

Optimal Sampling Gaps for Adaptive Submodular Maximization

Shaojie Tang,¹ Jing Yuan,²

¹Naveen Jindal School of Management, University of Texas at Dallas

² Department of Computer Science, University of Texas at Dallas
shaojie.tang@utdallas.edu, csyuanjing@gmail.com

Abstract

Running machine learning algorithms on large and rapidly growing volumes of data is often computationally expensive, one common trick to reduce the size of a data set, and thus reduce the computational cost of machine learning algorithms, is *probability sampling*. It creates a sampled data set by including each data point from the original data set with a known probability. Although the benefit of running machine learning algorithms on the reduced data set is obvious, one major concern is that the performance of the solution obtained from samples might be much worse than that of the optimal solution when using the full data set. In this paper, we examine the performance loss caused by probability sampling in the context of adaptive submodular maximization. We consider a simple probability sampling method which selects each data point with probability $r \in [0, 1]$. If we set the sampling rate $r = 1$, our problem reduces to finding a solution based on the original full data set. We define sampling gap as the largest ratio between the optimal solution obtained from the full data set and the optimal solution obtained from the samples, over independence systems. Our main contribution is to show that if the utility function is policywise submodular, then for a given sampling rate r , the sampling gap is both upper bounded and lower bounded by $1/r$. One immediate implication of our result is that if we can find an α -approximation solution based on a sampled data set (which is sampled at sampling rate r), then this solution achieves an αr approximation ratio against the optimal solution when using the full data set. We show that the property of policywise submodular can be found in a wide range of real-world applications, including pool-based active learning and adaptive viral marketing.

Introduction

Many machine learning methods are highly benefitted when they are fed with the right volume of data. One common approach to reduce the volume of a large data set is probability sampling, which generates a sampled data set by including each data point with a known probability. However, one major concern of running an algorithm on a sampled data set is that the performance of the sampling-based solution might be much worse than that of the optimal solution when using the full data set. In this paper, we examine the performance

loss caused by probability sampling in the context of adaptive submodular maximization over independence systems.

Due to the wide applicability of submodular functions, submodular maximization, whose objective is to select a group of items to maximize a submodular function on various types of independence systems, including matroid [Nemhauser, Wolsey, and Fisher 1978; Calinescu et al. 2007] and knapsack [Sviridenko 2004], has been extensively studied in the literature. Most of existing studies focus on the non-adaptive setting, where each item has a deterministic state and all items must be selected at once. However, the classic notation of submodularity can not capture the interactive nature of many applications, including active learning and experimental design, where one must adaptively select a group of items based on the stochastic observations collected from past selections. Recently, [Golovin and Krause 2011] introduce the notation of adaptive submodularity, which extends the classic notation of submodularity from sets to policies. In the problem of adaptive submodular maximization, each item has a particular state which is unknown initially. One must pick an item before observing its realized state. An adaptive policy can be represented using a decision tree which specifies which item to pick next based on the realizations observed so far.

In [Golovin and Krause 2011], they develop a simple adaptive greedy algorithm that achieves a $1 - 1/e$ approximation for the problem of cardinality constrained adaptive submodular maximization. Their algorithm starts with an empty set, and in each iteration, it selects an item with the largest marginal utility on top of the current observation. This algorithm requires $n \times k$ value oracle queries, where n is the size of the ground set and k is the cardinality constraint. However, evaluating the marginal utility of an item is expensive in many data intensive applications, making the standard greedy algorithm infeasible in practice. One natural idea to reduce the computational cost of any machine learning algorithms is to run them on a reduced ground set that is sampled from the full set. One major concern is that the output restricted to the sampled data set might be much worse than that of the optimal solution when using the full data set. This raises the *sampling gap* question:

What is the maximum ratio between the expected utility of the optimal solution (over independence systems) when using the full data set and the optimal solution when using

the sampled data set?

If this sampling gap is small then we can focus on finding a good solution based on the sampled data set while enjoying its benefits of reduced computational cost and small performance loss. In this work, we consider a simple sampling method that selects each item from the full set with probability $r \in [0, 1]$. Our objective is to examine the performance loss due to the probability sampling in the context of adaptive submodular maximization.

Overview of Results. We first introduce a class of stochastic functions, called *policywise submodular function*. Policywise submodularity refers to the property of diminishing returns over optimal policies, and we show that this property can be found in a wide range of real-world applications, including pool-based active learning and adaptive viral marketing. Our main contribution is to show that if the utility function is policywise submodular, then for a given sampling rate r , the sampling gap, i.e., the maximum ratio between the optimal solution based on the full data set and the optimal solution based on samples, over independence systems, is both upper bounded and lower bounded by $1/r$. One major implication of our result is that if we can find an α -approximation solution based on a sampled data set (which is sampled at sampling rate r), then this solution achieves an αr approximation ratio against the optimal solution when using the full data set.

Related Works

The adaptive variant of submodular maximization has been extensively studied in the literature [Chen and Krause 2013; Tang and Yuan 2020; Tang 2020; Yuan and Tang 2017; Fujii and Sakaue 2019; Gabillon et al. 2013; Golovin, Krause, and Ray 2010; Alaei, Makhdoomi, and Malekian 2021]. For the case of maximizing an adaptive monotone and adaptive submodular function subject to a cardinality constraint, [Golovin and Krause 2011] develops a simple adaptive greedy policy that achieves a tight $1 - 1/e$ approximation ratio. For maximizing a nonmonotone adaptive submodular function, [Tang 2021, 2022] develop the first constant approximation algorithms. Given the rapid growth of data volume, much recent research in submodular maximization has explored the possibility of developing fast and practical algorithms [Leskovec et al. 2007; Badanidiyuru and Vondrák 2014; Mirzasoleiman, Badanidiyuru, and Karbasi 2016; Ene and Nguyen 2018; Mirzasoleiman et al. 2015; Tang 2021; Feldman, Harshaw, and Karbasi 2017], and many of them have adopted the technique of random sampling to develop better approximation algorithms or reduce the computational cost for maximizing (non-adaptive) submodular functions subject to various constraints. The other line of work studies the problem of maximizing (non-adaptive) monotone submodular functions from samples [Balkanski, Rubinfeld, and Singer 2016]. Unlike our setting where the utility function (of the sampled data set) is known, they assume that the function optimized is not known a priori, it is learned from data. Summarizing, all aforementioned results focus on the non-adaptive setting and their goal is to investigate the impact of random sampling on maximizing submodular functions subject a particular algorithm or a particular constraint.

