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 $\lambda_{\min,\ell} = \min_\phi \min_{S \subseteq V \colon |S| \le \ell} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)$ and $\lambda_{\max,\ell} = \max_\phi \max_{S \subseteq V \colon |S| \le \ell} \lambda_{\max}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)$. 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 $\psi = \{(v_1, \phi(v_1)), \dots, (v_\ell, \phi(v_\ell))\}$ the partial realization observed so far, where $\{v_1, \dots, v_\ell\}$ 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 ψ_ℓ a set of observations of size ℓ.

```
1: $\psi_0 \leftarrow \emptyset$.
2: for $i = 1, \dots, \ell$ do
3:     $v \leftarrow \mathrm{argmax}_{v \in V} \Delta(v|\psi_{i-1})$.
4:     Observe $\phi(v)$ and let $\psi_i \leftarrow \psi_{i-1} \cup \{(v, \phi(v))\}$.
5: end for
6: 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

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

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. □

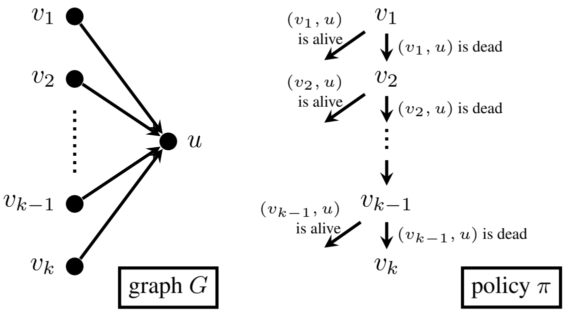


Figure 1. An example that implies the tightness of our bound.

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.

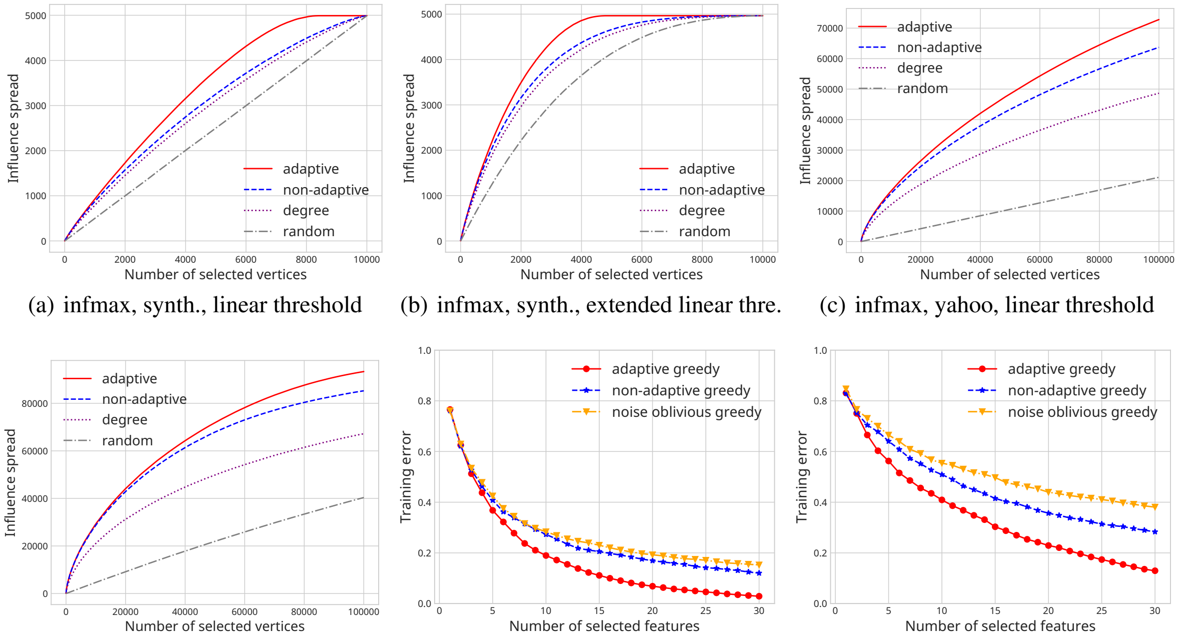
7.2. Bound of Adaptivity Gap

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*

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

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.

Appendices

A. Proof for Adaptive Submodularity Ratio

Proof of Proposition 1. First we deal with the "if" part. Let ψ_v be the partial realization just before v is selected in π. If there are multiple partial realizations ψ such that $\pi(\psi) = v$, we can duplicate v and take them to be different elements. From adaptive submodularity, for any partial realization ψ and policy π, we have

$$\begin{aligned}\Delta(\pi|\psi) &= \sum_{v\in V} \Pr(v \in E(\pi,\Phi)|\Phi \sim \psi)\Delta(v|\psi \cup \psi_v) \\ &\le \sum_{v\in V} \Pr(v \in E(\pi,\Phi)|\Phi \sim \psi)\Delta(v|\psi).\end{aligned}$$

Thus we can see $\gamma_{\psi,k} \ge 1$. Moreover, if π is a policy that selects a single element, the above inequality holds with equality. These two facts imply $\gamma_{\psi,k} = 1$.

Next we deal with the "only if" part. Let $\psi \subseteq \psi'$ be any partial realization such that $|\psi| + 1 = |\psi'|$ and $v \in V \setminus \mathrm{dom}(\psi')$ be any element. We define $u \in \mathrm{dom}(\psi') \setminus \mathrm{dom}(\psi)$ to be the additional element and y its state in ψ', i.e., $\psi' = \psi \cup \{(u, y)\}$. Let us consider a policy π that first selects u and, if $\phi(u) = y$, proceeds to select v. From the assumption, we have $\gamma_{\psi,2} = 1$, and thus $\Delta(\pi|\psi) \le \sum_{v\in V} \Pr(v \in E(\pi, \Phi))\Delta(v|\psi)$. We can calculate the left and right hand sides as follows:

$$\begin{aligned}\text{(LHS)} &= \Delta(u|\psi) + \Pr(\Phi(u) = y|\Phi \sim \psi)\Delta(v|\psi'), \\ \text{(RHS)} &= \Delta(u|\psi) + \Pr(\Phi(u) = y|\Phi \sim \psi)\Delta(v|\psi).\end{aligned}$$

Therefore, we obtain $\Delta(v|\psi') \le \Delta(v|\psi)$. By sequentially concatenating inequalities of this type, we can show that the statement holds for any $\psi \subseteq \psi'$. □

B. Proof for the Adaptive Greedy Algorithm

To prove Theorem 1, we introduce a lemma provided by Golovin & Krause (2011). Let $\pi'@\pi$ be a concatenated policy, i.e., a policy that executes π' as if from scratch after executing π. Adaptive monotonicity is known to be equivalent to the following condition:

Lemma 1 (Adopted from (Golovin & Krause, 2011, Lemma A.8))**.** *Fix $f : 2^V \times \mathcal{Y}^V \to \mathbb{R}_{\ge 0}$. Then we have $\Delta(v|\psi) \ge 0$ for all ψ with $p(\psi) > 0$ and all $v \in V$ if and only if for all policies π and π', we have $f_{\mathrm{avg}}(\pi) \le f_{\mathrm{avg}}(\pi'@\pi)$.*

Proof of Theorem 1. Let ψ be any possible partial realization that can appear while executing the adaptive greedy policy π. Since π stops after ℓ steps, we have $|\psi| \le \ell$. According to the definition of adaptive submodularity ratio, we have

$$\gamma_{\ell,k}\Delta(\pi^*|\psi) \le \sum_{v\in V} \Pr(v \in E(\pi^*, \Phi)|\Phi \sim \psi)\Delta(v|\psi) \le k \max_{v\in V} \Delta(v|\psi) \tag{A1}$$

since $\sum_{v\in V} \Pr(v \in E(\pi^*, \Phi)|\Phi \sim \psi) = \mathbb{E}[|E(\pi^*, \Phi)|] \le k$. Let Ψ be a random partial realization observed by executing $\pi_{[i]}$, where $\pi_{[i]}$ is a policy obtained by running π until it terminates or it selects i elements. Formally, Ψ conforms to the distribution $p_\Psi(\psi) \coloneqq \Pr(\Psi = \psi \mid \exists\phi,\ \psi = \{(v, \phi(v)) \mid v \in E(\pi_{[i]}, \phi)\})$. Then we can lower-bound the expected single step gain as follows:

