Optimal approximation for unconstrained non-submodular minimization

Marwa El Halabi Stefanie Jegelka

Abstract
Submodular function minimization is well studied, and existing algorithms solve it exactly or up to arbitrary accuracy. However, in many applications, such as structured sparse learning or batch Bayesian optimization, the objective function is not exactly submodular, but close. In this case, no theoretical guarantees exist. Indeed, submodular minimization algorithms rely on intricate connections between submodularity and convexity. We show how these relations can be extended to obtain approximation guarantees for minimizing non-submodular functions, characterized by how close the function is to submodular. We also extend this result to noisy function evaluations. Our approximation results are the first for minimizing non-submodular functions, and are optimal, as established by our matching lower bound.

1. Introduction
Many machine learning problems can be formulated as minimizing a set function $H$. This problem is in general NP-hard, and can only be solved efficiently with additional structure. One especially popular example of such structure is that $H$ is submodular, i.e., it satisfies the diminishing returns (DR) property: $H(A \cup \{i\}) - H(A) \geq H(B \cup \{i\}) - H(B)$, for all $A \subseteq B$, $i \in V \setminus B$. Several existing algorithms minimize a submodular $H$ in polynomial time, exactly or within arbitrary accuracy. Submodularity is a natural model for a variety of applications, such as image segmentation (Boykov & Kolmogorov 2004), data selection (Lin & Bilmes 2010), or clustering (Narasimhan et al. 2006). But, in many other settings, such as structured sparse learning, Bayesian optimization, and column subset selection, the objective function is not exactly submodular. Instead, it satisfies a weaker version of the diminishing returns property. An important class of such functions are $\alpha$-weakly DR-submodular functions, introduced in (Lehmann et al. 2006). The parameter $\alpha$ quantifies how close the function is to being submodular (see Section 2 for a precise definition). Furthermore, in many cases, only noisy evaluations of the objective are available. Hence, we ask: Do submodular minimization algorithms extend to such non-submodular noisy functions?

Non-submodular maximization, under various notions of approximate submodularity, has recently received a lot of attention (Das & Kempe 2011; Elenberg et al. 2018; Sakae 2019; Bian et al. 2017; Chen et al. 2017; Gattinir & Gomez-Rodriguez 2019; Harshaw et al. 2019; Kuhnle et al. 2018; Horel & Singer 2016; Hassidim & Singer 2018). In contrast, only few studies consider minimization of non-submodular set functions. Recent works have studied the problem of minimizing the ratio of two set functions, where one (Bai et al. 2016; Qian et al. 2017a) or both (Wang et al. 2019) are non-submodular. The ratio problem is related to constrained minimization, which does not admit a constant factor approximation even in the submodular case (Svitkina & Fleischer 2011). If the objective is approximately modular, i.e., it has bounded curvature, algorithmic techniques related to those for submodular maximization achieve optimal approximations for constrained minimization (Sviridenko et al. 2017; Iyer et al. 2013). Algorithms for minimizing the difference of two submodular functions were proposed in (Iyer & Bilmes 2012; Kawahara et al. 2015), but no approximation guarantees were provided. In this paper, we study the unconstrained non-submodular minimization problem

$$\min_{S \subseteq V} H(S) := F(S) - G(S),$$

where $F$ and $G$ are monotone (i.e., non-decreasing or non-increasing) functions, $F$ is $\alpha$-weakly DR-submodular, and $G$ is $\beta$-weakly DR-supermodular, i.e., $-G$ is $\beta^{-1}$-weakly DR-submodular. The definitions of weak DR-sub/supermodularity only hold for monotone functions, and thus do not directly apply to $H$. We show that, perhaps surprisingly, any set function $H$ can be decomposed into functions $F$ and $G$ that satisfy these assumptions, albeit with properties leading to weaker approximations when the function is far from being submodular.
A key strategy for minimizing submodular functions exploits a tractable tight convex relaxation that enables the use of convex optimization algorithms. But, this relies on the equivalence between the convex closure of a submodular function and the polynomial-time computable Lovász extension. In general, the convex closure of a set function is NP-hard to compute, and the Lovász extension is convex if and only if the set function is submodular. Thus, the optimization delicately relies on submodularity; generally, a tractable tight convex relaxation is impossible. Yet, in this paper, we show that for approximately submodular functions, the Lovász extension can be approximately minimized using a projected subgradient method (PGM). In fact, this strategy is guaranteed to obtain an approximate solution to Problem (1). This insight broadly expands the scope of submodular minimization techniques. In short, our main contributions are:

- the first approximation guarantee for unconstrained non-submodular minimization characterized by closeness to submodularity: PGM achieves a tight approximation of $H(S) \leq F(S^*)/\alpha - \beta G(S^*) + \epsilon$;
- an extension of this result to the case where only a noisy oracle of $H$ is accessible;
- a hardness result showing that improving on this approximation guarantee would require exponentially many queries in the value oracle model;
- applications to structured sparse learning and variance reduction in batch Bayesian optimization, implying the first approximation guarantees for these problems;
- experiments demonstrating the robustness of classical submodular minimization algorithms against noise and non-submodularity, reflecting our theoretical results.

2. Preliminaries

We begin by introducing our notation, the definitions of weak DR-submodularity/supermodularity, and by reviewing some facts about classical submodular minimization.

**Notation** Let $V = \{1, \cdots, d\}$ be the ground set. Given a set function $F : 2^V \to \mathbb{R}$, we denote the marginal gain of adding an element $i$ to a set $A$ by $F(i|A) = F(A \cup \{i\}) - F(A)$. Given a vector $x \in \mathbb{R}^d$, $x_i$ is its $i$-th entry and $\text{supp}(x) = \{i \in V | x_i \neq 0\}$ is its support set; $x$ also defines a modular set function as $x(A) = \sum_{i \in A} x_i$.

