Beyond Adaptive Submodularity: Approximation Guarantees of Greedy Policy with Adaptive Submodularity Ratio

Kaito Fujii [1] **Shinsaku Sakaue** [2]

Abstract

We propose a new concept named *adaptive submodularity ratio* to study the greedy policy for sequential decision making. While the greedy policy is known to perform well for a wide variety of adaptive stochastic optimization problems in practice, its theoretical properties have been analyzed only for a limited class of problems. We narrow the gap between theory and practice by using adaptive submodularity ratio, which enables us to prove approximation guarantees of the greedy policy for a substantially wider class of problems. Examples of newly analyzed problems include important applications such as adaptive influence maximization and adaptive feature selection. Our adaptive submodularity ratio also provides bounds of *adaptivity gaps*. Experiments confirm that the greedy policy performs well with the applications being considered compared to standard heuristics.

1. Introduction

Sequential decision making plays a crucial role in machine learning. In various scenarios, we must design an effective policy that repeatedly decides the next action to be taken by using the feedback obtained so far. The greedy policy is a simple but empirically effective approach to sequential decision making. At each step, it myopically makes a decision that seems the most beneficial among feasible choices.

Adaptive submodularity (Golovin & Krause, 2011) is a well-established framework for analyzing greedy algorithms for sequential decision making. It extends *submodularity*, which is a diminishing returns property of set functions, to the setting of adaptive decision making. This framework has successfully provided theoretical guarantees for greedy algorithms for active learning (Golovin et al., 2010), recommendation (Gabillon et al., 2013), and touch-based localization in robotics (Javdani et al., 2014).

However, adaptive submodularity is not omnipotent. While the greedy policy works well for various sequential decision making problems, many of these problems do not have adaptive submodularity. In fact, even if an objective function is submodular in the non-adaptive setting, its adaptive version does not always have adaptive submodularity. *Adaptive influence maximization* is one such example. In this problem, a decision maker aims at spreading information about a product by selecting several advertisements. She repeatedly alternates between selecting an advertisement and observing its effect. The objective function of this problem is known to have adaptive submodularity in the independent cascade model (Golovin & Krause, 2011), but not in a more general diffusion model called the *triggering model* (Kempe et al., 2003), which is extensively studied as an important class of diffusion models (Leskovec et al., 2007; Tang et al., 2014). Note that this objective function satisfies submodularity in the non-adaptive setting, while it does not satisfy adaptive submodularity in the adaptive setting. Examples of other problems lacking adaptive submodularity appear in many applications such as feature selection and active learning. Therefore, we are waiting for an analysis framework that goes beyond adaptive submodularity.

In the non-adaptive setting, *submodularity ratio* (Das & Kempe, 2011) is a prevalent tool for handling non-submodular functions (Khanna et al., 2017; Elenberg et al., 2017). Intuitively, it is a parameter of monotone set functions that measures their distance to submodular functions. An adaptive variant of submodularity ratio would be a promising approach to handling functions that lack adaptive submodularity, but how to define it is quite non-trivial since there is a large discrepancy between the non-adaptive and adaptive settings as exemplified above. In particular, success in defining an adaptive version of submodularity ratio involves meeting the following two requirements: it must yield an approximation guarantee of the greedy policy, and it must be bounded in various important applications such as the adaptive influence maximization and adaptive feature selection. Previous works (Kusner, 2014; Yong et al., 2017) tried to define similar notions, but none of them meet the

[1] University of Tokyo [2] NTT Communication Science Laboratories. Correspondence to: Kaito Fujii <kaito_fujii@mist.i.u-tokyo.ac.jp>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Table 1. Summary of our theoretical results about adaptive bipartite influence maximization and adaptive feature selection. We show lower bounds for the adaptive submodularity ratios, the approximation ratios of the adaptive greedy algorithm, and the adaptivity gaps. Let $\lambda_{\min,\ell} = \min_{\phi} \min_{S \subseteq V:\ |S| \leq \ell} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)$ and $\lambda_{\max,\ell} = \max_{\phi} \max_{S \subseteq V:\ |S| \leq \ell} \lambda_{\max}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)$. Parameters q and d are determined by the diffusion model and the underlying graph structure. The results of (Golovin & Krause, 2011) are indicated by †.

Problem	Adaptive submodularity ratio	Adaptive greedy	Adaptivity gaps
Linear threshold	$(k+1)/2k$	$1 - \exp(-(k+1)/2k)$	$(k+1)/2k$
Independent cascade	$1^\dagger$	$1 - 1/\mathrm{e}^\dagger$	$(1-q)^{\min\{d,k\}-1}$
Triggering	$(k+1)/2k$	$1 - \exp(-(k+1)/2k)$	
Feature selection	$\lambda_{\min,k+\ell}$	$1 - \exp(-\lambda_{\min,k+\ell})$	$\lambda_{\min,k}/\lambda_{\max,k}$

requirements.

Our Contribution. We propose an analysis framework, *adaptive submodularity ratio*, that meets the aforementioned requirements. An advantage of our proposal is that it has the potential to yield various theoretical results as in Table 1. Below we summarize our main contributions.

- We propose the definition of the adaptive submodularity ratio and, by using it, we prove an approximation guarantee of the adaptive greedy algorithm.
- We give a bound on the *adaptivity gap*[1], which represents the superiority of adaptive policies over non-adaptive policies, through the lens of the adaptive submodularity ratio.
- We provide lower-bounds of adaptive submodularity ratio for two important applications: adaptive influence maximization on bipartite graphs in the triggering model and adaptive feature selection. Regarding the former one, we show that our result is tight.
- Experiments confirm that the greedy policy performs well for the considered applications.