$$\begin{aligned}f_{\mathrm{avg}}(\pi_{[i+1]}) - f_{\mathrm{avg}}(\pi_{[i]}) &= \mathbb{E}\left[\max_{v\in V} \Delta(v|\Psi)\right] && \text{(due to the property of the adaptive greedy algorithm)} \\ &\ge \mathbb{E}\left[\frac{\gamma_{\ell,k}}{k}\Delta(\pi^*|\Psi)\right] && \text{(due to (A1))} \\ &= \frac{\gamma_{\ell,k}}{k}\left(f_{\mathrm{avg}}(\pi_{[i]}@\pi^*) - f_{\mathrm{avg}}(\pi_{[i]})\right) \\ &\ge \frac{\gamma_{\ell,k}}{k}\left(f_{\mathrm{avg}}(\pi^*) - f_{\mathrm{avg}}(\pi_{[i]})\right). && \text{(due to adaptive monotonicity and Lemma 1)}\end{aligned}$$

Let $\Delta_i \coloneqq f_{\mathrm{avg}}(\pi^*) - f_{\mathrm{avg}}(\pi_{[i]})$. The above inequality can be rewritten as $\Delta_i - \Delta_{i+1} \ge \gamma_{\ell,k}\Delta_i/k$, which implies $\Delta_{i+1} \le (1 - \gamma_{\ell,k}/k)\Delta_i$. By repeatedly using this inequality, we obtain $\Delta_\ell \le (1 - \gamma_{\ell,k}/k)^\ell\Delta_0 \le \exp(-\gamma_{\ell,k}\ell/k)f_{\mathrm{avg}}(\pi^*)$. Consequently, we have $f_{\mathrm{avg}}(\pi) \ge (1 - \exp(-\gamma_{\ell,k}\ell/k))f_{\mathrm{avg}}(\pi^*)$. □

C. Proofs for Adaptivity Gaps

Proof of Theorem 2. Let π^*_{non} be an optimal non-adaptive policy and π^* be an optimal adaptive policy. Since π^*_{non} is a non-adaptive policy, it selects the same subset for all ϕ, i.e., $E(\pi^*_{\mathrm{non}}, \phi) = E(\pi^*_{\mathrm{non}}, \phi')$ for all ϕ and ϕ'. Let $M \in \operatorname{argmax} \sum_{v \in M\colon |M| \le k} \Delta(v|\emptyset)$ and π^M_{non} the non-adaptive policy that selects M. From the optimality of π^*_{non}, we have

$$f_{\mathrm{avg}}(\pi^*_{\mathrm{non}}) \ge f_{\mathrm{avg}}(\pi^M_{\mathrm{non}}).$$

By the definition of the supermodularity ratio, we have

$$\Delta(\pi^M_{\mathrm{non}}|\emptyset) \ge \beta_{\emptyset,k} \sum_{v \in M} \Delta(v|\emptyset).$$

Note that $\sum_{v \in V} \Pr(v \in E(\pi^*, \Phi)) \le k$ and $\Pr(v \in E(\pi^*, \Phi)) \le 1$ for each $v \in V$. Due to the definition of M, we have

$$\sum_{v \in M} \Delta(v|\emptyset) \ge \sum_{v \in V} \Pr(v \in E(\pi^*, \Phi)) \Delta(v|\emptyset).$$

From the definition of adaptive submodularity ratio, we have

$$\sum_{v \in V} \Pr(v \in E(\pi^*, \Phi)) \Delta(v|\emptyset) \ge \gamma_{\emptyset,k} \Delta(\pi^*|\emptyset).$$

Combining these inequalities, we have

$$\begin{aligned} f_{\mathrm{avg}}(\pi^*_{\mathrm{non}}) &\ge \mathbb{E}_\Phi[f(\emptyset, \Phi)] + \Delta(\pi^M_{\mathrm{non}}|\emptyset) \\ &\ge \beta_{\emptyset,k}\gamma_{\emptyset,k}(\mathbb{E}_\Phi[f(\emptyset, \Phi)] + \Delta(\pi^*|\emptyset)) \\ &= \beta_{\emptyset,k}\gamma_{\emptyset,k} f_{\mathrm{avg}}(\pi^*). \end{aligned}$$

□

Proof of Corollary 1. From the approximation ratio, we have

$$f_{\mathrm{avg}}(\pi_{\mathrm{non}}) \ge \alpha f_{\mathrm{avg}}(\pi^*_{\mathrm{non}}).$$

From Theorem 2, we have

$$f_{\mathrm{avg}}(\pi^*_{\mathrm{non}}) \ge \beta_{\emptyset,k}\gamma_{\emptyset,k} f_{\mathrm{avg}}(\pi^*).$$

The above two inequalities imply the statement. □

From the following example, we can see that Theorem 2 is tight, i.e., for any rationals β and γ in $(0, 1]$, there exist f and p such that the equality holds.

Example 2. Let $V = \{u\} \cup \bigcup_{i=1}^{M} V_i$ be the ground set, where $V_i = \{v_i^1, \cdots, v_i^k\}$. Let $V_0 = \emptyset$. Let $\mathcal{Y} = \{0, 1, \cdots, M\}$. We define the probability distribution p such that $\phi(u) = y \in \mathcal{Y}$ with probability $p \in [0, 1/M]$ for each $y \ne 0$ and $\phi(u) = 0$ with probability $1 - pM$. Other elements always in state 0, i.e., $\phi(v) = 0$ with probability 1 for all $v \in V \setminus \{u\}$. We define the objective function f as

$$f(S, \phi) = \begin{cases} 1 + a|S \cap V_{\phi(u)}| & (u \in S) \\ 1 + ap(|S| - 1) & (u \notin S \text{ and } |S| \ge 1) \\ 0 & (S = \emptyset), \end{cases}$$

where $a \in \mathbb{R}_{\ge 0}$ is the parameter specified later. We have $\Delta(v|\emptyset) = 1$ for all $v \in V$. The supermodularity ratio $\beta_{\emptyset,k}$ of $\mathbb{E}[f(\cdot, \Phi)]$ is

$$\beta_{\emptyset,k} = \frac{1 + (k-1)ap}{k}.$$

The adaptive submodularity ratio $\gamma_{\emptyset,k}$ is

$$\gamma_{\emptyset,k} = \frac{k}{1 + (k-1)apM}.$$

The adaptivity gap is

$$\mathsf{GAP}_k(f, p) = \frac{1 + (k-1)ap}{1 + (k-1)apM}.$$

For any rationals $\beta \in (0, 1]$ and $\gamma \in (0, 1]$, there exist some k, a, M such that $\gamma_{\emptyset,k} = \gamma$ and $\beta_{\emptyset,k} = \beta$.

