Optimization from Structured Samples for Coverage Functions

Wei Chen [1] **Xiaoming Sun** [2 3] **Jialin Zhang** [2 3] **Zhijie Zhang** [2 3]

Abstract

We revisit the *optimization from samples* (OPS) model, which studies the problem of optimizing objective functions directly from the sample data. Previous results showed that we cannot obtain a constant approximation ratio for the maximum coverage problem using polynomially many independent samples of the form $\{S_i, f(S_i)\}_{i=1}^t$ (Balkanski et al., 2017), even if coverage functions are $(1-\epsilon)$-PMAC learnable using these samples (Badanidiyuru et al., 2012), which means most of the function values can be approximately learned very well with high probability. In this work, to circumvent the impossibility result of OPS, we propose a stronger model called *optimization from structured samples* (OPSS) for coverage functions, where the data samples encode the structural information of the functions. We show that under three general assumptions on the sample distributions, we can design efficient OPSS algorithms that achieve a constant approximation for the maximum coverage problem. We further prove a constant lower bound under these assumptions, which is tight when not considering computational efficiency. Moreover, we also show that if we remove any one of the three assumptions, OPSS for the maximum coverage problem has no constant approximation.

1. Introduction

Traditional optimization problems in the textbook are often formulated as mathematical models with specified parameters. The computational task is to optimize an objective function given parameters of the model. One such example is the maximum coverage problem. Given a family of subsets $T_1, T_2, \cdots, T_n$ of a ground set N and a positive integer k, the problem asks to find k subsets whose union contains the most number of elements in N. In practice, however, parameters of the model are often hidden in the complex real world and we cannot observe them directly. Instead, we can only learn information about the model from the passively observed sample data. Back to the maximum coverage problem, in this case we may not know the exact elements contained in every subset T_i, but only observe samples of subsets T_i's, and for each sample we only observe the number of elements it covers. An immediate question, recently raised by Balkanski et al. (2017), asks to what extent we can optimize objective functions based on the sample data that we use to learn them. More specifically, given samples $\{S_i, f(S_i)\}_{i=1}^t$ where S_i's are drawn i.i.d. from some distribution $\mathcal{D}$ on the subsets of N, $f : 2^N \to \mathbb{R}$ is an unknown objective function, and $t \in \text{poly}(|N|)$, can we solve $\max_{|S|\le k} f(S)$? For maximum coverage, S_i would be a collection of some subsets T_i's, function f would be the number of elements covered by such collections. Such problems form a new approach to optimization called *optimization from samples* (OPS) (Balkanski et al., 2017).

A reasonable and perhaps the most natural approach is to first learn a surrogate function $\tilde{f} : 2^N \to \mathbb{R}$ which approximates well the original function f and then optimize $\tilde{f}$ instead of f. One may expect that if we can approximate a function well, then we can also optimize it well. Standard frameworks of learnability in the literature include PAC learnability for boolean functions due to Valiant (1984) and PMAC learnability for real-valued set functions due to Balcan and Harvey (2011).

Unfortunately, the learning-and-then-optimization approach does not work in general. Indeed, Balkanski et al. (2017) show the striking result that the maximum coverage problem cannot be approximated within a ratio better than $2^{-\Omega(\sqrt{\log |N|})}$ using only polynomially many samples drawn i.i.d. from *any* distribution, even though (a) for any constant $\epsilon > 0$, coverage functions are $(1-\epsilon)$-PMAC learnable over *any* distribution (Badanidiyuru et al., 2012), which means most of the function values can be approximately learned very well with high probability; and (b) maximum

[1]Microsoft Research Asia, Beijing, China. [2]CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [3]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China.. Correspondence to: Wei Chen <weic@microsoft.com>, Xiaoming Sun, Jialin Zhang, Zhijie Zhang <{sunxiaoming, zhangjialin, zhangzhijie}@ict.ac.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

coverage problem as a special case of submodular function maximization has a $1 - 1/e$ approximation given a value oracle to the coverage function (Nemhauser et al., 1978).

The impossibility result by Balkanski et al. (2017) uses coverage functions defined over a partition of the ground set, which ensure the "good" and "bad" parts of the partition cannot be distinguished from the samples. In other words, the impossibility result arises because the samples do not provide information on the structure of coverage functions.

To circumvent the above impossibility result, we propose a stronger model called *optimization from structured samples* (OPSS) for coverage functions, which encodes structural information of the coverage functions into the samples. In many real-world applications, such structural information are often revealed in the data, for example, a crowd-sourcing platform records the crowd-workers' coverage on the tasks they took, a document analysis application records the keywords coverage on the documents they appear, etc. Thus the OPSS model is reasonable in practice. However, even in the stronger OPSS model, not all sample distributions will allow a constant approximation for the maximum coverage problem. In this paper, we study the assumptions that enable constant approximation in the OPSS model and its related algorithmic and hardness results. We now state our model and results in more detail.

1.1. Model

For sake of comparison, we first state the definition of optimization from samples (Balkanski et al., 2017) for general set functions.

Definition 1 (Optimization from samples (OPS))**.** *Let $\mathcal{F}$ be a class of set functions defined on the ground set L. $\mathcal{F}$ is α-optimizable from samples in constraint $\mathcal{M} \subseteq 2^L$ over distribution $\mathcal{D}$ on 2^L, if there exists a (not necessarily polynomial time) algorithm such that, given any parameter $\delta > 0$ and sufficiently large L, there exists some integer $t_0 \in poly(|L|, 1/\delta)$, for all $t \geq t_0$, for any set of samples $\{S_i, f(S_i)\}_{i=1}^t$ with $f \in \mathcal{F}$ and S_i's drawn i.i.d. from $\mathcal{D}$, the algorithm takes samples $\{S_i, f(S_i)\}_{i=1}^t$ as the input and returns $S \in \mathcal{M}$ such that*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[\mathbf{E}[f(S)] \geq \alpha \cdot \max_{T\in\mathcal{M}} f(T)] \geq 1 - \delta,$$

where the expectation is taken over the randomness of the algorithm.

Next we state the definition of coverage functions in terms of bipartite graphs as well as the definition of optimization from structured samples for coverage functions.

Definition 2 (Coverage functions)**.** *Assume there is a bipartite graph $G = (L, R, E)$. For node $u \in L \cup R$, let $N_G(u)$ denote its neighbors in G. The neighbors of a subset $S \subseteq L$ or $S \subseteq R$ is $N_G(S) = \cup_{u\in S} N_G(u)$. The coverage function $f_G : 2^L \to \mathbb{R}_+$ is the number of neighbors covered by a set $S \subseteq L$, i.e. $f_G(S) = |N_G(S)|$.*

Definition 3 (Optimization from structured samples (OPSS))**.** *Let $\mathcal{F}$ be the class of coverage functions defined on all bipartite graphs $\{G = (L, R, E)\}$ with two components L and R. $\mathcal{F}$ is α-optimizable under OPSS in constraint $\mathcal{M} \subseteq 2^L$ over distribution $\mathcal{D}$ on 2^L, if there exists a (not necessarily polynomial time) algorithm such that, given any parameter $\delta > 0$ and sufficiently large L, there exists some integer $t_0 \in poly(|L|, |R|, 1/\delta)$, for all $t \geq t_0$, for any set of samples $\{S_i, N_G(S_i)\}_{i=1}^t$ with $f_G \in \mathcal{F}$ and S_i's drawn i.i.d. from $\mathcal{D}$, the algorithm takes samples $\{S_i, N_G(S_i)\}_{i=1}^t$ as the input and returns $S \in \mathcal{M}$ such that*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[\mathbf{E}[f_G(S)] \geq \alpha \cdot \max_{T\in\mathcal{M}} f_G(T)] \geq 1 - \delta,$$

where the expectation is taken over the randomness of the algorithm.

