Guarantees for Greedy Maximization of Non-submodular Functions with Applications

Andrew An Bian [1] **Joachim M. Buhmann** [1] **Andreas Krause** [1] **Sebastian Tschiatschek** [1]

Abstract

We investigate the performance of the standard GREEDY algorithm for cardinality constrained maximization of non-submodular nondecreasing set functions. While there are strong theoretical guarantees on the performance of GREEDY for maximizing submodular functions, there are few guarantees for non-submodular ones. However, GREEDY enjoys strong empirical performance for many important non-submodular functions, e.g., the Bayesian A-optimality objective in experimental design. We prove theoretical guarantees supporting the empirical performance. Our guarantees are characterized by a combination of the (generalized) *curvature* α and the *submodularity ratio* γ. In particular, we prove that GREEDY enjoys a *tight* approximation guarantee of $\frac{1}{\alpha}(1-e^{-\gamma\alpha})$ for cardinality constrained maximization. In addition, we bound the submodularity ratio and curvature for several important real-world objectives, including the Bayesian A-optimality objective, the determinantal function of a square submatrix and certain linear programs with combinatorial constraints. We experimentally validate our theoretical findings for both synthetic and real-world applications.

1. Introduction

Many important problems, such as experimental design and sparse modeling, are naturally formulated as a subset selection problem, where a set function $F(S)$ over a K-cardinality constraint is maximized, i.e.,

$$\max_{S\subseteq\mathcal{V},|S|\leq K} F(S), \tag{P}$$

where $\mathcal{V}=\{v_1,\dots,v_n\}$ is the ground set. Specifically, in experimental design, the goal is to select a set of experiments to perform such that some statistical criterion is optimized. This problem arises naturally in domains where performing experiments is costly. In sparse modeling, the task is to identify sparse representations of signals, enabling interpretability and robustness in high-dimensional statistical problems—properties that are crucial in modern data analysis.

Frequently, the standard GREEDY algorithm (Alg. 1) is used to (approximately) solve (P). For the case that $F(S)$ is a monotone nondecreasing *submodular* set function[1], the GREEDY algorithm enjoys the multiplicative approximation guarantee of $(1-1/e)$ (Nemhauser et al., 1978; Vondrák, 2008; Krause & Golovin, 2014). This constant factor can be improved by refining the characterization of the objective using the *curvature* (Conforti & Cornuéjols, 1984; Vondrák, 2010; Iyer et al., 2013), which informally quantifies how *close* a submodular function is to being modular (i.e., $F(S)$ and $-F(S)$ are submodular).

Algorithm 1: The GREEDY Algorithm

Input: Ground set $\mathcal{V}$, set function $F:2^{\mathcal{V}}\to\mathbb{R}_+$, budget K
$S^0\leftarrow\emptyset$
for $t=1,\dots,K$ **do**
  $v^*\leftarrow\arg\max_{v\in\mathcal{V}\setminus S^{t-1}} F(S^{t-1}\cup\{v\})-F(S^{t-1})$
  $S^t\leftarrow S^{t-1}\cup\{v^*\}$
Output: S^K

However, for many applications, including experimental design and sparse Gaussian processes (Lawrence et al., 2003), $F(S)$ is in general not submodular (Krause et al., 2008) and the above guarantee does not hold. In practice, however, the standard GREEDY algorithm often achieves very good performance on these applications, e.g., in subset selection with the R^2 (squared multiple correlation) ob-

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland. Correspondence to: Joachim M. Buhmann <jbuhmann@inf.ethz.ch>, Andreas Krause <krausea@ethz.ch>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]$F(\cdot)$ is monotone nondecreasing if $\forall A\subseteq\mathcal{V}, v\in\mathcal{V}$, $F(A\cup\{v\})\geq F(A)$. $F(\cdot)$ is submodular iff it satisfies the diminishing returns property $F(A\cup\{v\})-F(A)\geq F(B\cup\{v\})-F(B)$ for all $A\subseteq B\subseteq\mathcal{V}\setminus\{v\}$. Assume wlog. that $F(\cdot)$ is normalized, i.e., $F(\emptyset)=0$.

jective (Das & Kempe, 2011). To explain the good empirical performance, Das & Kempe (2011) proposed the *submodularity ratio*, a quantity characterizing how *close* a set function is to being submodular.

Another important class of non-submodular set functions comes as the auxiliary function when optimizing a continuous function $f(\boldsymbol{x})$ s.t. combinatorial constraints, i.e., $\min_{\boldsymbol{x}\in\mathcal{C},\text{supp}(\boldsymbol{x})\in\mathcal{I}} f(\boldsymbol{x})$, where $\text{supp}(\boldsymbol{x}) := \{i \mid x_i \neq 0\}$ is the support set of $\boldsymbol{x}$, $\mathcal{C}$ is a convex set, and $\mathcal{I}$ is the independent sets of the combinatorial structure. One of the most popular ways to solve this problem is to use the GREEDY algorithm to maximize the auxiliary function $F(S) := \max_{\boldsymbol{x}\in\mathcal{C},\text{supp}(\boldsymbol{x})\subseteq S} -f(\boldsymbol{x})$. This setting covers various important applications, to name a few, feature selection (Guyon & Elisseeff, 2003), sparse approximation (Das & Kempe, 2008; Krause & Cevher, 2010), sparse recovery (Candes et al., 2006), sparse M-estimation (Jain et al., 2014), linear programming (LP) with combinatorial constraints, and column subset selection (Altschuler et al., 2016). Recently, Elenberg et al. (2016) proved that if $f(\boldsymbol{x})$ has L-restricted smoothness and m-restricted strong convexity, then the submodularity ratio of $F(S)$ is lower bounded by m/L. This result significantly enlarges the domain where the GREEDY algorithm can be applied.

In this paper, we combine and generalize the ideas of *curvature* and *submodularity ratio* to derive improved constant factor approximation guarantees of the GREEDY algorithm. Our guarantees allow us to better characterize the empirical success of applying GREEDY on a significantly larger class of non-submodular functions. Furthermore, we bound these characteristics for important applications, rendering the usage of GREEDY a principled choice rather than a mere heuristic. Our main contributions are:

- We prove the *first tight* constant-factor approximation guarantees for GREEDY on maximizing non-submodular nondecreasing set functions s.t. a cardinality constraint, characterized by a novel *combination* of the (generalized) notions of submodularity ratio γ and curvature α.
- By theoretically bounding parameters (γ, α) for several important objectives, including Bayesian A-optimality in experimental design, the determinantal function of a square submatrix and maximization of LPs with combinatorial constraints, our theory implies the *first* guarantees for them.
- Lastly, we experimentally validate our theory on several real-world applications. It is worth noting that for the Bayesian A-optimality objective, GREEDY generates comparable solutions as the classically used semidefinite programming (SDP) based method, but is usually two orders of magnitude faster.

