Robust Maximization of Non-Submodular Objectives

Ilija Bogunovic†
LIONS, EPFL
ilija.bogunovic@epfl.ch

Junyao Zhao†
LIONS, EPFL
zhaoju@student.ethz.ch

Volkan Cevher
LIONS, EPFL
volkan.cevher@epfl.ch

Abstract

We study the problem of maximizing a monotone set function subject to a cardinality constraint k in the setting where some number of elements τ is deleted from the returned set. The focus of this work is on the worst-case adversarial setting. While there exist constant-factor guarantees when the function is submodular [1, 2], there are no guarantees for non-submodular objectives. In this work, we present a new algorithm OBLIVIOUS-GREEDY and prove the first constant-factor approximation guarantees for a wider class of non-submodular objectives. The obtained theoretical bounds are the *first* constant-factor bounds that also hold in the *linear regime*, i.e. when the number of deletions τ is linear in k. Our bounds depend on established parameters such as the *submodularity ratio* and some novel ones such as the *inverse curvature*. We bound these parameters for two important objectives including support selection and variance reduction. Finally, we numerically demonstrate the robust performance of OBLIVIOUS-GREEDY for these two objectives on various datasets.

1 Introduction

A wide variety of important problems in machine learning can be formulated as the maximization of a *monotone*[1] set function $f : 2^V \to \mathbb{R}_+$ under the cardi-

nality constraint k, i.e.

$$\max_{S \subseteq V, |S| \leq k} f(S), \quad (1)$$

where $V = \{v_1, \cdots v_n\}$ is the ground set of items. However, in many applications, we might require robustness of the solution set, meaning that the objective value should deteriorate as little as possible after a subset of elements is deleted.

For example, an important problem in machine learning is feature selection, where the goal is to extract a subset of features that are informative w.r.t. a given task (e.g. classification). For some tasks, it is of great importance to select features that exhibit robustness against deletions. This is particularly important in domains with non-stationary feature distributions or with input sensor failures [3]. Another important example is the optimization of an unknown function from point evaluations that require performing costly experiments. When the experiments can fail, protecting against worst-case failures becomes important.

In this work, we consider the following robust variant of Problem (1):

$$\max_{S \subseteq V, |S| \leq k} \min_{E \subseteq S, |E| \leq \tau} f(S \setminus E), \quad (2)$$

where[2] τ is the size of subset E that is removed from the solution set S. When the objective function exhibits *submodularity*, a natural notion of diminishing returns[3], a constant factor approximation guarantee can be obtained for the robust Problem 2 [1, 2]. However, in many applications such as the above mentioned feature selection problem, the objective function $f(\cdot)$ is not submodular and the obtained guarantees are not applicable.

Background and related work. When the objective function is submodular, the simple GREEDY algorithm [4] achieves a $(1 - 1/e)$-multiplicative approximation guarantee for Problem 1. The constant factor

---

†Equal contribution.
[1]Non-negative and normalized (i.e. $f(\emptyset) = 0$) $f(\cdot)$ is *monotone* if for any sets $X \subseteq Y \subseteq V$ it holds $f(X) \leq f(Y)$.

[2]When $\tau = 0$, Problem (2) reduces to Problem (1).
[3]$f(\cdot)$ is submodular if for any sets $X \subseteq Y \subseteq V$ and any element $e \in V \setminus Y$, it holds that $f(X \cup \{e\}) - f(X) \geq f(Y \cup \{e\}) - f(Y)$.

can be further improved by exploiting the properties of the objective function, such as the *closeness* to being modular captured by the notion of *curvature* [5, 6, 7].

In many cases, the GREEDY algorithm performs well empirically even when the objective function deviates from being submodular. An important class of such objectives are γ-*weakly* submodular functions. Simply put, *submodularity ratio* γ is a quantity that characterizes how *close* the function is to being submodular. It was first introduced in [8], where it was shown that for such functions the approximation ratio of GREEDY for Problem (1) degrades slowly as the submodularity ratio decreases i.e. as $(1 - e^{-\gamma})$. In [9], the authors obtain the approximation guarantee of the form $\alpha^{-1}(1 - e^{-\gamma\alpha})$, that further depends on the curvature α.

When the objective is submodular, the GREEDY algorithm can perform arbitrarily badly when applied to Problem (2) [1, 2]. A submodular version of Problem (2) was first introduced in Krause *et al.* [10], while the first efficient algorithm and constant factor guarantees were obtained in Orlin *et al.* [1] for $\tau = o(\sqrt{k})$. In Bogunovic *et al.* [2], the authors introduce the PRO-GREEDY algorithm that attains the same 0.387-guarantee but it allows for greater robustness, i.e. the allowed number of removed elements is $\tau = o(k)$. It is not clear how the obtained guarantees generalize for non-submodular functions.

One important class of non-submodular functions that we consider in this work are those used for support selection:

$$f(S) := \max_{\mathbf{x} \in \mathcal{X}, \text{supp}(\mathbf{x}) \subseteq S} l(\mathbf{x}), \tag{3}$$

where $l(\cdot)$ is a continuous function, $\mathcal{X}$ is a convex set and $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$. A popular way to solve the problem of finding a k-sparse vector that maximizes l, i.e. $\mathbf{x} \in \arg\max_{\mathbf{x} \in \mathcal{X}, \|\mathbf{x}\|_0 \leq k} l(\mathbf{x})$ is to maximize the auxiliary set function in (3) subject to the cardinality constraint k. This setting and its variants have been used in various applications, for example, sparse approximation [8, 11], feature selection [12], sparse recovery [13], sparse M-estimation [14] and column subset selection problems [15]. An important result from [16] states that if $l(\cdot)$ is (m, L)-(strongly concave, smooth) then $f(S)$ is weakly submodular with submodularity ratio $\gamma \geq \frac{m}{L}$. Consequently, this result enlarges the number of problems where GREEDY comes with guarantees. In this work, we consider the robust version of this problem, where the goal is to protect against the worst-case adversarial deletions of features.

Deletion robust submodular maximization in the streaming setting has been considered in [17, 18, 19]. Other versions of robust submodular optimization

problems have also been studied. In [10], the goal is to select a set of elements that is robust against the worst possible objective from a given finite set of monotone submodular functions. The same problem with different types of constraints is considered in [20]. It was further studied in the domain of influence maximization [21, 22]. The robust version of the budget allocation problem was considered in [23].In [24], the authors study the problem of maximizing a monotone submodular function under adversarial noise. We conclude this section by noting that very recently a couple of different works have further studied robust submodular problems [25, 26, 27, 28].

