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 \( \tau \) 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 \( \tau \) 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 cardinality constraint \( k \), i.e.

\[
\max_{S \subseteq V, |S| \leq k} f(S),
\]

where \( V = \{v_1, \ldots, v_n\} \) 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):

\[
\max_{S \subseteq V, |S| \leq k} \min_{E \subseteq S, |E| \leq \tau} f(S \setminus E),
\]

where\(^2\) \( \tau \) 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

\[^1\]Equal contribution.
\[^2\]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 \( \gamma \)-weakly submodular functions. Simply put, submodularity ratio \( \gamma \) 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 \(\alpha^{-1}(1 - e^{-\gamma})\), that further depends on the curvature \(\alpha\).

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 \(\tau = \Theta(\sqrt{k})\). 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:

\[
f(S) := \max_{x \in \mathcal{X}, \text{supp}(x) \subseteq S} l(x),
\]

where \(l(\cdot)\) is a continuous function, \(\mathcal{X}\) is a convex set and \(\text{supp}(x) = \{i : x_i \neq 0\}\). A popular way to solve the problem of finding a \(k\)-sparse vector that maximizes \(l\), i.e. \(x \in \arg \max_{x \in \mathcal{X}, \|x\| \leq k} l(x)\) 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 \(l(\cdot) = \langle m, L \rangle\)-strongly concave, smooth) then \(f(S)\) is weakly submodular with submodularity ratio \(\gamma \geq \frac{\alpha}{2}\). 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.

\[
\sum_{i \in \Omega} \frac{f(\{i\})}{f(\{i\} \cup i)} \geq \gamma, \quad \forall \text{disjoint } S, \Omega \subseteq V.
\]  

(4)

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

\[
\sum_{i \in \Omega} \frac{f(\{i\})}{f(\{i\} \cup i)} \geq \bar{\gamma}, \quad \forall \text{disjoint } S, \Omega \subseteq V.
\]  

(5)

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

**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.

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

(6)

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

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

(7)

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

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

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

(8)

The superadditivity ratio is the largest scalar \( \bar{\nu} \in [0,1] \) such that

\[
\sum_{i \in S} f(\{i\}) \geq \frac{f(S)}{\bar{\nu}}, \quad \forall S \subseteq V.
\]  

(9)

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

The following proposition captures the relation between the above quantities.

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

\[
\nu \geq \gamma \geq 1 - \bar{\alpha} \quad \text{and} \quad \bar{\nu} \geq \bar{\gamma} \geq 1 - \alpha.
\]

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.

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

(10)

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

**Greedy guarantee.** Different works [8, 9] have studied the performance of the Greedy algorithm [4] for Problem 1 when the objective is \( \gamma \)-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 \( f: 2^V \rightarrow \mathbb{R}_+ \), 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

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

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

## 3 Algorithm and its Guarantees

We present our Oblivious-Greedy algorithm in Algorithm 1. The algorithm requires a non-negative monotone set function \( f: 2^V \rightarrow \mathbb{R}_+ \), 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., \( \lfloor \beta \tau \rfloor \) items with the individually highest objective values are selected. Here, \( \beta \in \mathbb{R}_+ \) is an input parameter, that together with \( \tau \), determines the size of \( S_0 \) \( \lfloor \beta \tau \rfloor \leq k \). We provide more information on this parameter in the next section. The second set \( S_1 \), of size \( k \), 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 \( \tau \) 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 \( \tau \). In contrast, Oblivious-Greedy always constructs two sets, where the first set is obtained by the fast Oblivious selection.
Let $\beta \leq k$, and applies to any non-negative monotone set function $f$. The next lemma generalizes the result obtained in [1, 2], and states that

$$\mu f(\cdot)$$

is a constant factor approximation (Lemma 1), while OBLIVIOUS selection achieves $(\gamma/k)$-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 $f(S \setminus E^*_k)$, where $S$ is the set of size $k$ returned by OBLIVIOUS-GREEDY, and $E^*_k$ is the set of size $\tau$ chosen by the adversary, i.e., $E^*_k \in \arg\min_{E \subseteq S, |E| \leq \tau} f(S \setminus E)$. Let $\text{OPT}_{(k-\tau,V \setminus E^*_k)}$ denote the optimal solution, of size $k-\tau$, when the ground set is $V \setminus E^*_k$. The goal in this section is to compare $f(S \setminus E^*_k)$ to $f(\text{OPT}_{(k-\tau,V \setminus E^*_k)})$. 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 $f(S \setminus E^*_k)$. For the returned set $S = S_0 \cup S_1$, we let $E_0$ denote elements removed from $S_0$, i.e., $E_0 := E^*_k \setminus S_0$ and similarly $E_1 := E^*_k \setminus S_1$. The first lemma is borrowed from [2], and states that $f(S \setminus E^*_k)$ 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 $f(E_i \mid (S \setminus E^*_k)) = \mu f(S_1)$ holds. Then, $f(S \setminus E^*_k) \geq (1 - \mu) f(S_1)$.

