) and (2), we have the proof.

By combining Lemma 4 and 7, we can follow the proof of Theorem 1 to derive the approximation guarantee of LDGM under noise, i.e., Theorem 4.

**Theorem 4.** The LDGM algorithm under noise finds a solution \( x_t \in P \) with

\[ f(x_t) \geq (1 - 1/e) \cdot OPT - \frac{mDL(1 - 1/e)^2}{l} - \frac{1}{l} \sum_{t=0}^{l-1} (1 - 1/l)^{l-1-t} \sum_{k=1}^{l/\gamma} (\epsilon_{t,i_k} - \epsilon_{t,i_k}). \]

To compare FW with LDGM under noise, we compare the last term of their approximation guarantees in Theorems 3 and 4 which are incurred by noise. Let \( Err_{t}^{FW} := \frac{1}{l} \sum_{t=0}^{l-1} (\epsilon_{t,i_k} - \epsilon_{t,i_k}) \) and \( Err_{t}^{LDGM} := \frac{1}{l} \sum_{t=0}^{l/\gamma} (\epsilon_{t,i_k} - \epsilon_{t,i_k}) \). That is, we are to compare \( Err_{t}^{FW} \) with \( Err_{t}^{LDGM} \). As LDGM uses the function value and FW uses the gradient, they are not directly comparable. Thus, we consider the case where only noisy function value can be obtained, and we use forward difference to estimate the gradient while keeping the two algorithms run with a similar time complexity.

Let \( f(x) \) and \( F(x) \) denote the exact and noisy function value, respectively. We assume that the additive noise model: \( f(x) - \epsilon \leq F(x) \leq f(x) + \epsilon \). For LDGM, we easily have

\[ |Err_{t}^{LDGM}| \leq 2\epsilon/\gamma \]

For FW, we use forward difference as the estimate of gradient: for each \( i \in [n] \),

\[ \frac{\partial f(x)}{\partial x_i} \approx \left\lceil \frac{f(x + \alpha \chi_i) - f(x)}{\alpha} \right\rceil = \frac{F(x + \alpha \chi_i) - F(x)}{\alpha}. \]

Since

\[ \left| \frac{F(x + \alpha \chi_i) - F(x)}{\alpha} - \frac{f(x + \alpha \chi_i) - f(x)}{\alpha} \right| \leq \frac{2\epsilon}{\alpha} \]

and

\[ \left| \frac{f(x + \alpha \chi_i) - f(x)}{\alpha} - \frac{\partial f(x)}{\partial x_i} \right| = \left| \frac{\partial f(x)}{\partial x_i} - \frac{\partial f(x)}{\partial x_i} \right| \leq L \left| \xi - x \right| < La, \]

where \( L \) is the Lipschitz condition parameter, we have \( |(\epsilon_{k})_i| = \left| \frac{\partial f(x)}{\partial x_i} - \frac{F(x + \alpha \chi_i) - F(x)}{\alpha} \right| < \frac{2\epsilon}{\alpha} + La < 2\sqrt{2\epsilon L} \). Thus,

\[ |Err_{t}^{FW}| = \left| \frac{1}{l} \sum_{k=1}^{n} (u_{t_k}^* - u_t)_i \cdot (\epsilon_{k})_i \right| < \frac{2nD\sqrt{2\epsilon L}}{l}. \]

We can see that as \( \epsilon \) decreases, \( Err_{t}^{FW} \) shrinks with \( O(\sqrt{\epsilon}) \) speed, while \( Err_{t}^{LDGM} \) shrinks with \( O(\epsilon) \) speed. Also, by setting large \( \gamma \), we can reduce impacts from noises. This implies that LDGM may be more robust to noise.
5 Empirical Study