Samples in OPSS are *structured* in that the exact members covered by a set $S \subseteq L$ are revealed, instead of only the number of covered members being revealed as in OPS. In this paper we focus on the cardinality constraint $\mathcal{M}_{\leq k} = \{S \subseteq L \mid |S| \leq k\}$. Maximizing coverage functions under this constraint is known as the *maximum coverage problem*.

Our OPSS model is defined so far only for coverage functions. One reason is that the impossibility of OPS given by Balkanski et al. (2017) is on the coverage functions, which is striking because coverage functions admit a simple constant approximation algorithm with the value oracle and is $(1 - \epsilon)$-PMAC learnable as mentioned before. Thus coverage function is the first to consider for circumventing the impossibility result for OPS. Another reason is that coverage functions exhibit natural structures via the bipartite graph representation. Other set functions may exhibit different combinatorial structures and thus the OPSS problem may need to be defined accordingly to reflect the specific structural information for other set functions.

1.2. Our Results

One of our main results is to provide a set of three general assumptions on the sample distribution together with an algorithm and show that the algorithm achieves a constant approximation ratio for the maximum coverage problem in OPSS under the assumption. The general assumption is summarized below.

Assumption 1. *We assume that the distribution $\mathcal{D}$ on 2^L satisfy the following three assumptions:*

1.1 ***Feasibility.*** *A sample $S \sim \mathcal{D}$ is always feasible, i.e. $|S| \leq k$.*

1.2 ***Polynomial bounded sample complexity.*** *For any $u \in L$, the probability $p_u = \Pr_{S\sim\mathcal{D}}[u \in S]$ satisfies $p_u \geq 1/|L|^c$ for some constant c.*

1.3 ***Negative correlation.*** *The random variables $X_u = \mathbf{1}_{u\in S}$ are "negatively correlated" (see Definition 4) over distribution $\mathcal{D}$.*

All three assumptions above are natural. In particular, Assumption 1.2 means that all elements in the ground set have sufficient probability to be sampled, and Assumption 1.3 means informally that the appearance of one element in the sampled set S would reduce the probability of the appearance of another element in S. In fact, typical distributions over $\mathcal{M}_{\leq k}$, such as uniform distribution $\mathcal{D}_{\leq k}$ over all subsets in $\mathcal{M}_{\leq k}$ or uniform distribution $\mathcal{D}_k$ over all subsets of exact size k, all satisfy these assumptions. Our result based on the above assumption is summarized by the following theorem.

Theorem 1. *If a distribution $\mathcal{D}$ satisfies Assumption 1, given any α-approximation algorithm A for the standard maximum coverage problem, coverage functions are $\frac{\alpha}{2}$-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}$ for any $k \leq |L|$. Furthermore, the OPSS algorithm uses a polynomial number of arithmetic operations and one call of algorithm A.*

The general approximation ratio α is to cover both polynomial-time and non-polynomial-time algorithms. If we need a polynomial-time algorithm, then we know that the best ratio we can achieve is $1 - 1/e$ if NP≠P (Nemhauser et al., 1978; Feige, 1998). Thus our OPSS algorithm achieves $\frac{1}{2}(1 - 1/e)$ approximation. If running time is not our concern, then we can use $\alpha = 1$ by an exhaustive search algorithm, and our OPSS algorithm achieves $\frac{1}{2}$ approximation.

We further show that if the distribution is $\mathcal{D}_k$, i.e. the uniform distribution over all subsets of exact size k, we have another OPSS algorithm to achieve $(\alpha - \epsilon)$ approximation, as shown below. This implies that our OPSS algorithm (almost) matches the approximation ratio of any algorithm for the standard maximum coverage problem.

Theorem 2. *For any constant $\epsilon > 0$, given any α-approximation algorithm A for the standard maximum coverage problem, coverage functions are $(\alpha - \epsilon)$-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}_k$, assuming that $\ln^2 |L| \leq k \leq |L|/2$ and $|R| \leq \frac{\epsilon}{2}|L|^{(\epsilon \ln |L|)/8}$. Furthermore, the OPSS algorithm uses a polynomial number of arithmetic operations and one call of algorithm A.*

Next, we prove a hardness result showing that the approximation ratio of $\frac{1}{2}$ is unavoidable for some distributions, which means that when efficiency is not the concern, our upper and lower bounds are tight.

Theorem 3. *There is a distribution $\mathcal{D}$ satisfying Assumption 1 such that coverage functions are not α-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}$ for any $\alpha > \frac{1}{2} + o(1)$.*

Finally, we also show that the three conditions given in Assumption 1 are necessary, in the sense that dropping any one of them would result in no constant approximation for the OPSS problem. This demonstrates that our three conditions need to work together to make OPSS solvable.

Theorem 4. *By dropping any one of the conditions in Assumption 1, there is a distribution $\mathcal{D}$ such that coverage functions are not α-optimizable under OPSS for any constant α in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}$.*

To summarize, in this paper we investigate the structural information on coverage functions that could allow us to circumvent the impossibility result in (Balkanski et al., 2017). We show that when the samples could reveal the covered elements rather than just the count, under certain reasonable assumptions on the sample distribution (Assumption 1), we could design an OPSS algorithm that achieves $\alpha/2$ approximation, where α is the approximation ratio of a standard maximum coverage problem. Moreover, for the uniform distribution on subsets of size k, we provide an efficient algorithm that achieves tight $\alpha - \epsilon$ approximation, matching the performance of any algorithm for the standard maximum coverage problem. On the lower bound side, we show that the approximation ratio of $1/2$ is unavoidable, which matches the upper bound when not considering computational complexity. Finally, we show that removing any one of the three conditions in Assumption 1, we cannot achieve constant approximation for OPSS. Our study opens up the possibility of studying structural information for achieving optimization from samples, which is needed in many applications in the big data era.

1.3. Related Work

The study of optimization from samples (OPS) was initiated by Balkanski et al. (2017). They proved that no algorithm can achieve an approximation ratio better than $2^{-\Omega(\sqrt{\log n})}$ for the maximum coverage problem under OPS. The same set of authors showed there is an optimal $(1-c)/(1+c-c^2)$ approximation algorithm for maximizing monotone submodular functions with curvature c subject to a cardinality constraint over uniform distributions under OPS (Balkanski et al., 2016). For submodular function minimization, it was proved in (Balkanski & Singer, 2017) that no algorithm can obtain an approximation strictly better than $2 - o(1)$ under OPS. And this is tight via a trivial 2-approximation algorithm. Rosenfeld et al. (2018) defined a weaker variant of OPS called *distributionally optimization from samples* (DOPS). They showed that a class of set functions is optimizable under DOPS if and only if it is PMAC-learnable.