Our study complements the existing studies by establishing a general framework for measuring the performance loss of the optimal solution caused by the probability sampling in the context of adaptive submodular maximization. Our results are not restricted to any particular algorithms or constraints, and the utility function is not necessarily monotone and it might take on negative values.

Preliminaries

We start by introducing some important notations. In the rest of this paper, we use $|X|$ to denote the cardinality of a set X .

Independence System

An *independence system* \mathcal{I} on the set V is a collection of subsets of V such that:

1. $\emptyset \in \mathcal{I}$;
2. \mathcal{I} , which is called the *independent sets*, is downward-closed, that is, $A \in \mathcal{I}$ and $B \subseteq A$ implies that $B \in \mathcal{I}$.

Examples of independence systems include matroid, knapsack, matching, and independent set. The *upper rank* $\text{rank}(\mathcal{I})$ of an independence system \mathcal{I} on V is defined as the size of the largest subset from \mathcal{I} , i.e., $\text{rank}(\mathcal{I}) = \max_{I \in \mathcal{I}} |I|$. For any two sets $S \in \mathcal{I}$ and $R \subseteq V$ such that $S \cap R = \emptyset$, let $\mathcal{I}_S^R = \{A \mid A \cup S \in \mathcal{I}, A \subseteq R\}$.

We next present three useful properties of any independence system (V, \mathcal{I}) . These properties will be used later to derive the main results of this paper.

Lemma 1 Consider any two sets $S \in \mathcal{I}$ and $R \subseteq V$ such that $S \cap R = \emptyset$, we have (V, \mathcal{I}_S^R) is an independence system and $\mathcal{I}_S^R \subseteq \mathcal{I}$.

Lemma 2 Assume $\text{rank}(\mathcal{I}) > 0$. For any item $e \in V$ such that $\{e\} \in \mathcal{I}$, we have $(V, \mathcal{I}_{\{e\}}^{V \setminus \{e\}})$ is an independence system and $\text{rank}(\mathcal{I}_{\{e\}}^{V \setminus \{e\}}) < \text{rank}(\mathcal{I})$.

Lemma 3 Consider any three sets $S \in \mathcal{I}$, $S' \in \mathcal{I}$, and $R \subseteq V$ such that $S' \subseteq S$ and $S \cap R = \emptyset$, we have $\mathcal{I}_S^R \subseteq \mathcal{I}_{S'}^R$.

Items and States

We consider a set V of n items, where each item is in a particular state, which is unknown initially, from O . Each item $e \in V$ has a random state $\Phi(e) \in O$. Let $\phi(e) \in O$ denote a realization of $\Phi(e)$. Thus, a *realization* ϕ is a mapping function that maps items to states: $\phi : V \rightarrow O$. In the example of experimental design, the item e may represent a test, such as the heart rate, and $\Phi(e)$ is the outcome of the test, such as, 70 per minute. There is a known prior probability distribution $p(\phi) = \{\Pr[\Phi = \phi] : \phi \in O^V\}$ over realizations. Note that when realizations are independent, the distribution p completely factorizes. However, this independent assumption may not hold in many real-world applications such as experimental design and active learning. We assume that one must select an item $e \in V$ before observing the value of $\Phi(e)$. For any subset of items $S \subseteq V$, we use $\psi : S \rightarrow O$ to denote a *partial realization* and $\text{dom}(\psi) = S$ is called the *domain* of ψ . A partial realization ψ is said to be consistent

with a realization ϕ , denoted $\phi \sim \psi$, if they are equal everywhere in $\text{dom}(\psi)$. A partial realization ψ is said to be a *sub-realization* of another partial realization ψ' , denoted $\psi \subseteq \psi'$, if $\text{dom}(\psi) \subseteq \text{dom}(\psi')$ and they are equal everywhere in the domain $\text{dom}(\psi)$ of ψ . Given a partial realization ψ , denote by $p(\phi \mid \psi)$ the conditional distribution over realizations conditioned on ψ : $p(\phi \mid \psi) = \Pr[\Phi = \phi \mid \Phi \sim \psi]$.

Policies

Any adaptive policy can be represented as a function π that maps a set of observations to a distribution $\mathcal{P}(V)$ of V : $\pi : 2^V \times O^V \rightarrow \mathcal{P}(V)$. It specifies which item to select next based on the past outcomes. There is a utility function $f : 2^V \times O^V \rightarrow \mathbb{R}$ from a subset of items and their states to a real number.

Definition 1 (Conditional Marginal Utility of an Item)

Given a utility function $f : 2^V \times O^V \rightarrow \mathbb{R}$, the conditional expected marginal utility of any item $e \in V$ on top of any partial realization ψ is defined as $f_{\text{avg}}(e \mid \psi) = \mathbb{E}_{\Phi}[f(\text{dom}(\psi) \cup \{e\}, \Phi) - f(\text{dom}(\psi), \Phi) \mid \Phi \sim \psi]$, where the expectation is taken over Φ with respect to $p(\phi \mid \psi)$.

Let the random variable $V(\pi, \phi)$ denote the subset of items selected by a policy π under realization ϕ . The expected utility $f_{\text{avg}}(\pi)$ of a policy π can be written as

$$f_{\text{avg}}(\pi) = \mathbb{E}_{\Phi \sim p(\phi), \Pi} f(V(\pi, \Phi), \Phi)$$

The expectation is taken over the realization and the internal randomness of the policy.