Main contributions:

- We initiate the study of the robust optimization Problem (2) for a wider class of monotone non-submodular functions. We present a new algorithm OBLIVIOUS-GREEDY and prove the first constant factor approximation guarantees for Problem (2). When the function is submodular and under mild conditions, we recover the approximation guarantees obtained in the previous works [1, 2].

- For both non-submodular and submodular case, we obtain the *first* constant factor approximation guarantees for the *linear regime*, i.e. when $\tau = ck$ for some $c \in (0, 1)$.

- Our theoretical bounds are expressed in terms of parameters that further characterize a set function. Some of them have been used in previous works, e.g. submodularity ratio, and some of them are novel, such as the *inverse curvature*. We prove some interesting relations between these parameters and obtain theoretical bounds for them in two important applications: (i) support selection and (ii) variance reduction objective used in batch Bayesian optimization. This allows us to obtain the *first robust* guarantees for these two important objectives.

- Finally, we experimentally validate the robustness of OBLIVIOUS-GREEDY in several scenarios, and demonstrate that it outperforms other robust and non-robust algorithms.

2 Preliminaries

Set function ratios. In this work, we consider a normalized monotone set function $f : 2^V \to \mathbb{R}_+$; we proceed by defining several quantities that characterize it. Some of the quantities were introduced and used in various different works, while the novel ones that

we consider are *inverse curvature, bipartite supermodularity ratio and (super/sub)additivity ratio.*

Definition 1 (Submodularity [8] and Supermodularity ratio). *The submodularity ratio of $f(\cdot)$ is the largest scalar $\gamma \in [0,1]$ s.t.*

$$\frac{\sum_{i \in \Omega} f(\{i\}|S)}{f(\Omega|S)} \geq \gamma, \quad \forall \text{ disjoint } S, \Omega \subseteq V. \quad (4)$$

while the supermodularity ratio is the largest scalar $\check{\gamma} \in [0,1]$ s.t.

$$\frac{f(\Omega|S)}{\sum_{i \in \Omega} f(\{i\}|S)} \geq \check{\gamma}, \quad \forall \text{ disjoint } S, \Omega \subseteq V. \quad (5)$$

The funciton $f(\cdot)$ is submodular (supermodular) *iff* $\gamma = 1$ ($\check{\gamma} = 1$). Hence, the submodularity/supermodularity ratio measures to what extent the function has submodular/supermodular properties. While $f(\cdot)$ is modular *iff* $\gamma = \check{\gamma} = 1$, in general, γ can be different from $\check{\gamma}$.

Definition 2 (Generalized curvature [6, 9] and inverse generalized curvature). *The generalized curvature of $f(\cdot)$ is the smallest scalar $\alpha \in [0,1]$ s.t.*

$$\frac{f(\{i\}|S \setminus \{i\} \cup \Omega)}{f(\{i\}|S \setminus \{i\})} \geq 1 - \alpha, \quad \forall S, \Omega \subseteq V, i \in S \setminus \Omega, \quad (6)$$

while the inverse generalized curvature is the smallest scalar $\check{\alpha} \in [0,1]$ s.t.

$$\frac{f(\{i\}|S \setminus \{i\})}{f(\{i\}|S \setminus \{i\} \cup \Omega)} \geq 1 - \check{\alpha}, \quad \forall S, \Omega \subseteq V, i \in S \setminus \Omega. \quad (7)$$

The function $f(\cdot)$ is submodular (supermodular) *iff* $\check{\alpha} = 0$ ($\alpha = 0$). The function is modular *iff* $\alpha = \check{\alpha} = 0$. In general, α can be different from $\check{\alpha}$.

Definition 3 (sub/superadditivity ratio). *The subadditivity ratio of $f(\cdot)$ is the largest scalar $\nu \in [0,1]$ such that*

$$\frac{\sum_{i \in S} f(\{i\})}{f(S)} \geq \nu, \quad \forall S \subseteq V. \quad (8)$$

The superadditivity ratio is the largest scalar $\check{\nu} \in [0,1]$ such that

$$\frac{f(S)}{\sum_{i \in S} f(\{i\})} \geq \check{\nu}, \quad \forall S \subseteq V. \quad (9)$$

If the function is submodular (supermodular) then $\nu = 1$ ($\check{\nu} = 1$).

The following proposition captures the relation between the above quantities.

Proposition 1. *For any $f(\cdot)$, the following relations hold:*

$$\nu \geq \gamma \geq 1 - \check{\alpha} \quad \text{and} \quad \check{\nu} \geq \check{\gamma} \geq 1 - \alpha.$$

We also provide a more general definition of the bipartite subadditivity ratio used in [12].

Definition 4 (Bipartite subadditivity ratio). *The bipartite subadditivity ratio of $f(\cdot)$ is the largest scalar $\theta \in [0,1]$ s.t.*

$$\frac{f(A) + f(B)}{f(S)} \geq \theta, \quad \forall S \subseteq V, A \cup B = S, A \cap B = \emptyset. \quad (10)$$

Remark 1. *For any $f(\cdot)$, it holds that $\theta \geq \check{\nu}\nu$.*

Greedy guarantee. Different works [8, 9] have studied the performance of the GREEDY algorithm [4] for Problem 1 when the objective is γ-weakly submodular. In our analysis, we are going to make use of the following important result from [8].

Lemma 1. *For a monotone normalized set function $f : 2^V \to \mathbb{R}_+$, with submodularity ratio $\gamma \in [0,1]$ the GREEDY algorithm when run for l steps returns a set S_l of size l such that*

$$f(S_l) \geq \left(1 - e^{-\gamma \frac{l}{k}}\right) f(\text{OPT}_{(k,V)}),$$

where $\text{OPT}_{(k,V)}$ is used to denote the optimal set of size k, i.e., $\text{OPT}_{(k,V)} \in \arg\max_{S \subseteq V, |S| \leq k} f(S)$.