2. Concepts and Tools

We first discuss the definition of negative correlation. Negative dependence among random variables has been extensively studied in the literature and there are a lot of qualitative versions of this concept (Jogdeo & Patil, 1975; Karlin & Rinott, 1980; Ghosh, 1981; Block et al., 1982; Joag-Dev & Proschan, 1983). Among them, the most widely accepted one is the *negative association* (NA) defined in (Joag-Dev & Proschan, 1983). However, in this paper, we only use a weaker version of NA. Thus, more distributions satisfy our definition of negative correlation. It is also easy to see that the uniform distributions $\mathcal{D}_k$ and $\mathcal{D}_{\leq k}$ both satisfy this definition.

Definition 4 (Negative correlation)**.** *A set of* 0-1 *random variables $X_1, \cdots, X_n$ is negative correlated, if for any disjoint subsets $I, J \subseteq [n] := \{1, \cdots, n\}$,*

$$\mathbf{E}\Big[\prod_{i\in I\cup J}(1-X_i)\Big] \leq \mathbf{E}\Big[\prod_{i\in I}(1-X_i)\Big]\mathbf{E}\Big[\prod_{j\in J}(1-X_j)\Big].$$

Then we prove the following lemma, which shows that the occurrence of an event would reduce the probability of occurrences of other events.

Lemma 1. *Assume that $X_1, \cdots, X_n$ are negatively correlated* 0-1 *random variables. Then for any $I \subseteq [n]$ and $j \notin I$,*

$$\Pr[\vee_{i\in I}(X_i = 1) \mid X_j = 1] \leq \Pr[\vee_{i\in I}(X_i = 1)].$$

Proof. Since $X_1, \cdots, X_n$ are negatively correlated,

$$\Pr[\wedge_{i\in I\cup\{j\}}(X_i = 0)] \leq \Pr[\wedge_{i\in I}(X_i = 0)]\Pr[X_j = 0],$$

which is equivalent to

$$\begin{aligned}&\Pr[\wedge_{i\in I}(X_i = 0)] - \Pr[\wedge_{i\in I}(X_i = 0), X_j = 1]\\ &\leq \Pr[\wedge_{i\in I}(X_i = 0)]\Pr[X_j = 0].\end{aligned}$$

Rearranging the last inequality, we have

$$\begin{aligned}&\Pr[\wedge_{i\in I}(X_i = 0)]\Pr[X_j = 1]\\ &\leq \Pr[\wedge_{i\in I}(X_i = 0), X_j = 1],\end{aligned}$$

which is equivalent to

$$\begin{aligned}&(1 - \Pr[\vee_{i\in I}(X_i = 1)])\Pr[X_j = 1]\\ &\leq \Pr[X_j = 1] - \Pr[\vee_{i\in I}(X_i = 1), X_j = 1].\end{aligned}$$

Rearranging the last inequality, we have

$$\begin{aligned}&\Pr[\vee_{i\in I}(X_i = 1), X_j = 1]\\ &\leq \Pr[\vee_{i\in I}(X_i = 1)]\Pr[X_j = 1].\end{aligned}$$

This concludes the proof. □

Algorithm 1 OPSS algorithm for the general Assumption 1

Input: Samples $\{S_i, N_G(S_i)\}_{i=1}^t$ and $k \in \mathbb{N}_+$

1: Let $T_1 = S_1$
2: Construct a surrogate bipartite graph $\tilde{G} = (L, R, \tilde{E})$ such that for each $u \in L$, $N_{\tilde{G}}(u) = \cap_{i:u\in S_i} N_G(S_i)$
3: Let $T_2 = A(\tilde{G}, k)$
4: **return** T_1 with probability $1/2$; and T_2 otherwise

Next is Chernoff bound used in the analysis of probability concentration.

Lemma 2 (Chernoff bound, (Mitzenmacher & Upfal, 2005))**.** *Let $X_1, X_2, \cdots, X_n$ be independent random variables in $\{0, 1\}$ with $\Pr[X_i = 1] \geq p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mathbf{E}[X] = \mu \geq \mu_L = \sum_{i=1}^n p_i$. Then, for $0 < \delta < 1$,*

$$\Pr[X \leq (1-\delta)\mu_L] \leq e^{-\mu_L\delta^2/2}.$$

3. Constant Approximations for OPSS

In this section, we present two constant approximation algorithms for OPSS and their results: one for the general distributions satisfying Assumption 1 (Theorem 1) and the other for the uniform distribution $\mathcal{D}_k$ (Theorem 2).

3.1. A Constant Approximation under Assumption 1

The algorithm is shown in Algorithm 1. It returns one of the two solutions T_1 and T_2 with equal probability, where T_1 is just the first sample, and T_2 is the solution of an α-approximation algorithm A on a constructed surrogate bipartite graph $\tilde{G}$ for the standard maximum coverage problem. The parameters of algorithm A denote the graph and the constraint respectively. The surrogate graph $\tilde{G} = (L, R, \tilde{E})$ is constructed from samples $\{S_i, N_G(S_i)\}_{i=1}^t$ such that for each node $u \in L$, we construct u's coverage in R as $N_{\tilde{G}}(u) = \cap_{i:u\in S_i} N_G(S_i)$, which is an estimate of $N_G(u)$. The intuition is as follows. If some singleton $\{u\}$ is drawn from $\mathcal{D}$, the knowledge about $N_G(u)$ is completely revealed. However, it might be the case that $\mathcal{D}$ always returns a large set S, and the exact knowledge about $N_G(u)$ for $u \in S$ is hidden behind $N_G(S)$. Thus to reveal as much knowledge about $N_G(u)$ as possible, it is natural to use the intersection of samples that contain u as an estimate.

The difficulty in the analysis is that $N_{\tilde{G}}(u)$ is always an overestimate of $N_G(u)$, and it is impossible to show that $N_{\tilde{G}}(u)$ is a good approximation of $N_G(u)$. One extreme example is that suppose for some $v \in L$, $\Pr_{S\sim\mathcal{D}}[v \in S] = 1$, then we have that $N_{\tilde{G}}(u)$ always contains all elements in $N_G(u) \cup N_G(v)$, which might be much larger than $N_G(u)$ itself. Thus T_2 itself might not be a good solution on the original graph G. To circumvent this difficulty, the key step is to show that for any $S \sim \mathcal{D}$, $N_{\tilde{G}}(T_2)\backslash N_G(T_2) \subseteq \cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u)) \subseteq N_G(S)$ with high probability

(Lemma 3). Consequently, $N_{\tilde{G}}(T_2) \subseteq N_G(T_1 \cup T_2)$ and we can obtain a constant approximation ratio by combining a random sample T_1 with T_2 as in Algorithm 1. Note that T_1 and T_2 may be correlated since they are both dependent on S_1, but this is not an issue based on our analysis.

Lemma 3. *For a given $\delta > 0$, suppose that the number of samples $t \geq \frac{4|L|^c|R|}{\delta} \ln \frac{4|L||R|}{\delta}$, where c is the constant in Assumption 1.2. Under Assumption 1, we have*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[\cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u)) \subseteq N_G(S_1)] \geq 1-\delta.$$

The proof of Lemma 3 is delayed to Section 3.1.1. For now, we use it to prove Theorem 5, which is a more concrete version of Theorem 1.