Definition 2 (Conditional Marginal Utility of a Policy)

Given a utility function $f : 2^V \times O^V \rightarrow \mathbb{R}$, the conditional expected marginal utility $f_{\text{avg}}(\pi \mid \psi)$ of a policy π on top of a partial realization ψ is $f_{\text{avg}}(\pi \mid \psi) = \mathbb{E}_{\Phi, \Pi}[f(\text{dom}(\psi) \cup V(\pi, \Phi), \Phi) - f(\text{dom}(\psi), \Phi) \mid \Phi \sim \psi]$, where the expectation is taken over realizations Φ with respect to $p(\phi \mid \psi)$ and the internal randomness of the policy.

Definition 3 (\mathcal{I}_S^R -restricted Policy) Consider two sets $S \in \mathcal{I}$ and $R \subseteq V$ such that $S \cap R = \emptyset$. A policy π is \mathcal{I}_S^R -restricted if $\forall \phi : V(\pi, \phi) \in \mathcal{I}_S^R$. Let $\Omega(\mathcal{I}_S^R)$ denote the set of all \mathcal{I}_S^R -restricted policies, i.e., $\Omega(\mathcal{I}_S^R) = \{\pi \mid \forall \phi : V(\pi, \phi) \in \mathcal{I}_S^R\}$.

Definition 4 (Optimal \mathcal{I}_S^R -restricted Policy on top of ψ) Consider two sets $S \in \mathcal{I}$ and $R \subseteq V$ such that $S \cap R = \emptyset$, and a partial realization ψ . Define $\pi^*(\mathcal{I}_S^R, \psi)$ as the optimal \mathcal{I}_S^R -restricted policy on top of ψ , i.e.,

$$\pi^*(\mathcal{I}_S^R, \psi) \in \arg \max_{\pi \in \Omega(\mathcal{I}_S^R)} f_{\text{avg}}(\pi \mid \psi)$$

With the above notation, $\pi^*(\mathcal{I}_\emptyset^V, \emptyset)$ represents the optimal policy over an independence system (V, \mathcal{I}) , i.e., $\pi^*(\mathcal{I}_\emptyset^V, \emptyset) \in \arg \max_{\pi \in \Omega(\mathcal{I}_\emptyset^V)} f_{\text{avg}}(\pi \mid \emptyset)$ or equivalently, $\pi^*(\mathcal{I}_\emptyset^V, \emptyset) \in \arg \max_{\pi \in \Omega(\mathcal{I}_\emptyset^V)} f_{\text{avg}}(\pi)$.

Policywise Submodularity and Sampling Gap

In this paper, we propose a new class of stochastic utility functions, *policywise submodular functions*. We define policywise submodularity as the diminishing return property about the expected marginal gain of the optimal policy over independence systems.

Definition 5 (Policywise Submodularity) A function $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to a prior $p(\phi)$ and an independence system (V, \mathcal{I}) if for any two partial realizations ψ^a and ψ^b such that $\psi^a \subseteq \psi^b$ and $\text{dom}(\psi^b) \in \mathcal{I}$, and any $R \subseteq V$ such that $R \cap \text{dom}(\psi^b) = \emptyset$, we have

$$f_{\text{avg}}(\pi^*(\mathcal{I}_{\text{dom}(\psi^b)}^R, \psi^a) \mid \psi^a) \geq f_{\text{avg}}(\pi^*(\mathcal{I}_{\text{dom}(\psi^b)}^R, \psi^b) \mid \psi^b)$$

Later we show that the above property can be found in many real-world applications. For example, a variety of objective functions, including generalized binary search [Golovin and Krause 2011], EC² [Golovin, Krause, and Ray 2010], ALuMA [Gonen, Sabato, and Shalev-Shwartz 2013], and the maximum Gibbs error criterion [Cuong et al. 2013], used in active learning are policywise submodular. This property can also be found in other applications wherever the states of the items are independent [Asadpour and Nazarzadeh 2016]. Moreover, we prove that the utility function of the adaptive viral marketing [Golovin and Krause 2011] is also policywise submodular.

Next we introduce the concept of *sampling gap*, which is defined as the ratio of the optimal solution obtained from the full data set and the optimal solution obtained from the sampled data set, over independence systems. The sampling gap measures the performance loss of the optimal solution due to the probability sampling. Intuitively, for a given sampling rate r , a smaller sampling gap indicates less performance loss due to probability sampling. In the rest of this paper, for any $R \subseteq V$, let π_R^* denote $\pi^*(\mathcal{I}_\emptyset^R, \emptyset)$ for short, i.e., π_R^* represents the optimal policy that selects items only from R . Hence, π_V^* represents the optimal policy using the full data set.

Definition 6 (Sampling Gap) Let T be a random subset of V where each item is included in T with probability $r \in [0, 1]$ (not necessarily independently). Define the sampling gap at sampling rate r as the largest (worst-case instance of $(f, p(\phi), V, \mathcal{I})$) ratio of the optimal policy when using the full ground set V and optimal policies when using the sampled set T , i.e.,

$$\text{sampling gap} = \max_{(f, p(\phi), V, \mathcal{I})} \frac{f_{\text{avg}}(\pi_V^*)}{\mathbb{E}_T[f_{\text{avg}}(\pi_T^*)]}$$

In this paper, we restrict our attention to the case of maximizing a policywise submodular function over independence systems. I.e., our goal is to provide an answer to the following question:

What is the value of $\max_{(f, p(\phi), V, \mathcal{I})} \frac{f_{\text{avg}}(\pi_V^*)}{\mathbb{E}_T[f_{\text{avg}}(\pi_T^*)]}$ given that $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and (V, \mathcal{I}) ?

Sampling Gaps for a Policywise Submodular Function

In this section we provide our main result, the optimal sampling gap for policywise submodular functions over independence systems. We prove the upper bound and lower bound of Theorem 1 in the following two subsections respectively. All missing proofs are deferred to the full version [Tang and Yuan 2021].