Organization. The rest of this paper is organized as follows. Section 2 provides the basic concepts and definitions. In Section 3, we formally define the adaptive submodularity ratio, which is the key concept of this study. In Sections 4 and 5, we provide bounds on the approximation ratio of the adaptive greedy algorithm and adaptivity gaps, respectively, by using the adaptive submodularity ratio. In Sections 6 and 7, we apply the frameworks developed in Sections 4 and 5 to two applications: adaptive influence maximization and adaptive feature selection. In Section 8, we experimentally check the performance of the adaptive greedy algorithm in several applications. In Section 9 we review related work.

2. Preliminaries

Adaptive Stochastic Optimization. Adaptive stochastic optimization is a general framework for handling problems of sequentially selecting elements, where we can observe the states of only the selected elements. Let V be the ground set consisting of a finite number of elements. Suppose every element $v \in V$ is assigned to some state in $\mathcal{Y}$, which is the set of all possible states. We let $\phi\colon V \to \mathcal{Y}$ be a map that associates each element, $v \in V$, with a state, $\phi(v) \in \mathcal{Y}$. We consider the Bayesian setting where ϕ is generated from a known prior distribution $p(\phi)$. Let Φ be a random variable representing the randomness of the realization ϕ.

A decision maker can select one element $v \in V$ at each step. After selecting v, she can observe the state $\phi(v)$ of v. She repeatedly selects an element and then observes its state. The important point is that she can utilize the information about the states observed so far for selecting the next element. We denote by $\psi = \{(v_1, \phi(v_1)), \ldots, (v_\ell, \phi(v_\ell))\}$ the partial realization observed so far, where $\{v_1, \ldots, v_\ell\}$ is the set of selected elements. The decision maker's strategy can be described as a *policy tree*, or simply *policy*. A policy is a decision tree that determines the element to be selected next. Formally, a policy π is a partial map that returns an element $v \in V$ to be selected next given partial realization ψ observed so far.

The goal of the decision maker is to maximize the expected value of the objective function $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$. The objective function value $f(S, \phi)$ depends on the set S of selected elements and the states ϕ of all elements. At the beginning, she does not know ϕ, but she can get partial information of ϕ by observing state $\phi(v)$ of selected v. In parallel, she must select elements to construct S that has high utility under the realization ϕ. Let $E(\pi, \phi) \subseteq V$ be the set selected by policy π under realization ϕ. The expected value achieved by policy π is

$$f_{\mathrm{avg}}(\pi) = \mathbb{E}_{\Phi}[f(E(\pi, \Phi), \Phi)],$$

where the expectation is taken with regard to the random variable Φ generated from p.

[1] The adaptivity gap is a different concept from *adaptive complexity* (Balkanski & Singer, 2018).

Adaptive Submodularity and Adaptive Monotonicity. Adaptive submodularity, which is an adaptive extension of submodularity, is a diminishing returns property of the expected marginal gain. The expected marginal gain of $v \in V$ when ψ has been observed so far is defined as

$$\begin{aligned}&\Delta(v|\psi)\\&:= \mathbb{E}[f(\mathrm{dom}(\psi) \cup \{v\}, \Phi) - f(\mathrm{dom}(\psi), \Phi)|\Phi \sim \psi],\end{aligned}$$

where $\mathrm{dom}(\psi) := \{v \in V \mid \exists y \in \mathcal{Y},\ (v, y) \in \psi\}$. We write $\Phi \sim \psi$ if Φ is generated from the posterior distribution $p(\phi|\psi)$. Given current realization ψ, the expected marginal gain, $\Delta(v|\psi)$, represents the expected increase in the objective value yielded by selecting v. Adaptive submodularity is defined as follows:

Definition 1 (Adaptive submodularity (Golovin & Krause, 2011))**.** Let $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ be a set function and p a distribution of ϕ. We say f is adaptive submodular with respect to p if for any partial realization $\psi \subseteq \psi'$ and any element $v \in V \setminus \mathrm{dom}(\psi')$, it holds that

$$\Delta(v|\psi) \geq \Delta(v|\psi').$$

The monotonicity can also be extended to the adaptive setting as follows:

Definition 2 (Adaptive monotonicity (Golovin & Krause, 2011))**.** Let $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ be a set function and p a distribution of ϕ. We say f is adaptive monotone with respect to p if for any partial realization ψ and any element $v \in V \setminus \mathrm{dom}(\psi)$, it holds that

$$\Delta(v|\psi) \geq 0.$$

Other Notations for Adaptive Stochastic Optimization. The expected marginal gain of policy π with partial realization ψ is defined as

$$\begin{aligned}&\Delta(\pi|\psi)\\&:= \mathbb{E}[f(\mathrm{dom}(\psi) \cup E(\pi, \Phi), \Phi) - f(\mathrm{dom}(\psi), \Phi)|\Phi \sim \psi].\end{aligned}$$

Similarly, the expected marginal gain of set $S \subseteq V$ with partial realization ψ is defined as

$$\Delta(S|\psi) := \mathbb{E}[f(\mathrm{dom}(\psi) \cup S, \Phi) - f(\mathrm{dom}(\psi), \Phi)|\Phi \sim \psi].$$

Let $\Pi_k := \{\pi \mid \forall\phi,\ |E(\pi, \phi)| \leq k\}$ be the set of all policies whose heights do not exceed k.

Submodularity Ratio and Supermodularity Ratio. The submodularity ratio of a monotone non-negative set function $f\colon 2^V \to \mathbb{R}_{\geq 0}$ with respect to set $U \subseteq V$ and parameter $k \geq 1$ is defined to be

$$\gamma_{U,k}(f) := \min_{L \subseteq U,\ S\colon |S| \leq k} \frac{\sum_{v \in S} f(v|L)}{f(S|L)},$$

where $f(v|L) := f(L \cup \{v\}) - f(L)$ and $f(S|L) := f(L \cup S) - f(L)$. If the numerator and denominator are both 0, the submodularity ratio is considered to be 1. We have $\gamma_{U,k} \in [0, 1]$, and a monotone set function f is submodular if and only if $\gamma_{U,k} = 1$ for every $U \subseteq V$ and $k \geq 1$.