Theorem 5. *If a distribution $\mathcal{D}$ satisfies Assumption 1, given any α-approximation algorithm A for the standard maximum coverage problem, coverage functions are $\frac{\alpha}{2}$-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}$ for any $k \leq |L|$. More precisely, for any $\delta > 0$, suppose that the number of samples $t \geq \frac{4|L|^c|R|}{\delta} \ln \frac{4|L||R|}{\delta}$, where c is the constant in Assumption 1.2. Let ALG be the solution returned by Algorithm 1 and OPT be the optimal solution on the original graph G. Then under Assumption 1, we have*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}\left[\mathbf{E}[f_G(ALG)] \geq \frac{\alpha}{2} f_G(OPT)\right] \geq 1-\delta.$$

Proof. By the construction of $\tilde{G}$, $N_G(u) \subseteq N_{\tilde{G}}(u)$ for any $u \in L$. Therefore, G is a subgraph of $\tilde{G}$ and $f_{\tilde{G}}(OPT) \geq f_G(OPT)$. Since A is an α approximation algorithm,

$$f_{\tilde{G}}(T_2) \geq \alpha f_{\tilde{G}}(OPT) \geq \alpha f_G(OPT).$$

On the other hand, it holds that $N_{\tilde{G}}(T_2)\backslash N_G(T_2) \subseteq \cup_{u\in T_2}(N_{\tilde{G}}(u)\backslash N_G(u)) \subseteq \cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u))$. Since $T_1 = S_1$, by Lemma 3, it holds with probability $1-\delta$ that $N_{\tilde{G}}(T_2)\backslash N_G(T_2) \subseteq N_G(T_1)$, and

$$\begin{aligned} f_{\tilde{G}}(T_2) &= |N_G(T_2) \cup (N_{\tilde{G}}(T_2)\backslash N_G(T_2))| \\ &\leq |N_G(T_2)| + |N_{\tilde{G}}(T_2)\backslash N_G(T_2)| \\ &\leq |N_G(T_2)| + |N_G(T_1)| \\ &= f_G(T_2) + f_G(T_1). \end{aligned}$$

Therefore, with probability $1-\delta$,

$$\begin{aligned} \mathbf{E}[f_G(ALG)] &= \mathbf{E}\left[\frac{1}{2}\cdot f_G(T_1) + \frac{1}{2}\cdot f_G(T_2)\right] \\ &\geq \mathbf{E}\left[\frac{1}{2} f_{\tilde{G}}(T_2)\right] \geq \frac{\alpha}{2} f_G(OPT). \end{aligned}$$

□

For common distributions, the constant c in Assumption 1.2 is usually small, thus Algorithm 1 requires moderately small number of samples. For instance, for distributions $\mathcal{D}_k$ and $\mathcal{D}_{\leq k}$, $\Pr_{S\sim\mathcal{D}_k}[u \in S] = k/|L|$ and $\Pr_{S\sim\mathcal{D}_{\leq k}}[u \in S] \geq 1/|L|$. Thus both distributions require only $O(\frac{|L||R|}{\delta} \ln \frac{|L||R|}{\delta})$ samples.

3.1.1. Proof of Lemma 3

We first introduce some notations. Let $|L| = n$, $|R| = m$ and $\bar{t} = \frac{2m}{\delta} \ln \frac{4mn}{\delta}$. For any node $u \in L$, let $t_u = |\{i : u \in S_i\}|$ be the number of samples where u appears. For any node $v \in R$, let $q_v = \Pr_{S\sim\mathcal{D}}[v \in N_G(S)]$ be the probability that v is covered by a sample $S \sim \mathcal{D}$. Our analysis starts with partitioning R into two subsets R_1 and R_2, where $R_1 = \{v \in R \mid q_v \leq 1 - \frac{\delta}{2m}\}$ and $R_2 = R\backslash R_1$. In general, we will show that nodes in R_1 will not appear in $\cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u))$ with high probability (Lemma 7) and R_2 will be covered by any sample $S \sim \mathcal{D}$ with high probability (Lemma 8). These facts together suffice to prove Lemma 3.

Lemma 4. *Assume that $t \geq 2n^c \cdot \bar{t}$. For fixed $u \in L$, $\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[t_u \leq \bar{t}] \leq \delta/(4mn)$.*

Proof. For fixed $u \in L$, let $X_i = 1$ if $u \in S_i$ and 0 otherwise. Then $t_u = \sum_{i=1}^t X_i$. By Assumption 1.2, $p_u = \Pr_{S\sim\mathcal{D}}[u \in S] \geq 1/n^c$. Thus $\mathbf{E}[t_u] \geq t/n^c \geq 2\bar{t}$. By Chernoff bound (Lemma 2),

$$\Pr[t_u \leq \bar{t}] = \Pr\left[t_u \leq \left(1-\frac{1}{2}\right)\cdot 2\bar{t}\right] \leq e^{-\bar{t}/4} \leq \frac{\delta}{4mn}.$$

The last inequality needs $m \geq 2\delta$, which is satisfied for all nontrivial instances. □

Lemma 5. *For any $u \in L$ and $v \in R$ such that $(u,v) \notin E$, $\Pr_{S\sim\mathcal{D}}[v \in N_G(S), u \in S] \leq \Pr_{S\sim\mathcal{D}}[v \in N_G(S)] \Pr_{S\sim\mathcal{D}}[u \in S]$.*

Proof. Just note that the event $\{v \in N_G(S)\}$ is equivalent to $\{\cup_{u'\in N_G(v)}(u' \in S)\}$. The lemma follows directly from Lemma 1. □

Lemma 6. *For any $u \in L$ and $v \in R$ such that $(u,v) \notin E$, $\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[v \in N_{\tilde{G}}(u)\backslash N_G(u), t_u = \ell] \leq q_v^l \cdot \Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[t_u = \ell]$, for any $\ell \in \mathbb{N}$.*

Proof. By the law of total probability, the formula on the left-hand side is equal to $\sum_{I\subseteq[t]:|I|=\ell} \Pr[v \in N_{\tilde{G}}(u)\backslash N_G(u), u \in \cap_{i\in I} S_i, u \notin \cup_{j\notin I} S_j]$. Since S_i's are independent samples, by construction of $N_{\tilde{G}}(u)$ and Lemma 5, we have

$$\begin{aligned} &\Pr[v \in N_{\tilde{G}}(u)\backslash N_G(u), u \in \cap_{i\in I} S_i, u \notin \cup_{j\notin I} S_j] \\ &= \Pr[v \in \cap_{i\in I} N_G(S_i), u \in \cap_{i\in I} S_i, u \notin \cup_{j\notin I} S_j] \end{aligned}$$

$$
\begin{aligned}
&= \prod_{i\in I}\Pr[v\in N_G(S_i), u\in S_i]\prod_{j\notin I}\Pr[u\notin S_j]\\
&\le \prod_{i\in I}(\Pr[v\in N_G(S_i)]\Pr[u\in S_i])\prod_{j\notin I}\Pr[u\notin S_j]\\
&= \prod_{i\in I}\Pr[v\in N_G(S_i)]\prod_{i\in I}\Pr[u\in S_i]\prod_{j\notin I}\Pr[u\notin S_j]\\
&= q_v^\ell\cdot\Pr[u\in\cap_{i\in I}S_i, u\notin\cup_{j\notin I}S_j].
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\Pr[v\in N_{\tilde G}(u)\backslash N_G(u), t_u=\ell]\\
&\le q_v^\ell\sum_{I\subseteq[t]:|I|=\ell}\Pr[u\in\cap_{i\in I}S_i, u\notin\cup_{j\notin I}S_j]\\
&= q_v^\ell\cdot\Pr[t_u=\ell].
\end{aligned}
$$

□

Lemma 7. *Assume that* $t\ge 2n^c\bar t$. *Then* $\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}[R_1\cap(\cup_{u\in L}(N_{\tilde G}(u)\backslash N_G(u)))=\emptyset]\ge 1-\delta/2$.