Theorem 1 *The sampling gap at sampling rate r for maximizing a policywise submodular function over independence systems is exactly $1/r$.*

Upper Bound of $1/r$

We first present an upper bound of the sampling gap over independence systems. Note that for an arbitrary partial realization ψ such that $\text{dom}(\psi) \in \mathcal{I}$, Lemma 1 implies that $(V, \mathcal{I}_{\text{dom}(\psi)}^{V \setminus \text{dom}(\psi)})$ is an independence system. Before presenting the main theorem, we first prove a useful technical lemma. We introduce a new function $f(\cdot|\psi) : 2^V \times O^V \rightarrow \mathbb{R}$ for any partial realization ψ such that $f(S, \phi|\psi) = f(S \cup \text{dom}(\psi), \phi) - f(\text{dom}(\psi), \phi)$ for any $S \subseteq V$ and realization ϕ such that $\phi \sim \psi$.

Lemma 4 *Suppose $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to a prior $p(\phi)$ and an independence system (V, \mathcal{I}) . For an arbitrary partial realization ψ such that $\text{dom}(\psi) \in \mathcal{I}$, $f(\cdot|\psi) : 2^V \times O^V \rightarrow \mathbb{R}$ is also policywise submodular with respect to a prior $p(\phi|\psi)$ and an independence system $(V, \mathcal{I}_{\text{dom}(\psi)}^{V \setminus \text{dom}(\psi)})$.*

Now we are ready to present the main theorem of this paper. For convenience, let $f(\emptyset)$ denote the expected utility of an empty set for short, i.e., $f(\emptyset) = \mathbb{E}_{\phi \sim p(\phi)}[f(\emptyset, \Phi)]$.

Theorem 2 *Let T be a random subset of V where each item is included in T with probability $r \in [0, 1]$. Suppose $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to a prior $p(\phi)$ and an independence system (V, \mathcal{I}) ,*

$$\mathbb{E}_T[f_{\text{avg}}(\pi_T^*)] \geq (1-r)f(\emptyset) + rf_{\text{avg}}(\pi_V^*)$$

Proof: We prove this lemma through the induction on the upper rank $\text{rank}(\mathcal{I})$ of the independence system (V, \mathcal{I}) .

For the base case when $\text{rank}(\mathcal{I}) = 0$, we have $f_{\text{avg}}(\pi_T^*) = f(\emptyset)$ for any T and $f_{\text{avg}}(\pi_V^*) = f(\emptyset)$. Hence, $\mathbb{E}_T[f_{\text{avg}}(\pi_T^*)] = f(\emptyset) \geq (1-r)f(\emptyset) + rf(\emptyset) = (1-r)f(\emptyset) + rf_{\text{avg}}(\pi_V^*)$.

Assume the statement holds for all independence systems (V, \mathcal{I}) such that $\text{rank}(\mathcal{I}) \leq l-1$, we next prove that it holds for all independence systems (V, \mathcal{I}) when $\text{rank}(\mathcal{I}) = l$. Assume $s \in V$ is the root of the decision tree of π_V^* , i.e., s is the first item selected by π_V^* , we next construct a policy π_T such that

- If $s \in T$, π_T first selects s , then adopts $\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))$ which is the optimal $\mathcal{I}_{\{s\}}^{T \setminus \{s\}}$ -restricted policy on top of $\Phi(s)$.
- If $s \notin T$, π_T adopts $\pi^*(\mathcal{I}_{\{s\}}^T, \emptyset)$ which is the optimal $\mathcal{I}_{\{s\}}^T$ -restricted policy on top of \emptyset .

We first show that π_T is a feasible \mathcal{I}_\emptyset^T -restricted policy.

Lemma 5 *π_T is a feasible \mathcal{I}_\emptyset^T -restricted policy.*

Because π_T^* is an optimal \mathcal{I}_\emptyset^T -restricted policy and π_T is a feasible \mathcal{I}_\emptyset^T -restricted policy (Lemma 5), we have $f_{\text{avg}}(\pi_T^*) \geq f_{\text{avg}}(\pi_T)$ for any T . Hence, $\mathbb{E}_T[f_{\text{avg}}(\pi_T)] \leq \mathbb{E}_T[f_{\text{avg}}(\pi_T^*)]$. To prove this theorem, it suffices to show that

$$\mathbb{E}_T[f_{\text{avg}}(\pi_T)] \geq (1-r)f(\emptyset) + rf_{\text{avg}}(\pi_V^*) \quad (1)$$

Hence, we next focus on proving (1). We first compute the expected utility of π_T for a given T . In the case of $s \in T$, we have

$$\begin{aligned} f_{\text{avg}}(\pi_T) &= f(\emptyset) + f_{\text{avg}}(s|\emptyset) \\ &\quad + \mathbb{E}_{\Phi(s)}[f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s))] \end{aligned} \quad (2)$$

In the case of $s \notin T$, we have

$$\begin{aligned} f_{\text{avg}}(\pi_T) &= f(\emptyset) + f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^T, \emptyset)|\emptyset) \\ &\geq f(\emptyset) + \mathbb{E}_{\Phi(s)}[f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^T, \Phi(s))|\Phi(s))] \end{aligned} \quad (3)$$

The inequality is due to $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and (V, \mathcal{I}) . Taking the expectation over T , we next bound the expected utility of π_T .

$$\begin{aligned} \mathbb{E}_T[f_{\text{avg}}(\pi_T)] &= \\ &r \left(f(\emptyset) + f_{\text{avg}}(s|\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s)) | s \in T] \right) \\ &\quad + (1-r) \left(f(\emptyset) + \mathbb{E}_T [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^T, \emptyset)|\emptyset) | s \notin T] \right) \\ &\geq r \left(f(\emptyset) + f_{\text{avg}}(s|\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s)) | s \in T] \right) \\ &\quad + (1-r) \left(f(\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^T, \Phi(s))|\Phi(s)) | s \notin T] \right) \\ &= r \left(f(\emptyset) + f_{\text{avg}}(s|\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s))] \right) \\ &\quad + (1-r) \left(f(\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s))] \right) \\ &= (1-r)f(\emptyset) + r \left(f(\emptyset) + f_{\text{avg}}(s|\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s), T} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s))|\Phi(s))] \right) \\ &\geq (1-r)f(\emptyset) + r \left(f(\emptyset) + f_{\text{avg}}(s|\emptyset) \right. \\ &\quad \left. + \mathbb{E}_{\Phi(s)} [f_{\text{avg}}(\pi^*(\mathcal{I}_{\{s\}}^{V \setminus \{s\}}, \Phi(s))|\Phi(s))] \right) \\ &= (1-r)f(\emptyset) + rf_{\text{avg}}(\pi_V^*) \end{aligned} \quad (5)$$