As an opposite concept of the submodularity ratio, the *supermodularity ratio*, was considered in Bogunovic et al. (2018), which is defined as follows:

$$\beta_{U,k}(f) := \min_{L \subseteq U,\ S\colon |S| \leq k} \frac{f(S|L)}{\sum_{v \in S} f(v|L)},$$

where we regard $0/0 = 1$. We have $\beta_{U,k} \in [1/k, 1]$, and f is supermodular if and only if $\beta_{U,k} = 1$ for every $U \subseteq V$ and $k \geq 1$. We omit f from $\gamma_{U,k}(f)$ and $\beta_{U,k}(f)$ if it is clear from the context.

3. Adaptive Submodularity Ratio

In this section, we provide a precise definition of the adaptive submodularity ratio, which extends the submodularity ratio from the non-adaptive setting to the adaptive setting. We need to define it carefully so that it can yield an approximation guarantee of the greedy policy. An important point is to generalize subset S of size at most k, used to define the submodularity ratio, to policy π of height at most k.

Definition 3 (Adaptive submodularity ratio)**.** Suppose that $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ is adaptive monotone w.r.t. a distribution p. Adaptive submodularity ratio $\gamma_{\psi,k} \in [0, 1]$ of f and p with respect to partial realization ψ and parameter $k \in \mathbb{Z}_{\geq 0}$ is defined to be

$$\begin{aligned}&\gamma_{\psi,k}(f, p) :=\\&\min_{\psi' \subseteq \psi,\ \pi \in \Pi_k} \frac{\sum_{v \in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi')\Delta(v|\psi')}{\Delta(\pi|\psi')}.\end{aligned}$$

We omit f and p if they are clear from the context. We also define $\gamma_{\ell,k} := \min_{\psi\colon|\psi| \leq \ell} \gamma_{\psi,k}$.

Intuitively, the adaptive submodularity ratio indicates the distance between (f, p) and the class of adaptive submodular functions. As with the non-adaptive setting, $\gamma_{\psi,k}(f, p) = 1$ implies the adaptive submodularity of f, which can formally be written as follows:

Proposition 1. *It holds that $\gamma_{\psi,k}(f, p) = 1$ for any partial realization ψ and $k \in \mathbb{Z}_{\geq 0}$ if and only if f is adaptive submodular with respect to p.*

The proof is given in Appendix A.

4. Adaptive Greedy Algorithm

In this section, we present a new approximation guarantee for the adaptive greedy algorithm based on the adaptive

Algorithm 1 Adaptive greedy algorithm (Golovin & Krause, 2011)

```
Input  The value oracle for the expected marginal gain $\Delta(\cdot|\cdot)$ associated with $f\colon 2^V \times \mathcal{Y}^V$ and $p \in \triangle^{\mathcal{Y}^V}$, a cardinality constraint $\ell \in \mathbb{Z}_{\geq 0}$.
Output  $\psi_\ell$ a set of observations of size $\ell$.
$\psi_0 \leftarrow \emptyset$.
for $i = 1, \dots, \ell$ do
    $v \leftarrow \mathrm{argmax}_{v \in V} \Delta(v|\psi_{i-1})$.
    Observe $\phi(v)$ and let $\psi_i \leftarrow \psi_{i-1} \cup \{(v, \phi(v))\}$.
end for
return $\psi_\ell$.
```

submodularity ratio. Thanks to this result, once the adaptive submodularity ratio is bounded, we can obtain approximation guarantees of the adaptive greedy algorithm for various applications. The adaptive greedy algorithm is an algorithm that starts with an empty set and repeatedly selects the element with the largest expected marginal gain. The detailed description is given in Algorithm 1. Golovin & Krause (2011) have shown that this algorithm achieves $(1 - 1/\mathrm{e})$-approximation to the expected objective value of an optimal policy if f is adaptive submodular w.r.t. p. Here we extend their result and show that the adaptive greedy algorithm achieves $(1 - \exp(-\gamma_{\ell,\ell}))$-approximation, where ℓ is the number of selected elements. More precisely, we can bound the approximation ratio relative to any policy π^* of height k as follows:

Theorem 1. *Suppose $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}_{\geq 0}$ is adaptive monotone with respect to p. Let π be a policy representing the adaptive greedy algorithm until ℓ step. Then, for any policy $\pi^* \in \Pi_k$, it holds that*

$$f_{\mathrm{avg}}(\pi) \geq \left(1 - \exp\left(-\frac{\gamma_{\ell,k}\ell}{k}\right)\right) f_{\mathrm{avg}}(\pi^*),$$

where $\gamma_{\ell,k}$ is the adaptive submodularity ratio of f w.r.t. p.

We provide the proof in Appendix B.

5. Non-adaptive Policies and Adaptivity Gaps

We show that the adaptive submodularity ratio is also useful for theoretically comparing the performances of adaptive and non-adaptive policies. More precisely, we present a lower-bound of the *adaptivity gap*, which represents the performance gap between adaptive and non-adaptive polices, by using the adaptive submodularity ratio. The adaptivity gap is defined as follows:

Definition 4 (Adaptivity gaps)**.** The adaptivity gap $\mathsf{GAP}_k(f, p)$ of an objective function $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}_{\geq 0}$ and a probability distribution p of $\phi\colon V \to \mathcal{Y}$ is defined as the ratio between an optimal adaptive policy and an optimal non-adaptive policy, i.e.,

$$\mathsf{GAP}_k(f, p) = \frac{\max_{M\colon |M| \leq k} \mathbb{E}_\Phi[f(M, \Phi)]}{\max_{\pi^* \in \Pi_k} f_{\mathrm{avg}}(\pi^*)},$$

where k is the height of adaptive and non-adaptive policies.