Proof. For node $v\in R_1$ and node $u\in L$ such that $(u,v)\notin E$, we have

$$
\begin{aligned}
&\Pr[v\in N_{\tilde G}(u)\backslash N_G(u)]\\
&= \sum_{\ell\ge 0}\Pr[v\in N_{\tilde G}(u)\backslash N_G(u), t_u=\ell]\\
&\le \sum_{\ell\ge 0}\Pr[t_u=\ell]\cdot q_v^\ell\\
&\le \sum_{\ell\le\bar t}\Pr[t_u=\ell]\cdot 1+\sum_{\ell>\bar t}\Pr[t_u=\ell]\cdot q_v^{\bar t}\\
&= \Pr[t_u\le\bar t]+\Pr[t_u>\bar t]\cdot q_v^{\bar t}\\
&\le \frac{\delta}{4mn}+\left(1-\frac{\delta}{2m}\right)^{\frac{2m}{\delta}\ln\frac{4mn}{\delta}}\\
&\le \frac{\delta}{4mn}+\frac{\delta}{4mn}=\frac{\delta}{2mn}.
\end{aligned}
$$

The first inequality holds due to Lemma 6. The second to last inequality holds due to Lemma 4, the fact that $q_v\le 1-\frac{\delta}{2m}$ for all $v\in R_1$ and $\bar t=\frac{2m}{\delta}\ln\frac{4mn}{\delta}$. Finally, by union bound, we have

$$
\begin{aligned}
&\Pr[R_1\cap(\cup_{u\in L}(N_{\tilde G}(u)\backslash N_G(u)))\neq\emptyset]\\
&= \Pr[\exists\, v\in R_1, u\in L \text{ s.t. } v\in N_{\tilde G}(u)\backslash N_G(u)]\\
&\le \sum_{v\in R_1,u\in L}\Pr[v\in N_{\tilde G}(u)\backslash N_G(u)]\\
&\le \sum_{v\in R_1,u\in L}\frac{\delta}{2mn}\le\frac{\delta}{2}.
\end{aligned}
$$

The proof is completed. □

Lemma 8. $\Pr_{S_1\sim\mathcal{D}}[R_2\subseteq N_G(S_1)]\ge 1-\delta/2$.

Algorithm 2 Tight OPSS algorithm under $\mathcal{D}_k$

Input: Samples $\{S_i, N_G(S_i)\}_{i=1}^t$, $k\in\mathbb{N}_+$, $\epsilon\in(0,1)$

1: Draw a set T_1 from $\mathcal{D}_{\epsilon k/2}$.
2: Construct a surrogate bipartite graph $\tilde G=(L,R,\tilde E)$ such that for each $u\in L$, $N_{\tilde G}(u)=\cap_{S_i:u\in S_i}N_G(S_i)$
3: Let $T_2=A(\tilde G,(1-\epsilon/2)k)$
4: **return** $T_1\cup T_2$

Proof. For a node $v\in R_2$, by definition, $\Pr_{S\sim\mathcal{D}}[v\notin N_G(S)]=1-q_v\le\frac{\delta}{2m}$. By union bound, we have $\Pr_{S\sim\mathcal{D}}[\exists\, v\in R_2 \text{ s.t. } v\notin N_G(S)]\le\delta/2$. That is, $\Pr_{S\sim\mathcal{D}}[R_2\subseteq N_G(S)]\ge 1-\delta/2$. □

Proof of Lemma 3. By Lemma 7, with probability $1-\delta/2$, $R_1\cap(\cup_{u\in L}(N_{\tilde G}(u)\backslash N_G(u)))=\emptyset$ and therefore $\cup_{u\in L}(N_{\tilde G}(u)\backslash N_G(u))\subseteq R_2$. On the other hand, by Lemma 8, with probability $1-\delta/2$, $R_2\subseteq N_G(S_1)$. Finally, by union bound, $\cup_{u\in L}(N_{\tilde G}(u)\backslash N_G(u))\subseteq N_G(S_1)$ with probability $1-\delta$. □

3.2. A Tight Algorithm for OPSS under $\mathcal{D}_k$

In this section, we present a tight algorithm for OPSS under distribution $\mathcal{D}_k$, the uniform distribution over all subsets of size k. Compared with Algorithm 1, Algorithm 2 takes an additional input $\epsilon\in(0,1)$ and has two other modifications. First, when constructing T_2, the constraint is replaced by $|S|\le(1-\epsilon/2)k$, which only incurs little loss in the approximation ratio. Second, instead of assigning a sample $S\sim\mathcal{D}_k$ to T_1, the algorithm picks a set uniformly at random from all subsets of size $\epsilon k/2$ and assigns it to T_1. The key observation is that under distribution $\mathcal{D}_k$, although T_1 is quite small, it suffices to cover nodes in $N_{\tilde G}(T_2)\backslash N_G(T_2)$ with high probability. However, this is not true for general distributions. As a result, $T_1\cup T_2$ yields an $\alpha-\epsilon$ approximation for the problem, and it is also feasible.

We begin the analysis with some notations. Let $|L|=n$, $|R|=m$ and $\bar t=\left(\frac{2m}{\epsilon}\right)^{\frac{8}{\epsilon}}\ln\frac{2mn}{\delta}$. In the analysis, we assume that $\ln^2 n\le k\le n/2$ and $m\le\frac{\epsilon}{2}n^{(\epsilon\ln n)/8}$. This is a sufficient condition for a key inequality, as we will further explain after Theorem 6. For any node $u\in L$, let $t_u=|\{i: u\in S_i\}|$ be the number of samples where u appears. For any node $v\in R$, let $q_v=\Pr_{S\sim\mathcal{D}_k}[v\in N_G(S)]$ be the probability that v is covered by a sample $S\sim\mathcal{D}_k$. Let $d(v)=|N(v)|$ denote the number of v's neighbors. Partition R into two subsets R_1 and R_2, where $R_1=\{v\in R\mid d(v)<\frac{2n}{\epsilon k}\ln\frac{2m}{\epsilon}\}$ and $R_2=R\backslash R_1$. While in the general case discussed in previous section, R is partitioned according to the value of q_v, here we partition R according to the value of $d(v)$. The reason is that $\mathcal{D}_k$ is a uniform distribution. Thus for $v\in R$, the more neighbors it has, the higher probability it will be covered by a sample

$S \sim \mathcal{D}_k$. The observation is further formulated as Lemma 9. Based on it, we can show that with high probability nodes in R_1 will not appear in $\cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u))$ (Lemma 10). Besides, q_v increases exponentially with respect to $d(v)$. Thus instead of picking a sample from $\mathcal{D}_k$, drawing a set T_1 from $\mathcal{D}_{\epsilon k/2}$ suffices to cover nodes in R_2 (Lemma 11).