D. Proof for Adaptive Influence Maximization

In this section, we provide the full proof for Theorem 3. For the readability, we first give a proof for the case of the linear threshold model, which is a special case of the triggering model. After that, we give a proof for the case of the triggering model.

D.1. Proof for the Linear Threshold Model

Proof of Theorem 3 in the case of the linear threshold model. Let V be the source vertices, U the sink vertices, and $A \subseteq V \times U$ the directed edges. For notational simplicity, assume that $G = (V \cup U, A)$ is a complete bipartite graph, i.e., $A = V \times U$. By setting $b(a) = 0$ for all edges $a \in A$ that originally do not exist, we can assume this without loss of generality. Fix any $\psi' \subseteq \psi$ and $\pi \in \Pi_k$. It suffices to prove

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

Let $\Delta_u(\cdot|\psi')$ be the expected marginal gain obtained by activating $u \in U$. Below we explain that the above inequality can be separated for each $u \in U$; i.e., it is enough to prove the above inequality for the case where $w(u) > 0$ for just one vertex $u \in U$ and 0 for the others. The objective function is the linear sum of the one for each $u \in U$: $\Delta(\cdot|\psi') = \sum_{u\in U} \Delta_u(\cdot|\psi')$. Therefore, the above inequality is decomposed into the sum of

$$\Delta_u(\pi|\psi') \le \frac{2k}{k+1} \sum_{v\in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi')\Delta_u(v|\psi') \tag{A2}$$

for each $u \in U$. Note that the states of any $(v, u) \in A$ and $(v', u') \in A$ are independent of each other if $u \neq u'$. Since the feedback about any $u' \in U$ such that $u' \neq u$ is never correlated with the states of edges pointing to u, we can regard the feedback about u' as an independent random factor when considering (A2). Thus we can see that it is sufficient to consider the case of one sink vertex. Note that a randomized policy can be expressed as a linear sum of deterministic policies. Therefore, it is enough to consider the case where π is a deterministic policy. Below we fix $u \in U$ and use Δ instead of Δ_u for notational ease. We can assume $w(u) = 1$ without loss of generality. If u has been already activated in ψ', both sides of (A2) are equal to zero; thus it holds trivially. We then consider the case where u is not activated in ψ'.

Let ψ_v be the partial realization just before v is selected in π. If there are multiple partial realizations ψ such that $\pi(\psi) = v$, we can duplicate v and consider them to be different elements. We can decompose $\Delta(\pi|\psi')$ as

$$\Delta(\pi|\psi') = \sum_{v\in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi')\Delta(v|\psi' \cup \psi_v).$$

The inequality that we aim to prove can be written as

$$\sum_{v\in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi') \left\{ \frac{2k}{k+1}\Delta(v|\psi') - \Delta(v|\psi' \cup \psi_v) \right\} \ge 0.$$

Since π is a deterministic policy that observes only states of edges pointing to u, there exists a path in policy tree π wherein u remains inactive; in Figure 3 such a path is colored in thin gray. Let $P = \{v_1, \cdots, v_m\} \subseteq V$ be the path, where $m \le k$ and policy π selects the vertices $v_1, \cdots, v_m$ in this order. We consider proving the above inequality for P and $V\backslash P$ separately. We can easily see that $\Delta(v|\psi' \cup \psi_v) = 0$ holds for all $v \in V \setminus P$ since u is already activated there. Therefore, it is enough to prove

$$\sum_{v\in P} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi') \left\{ \frac{2k}{k+1}\Delta(v|\psi') - \Delta(v|\psi' \cup \psi_v) \right\} \ge 0. \tag{A3}$$

Now we calculate the left hand side of this inequality, which we denote by C. Since u has not been activated yet in ψ', all edges (s, u) are dead for all $s \in \mathrm{dom}(\psi')$. In the linear threshold model, we can define $p_i := b(v_i u)/(1 - \sum_{t\in V\setminus \mathrm{dom}(\psi')} b(tu))$ to be the posterior probability that edge (v, u) is alive under observations ψ' for each $i = 1, \ldots, m$. Now we have $\Pr(v_i \in E(\pi, \Phi)|\Phi \sim \psi') = \Pr(\Phi \sim \psi' \cup \psi_{v_i}|\Phi \sim \psi') = 1 - \sum_{j=1}^{i-1} p_j$. In addition, we have $\Delta(v_i|\psi') = p_i$

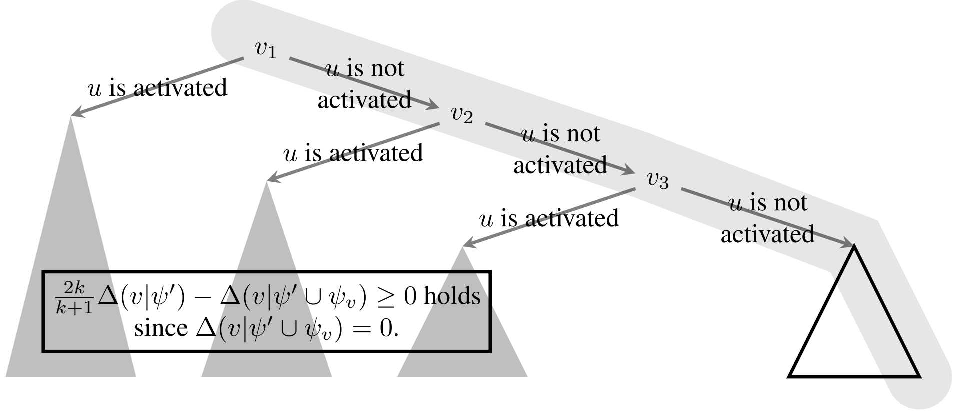


Figure 3. A description of our proof method. We can decompose the policy tree into the path wherein u is not activated and the rest.