Theorem 2. *Let $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}_{\geq 0}$ be an objective function and p a probability distribution of $\phi\colon V \to \mathcal{Y}$. Let $\gamma_{\emptyset,k}$ be the adaptive submodularity ratio of f w.r.t. p. Let $\beta_{\emptyset,k}$ be the supermodularity ratio of the set function $\mathbb{E}_\Phi[f(\cdot, \Phi)]$ of non-adaptive policies. We have*

$$\mathsf{GAP}_k(f, p) \geq \beta_{\emptyset,k}\gamma_{\emptyset,k}.$$

Therefore, given any non-adaptive α-approximation algorithm, we can evaluate its performance relative to an optimal adaptive policy as follows:

Corollary 1. *Let $\pi_{\mathrm{non}} \in \Pi_k$ be a non-adaptive policy that achieves α-approximation to an optimal non-adaptive policy π^*_{non}. Let $\gamma_{\emptyset,k}$ be the adaptive submodularity ratio of f w.r.t. p. Let $\beta_{\emptyset,k}$ be the supermodularity ratio of the non-adaptive objective function $\mathbb{E}_\Phi[f(\cdot, \Phi)]$. Let π^* be an optimal adaptive policy. We have*

$$f_{\mathrm{avg}}(\pi_{\mathrm{non}}) \geq \alpha\beta_{\emptyset,k}\gamma_{\emptyset,k} f_{\mathrm{avg}}(\pi^*).$$

Proofs are given in Appendix C.

6. Adaptive Influence Maximization

In this section, we consider adaptive influence maximization on bipartite graphs. We provide a bound on the adaptive submodularity ratio in the case of the triggering model, and we show that this result is tight. We also present bounds on the adaptivity gaps in the case of the independent cascade and linear threshold models by using the adaptive submodularity ratio.

Let $G = (V \cup U, A)$ be a directed bipartite graph with source vertices V, sink vertices U, and directed edges $A \subseteq V \times U$. In the case of bipartite influence model (Alon et al., 2012), this graph represents the relationship between advertisements V and customers U. We consider the problem of selecting several advertisements $S \subseteq V$ to make as much influence as possible on the customers. Here, each edge is determined to be alive or dead according to a certain distribution, and influence can be spread only through live edges. Given vertex weights $w\colon U \to \mathbb{R}_{\geq 0}$, the objective function to be maximized is $f(X) = \sum_{u \in \bigcup_{v \in X} R(v)} w(u)$, where, for each $v \in V$, $R(v) \subseteq U$ represents a set of vertices that are reachable from v by going through only live edges. In the adaptive version of influence maximization, at each step, we select a vertex $v \in V$ and observe the states of all outgoing edges $(v, u) \in A$, while, in the non-adaptive setting, we select $S \subseteq V$ before observing the states of any edges.

We consider a general diffusion model called the *triggering model* (Kempe et al., 2003), which includes various important models such as the independent cascade model and the linear threshold model as special cases. In the triggering model, each vertex $v \in V$ is associated with some known probability distribution over the power set of incoming edges. According to this distribution, a subset of incoming live edges is determined. A vertex gets activated if and only if it is reachable from some selected vertex (or seed vertex) through only live edges. We aim to maximize the total weight of activated vertices by appropriately selecting seed vertices. Note that this objective function is submodular in the non-adaptive setting.

For later use, we explain the linear threshold model, a special case of the triggering model. In this model, the probability distribution on the incoming edges of each vertex is restricted so that each vertex has at most one live edge in any realization. In other words, there exists $b\colon A \to \mathbb{R}_{\geq 0}$ such that, for each $v \in V$, we have $\sum_{a \in \delta_-(v)} b(a) \leq 1$, where $\delta_-(v)$ is the full set of edges pointing to v, and $a \in A$ is alive with probability $b(a)$ exclusively over $\delta_-(v)$. In contrast to the linear threshold model, the triggering model accepts any distribution over the power set of $\delta_-(v)$.

6.1. Bound of Adaptive Submodularity Ratio

We first present the bound of adaptive submodularity ratio. Here we provide a proof sketch, and the full proof is given in Appendices D.1 and D.2.

Theorem 3. *Let G be an arbitrary directed bipartite graph and w be any weight function. For any $k \in \mathbb{Z}_{\geq 0}$ and partial realization ψ, the adaptive submodularity ratio $\gamma_{\psi,k}$ of the objective function and the distribution of the adaptive influence maximization in the triggering model is lower-bounded as follows:*

$$\gamma_{\psi,k} \geq \frac{k+1}{2k}.$$

Proof sketch of Theorem 3. Since the objective function and the probability distribution of edge states can be decomposed into those defined for each vertex $u \in U$, it is sufficient to consider the case where $|U| = 1$.

Our goal is to prove

$$\Delta(\pi|\psi') \leq \frac{2k}{k+1} \sum_{v \in V} \Pr(v \in E(\pi, \Phi) | \Phi \sim \psi') \Delta(v|\psi')$$

for any observation ψ' and policy $\pi \in \Pi_k$. By duplicating $v \in V$ that appears multiple times in policy tree π, we can write the above inequality as

$$\sum_{v \in V} \mathrm{P}_{v,\pi} \left(\frac{2k}{k+1} \Delta(v|\psi') - \Delta(v|\psi' \cup \psi_v) \right) \geq 0,$$

where $\mathrm{P}_{v,\pi}$ is a shorthand for $\Pr(v \in E(\pi, \Phi) | \Phi \sim \psi')$ and ψ_v is the observation just before v is selected. We decompose the policy tree into the path wherein u remains inactive and the rest, and prove the inequality for each part separately. □

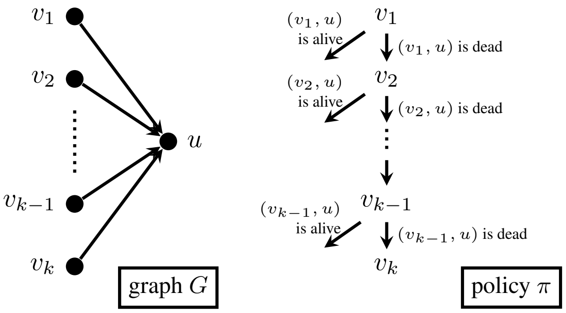


Figure 1. An example that implies the tightness of our bound.