3 Algorithm and its Guarantees

We present our OBLIVIOUS-GREEDY algorithm in Algorithm 1. The algorithm requires a non-negative monotone set function $f : 2^V \to \mathbb{R}_+$, and the ground set of items V. It constructs two sets S_0 and S_1. The first set S_0 is constructed via oblivious selection, i.e. $\lceil \beta \tau \rceil$ items with the individually highest objective values are selected. Here, $\beta \in \mathbb{R}_+$ is an input parameter, that together with τ, determines the size of S_0 ($|S_0| = \lceil \beta \tau \rceil \leq k$). We provide more information on this parameter in the next section. The second set S_1, of size $k - |S_0|$, is obtained by running the GREEDY algorithm on the remaining items $V \setminus S_0$. Finally, the algorithm outputs the set $S = S_0 \cup S_1$ of size k that is robust against the worst-case removal of τ elements.

Intuitively, the role of S_0 is to ensure robustness, as its elements are selected independently of each other and have high marginal values, while S_1 is obtained greedily and it is near-optimal on the set $V \setminus S_0$.

OBLIVIOUS-GREEDY is simpler than the submodular algorithms PRO-GREEDY [2] and OSU [1]. Both of these algorithms construct multiple sets (buckets) whose number and size depend on the input parameters k and τ. In contrast, OBLIVIOUS-GREEDY always constructs two sets, where the first set is obtained by the fast OBLIVIOUS selection.

Algorithm 1 OBLIVIOUS-GREEDY algorithm

Require: Set V, k, τ, $\beta \in \mathbb{R}_+$ and $\lceil \beta\tau \rceil \le k$
Ensure: Set $S \subseteq V$ such that $|S| \le k$
1: $S_0, S_1 \leftarrow \emptyset$
2: **for** $i \leftarrow 0$ **to** $\lceil \beta\tau \rceil$ **do**
3:      $v \leftarrow \arg\max_{v \in V \setminus S_0} f(\{v\})$
4:      $S_0 \leftarrow S_0 \cup \{v\}$
5: $S_1 \leftarrow$ GREEDY$(k - |S_0|, (V \setminus S_0))$
6: $S \leftarrow S_0 \cup S_1$
7: **return** S

For Problem (1) and the weakly submodular objective, the GREEDY algorithm achieves a constant factor approximation (Lemma 1), while OBLIVIOUS selection achieves (γ/k)-approximation [12]. For the harder Problem (2), GREEDY can fail arbitrarily badly [2]. Interestingly enough, the combination of these two algorithms reflected in OBLIVIOUS-GREEDY leads to a constant factor approximation for Problem (2).

3.1 Approximation guarantee

The quantity of interest in this section is the remaining utility after the adversarial removal of elements $f(S \setminus E_S^*)$, where S is the set of size k returned by OBLIVIOUS-GREEDY, and E_S^* is the set of size τ chosen by the adversary, i.e., $E_S^* \in \arg\min_{E \subset S, |E| \le \tau} f(S \setminus E)$. Let $\text{OPT}_{(k-\tau, V \setminus E_S^*)}$ denote the optimal solution, of size $k - \tau$, when the ground set is $V \setminus E_S^*$. The goal in this section is to compare $f(S \setminus E_S^*)$ to $f(\text{OPT}_{(k-\tau, V \setminus E_S^*)})$.[4] All the omitted proofs from this section can be found in the supplementary material.

Intermediate results. Before stating our main result, we provide three lower bounds on $f(S \setminus E_S^*)$. For the returned set $S = S_0 \cup S_1$, we let E_0 denote elements removed from S_0, i.e., $E_0 := E_S^* \cap S_0$ and similarly $E_1 := E_S^* \cap S_1$. The first lemma is borrowed from [2], and states that $f(S \setminus E_S^*)$ is at least some constant fraction of the utility of the elements obtained greedily in the second stage.

Lemma 2. *For any $f(\cdot)$ (not necessarily submodular), let $\mu \in [0, 1]$ be a constant such that $f(E_1 \mid (S \setminus E_S^*)) = \mu f(S_1)$ holds. Then, $f(S \setminus E_S^*) \ge (1 - \mu)f(S_1)$.*

The next lemma generalizes the result obtained in [1, 2], and applies to any non-negative monotone set function with bipartite subadditivity ratio θ.

Lemma 3. *Let $\theta \in [0, 1]$ be a bipartite subadditivity ratio defined in Eq. (10). Then $f(S \setminus E_S^*)$ is at least*

$$\theta f(\text{OPT}_{(k-\tau, V \setminus E_S^*)}) - (1 - e^{-\frac{k - |S_0|}{k - \tau}})^{-1} f(S_1).$$

---

[4]As shown in [1], $f(\text{OPT}_{(k-\tau, V \setminus E_S^*)}) \ge f(\text{OPT} \setminus E_{\text{OPT}}^*)$, where OPT is the optimal solution to Problem (2).

In other words, if $f(S_1)$ is *small* compared to the utility of the optimal solution, then $f(S \setminus E_S^*)$ is at least a constant factor away from the optimal solution.

Next, we present our key lemma that further relates $f(S \setminus E_S^*)$ to the utility of the set S_1 with no deletions.

Lemma 4. *Let β be a constant such that $|S_0| = \lceil \beta\tau \rceil$ and $|S_0| \le k$, and let $\check{\nu}, \check{\alpha} \in [0, 1]$ be a superadditivity ratio and generalized inverse curvature (Eq. (9) and Eq. (7), respectively). Finally, let μ be a constant defined as in Lemma 2. Then,*

$$f(S \setminus E_S^*) \ge (\beta - 1)\check{\nu}(1 - \check{\alpha})\mu f(S_1).$$

Proof. We have:

$$f(S \setminus E_S^*) \ge f(S_0 \setminus E_0)$$
$$\ge \check{\nu} \sum_{e_i \in S_0 \setminus E_0} f(\{e_i\}) \tag{11}$$
$$\ge \frac{|S_0 \setminus E_0|}{|E_1|} \check{\nu} \sum_{e_i \in E_1} f(\{e_i\}) \tag{12}$$
$$\ge \frac{(\beta - 1)\tau}{\tau} \check{\nu} \sum_{e_i \in E_1} f(\{e_i\}) \tag{13}$$
$$\ge (\beta - 1)\check{\nu}(1 - \check{\alpha})$$
$$\times \sum_{i=1}^{|E_1|} f\left(\{e_i\} | (S \setminus E_S^*) \cup E_1^{(i-1)}\right) \tag{14}$$
$$= (\beta - 1)\check{\nu}(1 - \check{\alpha})f\left(E_1 | (S \setminus E_S^*)\right) \tag{15}$$
$$= (\beta - 1)\check{\nu}(1 - \check{\alpha})\mu f(S_1). \tag{16}$$

Eq. (11) follows by the superadditivity. Eq. (12) follows from the way S_0 is constructed, i.e. via OBLIVIOUS selection that ensures $f(\{i\}) \ge f(\{j\})$ for every $i \in S_0 \setminus E_0$ and $j \in E_1$. Eq. (13) follows from $|S_0 \setminus E_0| = \lceil \beta\tau \rceil - |E_0| \ge \beta\tau - \tau = (\beta - 1)\tau$, and $|E_1| \le \tau$.