Notation. We use boldface letters, e.g., $\boldsymbol{x}$, to represent vectors, and capital boldface letters, e.g., $\mathbf{A}$, to denote matrices. x_i is the i^{th} entry of the vector $\boldsymbol{x}$. We refer to $\mathcal{V} = \{v_1, ..., v_n\}$ as the ground set. We use $f(\cdot)$ to denote a continuous function, and $F(\cdot)$ to represent a set function. $\text{supp}(\boldsymbol{x}) := \{i \in \mathcal{V} \mid x_i \neq 0\}$ is the support set of the vector $\boldsymbol{x}$, and $[n] := \{1, ..., n\}$ for an integer $n \geq 1$. We denote the marginal gain of a set $\Omega \subseteq \mathcal{V}$ in context of a set $S \subseteq \mathcal{V}$ as $\rho_\Omega(S) := F(\Omega \cup S) - F(S)$. For $v \in \mathcal{V}$, we use the shorthand $\rho_v(S)$ for $\rho_{\{v\}}(S)$.

2. Submodularity Ratio and Curvature

In this section we provide the *submodularity ratio* and *curvature* for general, not necessarily submodular functions[2], they are natural extensions of the classical ones. Let $S^0 = \emptyset$, $S^t = \{j_1, ..., j_t\}, t = 1, ..., K$ be the successive sets chosen by GREEDY. For brevity, let $\rho_t := \rho_{j_t}(S^{t-1})$ be the marginal gain of GREEDY in step t.

Definition 1 (Submodularity ratio (Das & Kempe, 2011))**.** *The* submodularity ratio *of a non-negative set function $F(\cdot)$ is the largest scalar γ s.t.*

$$\sum\nolimits_{\omega\in\Omega\setminus S} \rho_\omega(S) \geq \gamma\rho_\Omega(S), \forall\, \Omega, S \subseteq \mathcal{V}.$$

The greedy submodularity ratio *is the largest scalar γ^G s.t.*

$$\sum_{\omega\in\Omega\setminus S^t} \rho_\omega(S^t) \geq \gamma^G\rho_\Omega(S^t), \forall |\Omega| = K, t = 0, \ldots, K-1.$$

It is easy to see that $\gamma^G \geq \gamma$. The submodularity ratio measures to what extent $F(\cdot)$ has submodular properties. We make the following observations:

Remark 1. *For a nondecreasing function $F(\cdot)$, it holds a) $\gamma, \gamma^G \in [0, 1]$; b) $F(\cdot)$ is submodular iff $\gamma = 1$.*

Definition 2 (Generalized curvature)**.** *The* curvature *of a non-negative function $F(\cdot)$ is the smallest scalar α s.t.*

$$\rho_i(S \setminus \{i\} \cup \Omega) \geq (1-\alpha)\rho_i(S \setminus \{i\}),$$
$$\forall\, \Omega, S \subseteq \mathcal{V}, i \in S\setminus\Omega.$$

The greedy curvature *is the smallest scalar $\alpha^G \geq 0$ s.t.*

$$\rho_{j_i}(S^{i-1} \cup \Omega) \geq (1-\alpha^G)\rho_{j_i}(S^{i-1}),$$
$$\forall\, \Omega : |\Omega| = K, i : j_i \in S^{K-1}\setminus\Omega.$$

[2]Curvature is commonly defined for submodular functions. Sviridenko et al. (2013) presented a notion of curvature for monotone non-submodular functions. We show in Appendix C the details of these notions and the relations to ours. Additionally, we prove in Remark 3 of Appendix C.2 that our combination of curvature and submodularity ratio is more expressive than that of Sviridenko et al. (2013) in characterizing the maximization of problem (P) using standard GREEDY.

When $K = n$ or 1, $S^{K-1} \backslash \Omega = \emptyset$, it is natural to define $\alpha^G = 0$. It is easy to observe that $\alpha^G \leq \alpha$. Note that the classical *total* curvature is $\alpha^{\text{total}} := 1 - \min_{i \in \mathcal{V}} \frac{\rho_i(\mathcal{V} \backslash \{i\})}{\rho_i(\emptyset)}$.

Remark 2. *For a nondecreasing function $F(\cdot)$, it holds: a) $\alpha, \alpha^G \in [0, 1]$; b) $F(\cdot)$ is* supermodular *iff $\alpha = 0$; c) If $F(\cdot)$ is submodular, then $\alpha^G \leq \alpha = \alpha^{\text{total}}$.*

So for a submodular function, our notion of curvature is consistent with α^{total}. Notably, α^G usually characterizes the problem better than α^{total}, as will be validated in Section 5.

3. Approximation Guarantee

We present approximation guarantee of GREEDY in Theorem 1. Note that both versions of the submodularity ratio and curvature apply in the proof. For brevity, we use γ and α to refer to any of these versions in the sequel. In Section 3.3 we prove tightness of the approximation guarantees. All omitted proofs are given in Appendix B.

Theorem 1. *Let $F(\cdot)$ be a non-negative nondecreasing set function with submodularity ratio $\gamma \in [0, 1]$ and curvature $\alpha \in [0, 1]$. The* GREEDY *algorithm enjoys the following approximation guarantee for solving problem* (P)*:*

$$F(S^K) \geq \frac{1}{\alpha}\left[1 - \left(\frac{K - \alpha\gamma}{K}\right)^K\right] F(\Omega^*) \geq \frac{1}{\alpha}(1 - e^{-\alpha\gamma}) F(\Omega^*), \quad (1)$$

where Ω^* is the optimal solution of (P) *and S^K the output of the* GREEDY *algorithm.*[3]

3.1. Interpreting Theorem 1

Before proving the theorem, we want to give the reader an intuition of the results and show how our results recover and extend several classical guarantees for the GREEDY algorithm. For the case $\alpha = 0$ (i.e., $F(\cdot)$ is *supermodular*), the approximation guarantee is $\lim_{\alpha \to 0} \frac{1}{\alpha}(1 - e^{-\alpha\gamma}) = \gamma$, which gives the first guarantee of greedily maximizing a nondecreasing supermodular function with bounded γ. When $\gamma = 1$, (i.e., $F(\cdot)$ is submodular), we recover the guarantee of $\alpha^{-1}(1 - e^{-\alpha})$ (Conforti & Cornuéjols, 1984). For the case $\alpha = 1$, we have a guarantee of $(1 - e^{-\gamma})$ (Das & Kempe, 2011). For the case $\alpha = 1, \gamma = 1$, we recover the classical guarantee of $(1 - 1/e)$ (Nemhauser et al., 1978). We plot the constant-factor approximation guarantees for different values of γ and α in Fig. 1. One interesting phenomenon is that γ and α play different roles: Looking at $\gamma = 0$, the approximation factor is always 0, independent of the value α takes. In contrast, for $\alpha = 0$, the approximation guarantee is $(1 - e^{-\gamma})$. This can be interpreted as the curvature *boosting* the guarantees.