We can see that the above bound is tight even for the linear threshold model by considering the following example.

Example 1. Let G be a bipartite directed graph with $V = \{v_1, \ldots, v_k\}$, $U = \{u\}$, and $A = \{(v_i, u) \mid i \in [k]\}$. Let w be the vertex weight such that $w(u) = 1$. We consider the linear threshold model in which an edge selected out of A uniformly at random is alive and the other edges are dead. We consider a simple policy π that selects all vertices one by one until u is activated. These graph and policy are illustrated in Figure 1. Since π finally activates u, the expected gain of π is $\Delta(\pi|\emptyset) = 1$. The probability that π selects each vertex is $\Pr(v_i \in E(\pi, \Phi)) = (k - i + 1)/k$. The expected marginal gain of v_i is $\Delta(v_i|\emptyset) = 1/k$. The adaptive submodularity ratio can be upper-bounded as

$$\begin{aligned}\gamma_{\emptyset,k} &\leq \frac{\sum_{v \in V} \Pr(v \in E(\pi, \Phi)) \Delta(v|\emptyset)}{\Delta(\pi|\emptyset)} \\ &\leq \sum_{i=1}^{k} \frac{k-i+1}{k} \cdot \frac{1}{k} \\ &\leq \frac{k+1}{2k}.\end{aligned}$$

Hence the lower-bound in Theorem 3 is tight.

The assumption that G is bipartite, considered in Theorem 3, may seem excessively strong, but it is actually a vital assumption. We show that, if G is not a bipartite graph, the adaptive submodularity ratio can be arbitrarily small; in fact, such an example can be constructed with the linear threshold model on a very simple graph G. We describe the details in Appendix D.3.

6.2. Bound of Adaptivity Gap

Next we provide a bound on the adaptivity gaps of bipartite influence maximization problems by using the adaptive submodularity ratio. First we consider the independent cascade model. Since the adaptive submodularity holds for the independent cascade model (Golovin & Krause, 2011), the adaptive submodularity ratio of its objective function is 1 by Proposition 1. In addition, by using a bound of the curvature (Maehara et al., 2017) and an inequality between the supermodularity ratio and the curvature (Bogunovic et al., 2018), we obtain $\beta_{\emptyset,k} \geq (1-q)^{\min\{k,d\}-1}$, where q is an upper bound of the probability that each edge is alive and d is the largest degree of the vertex in V. From Theorem 2, we obtain the following result.

Proposition 2. *Let f be the objective function and p the probability distribution of bipartite influence maximization in the independent cascade model. We have*

$$\mathsf{GAP}_k(f,p) \geq (1-q)^{\min\{k,d\}-1}.$$

We can derive a similar bound for the linear threshold model. Since the expected objective function is a linear function, its supermodularity ratio is 1. As a special case of Theorem 3, we have $\gamma_{\emptyset,k} \geq \frac{k+1}{2k}$. Combining these bounds with Theorem 2, we obtain the following result.

Proposition 3. *Let f be the objective function and p the probability distribution of bipartite influence maximization in the linear threshold model. We have*

$$\mathsf{GAP}_k(f,p) \geq \frac{k+1}{2k}.$$

7. Adaptive Feature Selection

In this section, we consider an adaptive variant of feature selection for sparse regression. All proofs related to this section are presented in Appendix E.

Let us consider the following scenario. A learner has all feature vectors in advance, but they are not accurate due to sensing noise. Here each sensor corresponds to a single feature vector. The learner can obtain accurate feature vectors by replacing inaccurate sensors with high-quality sensors, but the number of high-quality sensors is limited to k. The learner selects k features for observing their accurate feature vectors.

We formalize this scenario as the following problem. At the beginning, a learner knows a response vector $\mathbf{b} \in \mathbb{R}^m$ and a prior distribution over the features, but does not know the features themselves. Namely, we regard the inaccurate feature vectors obtained with noisy sensors as prior distributions on accurate feature vectors. A random variable Φ indicates the uncertainty over the observed feature vectors. From the noisy sensors, we can know only a prior distribution of Φ but not the true ϕ. Let $V = [n]$ be the set of features. At each step, the learner can query a feature $v \in V$ and observe its feature vector $\phi(v) \in \mathbb{R}^m$. We assume the noise of sensors are independent of each other; i.e., there exists a distribution $p_v(\phi(v))$ for each $v \in V$ and we can factorize p as $p(\phi) = \prod_{v \in V} p_v(\phi(v))$.

Let $\mathbf{A}(\phi) = (\phi(1) \cdots \phi(n))$ be the realized feature matrix under realization ϕ. The objective function to be maximized is defined as $f(S,\phi) = \|\mathbf{b}\|_2^2 - \min_{\mathbf{w} \in \mathbb{R}^S} \|\mathbf{b} - \mathbf{A}(\phi)_S \mathbf{w}\|_2^2$.

7.1. Bound of Adaptive Submodularity Ratio

To bound the adaptive submodularity ratio of adaptive feature selection, we give a general lower bound of the adaptive submodularity ratio by using (non-adaptive) submodularity ratios of all realizations.