In this section, we empirically investigate the performance of LDGM on the budget allocation problem with continuous assignments [6]. Let a bipartite graph $G = (V; T; E)$ represent a social network, where each source node in $V$ is a marketing channel, each target node in $T$ is a customer, and $E \subseteq V \times T$ is the edge set. The goal of budget allocation is to distribute the budget $k$ among the source nodes such that the expected number of target nodes that get activated is maximized. The allocation of the budget can be represented by a vector $x \in \mathcal{X}$, where $x_i$ is the budget allocated on $v_i \in V$. Each source node $v_i$ ($i \in [n]$) has a probability $p_i \in [0, 1]$, and the probability that a target node $t \in T$ gets activated is $f_t(x) = 1 - \prod_{(v_i, t) \in E} (1 - p_i)^{x_i}$. By the linearity of expectation, the expected number of active target nodes is $\sum_{t \in T} f_t(x)$, which is monotone and DR-submodular [6].

The traditional budget setting is a size constraint, i.e., $|x| \leq B$ and $x \geq 0$, which is a simple polytope. We here generalize it to a convex polytope $P$.

We compare LDGM with the gradient-based algorithms, FW [6] and SCG [17]. The number of iterations is set to 40 for all algorithms, as using more than 40 iterations will not bring much improvement. For SCG, the step size at $t$-th iteration is set to be $\frac{4}{(t+8)^{\gamma}}$. To empirically compare these algorithms on budget allocation, we use a bipartite graph $G = (V; T; E)$ with 50 source nodes, 80 target nodes and 120 edges. For the polytope constraint, we first uniformly randomly choose the vertex set $E$ with 80 vertices, and then let the polytope $P = k \cdot \text{conv}(E)$, where $k$ is a parameter we use to control how large is $P$. The 120 edges between source nodes and target nodes are randomly generated, and we select each $p_i$ uniformly from $[0, 0.4]$.

We then consider two different settings. For the noiseless case (a), LDGM has access to the exact function value, and FW and SCG have access to the exact gradient. For LDGM, the lookahead parameter $\gamma$ is set to be 1. We also enumerate vertices to choose the best vertex as a baseline. For the noisy case (b), we focus on the situation that only noisy function values can be obtained. We assume additive noise $\epsilon$ uniformly chosen in $[-0.5, 0.5]$, i.e., instead of returning $f(x)$, a function call returns $f(x) + \epsilon$. As we want to keep the algorithms with the same computation time, we use forward difference with step size $a$ to estimate the gradient. Note that the choice of $a$ is crucial to have a good estimation, and the tuning process is time-consuming. We tuned $a$ to be 2.8 to let SCG and FW perform nearly the best. In the noise environment, the lookahead parameter $\gamma$ is set to be 8 in LDGM.

The results are plotted in Figure 1. We can observe that if without noise, LDGM, FW and SCG perform nearly the same, outperforming the best vertex baseline. When noise is introduced, LDGM is better than all the other algorithms. Note that SCG is always better than FW under noise, which is due to the averaging technique in the update procedure of SCG.

6 Conclusion

In this paper, we propose a derivative-free algorithm LDGM for maximizing monotone (weakly) DR-submodular continuous functions under a convex polytope constraint. We prove that LDGM can achieve the same approximation guarantee as the best previous gradient-based algorithm. We
also show that LDGM can be more robust to noise than gradient-based algorithms. Experiments on budget allocation show the effectiveness of LDGM.

References

[1] N. Alon, I. Gamzu, and M. Tennenholtz. Optimizing budget allocation among channels and influencers. In Proceedings of the 21st International Conference on World Wide Web (WWW’12), pages 381–388, Lyon, France, 2012.

[2] D. Avis and K. Fukuda. Reverse search for enumeration. *Discrete Applied Mathematics, 65*(1-3):21–46, 1996.

[3] F. Bach. Submodular functions: From discrete to continuous domains. *arXiv:1511.00394*, 2015.

[4] F. Bach. Efficient algorithms for non-convex isotonic regression through submodular optimization. *arXiv:1707.09157*, 2017.