[3] For the setting that GREEDY is allowed to pick more than K elements, e.g., pick $K' > K$ elements, our theory can be easily extended to show that $F(S^{K'}) \geq \alpha^{-1}(1 - e^{-\alpha\gamma K'/K}) F(\Omega^*)$.

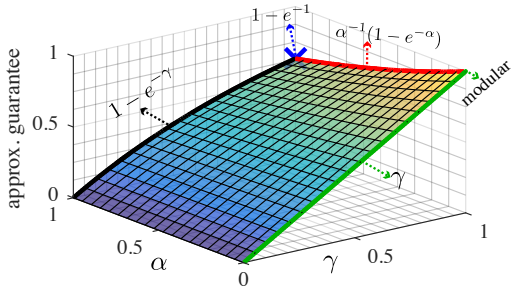


Figure 1: Approximation guarantee $\frac{1}{\alpha}(1 - e^{-\alpha\gamma})$. The blue cross marks the classical $(1 - 1/e)$-guarantee of GREEDY. The *red* line illustrates the influence of the curvature on the guarantees for submodular functions, and the *black* line illustrates the influence of γ on the guarantees for the worst-case curvature $\alpha = 1$. The *green* line is the guarantees for K-cardinality constrained *supermodular* maximization.

3.2. Proof of Theorem 1

The high-level proof framework is based on Conforti & Cornuéjols (1984) (where they derive the approximation guarantee for maximizing a nondecreasing *submodular* function with bounded curvature). However, adapting the proof to non-submodular functions requires several changes detailed in Section 6.

Proof overview. Let us denote all problem instances of maximizing a non-negative nondecreasing function $F(\cdot)$ s.t. K-cardinality constraint ($\max_{|S| \leq K} F(S)$) to be $\mathcal{P}_{K,\alpha,\gamma}$, where $F(\cdot)$ is parametrized by submodularity ratio γ and curvature α. Let $P_{\Omega^*, S^K} \in \mathcal{P}_{K,\alpha,\gamma}$ denote those problem instances with optimal solution Ω^* and greedy solution S^K. We group all problem instances $\mathcal{P}_{K,\alpha,\gamma}$ according to the set $\Omega^* \cap S^K := \{l_1 = j_{m_1}, l_2 = j_{m_2}, \ldots, l_s = j_{m_s}\}$, where $j_{m_1}, \ldots, j_{m_s}$ are consistent with the order of greedy selection. Let us denote the problem instances with $\Omega^* \cap S^K = \{l_1, \ldots, l_s\}$ as the group $\mathcal{P}_{K,\alpha,\gamma}(\{l_1, \ldots, l_s\})$.

The *main* idea of the proof is to investigate the worst-case approximation ratio of each group of the problem instances $\mathcal{P}_{K,\alpha,\gamma}(\{l_1, \ldots, l_s\}), \forall \{l_1, \ldots, l_s\} \subseteq S^K$. We do this by constructing LPs based on the properties of the problem instances. By studying the structures of these LPs, we will prove that the worst-case approximation ratio of all problem instances occurs when $\Omega^* \cap S^K = \emptyset$. Thus the desired approximation guarantee corresponds to the worst-case approximation ratio of $\mathcal{P}_{K,\alpha,\gamma}(\emptyset)$.

The proof. When $\gamma = 0$ or $F(\Omega^*) = 0$, (1) holds naturally. In the following, let $\gamma \in (0, 1]$ and $F(\Omega^*) > 0$. First, we present Lemma 1, which will be used to construct the LPs.

Lemma 1. *For any $\Omega \subseteq \mathcal{V}$ with $|\Omega| = K$ and any $t \in \{0, \dots, K-1\}$, let $w^t := |S^t \cap \Omega|$. It holds that*

$$\alpha \sum_{i:j_i \in S^t \setminus \Omega} \rho_i + \sum_{i:j_i \in S^t \cap \Omega} \rho_i + \gamma^{-1}(K - w^t)\rho_{t+1} \geq F(\Omega).$$

We now specify the constructing of the LPs: For any problem instance $P_{\Omega^*, S^K} \in \mathcal{P}_{K,\alpha,\gamma}(\{l_1, \dots, l_s\})$, we know that $F(S^K) = \sum_{i=1}^K \rho_i$ (telescoping sum). Hence, the approximation ratio is $\frac{F(S^K)}{F(\Omega^*)} = \sum_i \frac{\rho_i}{F(\Omega^*)}$, which we denote as $R(\{l_1, \dots, l_s\}) = \sum_i \frac{\rho_i}{F(\Omega^*)}$. Define $x_i := \frac{\rho_i}{F(\Omega^*)}, i \in [K]$. Since F is nondecreasing, $x_i \geq 0$. Plugging $\Omega = \Omega^*$ into Lemma 1, and considering $t = 0, \dots, K-1$, we have in total K constraints over the variables x_i, which constitute the constraints of the LP. So the worst-case approximation ratio of the group $\mathcal{P}_{K,\alpha,\gamma}(\{l_1, \dots, l_s\})$ is:

$$\underline{R}(\{l_1, \dots, l_s\}) = \min \sum_{i=1}^K x_i, \text{ s.t. } x_i \geq 0 \text{ and,}$$

$$\begin{array}{l} \text{row } (0) \\ \text{row } (1) \\ \vdots \\ \text{row } (l_1 - 1) \\ \text{row } (l_2 - 1) \\ \text{row } (q = l_r) \\ \vdots \\ \text{row } (l_s - 1) \\ \vdots \\ \text{row } (K-1) \end{array} \begin{bmatrix} \frac{K}{\gamma} & & & & & & & \\ \alpha & \frac{K}{\gamma} & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ \alpha & \alpha & \cdots & \frac{K}{\gamma} & & & & \mathbf{0} \\ \alpha & \alpha & \cdots & 1 & \frac{K-1}{\gamma} & & & \\ \alpha & \alpha & \cdots & 1 & 1 & \frac{K-r}{\gamma} & & \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \\ \alpha & \alpha & \cdots & 1 & 1 & \alpha & \cdots & \frac{K-s+1}{\gamma} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \alpha & \alpha & \cdots & 1 & 1 & \alpha & \cdots & 1 \end{bmatrix} \tag{2}$$

The following Lemma presents the key structure of the constructed LPs, which will be used to deduce the relation between the LPs of different problem instance groups.