Theorem 4. *Let $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ be adaptive monotone w.r.t. distribution $p(\phi)$. Assume the value of $f(S,\phi)$ depends only on $(\phi(v))_{v \in S}$ not on $(\phi(v))_{v \in V \setminus S}$, i.e., $f(S,\phi) = f(S,\phi')$ for all ϕ and ϕ' such that $\phi(v) = \phi(v)$ for all $v \in S$. We also assume $p(\phi)$ can be factorized to distributions $p_v(\phi(v))$ of states of each $v \in V$, i.e., $p(\phi) = \prod_{v \in V} p_v(\phi(v))$. Let $\gamma^{\phi}_{X,k}$ be the submodularity ratio of $f(\cdot,\phi)$ for each realization ϕ. For any distribution p_v of $\phi(v)$, the adaptive submodularity ratio $\gamma_{\psi,k}$ can be bounded as*

$$\gamma_{\psi,k} \geq \min_{\phi \sim \psi} \gamma^{\phi}_{\mathrm{dom}(\psi),k}.$$

By using Theorem 4 and the result of (Das & Kempe, 2011), we obtain the following lower bound of the adaptive submodularity ratio.

Corollary 2. *Assume each column of $\mathbf{A}(\phi)$ is normalized. For any $\mathbf{b} \in \mathbb{R}^n$ and any distribution p_v of each $\phi(v)$, the adaptive submodularity ratio $\gamma_{\ell,k}$ can be bounded as*

$$\gamma_{\ell,k} \geq \min_{\phi} \min_{S \subseteq V : |S| \leq k+\ell} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S),$$

where $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue.

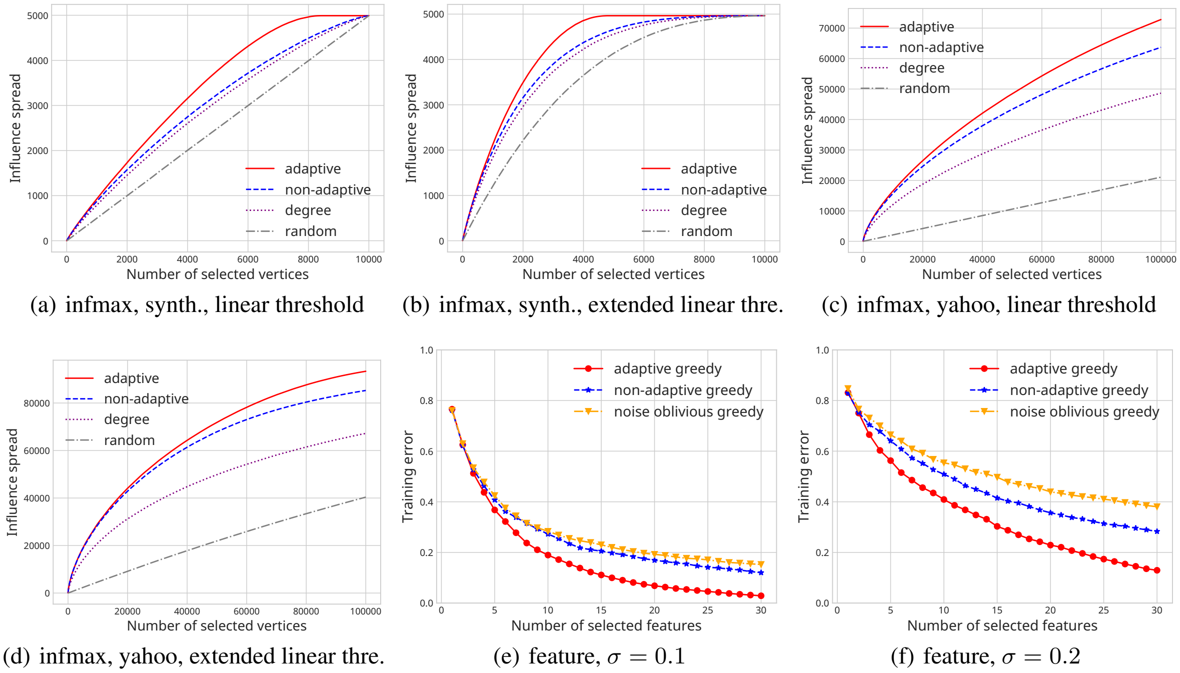
7.2. Bound of Adaptivity Gap

We can also obtain a bound on the adaptivity gap of adaptive feature selection as follows:

Proposition 4. *Let $f(S,\phi) = \|\mathbf{b}\|_2^2 - \min_{\mathbf{w} \in \mathbb{R}^S} \|\mathbf{b} - \mathbf{A}(\phi)_S \mathbf{w}\|_2^2$ and suppose that $p(\phi)$ can be factorized as $p(\phi) = \prod_{v \in V} p_v(\phi(v))$. We have*

$$\mathsf{GAP}_k \geq \frac{\min_{\phi} \min_{S \subseteq V\colon |S| \leq k} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)}{\max_{\phi} \max_{S \subseteq V\colon |S| \leq k} \lambda_{\max}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)}.$$

Remark 1. These results on the adaptive submodularity ratio and adaptivity gap can be extended to more general loss functions with restricted strong concavity and restricted smoothness as in (Elenberg et al., 2018).



(a) infmax, synth., linear threshold

(b) infmax, synth., extended linear thre.

(c) infmax, yahoo, linear threshold

(d) infmax, yahoo, extended linear thre.

(e) feature, $\sigma = 0.1$

(f) feature, $\sigma = 0.2$

Figure 2. Experimental results on adaptive influence maximization (a)–(d) and adaptive feature selection (e)–(f). (a) and (b) are the results on synthetic datasets with the linear threshold model and extended linear threshold model, respectively. (c) and (d) are the results on Yahoo! dataset (Yah) with the linear threshold model and extended linear threshold model, respectively. (e) and (f) are the results on synthetic datasets with uniform noise distribution on $[-\sigma, \sigma]$ with $\sigma = 0.1, 0.2$, respectively.