To prove Eq. (14), let $E_1 = \{e_1, \cdots e_{|E_1|}\}$, and let $E_1^{(i-1)} \subseteq E_1$ denote the set $\{e_1, \cdots, e_{i-1}\}$. Also, let $E_1^{(0)} = \emptyset$. Eq. (14) then follows from

$$f(\{e_i\}) \ge (1 - \check{\alpha})f\left(\{e_i\} | (S \setminus E_S^*) \cup E_1^{(i-1)}\right),$$

which in turns follows from (7) by setting $S = \{e_i\}$ and $\Omega = (S \setminus E_S^*) \cup E_1^{(i-1)}$.

Finally, Eq. (15) follows from $f(E_1|(S \setminus E_S^*)) = \sum_{e_i \in E_1} f\left(\{e_i\}|(S \setminus E_S^*) \cup E_1^{(i-1)}\right)$ (telescoping sum) and Eq. (16) follows from the definition of μ. $\qquad\square$

Main result. We obtain the main result by examining the maximum of the obtained lower bounds in Lemma 2, 3 and 4. Note, that all three obtained lower bounds
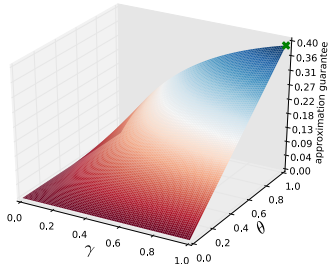
Ilija Bogunovic[†], Junyao Zhao[†], Volkan Cevher

Figure 1: Approximation guarantee obtained in Remark 2. The green cross represents the approximation guarantee when f is submodular ($\gamma = \theta = 1$).

depend on $f(S_1)$. In Lemma 3, we benefit from $f(S_1)$ being small while the opposite is true for Lemma 2 and 4 (both bounds are increasing in $f(S_1)$). By examining the latter two, we observe that in Lemma 2 we benefit from μ being small (i.e. the utility that we lose due to E_1 is small compared to the utility of the whole set S_1) while the opposite is true for Lemma 4. By carefully balancing between these cases (see Appendix C for details) we arrive at our main result.

Theorem 1. *Let $f : 2^V \to \mathbb{R}_+$ be a normalized, monotone set function with submodularity ratio γ, bipartite subadditivity ratio θ, inverse curvature $\check{\alpha}$ and superadditivity ratio $\check{\nu}$, every parameter in $[0, 1]$. For a given budget k and $\tau = \lceil ck \rceil$, for some $c \in (0, 1)$, the* OBLIVIOUS-GREEDY *algorithm with β s.t. $\lceil \beta \tau \rceil \leq k$ and $\beta > 1$, , returns a set S of size k such that when $k \to \infty$ we have*

$$f(S \setminus E_S^*) \geq \frac{\theta P \left(1 - e^{-\gamma \frac{1-\beta c}{1-c}}\right)}{1 + P \left(1 - e^{-\gamma \frac{1-\beta c}{1-c}}\right)} f(\mathrm{OPT}_{(k-\tau, V \setminus E_S^*)}).$$

where P is used to denote $\frac{(\beta-1)\check{\nu}(1-\check{\alpha})}{1+(\beta-1)\check{\nu}(1-\check{\alpha})}$.

Remark 2. *Consider $f(\cdot)$ from Theorem 1 with $\check{\nu} \in (0, 1]$ and $\check{\alpha} \in [0, 1)$. When $\tau = o\left(\frac{k}{\beta}\right)$ and $\beta \geq \log k$ we have:*

$$f(S \setminus E_S^*) \geq \left(\theta \frac{1 - e^{-\gamma}}{2 - e^{-\gamma}} + o(1)\right) f(\mathrm{OPT}_{(k-\tau, V \setminus E_S^*)}).$$

Interpretation. An open question from [2] is whether a constant factor approximation guarantee is possible in the *linear regime*, i.e. when the number of removals is $\tau = \lceil ck \rceil$ for some constant $c \in (0, 1)$ [2]. In Theorem 1 we obtain the first asymptotic constant factor approximation in this regime.

Additionally, when f is submodular, all the parameters in the obtained bound are fixed ($\check{\alpha} = 0$ and $\gamma = \theta = 1$ due to submodularity) except the superadditivity ratio $\check{\nu}$ which can take any value in $[0, 1]$. The approximation factor improves for greater $\check{\nu}$, i.e. the

closer the function is to being superadditive. On the other hand, if f is supermodular then $\check{\nu} = 1$ while $\check{\alpha}, \theta, \gamma$ are in $[0, 1]$, and the approximation factor improves for larger θ and γ, and smaller $\check{\alpha}$.

From Remark 2, when f is submodular, OBLIVIOUS-GREEDY achieves an asymptotic approximation factor of at least 0.387. This matches the approximation guarantee obtained in [2, 1], while it allows for a greater number of deletions $\tau = o\left(\frac{k}{\log k}\right)$ in comparison to $\tau = o\left(\frac{k}{\log^3 k}\right)$ and $\tau = o(\sqrt{k})$ obtained in [2] and [1], respectively. Most importantly, our result holds for a wider range of non-submodular functions. In Figure 1 we show how the asymptotic approximation factor changes as a function of γ and θ.

We also obtain an alternative formulation of our main result, which we present in the following corollary.