The first inequality is due to (2) and (4). The second inequality is due to the observation that for any $\Phi(s)$,

$$\begin{aligned} \mathbb{E}_T[f_{avg}(\pi^*(\mathcal{I}_{\{s\}}^{T \setminus \{s\}}, \Phi(s)) | \Phi(s))] \\ \geq r f_{avg}(\pi^*(\mathcal{I}_{\{s\}}^{V \setminus \{s\}}, \Phi(s)) | \Phi(s)) \end{aligned} \quad (6)$$

(6) follows from the inductive assumption based on the following three facts: (1) $\text{rank}(\mathcal{I}_{\{s\}}^{V \setminus \{s\}}) < \text{rank}(\mathcal{I}) = l$ (Lemma 2), (2) $f(\cdot | \Phi(s))$ is policywise submodular with respect to a prior $p(\phi | \Phi(s))$ and an independence system $(V, \mathcal{I}_{\{s\}}^{V \setminus \{s\}})$ for any $\Phi(s)$ (Lemma 4), and (3) $f_{avg}(\emptyset | \Phi(s)) = 0$ for any $\Phi(s)$. \square

Assume $f(\emptyset) \geq 0$. Theorem 2 implies the following two corollaries.

Corollary 1 *Given that $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and (V, \mathcal{I}) , then the sampling gap at sampling rate r is upper bounded by $1/r$, i.e., $\max_{(f, p(\phi), V, \mathcal{I})} \frac{f_{avg}(\pi_V^*)}{\mathbb{E}_T[f_{avg}(\pi_T^*)]} \leq \frac{1}{r}$.*

The following corollary shows that if we can find an approximation solution over a sampled data set, then this solution achieves a bounded approximation ratio for the original problem when using the full data set.

Corollary 2 *If $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to a prior $p(\phi)$ and an independence system (V, \mathcal{I}) , and there exists an α -approximation \mathcal{I}_\emptyset^T -restricted policy π_T^α for every T , i.e., $\forall T, \exists \pi_T^\alpha \in \Omega(\mathcal{I}_\emptyset^T), f_{avg}(\pi_T^\alpha) \geq \alpha f_{avg}(\pi_T^*)$, then $\frac{f_{avg}(\pi_V^*)}{\mathbb{E}_T[f_{avg}(\pi_T^*)]} \leq \frac{1}{\alpha r}$.*

Lower Bound of $1/r$

In this section we construct a policywise submodular function and a simple cardinality constraint to show that the sampling gap is lower bounded by $1/r$. Assume $V = \{e\}$, i.e., the ground set contains only one item, and $O = \{o\}$, i.e., there is only one state. Define $f(\emptyset, (e, o)) = 0$ and $f(\{e\}, (e, o)) = 1$. Moreover, $\mathcal{I} = \{\{e\}, \emptyset\}$. We first show that $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and \mathcal{I} . Because $\mathcal{I}_{\{e\}}^\emptyset = \{\emptyset\}$, we have $f_{avg}(\pi^*(\mathcal{I}_{\{e\}}^\emptyset, \emptyset) | \emptyset) = f_{avg}(\pi^*(\mathcal{I}_{\{e\}}^\emptyset, \{o\}) | (e, o)) = 0$, which implies that $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and \mathcal{I} . We next show that the sampling gap at sampling rate r is at least $1/r$. Let T be a random set where e appears with probability r . In the case of $T = \{e\}$, π_T^* selects e , thus, $\mathbb{E}_T[f_{avg}(\pi_T^*) | e \in T] = 1$. In the case of $T = \{\emptyset\}$, π_T^* selects \emptyset , thus, $\mathbb{E}_T[f_{avg}(\pi_T^*) | e \notin T] = 0$. Thus, $\mathbb{E}_T[f_{avg}(\pi_T^*)] = r \mathbb{E}_T[f_{avg}(\pi_T^*) | e \in T] + (1 - r) \mathbb{E}_T[f_{avg}(\pi_T^*) | e \notin T] = r$. Moreover, π_V^* always selects e , thus, $f_{avg}(\pi_V^*) = 1$. Hence, the sampling gap at sampling rate r is lower bounded by $1/r$.

Applications

We next show that the property of policywise submodularity can be found in a wide range of real-world applications. We start with introducing two well-studied classes of stochastic functions. Then we build a relation between these two classes of functions and the property of policywise submodularity.

Definition 7 [Fujii and Kashima 2016][Policy-adaptive Submodularity] *A function $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policy-adaptive submodular with respect to $p(\phi)$, if for any two partial realizations ψ^a and ψ^b such that $\psi^a \subseteq \psi^b$, and any policy π such that $\forall \Phi, V(\pi, \Phi) \subseteq V \setminus \text{dom}(\psi^b)$, we have $f_{avg}(\pi | \psi^a) \geq f_{avg}(\pi | \psi^b)$.*

Definition 8 [Golovin and Krause 2011][Adaptive Submodularity] *A function $f : 2^V \times O^V \rightarrow \mathbb{R}$ is adaptive submodular with respect to $p(\phi)$, if for any two partial realizations ψ^a and ψ^b such that $\psi^a \subseteq \psi^b$, and any item $e \notin \text{dom}(\psi^b)$, we have $f_{avg}(e | \psi^a) \geq f_{avg}(e | \psi^b)$.*

Our next lemma shows that policy-adaptive submodularity implies both adaptive submodularity and policywise submodularity. Later, we show that policy-adaptive submodularity is a strictly stronger condition than policywise submodularity.