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

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

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

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

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

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

$$f(S \setminus E^*_k) \geq (\beta - 1)\nu(1 - \alpha) f(S_1).$$

**Proof.** We have:

$$f(S \setminus E^*_k) \geq f(S_0 \setminus E_0) \geq \nu \sum_{e_i \in E_0 \setminus S_0} f(\{e_i\}) \geq \beta \nu \sum_{e_i \in E_0 \setminus S_0} f(\{e_i\}) \geq (\beta - 1)\nu(1 - \alpha) f(S_1).$$

Eq. (11) follows by the superadditivity. Eq. (12) follows from the way $S_0$ is constructed, i.e. via OBLIVIOUS selection that ensures $f(\{i\}) \geq f(\{j\})$ for every $i \in S_0 \setminus S_0$ and $j \in E_1$. Eq. (13) follows from $|S_0 \setminus E_0| = \lceil \beta \tau \rceil - |E_0| \geq \beta \tau - \tau = (\beta - 1)\tau$, and $|E_1| \leq \tau$. To prove Eq. (14), let $E_1 = \{e_1, \ldots, e_{|E_1|}\}$, and let $E^{(i-1)} \subseteq E_1$ denote the set $\{e_1, \ldots, e_{i-1}\}$. Also, let $E^{(0)} = \emptyset$. Eq. (14) then follows from

$$f(\{e_i\}) \geq (1 - \alpha) f(\{e_i\} \binom{S \setminus E^*_k} \cup E^{(i-1)}),$$

which in turns follows from (7) by setting $S = \{e_i\}$ and $E = (S \setminus E^*_k) \cup E^{(i-1)}$. Finally, Eq. (15) follows from $f(E_1 \mid (S \setminus E^*_k)) = \sum_{e_i \in E_1} f(\{e_i\} \binom{S \setminus E^*_k} \cup E^{(i-1)})$ (telescoping sum) and Eq. (16) follows from the definition of $\mu$.

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
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 $\mu$ 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 $f: 2^V \rightarrow \mathbb{R}_+$ be a normalized, monotone set function with submodularity ratio $\gamma$, bipartite subadditivity ratio $\theta$, inverse curvature $\nu$ and superadditivity ratio $\nu \dagger$, 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 $\beta$ s.t. $\beta \tau \leq k$ and $\beta > 1$, returns a set $S$ of size $k$ such that when $k \rightarrow \infty$ we have

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

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

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

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

**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 ($\alpha = 0$ and $\gamma = \theta = 1$ due to submodularity) except the superadditivity ratio $\nu$ which can take any value in $[0,1]$. The approximation factor improves for greater $\nu$, i.e. the closer the function is to being superadditive. On the other hand, if $f$ is supermodular then $\nu = 1$ while $\alpha, \theta, \gamma$ are in $[0,1]$, and the approximation factor improves for larger $\theta$ and $\gamma$, and smaller $\alpha$.

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 $\tau = o\left(\frac{k}{\log k}\right)$ 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 $\gamma$ and $\theta$.

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 $P := \frac{(\beta-1)\nu}{1 + (\beta-1)\nu(1-\alpha)}$. Then we have

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

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

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

The key observation is that the approximation factor depends on $\nu$ instead of inverse curvature $\alpha$. 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 subadditivity 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: \( \max_{x \in \mathcal{X}} \{ l(x) \} \). As in [16], we let \( \text{supp}(x) = \{ i : x_i \neq 0 \} \), and consider the associated normalized monotone set function

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

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

\[
    -\frac{m}{2} \| y - x \|^2 \geq T_l(x, y) \geq -\frac{L}{2} \| y - x \|^2,
\]

and \( f \)'s submodularity ratio \( \gamma \) is lower bounded by \( \frac{m}{2} \). Subsequently, in [12] it is shown that \( \theta \) can also be lower bounded by the same ratio \( \frac{m}{2} \).

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 \( S \subseteq [d] \):

\[
    \min_{|S| \leq \tau, E_S \subseteq S} \max_{\text{supp}(x) \subseteq S \setminus E_S} l(x) - l(0).
\]

By inspecting the bound obtained in Corollary 1, it remains to bound the (super/sub)additive ratio \( \nu \) and \( \tilde{\nu} \). The first bound follows by combining the result \( \gamma \geq \frac{m}{2} \) with Proposition 1: \( \nu \geq \gamma \geq \frac{m}{2} \). To prove the second bound, we make use of the previous result.

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

The second bound follows by combining the result in Proposition 2 and Proposition 1: \( \tilde{\nu} \geq \bar{\gamma} \geq \frac{m}{2} \).

### 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 \( f(x) \) be an unknown function defined over a finite domain \( \mathcal{X} = \{ x_1, \ldots, x_n \} \), where \( x_i \in \mathbb{R}^d \). Once we evaluate the function at some point \( x_i \in \mathcal{X} \), we receive a noisy observation \( y_i = f(x_i) + z \), 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 \( k(x,x') \), i.e. \( f \sim \mathcal{GP}(0, k(x,x')) \). Let \( S = \{ e_1, \ldots, e_{\lvert S \rvert} \} \subseteq [n] \) denote the set of points, and \( X_S := \{ x_{e_1}, \ldots, x_{e_{\lvert S \rvert}} \} \in \mathbb{R}^{\lvert S \rvert \times d} \) and \( y_S := \{ y_{e_1}, \ldots, y_{\lvert S \rvert} \} \) denote the corresponding data matrix and observations, respectively. The posterior distribution of \( f \) given the points \( X_S \) and observations \( y_S \) is again a GP, with the posterior variance given by:

\[
    \sigma^2_{x,S} = k(x,x) - k(x,X_S) (k(X_S, X_S) + \sigma^2 I_{\lvert S \rvert})^{-1} \times k(X_S,x).
\]

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

\[
    F_M(S) := \sum_{x \in X_M} \sigma^2_x - \sigma^2_{x,S}, \quad (17)
\]

where \( \sigma^2_x = k(x,x) \). 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

\[
    \min_{|S| \leq \tau, E_S \subseteq S} \sum_{x \in X_M} \sigma^2_x - \sigma^2_{x,S \setminus E_S}.
\]

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

We can combine this result with Proposition 1, to obtain \( \nu \geq \gamma \geq \frac{\sigma^4}{\sigma^2 + k_{\max}} \) and \( \tilde{\nu} \geq \bar{\gamma} \geq \frac{\sigma^2}{\sigma^2 + k_{\max}} \). Also, we have \( \theta \geq \frac{1}{(\sigma^2 + k_{\max})^2} \) (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
\[\min_{E \subseteq S, |E| \leq \tau} f(S \setminus E).\] 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 \(\tau\) 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:** In order to introduce randomness in the removal process we consider (iii) **Random Greedy Min** — iteratively selects a random element from the top \(\tau\) 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 \((|S|/\tau) \log(1/e)\) 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.