Corollary 1. *Consider the setting from Theorem 1 and let $P := \frac{(\beta-1)\check{\nu}\nu}{1+(\beta-1)\check{\nu}(1-\nu)}$. Then we have*

$$f(S \setminus E_S^*) \geq \frac{\theta^2 P \left(1 - e^{-\gamma \frac{1-\beta c}{1-c}}\right)}{1 + \theta P \left(1 - e^{-\gamma \frac{1-\beta c}{1-c}}\right)} f(\mathrm{OPT}_{(k-\tau, V \setminus E_S^*)}).$$

Additionally, consider $f(\cdot)$ with $\check{\nu}, \nu \in (0, 1]$. When $\tau = o\left(\frac{k}{\beta}\right)$ and $\beta \geq \log k$, as $k \to \infty$, we have that $f(S \setminus E_S^*)$ is at least

$$\left(\frac{\theta^2(1 - e^{-\gamma})}{1 + \theta(1 - e^{-\gamma})} + o(1)\right) f(\mathrm{OPT}_{(k-\tau, V \setminus E_S^*)}).$$

The key observation is that the approximation factor depends on ν instead of inverse curvature $\check{\alpha}$. The asymptotic approximation ratio is slightly worse here compared to Theorem 1. However, depending on the considered application, it might be significantly harder to provide bounds for the inverse curvature than bipartite subadditivty ratio, and hence in such cases, this formulation might be more suitable.

4   Applications

In this section, we consider two important real-world applications where deletion robust optimization is of interest. We show that the parameters used in the statement of our main theoretical result can be explicitly characterized, which implies that the obtained guarantees are applicable.

4.1   Robust Support Selection

We first consider the recent results that connect submodularity with concavity [16, 12]. In order to obtain

bounds for robust support selection for general concave functions, we make use of the theoretical bounds obtained for Oblivious-Greedy in Corollary 1.

Given a differentiable concave function $l : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set, and $k \leq d$, the support selection problem is: $\max_{\|\mathbf{x}\|_0 \leq k} l(\mathbf{x})$. As in [16], we let $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$, and consider the associated normalized monotone set function

$$f(S) := \max_{\text{supp}(\mathbf{x}) \subseteq S, \mathbf{x} \in \mathcal{X}} l(\mathbf{x}) - l(\mathbf{0}).$$

Let $T_l(\mathbf{x}, \mathbf{y}) := l(\mathbf{y}) - l(\mathbf{x}) - \langle \nabla l(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$. An important result from [16] can be rephrased as follows: if $l(\cdot)$ is L-smooth and m-strongly concave then for all $\mathbf{x}, \mathbf{y} \in \text{dom}(l)$, it holds

$$-\frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \geq T_l(\mathbf{x}, \mathbf{y}) \geq -\frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2,$$

and f's submodularity ratio γ is lower bounded by $\frac{m}{L}$. Subsequently, in [12] it is shown that θ can also be lower bounded by the same ratio $\frac{m}{L}$.

In this paper, we consider the robust support selection problem, that is, finding a set of features $S \subseteq [d]$ of size k that is robust against the deletion of limited number of features. More formally, the goal is to maximize the following objective over all $S \subseteq [d]$:

$$\min_{|E_S| \leq \tau, E_S \subseteq S} \max_{\text{supp}(\mathbf{x}) \subseteq S \backslash E_S} l(\mathbf{x}) - l(\mathbf{0}).$$

By inspecting the bound obtained in Corollary 1, it remains to bound the (super/sub)additive ratio ν and $\check{\nu}$. The first bound follows by combining the result $\gamma \geq \frac{m}{L}$ with Proposition 1: $\nu \geq \gamma \geq \frac{m}{L}$. To prove the second bound, we make use of the following result.

Proposition 2. *The supermodularity ratio $\check{\gamma}$ of the considered objective $f(\cdot)$ can be lower bounded by $\frac{m}{L}$.*

The second bound follows by combining the result in Proposition 2 and Proposition 1: $\check{\nu} \geq \check{\gamma} \geq \frac{m}{L}$.

4.2 Variance Reduction in Robust Batch Bayesian Optimization

In batch Bayesian optimization, the goal is to optimize an unknown non-convex function from *costly* concurrent function evaluations [29, 30, 31]. Most often, the concurrent evaluations correspond to running an expensive batch of experiments. In the case where experiments can fail, it is beneficial to select a set of experiments in a robust way.

Different acquisition (i.e. auxiliary) functions have been proposed to evaluate the utility of candidate

points for the next evaluations of the unknown function [32]. Recently in [33], the *variance reduction* objective was used as the acquisition function – the unknown function is evaluated at the points that maximally reduce variance of the posterior distribution over the given set of points that represent *potential maximizers*. We formalize this as follows.

Setup. Let $f(\mathbf{x})$ be an unknown function defined over a finite domain $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$. Once we evaluate the function at some point $\mathbf{x}_i \in \mathcal{X}$, we receive a noisy observation $y_i = f(\mathbf{x}_i) + z$, where $z \sim \mathcal{N}(0, \sigma^2)$. In Bayesian optimization, f is modeled as a sample from a Gaussian process. We use a Gaussian process with zero mean and kernel function $k(\mathbf{x}, \mathbf{x}')$, i.e. $f \sim GP(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$. Let $S = \{e_1, \cdots, e_{|S|}\} \subseteq [n]$ denote the set of points, and $\mathbf{X}_S := [\mathbf{x}_{e_1}, \cdots, \mathbf{x}_{e_{|S|}}] \in \mathbb{R}^{|S| \times d}$ and $\mathbf{y}_S := [y_1, \cdots, y_{|S|}]$ denote the corresponding data matrix and observations, respectively. The posterior distribution of f given the points $\mathbf{X}_S$ and observations $\mathbf{y}_S$ is again a GP, with the posterior variance given by:

$$\sigma_{\mathbf{x}|S}^2 = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X}_S) \left(k(\mathbf{X}_S, \mathbf{X}_S) + \sigma^2 \mathbf{I}_{|S|}\right)^{-1} \\ \times k(\mathbf{X}_S, \mathbf{x}).$$

For a given set of potential maximizers $M \subseteq [n]$, the variance reduction objective is defined as follows:

$$F_M(S) := \sum_{\mathbf{x} \in X_M} \sigma_{\mathbf{x}}^2 - \sigma_{\mathbf{x}|S}^2, \qquad (17)$$

where $\sigma_{\mathbf{x}}^2 = k(\mathbf{x}, \mathbf{x})$. We show in Appendix D.2.1 that this objective is *not* submodular in general.