[5] A. A. Bian, K. Levy, A. Krause, and J. M. Buhmann. Continuous DR-submodular maximization: Structure and algorithms. In *Advances in Neural Information Processing Systems 30 (NIPS’17)*, pages 486–496, Long Beach, CA, 2017.

[6] A. A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS’17)*, pages 111–120, Fort Lauderdale, FL, 2017.

[7] E. M. Bronstein. Approximation of convex sets by polytopes. *Journal of Mathematical Sciences, 153*(6):727–762, 2008.

[8] C. Chekuri, T.S. Jayram, and J. Vondráč. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS’15)*, pages 201–210, Rehovot, Israel, 2015.

[9] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[10] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM, 45*(4):634–652, 1998.

[11] P. M. Gruber. Aspects of approximation of convex bodies. In *Handbook of Convex Geometry, Part A*, pages 319–345. 1993.

[12] H. Hassani, M. Soltanolkotabi, and A. Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems 30 (NIPS’17)*, pages 5843–5853, Long Beach, CA, 2017.

[13] M. Kapralov, I. Post, and J. Vondráč. Online submodular welfare maximization: Greedy is optimal. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA’13)*, pages 1216–1225, New Orleans, LA, 2013.

[14] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’03)*, pages 137–146, Washington, DC, 2003.

[15] A. Krause and D. Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, February 2014.

[16] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research, 9*:235–284, 2008.

[17] A. Mokhtari, H. Hassani, and A. Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS’18)*, pages 1886–1895, Playa Blanca, Spain, 2018.

[18] G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research, 3*(3):177–188, 1978.

[19] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming, 14*(1):265–294, 1978.
Algorithm 2

Algorithm 2 LDGM with the Generalized Greedy Algorithm

Input: A monotone function $f: \mathcal{X} \to \mathbb{R}$, and an orthogonal set $\mathcal{E} = \{ \frac{1}{k}x \mid x \in \text{Frontier}(E) \}$

Parameter: Number $l$ of steps

Output: $x \in \mathcal{P}$, where $\mathcal{P} = \text{conv}(E)$

1: $x_0 := 0$ and $t := 0$
2: $x^* := \arg\max_{e \in \mathcal{E}} f(l e)$
3: while True do
4:   $(e^*, k) := \arg\max_{(e,j)} \left\{ \frac{f(x_t + j e) - f(x_t)}{j} \mid e \in \mathcal{E}, j \in \mathbb{Z}_+ : j \leq l - \frac{\langle x_t, e \rangle}{\langle e, e \rangle} \right\}$
5:   if $t + k > l$ then
6:      $x_t := x_t + (l-t)e^*$
7:      return $\arg\max_{x \in \{x_t, x^*\}} f(x)$
8:   else
9:      $x_{t+k} := x_t + ke^*$
10:     $t := t + k$
11: end if
12: end while

To prove Theorem 2, we need the following lemma, which is about the average gain by adding multiple copies of one vector from $\mathcal{E}$ in each iteration of the algorithm.

Lemma 8. Let $v^*$ be the best solution one can achieve using $l$ vectors from $\mathcal{E}$ which is an orthogonal set. Let $x$ be the solution obtained in one iteration of Algorithm 2 and let $(e^*, k)$ denote the combination obtained at line 4 in the next iteration. Then, we have