**Set function classes** The function $F$ is normalized if $F(\emptyset) = 0$, and non-decreasing (non-increasing) if $F(A) \leq F(B)$ ($F(A) \geq F(B)$) for all $A \subseteq B$. $F$ is submodular if it has diminishing marginal gains: $F(i|A) \geq F(i|B)$ for all $A \subseteq B$, $i \in V \setminus B$, modular if the inequality holds as an equality, and supermodular if $F(i|A) \leq F(i|B)$. Relaxing these inequalities leads to the notions of weak DR-submodularity/supermodularity introduced in (Lehmann et al., 2006) and (Bian et al., 2017), respectively.

**Definition 1** (Weak DR-sub/supermodularity). A set function $F$ is $\alpha$-weakly DR-submodular, with $\alpha > 0$, if

$$F(i|A) \geq \alpha F(i|B), \text{ for all } A \subseteq B, i \in V \setminus B.$$ 

Similarly, $F$ is $\beta$-weakly DR-supermodular, with $\beta > 0$, if

$$F(i|B) \geq \beta F(i|A), \text{ for all } A \subseteq B, i \in V \setminus B.$$ 

We say that $F$ is $(\alpha, \beta)$-weakly DR-modular if it satisfies both properties.

If $F$ is non-decreasing, then $\alpha, \beta \in (0, 1]$, and if it is non-increasing, then $\alpha, \beta \geq 1$. $F$ is submodular (supermodular) iff $\alpha = 1$ ($\beta = 1$) and modular iff both $\alpha = \beta = 1$.

The parameters $1-\alpha$ and $1-\beta$ are referred to as generalized inverse curvature (Bogunovic et al., 2018) and generalized curvature (Bian et al., 2017), respectively. They extend the notions of inverse curvature and curvature (Conforti & Cornuéjols, 1984) commonly defined for supermodular and submodular functions. These notions are also related to weakly sub-/supermodular functions (Das & Kempe, 2011; Bogunovic et al., 2018). Namely, the classes of weakly DR-sub-/super-/modular functions are respective subsets of the classes of weakly sub-/super-/modular functions (El Halabi et al., 2018, Prop. 8), (Bogunovic et al., 2018, Prop. 1), as illustrated in Figure 2. For a survey of other notions of approximate submodularity, we refer the reader to (Bian et al., 2017, Sect. 6).

**Submodular minimization** Minimizing a submodular set function $F$ is equivalent to minimizing a non-smooth convex function that is given by a continuous extension of $F$, i.e., a continuous interpolation of $F$ on the full hypercube $[0, 1]^d$. This extension, called the Lovász extension (Lovász, 1983), is convex if and only if $F$ is submodular.

**Definition 2** (Lovász extension). Given any normalized set function $F$, its Lovász extension $f_L : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f_L(s) = \sum_{k=1}^d s_{j_k} F(j_k|S_{k-1}),$$

where $s_{j_1} \geq \cdots \geq s_{j_k}$ are the sorted entries of $s$ in decreasing order, and $S_k = \{j_1, \cdots, j_k\}$.
Minimizing $f_L$ is equivalent to minimizing $F$. Moreover, when $F$ is submodular, a subgradient $\kappa$ of $f_L$ at any $s \in \mathbb{R}^d$ can be computed efficiently by sorting the entries of $s$ in decreasing order and taking $\kappa_k = F(S_k \setminus S_{k-1})$ for all $k \in V$ (Edmonds, 2003). This relation between submodularity and convexity allows for generic convex optimization algorithms to be used for minimizing $F$. However, it has been unclear how these relations are affected if the function is only approximately submodular. In this paper, we give an answer to this question.

3. Approximately submodular minimization

We consider set functions $H : 2^V \to \mathbb{R}$ of the form $H(S) = F(S) - G(S)$, where $F$ is $\alpha$-weakly DR-submodular, $G$ is $\beta$-weakly DR-supermodular, and both $F$ and $G$ are normalized non-decreasing functions. We later extend our results to non-increasing functions. We assume a value oracle access to $H$; i.e., there is an oracle that, given a set $S \subseteq V$, returns the value $H(S)$. Note that $H$ itself is in general not weakly DR-submodular. Interestingly, any set function can be decomposed in this form.

**Proposition 1.** Given a set function $H$, and $\alpha, \beta \in (0, 1]$ such that $\alpha \beta < 1$, there exists a non-decreasing $\alpha$-weakly DR-submodular function $F$, and a non-decreasing $(\alpha, \beta)$-weakly DR-modular function $G$, such that $H(S) = F(S) - G(S)$ for all $S \subseteq V$.

**Proof sketch.** This decomposition builds on the decomposition of $H$ into the difference of two non-decreasing submodular functions (Iyer & Bilmes, 2012). We start by choosing any function $G'$ which is non-decreasing $(\alpha, \beta)$-weakly DR-modular, and is strictly $\alpha$-weakly DR-submodular, i.e., $\epsilon_{G'} = \min_{i \in V, A \subseteq B \subseteq V \setminus i} G'(i|A) - \alpha G'(i|B) > 0$. It is not possible to choose $G'$ such that $\alpha = \beta = 1$ (this would imply $G'(i|B) \geq G'(i|A) > G'(i|B)$). We then construct $F$ and $G$ based on $G'$.

Let $\epsilon_H = \min_{i \in V, A \subseteq B \subseteq V \setminus i} H(i|A) - \alpha H(i|B) < 0$ be the violation of $\alpha$-weak DR-submodularity of $H$; we may use a lower bound $\epsilon_H \leq \epsilon_H$. We define $F'(S) = H(S) + \frac{\epsilon_H}{\epsilon_{G'}} G'(S)$. $F'$ is not necessarily non-decreasing. To correct for that, let $V^- = \{ i : F'(i|V \setminus i) < 0 \}$ and define $F(S) = F'(S) - \sum_{i \in V^-} F'(i|V \setminus i)$. We can show that $F$ is non-decreasing $\alpha$-weakly DR-submodular.