Finally, our goal is to find a set of points S of size k that maximizes

$$\min_{|E_S| \leq \tau, E_S \subseteq S} \sum_{\mathbf{x} \in X_M} \sigma_{\mathbf{x}}^2 - \sigma_{\mathbf{x}|S \backslash E_S}^2.$$

Proposition 3. *Assume the kernel function is such that $k(\mathbf{x}_i, \mathbf{x}_i) \leq k_{max}$, for every $i \in [n]$. The objective function in (17) is normalized and monotone, and both its curvature α and inverse curvature $\check{\alpha}$ can be upper bounded by $\frac{k_{max}}{\sigma^2 + k_{max}}$.*

We can combine this result with Proposition 1, to obtain $\nu \geq \gamma \geq \frac{\sigma^2}{\sigma^2 + k_{\max}}$ and $\check{\nu} \geq \check{\gamma} \geq \frac{\sigma^2}{\sigma^2 + k_{\max}}$. Also, we have $\theta \geq \frac{\sigma^4}{(\sigma^2 + k_{\max})^2}$ (Remark 1). Consequently, all the parameters from Theorem 1 are explicitly characterized.

5 Experimental Results

Optimization performance. For a returned set S, we measure the performance in terms of

Ilija Bogunovic[†], Junyao Zhao[†], Volkan Cevher

$\min_{E \subseteq S, |E| \leq \tau} f(S \setminus E)$. Note that $f(S \setminus E)$ is a submodular function in E. Finding the minimizer E s.t. $|E| \leq \tau$ is NP-hard even to approximate [34]. We rely on the following methods in order to find E of size τ that degrades the solution as much as possible:

– Greedy adversaries: (i) *Greedy Min* – iteratively removes elements to reduce the objective value $f(S \setminus E)$ as much as possible, and (ii) *Greedy Max* – iteratively adds elements from S to maximize the objective $f(E)$.

– Random Greedy adversaries:[5] In order to introduce randomness in the removal process we consider (iii) *Random Greedy Min* – iteratively selects a random element from the top τ elements whose marginal gains are the highest in terms of reducing the objective value $f(S \setminus E)$ and (iv) *Stochastic Greedy Min* – iteratively selects an element, from a random set $R \subseteq V$, with the highest marginal gain in terms of reducing $f(S \setminus E)$. At every step, R is obtained by subsampling $(|S|/\tau) \log(1/\epsilon)$ elements from S.

The minimum objective value $f(S \setminus E)$ among all obtained sets E is reported. Most of the time, for all the considered algorithms, *Greedy Min* finds E that reduces utility the most.
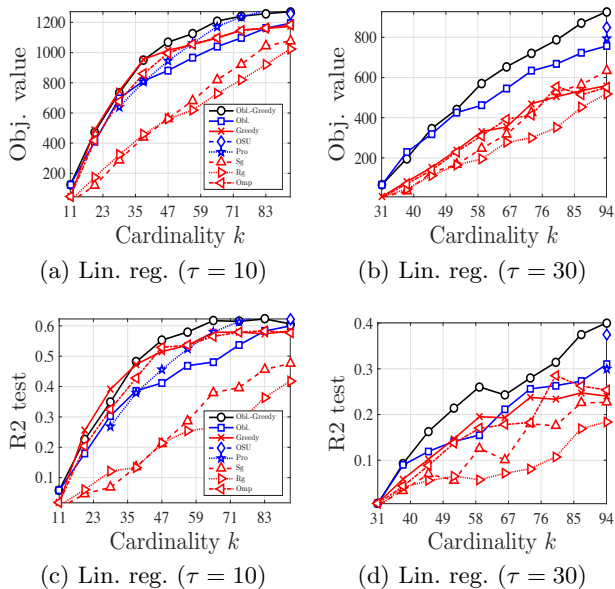

(a) Lin. reg. ($\tau = 10$) (b) Lin. reg. ($\tau = 30$)

(c) Lin. reg. ($\tau = 10$) (d) Lin. reg. ($\tau = 30$)

Figure 2: Comparison of the algorithms on the linear regression task.

5.1 Robust Support Selection

Linear Regression. Our setup is similar to the one in [12]. Each row of the design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is generated by an autoregressive process,

$$X_{i,t+1} = \sqrt{1 - \alpha^2} X_{i,t} + \alpha \epsilon_{i,t}, \qquad (18)$$

---

[5]The random adversaries are inspired by [35] and [36].

(a) Log. synthetic ($\tau = 10$) (b) Log. synthetic ($\tau = 30$)

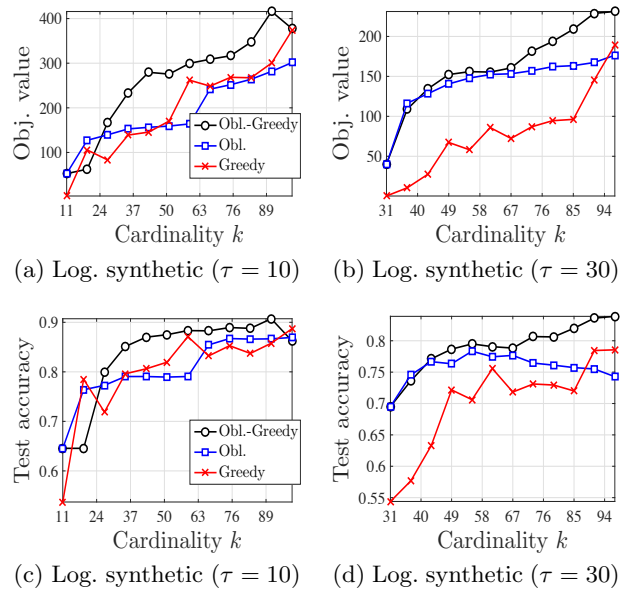(c) Log. synthetic ($\tau = 10$) (d) Log. synthetic ($\tau = 30$)

Figure 3: Logistic regression task with synth. dataset.

where $\epsilon_{i,t}$ is i.i.d. standard Gaussian with variance $\alpha^2 = 0.5$. We use $n = 800$ training data points and $d = 1000$. An additional 2400 points are used for testing. We generate a 100-sparse regression vector by selecting random entries of ω and set them $\omega_s = (-1)^{\text{Bern}(1/2)} \times \left(5\sqrt{\frac{\log d}{n}} + \delta_s \right)$, where δ_s is a standard i.i.d. Gaussian noise. The target is given by $\mathbf{y} = \mathbf{X}\omega + \mathbf{z}$, where $\forall i \in [n]$, $z_i \sim \mathcal{N}(0, 5)$. We compare the performance of OBLIVIOUS-GREEDY against: (i) robust algorithms (in blue) such as OBLIVIOUS, PRO-GREEDY [2], OSU [1], (ii) greedy-type algorithms (in red) such as GREEDY, STOCHASTIC-GREEDY [36], RANDOM-GREEDY [35], ORTHOGONAL-MATCHING-PURSUIT. We require $\beta > 1$ for our asymptotic results to hold, but we found out that in practice (small k regime) $\beta \leq 1$ usually gives the best performance. We use OBLIVIOUS-GREEDY with $\beta = 1$ unless stated otherwise.