Lemma 9. *For any $v \in R_1$, $q_v \le 1 - \left(\frac{\epsilon}{2m}\right)^{8/\epsilon}$.*

Proof. It is easy to verify that when $\ln^2 n \le k$ and $m \le \frac{\epsilon}{2} n^{(\epsilon \ln n)/8}$, we have $\frac{2n}{\epsilon k} \ln \frac{2m}{\epsilon} \le n/4$. Thus for any $v \in R_1$, $d(v) < \frac{2n}{\epsilon k} \ln \frac{2m}{\epsilon} \le n/4$. Together with $k \le n/2$, we have

$$\begin{aligned}
1 - q_v &= \frac{\binom{n-d(v)}{k}}{\binom{n}{k}} \\
&= \frac{(n-d(v))\cdots(n-d(v)-k+1)}{n\cdots(n-k+1)} \\
&\ge \left(1 - \frac{d(v)}{n-k+1}\right)^k \\
&\ge \left(1 - \frac{d(v)}{n/2}\right)^k \\
&\ge \exp\left(-\frac{4kd(v)}{n}\right) \\
&\ge \exp\left(-(8/\epsilon)\ln(2m/\epsilon)\right) \\
&= (\epsilon/2m)^{8/\epsilon}.
\end{aligned}$$

The third inequality holds since $1 - x \ge e^{-2x}$ for $x \in [0, 1/2]$. The last inequality holds since $d(v) < \frac{2n}{\epsilon k} \ln \frac{2m}{\epsilon}$ for $v \in R_1$. □

Similar to Lemmas 7 and 8, we show the following lemmas. The proofs are included in Section 3.2.1.

Lemma 10. *Assume that $t \ge 2(n/k) \cdot \bar{t}$. We have*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}_k}[R_1 \cap (\cup_{u\in L}(N_{\tilde{G}}(u)\backslash N_G(u))) = \emptyset] \ge 1 - \delta.$$

Lemma 11. $\Pr_{T_1\sim\mathcal{D}_{\epsilon k/2}}[R_2 \subseteq N_G(T_1)] \ge 1 - \epsilon/2$.

Now we prove Theorem 6, which is a more concrete version of Theorem 2.

Theorem 6. *For any constant $\epsilon > 0$, given any α-approximation algorithm A for the standard maximum coverage problem, coverage functions are $(\alpha - \epsilon)$-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\le k}$ over $\mathcal{D}_k$, assuming that $\ln^2 |L| \le k \le |L|/2$ and $|R| \le \frac{\epsilon}{2}|L|^{(\epsilon \ln |L|)/8}$. More precisely, for any $\delta > 0$, suppose that the number of samples $t \ge \frac{2|L|}{k}\left(\frac{2|R|}{\epsilon}\right)^{\frac{8}{\epsilon}} \ln \frac{2|L||R|}{\delta}$. Let ALG be the solution returned by Algorithm 2 and OPT be the optimal solution on the original graph G. Then*

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}_k}[\mathbf{E}[f_G(ALG)] \ge (\alpha - \epsilon) f_G(OPT)] \ge 1 - \delta.$$

Proof. By the construction of $\tilde{G}$, $N_G(u) \subseteq N_{\tilde{G}}(u)$ for any $u \in L$. Therefore, G is a subgraph of $\tilde{G}$ and $f_{\tilde{G}}(OPT) \ge f_G(OPT)$. Let OPT_k be the optimal solution when selecting k elements. Since A is an α approximation algorithm and $|T_2| \le (1 - \epsilon/2)k$,

$$\begin{aligned}
&f_{\tilde{G}}(T_2) \ge \alpha f_{\tilde{G}}(OPT_{(1-\epsilon/2)k}) \\
&\ge \alpha(1 - \epsilon/2) f_{\tilde{G}}(OPT_k) \ge \alpha(1 - \epsilon/2) f_G(OPT),
\end{aligned}$$

where the second inequality above utilizes the submodularity of the coverage functions.

Let $\mathcal{E}$ be the event $R_1 \cap (\cup_{u\in L} N_{\tilde{G}}(u)\backslash N_G(u)) = \emptyset$. By Lemma 10, $\Pr_{S_1,\cdots,S_t\sim\mathcal{D}_k}[\mathcal{E}] \ge 1 - \delta$.

We now assume that event $\mathcal{E}$ holds. In this case, we first have $N_{\tilde{G}}(T_2)\backslash N_G(T_2) \subseteq \cup_{u\in L} N_{\tilde{G}}(u)\backslash N_G(u) \subseteq R_2$. Next, conditioned on $\mathcal{E}$, we still have the claim in Lemma 11 because the sampling of T_1 is independent of the sampling of $S_1, \ldots, S_t$. Therefore, when $\mathcal{E}$ holds, we have

$$\begin{aligned}
&\mathbf{E}[f_G(ALG)] = \mathbf{E}[f_G(T_1 \cup T_2)] \\
&\ge \Pr[R_2 \subseteq N_G(T_1)]\mathbf{E}[f_G(T_1 \cup T_2) \mid R_2 \subseteq N_G(T_1)] \\
&\ge \Pr[R_2 \subseteq N_G(T_1)]\mathbf{E}[|N_{\tilde{G}}(T_2)| \mid R_2 \subseteq N_G(T_1)] \\
&\ge \alpha(1 - \epsilon/2)^2 f_G(OPT) \ge (\alpha - \epsilon) f_G(OPT).
\end{aligned}$$

This concludes the proof. □

We remark that Lemma 9 (and thus Theorem 6) holds as long as $k \le |L|/2$ and $|R| \le \frac{\epsilon}{2} e^{(\epsilon k)/8}$. The technical condition $\ln^2 |L| \le k \le |L|/2$ and $|R| \le \frac{\epsilon}{2}|L|^{(\epsilon \ln |L|)/8}$ is indeed a relaxed sufficient condition by setting a lower bound for k. However, it provides a reasonable asymptotic requirement on k and $|R|$ in terms of $|L|$.

3.2.1. Proof of Lemmas 10 and 11

Lemma 12. *Assume that $t \ge 2(n/k) \cdot \bar{t}$, where $\bar{t} = \left(\frac{2m}{\epsilon}\right)^{\frac{8}{\epsilon}} \ln \frac{2mn}{\delta}$. For fixed $u \in L$, $\Pr[t_u \le \bar{t}] \le \delta/(2mn)$.*

Proof. For fixed $u \in L$, let $X_i = 1$ if $u \in S_i$ and 0 otherwise. Then $t_u = \sum_{i=1}^t X_i$. Since $p_u = \Pr_{S\sim\mathcal{D}_k}[u \in S] \ge k/n$, $\mathbf{E}[t_u] \ge tk/n = 2\bar{t}$. By Chernoff bound (Lemma 2),

$$\Pr[t_u \le \bar{t}] = \Pr\left[t_u \le \left(1 - \frac{1}{2}\right) \cdot 2\bar{t}\right] \le e^{-\bar{t}/4} \le \frac{\delta}{2mn}.$$