We also define $G(S) = \frac{\epsilon_{G'}}{\epsilon_H} G'(S) - \sum_{i \in S \cap V^+} F'(i|V \setminus i)$, then $G$ is non-decreasing $(\alpha, \beta)$-weakly DR-modular, and $H(S) = F(S) - G(S)$.

**Proposition 1** generalizes the result of (Cunningham, 1983; Theorem 18) showing that any submodular function can be decomposed into the difference of a non-decreasing submodular function and a non-decreasing modular function. When $H$ is submodular, the decomposition in Proposition 1 recovers the one from (Cunningham, 1983) by simply choosing $\alpha = \beta = 1$. The resulting violation of submodularity is $\epsilon_H = 0$, and $G'$ is not needed.

Computing such a decomposition is not required to run PGM for minimization; it is only needed to evaluate the corresponding approximation guarantee. The construction in the above proof uses the maximum violation $\epsilon_H$ of $\alpha$-weak DR-submodularity of $H$, which is NP-hard in general. However, when $\epsilon_H$ or a lower bound of it is known, $F$ and $G$ can be obtained in polynomial time, for a suitable choice of $G'$. Proposition 2 provides a valid choice of $G'$ for $\alpha = 1$.

**Proposition 2.** Given $\beta \in (0, 1)$, let $G'(S) = g(|S|)$ where $g(x) = \frac{1}{2}ax^2 + (1 - \frac{1}{2}a)x$ with $a = \frac{\beta + 1}{\beta}$. Then $G'$ is non-decreasing $(1, \beta)$-weakly DR-modular, and is strictly submodular, with $\epsilon_{G'} = \min_{i \in V, A \subseteq B \subseteq V \setminus i} G'(i|A) - G'(i|B) = -a > 0$.

The lower bound on $\epsilon_H$ and the choice of $\alpha, \beta$ and $G'$ will affect the approximation guarantee on $H$, as we clarify later. When $H$ is far from being submodular, it may not be possible to choose $G'$ to obtain a non-trivial guarantee. However, many important non-submodular functions do admit a decomposition which leads to non-trivial bounds. We call such functions approximately submodular, and provide some examples in Section 4.

In what follows, we establish a connection between approximate submodularity and approximate convexity, which allows us to derive a tight approximation guarantee for PGM on Problem (1). All omitted proofs are in the Supplement.

3.1. Convex relaxation

When $H$ is not submodular, the connections between its Lovász extension and tight convex relaxation for exact minimization, outlined in Section 2, break down. However, Problem (1) can still be converted to a non-smooth convex optimization problem, via a different convex extension. Given a set function $H$, its convex closure $h^-$ is the pointwise largest convex function from $[0, 1]^d$ to $\mathbb{R}$ that always lower bounds $H$. Intuitively, $h^-$ is the tightest convex extension of $H$ on $[0, 1]^d$. The following equivalence holds (Dughmi, 2009; Prop. 3.23):

$$\min_{S \subseteq V} H(S) = \min_{s \in [0, 1]^d} h^-(s). \tag{2}$$

Unfortunately, evaluating and optimizing $h^-$ for a general set function is NP-hard (Vondrák, 2007). The key property that makes Problem (2) efficient to solve when $H$ is submodular is that its convex closure then coincides with its tractable Lovász extension, i.e., $h^- = h_L$. This equivalence no longer holds if $H$ is only approximately submodular. But, in this case, a weaker key property holds: Lemma 1.
shows that the Lovász extension approximates the convex closure \( h^+ \), and that the same vectors that served as its subgradients in the submodular case can serve as approximate subgradients to \( h^- \).

**Lemma 1.** Given a vector \( s \in [0, 1]^d \) such that \( s_{j_1} \geq \cdots \geq s_{j_d} \), we define \( \kappa \) such that \( \kappa_{j_k} = H(j_k|S_{k-1}) \) where \( S_k = \{j_1, \ldots, j_k\} \). Then, \( h_L(s) = \kappa^T s \geq h^-(s) \), and

\[
\kappa(A) = \frac{1}{2} F(A) - \beta G(A) \quad \text{for all} \ A \subseteq V, \\
\kappa^T s' \leq \frac{1}{2} f^-((s')) + \beta(-g)^-(s') \quad \text{for all} \ s' \in [0, 1]^d.
\]

To prove Lemma 1, we use a specific formulation of the convex closure \( h^- \) (El Halabi 2018, Def. 20):

\[
h^- = \max_{\kappa \in \mathbb{R}^d, \rho \in \mathbb{R}} \{ \kappa^T s + \rho : \kappa(A) + \rho \leq H(A), \forall A \subseteq V \},
\]

and build on the proof of Edmonds’ greedy algorithm [Edmonds 2003]. We can view the vector \( \kappa \) in Lemma 1 as an approximate subgradient of \( h^- \) at \( s \) in the following sense:

\[
\frac{1}{2} f^-((s')) + \beta(-g)^-(s') \geq h^-((s)) + (\kappa^T - s), \quad \forall s' \in [0, 1]^d.
\]

Lemma 1 also implies that the Lovász extension \( h_L \) approximates the convex closure \( h^- \) in the following sense:

\[
h^-((s)) \leq h_L((s)) \leq \frac{1}{2} f^-((s)) + \beta(-g)^-(s), \quad \forall s \in [0, 1]^d.
\]

We can thus say that \( h_L \) is approximately convex in this case. This key insight allows us to approximately minimize \( h^- \) via convex optimization algorithms.