Lemma 6 *If $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policy-adaptive submodular with respect to $p(\phi)$, then $f : 2^V \times O^V \rightarrow \mathbb{R}$ is both adaptive submodular with respect to $p(\phi)$ and policywise submodular with respect to $p(\phi)$ and any independence system (V, \mathcal{I}) .*

Thanks to the recent progress in adaptive submodular maximization [Golovin and Krause 2011; Tang 2021], there exists efficient solutions for maximizing an adaptive submodular function subject to many practical constraints, including matroid and knapsack constraints. Hence, Lemma 6, together with Corollary 2, implies that if a function $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policy-adaptive submodular with respect to $p(\phi)$, then running existing algorithms on the sampled ground set has comparable performance to running them on the full set.

We next discuss three representative applications whose objective function satisfies the policywise submodularity. The objective functions in the first two applications are policy-adaptive submodular, which implies both adaptive submodularity and policywise submodularity. Although the objective function in the third application does not satisfy the policy-adaptive submodularity, it is still adaptive submodular and policywise submodular. This implies that policy-adaptive submodularity is a strictly stronger condition than policywise submodularity.

Application 1: Pool-based Active Learning [Golovin and Krause 2011]. We use \mathcal{H} to denote the set of candidate hypothesis. Each hypothesis $h \in \mathcal{H}$ represents some realization, i.e., $h : V \rightarrow \Phi$, where each item $e \in V$ can be viewed as a data point and the state $\Phi(e)$ of a data point e can be viewed as the label of e . Let p_H be a prior distribution over hypotheses. Define $p_H(\mathcal{H}') = \sum_{h \in \mathcal{H}'} p_H(h)$ for any $\mathcal{H}' \subseteq \mathcal{H}$. Then the prior distribution over realizations can be represented as $p(\phi) = p_H(h | \phi \sim h)$. The version space under observations ψ is defined to be $\mathcal{H}(\psi) = \{h \in \mathcal{H} | h \sim \psi\}$, i.e., $\mathcal{H}(\psi)$ contains all hypothesis whose labels are consistent with ψ in the domain of ψ . Given a realization ϕ and a group of labeled data points S , the utility function of generalized binary search under the Bayesian setting is

$$f(S, \phi) = 1 - p_H(\mathcal{H}(\phi(S))) \quad (7)$$

where $\phi(S) = \{(e, \phi(e)) \mid e \in S\}$.

(7) measures the reduction in version space mass after obtaining the states (a.k.a. labels) about S . Intuitively, our objective is to select (a.k.a. query the labels of) a group of data points to maximize the expected reduction in version space mass. It has been shown that the above utility function is both adaptive submodular with respect to $p(\phi)$ (Section 9 in [Golovin and Krause 2011]) and policy-adaptive submodular with respect to $p(\phi)$ (Proposition A.1 in [Fujii and Kashima 2016]). This, together with Lemma 6, implies the following proposition.

Proposition 1 *The utility function $f : 2^V \times O^V \rightarrow \mathbb{R}$ of pool-based active learning is policywise submodular with respect to $p(\phi)$ and any independence system (V, \mathcal{I}) .*

Moreover, many other types of objective functions of active learning, including EC² [Golovin, Krause, and Ray 2010], ALuMa [Gonen, Sabato, and Shalev-Shwartz 2013], and the maximum Gibbs error criterion [Cuong et al. 2013], are both adaptive submodular and policywise submodular.

Application 2: The Case of Independent Items [Asadpour and Nazerzadeh 2016]. The property of policywise submodularity can also be found in any applications in which the states of items are independent of each other. One such application is sensor selection [Golovin and Krause 2011]. The following proposition follows from Lemma 6 and the fact that if $\Phi(1), \Phi(2), \dots, \Phi(n)$ are independent and f is adaptive submodular with respect to $p(\phi)$, then f is policy-adaptive submodular with respect to $p(\phi)$ (Proposition A.5 in [Fujii and Kashima 2016]).

Proposition 2 *If $\Phi(1), \Phi(2), \dots, \Phi(n)$ are independent and $f : 2^V \times O^V \rightarrow \mathbb{R}$ is adaptive submodular with respect to $p(\phi)$, then $f : 2^V \times O^V \rightarrow \mathbb{R}$ is policywise submodular with respect to $p(\phi)$ and any independence system (V, \mathcal{I}) .*

Application 3: Adaptive Viral Marketing [Golovin and Krause 2011]. The third application is the adaptive variant of viral marketing [Golovin and Krause 2011]. We use a directed graph $G = (V, E)$ to represent a social network, where V represents a set of individuals and E represents a set of edges. Under the Independent Cascade Model [Kempe and Mahdian 2008], each edge $(u, v) \in E$ is associated with a propagation probability $p_{uv} \in [0, 1]$. In step 0, we activate a set of *seeds*. Then, in each subsequent step t , each individual u , that is newly activated, has a single chance to activate each of its neighbors v ; it succeeds with a probability p_{uv} . If u succeeds, then v becomes activated in step $t + 1$. This process is iterated till no more individuals are newly activated. Under the adaptive setting [Golovin and Krause 2011], we model the state $\phi(u)$ of u as a function $\phi(u) : E \rightarrow \{0, 1, ?\}$, where $\phi(u)((u, v)) = 0$ means that selecting u reveals that (u, v) is blocked (i.e., u fails to activate v), $\phi(u)((u, v)) = 1$ means that selecting u reveals that (u, v) is live (i.e., u succeeds in activating v), and $\phi(u)((u, v)) = ?$ means that selecting u can not reveal the status of (u, v) . We assume that selecting a seed u can reveal the status (live or block) of every out-going edge of every individual that can be reached by u through a path composed of live edges. For a given set of seeds S and a realization

ϕ , we define the utility $f(S, \phi)$ as the number of individuals that can be reached by at least one seed from S through live edges (including S), i.e.,

$$f(S, \phi) = |\{v \mid \exists u \in S, w \in V, \phi(u)((w, v)) = 1\}| + |S| \quad (8)$$

It has been shown that the above utility function is adaptive submodular (Section 8 in [Golovin and Krause 2011]). We next show that this function is also policywise submodular, however, it is not policy-adaptive submodular.