The results are shown in Fig. 6. Since PRO-GREEDY and OSU only make sense in the regime where τ is relatively small, the plots show their performance only for feasible values of k. It can be observed that OBLIVIOUS-GREEDY achieves the best performance among all the methods in terms of both training error and test score. Also, the greedy-type algorithms become less robust for larger values of τ.

Logistic Regression. We compare the performance of OBLIVIOUS-GREEDY vs. GREEDY and OBLIVIOUS selection on both synthetic and real-world data.

– *Synthetic data:* We generate a 100-sparse ω by letting $\omega_s = (-1)^{\text{Bern}(1/2)} \times \delta_s$, with $\delta_s \sim \text{Unif}([-1, 1])$. The design matrix $\mathbf{X}$ is generated as in (18), with

$\alpha^2 = 0.09$. We set $d = 200$, and use $n = 600$ points for training and additional 1800 points for testing. The label of the i-th data point $\mathbf{X}_{(i,\cdot)}$ is set to 1 if $1/(1 + \exp(\mathbf{X}_{(i,\cdot)}\boldsymbol{\beta})) > 0.5$ and 0 otherwise. The results are shown in Fig. 3. We can observe that OBLIVIOUS-GREEDY outperforms other methods both in terms of the achieved objective value and generalization error. We also note that the performance of GREEDY decays significantly when τ increases.

– *MNIST:* We consider the 10-class logistic regression task on the MNIST [37] dataset. In this experiment, we set $\beta = 0.5$ in OBLIVIOUS-GREEDY, and we sample 200 images for each digit for the training phase and 100 images of each for testing. The results are shown in Fig. 4. It can be observed that OBLIVIOUS-GREEDY has a distinctive advantage over GREEDY and OBLIVIOUS, while when τ increases the performance of GREEDY decays significantly and more robust OBLIVIOUS starts to outperform it.
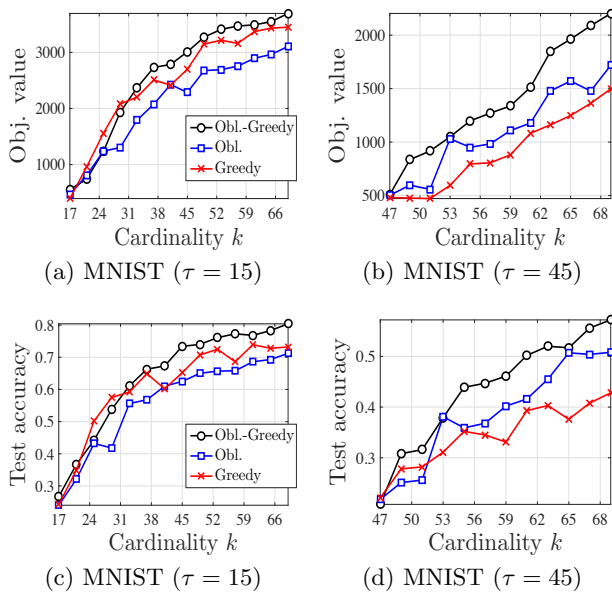

Figure 4: Logistic regression with MNIST dataset.

5.2 Robust Batch Bayesian Optimization via Variance Reduction

Setup. We conducted the following synthetic experiment. A design matrix X of size 600×20 is obtained via the autoregressive process from (18). The function values at these points are generated from a GP with $3/2$-Mátern kernel [38] with both lengthscale and output variance set to 1.0. The samples of this function are corrupted by Gaussian noise, $\sigma^2 = 1.0$. Objective function used is the variance reduction (Eq. (17)). Finally, half of the points randomly chosen are selected in the set M, while the other half is used in the selection process. We use $\beta = 0.5$ in our algorithm.
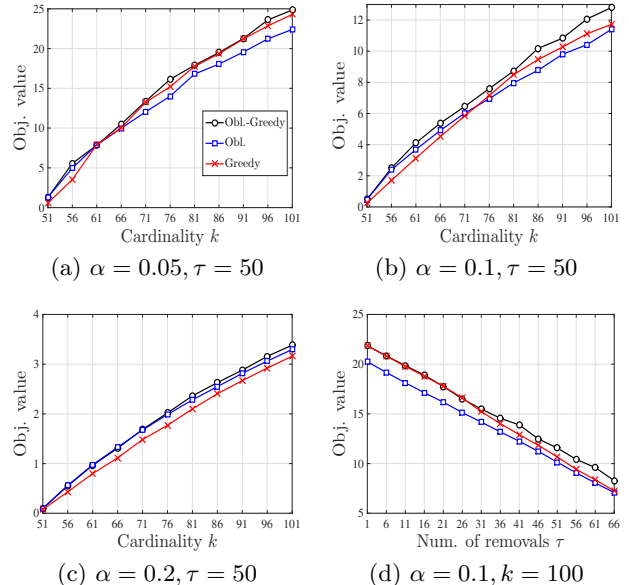

Figure 5: Comparison of the algorithms on the variance reduction task.

Results. In Figure 5 (a), (b), (c), the performance of all three algorithms is shown when τ is fixed to 50. Different figures correspond to different α values. We observe that when $\alpha = 0.1$, GREEDY outperforms OBLIVIOUS for most values of k, while OBLIVIOUS clearly outperforms GREEDY when $\alpha = 0.2$. For all presented values of α, OBLIVIOUS-GREEDY outperforms both GREEDY and OBLIVIOUS selection. For larger values of α, the correlation between the points becomes small and consequently so do the objective values. In such cases, all three algorithms perform similarly. In Figure 5 (d), we show how the performance of all three algorithms decreases as the number of removals increases. When the number of removals is small both GREEDY and our algorithm perform similarly, while as the number of removals increases the performance of GREEDY drops more rapidly.

6 Conclusion

We have presented a new algorithm OBLIVIOUS-GREEDY that achieves constant-factor approximation guarantees for the robust maximization of monotone non-submodular objectives. The theoretical guarantees hold for general $\tau = ck$ for some $c \in (0, 1)$, which resolves the important question posed in [1, 2]. We have also obtained the first robust guarantees for support selection and variance reduction objectives. In various experiments, we have demonstrated the robust performance of OBLIVIOUS-GREEDY by showing that it outperforms both OBLIVIOUS selection and GREEDY, and hence achieves the best of both worlds.