and $\Delta(v_i|\psi' \cup \psi_{v_i}) = p_i/(1 - \sum_{j=1}^{i-1} p_j)$, and hence

$$C = \sum_{i=1}^{m} \left(1 - \sum_{j=1}^{i-1} p_j\right) \left\{ \frac{2k}{k+1} p_i - \frac{p_i}{1 - \sum_{j=1}^{i-1} p_j} \right\}.$$

In the case of $m = 1$, we have $C = (k-1)/(k+1)p_i \geq 0$. For $m \geq 2$, we obtain

$$\begin{aligned} C &= \sum_{i=1}^{m} \frac{2k}{k+1} p_i \left(1 - \sum_{j=1}^{i-1} p_j\right) - \sum_{i=1}^{m} p_i \\ &= \frac{k-1}{k+1} \sum_{i=1}^{m} p_i - \frac{2k}{k+1} \sum_{i=1}^{m} \left(p_i \sum_{j=1}^{i-1} p_j\right) \\ &= \frac{k-1}{k+1} \left\{ \sum_{i=1}^{m} p_i - \frac{2k}{k-1} \sum_{i=1}^{m} \left(p_i \sum_{j=1}^{i-1} p_j\right) \right\}. \end{aligned}$$

The right hand side can be bounded from below as

$$\begin{aligned} &\frac{k-1}{k+1} \left\{ \sum_{i=1}^{m} p_i - \frac{2k}{k-1} \sum_{i=1}^{m} \left(p_i \sum_{j=1}^{i-1} p_j\right) \right\} \\ &= \frac{k-1}{k+1} \left\{ \mathbf{1}^\top \mathbf{p} - \frac{k}{k-1} \mathbf{p}^\top (\mathbf{1}\mathbf{1}^\top - \mathbf{I}) \mathbf{p} \right\} \\ &\geq \frac{k-1}{k+1} \left\{ \mathbf{1}^\top \mathbf{p} - \frac{m}{m-1} \mathbf{p}^\top (\mathbf{1}\mathbf{1}^\top - \mathbf{I}) \mathbf{p} \right\}, \end{aligned}$$

where $\mathbf{p} = (p_1, \ldots, p_m)^\top$ and $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix. The inequality comes from $2 \leq m \leq k$ and $\mathbf{p}^\top(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{p} \geq 0$. Since each entry of $\mathbf{p}$ represents a probability, we have $\mathbf{0} \leq \mathbf{p} \leq \mathbf{1}$ and $0 \leq \mathbf{1}^\top \mathbf{p} \leq 1$. From Lemma 2 proved below, we can see that this is non-negative. Therefore, we conclude that (A3) holds. □

In the above proof, we used the following lemma.

Lemma 2. *Let $m \geq 2$ and $\mathbf{p} \in \mathbb{R}^m$ be an arbitrary vector such that $\mathbf{0} \leq \mathbf{p} \leq \mathbf{1}$ and $0 \leq \mathbf{1}^\top \mathbf{p} \leq 1$, then we have*

$$\mathbf{1}^\top \mathbf{p} - \frac{m}{m-1} \mathbf{p}^\top (\mathbf{1}\mathbf{1}^\top - \mathbf{I}) \mathbf{p} \geq 0.$$

Proof. Let $\mathbf{U} = (\mathbf{u}_1 \cdots \mathbf{u}_m) \in \mathbb{R}^{m \times m}$ be an orthonormal matrix whose first column is defined as $\mathbf{u}_1 = \mathbf{1}/\sqrt{m}$; we can write $\mathbf{p} = \mathbf{U}\mathbf{q}$ with some vector $\mathbf{q} = (q_1, \dots, q_m)^\top$. Since $\mathbf{u}_1^\top \mathbf{u}_i = 0$ for all $i \neq 1$, we obtain $\mathbf{U}^\top \mathbf{1} = (\sqrt{m}, 0, \dots, 0)^\top$. Hence the left hand side of the target inequality can be rewritten as

$$\begin{aligned}\mathbf{1}^\top \mathbf{p} - \frac{m}{m-1}\mathbf{p}^\top(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{p} &= \mathbf{1}^\top \mathbf{U}\mathbf{q} - \frac{m}{m-1}\mathbf{q}^\top \mathbf{U}^\top(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{U}\mathbf{q} \\ &= \frac{m}{m-1}(\|\mathbf{q}\|_2^2 - mq_1^2) + \sqrt{m}q_1 \\ &= \sqrt{m}q_1(1 - \sqrt{m}q_1) + \frac{m}{m-1}(q_2^2 + \cdots + q_m^2).\end{aligned}$$

Since we have $0 \leq \mathbf{1}^\top \mathbf{p} = \sqrt{m}q_1 \leq 1$, this value is non-negative. □

D.2. Proof for the Triggering Model

Proof of Theorem 3. The outline of the proof for the triggering model is the same as the one for the linear threshold model. In the case of the triggering model, we can write C as follows:

$$C = \sum_{i=1}^{m} \Pr\left(\bigwedge_{j=1}^{i-1} X_j = 0 \middle| \Phi \sim \psi'\right)\left\{\frac{2k}{k+1}\Pr\left(X_i = 1 | \Phi \sim \psi'\right) - \Pr\left(X_i = 1 \middle| \Phi \sim \psi', \bigwedge_{j=1}^{i-1} X_j = 0\right)\right\},$$

where X_i is an event in which edge (v_i, u) is alive. Different from the linear threshold model, we cannot express C explicitly with parameters. Hence we define

$$\begin{aligned} p_i &\coloneqq \Pr(X_i = 1 | \Phi \sim \psi') \quad \text{for } i = 1, \cdots, m, \\ a_i &\coloneqq \Pr\left(X_i = 1 \wedge \left\{\textstyle\bigwedge_{j=1}^{i-1} X_i = 0\right\} \middle| \Phi \sim \psi'\right) \quad \text{for } i = 1, \cdots, m, \\ \text{and} \quad h_i &\coloneqq \Pr\left(\left\{\textstyle\bigwedge_{j=1}^{i} X_i = 0\right\} \middle| \Phi \sim \psi'\right) \quad \text{for } i = 0, \cdots, m. \end{aligned}$$

With these definitions, we can calculate C as

$$\begin{aligned} C &= \sum_{i=1}^{m} h_{i-1}\left(\frac{2k}{k+1}p_i - \frac{a_i}{h_{i-1}}\right) \\ &= \frac{2k}{k+1}\sum_{i=1}^{m} p_i h_{i-1} - \sum_{i=1}^{m} a_i. \end{aligned}$$

Our goal is to prove that

$$\frac{2k}{k+1}\sum_{i=1}^{m} p_i h_{i-1} - \sum_{i=1}^{m} a_i \geq 0.$$

Note that we have

$$0 \leq a_i \leq p_i \leq 1 \qquad \text{and} \qquad 0 \leq h_i \leq 1 \qquad \text{for } i = 1, \dots, m,$$

where $a_1 = p_1$ and $h_0 = 1$. Therefore, if $m = 1$, we have

$$\begin{aligned} \frac{2k}{k+1}\sum_{i=1}^{m} p_i h_{i-1} - \sum_{i=1}^{m} a_i &= \frac{2k}{k+1}p_1 h_0 - a_1 \\ &= \frac{k-1}{k+1}p_1 \\ &\geq 0. \end{aligned}$$

Furthermore, it holds that

$$h_i + \textstyle\sum_{j=1}^{i} a_j = \Pr\left(\left\{\bigwedge_{j=1}^{i} X_i = 0\right\}\right) + \sum_{j=1}^{i}\Pr\left(X_i = 1 \wedge \left\{\bigwedge_{j=1}^{i-1} X_i = 0\right\}\right) = 1$$

for $i = 0, \dots, m$, where $\sum_{j=1}^{0} a_j = 0$. By combining this equality with $0 \leq h_i \leq 1$, we obtain

$$0 \leq \sum_{j=1}^{i} a_j \leq 1$$

for $i = 0, \dots, m$. With these inequalities, the LHS of the target inequality can be lower-bounded as

$$\begin{aligned}
\frac{2k}{k+1} \sum_{i=1}^{m} p_i h_{i-1} - \sum_{i=1}^{m} a_i &= \frac{2k}{k+1} \sum_{i=1}^{m} p_i \left(1 - \sum_{j=1}^{i-1} a_j \right) - \sum_{i=1}^{m} a_i \\
&\geq \frac{2k}{k+1} \sum_{i=1}^{m} a_i \left(1 - \sum_{j=1}^{i-1} a_j \right) - \sum_{i=1}^{m} a_i \\
&= \frac{k-1}{k+1} \sum_{i=1}^{m} a_i - \frac{2k}{k+1} \sum_{i>j} a_i a_j \\
&= \frac{k-1}{k+1} \left(\mathbf{1}^\top \mathbf{a} - \frac{k}{k-1} \mathbf{a}^\top (\mathbf{1}\mathbf{1}^\top - \mathbf{I}) \mathbf{a} \right) \\
&\geq \frac{k-1}{k+1} \left(\mathbf{1}^\top \mathbf{a} - \frac{m}{m-1} \mathbf{a}^\top (\mathbf{1}\mathbf{1}^\top - \mathbf{I}) \mathbf{a} \right),
\end{aligned}$$

which is non-negative from Lemma 2; this completes the proof as with the case of the linear threshold model. □

D.3. Example for the Case of General Graphs

In this subsection, we provide a problem instance of a general graph in which the adaptive submodularity ratio can be very small.