The results are shown in Fig. 6. Since **PRO-GREEDY** and OSU only make sense in the regime where \(\tau\) 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 \(\tau\).

The results are shown in Fig. 3. Logistic regression task with synth. dataset.

**Logistic Regression.** 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 \(\tau\) 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 \(\tau\).
\( \alpha^2 = 0.09 \). 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 \( X_{(i)} \) is set to 1 if \( 1/(1 + \exp(X_{(i)}\beta)) > 0.5 \) 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 \( \tau \) 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 \( \tau \) increases the performance of GREEDY decays significantly and more robust OBLIVIOUS starts to outperform it.

![Figure 4: Logistic regression with MNIST dataset.](image)

**5.2 Robust Batch Bayesian Optimization via Variance Reduction**

**Setup.** We conducted the following synthetic experiment. A design matrix \( X \) of size \( 600 \times 20 \) 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.](image)

**Results.** In Figure 5 (a), (b), (c), the performance of all three algorithms is shown when \( \tau \) is fixed to 50. Different figures correspond to different \( \alpha \) 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 \( \alpha \), OBLIVIOUS-GREEDY outperforms both GREEDY and OBLIVIOUS selection. For larger values of \( \alpha \), 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.
Acknowledgement

The authors would like to thank Jonathan Scarlett and Slobodan Mitrovic for useful discussions. IB and VC’s work was supported in part by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement number 725594), in part by the Swiss National Science Foundation (SNF), project 407540_167319/1, in part by the NCCR MARVEL, funded by the Swiss National Science Foundation.

References


Appendix

Robust Maximization of Non-Submodular Objectives
(Ilija Bogunovic, Junyao Zhao and Volkan Cevher, AISTATS 2018)

A Organization of the Appendix

- Appendix B: Proofs from Section 2
- Appendix C: Proofs of the Main Result (Section 3)
- Appendix D: Proofs from Section 4
- Appendix E: Additional experiments

B Proofs from Section 2

B.1 Proof of Proposition 1

Proof. We prove the following relations:

- \( \nu \geq \gamma, \tilde{\nu} \geq \tilde{\gamma} \):

  By setting \( S = \emptyset \) in both Eq. (4) and Eq. (5), we obtain \( \forall S \subseteq V : \)

  \[
  \sum_{i \in S} f(\{i\}) \geq \gamma f(S),
  \]

  and

  \[
  f(S) \geq \tilde{\gamma} \sum_{i \in S} f(\{i\}).
  \]

  The result follows since, by definition of \( \nu \) and \( \tilde{\nu} \), they are the largest scalars such that Eq. (19) and Eq. (20) hold, respectively.

- \( \gamma \geq 1 - \tilde{\alpha}, \tilde{\gamma} \geq 1 - \alpha \):

  Let \( S, \Omega \subseteq V \) be two arbitrary disjoint sets. We arbitrarily order elements of \( \Omega = \{e_1, \cdots, e_{|\Omega|}\} \) and we let \( \Omega_{j-1} \) denote the first \( j - 1 \) elements of \( \Omega \). We also let \( \Omega_0 \) be an empty set.

  By the definition of \( \tilde{\alpha} \) (see Eq. (7)) we have:

  \[
  \sum_{j=1}^{|\Omega|} f(\{e_j\}|S) = \sum_{j=1}^{|\Omega|} f(\{e_j\}|S \cup \{e_j\} \setminus \{e_j\}) \geq \sum_{j=1}^{|\Omega|} (1 - \tilde{\alpha}) f(\{e_j\}|S \cup \{e_j\} \setminus \{e_j\} \cup \Omega_{j-1}) = (1 - \tilde{\alpha}) f(\Omega|S),
  \]

  where the last equality is obtained via telescoping sums.

  Similarly, by the definition of \( \alpha \) (see Eq. (6)) we have:

  \[
  (1 - \alpha) \sum_{j=1}^{|\Omega|} f(\{e_j\}|S) = \sum_{j=1}^{|\Omega|} (1 - \alpha) f(\{e_j\}|S \cup \{e_j\} \setminus \{e_j\}) \leq \sum_{j=1}^{|\Omega|} f(\{e_j\}|S \cup \{e_j\} \setminus \{e_j\} \cup \Omega_{j-1}) = f(\Omega|S).
  \]

  Because \( S \) and \( \Omega \) are arbitrary disjoint sets, and both \( \gamma \) and \( \tilde{\gamma} \) are the largest scalars such that for all disjoint sets \( S, \Omega \subseteq V \) the following holds \( \sum_{j=1}^{|\Omega|} f(\{e_j\}|S) \geq \gamma f(\Omega|S) \) and \( \tilde{\gamma} \sum_{j=1}^{|\Omega|} f(\{e_j\}|S) \leq f(\Omega|S) \), it follows from Eq. (21) and Eq. (22), respectively, that \( \gamma \geq 1 - \tilde{\alpha} \) and \( \tilde{\gamma} \geq 1 - \alpha \).
B.2 Proof of Remark 1