### 3.2. Algorithm and approximation guarantees

Equipped with the approximate subgradients of \( h^- \), we can now apply an approximate projected subgradient method (PGM). Starting from an arbitrary \( s_0 \in [0, 1]^d \), PGM iteratively updates \( s_{t+1} = \Pi_{[0, 1]^d}(s_t - \eta \kappa^t) \), where \( \kappa^t \) is the approximate subgradient at \( s_t \) from Lemma 1 and \( \Pi_{[0, 1]^d} \) is the projection onto \([0, 1]^d\). We set the step size \( \eta = \frac{R}{L \sqrt{T}} \), where \( L = F(V) + G(V) \) is the Lipschitz constant, i.e., \( \| \kappa^t \|_2 \leq L \) for all \( t \), and \( R = 2\sqrt{d} \) is the domain radius \( \| s - s^* \|_2 \leq R \).

**Theorem 1.** After \( T \) iterations of PGM, \( \hat{s} \in \arg \min_{t \in \{1, \ldots, T\}} h_L(s^t) \) satisfies:

\[
h^-((\hat{s})) \leq h_L((\hat{s})) \leq \frac{1}{2} f^-((s^*)) + \beta(-g)^-(s^*) + \frac{RL}{\sqrt{T}},
\]

where \( s^* \) is an optimal solution of \( \min_{s \in [0, 1]^d} h^-(s) \).

Importantly, the algorithm does not need to know the parameters \( \alpha \) and \( \beta \), which can be hard to compute in practice. In fact, its iterates are exactly the same as in the submodular case. Theorem 1 provides an approximate fractional solution \( \hat{s} \in [0, 1]^d \). To round it to a discrete solution, Corollary 1 shows that it is sufficient to pick the superlevel set of \( \hat{s} \) with the smallest \( H \) value.

**Corollary 1.** Given the fractional solution \( \hat{s} \) in Theorem 1 let \( \hat{S}_k = \{j_1, \ldots, j_k\} \) such that \( \hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_d} \), and \( S_0 = \emptyset \). Then, \( \hat{S} = \arg \min_{k \in \{0, \ldots, d\}} H(\hat{S}_k) \) satisfies

\[
H(\hat{S}) \leq \frac{1}{2} F(S^*) - \beta G(S^*) + \frac{RL}{\sqrt{T}},
\]

where \( S^* \) is an optimal solution of Problem 1.

To obtain a set that satisfies \( H(\hat{S}) \leq F(S^*)/\alpha - \beta G(S^*) + \epsilon \), we thus need at most \( O(dL^2/\epsilon^2) \) iterations of PGM, where the time per iteration is \( O(d \log d + d \text{ EO}) \), with EO being the time needed to evaluate \( H \) on any set. Moreover, the techniques from [Chakrabarty et al. 2017, Axelrod et al. 2019] for accelerating the runtime of stochastic PGM to \( O(d \text{ EO}/\epsilon^2) \) can be extended to our setting.

If \( F \) is regarded as a cost and \( G \) as a revenue, this guarantee states that the returned solution achieves at least a fraction \( \beta \) of the revenue of the optimal solution, by paying at most a \( 1/\alpha \)-multiple of the cost. The quality of this guarantee depends on \( F, G \) and their parameters \( \alpha, \beta \); it becomes vacuous when \( F(S^*)/\alpha \geq \beta G(S^*) \). If \( F \) is submodular, Problem 1 reduces to submodular minimization and Corollary 1 recovers the guarantee \( H(\hat{S}) \leq H(S^*) + RL/\sqrt{T} \).

**Remark 1.** The upper bound in Corollary 1 still holds if the worst case parameters \( \alpha, \beta \) are instead replaced by \( \alpha_T = \frac{1}{T} \sum_{t=1}^{T} F(S^*_t)/\alpha \) and \( \beta_T = \frac{1}{T} \sum_{t=1}^{T} G(S^*_t)/\beta \), where \( (\kappa^t)_{j_k} = F(j_k|S^*_{k-1}) \) and \( (\kappa^t)_{j_k} = G(j_k|S^*_{k-1}) \). This refined upper bound yields improvements if only few of the relevant submodularity inequalities are violated.

All results in this section extend to the case where \( F \) and \( G \) are non-increasing functions.

**Corollary 2.** Given \( H(S) = F(S) - G(S) \), where \( F \) and \( G \) are non-increasing functions with \( F(V) = G(V) = 0 \), we run PGM with \( H(S) = H(V \backslash S) \) for \( T \) iterations. Let \( \tilde{s} \in \arg \min_{t \in \{1, \ldots, T\}} h_L(s^t) \) and \( \tilde{S} = V \backslash \tilde{s} \), where \( \tilde{S} \) is the superlevel set of \( \tilde{s} \) with the smallest \( H \) value, then

\[
H(\tilde{S}) \leq \frac{1}{2} F(S^*) - \frac{1}{\beta} G(S^*) + \frac{RL}{\sqrt{T}},
\]

where \( S^* \) is an optimal solution of Problem 1.

For a general set function \( H \), using \( F \) and \( G \) from the decomposition in Proposition 1, yields in Corollary 1

\[
H(\tilde{S}) \leq \frac{1}{2} H(S^*) + (\frac{1}{\alpha} - \beta) \left( \frac{\rho}{c_G} G'(S^*) - \sum_{i \in S^*} F'(i(V \backslash i)) \right) + \epsilon,
\]

where \( \rho \) is a lower bound on the violation of \( \alpha \)-weak DR-submodularity of \( H \), \( F' \) and \( G' \) are the auxiliary functions used to construct \( F \) and \( G \), and \( c_G \) is the strict \( \alpha \)-weak DR-submodularity of \( G' \) (see proof of Proposition 1 for precise definitions). It is clear that a larger lower bound \( \rho \)
worsens the upper bound on $H(\hat{S})$. Moreover, the choice of $G'$ affects the bound: ideally, we want to choose $G'$ to minimize $G'(|S^*|)$, and maximize the quantities $\alpha$, $c_{G'}$, and $\beta$, which characterize how submodular and supermodular $G'$ is, respectively. However, a larger $\alpha$ leads to a larger $|c_{G'}|$ and smaller $c_{G'}$, and a larger $c_{G'}$ would result in a smaller $\beta$, and vice versa. The best choice of $G'$ will depend on $H$. In Appendix [B.4] we provide an example showing that the approximation guarantees in Corollary [1] and 3 are tight, i.e., they cannot be improved for PGM, even if $F$ and $G$ are weakly DR-modular. Furthermore, in Section 3.4 we show that these approximation guarantees are optimal in general. Apart from the above results for general unconstrained minimization, our results also imply approximation guarantees for generalizing constrained submodular minimization to weakly DR-submodular functions. We discuss this extension in Appendix [A].