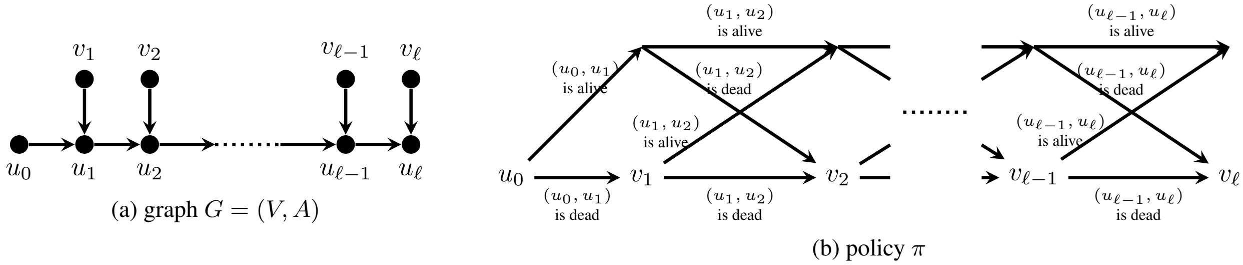
Before that, we briefly describe the problem setting of adaptive influence maximization in general graphs. Let $G = (V', A)$ be a general directed graph and $V \subseteq V'$ be a set of vertices that can be selected. At each step, the algorithm selects one vertex $v \in V$, then the influence spreads from v according to some stochastic diffusion process such as the independent cascade model or the linear threshold model. After that, the algorithm observes the diffusion from this vertex v under some feedback model. This problem includes the bipartite influence maximization as a special case where $G = (V \cup U, A)$ is a directed bipartite graph with $A \subseteq V \times U$ and $w(v) = 0$ for all $v \in V$.

There are two standard feedback models, both of which are proposed by Golovin & Krause (2011). Note that these two feedback models are equivalent in bipartite graphs. In the first feedback model called the *myopic feedback model*, the algorithm observes the states of all edges outgoing from v. Golovin & Krause (2011) proved that the adaptive submodularity does not hold in this case by giving a simple example. This analysis can be applied to both the independent cascade and linear threshold models. With this example instance, we can readily see that the adaptive submodularity ratio can be very small under the myopic feedback model. These facts imply that the myopic feedback model is typically too harsh to deal with.

In the second feedback model called the *full-adoption feedback model*, the algorithm observes the states of all edges outgoing from any vertex $u \in R(v)$ when selecting v, where $R(v)$ is the set of all vertices reachable from v only through live edges. Golovin & Krause (2011) showed that, even if graphs are general (non-bipartite), the objective function satisfies adaptive submodularity under the independent cascade model with the full-adaption feedback.

Below we show that, even under the linear threshold model with the full-adoption feedback, the adaptive submodularity ratio can be arbitrarily small if the graph is non-bipartite. This fact implies that the assumption of bipartiteness, which we imposed to obtain the bound on the adaptive submodularity ratio, is almost inevitable.

Example 3. Let G be a directed graph with vertices $V = \{v_1, \dots, v_\ell\} \cup \{u_0, u_1, \dots, u_\ell\}$ and directed edges $A = \{(u_{i-1}, u_i) \mid i = 1, \dots, \ell\} \cup \{(v_i, u_i) \mid i = 1, \dots, \ell\}$. Let w be the vertex weight such that $w(v) = 1$ for all $v \in V$. We consider the following linear threshold model: for each $i \in [\ell]$, only one of the two edges, (v_i, u_i) and (u_{i-1}, u_i), entering u_i is alive with probability ϵ and $1 - \epsilon$, respectively.



(a) graph $G = (V, A)$

(b) policy π

Figure 4. An instance with a non-bipartite graph such that the adaptive submodularity ratio can be arbitrarily small. Since the space is limited, nodes of π that have the same subtree are indicated by a single node.

Let π be a policy defined as follows. π first selects u_0. Then the realized states of some edges are revealed under the full-adoption feedback model and we can observe which vertices are activated. If u_ℓ is activated, π stops. Otherwise, there exists some $i \in [\ell]$ such that u_{i-1} is activated but u_i is not. Then π proceeds to select v_i. Repeat this procedure until u_ℓ is activated. The graph and policy are illustrated in Figure 4.

First we consider the probability $\Pr(v_i \in E(\pi, \Phi))$ for each $i \in [\ell]$. We can see that π selects v_i if and only if the edge (u_{i-1}, u_i) is dead, which yields $\Pr(v_i \in E(\pi, \Phi)) = \epsilon$. We can easily confirm that π finally activates all $u_0, \dots, u_\ell$ for every realization and each v_i is selected with probability ϵ, therefore $\Delta(\pi|\emptyset) = \ell + 1 + \epsilon\ell$. On the other hand, the numerator of the definition of the adaptive submodularity ratio can be calculated as follows. The expected marginal gain of v_i is

$$\Delta(v_i|\emptyset) = 1 + \sum_{j=i}^{\ell} \epsilon(1-\epsilon)^{j-i} = 2 - (1-\epsilon)^{\ell-i+1}.$$

Similarly, we have $\Delta(u_0|\emptyset) = \frac{1}{\epsilon}\{1 - (1-\epsilon)^{\ell+1}\}$. Finally, we can compute the adaptive submodularity ratio as

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

By setting $\epsilon = 1/\sqrt{\ell}$ and taking $\ell \to \infty$, we can see $\gamma_{\emptyset,\ell} \to 0$. To conclude, the adaptive submodularity ratio can become arbitrarily small if the graph is non-bipartite.