Proof. Consider any set $S \subseteq V$, and $A$ and $B$ such that $A \cup B = S$, $A \cap B = \emptyset$. We have
\[
\frac{f(A) + f(B)}{f(S)} \geq \frac{\bar{\nu} \sum_{i \in A} f(i) + \bar{\nu} \sum_{i \in B} f(i)}{f(S)} = \frac{\bar{\nu} \sum_{i \in S} f(i)}{f(S)} \geq \nu \bar{\nu},
\]
where the first and second inequality follow by the definition of $\nu$ and $\bar{\nu}$ (Eq. (8) and Eq. (9)), respectively.
By the definition (see Eq. (10)), $\theta$ is the largest scalar such that $f(A) + f(B) \geq \theta f(S)$ holds, hence, it follows $\theta \geq \nu \bar{\nu}$. \[\square\]

C Proofs of the Main Result (Section 3)

C.1 Proof of Lemma 2

We reproduce the proof from [2] for the sake of completeness.

Proof.
\[
f(S \setminus E_S^0) = f(S) - f(S) + f(S \setminus E_S^0)
\]
\[
= f(S_0 \cup S_1) + f(S \setminus E_0) - f(S \setminus E_0) - f(S) + f(S \setminus E_S^0)
\]
\[
= f(S_1) + f(S \setminus E_0) - f(S) + f(S \setminus E_S^0)
\]
\[
= f(S_1) + f(S_0 | (S \setminus S_0)) - f(S) + f(S \setminus E_S^0)
\]
\[
= f(S_1) - f(E_1 | S \setminus E_S^0) + f(S_0 | (S \setminus S_0)) - f(E_0 | (S \setminus E_0))
\]
\[
\geq (1 - \mu) f(S_1),
\]
where we used $S = S_0 \cup S_1$, $E_S^0 = E_0 \cup E_1$, and (23) follows from monotonicity, i.e., $f(S_0 | (S \setminus S_0)) - f(E_0 | (S \setminus E_0)) \geq 0$ (due to $E_0 \subseteq S_0$ and $S \setminus S_0 \subseteq S \setminus E_0$), along with the definition of $\mu$. \[\square\]

C.2 Proof of Lemma 3

Proof. We start by defining $S_0' := \text{OPT}_{(k-\tau, V \setminus E_0)} \cap (S_0 \setminus E_0)$ and $X := \text{OPT}_{(k-\tau, V \setminus E_0)} \setminus S_0'$.
\[
f(S_0 \setminus E_0) + f(\text{OPT}_{(k-\tau, V \setminus S_0)}) \geq f(S_0') + f(X)
\]
\[
\geq \theta f(\text{OPT}_{(k-\tau, V \setminus E_0)})
\]
\[
\geq \theta f(\text{OPT}_{(k-\tau, V \setminus E_S^0)}),
\]
where (24) follows from monotonicity as $S_0' \subseteq (S_0 \setminus E_0)$ and $(V \setminus S_0) \subseteq (V \setminus E_0)$. Eq. (25) follows from the fact that $\text{OPT}_{(k-\tau, V \setminus E_0)} = S_0' \cup X$ and the bipartite subadditive property (10). The final equation follows from the definition of the optimal solution and the fact that $E_S^0 = E_0 \cup E_1$.

By rearranging and noting that $f(S \setminus E_S^0) \geq f(S_0 \setminus E_0)$ due to $(S_0 \setminus E_0) \subseteq (S \setminus E_S^0)$ and monotonicity, we obtain
\[
f(S \setminus E_S^0) \geq \theta f(\text{OPT}_{(k-\tau, V \setminus E_S^0)}) - f(\text{OPT}_{(k-\tau, V \setminus S_0)}).
\]
\[\square\]

C.3 Proof of Theorem 1

Before proving the theorem we outline the following auxiliary lemma:
Lemma 5 (Lemma D.2 in [2]). For any set function \( f \), sets \( A, B \), and constant \( \alpha > 0 \), we have

\[
\max\{\alpha f(A), \beta f(B) - f(A)\} \geq \left(\frac{\alpha}{1 + \alpha}\right) \beta f(B).
\]

Next, we prove the main theorem.

Proof. First we note that \( \beta \) should be chosen such that the following condition holds \( |S_0| = \lfloor \beta \tau \rfloor \leq k \). When \( \tau = \lfloor ck \rfloor \) for \( c \in (0, 1) \) and \( k \to \infty \) the condition \( \beta < \frac{1}{c} \) suffices.

We consider two cases, when \( \mu = 0 \) and \( \mu \neq 0 \). When \( \mu = 0 \), from Lemma 2 we have

\[
f(S \setminus E^*_S) \geq f(S_1)
\]

On the other hand, when \( \mu \neq 0 \), by Lemma 2 and 4 we have

\[
f(S \setminus E^*_S) \geq \max\{(1 - \mu) f(S_1), (\beta - 1) \bar{\nu}(1 - \bar{\alpha}) f(S_1)\} \geq \frac{(\beta - 1) \bar{\nu}(1 - \bar{\alpha})}{1 + (\beta - 1) \bar{\nu}(1 - \bar{\alpha})} f(S_1).
\]

By denoting \( P := \frac{(\beta - 1) \bar{\nu}(1 - \bar{\alpha})}{1 + (\beta - 1) \bar{\nu}(1 - \bar{\alpha})} \) we observe that \( P \in [0, 1] \) once \( \beta \geq 1 \). Hence, by setting \( \beta \geq 1 \) and taking the minimum between two bounds in Eq. (29) and Eq. (28) we conclude that Eq. (29) holds for any \( \mu \in [0, 1] \).