Lemma 2. *Assume that the optimal solution of the constructed LP is $\boldsymbol{x}^* \in \mathbb{R}_+^K$ and that $s = |\Omega^* \cap S^K| \geq 1$. For all $1 \leq r \leq s$ it holds that $x_q^* \leq x_{q+1}^*$, where $q = l_r$.*

Proof sketch of Lemma 2. Assume by virture of creating a contradiction that $x_q^* > x_{q+1}^*$. We can always create a new feasible solution $\boldsymbol{y}^* \in \mathbb{R}_+^K$ by decreasing x_q^* by some $\epsilon > 0$, while increasing all the x_{q+1}^* to x_K^* by some proper values, s.t. $\boldsymbol{y}^*$ has smaller LP objective value. Specifically, we define $\boldsymbol{y}^*$ as: for $k = 1, \dots, q-1, y_k^* := x_k^*$; $y_q^* := x_q^* - \epsilon$; for $k = q+1, \dots, K, y_k^* := x_k^* + \epsilon_k$ where ϵ_ks are defined recursively as: $\epsilon_{q+1} = \epsilon \frac{\gamma}{K-r}$, and

$$\epsilon_{q+1+u} = \epsilon_{q+u} \frac{K - r - u + 1 - \gamma}{K - r - u}, 1 \leq u \leq K - q - 1.$$

Claim 1. *a) The new solution $\boldsymbol{y}^* \geq 0$; b) All of the constraints in (2) are still feasible for $\boldsymbol{y}^*$.*

After that the change of the LP objective is,

$$\Delta_{LP} = -\epsilon + \epsilon_{q+1} + \epsilon_{q+2} + \dots + \epsilon_K.$$

One can prove that the LP objective decreases:

Claim 2. *For all $K \geq 1$, $1 \leq r \leq q < K$, it holds that $\Delta_{LP} \leq 0, \forall \gamma \in (0, 1]$. Equality is achieved when $r = q$ and $\gamma = 1$.*

Therefore we reach the contradiction that $\boldsymbol{x}^*$ is an optimal solution of the constructed LP. □

Given Lemma 2, we prove in the following Lemma, which states that the worst-case approximation ratio of all problem instances occurs when $\Omega^* \cap S^K = \emptyset$.

Lemma 3. *For all $\{l_1, \dots, l_s\} \subseteq S^K$, it holds that $\underline{R}(\{l_1, \dots, l_s\}) \geq \underline{R}(\emptyset) = \frac{1}{\alpha}\left[1 - \left(\frac{K - \alpha\gamma}{K}\right)^K\right]$.*

So the greedy solution has objective $F(S^K) \geq \frac{1}{\alpha}\left[1 - \left(\frac{K-\alpha\gamma}{K}\right)^K\right] F(\Omega^*) \geq \frac{1}{\alpha}(1 - e^{-\alpha\gamma}) F(\Omega^*)$.

$$\begin{bmatrix} \frac{K}{\gamma} & & & & & & \\ \alpha & \frac{K}{\gamma} & & & & & \\ \vdots & \vdots & \ddots & & & & \\ \alpha & \alpha & \cdots & \frac{K}{\gamma} & & & \\ \alpha & \alpha & \cdots & 1 & \frac{K-1}{\gamma} & & \\ \alpha & \alpha & \cdots & 1 & 1 & \frac{K-r}{\gamma} & \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \\ \alpha & \alpha & \cdots & 1 & 1 & \alpha & \cdots & \frac{K-s}{\gamma} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{l_1} \\ x_{l_2} \\ x_{q+1} \\ \vdots \\ x_{l_s} \\ \vdots \\ x_K \end{bmatrix} \geq \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

3.3. Tightness Result

We demonstrate that the approximation guarantee in Theorem 1 is tight, i.e., for every submodularity ratio γ and every curvature α, there exist set functions that achieve the bound exactly.

Assume the ground set $\mathcal{V}$ contains the elements in $S := \{j_1, \dots, j_K\}$ and the elements in $\Omega := \{\omega_1, \dots, \omega_K\}$ ($S \cap \Omega = \emptyset$) and $n - 2K$ dummy elements. The objective function we are going to construct will not depend on these dummy elements, i.e., the objective value of a set does not change if dummy elements are removed from or added to that set. Consequently, the dummy elements will not affect the submodularity ratio and the curvature. For the constants $\alpha \in [0, 1], \gamma \in (0, 1]$, we define the objective function as,

$$F(T) := \frac{f(|\Omega \cap T|)}{K}\left(1 - \alpha\gamma \sum_{i:j_i \in S \cap T} \xi_i\right) + \sum_{i:j_i \in S \cap T} \xi_i, \tag{3}$$

where $\xi_i := \frac{1}{K}\left(\frac{K - \gamma\alpha}{K}\right)^{i-1}, i \in [K]$; $f(x) = \frac{\gamma^{-1} - 1}{K - 1}x^2 + \frac{K - \gamma^{-1}}{K-1}x$. Note that $f(x)$ is convex nondecreasing over $[0, K]$, and that $f(0) = 0, f(1) = 1, f(K) = K/\gamma$. It is clear that $F(\emptyset) = 0$ and $F(\cdot)$ is monotone nondecreasing. The following lemma shows that it is generally non-submodular and non-supermodular.

Lemma 4. *For the objective in (3): a) When $\alpha = 0$, it is supermodular; b) When $\gamma = 1$, it is submodular; c) $F(T)$ has submodularity ratio γ and curvature α.*

Considering the problem of $\max_{|T|\leq K} F(T)$, we claim that the GREEDY algorithm may output S. This can be proved by induction. One can see that $\rho_{j_1}(\emptyset) = \xi_1 = \rho_{\omega_1}(\emptyset)$, so GREEDY can choose j_1 in the first step. Assume in step $t-1$ GREEDY has chosen $S^{t-1} = \{j_1, \ldots, j_{t-1}\}$, one can verify that the marginal gains coincide, i.e., $\rho_{j_t}(S^{t-1}) = \xi_t = \rho_{\omega_t}(S^{t-1})$. However, the optimal solution is actually Ω with function value as $F(\Omega) = \frac{1}{\gamma}$. So the approximation ratio is $\frac{F(S)}{F(\Omega)} = \frac{1}{\alpha}\left[1 - \left(\frac{K-\alpha\gamma}{K}\right)^K\right]$, which matches our approximation guarantee in Theorem 1.

4. Applications

We consider several important real-world applications and their corresponding objective functions. We show that the submodularity ratio and the curvature of these functions can be bounded and, hence, the approximation guarantees from our theoretical results are applicable. All the omitted proofs are provided in Appendix D.