3.3. Extension to noisy evaluations

In many real-world applications, we do not have access to the objective function itself, but rather to a noisy version of it. Several works have considered maximizing noisy oracles of submodular (Horel & Singer, 2016; Singla et al., 2016; Hassidim & Singer, 2017, 2018) and weakly submodular (Qian et al., 2017b) functions. In contrast, to the best of our knowledge, minimizing noisy oracles of submodular functions was only studied in (Blais et al., 2018).

We address a more general setup where the underlying function $H$ is not necessarily submodular. We assume again that $F$ and $G$ are normalized and non-decreasing. The results easily extend to non-increasing functions as in Corollary [3]. We show in Proposition [3] that our approximation guarantee for Problem (1) continues to hold when we only have access to an approximate oracle $\hat{H}$. Essentially, $\hat{H}$ still allows to obtain approximate subgradients of $h^-$ in the sense of Lemma (1) but now with an additional additive error.

**Proposition 3.** Assume we have an approximate oracle $\hat{H}$ with input parameters $\epsilon, \delta \in (0, 1)$, such that for every $S \subseteq V$, $|H(\hat{S}) - H(S)| \leq \epsilon$ with probability $1 - \delta$. We run PGM with $H$ for $T$ iterations. Let $\hat{s} = \arg\min_{s \in \{1, \ldots, T\}} \hat{h}_L(s^\epsilon)$, and $\hat{S}_k = \{j_1, \ldots, j_k\}$ such that $\hat{s}_{j_1} \geq \cdots \geq \hat{s}_{j_k}$. Then $\hat{S} = \arg\min_{s \in \{0, \ldots, d\}} \hat{H}(\hat{S}_k) \in H(\hat{S}) \leq \frac{1}{\alpha} F(S^*) - \beta G(S^*) + \epsilon'$, with probability $1 - \delta'$, by choosing $\epsilon' = \frac{\epsilon}{8d}$, $\delta = \frac{8d\epsilon^2}{32d^2\alpha}$, and using $2Td \geq \epsilon H(\hat{S})$. Blais et al. (2018) consider the same setup for the special case of submodular $H$, and use the cutting plane method of Lee et al. (2015). Their runtime has better dependence $O(\log(1/\epsilon'))$ on the error $\epsilon'$, but worse dependence $O(d^3)$ on the dimension $d = |V|$, and their result needs oracle accuracy $\epsilon = O(\epsilon'^2/d^3)$. Hence, for large ground set sizes $d$, Proposition [3] is preferable. This proposition allows us, in particular, to handle multiplicative and additive noise in $H$.

**Proposition 4.** Let $\hat{H} = \xi H$ where the noise $\xi \geq 0$ is bounded by $|\xi| \leq \omega$ and is independently drawn from a distribution $D$ with mean $\mu > 0$. We define the function $\hat{H}_m$ as the mean of $m$ queries to $\hat{H}(S)$. $\hat{H}_m$ is then an approximate oracle to $\mu H$. In particular, for every $\delta, \epsilon \in (0, 1)$, taking $m = (\omega \mu H_{\text{max}}/\epsilon)^2 \ln(1/\delta)$ where $\mu H_{\text{max}} = \max_{S \subseteq V} H(S)$, we have for every $S \subseteq V$, $|\hat{H}_m(S) - \mu H(S)| \leq \epsilon$ with probability at least $1 - \delta$.

Propositions [3] and 4 imply that by using PGM with $\hat{H}_m$ and picking the superlevel set with the smallest $\hat{H}_m$ value, we can find a set $\hat{S}$ such that $H(\hat{S}) \leq F(S^*)/\alpha - \beta G(S^*) + \epsilon'$ with probability $1 - \delta'$, using $m = O((\omega \mu H_{\text{max}})^2 \ln(\frac{d^2}{\delta \epsilon^{\alpha\beta}}))$ samples, after $T = O((\sqrt{d} \mu H_{\text{max}}/\epsilon')^2)$ iterations, with $O(\frac{\omega}{\mu} (H_{\text{max}})^4 \ln(\frac{d^2}{\delta \epsilon^{\alpha\beta}}))$ total calls to $\hat{H}$. Note that $H_{\text{max}}$ is upper bounded by $F(V)$. This result provides a theoretical upper bound of the number of samples needed to be robust to bounded multiplicative noise. Much fewer samples are actually needed in practice, as illustrated in our experiments (Section 3.4). Using similar arguments, our results also extend to additive noise oracles $\bar{H} = H + \xi$.

3.4. Inapproximability Result

By Proposition [1], Problem (1) is equivalent to general set function minimization. Thus, solving it optimally or within any multiplicative approximation factor, i.e., $H(\hat{S}) \leq \gamma(d) H(S^*)$ for some positive polynomial time computable function $\gamma(d)$ of $d$, is NP-Hard (Trevisan, 2004; Iyer & Bilmes, 2012). Moreover, in the value oracle model, it is impossible to obtain any multiplicative constant factor approximation within a subexponential number of queries (Iyer & Bilmes, 2012). Hence, it is necessary to consider bicriteria-like approximation guarantees as we do.