8. Experiments

We conduct experiments on two applications: adaptive influence maximization and adaptive feature selection. For each setting, we conduct 20 trials and plot their mean values.

8.1. Adaptive Influence Maximization

Datasets. We conduct experiments on two datasets of adaptive influence maximization. The first dataset is a synthetic bipartite graph generated randomly according to Erdös–Renyi rule. We set the number of source and sink vertices to 10000, i.e., $|V| = |U| = 10000$. For each pair $(v, u) \in V \times U$, we add an edge between v and u with probability 0.001. The second dataset is Yahoo! Search Marketing Advertiser–Phrase Bipartite Graph (Yah), which is a bipartite graph representing relationships between advertisers and search phrases; we have $|V| = 459678$, $|U| = 193582$, and $|A| = 2278448$. For both datasets, the weight of each vertex in U is drawn from the uniform distribution on $[0, 1]$.

Diffusion Model. We consider two diffusion models. The first one is the linear threshold model. The probability that each edge $(v, u) \in A$ is alive is set to the reciprocal of the degree of the sink vertex, that is, $1/|\delta_-(v)|$. As the second diffusion model, we consider an extended version of the linear threshold model, which is also a special case of the triggering model. In this model, for each sink vertex v, the subset of incoming live edges is determined as follows. We sample t edges with replacement from $\delta_-(v)$ uniformly at random, and an edge turns alive if it is sampled at least once. In our experiments, parameter t is set to 3.

Benchmarks. We compare the adaptive greedy algorithm with three non-adaptive benchmarks. The first benchmark is the non-adaptive greedy algorithm, called non-adaptive, which is a standard greedy algorithm (Nemhauser et al., 1978) for maximizing the expected value of the objective function $\mathbb{E}_\Phi[f(\cdot, \Phi)]$. The second benchmark is Degree, which selects the set of vertices with the top-k largest degree. The third benchmark is Random, which selects a random subset of size k.

Results. Objective values achieved by the algorithms are shown in Figures 2(a) to 2(d). In all settings, the adaptive greedy algorithm outperforms all the benchmarks.

8.2. Adaptive Feature Selection

Datasets. We use synthetic datasets generated randomly as follows. First we determine the mean $\mathbb{E}_\Phi[\mathbf{A}(\Phi)] \in \mathbb{R}^{m \times n}$ according to the uniform distribution on $[0, 1]$. After that, each column is normalized so that its mean is 0 and its standard deviation is 1. We obtain $\mathbf{A}(\phi)$ by adding

$\epsilon \in \mathbb{R}^{m \times n}$ to $\mathbb{E}_\Phi[\mathbf{A}(\Phi)]$, where each element of ϵ is drawn from the uniform distribution on $[-\sigma, \sigma]$. We consider two settings: $\sigma = 0.1$ and 0.2. We select a random sparse subset S^* of features such that $|S^*| = 30$, and we let $\mathbf{y} = \mathbf{A}(\phi)_{S^*}\mathbf{w}$ be the response vector, where each element of $\mathbf{w} \in \mathbb{R}^S$ is drawn from the standard normal distribution. In all settings, we set $n = 1000$ and $m = 100$.

Benchmarks. We compare the adaptive greedy algorithm with two benchmarks. The first benchmark is the non-adaptive greedy algorithm. Regarding the adaptive and non-adaptive greedy algorithms, it is hard to evaluate the exact values of the objective functions, and so we approximately evaluate them by sampling $\mathbf{A}(\Phi)$ randomly according to posterior distributions. The second benchmark is the noise-oblivious greedy algorithm, a non-adaptive algorithm that greedily selects a subset based on the mean, $\mathbb{E}_\Phi[\mathbf{A}(\Phi)]$.

Results. The results are shown in Figures 2(e) and 2(f). In both settings, the adaptive greedy algorithm outperforms the two benchmarks.

9. Related Work

Comparison with (Kusner, 2014). To our knowledge, the first attempt to generalize submodularity ratio to the adaptive setting is (Kusner, 2014). They defined *approximate adaptive submodularity*, a notion that is similar to ours, as follows:

$$\gamma = \min_{S \subseteq V, \psi} \frac{\sum_{v \in S} \Delta(v|\psi)}{\Delta(S|\psi)}.$$

The key difference is that they did not replace subset S with policy π. In Appendix F, we show that the approximate adaptive submodularity is not sufficient for providing an approximation guarantee of the adaptive greedy algorithm.

Comparison with (Yong et al., 2017). Another attempt to relax adaptive submodularity is presented in (Yong et al., 2017). They introduced *ζ-weakly adaptive submodular functions* as follows:

Definition 5 (ζ-weak adaptive submodularity)**.** Let $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ be a set function and p be a distribution of ϕ. For any $\zeta \geq 1$, we say f is adaptive submodular with respect to p if for any partial realization $\psi \subseteq \psi'$ and any element $v \in V \setminus \mathrm{dom}(\psi')$, it holds $\zeta\Delta(v|\psi) \geq \Delta(v|\psi')$. Let ζ^* be the infimum of ζ satisfying the above inequality.

Analogous to our adaptive submodularity ratio, one can readily see that 1-weak adaptive submodularity is equivalent to the adaptive submodularity. In general, however, there is a difference between the two notions; the adaptive submodularity ratio can be bounded from below by $1/\zeta^*$, implying that it is more demanding to bound the value of ζ^* than that of the adaptive submodularity ratio.