4.1. Bayesian A-optimality in Experimental Design

In Bayesian experimental design (Chaloner & Verdinelli, 1995), the goal is to select a set of experiments to perform s.t. some statistical criterion is optimized, e.g., the variance of certain parameter estimates is minimized. Krause et al. (2008) investigated several criteria for this purpose, amongst others the Bayesian A-optimality criterion. This criterion is used to maximally reduce the variance in the posterior distribution over the parameters. In general, the criterion is *not* submodular as shown in Krause et al. (2008, Section 8.4).

Formally, assume there are n experimental stimuli $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, each $\boldsymbol{x}_i \in \mathbb{R}^d$, which constitute the data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$. Let us arrange a set $S \subseteq \mathcal{V}$ of stimuli as a matrix $\mathbf{X}_S := [\boldsymbol{x}_{v_1}, \ldots, \boldsymbol{x}_{v_s}] \in \mathbb{R}^{d\times|S|}$. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be the parameter vector in the linear model $\boldsymbol{y}_S = \mathbf{X}_S^\top \boldsymbol{\theta} + \boldsymbol{w}$, where $\boldsymbol{w}$ is the Gaussian noise with zero mean and variance σ^2, i.e., $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, and $\boldsymbol{y}_S$ is the vector of dependent variables. Suppose the prior takes the form of an isotropic Gaussian, i.e., $\boldsymbol{\theta} \sim \mathcal{N}(0, \boldsymbol{\Lambda}^{-1}), \boldsymbol{\Lambda} = \beta^2\mathbf{I}$. Then,

$$\begin{bmatrix}\boldsymbol{y}_S \\ \boldsymbol{\theta}\end{bmatrix} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = \begin{bmatrix}\sigma^2\mathbf{I} + \mathbf{X}_S^\top\boldsymbol{\Lambda}^{-1}\mathbf{X}_S & \mathbf{X}_S^\top\boldsymbol{\Lambda}^{-1} \\ \boldsymbol{\Lambda}^{-1}\mathbf{X}_S & \boldsymbol{\Lambda}^{-1}\end{bmatrix}.$$

This implies that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}_S} = (\boldsymbol{\Lambda} + \sigma^{-2}\mathbf{X}_S\mathbf{X}_S^\top)^{-1}$. The A-optimality objective is defined as,

$$\begin{aligned} F_A(S) &:= \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}) - \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}_S}) \\ &= \mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - \mathrm{tr}((\boldsymbol{\Lambda} + \sigma^{-2}\mathbf{X}_S\mathbf{X}_S^\top)^{-1}). \end{aligned} \tag{4}$$

The following Proposition gives bounds on the submodularity ratio and curvature of (4).

Proposition 1. *Assume normalized stimuli, i.e., $\|\boldsymbol{x}_i\| = 1, \forall i \in \mathcal{V}$. Let the spectral norm of $\mathbf{X}$ be $\|\mathbf{X}\|$.*[4] *Then, a) The objective in (4) is monotone nondecreasing. b) Its submodularity ratio γ can be lower bounded by $\frac{\beta^2}{\|\mathbf{X}\|^2(\beta^2+\sigma^{-2}\|\mathbf{X}\|^2)}$, and its curvature α can be upper bounded by $1 - \frac{\beta^2}{\|\mathbf{X}\|^2(\beta^2+\sigma^{-2}\|\mathbf{X}\|^2)}$.*

4.2. The Determinantal Function

The determinantal function of a square submatrix is widely used in many areas, e.g., in determinantal point processes (Kulesza & Taskar, 2012) and active set selection for sparse Gaussian processes. Monotone nondecreasing determinantal functions appear in the second problem. Assume $\boldsymbol{\Sigma}$ is the covariance matrix parameterized by a positive definite kernel. In the Informative Vector Machine (Lawrence et al., 2003), the information gain of a subset of points $S \subseteq \mathcal{V}$ is $\frac{1}{2}\log F(S)$, where

$$F(S) := \det(\mathbf{I} + \sigma^{-2}\boldsymbol{\Sigma}_S), \tag{5}$$

where σ is the noise variance in the Gaussian process model, $\boldsymbol{\Sigma}_S$ is the square submatrix with both its rows and columns indexed by S. Although $\log F(S)$ is submodular, $F(S)$ is in general not submodular. The approximation guarantee of GREEDY for maximizing $\log F(S)$ does not translate to a guarantee for maximizing $F(S)$. The following Proposition characterizes (5).

Proposition 2. *a) $F(S)$ in (5) is supermodular, its curvature is 0; b) Let the eigenvalues of $\mathbf{A} := \mathbf{I} + \sigma^{-2}\boldsymbol{\Sigma}$ be $\lambda_1 \geq \cdots \geq \lambda_n > 1$. The greedy submodularity ratio of $F(S)$ can be lower bounded by $\frac{K(\lambda_n - 1)}{(\prod_{j=1}^K \lambda_j) - 1}$.*

4.3. LPs with Combinatorial Constraints

LPs with combinatorial constraints appear frequently in practice. Consider the following example: Suppose that $\mathcal{V}$ is the set of all products a company can produce. Given budget constraints on the raw materials needed, companies consider the LP $\max_{\boldsymbol{x}\in\mathcal{P}}\langle \boldsymbol{d}, \boldsymbol{x}\rangle$, where $\boldsymbol{d}$ is the vector of profits for the individual products and where $\mathcal{P}$ is a polytope representing the continuous constraints. The above LP can be used to assess the profit maximizing production plan. Usually the company needs to consider *combinatorial* constraints as well. For instance, the company has at most K production lines, thus they have to select a subset of K products to produce. Often this kind of problems can be formalized as $\max_{\boldsymbol{x}\in\mathcal{P}, \mathrm{supp}(\boldsymbol{x})\in\mathcal{I}}\langle \boldsymbol{d}, \boldsymbol{x}\rangle$, where $\mathcal{I}$ is the independent set of the combinatorial structure. Hence, a natural auxiliary set function is,

$$F(S) := \max_{\mathrm{supp}(\boldsymbol{x})\subseteq S,\ \boldsymbol{x}\in\mathcal{P}}\langle \boldsymbol{d}, \boldsymbol{x}\rangle,\ \forall S \subseteq \mathcal{V}. \tag{6}$$

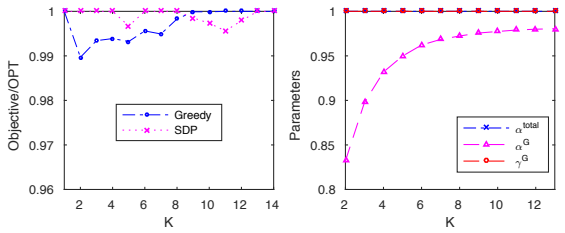
[4] By Weyl's inequality, a naive upper bound is $\|\mathbf{X}\| \leq \sqrt{n}$.