E. Proof for Adaptive Feature Selection

Proof of Theorem 4. Let ψ be any partial realization and $\pi \in \Pi_k$ be any policy of height at most k. Fix an arbitrary subset $\psi' \subseteq \psi$. Note that we have $f(\mathrm{dom}(\psi), \phi) = f(\mathrm{dom}(\psi), \phi')$ for any $\phi, \phi' \supseteq \psi$ due to the assumption that $f(S, \phi)$ depends only on $(\phi(v))_{v \in S}$; considering this, we abuse the notation and define $f(\psi) \coloneqq f(\mathrm{dom}(\psi), \phi)$ for any $\phi \supseteq \psi$. Let ψ_v be the partial realization just before v is selected in π. If there are multiple partial realizations ψ such that $\pi(\psi) = v$, we can duplicate v and consider them to be different elements. Now we can transform the numerator of the adaptive submodularity ratio as

$$\begin{aligned}&\sum_{v \in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi')\Delta(v|\psi') \\ &= \sum_{v \in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi')\mathbb{E}\left[f(\psi' \cup \{(v, \Phi(v))\}) - f(\psi')|\Phi \sim \psi'\right] \\ &= \sum_{v \in V} \Pr(v \in E(\pi, \Phi)|\Phi \sim \psi') \sum_{y \in \mathcal{Y}} \Pr(\Phi(v) = y|\Phi \sim \psi')\Big\{f(\psi' \cup \{(v, y)\}) - f(\psi')\Big\}\end{aligned}$$

$$= \sum_{v\in V} \Pr(v \in E(\pi,\Phi)|\Phi \sim \psi') \sum_{y\in\mathcal{Y}} \Pr(\Phi(v) = y|\Phi \sim \psi' \cup \psi_v)\Big\{f(\psi' \cup \{(v,y)\}) - f(\psi')\Big\}$$

(due to the independence of $\phi(v)$ from $(\phi(u))_{u\in\mathrm{dom}(\psi_v)}$)

$$= \sum_{v\in V} \Pr(v \in E(\pi,\Phi)|\Phi \sim \psi')\mathbb{E}\left[f(\psi' \cup \{(v,\Phi(v))\}) - f(\psi')|\Phi \sim \psi' \cup \psi_v\right]$$

$$= \sum_{v\in V} \Pr(\Phi \sim \psi' \cup \psi_v|\Phi \sim \psi')\mathbb{E}\left[f(\mathrm{dom}(\psi') \cup \{v\},\Phi)) - f(\mathrm{dom}(\psi'),\Phi)|\Phi \sim \psi' \cup \psi_v\right]$$

$$= \mathbb{E}\left[\sum_{v\in E(\pi,\Phi)} \Big\{f(\mathrm{dom}(\psi') \cup \{v\},\Phi) - f(\mathrm{dom}(\psi'),\Phi)\Big\}\middle|\Phi \sim \psi'\right].$$

From the above equality, we get

$$\begin{aligned}
&\min_{\phi\sim\psi} \gamma^{\phi}_{\mathrm{dom}(\psi),k}\Delta(\pi|\psi')\\
&= \min_{\phi\sim\psi} \gamma^{\phi}_{\mathrm{dom}(\psi),k}\mathbb{E}\left[f(\mathrm{dom}(\psi') \cup E(\pi,\Phi),\Phi) - f(\mathrm{dom}(\psi'),\Phi)|\Phi \sim \psi'\right]\\
&\le \mathbb{E}\left[\gamma^{\Phi}_{\mathrm{dom}(\psi),k}\Big\{f(\mathrm{dom}(\psi') \cup E(\pi,\Phi),\Phi) - f(\mathrm{dom}(\psi'),\Phi)\Big\}\middle|\Phi \sim \psi'\right]\\
&\le \mathbb{E}\left[\sum_{v\in E(\pi,\Phi)} \Big\{f(\mathrm{dom}(\psi') \cup \{v\},\Phi) - f(\mathrm{dom}(\psi'),\Phi)\Big\}\middle|\Phi \sim \psi'\right] \qquad \text{(From the definition of submodularity ratio)}\\
&= \sum_{v\in V} \Pr(v \in E(\pi,\phi)|\Phi \sim \psi')\Delta(v|\psi').
\end{aligned}$$

This inequality holds for any ψ and $\pi \in \Pi_k$. To conclude, we obtain $\gamma_{\psi,k} \ge \min_{\phi\sim\psi} \gamma^{\phi}_{\mathrm{dom}(\psi),k}$. □

To prove Corollary 2, we use the following bound on the submodularity ratio provided by Das & Kempe (2011).

Theorem 5 (Adopted from (Das & Kempe, 2011, Lemma 2.4))**.** *Assume each column of* $\mathbf{A}(\phi)$ *is normalized. Then*

$$\gamma_{U,k} \ge \min_{S\subseteq V:\ |S|\le k+|U|} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S).$$

Proof of Corollary 2. From the definition, we can see $f(S,\phi)$ depends only on selected columns $(\phi(v))_{v\in S}$ and not on the other columns $(\phi(v))_{v\in V\setminus S}$.

We can show

$$\begin{aligned}
f(S,\phi) &= \|\mathbf{b} - \mathbf{0}\|^2 - \min_{\mathrm{supp}(\mathbf{w})\subseteq S} \|\mathbf{b} - \mathbf{A}(\phi)\mathbf{w}\|^2\\
&\le \|\mathbf{b} - \mathbf{0}\|^2 - \min_{\mathrm{supp}(\mathbf{w})\subseteq T} \|\mathbf{b} - \mathbf{A}(\phi)\mathbf{w}\|^2 = f(T,\phi)
\end{aligned}$$

for all $S \subseteq T$. From this property, called strong adaptive monotonicity, for any partial realization ψ and $v \in V \setminus \mathrm{dom}(\psi)$, we obtain

$$\begin{aligned}
\Delta(v|\psi) &= \mathbb{E}[f(\mathrm{dom}(\psi) \cup \{v\},\Phi) - f(\mathrm{dom}(\psi),\Phi)|\Phi \sim \psi]\\
&\ge 0,
\end{aligned}$$

from which the adaptive monotonicity of f w.r.t. p follows.