By combining Eq. (29) with Lemma 1 we obtain

\[
f(S \setminus E^*_S) \geq P \left(1 - e^{-\gamma \frac{k - \lfloor \beta \tau \rfloor}{k}}\right) f(OPT_{(k-\tau, V \setminus S_0)}).
\]

By further combining this with Lemma 3 we have

\[
f(S \setminus E^*_S) \geq \frac{\theta P \left(1 - e^{-\gamma \frac{k - \lfloor \beta \tau \rfloor}{k}}\right)}{1 + P \left(1 - e^{-\gamma \frac{k - \lfloor \beta \tau \rfloor}{k}}\right)} f(OPT_{(k-\tau, V \setminus E^*_S)}).
\]

where the second inequality follows from Lemma 5. By plugging in \( \tau = \lfloor ck \rfloor \) we further obtain

\[
f(S \setminus E^*_S) \geq \frac{\theta P \left(1 - e^{-\gamma \frac{k - \beta \lfloor ck \rfloor - 1}{k}}\right)}{1 + P \left(1 - e^{-\gamma \frac{k - \beta \lfloor ck \rfloor - 1}{k}}\right)} f(OPT_{(k-\tau, V \setminus E^*_S)})
\]

\[
\geq \theta \frac{P \left(1 - e^{-\gamma \frac{1 - \beta c - 1}{1 - c}}\right)}{1 + P \left(1 - e^{-\gamma \frac{1 - \beta c - 1}{1 - c}}\right)} f(OPT_{(k-\tau, V \setminus E^*_S)})
\]

Finally, Remark 2 follows from Eq. (30) when \( \tau \in o\left(\frac{k}{\beta}\right) \) and \( \beta \geq \log k \) (note that the condition \( |S_0| = \lfloor \beta \tau \rfloor \leq k \) is thus satisfied), as \( k \to \infty \), we have both \( \frac{k - \lfloor \beta \tau \rfloor}{k} \to 1 \) and \( P = \frac{(\beta - 1) \bar{\nu}(1 - \bar{\alpha})}{1 + (\beta - 1) \bar{\nu}(1 - \bar{\alpha})} \to 1 \), when \( \bar{\nu} \in (0, 1] \) and \( \bar{\alpha} \in [0, 1) \).

\[\square\]
C.4 Proof of Corollary 1

To prove this result we need the following two lemmas that can be thought of as the alternative to Lemma 2 and 4.

**Lemma 6.** Let $\mu' \in [0, 1]$ be a constant such that $f(E_1) = \mu' f(S_1)$ holds. Consider $f(\cdot)$ with bipartite subadditivity ratio $\theta \in [0, 1]$ defined in Eq. (4). Then

\[
f(S \setminus E_0^*) \geq (\theta - \mu') f(S_1). \tag{32}
\]

**Proof.** By the definition of $\theta$, $f(S_1 \setminus E_1) + f(E_1) \geq \theta f(S_1)$. Hence,

\[
f(S \setminus E_0^*) \geq f(S_1 \setminus E_1)
\geq \theta f(S_1) - f(E_1)
= (\theta - \mu') f(S_1).
\]

\[ \square \]

**Lemma 7.** Let $\beta$ be a constant such that $|S_0| \leq |\beta \tau|$ and $|S_0| \leq k$, and let $\hat{\nu}, \nu \in [0, 1]$ be superadditivity and subadditivity ratio (Eq. (9) and Eq. (8), respectively). Finally, let $\mu'$ be a constant defined as in Lemma 6. Then,

\[
f(S \setminus E_0^*) \geq (\beta - 1)\hat{\nu}\nu f(S_1). \tag{33}
\]

**Proof.** The proof follows that of Lemma 4, with two modifications. In Eq. (34) we used the subadditive property of $f(\cdot)$, and Eq. (35) follows by the definition of $\mu'$.

\[
f(S \setminus E_0^*) \geq f(S_0 \setminus E_0)
\geq \hat{\nu} \sum_{e_i \in S_0 \setminus E_0} f(\{e_i\})
\geq \frac{|S_0 \setminus E_0|}{|E_1|} \hat{\nu} \sum_{e_i \in E_1} f(\{e_i\})
\geq (\beta - 1)\tau \hat{\nu} \sum_{e_i \in E_1} f(\{e_i\})
\geq (\beta - 1)\hat{\nu}\nu f(E_1)
= (\beta - 1)\hat{\nu}\nu f(S_1). \tag{34}
\]

\[
f(S \setminus E_0^*) \geq (\beta - 1)\hat{\nu}\nu f(S_1).
\]

\[ \square \]

Next we prove the main corollary. The proof follows the steps of the proof from Appendix C.3, except that here we make use of Lemma 6 and 7.

**Proof.** We consider two cases, when $\mu' = 0$ and $\mu' \neq 0$. When $\mu' = 0$, from Lemma 6 we have

\[
f(S \setminus E_0^*) \geq \theta f(S_1).
\]

On the other hand, when $\mu' \neq 0$, by Lemma 6 and 7 we have

\[
f(S \setminus E_0^*) \geq \max\{((\theta - \mu') f(S_1), (\beta - 1)\hat{\nu}\nu f(S_1)\}
\geq \theta - (\beta - 1)\hat{\nu}\nu f(S_1).
\]

By denoting $P := \frac{(\beta - 1)\hat{\nu}\nu}{1 + (\beta - 1)\hat{\nu}\nu}$ and observing that $P \in [0, 1)$ once $\beta \geq 1$, we conclude that Eq. (36) holds for any $\mu' \in [0, 1]$ once $\beta \geq 1$.