Let $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \leq \boldsymbol{x} \leq \bar{\boldsymbol{u}}, \mathbf{A}\boldsymbol{x} \leq \boldsymbol{b}, \bar{\boldsymbol{u}} \in \mathbb{R}^n_+, \mathbf{A} \in \mathbb{R}^{m\times n}_+, \boldsymbol{b} \in \mathbb{R}^m_+\}$. In general $F(S)$ in (6) is non-submodular as illustrated by two examples in Appendix D.3. Upper bounding the curvature is equivalent to lower bounding $\frac{F(S\cup\Omega)-F(S\setminus\{i\}\cup\Omega)}{F(S)-F(S\setminus\{i\})}$, which can be 0 in the worst case. However, the submodularity ratio can be lower bounded by a non-zero scalar.

Proposition 3. *a) $F(S)$ in (6) is a normalized nondecreasing set function. b) With regular non-degeneracy assumptions (details in Appendix D.3.2), its submodularity ratio can be lower bounded by $\gamma_0 > 0$.*

4.4. More Applications

Many real-world applications can benefit from the theory in this work, for instance: subset selection using the R^2 objective, sparse modeling and the budget allocation problem with combinatorial constraints. Details on these applications are deferred to Appendix G.



(a) Objective values/OPT (b) Parameters

Figure 2: Results on the Boston Housing data.

5. Experimental Results

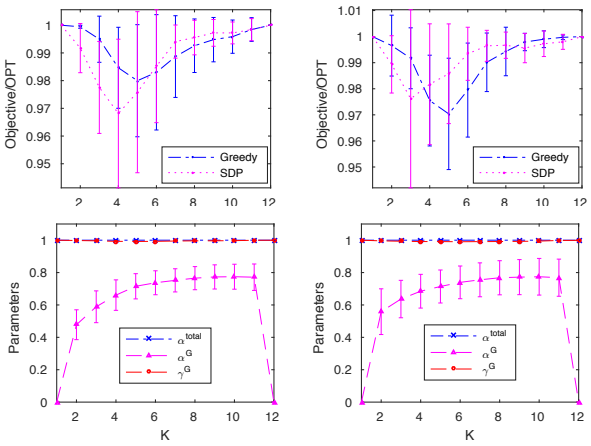
We empirically validated approximation guarantees characterized by the submodularity ratio and the curvature for several applications. Since it is too time consuming to calculate the full versions of α and γ using exhaustive search, we only calculated the *greedy* versions (α^G, γ^G). All averaged results are from 20 repeated experiments. Source code is available at https://github.com/bianan/non-submodular-max.[5] More results are put in Appendix H.

5.1. Bayesian Experimental Design

We considered the Bayesian A-optimality objective for both synthetic and real-world data. In all experiments, we normalized the data points to have unit ℓ_2-norm.

Real-world results: We used the Boston Housing Data.

[5] All experiments were implemented using Matlab. We used the SDP solver provided by CVX (Version 2.1).



(a) Correlation: 0.2 (b) Correlation: 0.6

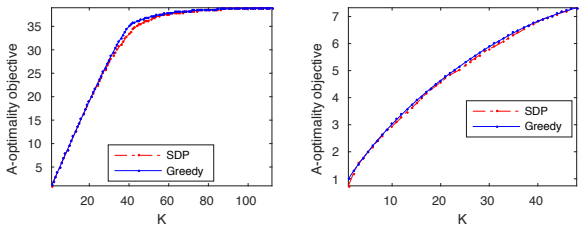
Figure 3: Results for A-optimality on synthetic data.

The dataset[6] has 14 features (e.g., crime rate, property tax rates, etc.) and 516 samples. To be able to quickly calculate the parameters and optimal solution by exhaustive search, the first $n = 14$ samples were used. As a baseline, we used an SDP-based algorithm (abbreviated as SDP, details are available in Appendix E). Results are shown in Fig. 2 for varying values of K. In Fig. 2a we can observe that both GREEDY and SDP compute near-optimal solutions. From Fig. 2b we can see that the greedy submodularity ratio γ^G is close to 1, and that the greedy curvature α^G is less than 1, while the classical curvature α^{total} is always 1 (the worst-case value). This implies that the classical total curvature α^{total} characterizes the considered maximization problems less accurate than the greedy curvature.

Synthetic results: We generated random observations from a multivariate Gaussian distribution with different correlations. To be able to assess the ground truth, we used $n = 12$ samples with $d = 6$ features. Fig. 3 shows the results with correlation 0.2 (first column) and 0.6 (second column), respectively: The first row shows the average objective values over the optimal value with error bars, and the second row shows the parameters. One can observe that GREEDY always obtains near-optimal solutions and that these solutions are roughly comparable with those obtained by the SDP. The classical curvature α^{total} is always close to 1, while α^G take smaller values, and γ^G takes values close to 1, thus characterize the performance of GREEDY better.

Medium-scale synthetic experiments: To compare the runtime of SDP and GREEDY, we considered *medium-scale* datasets (we cannot report results on larger datasets because of the huge computational demands of the SDP).

[6] https://archive.ics.uci.edu/ml/datasets/Housing



(a) $n = 112, d = 40$, corr. $= .5$ (b) $n = 48, d = 24$, corr. : .99

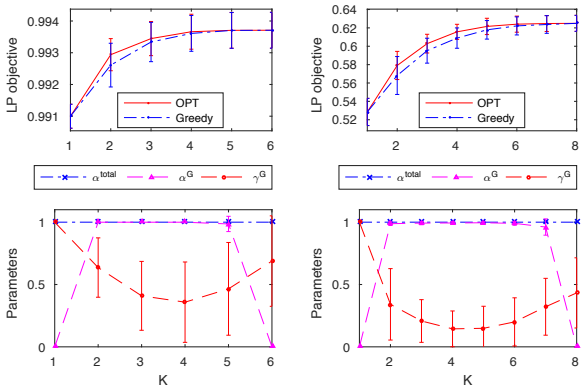
Figure 4: A-optimality on medium-scale problems

Fig. 4 shows the objective value achieved by GREEDY and SDP for different numbers of features d and numbers of samples n, as well as the correlations. We can observe that GREEDY computes solutions that are on par or superior to those of SDP. In Table 1 we summarize the runtime of GREEDY and SDP for different values of d and n, for correlation 0.5. Furthermore, we show the ratio of runtimes of the two algorithms. We can observe that GREEDY is usually two *orders* of magnitude faster than SDP.