We now show that our approximation results are optimal: in the value oracle model, no algorithm with a subexponential number of queries can improve on the approximation guarantees achieved by PGM, even when $G$ is weakly DR-modular.

**Theorem 2.** For any $\alpha, \beta \in (0, 1)$ such that $\alpha \beta < 1$, $d > 2$ and $\delta > 0$, there are instances of Problem (1) such that (no deterministic or randomized) algorithm, using less than exponentially many queries, can always find a solution $S \subseteq V$ of expected value at most $F(S^*)/\alpha - \beta G(S^*) - \delta$.

**Proof sketch.** Our proof technique is similar to Feige et al. (2011): We randomly partition the ground set into $V = C \cup D$, and construct a normalized set function $H$ whose
values depend only on \(k(S) = |S \cap C|\) and \(\ell(S) = |S \cap D|:\)
\[
H(S) = \begin{cases} 0 & \text{if } |k(S) - \ell(S)| \leq \epsilon d \\ \frac{2\epsilon}{\epsilon^2 - 1} & \text{otherwise}, \end{cases}
\]
for some \(\epsilon \in [1/d, 1/2]\). We use Proposition 1 to decompose \(H\) into the difference of a non-decreasing \(\alpha\)-weakly DR-submodular function \(F\), and a non-decreasing \((\alpha, \beta)\)-weakly DR-modular function \(G\). We argue that, with probability \(1 - 2\epsilon(\epsilon^2 - 1)\), any given query \(S\) will be “balanced”, i.e., \(|k(S) - \ell(S)| \leq \epsilon d\). Hence no algorithm can distinguish between \(H\) and the constant zero function, with subexponentially many queries. On the other hand, we have \(H(S^*) = \frac{2\epsilon}{\epsilon^2 - 1} < 0\), achieved at \(S^* = C\) or \(D\), and \(\frac{1}{\alpha} F(S^*) - \beta G(S^*) - \delta < 0\). Therefore, the algorithm cannot find a set with value \(H(S) \leq F(S^*)/\alpha - \beta G(S^*) - \delta\). \(\square\)

The approximation guarantees in Corollary 1 and 3 are thus optimal. In the above proof, \(G\) belongs to the smaller class of weakly DR-modular functions, but \(F\) not necessarily. Whether the approximation guarantee can be improved when \(F\) is also weakly DR-modular is left as an open question. Yet, the tightness result in Appendix A implies that such improvement cannot be achieved by PGM.

4. Applications

Several applications can benefit from the theory in this work. We discuss two examples here, where we show that the objective functions have the form of Problem 1 implying the first approximation guarantees for these problems. Other examples include column subset selection (Sviridenko et al., 2017) and Bayesian A-optimal experimental design (Bian et al., 2017), where \(F\) is the cardinality function, and \(G\) is weakly DR-supermodular with \(\beta\) depending on the inverse of the condition number of the data matrix.

4.1. Structured sparse learning

Structured sparse learning aims to estimate a sparse parameter vector whose support satisfies a particular structure, such as group-sparsity, clustering, tree-structure, or diversity (Obozinski & Bach, 2016; Kyrillidis et al., 2015). Such problems can be formulated as
\[
\min_{x \in \mathbb{R}^d} \ell(x) + \lambda F(\text{supp}(x)),
\]
where \(\ell\) is a convex loss function and \(F\) is a set function favoring the desirable supports. Existing convex methods propose to replace the discrete regularizer \(F(\text{supp}(x))\) by its “closest” convex relaxation (Bach, 2010; El Halabi & Cevher, 2015; Obozinski & Bach, 2016; El Halabi et al., 2018). For example, the cardinality regularizer \(|\text{supp}(x)|\) is replaced by the \(\ell_1\)-norm. This allows the use of standard convex optimization methods, but does not provide any approximation guarantee for the original objective function without statistical modeling assumptions. This approach is computationally feasible only when \(F\) is submodular (Bach, 2010) or can be expressed as an integral linear program (El Halabi & Cevher, 2015).

Alternatively, one may write Problem 1 as
\[
\min_{S \subseteq V} H(S) = \lambda F(S) - G(S),
\]
where \(G^\ell(S) = \ell(0) - \min_{\lambda \in \text{supp}(x)} \ell(x)\) is a normalized non-decreasing set function. Recently, it was shown that if \(\ell\) has restricted smoothness and strong convexity, \(G^\ell\) is weakly modular (Elenberg et al., 2018; Bogunovic et al., 2018; Sakaue, 2018). This allows for approximation guarantees of greedy algorithms to be applied to the constrained variant of Problem 1, but only for the special cases of a sparsity constraint (Das & Kempe, 2011; Elenberg et al., 2018) or some near-modular constraints (Sakaue, 2019).

In applications, however, the structure of interest is often better modeled by a non-modular regularizer \(F\), which may be submodular (Bach, 2010) or non-submodular (El Halabi & Cevher, 2015; El Halabi et al., 2018). Weak modularity of \(G^\ell\) is not enough to directly apply the result in Corollary 1, but, if the loss function \(\ell\) is smooth, strongly convex, and is generated from random data, then we show that \(G^\ell\) is also weakly DR-modular.

Proposition 5. Let \(\ell(x) = L(x) - z^T x\), where \(L\) is smooth and strongly convex, and \(z \in \mathbb{R}^d\) has a continuous density w.r.t the Lebesgue measure. Then there exist \(\alpha_G, \beta_G > 0\) such that \(G^\ell\) is \((\alpha_G, \beta_G)\)-weakly DR-modular, almost surely.

We prove Proposition 5 by first utilizing a result from (Elenberg et al., 2018), which relates the marginal gain of \(G^\ell\) to the marginal decrease of \(\ell\). We then argue that the minimizer of \(\ell\), restricted to any given support, has full support with probability one, and thus \(\ell\) has non-zero marginal decrease with probability one. The proof is given in Appendix C.