$$\frac{f(x + ke^*) - f(x)}{k} \geq \frac{\alpha}{l} (f(v^*) - f(x)).$$

Proof. Suppose $\mathcal{E} = \{e_1, ..., e_m\}$. Let us denote $e_{it}$ as the $t$-th entry at $e_i$. For any $e \in \mathcal{E} \subseteq \mathbb{R}^n_+$, we have $e \geq 0$, i.e., any entry of $e$ is not smaller than $0$. Since $\mathcal{E}$ is orthogonal (i.e., $\forall i, j : \langle e_i, e_j \rangle = 0$), we can get $\forall t \in [n], e_{it}$ and $e_{jt}$ cannot both be larger than $0$; otherwise $(e_i, e_j) = \sum_{t=1}^n e_{it} \cdot e_{jt} > 0$, where makes a contradiction. Thus, the coordinate-wise maximum ‘$\lor$’ is equivalent to ‘$+$’ for $e_i$ and $e_j$, i.e., $e_{i} \lor e_{j} = e_{i} + e_{j}$. Let $v = x \lor v^* - x$. As $x$ is obtained by line 9 of Algorithm 2, we can represent $v$ by $\sum_{i=1}^m l_i e_i$, where $l_i \in \mathbb{Z}_+$. Let $I = \{i \mid l_i > 0\}$, then we have $v = \sum_{i \in I} l_i e_i$. We use $I(1 : j)$ to denote the set of the first $j$ elements in $I$, and $I_j$ to denote the $j$-th element in $I$.

$$f(v^*) - f(x) \leq f(x \lor v^*) - f(x) = f\left(x + \sum_{i \in I} l_i e_i\right) - f(x)$$

$$= \sum_{j=1}^{\lfloor l \rfloor} f\left(x + \sum_{i \in I(1 : j)} l_i e_i\right) - f\left(x + \sum_{i \in I(1 : j-1)} l_i e_i\right)$$
We then analyze the margin
\[
\sum_{j=1}^{[|I|]} \frac{f(x + \sum_{i \in I(j)} l_i e_i) - f(x + \sum_{i \in I(j-1)} l_i e_i)}{l_{I_j}} - f(x + \sum_{i \in I(j-1)} l_i e_i)
\]
\[
\leq \frac{1}{\alpha} \sum_{j=1}^{[|I|]} \frac{f(x + l_{I_j} e_{I_j}) - f(x)}{l_{I_j}} \leq \frac{1}{\alpha} \sum_{j=1}^{[|I|]} \frac{f(x + ke^*) - f(x)}{k} \leq \frac{l}{\alpha} \frac{f(x + ke^*) - f(x)}{k},
\]
where the first inequality is by the monotonicity of \( f \), the second inequality is by the definition of submodularity ratio (i.e., Definition 2) and the orthogonality of \( E \), the third inequality is derived by
\[
\arg\max_{x} f(x + ke^*) \quad \text{for } x \quad \text{and } \quad \arg\max_{x} f(x + ke^*) - f(x) \quad \text{for } x.
\]
However, applying the latter one, we have
\[
\sum_{j=1}^{[|I|]} l_{I_j} \leq l.
\]
Note that Algorithm 1 terminates when line 5 is satisfied, thus we have
\[
\sum_{j=1}^{[|I|]} l_{I_j} \leq l.
\]
Let us denote \( x^{e} \) as the solution obtained by Algorithm 2 after \( s \) iterations. As \( x^{(0)} = 0 \). Let \( (e^{(s)}, k^{(s)}) \) be the combination returned by line 4 in the \( s \)-th iteration of Algorithm 2. Assume that the algorithm terminates after \( t+1 \) iterations. According to Lemma 8, we have, for any \( s \leq t \),
\[
\frac{f(x^{(s+1)}) - f(x^{(s)})}{k^{(s+1)}} \geq \frac{\alpha}{l} (f(v^*) - f(x^{(s)})).
\]
By rearranging the above inequality, we get
\[
f(x^{(s+1)}) \geq \frac{\alpha k^{(s+1)}}{l} f(x^{(s)}) + \left( 1 - \frac{\alpha k^{(s+1)}}{l} \right) f(x^{(s)}).