Table 1: Runtime in seconds of GREEDY and SDP. The last row shows the ratio of runtimes of SDP and GREEDY.

	d: 60 n: 80	d: 40 n: 112	d: 64 n: 128	d: 100 n: 200	d: 120 n: 250
GREEDY	0.278	0.360	0.765	4.666	10.56
SDP	95.2	115.2	205.4	1741.2	3883.5
$\frac{\text{SDP}}{\text{GREEDY}}$	341.7	319.9	268.7	373.2	367.7

5.2. LPs with Combinatorial Constraints



(a) $n = 6, m = 20$ (b) $n = 8, m = 30$

Figure 5: Results for LPs with K-cardinality constraints.

We generated synthetic LPs as follows: Firstly, we generated the matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}, A_{ij} \in [0, 1]$ by drawing all entries independently from a uniform distribution on $[0, 1]$. We set $\boldsymbol{b} = \boldsymbol{d} = \mathbf{1}$, and set $\bar{\boldsymbol{u}}$ as $\mathbf{1}$. The first row of Fig. 5 plots the optimal LP objective (calculated using exhaustive search) and the LP objective returned by GREEDY. The second row shows the curvature and submodularity ratio. The first column (Fig. 5a) presents the results for $n = 6, m = 20$, while the second column (Fig. 5b) presents that for $n = 8, m = 30$. Note the greedy submodularity ratio takes values between ~ 0.15 and 1, and that the curvature is close to the worst-case value of 1. These observations are consistent with the theory in Section 4.3.

5.3. Determinantal Functions Maximization

We experimented with synthetic and real-world data: For synthetic data, we generated random covariance matrices $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ with uniformly distributed eigenvalues in $[0, 1]$. We set $n = 10, \sigma = 2$. In Fig. 6 (left) we plot the optimal determinantal objective value and the value achieved by GREEDY. Fig. 6 (right) traces the greedy submodularity ratio γ^G. Since the determinantal objective is supermodular, so the approximation guarantee equals to γ^G. We can see that γ^G can reasonably predict the performance of GREEDY.

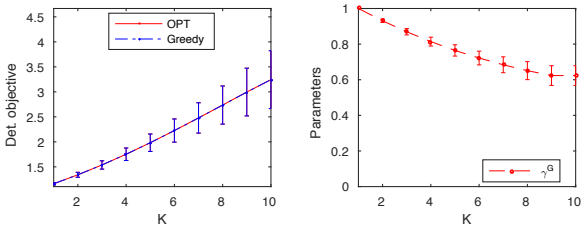


Figure 6: Synthetic result. Left: objective value, right: γ^G

For real-world data, we considered an active set selection task on the CIFAR-10[7] dataset. The first $n = 12$ images in the test set were used to calculate the covariance matrix with an squared exponential kernel ($k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/h^2)$, h was set to be 1). The results in Fig. 7 shows similar results as with the synthetic data.

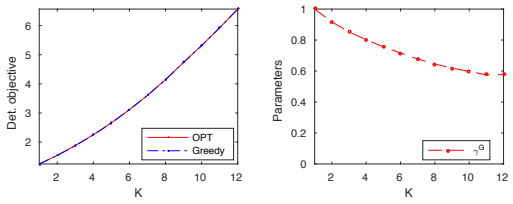


Figure 7: CIFAR-10 result. Left: objective value, right: γ^G

[7] https://www.cs.toronto.edu/~kriz/cifar.html

6. Related Work

In this section we briefly discuss related work on various notions of non-submodularity and the optimization of non-submodular functions (Further details in Appendix F).

Relation to Conforti & Cornuéjols (1984) in deriving approximation guarantees. In proving Theorem 1 we use the similar proof framework (i.e., utilizing LP formulations to analyze the worst-case approximation ratios of different groups of problem instances) as that in Conforti & Cornuéjols (1984), where they derive guarantees for maximizing submodular functions. However, since we are proving guarantees for non-submodular functions, the specific techniques on how to manipulate these LPs are different. Specifically, 1) The building block to construct LPs (Lemma 1) is different; 2) The technique to prove the structure of the LPs (which corresponds to Lemma 2) is significantly different for a submodular function and a non-submodular function, and Lemma 2 is the key to investigate the worst-case approximation ratios of different groups of problem instances. 3) The specific way to prove Lemma 3 is also different since the constraints of the LPs are different for submodular and non-submodular functions.

Submodularity ratio and curvature. Curvature is typically defined for submodular functions. Sviridenko et al. (2013) present a notion of curvature for monotone non-submodular functions. Appendix C provides details of that notion and relates it to our definition. Yoshida (2016) prove an improved approximation ratio for knapsack-constrained maximization of submodular functions with bounded curvature. Submodularity ratio (Das & Kempe, 2011) is a quantity characterizing how close a function is to being submodular.

Approximate submodularity. Krause et al. (2008) define *approximately submodular* functions with parameter $\epsilon \geq 0$ as those functions F that satisfy an approximate diminishing returns property, i.e., $\forall A \subseteq B \subseteq \mathcal{V} \setminus v$ it holds that $\rho_v(A) \geq \rho_v(B) - \epsilon$. GREEDY yields a solution with objective $F(S^K) \geq (1 - e^{-1})F(\Omega^*) - K\epsilon$, for maximizing a monotone F s.t. a K-cardinality constraint. Du et al. (2008) study the greedy maximization of non-submodular potential functions with *restricted submodularity* and *shifted submodularity*. Restricted submodularity refers to functions which are submodular only over some collection of subsets of $\mathcal{V}$, and shifted submodularity can be viewed as a special case of the approximate diminishing returns as defined above. Recently, Horel & Singer (2016) study *ϵ-approximately submodular* functions, which arised from their research on "noisy" submodular functions. A function $F(\cdot)$ is ϵ-approximately submodular if there exists a submodular function G s.t. $(1 - \epsilon)G(S) \leq F(S) \leq (1 + \epsilon)G(S), \forall S \subseteq \mathcal{V}$.

Weak submodularity. Borodin et al. (2014) study *weakly submodular* functions, i.e., montone, nomalized functions $F(\cdot)$ s.t. for any S, T, it holds $|T|F(S) + |S|F(T) \geq |S \cap T|F(S \cup T) + |S \cup T|F(S \cap T)$. For a function $F(\cdot)$, we show in Remark 4 that the following two facts do not imply each other: i) $F(\cdot)$ is weakly submodular; ii) The submodularity ratio of $F(\cdot)$ is strictly larger than 0, and its curvature is strictly smaller than 1.

Other notions of non-submodularity. Feige & Izsak (2013) introduce the *supermodular degree* as a complexity measure for set functions. They show that a greedy algorithm for the welfare maximization problem enjoys an approximation guarantee increasing linearly with the supermodular degree. Zhou & Spanos (2016) use the *submodularity index* to characterize the performance of the RANDOMGREEDY algorithm (Buchbinder et al., 2014) for maximizing a non-monotone function.