The last inequality holds as long as $\bar{t} \ge 4 \ln \frac{2mn}{\delta}$. □

Proof of Lemma 10. For node $v \in R_1$ and node $u \in L$ such that $(u, v) \notin E$, we have

$$\begin{aligned}
&\Pr[v \in N_{\tilde{G}}(u)\backslash N_G(u)] \\
&= \sum_{\ell \ge 0} \Pr[v \in N_{\tilde{G}}(u)\backslash N_G(u), t_u = \ell]
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{\ell \ge 0} \Pr[t_u = \ell] \cdot q_v^{\ell} \\
&\le \sum_{\ell \le \bar{t}} \Pr[t_u = \ell] \cdot 1 + \sum_{\ell > \bar{t}} \Pr[t_u = \ell] \cdot q_v^{\bar{t}} \\
&= \Pr[t_u \le \bar{t}] + \Pr[t_u > \bar{t}] \cdot q_v^{\bar{t}} \\
&\le \frac{\delta}{2mn} + \left(1 - \left(\frac{\epsilon}{2m}\right)^{\frac{8}{\epsilon}}\right)^{\left(\frac{2m}{\epsilon}\right)^{\frac{8}{\epsilon}} \ln \frac{2mn}{\delta}} \\
&\le \frac{\delta}{2mn} + \frac{\delta}{2mn} = \frac{\delta}{mn}.
\end{aligned}
$$

The first inequality holds due to Lemma 6. The second to last inequality holds due to Lemma 12 and Lemma 9, and the fact that $\bar{t} = \left(\frac{2m}{\epsilon}\right)^{\frac{8}{\epsilon}} \ln \frac{2mn}{\delta}$.

Finally, by union bound, we have

$$
\begin{aligned}
&\Pr[R_1 \cap (\cup_{u \in L} N_{\tilde{G}}(u) \backslash N_G(u)) \neq \emptyset] \\
&= \Pr[\exists\, v \in R_1, u \in L \text{ s.t. } v \in N_{\tilde{G}}(u) \backslash N_G(u)] \\
&\le \sum_{v \in R_1, u \in L} \Pr[v \in N_{\tilde{G}}(u) \backslash N_G(u)] \\
&\le \sum_{v \in R_1, u \in L} \delta/(mn) \le \delta.
\end{aligned}
$$

The proof is completed. □

Proof of Lemma 11. For node $v \in R_2$, $d(v) \ge \frac{2n}{\epsilon k} \ln \frac{2m}{\epsilon}$, then

$$
\begin{aligned}
&\Pr_{T_1 \sim \mathcal{D}_{\epsilon k/2}}[v \notin N_G(T_1)] = \frac{\binom{n-d(v)}{\epsilon k/2}}{\binom{n}{\epsilon k/2}} \\
&= \frac{(n-d(v)) \cdots (n-d(v)-\epsilon k/2+1)}{n \cdots (n-\epsilon k/2+1)} \\
&\le \left(1 - \frac{d(v)}{n}\right)^{\epsilon k/2} \le \exp\left(-\frac{\epsilon k d(v)}{2n}\right) \\
&\le \frac{\epsilon}{2m}.
\end{aligned}
$$

By union bound, we have

$$
\Pr_{T_1 \sim \mathcal{D}_{\epsilon k/2}}[\exists\, v \in R_2, v \notin N_G(T_1)] \le \epsilon/2.
$$

That is, $\Pr_{T_1 \sim \mathcal{D}_{\epsilon k/2}}[R_2 \subseteq N_G(T_1)] \ge 1 - \epsilon/2$. □

4. Hardness Results for OPSS

4.1. The $1/2$ Hardness for OPSS under Assumption 1

Theorem 3. *There is a distribution $\mathcal{D}$ satisfying Assumption 1 such that coverage functions are not α-optimizable under OPSS in the cardinality constraint $\mathcal{M}_{\le k}$ over $\mathcal{D}$ for any $\alpha > \frac{1}{2} + o(1)$.*

Proof. The distribution $\mathcal{D}$ is constructed as follows. Number nodes in L such that $L = \{u_1, \cdots, u_n\}$. Let L_1 contain the first $k-1$ nodes and $L_2 = L \backslash L_1$. Any sample S from $\mathcal{D}$ always contains the $k-1$ nodes in L_1. The last node in S is picked uniformly at random from L_2. It is easy to see that distribution $\mathcal{D}$ satisfies Assumption 1.

Next, we construct a class of graphs $G_1, \cdots, G_{k-1}$ as follows such that they cannot be distinguished from the samples. (a) For any $i \le k-1$ and $u, v \in L$, $N_{G_i}(u) \cap N_{G_i}(v) = \emptyset$; (b) for any $i, j \le k-1$ and $u \in L_2$, $|N_{G_i}(u)| = r$ and $N_{G_i}(u) = N_{G_j}(u)$; (c) for any $i \le k-1$ and $u \in L_1$ with $u \neq u_i$, $N_{G_i}(u) = \emptyset$; (d) for any $i \le k-1$, and u_i covers the same set of $(k-1)r$ nodes across different graph G_i's. Clearly, the optimal solution OPT_i of G_i contains node u_i and arbitrary $k-1$ nodes in L_2. Thus $f_{G_i}(OPT_i) = 2(k-1)r$.

We prove the desired ratio by a probabilistic argument. Let B be any (randomized) OPSS algorithm and T be the solution it returns. Let G be a graph drawn uniformly at random from $G_1, \cdots, G_{k-1}$. Since any sample of $\mathcal{D}$ always return the first $k-1$ nodes, and the union coverage of these $k-1$ nodes is always the same across different graphs G_i's, solution T is independent of the random choice of G, although it may be dependent on the random choices in the samples from nodes in L_2. Suppose the solution T of B is fixed. Let $x = |T \cap L_1|, 0 \le x \le k-1$. By the above argument that T and G are independent, we have that the expected number of nodes covered by T is

$$
\begin{aligned}
&\mathbf{E}_G[f_G(T) \mid \text{solution } T \text{ of } B \text{ are fixed}] \\
= \;& \frac{x}{k-1}(k-1)r + (k-x)r = kr.
\end{aligned}
$$

As a result, $\mathbf{E}_{B,G}[f_G(T)] = kr$, which implies there must be a G from $G_1, \cdots, G_{k-1}$ such that $\mathbf{E}_B[f_G(T)] \le kr$. Thus $\mathbf{E}_B[f_G(T)]/f_G(OPT) \le k/(2(k-1)) = 1/2 + o(1)$. □

4.2. Assumption 1 Is Necessary

In this section, we show that the three conditions in Assumption 1 are necessary, in the sense that dropping any one of them would result in no constant approximation for the OPSS problem. The necessity of Assumption 1.2 is relatively trivial, and we include it in the full version.

Assumption 1.1 Is Necessary. For the distribution which always returns L, no reasonable algorithm exists for the OPSS problem. Thus it is easy to see that we cannot drop Assumption 1.1 without any restriction. Instead, we show that even if we relax assumption 1.1 a little bit, no constant approximation algorithm exists.

Theorem 7. *Let $\mathcal{D}_r$ be the uniform distribution over all subsets of size $r = \omega(k \log^2 |L|)$. The coverage functions are not α-optimizable under OPSS for any constant α in the cardinality constraint $\mathcal{M}_{\le k}$ over $\mathcal{D}_r$.*

Proof. Clearly, $\mathcal{D}_r$ satisfies Assumption 1.2 and 1.3, but not Assumption 1.1. Let $|L| = n$, $|R| = m = \text{poly}(n)$, and $p = r/\log^2 n = \omega(k)$. We first construct a class of graphs $G_1, \cdots, G_p$ where $G_i = (L, R, E_i)$. Let L be partitioned into disjoint subsets $\{L_1, \cdots, L_p\}$, each with $q = n/p$ nodes. For graph G_i, L_i is *good* in that for any $u \in L_i$, $N_{G_i}(u) = R$; each L_j with $j \neq i$ is *bad* in that $N_{G_i}(L_j) = \emptyset$. Clearly, the optimal solution of any graph covers m nodes.