By applying Theorem 4, we obtain

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

where $\gamma^{\phi}_{X,k}$ is the submodularity ratio of $f(\cdot,\phi)$ for realization ϕ. From Theorem 5, we obtain the following lower bound:

$$\gamma^{\phi}_{\mathrm{dom}(\psi),k} \ge \min_{S\subseteq V:\ |S|\le k+|\psi|} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S).$$

Finally, we have

$$\begin{aligned}\gamma_{\ell,k} &= \min_{\psi\colon |\psi|\le\ell} \gamma_{\psi,k}\\ &\ge \min_{\psi\colon |\psi|\le\ell} \min_{\phi} \gamma^{\phi}_{\mathrm{dom}(\psi),k}\\ &\ge \min_{\psi\colon |\psi|\le\ell} \min_{\phi} \min_{S\subseteq V\colon |S|\le k+|\psi|} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)\\ &= \min_{\phi} \min_{S\subseteq V\colon |S|\le k+\ell} \lambda_{\min}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S).\end{aligned}$$

□

E.1. Proof for the Adaptivity Gap

Proof of Proposition 4. We can readily confirm that the objective function can be rewritten as follows:

$$f(S,\phi) = (\mathbf{A}(\phi)_S^\top \mathbf{b})^\top (\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)^+ (\mathbf{A}(\phi)_S^\top \mathbf{b}).$$

For any $S \subseteq V$ such that $|S| \le k$, we have

$$\begin{aligned}\mathbb{E}[f(S,\Phi)] &= \mathbb{E}\left[(\mathbf{A}(\Phi)_S^\top \mathbf{b})^\top (\mathbf{A}(\Phi)_S^\top \mathbf{A}(\Phi)_S)^+ (\mathbf{A}(\Phi)_S^\top \mathbf{b})\right]\\ &\ge \mathbb{E}\left[\lambda_{\min}((\mathbf{A}(\Phi)_S^\top \mathbf{A}(\Phi)_S)^+)\|\mathbf{A}(\Phi)_S^\top \mathbf{b}\|_2^2\right]\\ &\ge \mathbb{E}\left[\frac{\|\mathbf{A}(\Phi)_S^\top \mathbf{b}\|_2^2}{\max_\phi \max_{T\subseteq V\colon |T|\le k} \lambda_{\max}(\mathbf{A}(\phi)_T^\top \mathbf{A}(\phi)_T)}\right]\\ &= \frac{\mathbb{E}\left[\|\mathbf{A}(\Phi)_S^\top \mathbf{b}\|_2^2\right]}{\max_\phi \max_{T\subseteq V\colon |T|\le k} \lambda_{\max}(\mathbf{A}(\phi)_T^\top \mathbf{A}(\phi)_T)}\\ &= \frac{\sum_{v\in S} \mathbb{E}\left[(\mathbf{A}(\Phi)_v^\top \mathbf{b})^2\right]}{\max_\phi \max_{T\subseteq V\colon |T|\le k} \lambda_{\max}(\mathbf{A}(\phi)_T^\top \mathbf{A}(\phi)_T)}\\ &= \frac{\sum_{v\in S} \mathbb{E}[f(\{v\},\Phi)]}{\max_\phi \max_{T\subseteq V\colon |T|\le k} \lambda_{\max}(\mathbf{A}(\phi)_T^\top \mathbf{A}(\phi)_T)}.\end{aligned}$$

From this inequality, we can bound the supermodularity ratio $\beta_{\emptyset,k}$ of $\mathbb{E}_\Phi[f(\cdot,\Phi)]$ as

$$\beta_{\emptyset,k} \ge \frac{1}{\max_\phi \max_{S\subseteq V\colon |S|\le k} \lambda_{\max}(\mathbf{A}(\phi)_S^\top \mathbf{A}(\phi)_S)}.$$

Plugging it and the inequality of Corollary 2 into Theorem 2, we obtain

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

□

F. Counterexample to the Statement of (Kusner, 2014)

Kusner (2014) has defined *approximate adaptive submodularity* as follows:

Definition 6 (Adopted from (Kusner, 2014, Definition 2))**.** A set function $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ and a distribution p on $\mathcal{Y}^V$ is *approximately adaptive submodular* if for any subrealization ψ such that $p(\psi) > 0$ and any $S \subseteq V \setminus \mathrm{range}(\psi)$, we have

$$\sum_{v\in S} \Delta(v|\psi) \ge \gamma\Delta(S|\psi),$$

where $\gamma \in [0,1]$ represents the submodularity ratio of the non-adaptive function.

Below we present a counterexample to the statement of (Kusner, 2014), which says that a bounded γ yields a bounded approximation ratio of the adaptive greedy algorithm.

Let $\mathcal{Y} = \{0, 1, \ldots, M-1\}$ be the set of all possible states and $V = \{u\} \cup \{z_i \mid i \in [k]\} \cup \{v_i^y \mid i \in [k-1],\ y \in \mathcal{Y}\}$ be the ground set. We define $f \colon 2^V \times \mathcal{Y}^V \to \mathbb{R}$ as follows:

$$f(S, \phi) = |S \cap \{u\}| + (1+\epsilon)|S \cap \{z_1, \ldots, z_k\}| + M \sum_{y \in \mathcal{Y}, i \in [k-1]} \mathbf{1}_{\{\phi(v_i^y)=1 \text{ and } v_i^y \in S\}},$$

where $\epsilon > 0$ is any small constant. Note that this function is normalized and adaptive monotone. For each $y \in \mathcal{Y}$, we define ϕ_y as $\phi_y(u) = y$, $\phi_y(z_i) = 0$ for each $i \in [k]$, $\phi_y(v_i^y) = 1$ for each $i \in [k-1]$, and $\phi_y(v_i^{y'}) = 0$ for each $y' \in \mathcal{Y} \setminus \{y\}$ and $i \in [k-1]$. Let p be a distribution defined as

$$p(\phi) = \begin{cases} \frac{1}{|\mathcal{Y}|} & \text{if } \phi = \phi_y \text{ for some } y \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that f is approximately adaptive submodular with $\gamma = 1$ w.r.t. p because $\Delta(\cdot|\psi)$ is a linear function for any subrealization ψ. Note that f is not adaptive submodular w.r.t. p because $\Delta(v_1^1|\emptyset) = 1 < M = \Delta(v_1^1|\{(u,1)\})$.

Kusner (2014) stated that the adaptive greedy algorithm achieves $(1 - \mathrm{e}^{-\gamma})$-approximation for any normalized, adaptive monotone, and approximately adaptive submodular function. However, the adaptive greedy algorithm achieves only $(1+\epsilon)/M$-approximation for the above f and p as is explained below. The adaptive greedy algorithm selects $z_1, \ldots, z_k$ since their expected marginal gain is $1+\epsilon$ and the expected marginal gain of other elements is 1. On the other hand, the optimal policy first selects u and proceeds to select $\{v_1^{\phi(u)}, \ldots, v_{k-1}^{\phi(u)}\}$ according to the observed $\phi(u)$. The adaptive greedy policy achieves $k(1+\epsilon)$ and the optimal policy achieves $1 + (k-1)M$. Thus the approximation ratio gets close to $(1+\epsilon)/M$ as k increases, and it can be arbitrarily small since the number of possible states, M, is not bounded. Namely, even if γ is bounded by a constant, the approximation guarantee of the adaptive greedy algorithm can become arbitrarily bad in general, which contradicts the statement of (Kusner, 2014).

G. About Comparison with (Yong et al., 2017)

G.1. Proof for Comparison with (Yong et al., 2017)

Proof of Proposition 5. From the definition of ζ^*, we have $\zeta^* \Delta(v|\psi) \geq \Delta(v|\psi')$ for any $\psi \subseteq \psi'$ and $v \in V \setminus \mathrm{dom}(\psi')$. It is enough to show $\frac{1}{\zeta^*}\Delta(\pi|\psi) \leq \sum_{v \in V} \Pr(v \in E(\pi, \phi))\Delta(v|\psi)$ for arbitrary $\psi \subseteq \psi'$ and π. Let ψ_v be the partial realization just before v is selected in π. If there are multiple partial realizations ψ such that $\pi(\psi) = v$, we can duplicate v and take them to be different elements. Then we can write $\Delta(\pi|\psi) = \sum_{v \in V} \Pr(v \in E(\pi, \phi))\Delta(v|\psi_v)$. By applying the bound of weak adaptive submodularity, we have

$$\begin{aligned} \Delta(\pi|\psi) &= \sum_{v \in V} \Pr(v \in E(\pi, \phi))\Delta(v|\psi \cup \psi_v) \\ &\leq \zeta^* \sum_{v \in V} \Pr(v \in E(\pi, \phi))\Delta(v|\psi), \end{aligned}$$

which implies the statement. □

From this proposition, we can see that Theorem 1 is stronger than the result of (Yong et al., 2017) as follows. They showed that the adaptive greedy algorithm is guaranteed to achieve $(1 - \exp(-\ell/(\zeta^* k)))$-approximation in (Yong et al., 2017, Theorem 1). From Proposition 5, we always have $(1 - \exp(-\ell/(\zeta^* k))) \leq (1 - \exp(-\gamma_{\psi,k}\ell/k))$.

G.2. Counterexample to (Yong et al., 2017, Proposition 2)

In this subsection we provide an instance of group-based active diagnosis in which the weak adaptive submodularity cannot give a bound of the approximation ratio of the adaptive greedy algorithm.