Proposition 3 *The utility function $f : 2^V \times O^V \rightarrow \mathbb{R}$ of adaptive viral marketing is policywise submodular with respect to $p(\phi)$ and any independence system (V, \mathcal{I}) .*

Lemma 7 *Policy-adaptive submodularity is a strictly stronger condition than policywise submodularity.*

Performance Evaluation

We conduct experiments to evaluate the impact of probabilistic sampling based on two popular machine learning applications: adaptive viral marketing and pool-based active learning [Golovin and Krause 2011].

Adaptive Viral Marketing. A detailed description of this application can be found in Application 3 from the previous section. We capture the social network by a directed weighted graph and run experiments on four large-scale benchmark social networks: *Wikivote*, *NetHEPT*, *NetPHY* and *Epinions* (<http://snap.stanford.edu/data/>). *Wikivote* records 103,663 votes from 7,066 users participating in the elections from the Wikipedia community. *NetHEPT* is a large academic collaboration network extracted from the High Energy Physics Theory section of arXiv, including 15,233 nodes each representing an author and 62,774 edges each representing one paper co-authored by two nodes. *NetPHY* is another academic collaboration network extracted from the Physics section of arXiv, which contains 37,154 nodes and 231,584 edges. *Epinion* is a Who-trust-whom network of Epinions.com, containing 75,879 nodes and 508,837 edges. Each node represents a user and each edge represents a trust relationship. We adopt the Independent Cascade model as diffusion model and assign a uniform probability of 0.01 to each edge as discussed in [Kempe, Kleinberg, and Tardos 2003].

Pool-based Active Learning. A detailed description of this application can be found in Application 1 from the previous section. We consider 1,000 hypotheses, 80 unlabeled data points and 50 queries each covers one or two unlabeled data points. The probability of each hypothesis is drawn from $(0, 1)$ uniformly at random with normalization; each data point is assigned a value randomly selected from its possible set of labels.

Algorithms and Parameters. For the adaptive viral marketing application, we evaluate the performance of the following three algorithms under different sampling rates: adaptive greedy algorithm (AG), non-adaptive greedy algorithm (NG) and random algorithm (RDM). AG adaptively selects a node that introduces the largest expected marginal influence spread in each round based on prior observations until

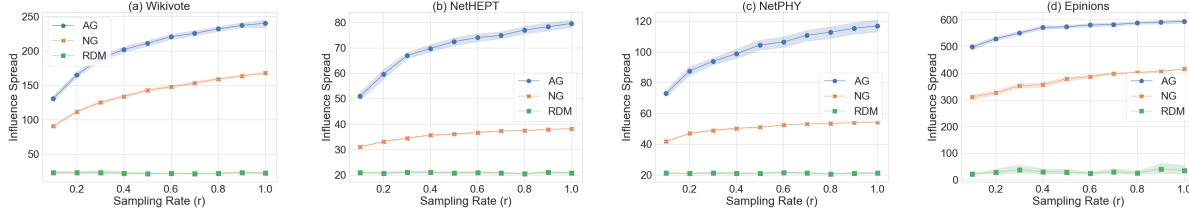


Figure 1: Adaptive Viral Marketing: Influence Spread vs. Sampling Rate.

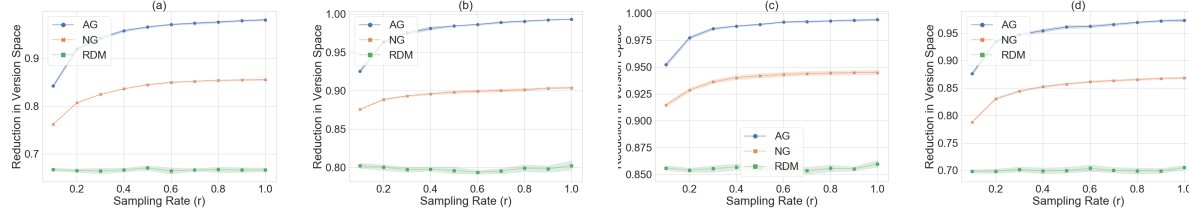


Figure 2: Pool-based Active Learning: Reduction in Version Space vs. Sampling Rate

k nodes are selected. Here prior observations refer to the realized influence propagation in the network triggered by the previously selected seeds. Although AG is not necessarily optimal, it can approximate the optimal policy within a factor of $1 - 1/e$ [Golovin and Krause 2011], i.e., the expected size of the influence generated by AG is at least $1 - 1/e$ times the optimal solution. Therefore, through examining the performance of AG under varies sampling rates, it can help us understand the impact of probability sampling on the optimal solution. The second algorithm NG is a non-adaptive version of AG. It adds the node that maximally increases the objective value before any observations take place. We run simulations 10,000 times and took the average in order to obtain reasonable estimates of the influence spread. For RDM, we randomly select a set of seed nodes of size k as an output.

We sample each node independently with sampling rate r . We vary r from 0.1 to 1. When $r = 1$, algorithms run on a full data set that includes all nodes in the network. A lower r indicates a smaller (in expectation) candidate set from which we select our seed nodes. For example, when $r = 0.5$, we obtain a sample by adding each node in the network with probability 0.5. For each sampling rate, we obtain 30 samples, and report the average performance of each algorithm over these samples, denoted as the point on the line, as well as the confidence interval of the result, denoted as the bar around the line.

For the pool-based active learning application, we evaluate the performance of AG, NG and RDM under various sampling rates. AG iteratively identifies a query with the highest marginal version space reduction based on prior observations until 5 queries have been selected. Here prior observations refer to the user’s answer to the query. NG is a non-adaptive version of AG. RDM randomly selects a set of 5 queries as an output. We use a similar sampling setting as above in our experiments.