Next we show that with high probability $G_1, \cdots, G_p$ cannot be distinguished from the samples. For L_i of G_i,

$$\Pr_{S\sim\mathcal{D}_r}[S \cap L_i = \emptyset] = \frac{\binom{n-q}{r}}{\binom{n}{r}} \leq \left(1 - \frac{q}{n}\right)^r \leq e^{-\log^2 n}.$$

Thus for $t = \text{poly}(|L|, |R|) = \text{poly}(n)$ samples, by the union bound,

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}_r}[\exists\, j \in [t] \text{ s.t. } S_j \cap L_i = \emptyset] \leq te^{-\log^2 n} = o(1).$$

Hence with probability $1 - o(1)$, $N_{G_i}(S_j) = R$ for all S_j and all G_i. Below we assume this is exactly the case for all G_i's, which means no algorithm can distinguish these G_i's from the samples.

We prove the desired ratio by a probabilistic argument. Let B be any (randomized) algorithm and T be the solution it returns. Let G be a graph drawn uniformly at random from $G_1, \cdots, G_p$ and L_g be the good part of G. Suppose the solution T of B is fixed. Since $|T| \leq k$, it can touch at most k L_j's. Thus $\Pr_G[T \cap L_g \neq \emptyset \mid$ the solution T of B is fixed$] \leq k/p = o(1)$. Since B cannot distinguish those graphs from the samples, the solution T of B is independent of the random graph G. As a result, $\Pr_{B,G}[T \cap L_g \neq \emptyset] = o(1)$ and $\mathbf{E}_{B,G}[f_G(T)] = o(1) \cdot m$, which implies there must be a G from $G_1, \cdots, G_p$ such that $\mathbf{E}_B[f_G(T)] = o(1) \cdot m$. Thus $\mathbf{E}_B[f_G(T)]/f_G(OPT) = o(1)$ with probability $1 - o(1)$. □

As a complement of Theorem 7, we show that as long as $r = O(k)$, we have a constant approximation algorithm for the OPSS problem. The proof is included in the full version.

Theorem 8. *Let $\mathcal{D}_r$ be the uniform distribution over all subsets of size $r = O(k)$. The coverage functions are α-optimizable under OPSS for some constant α in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}_r$.*

Assumption 1.3 Is Necessary. Assumption 1.3 plays a central role in the analysis of our algorithms. Thus it is reasonable to consider its necessity. In this section we show that this is exactly the case.

Theorem 9. *There is a distribution $\mathcal{D}$, which satisfies Assumption 1.1 and 1.2, but not Assumption 1.3, such that coverage functions are not α-optimizable under OPSS for any constant α in the cardinality constraint $\mathcal{M}_{\leq k}$ over $\mathcal{D}$.*

Proof. The distribution $\mathcal{D}$ is constructed as follows. Let L be partitioned into n/k disjoint subsets $L_1, \cdots, L_{n/k}$; each L_j contains exactly k nodes. A sample $S \sim \mathcal{D}$ is drawn uniformly at random from $L_1, \cdots, L_{n/k}$. Clearly, this distribution satisfies Assumption 1.1 and 1.2, but it is not negatively correlated.

Let G be a random graph constructed with the following properties: (a) $N_G(L_i) \cap N_G(L_j) = \emptyset$ for any $i \neq j$; (b) $|N_G(L_i)| = r$ for all $i \leq n/k$; (c) within each L_i, there is a node u^i such that $N_G(u^i) = N_G(L_i)$; (d) for node $u \in L_i$ with $u \neq u^i$, $N_G(u) = \emptyset$; (e) node u^i is determined by selecting a uniformly random node from L_i. All the possible outcomes of G form the graph class $\mathcal{G}$. It is easy to see graphs from $\mathcal{G}$ cannot be distinguished from the samples. The optimal solution of any graph from $\mathcal{G}$ covers kr nodes.

Now we prove the desired ratio by a probabilistic argument. Let B be any (randomized) algorithm and T be the solution it returns. Suppose the solution T of B is fixed. Then

$$\mathbf{E}_G[f_G(T) \mid \text{the solution } T \text{ of } B \text{ is fixed}]$$
$$= \sum_{j=1}^{n/k} \Pr_G[T \text{ contains } u^j \text{ of } L_j] \cdot r = \sum_{j=1}^{n/k} \frac{|T \cap L_j|}{k} \cdot r = r.$$

Since B cannot distinguish those graphs from the samples, the solution T of B is independent of the random graph G. As a result, $\mathbf{E}_{B,G}[f_G(T)] = r$, which implies there must be some fixed G in the graph class $\mathcal{G}$ such that $\mathbf{E}_B[f_G(T)] \leq r$. Thus $\mathbf{E}_B[f_G(T)]/f_G(OPT) \leq 1/k$. □

5. Future Work

One immediate question is to close the $[\frac{1}{2}(1 - e^{-1}), \frac{1}{2}]$ gap of polynomial time algorithms under Assumption 1 in our model. Besides, it is interesting to define suitable structured samples for other set functions and investigate the possibility of optimization for those functions. One concrete example of such functions is the probabilistic coverage function where each edge (u, v) in the bipartite graph $G = (L, R, E)$ has a probability indicating the probability that u covers v.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China Grants No. 61832003, 61761136014, 61872334, the 973 Program of China Grant No. 2016YFB1000201, K.C. Wong Education Foundation.

References

Badanidiyuru, A., Dobzinski, S., Fu, H., Kleinberg, R., Nisan, N., and Roughgarden, T. Sketching valuation functions. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pp. 1025–1035, 2012.

Balcan, M. and Harvey, N. J. A. Learning submodular functions. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pp. 793–802, 2011.

Balkanski, E. and Singer, Y. Minimizing a submodular function from samples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 814–822, 2017.

Balkanski, E., Rubinstein, A., and Singer, Y. The power of optimization from samples. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4017–4025, 2016.

Balkanski, E., Rubinstein, A., and Singer, Y. The limitations of optimization from samples. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 1016–1027, 2017.

Block, H. W., Savits, T. H., and Shaked, M. Some concepts of negative dependence. *The Annals of Probability*, 10 (3):765–772, 1982.

Feige, U. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

Ghosh, M. Multivariate negative dependence. *Communications in Statistics-Theory and Methods*, 10(4):307–337, 1981.

Joag-Dev, K. and Proschan, F. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

Jogdeo, K. and Patil, G. P. Probability inequalities for certain multivariate discrete distribution. *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 158–164, 1975.

Karlin, S. and Rinott, Y. Classes of orderings of measures and related correlation inequalities II. multivariate reverse rule distributions. *Journal of Multivariate Analysis*, 10 (4):499–516, 1980.

Mitzenmacher, M. and Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005. ISBN 978-0-521-83540-4.

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.

Rosenfeld, N., Balkanski, E., Globerson, A., and Singer, Y. Learning to optimize combinatorial functions. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 4371–4380, 2018.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27 (11):1134–1142, 1984.