By combining Eq. (36) with Lemma 1 we obtain

\[
f(S \setminus E_0^*) \geq \theta P \left(1 - e^{-\gamma \frac{1 - (\beta - 1)}{1 + (\beta - 1)\hat{\nu}\nu}}\right) f(OPT_{(k-\tau, \nu \setminus S_0)}).
\]

(37)
Moreover, for $B^*$ 

By further combining this with Lemma 3 we have

$$
\begin{align*}
    f(S \setminus E^*_B) \geq & \max \{ \theta f(\text{OPT}(k - r, V \setminus E^*_B)) - f(\text{OPT}(k - r, V \setminus S_0)), \theta P \left( 1 - e^{-\gamma \frac{k - \lfloor \beta k \rfloor}{k - r}} \right) f(\text{OPT}(k - r, V \setminus S_0)) \} \\
    \geq & \frac{\theta^2 P \left( 1 - e^{-\gamma \frac{k - \lfloor \beta k \rfloor}{k - r}} \right)}{1 + \theta P \left( 1 - e^{-\gamma \frac{k - \lfloor \beta k \rfloor}{k - r}} \right)} f(\text{OPT}(k - r, V \setminus E^*_B)),
\end{align*}
$$

(38)

where the second inequality follows from Lemma 5. By plugging in $\tau = \lceil ck \rceil$ in the last equation and by letting $k \to \infty$ we arrive at:

$$
    f(S \setminus E^*_B) \geq \frac{\theta^2 P \left( 1 - e^{-\gamma \frac{k - \lfloor \beta k \rfloor}{k - r}} \right)}{1 + \theta P \left( 1 - e^{-\gamma \frac{k - \lfloor \beta k \rfloor}{k - r}} \right)} f(\text{OPT}(k - r, V \setminus E^*_B)).
$$

Finally, from Eq. (38), when $\tau \in o \left( \frac{k}{\beta} \right)$ and $\beta \geq \log k$, as $k \to \infty$, we have both $\frac{k - \lfloor \beta k \rfloor}{k - r} \to 1$ and $P = \frac{(\beta - 1) \nu}{1 + (\beta - 1) \nu} \to 1$ (when $\nu, \nu \in (0, 1]$). It follows

$$
    f(S \setminus E^*_B) \xrightarrow{k \to \infty} \frac{\theta^2 (1 - e^{-\gamma})}{1 + \theta (1 - e^{-\gamma})} f(\text{OPT}(k - r, V \setminus E^*_B)).
$$

\[ \square \]

\section{D Proofs from Section 4}

\subsection{D.1 Proof of Proposition 2}

\textit{Proof.} The goal is to prove: $\hat{\gamma} \geq \frac{m}{T}$.

Let $S \subseteq [d]$ and $\Omega \subseteq [d]$ be any two disjoint sets, and for any set $A \subseteq [d]$ let $x^{(A)} = \arg\max_{\text{supp}(x) \subseteq A, x \in \mathcal{X}} l(x)$. Moreover, for $B \subseteq [d]$ let $x^{(A)}_B$ denote those coordinates of vector $x^{(A)}$ that correspond to the indices in $B$.

We proceed by upper bounding the denominator and lower bounding the numerator in (5). By definition of $x^{(S)}$ and strong concavity of $l(\cdot)$,

$$
    l(x^{(S \cup i)}) - l(x^{(S)}) \leq \langle \nabla l(x^{(S)}), x^{(S \cup i)} - x^{(S)} \rangle - \frac{m}{2} \left\| x^{(S \cup i)} - x^{(S)} \right\|^2
    \leq \max_{v, v^{(S \cup i)} \in \mathcal{X}} \langle \nabla l(x^{(S)}), v - x^{(S)} \rangle - \frac{m}{2} \left\| v - x^{(S)} \right\|^2
    = \frac{1}{2m} \left\| \nabla l(x^{(S)}) \right\|^2
$$

where the last equality follows by plugging in the maximizer $v = x^{(S)} + \frac{1}{m} \nabla l(x^{(S)})$. Hence,

$$
    \sum_{i \in \Omega} \left( l(x^{(S \cup i)}) - l(x^{(S)}) \right) \leq \sum_{i \in \Omega} \frac{1}{2m} \left\| \nabla l(x^{(S)}) \right\|^2 = \frac{1}{2m} \left\| \nabla l(x^{(S)}) \right\| \Omega^2.
$$

On the other hand, from the definition of $x^{(S \cup i)}$ and due to smoothness of $l(\cdot)$ we have

$$
    l(x^{(S \cup i)}) - l(x^{(S)}) \geq l(x^{(S)}) + \frac{1}{L} \nabla l(x^{(S)}) \Omega - l(x^{(S)})
    \geq \langle \nabla l(x^{(S)}), \frac{1}{L} \nabla l(x^{(S)}) \Omega \rangle - \frac{L}{2} \left\| \frac{1}{L} \nabla l(x^{(S)}) \right\|^2
    = \frac{1}{2L} \left\| l(x^{(S)}) \right\|^2.
$$
It follows that
\[
\frac{l(x(S\cup\Omega)) - l(x(S))}{\sum_{i\in\Omega} (l(x(S\cup\{i\})) - l(x(S)))} \geq \frac{m}{L}, \quad \forall \text{ disjoint } S, \Omega \subseteq [d]
\]