Experimental results. We present the results in Figure 1 and Figure 2. We observe that the performance trends of the algorithms are overall consistent between all datasets and application domains. For the adaptive viral marketing, we evaluate the performance of algorithms as measured by the influence spread with respect to the changes in the sampling rate on four datasets. We set $k = 10$ for Epinions, and $k = 20$ for other three datasets. We observe that among three algorithms examined, AG is the best in all settings. Surprisingly, on three larger datasets *NetHEPT*, *NetPHY* and *Epinions*, AG with $r = 0.1$ outperforms NG with $r = 1$, which demonstrates the power of adaptivity. As expected, RDM has the poorest performance across all settings. We also notice that when the sampling rate is larger than 0.6, the performance loss (as compared with when $r = 1$) of AG is within 10%. Therefore, we can choose $r = 0.6$ to reduce the computational cost significantly without much sacrifice in performance.

For the active learning application, we evaluate the performance of algorithms as measured by the yielded reduction in version space with respect to the changes in the sampling rate. The results are plotted in Figure 2. We first consider the scenario where each data point has the same number of possible labels. We set the number of labels to 2, 3, 4 and the results are shown in Figure 2(a), (b) and (c) respectively. To diversity the data points in our data set, we then randomly divide our 80 unlabeled data points into three groups. The first group contains 64 data points with binary labels. The second and third group each contains 8 data points with three and four possible labels respectively. The results are shown in Figure 2(d). AG is the best under all settings, and its performance tends to stabilize after r is greater than 0.6. In particular, the performance loss caused by probability sampling is within 10% for $r \geq 0.3$ and is within 5% for $r \geq 0.5$. This verifies the value of probability sampling in the context of pool-based active learning.

References

- Alaei, S.; Makhdoumi, A.; and Malekian, A. 2021. Maximizing sequence-submodular functions and its application to online advertising. *Management Science*.
- Asadpour, A.; and Nazerzadeh, H. 2016. Maximizing stochastic monotone submodular functions. *Management Science*, 62(8): 2374–2391.
- Badanidiyuru, A.; and Vondrák, J. 2014. Fast algorithms for maximizing submodular functions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, 1497–1514. SIAM.
- Balkanski, E.; Rubinstein, A.; and Singer, Y. 2016. The Power of Optimization from Samples. In *NIPS*, 4017–4025.
- Calinescu, G.; Chekuri, C.; Pál, M.; and Vondrák, J. 2007. Maximizing a submodular set function subject to a matroid constraint. In *International Conference on Integer Programming and Combinatorial Optimization*, 182–196. Springer.
- Chen, Y.; and Krause, A. 2013. Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization. *ICML (1)*, 28(160-168): 8–1.
- Cuong, N. V.; Lee, W. S.; Ye, N.; Chai, K. M.; and Chieu, H. L. 2013. Active learning for probabilistic hypotheses using the maximum Gibbs error criterion. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 1457–1465.
- Ene, A.; and Nguyen, H. L. 2018. Towards nearly-linear time algorithms for submodular maximization with a matroid constraint. *arXiv preprint arXiv:1811.07464*.
- Feldman, M.; Harshaw, C.; and Karbasi, A. 2017. Greed is good: Near-optimal submodular maximization via greedy optimization. In *Conference on Learning Theory*, 758–784. PMLR.
- Fujii, K.; and Kashima, H. 2016. Budgeted stream-based active learning via adaptive submodular maximization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 514–522.
- Fujii, K.; and Sakaue, S. 2019. Beyond Adaptive Submodularity: Approximation Guarantees of Greedy Policy with Adaptive Submodularity Ratio. In *International Conference on Machine Learning*, 2042–2051.
- Gabillon, V.; Kveton, B.; Wen, Z.; Eriksson, B.; and Muthukrishnan, S. 2013. Adaptive submodular maximization in bandit setting. In *Advances in Neural Information Processing Systems*, 2697–2705.
- Golovin, D.; and Krause, A. 2011. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42: 427–486.
- Golovin, D.; Krause, A.; and Ray, D. 2010. Near-Optimal Bayesian Active Learning with Noisy Observations. In *NIPS*.
- Gonen, A.; Sabato, S.; and Shalev-Shwartz, S. 2013. Efficient active learning of halfspaces: an aggressive approach. In *International Conference on Machine Learning*, 480–488. PMLR.
- Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.
- Kempe, D.; and Mahdian, M. 2008. A cascade model for externalities in sponsored search. In *International Workshop on Internet and Network Economics*, 585–596. Springer.
- Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429.
- Mirzasoleiman, B.; Badanidiyuru, A.; and Karbasi, A. 2016. Fast Constrained Submodular Maximization: Personalized Data Summarization. In *ICML*, 1358–1367.
- Mirzasoleiman, B.; Badanidiyuru, A.; Karbasi, A.; Vondrák, J.; and Krause, A. 2015. Lazier than lazy greedy. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions-I. *Mathematical programming*, 14(1): 265–294.
- Sviridenko, M. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1): 41–43.
- Tang, S. 2020. Price of dependence: stochastic submodular maximization with dependent items. *Journal of Combinatorial Optimization*, 39(2): 305–314.
- Tang, S. 2021. Beyond pointwise submodularity: Non-monotone adaptive submodular maximization in linear time. *Theoretical Computer Science*, 850: 249–261.
- Tang, S. 2022. Beyond Pointwise Submodularity: Non-monotone Adaptive Submodular Maximization Subject to Knapsack and k-System Constraints. In Le Thi, H. A.; Pham Dinh, T.; and Le, H. M., eds., *Modelling, Computation and Optimization in Information Systems and Management Sciences*, 16–27. Cham: Springer International Publishing. ISBN 978-3-030-92666-3.
- Tang, S.; and Yuan, J. 2020. Influence maximization with partial feedback. *Operations Research Letters*, 48(1): 24–28.
- Tang, S.; and Yuan, J. 2021. Optimal Sampling Gaps for Adaptive Submodular Maximization. *arXiv preprint arXiv:2104.01750*.
- Yuan, J.; and Tang, S.-J. 2017. Adaptive discount allocation in social networks. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 1–10.