Optimization of non-submodular functions. The submodular-supermodular procedure has been proposed to minimize the difference of two submodular functions (Narasimhan & Bilmes, 2005; Iyer & Bilmes, 2012). Jegelka & Bilmes (2011) present the problem of minimizing "cooperative cuts", which are non-submodular in general, and propose efficient algorithms for optimization. Kawahara et al. (2015) analyze unconstrained minimization of the sum of a submodular function and a tree-structured supermodular function. Bai et al. (2016) investigate the minimization of the ratio of two submodular functions, which can be solved with bounded approximation factor.

7. Conclusion

We analyzed the guarantees for greedy maximization of non-submodular nondecreasing set functions. By combining the (generalized) curvature α and submodularity ratio γ for generic set functions, we prove the *first* tight approximation bounds in terms of these definitions for greedily maximizing nondecreasing set functions. These approximation bounds significantly enlarge the domain where GREEDY has guarantees. Furthermore, we theoretically bounded the parameters α and γ for several non-trivial applications, and validate our theory in various experiments.

ACKNOWLEDGEMENTS

The authors would like to thank Adish Singla, Kfir Y. Levy and Aurelien Lucchi for valuable discussions. This research was partially supported by ERC StG 307036 and the Max Planck ETH Center for Learning Systems. This work was done in part while Andreas Krause was visiting the Simons Institute for the Theory of Computing.

References

Altschuler, Jason, Bhaskara, Aditya, Fu, Gang, Mirrokni, Vahab, Rostamizadeh, Afshin, and Zadimoghaddam, Morteza. Greedy column subset selection: New bounds and distributed algorithms. In *ICML*, pp. 2539–2548, 2016.

Bach, Francis. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

Bai, Wenruo, Iyer, Rishabh, Wei, Kai, and Bilmes, Jeff. Algorithms for optimizing the ratio of submodular functions. In *ICML*, pp. 2751–2759, 2016.

Bertsimas, Dimitris and Tsitsiklis, John. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

Bian, Andrew An, Mirzasoleiman, Baharan, Buhmann, Joachim M., and Krause, Andreas. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *AISTATS*, pp. 111–120, 2017.

Borodin, Allan, Le, Dai Tri Man, and Ye, Yuli. Weakly submodular functions. *arXiv preprint arXiv:1401.6697*, 2014.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Buchbinder, Niv, Feldman, Moran, Naor, Joseph, and Schwartz, Roy. Submodular maximization with cardinality constraints. In *SODA*, pp. 1433–1452, 2014.

Candes, Emmanuel J, Romberg, Justin K, and Tao, Terence. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

Chaloner, Kathryn and Verdinelli, Isabella. Bayesian experimental design: A review. *Statistical Science*, 10(3): 273–304, 1995.

Conforti, Michele and Cornuéjols, Gérard. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete Applied Mathematics*, 7(3): 251–274, 1984.

Das, Abhimanyu and Kempe, David. Algorithms for subset selection in linear regression. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pp. 45–54, 2008.

Das, Abhimanyu and Kempe, David. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *ICML*, pp. 1057–1064, 2011.

Du, Ding-Zhu, Graham, Ronald L, Pardalos, Panos M, Wan, Peng-Jun, Wu, Weili, and Zhao, Wenbo. Analysis of greedy approximations with nonsubmodular potential functions. In *SODA*, pp. 167–175, 2008.

Elenberg, Ethan R, Khanna, Rajiv, Dimakis, Alexandros G, and Negahban, Sahand. Restricted strong convexity implies weak submodularity. *arXiv preprint arXiv:1612.00804*, 2016.

Feige, Uriel and Izsak, Rani. Welfare maximization and the supermodular degree. In *Proceedings of the Fourth Conference on Innovations in Theoretical Computer Science*, pp. 247–256, 2013.

Guyon, Isabelle and Elisseeff, André. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Horel, Thibaut and Singer, Yaron. Maximization of approximately submodular functions. In *NIPS*, pp. 3045–3053. 2016.

Iyer, Rishabh and Bilmes, Jeff. Algorithms for approximate minimization of the difference between submodular functions, with applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 407–417, 2012.

Iyer, Rishabh K, Jegelka, Stefanie, and Bilmes, Jeff A. Curvature and optimal algorithms for learning and minimizing submodular functions. *NIPS*, pp. 2742–2750, 2013.

Jain, Prateek, Tewari, Ambuj, and Kar, Purushottam. On iterative hard thresholding methods for high-dimensional m-estimation. In *NIPS*, pp. 685–693, 2014.

Jegelka, Stefanie and Bilmes, Jeff. Submodularity beyond submodular energies: coupling edges in graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1897–1904, 2011.

Kawahara, Yoshinobu, Iyer, Rishabh K, and Bilmes, Jeff A. On approximate non-submodular minimization via tree-structured supermodularity. In *AISTATS*, pp. 444–452, 2015.

Krause, Andreas and Cevher, Volkan. Submodular dictionary selection for sparse representation. In *ICML*, pp. 567–574, 2010.

Krause, Andreas and Golovin, Daniel. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, February 2014.

Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

Kulesza, Alex and Taskar, Ben. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Lawrence, Neil, Seeger, Matthias, and Herbrich, Ralf. Fast sparse gaussian process methods: The informative vector machine. *NIPS*, pp. 625–632, 2003.

Narasimhan, Mukund and Bilmes, Jeff. A submodular-supermodular procedure with applications to discriminative structure learning. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 404–412, 2005.

Nemhauser, George L, Wolsey, Laurence A, and Fisher, Marshall L. An analysis of approximations for maximizing submodular set functions–i. *Mathematical Programming*, 14(1):265–294, 1978.

Soma, Tasuku, Kakimura, Naonori, Inaba, Kazuhiro, and Kawarabayashi, Ken-ichi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *ICML*, pp. 351–359, 2014.

Sviridenko, Maxim, Vondrák, Jan, and Ward, Justin. Optimal approximation for submodular and supermodular optimization with bounded curvature. *arXiv preprint arXiv:1311.4728*, 2013.

Vondrák, Jan. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pp. 67–74, 2008.

Vondrák, Jan. Submodularity and curvature: the optimal algorithm. *RIMS Kokyuroku Bessatsu B*, 23:253–266, 2010.

Yoshida, Yuichi. Maximizing a monotone submodular function with a bounded curvature under a knapsack constraint. *arXiv preprint arXiv:1607.04527*, 2016.

Zhou, Yuxun and Spanos, Costas J. Causal meets submodular: Subset selection with directed information. In *NIPS*, pp. 2649–2657. 2016.