The actual \(\alpha_G, \beta_G\) parameters depend on the conditioning of \(\ell\). Their positivity also relies on \(z\) being random, typically, data drawn from a distribution (Sakaue, 2018, Sect. A.1). In Section 5.2, we evaluate Proposition 5 empirically.

The approximation guarantee in Corollary 1 thus applies directly to Problem 5, whenever \(\ell\) has the form in Proposition 5 and \(F\) is \(\alpha\)-weakly DR-submodular. For example, this holds when \(\ell\) is the least squares loss with a nonsingular measurement matrix. Examples of structure-inducing regularizers \(F\) include submodular regularizers (Bach, 2010), and non-submodular ones such as the range cost function (Bach, 2010; El Halabi et al., 2018) \((\alpha = \frac{1}{d - 1})\), which favors interval supports, with applications in time-series and cancer diagnosis (Kapoor et al., 2008), and the cost function considered (Sakaue, 2019) \((\alpha = \frac{a + b - 1}{a + b - 1 + d - 1})\), where \(0 < 2a < b\) are cost parameters, which favors the selection of sparse and cheap features, with applications in healthcare.
4.2. Batch Bayesian optimization

The goal in batch Bayesian optimization is to optimize an unknown expensive-to-evaluate noisy function $f$ with as few batches of function evaluations as possible (Desautels et al., 2014; Gonzalez et al., 2016). For example, evaluations can correspond to performing expensive experiments. The evaluation points are chosen to maximize an acquisition function subject to a cardinality constraint. Several acquisition functions have been proposed for this purpose, amongst others the variance reduction function (Krause et al., 2008; Bogunovic et al., 2016). This function is used to maximally reduce the variance of the posterior distribution over potential maximizers of the unknown function.

Often, the unknown $f$ is modeled by a Gaussian process with zero mean and kernel function $k(x, x')$, and we observe noisy evaluations $y = f(x) + \epsilon$ of the function, where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \). Given a set $\mathcal{X} = \{x_1, \ldots, x_d\}$ of potential maximizers of $f$, each $x_i \in \mathbb{R}^n$, and a set $S \subseteq V$, let $y_S = [y_i]_{i \in S}$ be the corresponding observations at points $x_i, i \in S$. The posterior distribution of $f$ given $y_S$ is again a Gaussian process, with variance $\sigma^2_S(x) = k(x, x) - k_S(x)^T(K_S + \sigma^2 I)^{-1}k_S(x)$ where $k_S = [k(x_i, x_j)]_{i,j \in S}$, and $K_S = [k(x_i, x_j)]_{i, j \in S}$ is the corresponding submatrix of the positive definite kernel matrix $K$. The variance reduction function is defined as:

$$G(S) = \sum_{i \in V} \sigma^2(x_i) - \sigma^2_S(x_i),$$

where $\sigma^2(x_i) = k(x_i, x_i)$. We show that the variance reduction function is weakly DR-modular.

**Proposition 6.** The variance reduction function $G$ is non-decreasing $(\beta, \beta)$-weakly DR-modular, with $\beta = \sqrt{\frac{\lambda_{\text{min}}(K)}{\lambda_{\text{max}}(K)}} \frac{\lambda_{\text{min}}(K)}{\lambda_{\text{max}}(K)}$, where $\lambda_{\text{max}}(K)$ and $\lambda_{\text{min}}(K)$ are the largest and smallest eigenvalues of $K$.

To prove Proposition 6 we show that $G$ can be written as a noisy column subset selection objective, and prove that such an objective function is weakly DR-modular, generalizing the result of (Sviridenko et al., 2017). The proof is given in Appendix C.2. The variance reduction function can thus be maximized with a greedy algorithm to a $\beta$-approximation (Sviridenko et al., 2017), which follows from a stronger notion of approximate modularity.

Maximizing the variance reduction may also be phrased as an instance of Problem $[\mathbf{1}]$, with $G$ being the variance reduction function, and $F(S) = \lambda |S|$ an item-wise cost. This formulation easily allows to include nonlinear costs with (weak) decrease in marginal costs (economies of scale). For example, in the sensor placement application, the cost of placing a sensor in a hazardous environment may diminish if other sensors are also placed in similar environments. Unlike previous works, the approximation guarantee in Corollary $[\mathbf{1}]$ still applies to such cost functions, while maintaining the $\beta$-approximation with respect to $G$.

5. Experiments

We empirically validate our results on noisy submodular minimization and structured sparse learning. In particular, we address the following questions: (1) How robust are different submodular minimization algorithms, including PGM, to multiplicative noise? (2) How well can PGM minimize a non-submodular objective? Do the parameters $\alpha, \beta$ accurately characterize its performance?

All experiments were implemented in Matlab, and conducted on cluster nodes with 16 Intel Xeon E5 CPU cores and 64 GB RAM. Source code is available at https://github.com/marwash25/non-sub-min.

5.1. Noisy submodular minimization

First, we consider minimizing a submodular function $\mathcal{H}$ given a noisy oracle $\tilde{\mathcal{H}} = \xi \mathcal{H}$, where $\xi$ is independently drawn from a Gaussian distribution with mean one and standard deviation 0.1. We evaluate the performance of different submodular minimization algorithms, on two example problems, minimum cut and clustering. We use the Matlab code from http://www.di.ens.fr/~fbach/submodular/ and compare seven algorithms: the minimum-norm-point algorithm (MNP) (Fujishige & Isotani, 2011), the conditional gradient method (Jaggi, 2013) with fixed step-size (CG-2/(t+2)) and with line search (CGLS), PGM with fixed step-size (PGM-1/\sqrt{t}) and with the approximation of Polyak’s rule (PGM-polyak) (Bertsekas, 1995), the analytic center cutting plane method (Gottlieb & Vial, 1993) (ACCPM) and a variant of it that emulates the simplicial method (ACCPM-Kelley).