The formal problem statement of group-based active diagnosis can be described as follows. We have set V of tests and set $\mathcal{Y}$ of their possible outcomes. There are two random variables that uniquely specify the outcome of each test: the state x and the mode q. Let $\mathcal{X}$ be the set of all possible states and $\mathcal{Q}$ the set of all possible modes. We know the prior joint distribution $p(x, q)$ of x and q, but does not know their true values. Let $\mu(v, x, q) \in \mathcal{Y}$ be the unique outcome of test v when the true state is x and the true mode is q. We aim to determine x by sequentially conducting several tests out of V.

Yong et al. (2017) formulated this problem as the problem of maximizing the following objective function:

$$f(S, (x, q)) = 1 - \sum_{x' \in \mathcal{X}\colon \exists q' \in \mathcal{Q},\ \forall v \in S,\ \mu(v,x',q')=\mu(v,x,q')} \sum_{q'' \in \mathcal{Q}} p(x', q''),$$

where the first summation is about all possible $x' \in \mathcal{X}$ under the outcomes of tests S made so far. Proposition 2 of (Yong et al., 2017) claims that this objective function is ζ-weakly adaptive submodular for

$$\zeta \le \frac{|\mathcal{Q}|}{\min_{x \in \mathcal{X}, q \in \mathcal{Q}} p(x, q)}.$$

However, it does not hold in the following example.

Example 4. Let $\mathcal{X} = \{x_1, x_2\}$ be the set of states and $\mathcal{Q} = \{q_1, q_2, q_3\}$ the set of modes. For each $x \in \mathcal{X}$ and $q \in \mathcal{Q}$, we assume $p(x, q) = \frac{1}{6}$. We consider two actions v_1 and v_2, which yield the unique outcome out of $\mathcal{Y} = \{+1, -1\}$ indicated in Table 2 for each state $x \in \mathcal{X}$ and mode $q \in \mathcal{Q}$.

Table 2. Outcome

(x, q)	$\mu(v_1, x, q)$	$\mu(v_2, x, q)$
(x_1, q_1)	$+1$	$+1$
(x_1, q_2)	$+1$	-1
(x_1, q_3)	-1	$+1$
(x_2, q_1)	$+1$	$+1$
(x_2, q_2)	$+1$	-1
(x_2, q_3)	-1	-1

We first consider the expected marginal gain obtained by performing action v_2 at the beginning. In this situation, performing v_2 yields outcome $+1$ or -1 with probability $1/2$. If the outcome is $+1$, we can reject neither x_1 nor x_2. This is the case for outcome -1. Thus we have $\Delta(v_2|\emptyset) = 0$.

Next we assume the algorithm performs v_1 at the beginning and obtains the outcome of -1, i.e., $\psi = \{(v_1, -1)\}$. Now the possible pairs of the state and the mode are only (x_1, q_3) and (x_2, q_3). By performing action v_2, we obtain the outcome $+1$ or -1 with probability $\frac{1}{2}$ and reject x_2 or x_1, respectively. Thus the expected marginal gain is $\Delta(v_2|\psi) = \frac{1}{2}\mathbb{P}[x_2] + \frac{1}{2}\mathbb{P}[x_1] = \frac{1}{2}$.

From the definition of ζ, we must have $\Delta(v_2|\psi) \le \zeta\Delta(v_2|\emptyset)$, but no finite ζ satisfies this inequality. This contradicts Proposition 2 in (Yong et al., 2017), which claims ζ is finite.

G.3. Comparison in Adaptive Influence Maximization

We provide an instance of adaptive influence maximization such that the adaptive submodularity ratio yields an approximation ratio significantly better than that obtained with the weak adaptive submodularity (Yong et al., 2017).

Example 5. We use the same problem instance as Example 1. At the beginning, the expected marginal gain of v_k is $\Delta(v_k|\emptyset) = 1/k$. Let ψ be the observations obtained when $v_1, \dots, v_{k-1}$ are selected and all edges are turned out to be dead. In this case, since the edge (v_k, u) must be alive, the expected marginal gain is $\Delta(v_k|\psi) = 1$. The weak adaptive submodularity constant is lower-bounded as $\zeta \ge \Delta(v_k|\psi)/\Delta(v_k|\emptyset) = k$. This implies that the weak adaptive submodularity constant cannot yield a lower bound of the approximation ratio better than $1 - \exp(-\frac{1}{k}) = \mathrm{O}(\frac{1}{k})$, while the adaptive submodularity ratio provides a lower bound $1 - \exp(-(k+1)/2k) = \Omega(1)$.

G.4. Comparison in Adaptive Feature Selection

Regarding adaptive feature selection, we describe an advantage of the adaptive submodularity ratio in comparison with the weak adaptive submodularity (Yong et al., 2017). As detailed below, there exists an instance with the following condition: the approximation ratio obtained with the adaptive submodularity ratio is bounded, while that obtained with the weak adaptive submodularity is 0.

Example 6. We can make such an instance even if ϕ is deterministic. Let $\mathbf{A}(\phi) = (\phi(1) \cdots \phi(n))$ be the realized feature matrix under realization ϕ. The objective function is defined as

$$f(S, \phi) = \|\mathbf{b}\|_2^2 - \min_{\mathbf{w} \in \mathbb{R}^S} \|\mathbf{b} - \mathbf{A}(\phi)_S \mathbf{w}\|_2^2.$$

We here let

$$\mathbf{A}(\phi) = \begin{bmatrix} 1 & 1/\sqrt{2} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{2} & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{bmatrix} 0 \\ a \\ a \\ \vdots \\ a \end{bmatrix},$$

where $a > 0$ is an any positive real value. Let $S = \{3, \ldots, n\}$ and $T = \{2, \ldots, n\}$, which satisfy $S \subseteq T$. Then, we have

$$\min_{\mathbf{w} \in \mathbb{R}^S} \|\mathbf{b} - \mathbf{A}(\phi)_S \mathbf{w}\|_2^2 = \min_{\mathbf{w} \in \mathbb{R}^{S \cup \{1\}}} \|\mathbf{b} - \mathbf{A}(\phi)_{S \cup \{1\}} \mathbf{w}\|_2^2 = a^2$$

and

$$\min_{\mathbf{w} \in \mathbb{R}^T} \|\mathbf{b} - \mathbf{A}(\phi)_S \mathbf{w}\|_2^2 = \frac{a^2}{2} > \min_{\mathbf{w} \in \mathbb{R}^{T \cup \{1\}}} \|\mathbf{b} - \mathbf{A}(\phi)_{T \cup \{1\}} \mathbf{w}\|_2^2 = 0.$$

Therefore, we obtain

$$f(S \cup \{1\}, \phi) - f(S, \phi) = a^2 - a^2 = 0 \qquad \text{and} \qquad f(T \cup \{1\}, \phi) - f(T, \phi) = \frac{a^2}{2} - 0 = \frac{a^2}{2},$$

which implies that ζ cannot be bounded from above in general. On the other hand, the largest and smallest eigenvalues of the Hessian, $\mathbf{A}(\phi)^\top \mathbf{A}(\phi)$, are $1 + 1/\sqrt{2}$ and $1 - 1/\sqrt{2}$, respectively. Therefore, the condition number is bounded from above by $3 + 2\sqrt{2}$, which means the adaptive submodularity ratio is bounded from below by $1/(3 + 2\sqrt{2})$.