\]
Next we prove by induction that \( \forall 0 \leq s \leq t \) : \( f(x^{(s)}) \geq (1 - \frac{\alpha}{s+1} \frac{k^{(s)}}{l}) f(x^{(s)}) \). It trivially holds for \( s = 0 \). Assume the inequality holds for \( s \). Then, we have
\[
f(x^{(s+1)}) \geq \frac{\alpha k^{(s+1)}}{l} f(x^{(s)}) + \left( 1 - \frac{\alpha k^{(s+1)}}{l} \right) f(x^{(s)})
\]
\[
\geq \left( 1 - \frac{\alpha}{s+1} \frac{k^{(s+1)}}{l} \right) f(x^{(s)})
\]
\[
\geq \left( 1 - \frac{\alpha}{s+1} \frac{k^{(s+1)}}{l} \right) f(x^{(s)}).
\]
where the last inequality is derived by applying the AM-GM inequality. Thus, for \( x^{(t+1)} \), we have
\[
f(x^{(t+1)}) \geq \left( 1 - \frac{\alpha}{t+1} \frac{k^{(t)}}{l} \right) f(x^{*}).
\]
Note that Algorithm 2 terminates when line 5 is satisfied, thus we have \( \sum_{i=1}^{t} k^{(i)} \leq l \). By applying the latter one, we have
\[
f(x^{(t+1)}) \geq \left( 1 - \frac{\alpha}{t+1} \frac{k^{(t)}}{l} \right) f(x^{*}) \geq (1 - e^{-\alpha}) f(x^*).
\]
However, \( x^{(t+1)} \) is not a feasible solution. Note that the final output solution of Algorithm 2 is
\[
\arg\max_{x \in \{x^{(t)}, x^*\}} f(x), \quad \text{where } x^* = \arg\max_{x \in E} f(x).
\]
We can represent \( x^{(t+1)} \) as
\[
f(x^{(t+1)}) = f(x^{(t)}) + \left( f(x^{(t+1)}) - f(x^{(t)}) \right).
\]
We then analyze the margin \( f(x^{(t+1)}) - f(x^{(t)}) = f(x^{(t)}) + k^{(t+1)} e^{(t+1)} - f(x^{(t)}) \). As \( E \) is orthogonal here, \( x^{(t)} \) can be represented by one unique combination of \( e_i \in E \). Suppose \( x^{(t)} \) already had \( k' \) copies of \( e^{(t+1)} \), i.e., \( \langle x^{(t)} - k' e^{(t+1)}, e^{(t+1)} \rangle = 0 \). Thus, by applying the definition of \( \alpha \), we have
\[
f(x^{(t)} + k^{(t+1)} e^{(t+1)}) - f(x^{(t)}) \]
\[
\begin{align*}
\leq & \frac{1}{\alpha} \left( f(k^{(t+1)} + k')e^{(t+1)} - f(k'e^{(t+1)}) \right) \\
\leq & \frac{1}{\alpha} \left( f(le^{(t+1)}) - f(k'e^{(t+1)}) \right) \\
\leq & \frac{1}{\alpha} \left( f(x^*) - f(k'e^{(t+1)}) \right) \leq \frac{1}{\alpha} f(x^*),
\end{align*}
\]

where the second inequality is by \( k^{(t+1)} \leq l - \frac{\langle x_t, e(t+1) \rangle}{\alpha(t+1)} = l - k' \), and the third inequality is by \( x^* = \arg \max_{e \in E} f(le) \). Thus, we get

\[
f(x^{(t+1)}) \leq f(x^{(t)}) + \frac{1}{\alpha} f(x^*) \leq \frac{1}{\alpha} (f(x^{(t)}) + f(x^*)),
\]

where the last inequality is by \( \alpha \in [0, 1] \). Since we have shown that \( f(x^{(t+1)}) \geq (1 - e^{-\alpha}) f(x^*) \), we get

\[
f(x^{(t)}) + f(x^*) \geq \alpha (1 - e^{-\alpha}) f(x^*),
\]

which implies that \( \max\{f(x^{(t)}), f(x^*)\} \geq (\alpha/2)(1 - e^{-\alpha}) f(x^*) \). Therefore, by applying \( f(x^*) \geq OPT - \frac{mDL}{l} \), the theorem holds. \( \square \)