We finish the proof by noting that \(\tilde{\gamma}\) is the largest constant for the above statement to hold.

\[
\square
\]

### D.2 Variance Reduction in GPs

#### D.2.1 Non-submodularity of Variance Reduction

The goal of this section is to show that the GP variance reduction objective is not submodular in general. Consider the following PSD kernel matrix:

\[
K = \begin{bmatrix}
1 & \sqrt{1-z^2} & 0 \\
\sqrt{1-z^2} & 1 & z^2 \\
0 & z^2 & 1
\end{bmatrix}.
\]

We consider a single \(x = \{3\}\) (i.e. \(M\) is a singleton) that corresponds to the third data point. The objective is as follows:

\[
F(i|S) = \sigma^2_{\{3\}|S} - \sigma^2_{\{3\}|S\cup i}.
\]

The submodular property implies \(F(\{1\}) \geq F(\{1\}|\{2\})\). We have:

\[
F(\{1\}) = \sigma^2_{\{3\}} - \sigma^2_{\{3\}|\{1\}}
= 1 - K(\{3\}, \{3\}) - K(\{3\}, \{1\})(K(\{1\}, \{1\}) + \sigma^2)^{-1}K(\{1\}, \{3\})
= 1 - 1 + 0 = 0,
\]

and

\[
F(\{2\}) = \sigma^2_{\{1\}} - \sigma^2_{\{3\}|\{2\}}
= 1 - K(\{3\}, \{3\}) - K(\{3\}, \{2\})(K(\{2\}, \{2\}) + \sigma^2)^{-1}K(\{2\}, \{3\})
= 1 - (1 - \sigma^2(1 + \sigma^2)^{-1}z^2) = \frac{z^4}{1 + \sigma^2},
\]

and

\[
F(\{1, 2\}) = \sigma^2_{\{3\}} - \sigma^2_{\{3\}|\{1, 2\}}
= 1 - K(\{3\}, \{3\}) + [K(\{3\}, \{1\}), K(\{3\}, \{2\})] \begin{bmatrix} 1 + \sigma^2, K(\{2\}, \{1\}) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\
K(\{1\}, \{2\}), 1 + \sigma^2
\end{bmatrix}
= 1 - 1 + [0, z^2] \begin{bmatrix} 1 + \sigma^2, \sqrt{1-z^2} \\
\sqrt{1-z^2}, 1 + \sigma^2
\end{bmatrix}^{-1} \begin{bmatrix} 0 \\
z^2
\end{bmatrix}
= \frac{z^4(1 + \sigma^2)}{(1 + \sigma^2)^2 - (1 - z^2)},
\]

We obtain,

\[
F(\{1\}|\{2\}) = F(\{1, 2\}) - F(\{2\})
= \frac{z^4}{(1 + \sigma^2) - (1 - z^2)(1 + \sigma^2)^{-1}} - \frac{z^4}{1 + \sigma^2}.
\]

When \(z \in (0, 1)\), \(F(\{1\}|\{2\})\) is strictly greater than 0, and hence greater than \(F(\{1\})\). This is in contradiction with the submodular property which implies \(F(\{1\}) \geq F(\{1\}|\{2\})\).
D.2.2 Proof of Proposition 3

Proof. We are interested in lower bounding the following ratios: \( \frac{f((i)|S \setminus \{i\} \cup \Omega)}{f((i)|S \setminus \{i\})} \) and \( \frac{f((i)|S \setminus \{i\} \cup \Omega)}{f((i)|S \setminus \{i\})} \).

Let \( k_{\text{max}} \in \mathbb{R}_+ \) be the largest variance, i.e. \( k(x_i, x_i) \leq k_{\text{max}} \) for every \( i \). Consider the case when \( M \) is a singleton set:

\[
f(i|S) = \sigma^2_{x|S} - \sigma^2_{x|i,S}.\]

By using \( \Omega = \{i\} \) in Eq. (39), we can rewrite \( f(i|S) \) as

\[
f(i|S) = a_i^2 B_i^{-1},\]

where \( a_i, B_i \in \mathbb{R}_+ \), and are given by:

\[a_i = k(x, x_i) - k(x, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1}k(X_S, x_i)\]

and

\[B_i = \sigma^2 + k(x_i, x_i) - k(x_i, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1}k(X_S, x_i).\]

By using the fact that \( k(x_i, x_i) \leq k_{\text{max}}, \) for every \( i \) and \( S \), we can upper bound \( B_i \) by \( \sigma^2 + k_{\text{max}} \) (note that \( k(x_i, x_i) - k(x_i, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1}k(X_S, x_i) \geq 0 \) as variance cannot be negative), and lower bound by \( \sigma^2 \).