We replace the true oracle for $\mathcal{H}$ by the approximate oracle $\tilde{\mathcal{H}}_m(S) = \frac{1}{m} \sum_{i=1}^m \xi_i H(S)$, for all these algorithms, and test them on two datasets: Genmf-long, a min-cut/max-flow problem with $d = 575$ nodes and 2390 edges, and Two-moons, a synthetic semi-supervised clustering instance with $d = 400$ data points and 16 labeled points. For details of the algorithms and datasets, we refer the reader to (Bach, 2013) Sect. 12.1 for more details about the algorithms and datasets. We stopped each algorithm after 1000 iterations for the first dataset and after 400 iterations for the second one, or until the approximate duality gap reached $10^{-8}$. To compute the optimal value $H(S^*)$, we use MNP with the noise-free oracle $\tilde{\mathcal{H}}$.

Figure 2 shows the gap in discrete objective value for all algorithms on the two datasets, for increasing number of samples $m$ (top), and for two fixed values of $m$, as a function of iterations (middle and bottom). We plot the best value achieved so far. As expected, the accuracy improves with more samples. In fact, this improvement is faster than
the bounds in Proposition 4 and in (Blais et al. 2018). The objective values in the Two-moons data are smaller, which makes it easier to solve in the multiplicative noise setting (Prop. 4), as we indeed observe. Among the compared algorithms, ACCPM and MNP converge fastest, as also observed in (Bach, 2013) without noise, but they also seem to be the most sensitive to noise. In summary, these empirical results suggest that submodular minimization algorithms are indeed robust to noise, as predicted by our theory.

5.2. Structured sparse learning

Our second set of experiments is structured sparse learning, where we aim to estimate a sparse parameter vector \( x^* \in \mathbb{R}^d \) whose support is an interval. The range function \( H(S) \) (Bach, 2010). As discussed in Section 4.1, no prior method provides a guaranteed approximate solution to Problem 3 with such regularizers, with the exception of some statistical assumptions, under which \( x^* \) can be recovered using the tightest convex relaxation \( \Theta^r \) of \( F^r \) (El Halabi et al. 2018). Evaluating \( \Theta^r \) involves a linear program with constraints corresponding to all possible interval sets. Such exhaustive search is not feasible in more complex settings.

We consider a simple linear regression setting in which \( x^* \in \mathbb{R}^d \) has \( k \) consecutive ones and is zero otherwise. We observe \( y = Ax^* + \epsilon \), where \( A \in \mathbb{R}^{d \times n} \) is an i.i.d Gaussian matrix with normalized columns, and \( \epsilon \in \mathbb{R}^n \) is an i.i.d Gaussian noise vector with standard deviation 0.01. We set \( d = 250, k = 20 \) and vary the number of measurements \( n \) between \( d/4 \) and \( 2d \). We compare the solutions obtained by minimizing the least squares loss \( (x) = \frac{1}{2} \| y - Ax \|_2^2 \) with the three regularizers: The range function \( F^r \), where \( H \) is optimized via exhaustive search (OPT-Range), or via PGM (PGM-Range); the modified range function \( F^{mr} \), solved via exhaustive search (OPT-ModRange), or via PGM (PGM-ModRange); and the convex relaxation \( \Theta^r \) (CR-Range), solved using CVX (Grant & Boyd 2014). The marginal gains of \( G^t \) can be efficiently computed using rank-1 updates of the pseudo-inverse (Meyer, 1973).

Figure 5 (top) displays the best achieved support error in hamming distance, and estimation error \( \| \hat{x} - x^* \|_2 / \| x^* \|_2 \) on the regularization path, where \( \lambda \) was varied between \( 10^{-4} \) and 10. Figure 5 (middle and bottom) illustrates the objective value \( H = \lambda F^r - G^t \) for PGM-Range, CR-Range, and OPT-Range, and \( H = \lambda F^{mr} - G^t \) for PGM-ModRange, and OPT-ModRange, and the corresponding parameters \( \alpha_T, \beta_T \) defined in Remark 1 for two fixed values of \( n \). Results are averaged over 5 runs.

We observe that PGM minimizes the objective with \( F^{mr} \)
almost exactly as \( n \) grows. It performs a bit worse with \( F^r \), which is expected since \( F^r \) is not submodular. This is also reflected in the support and estimation errors. Moreover, \( \alpha_T, \beta_T \) here reasonably predict the performance of PGM; larger values correlate with closer to optimal objective values. They are also more accurate than the worst case \( \alpha, \beta \) in Definition 1. Indeed, the \( \alpha_T \) for the range function is much larger than the worst case \( \frac{1}{2} \). Similarly, \( \beta_T \) for \( G^\ell \) is quite large and approaches 1 as \( n \) grows, while in Proposition 5 the worst case \( \beta \) is only guaranteed to be non-zero when \( \ell \) is strongly convex. Finally, the convex approach with \( \Theta^r \) essentially matches the performance of OPT-Range when \( n \geq d \). In this regime, \( G^\ell \) becomes nearly modular, hence the convex objective \( \ell + \lambda \Theta^r \) starts approximating the convex closure of \( \lambda F^r - G^\ell \).

6. Conclusion

We established new links between approximate submodularity and convexity, and used them to analyze the performance of PGM for unconstrained, possibly noisy, non-submodular minimization. This yielded the first approximation guarantee for this problem, with a matching lower bound establishing its optimality. We experimentally validated our theory, and illustrated the robustness of submodular minimization algorithms to noise and non-submodularity.

Acknowledgments

This research was supported by a DARPA D3M award, NSF CAREER award 1553284, and NSF award 1717610. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

References


Optimal approximation for unconstrained non-submodular minimization


Obozinski, G. and Bach, F. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. working paper or preprint, December 2016. URL https://hal-enpc.archives-ouvertes.fr/hal-01412385