Proposition 5. *For any set function $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ and distribution p, we have $\frac{1}{\zeta^*} \leq \min_{k \in \mathbb{Z}_{\geq 0}, \psi} \gamma_{\psi,k}$.*

We provide a proof in Appendix G.1. Yong et al. (2017) studied a problem called *group-based active diagnosis* and gave a bound of ζ, but some vital assumptions seem to have been missed. In Appendix G.2, we provide a problem instance in which their bound does not hold. We also present instances of adaptive influence maximization and adaptive feature selection for which our framework provides strictly better approximation ratios than those obtained with the weak adaptive submodularity in Appendices G.3 and G.4.

Adaptive Submodularity. Adaptive submodularity was proposed by Golovin & Krause (2011). There are several attempts to adaptively maximize set functions that do not satisfy adaptive submodularity (e.g., (Kusner, 2014; Yong et al., 2017)). Chen et al. (2015) analyzed the greedy policy focusing on the maximization of mutual information, which does not have adaptive submodularity.

Submodularity Ratio. Submodularity ratio was proposed by Das & Kempe (2011) for sparse regression with squared ℓ_2 loss. Recently, Elenberg et al. (2018) extended this result to more general loss functions with restricted strong convexity and restricted smoothness. Bogunovic et al. (2018) proposed the notion of *supermodularity ratio*. Bian et al. (2017) provided a guarantee of the non-adaptive greedy algorithm for the case where the total curvature and submodularity ratio of objective functions are bounded.

Influence Maximization. Influence maximization was proposed by Kempe et al. (2003). An adaptive version of influence maximization was first considered by Golovin & Krause (2011). They showed that this objective function satisfies adaptive submodularity under the independent cascade model in general graphs. Influence maximization on a bipartite graph has been studied for applications to advertisement selection (Alon et al., 2012; Soma et al., 2014). This problem setting was extended to the adaptive setting by Hatano et al. (2016), but only the independent cascade model was considered. The curvature of its objective function was studied by Maehara et al. (2017).

Feature Selection. Kale et al. (2017) considered the problem called adaptive feature selection, but their problem setting is different from ours. In their setting, the learner solves feature selection problems multiple times. They studied the adaptivity among the multiple rounds, while we studied the adaptivity inside of a single round.

Acknowledgements

K.F. was supported by JSPS KAKENHI Grant Number JP 18J12405.

References

Yahoo! webscope dataset: G1 - Yahoo! Search Marketing Advertiser-Phrase Bipartite Graph, Version 1.0. URL https://webscope.sandbox.yahoo.com/.

Alon, N., Gamzu, I., and Tennenholtz, M. Optimizing budget allocation among channels and influencers. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012*, pp. 381–388, 2012.

Balkanski, E. and Singer, Y. The adaptive complexity of maximizing a submodular function. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pp. 1138–1151, 2018.

Bian, A. A., Buhmann, J. M., Krause, A., and Tschiatschek, S. Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pp. 498–507, 2017.

Bogunovic, I., Zhao, J., and Cevher, V. Robust maximization of non-submodular objectives. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, pp. 890–899, 2018.

Chen, Y., Hassani, S. H., Karbasi, A., and Krause, A. Sequential information maximization: When is greedy near-optimal? In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pp. 338–363, 2015.

Das, A. and Kempe, D. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 1057–1064, 2011.

Elenberg, E. R., Dimakis, A. G., Feldman, M., and Karbasi, A. Streaming weak submodularity: Interpreting neural networks on the fly. In *Advances in Neural Information Processing Systems 30*, pp. 4047–4057, 2017.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. Restricted strong convexity implies weak submodularity. *Ann. Statist.*, 46(6B):3539–3568, 2018.

Gabillon, V., Kveton, B., Wen, Z., Eriksson, B., and Muthukrishnan, S. Adaptive submodular maximization in bandit setting. In *Advances in Neural Information Processing Systems 26*, pp. 2697–2705, 2013.

Golovin, D. and Krause, A. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artif. Intell. Res.*, 42:427–486, 2011.

Golovin, D., Krause, A., and Ray, D. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems 23*, pp. 766–774, 2010.

Hatano, D., Fukunaga, T., and Kawarabayashi, K. Adaptive budget allocation for maximizing influence of advertisements. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 3600–3608, 2016.

Javdani, S., Chen, Y., Karbasi, A., Krause, A., Bagnell, D., and Srinivasa, S. S. Near optimal bayesian active learning for decision making. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, AISTATS 2014*, pp. 430–438, 2014.

Kale, S., Karnin, Z., Liang, T., and Pál, D. Adaptive feature selection: Computationally efficient online sparse linear regression under RIP. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pp. 1780–1788, 2017.

Kempe, D., Kleinberg, J. M., and Tardos, É. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003*, pp. 137–146, 2003.

Khanna, R., Elenberg, E. R., Dimakis, A. G., Negahban, S., and Ghosh, J. Scalable greedy feature selection via weak submodularity. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, pp. 1560–1568, 2017.

Kusner, M. J. Approximately adaptive submodular maximization. In *NIPS Workshop on Discrete and Combinatorial Problems in Machine Learning*, 2014.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J. M., and Glance, N. S. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2007*, pp. 420–429, 2007.

Maehara, T., Kawase, Y., Sumita, H., Tono, K., and Kawarabayashi, K. Optimal pricing for submodular valuations with bounded curvature. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 622–628, 2017.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions - I. *Math. Program.*, 14(1):265–294, 1978.

Soma, T., Kakimura, N., Inaba, K., and Kawarabayashi, K. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pp. 351–359, 2014.

Tang, Y., Xiao, X., and Shi, Y. Influence maximization: near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD 2014*, pp. 75–86, 2014.

Yong, S. Z., Gao, L., and Ozay, N. Weak adaptive submodularity and group-based active diagnosis with applications to state estimation with persistent sensor faults. In *2017 American Control Conference (ACC)*, pp. 2574–2581, 2017.