It follows that for every \( i \) and \( S \) we have:

\[
\frac{a_i^2}{\sigma^2 + k_{\text{max}}} \leq f(i|S) \leq \frac{a_i^2}{\sigma^2}.\]

Therefore,

\[
\frac{f((i)|S \setminus \{i\} \cup \Omega)}{f((i)|S \setminus \{i\})} \geq \frac{a_i^2/\sigma^2 + k_{\text{max}}}{a_i^2/\sigma^2} = \frac{\sigma^2}{\sigma^2 + k_{\text{max}}}, \quad \forall S, \Omega \subseteq V, i \in S \setminus \Omega,
\]

\[
\frac{f((i)|S \setminus \{i\})}{f((i)|S \setminus \{i\} \cup \Omega)} \geq \frac{a_i^2/\sigma^2 + k_{\text{max}}}{a_i^2/\sigma^2} = \frac{\sigma^2}{\sigma^2 + k_{\text{max}}}, \quad \forall S, \Omega \subseteq V, i \in S \setminus \Omega.
\]

It follows:

\[
(1 - \alpha) \geq \frac{\sigma^2}{\sigma^2 + k_{\text{max}}}, \quad \text{and}
\]

\[
(1 - \hat{\alpha}) \geq \frac{\sigma^2}{\sigma^2 + k_{\text{max}}}.
\]

The obtained result also holds for any set \( M \subseteq [n] \). \( \square \)

D.2.3 Alternative GP variance reduction form

Here, the goal is to show that the variance reduction can be written as

\[
F(\Omega|S) = \sigma^2_{x|S} - \sigma^2_{x|i,S,\Omega} = aB^{-1}a^T,
\]

where \( a \in \mathbb{R}_+^{1 \times |\Omega\setminus S|} \), \( B \in \mathbb{R}_+^{|\Omega\setminus S| \times |\Omega\setminus S|} \) and are given by:

\[a := k(x, X_{\Omega\setminus S}) - k(x, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1}k(X_S, X_{\Omega\setminus S}),\]

and

\[B := \sigma^2 I + k(X_{\Omega\setminus S}, X_{\Omega\setminus S}) - k(X_{\Omega\setminus S}, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1}k(X_S, X_{\Omega\setminus S}).\]

This form is used in the proof in Appendix D.2.2.
Proof. Recall the definition of the posterior variance:

\[ \sigma^2_{x|S} = k(x, x) - k(x, X_S) \left( k(X_S, X_S) + \sigma^2 I_{|S|} \right)^{-1} k(X_S, x). \]

We have

\[
F(\Omega|S) = \sigma^2_{x|S} - \sigma^2_{x|S\cup\Omega} \\
= k(x, X_{S\cup\Omega}) \left( k(X_{S\cup\Omega}, X_{S\cup\Omega}) + \sigma^2 I_{|S\cup\Omega|} \right)^{-1} k(X_{S\cup\Omega}, x) - k(x, X_S) \left( k(X_S, X_S) + \sigma^2 I_{|S|} \right)^{-1} k(X_S, x)
\]

where we use the following notation:

\[
m_1 := k(x, X_S), \\
m_2 := k(x, X_{\Omega\setminus S}), \\
A_{11} := k(X_S, X_S) + \sigma^2 I_{|S|}, \\
A_{12} := k(X_S, X_{\Omega\setminus S}), \\
A_{21} := k(X_{\Omega\setminus S}, X_S), \\
A_{22} := k(X_{\Omega\setminus S}, X_{\Omega\setminus S} + \sigma^2 I_{|\Omega\setminus S|}).
\]

By using the inverse formula [39, Section 9.1.3] we obtain:

\[
F(\Omega|S) = [m_1, m_2] \left[ \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1} A_{12} B^{-1} A_{21} \end{bmatrix}, -B^{-1} A_{21} A_{11}^{-1} \right] \begin{bmatrix} m_1^T \\ m_2^T \end{bmatrix} - m_1 A_{11}^{-1} m_1^T,
\]

where

\[
B := A_{22} - A_{21} A_{11}^{-1} A_{12}.
\]

Finally, we obtain:

\[
F(\Omega|S) = m_1 A_{11}^{-1} m_1^T + m_1 A_{11}^{-1} A_{12} B^{-1} A_{21} A_{11}^{-1} m_1^T - m_2 B^{-1} A_{21} A_{11}^{-1} m_1^T \\
- m_1 A_{11}^{-1} A_{12} B^{-1} m_2^T + m_2 B^{-1} m_2^T - m_1 A_{11}^{-1} m_1^T \\
= m_1 A_{11}^{-1} A_{12} B^{-1} (A_{21} A_{11}^{-1} m_1^T - m_2^T) - m_2 B^{-1} (A_{21} A_{11}^{-1} m_1^T - m_2^T) \\
= (m_1 A_{11}^{-1} A_{12} - m_2) B^{-1} (A_{21} A_{11}^{-1} m_1^T - m_2^T) \\
= (m_2 - m_1 A_{11}^{-1} A_{12}) B^{-1} (m_2^T - A_{21} A_{11}^{-1} m_1^T).
\]

By setting

\[
a := m_2 - m_1 A_{11}^{-1} A_{12} \\
= k(x, X_{\Omega\setminus S}) - k(x, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1} k(X_S, X_{\Omega\setminus S})
\]

and

\[
a^T := m_2^T - A_{21} A_{11}^{-1} m_1^T \\
= k(X_{\Omega\setminus S}, x) - k(X_{\Omega\setminus S}, X_S)(k(X_S, X_S) + \sigma^2 I)^{-1} k(X_S, x)
\]

we have

\[
F(\Omega|S) = a B^{-1} a^T,
\]

where

\[
B = \sigma^2 I_{|\Omega\setminus S|} + k(X_{\Omega\setminus S}, X_{\Omega\setminus S}) - k(X_{\Omega\setminus S}, X_S)(k(X_S, X_S) + \sigma^2 I_{|S|})^{-1} k(X_S, X_{\Omega\setminus S}).
\]

\[\square\]
E Additional Experiments

![Graphs showing additional experiments for comparison of the algorithms on support selection task.](image)

Figure 6: Additional experiments for comparison of the algorithms on support selection task.