Ilija Bogunovic[†], Junyao Zhao[†], Volkan Cevher

References

[1] J. B. Orlin, A. S. Schulz, and R. Udwani, "Robust monotone submodular function maximization," in *Int. Conf. on Integer Programming and Combinatorial Opt. (IPCO)*, Springer, 2016.

[2] I. Bogunovic, S. Mitrović, J. Scarlett, and V. Cevher, "Robust submodular maximization: A non-uniform partitioning approach," in *Int. Conf. on Machine Learning (ICML)*, August 2017.

[3] A. Globerson and S. Roweis, "Nightmare at test time: robust learning by feature deletion," in *Int. Conf. Machine Learning (ICML)*, 2006.

[4] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[5] M. Conforti and G. Cornuéjols, "Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the rado-edmonds theorem," *Discrete applied mathematics*, vol. 7, no. 3, pp. 251–274, 1984.

[6] J. Vondrák, "Submodularity and curvature: The optimal algorithm (combinatorial optimization and discrete algorithms)," *Kokyuroku Bessatsu*, p. 23:253–266, 2010.

[7] R. K. Iyer, S. Jegelka, and J. A. Bilmes, "Curvature and optimal algorithms for learning and minimizing submodular functions," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, pp. 2742–2750, 2013.

[8] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Proc. of Int. Conf. on Machine Learning, ICML*, pp. 1057–1064, 2011.

[9] A. A. Bian, J. Buhmann, A. Krause, and S. Tschiatschek, "Guarantees for greedy maximization of non-submodular functions with applications," in *Proc. Int. Conf. on Machine Learning (ICML)*, August 2017.

[10] A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta, "Robust submodular observation selection," *Journal of Machine Learning Research*, vol. 9, no. Dec, pp. 2761–2801, 2008.

[11] V. Cevher and A. Krause, "Greedy dictionary selection for sparse representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 979–988, 2011.

[12] R. Khanna, E. Elenberg, A. G. Dimakis, S. Negahban, and J. Ghosh, "Scalable greedy feature selection via weak submodularity," in *Proc. of Int. Conf. on Artificial Intelligence and Statistics, AISTATS*, pp. 1560–1568, 2017.

[13] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[14] P. Jain, A. Tewari, and P. Kar, "On iterative hard thresholding methods for high-dimensional m-estimation," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, pp. 685–693, 2014.

[15] J. Altschuler, A. Bhaskara, G. Fu, V. Mirrokni, A. Rostamizadeh, and M. Zadimoghaddam, "Greedy column subset selection: New bounds and distributed algorithms," in *Int. Conf. on Machine Learning (ICML)*, pp. 2539–2548, 2016.

[16] E. R. Elenberg, R. Khanna, A. G. Dimakis, and S. Negahban, "Restricted strong convexity implies weak submodularity," *CoRR*, vol. abs/1612.00804, 2016.

[17] S. Mitrovic, I. Bogunovic, A. Norouzi-Fard, J. M. Tarnawski, and V. Cevher, "Streaming robust submodular maximization: A partitioned thresholding approach," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, pp. 4560–4569, 2017.

[18] B. Mirzasoleiman, A. Karbasi, and A. Krause, "Deletion-robust submodular maximization: Data summarization with "the right to be forgotten"," in *Int. Conf. Mach. Learn. (ICML)*, pp. 2449–2458, 2017.

[19] E. Kazemi, M. Zadimoghaddam, and A. Karbasi, "Deletion-robust submodular maximization at scale," *arXiv preprint arXiv:1711.07112*, 2017.

[20] T. Powers, J. Bilmes, S. Wisdom, D. W. Krout, and L. Atlas, "Constrained robust submodular optimization." NIPS OPT2016 workshop, 2016.

[21] X. He and D. Kempe, "Robust influence maximization," in *Int. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 885–894, 2016.

[22] W. Chen, T. Lin, Z. Tan, M. Zhao, and X. Zhou, "Robust influence maximization," *arXiv preprint arXiv:1601.06551*, 2016.

[23] M. Staib and S. Jegelka, "Robust budget allocation via continuous submodular functions," in *Proc. of Int. Conf. on Machine Learning (ICML)*, pp. 3230–3240, 2017.

[24] A. Hassidim and Y. Singer, "Submodular optimization under noise," in *Proc. of Conf. on Learning Theory, COLT*, pp. 1069–1122, 2017.

[25] R. Udwani, "Multi-objective maximization of monotone submodular functions with cardinality constraint," *arXiv preprint arXiv:1711.06428*, 2017.

[26] B. Wilder, "Equilibrium computation for zero sum games with submodular structure," *arXiv preprint arXiv:1710.00996*, 2017.

[27] N. Anari, N. Haghtalab, S. Pokutta, M. Singh, A. Torrico, *et al.*, "Robust submodular maximization: Offline and online algorithms," *arXiv preprint arXiv:1710.04740*, 2017.

[28] R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis, "Robust optimization for non-convex objectives," in *Adv. in Neur. Inf. Proc. Sys.*, pp. 4708–4717, 2017.

[29] T. Desautels, A. Krause, and J. W. Burdick, "Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3873–3923, 2014.

[30] J. González, Z. Dai, P. Hennig, and N. Lawrence, "Batch bayesian optimization via local penalization," in *Artificial Intelligence and Statistics*, pp. 648–657, 2016.

[31] J. Azimi, A. Jalali, and X. Z. Fern, "Hybrid batch bayesian optimization," in *Proc. of Int. Conf. on Machine Learning, ICML*, 2012.

[32] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[33] I. Bogunovic, J. Scarlett, A. Krause, and V. Cevher, "Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation," in *Adv. in Neur. Inf. Proc. Sys.*, pp. 1507–1515, 2016.

[34] Z. Svitkina and L. Fleischer, "Submodular approximation: Sampling-based algorithms and lower bounds," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1715–1737, 2011.

[35] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz, "Submodular maximization with cardinality constraints," in *Proc. of ACM-SIAM symposium on Discrete algorithms*, pp. 1433–1452, Society for Industrial and Applied Mathematics, 2014.

[36] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause, "Lazier than lazy greedy," in *Proc. Conf. Art. Intell. (AAAI)*, 2015.

[37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

[39] K. B. Petersen, M. S. Pedersen, *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, p. 15, 2008.