The Limitations of Optimization from Samples[*]

Eric Balkanski
Harvard University
ericbalkanski@g.harvard.edu

Aviad Rubinstein
UC Berkeley
aviad@eecs.berkeley.edu

Yaron Singer
Harvard University
yaron@seas.harvard.edu

Abstract

In this paper we consider the following question: can we optimize objective functions from the training data we use to learn them? We formalize this question through a novel framework we call *optimization from samples* (`OPS`). In `OPS`, we are given sampled values of a function drawn from some distribution and the objective is to optimize the function under some constraint.

While there are interesting classes of functions that can be optimized from samples, our main result is an impossibility. We show that there are classes of functions which are statistically learnable and optimizable, but for which no reasonable approximation for optimization from samples is achievable. In particular, our main result shows that there is no constant factor approximation for maximizing coverage functions under a cardinality constraint using polynomially-many samples drawn from any distribution.

We also show tight approximation guarantees for maximization under a cardinality constraint of several interesting classes of functions including unit-demand, additive, and general monotone submodular functions, as well as a constant factor approximation for monotone submodular functions with bounded curvature.

[*] This paper appeared in STOC 2017

1 Introduction

The traditional approach in optimization typically assumes there is an underlying model known to the algorithm designer, and the goal is to optimize an objective function defined through the model. In a routing problem, for example, the model can be a weighted graph which encodes roads and their congestion, and the objective is to select a route that minimizes expected travel time from source to destination. In influence maximization, we are given a weighted graph which models the likelihood of individuals forwarding information, and the objective is to select a subset of nodes to spread information and maximize the expected number of nodes that receive information [41].

In many applications like influence maximization or routing, we do not actually know the objective functions we wish to optimize since they depend on the behavior of the world generating the model. In such cases, we gather information about the objective function from past observations and use that knowledge to optimize it. A reasonable approach is to learn a surrogate function that approximates the function generating the data (e.g. [33, 18, 15, 20, 19, 49]), and optimize the surrogate. In routing, we may observe traffic, fit weights to a graph that represents congestion times, and optimize for the shortest path on the weighted graph learned from data. In influence maximization, we can observe information spreading in a social network, fit weights to a graph that encodes the influence model, and optimize for the k most influential nodes. But what are the guarantees we have?

One problem with optimizing a surrogate learned from data is that it may be inapproximable. For a problem like influence maximization, for example, even if a surrogate $\widetilde{f} : 2^N \to \mathbb{R}$ approximates a submodular influence function $f : 2^N \to \mathbb{R}$ within a factor of $(1 \pm \epsilon)$ for sub-constant $\epsilon > 0$, in general there is no polynomial-time algorithm that can obtain a reasonable approximation to $\max_{S:|S| \leq k} \widetilde{f}(S)$ or $\max_{S:|S| \leq k} f(S)$ [37]. A different concern is that the function learned from data may be approximable (e.g. if the surrogate remains submodular), but its optima are very far from the optima of the function generating the data. In influence maximization, even if the weights of the graph are learned within a factor of $(1 \pm \epsilon)$ for sub-constant $\epsilon > 0$ the optima of the surrogate may be a poor approximation to the true optimum [49, 39]. The sensitivity of optimization to the nuances of the learning method therefore raises the following question:

Can we actually optimize objective functions from the training data we use to learn them?

Optimization from samples. In this paper we consider the following question: given an unknown objective function $f : 2^N \to \mathbb{R}$ and samples $\{S_i, f(S_i)\}_{i=1}^m$ where S_i is drawn from some distribution $\mathcal{D}$ and $m \in \mathsf{poly}(|N|)$, is it possible to solve $\max_{S:|S| \leq k} f(S)$? More formally:

Definition. *A class of functions $\mathcal{F} : 2^N \to \mathbb{R}$ is* **α-optimizable in $\mathcal{M}$ from samples over distribution** $\mathcal{D}$ *if there exists a (not necessarily polynomial time) algorithm whose input is a set of samples $\{S_i, f(S_i)\}_{i=1}^m$, where $f \in \mathcal{F}$ and S_i is drawn i.i.d. from $\mathcal{D}$, and returns $S \in \mathcal{M}$ s.t.:*

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

where the expectation is over the decisions of the algorithm, $m \in \mathsf{poly}(|N|)$, $\delta \in [0, 1)$ is a constant.

An algorithm with the above guarantees is an α-`OPS` algorithm. In this paper we focus on the simplest constraint, where $\mathcal{M} = \{S \subseteq N : |S| \leq k\}$ is a cardinality constraint. For a class of functions $\mathcal{F}$ we say that optimization from samples is *possible* when there exists some constant $\alpha \in (0, 1]$ and any distribution $\mathcal{D}$ s.t. $\mathcal{F}$ is α-optimizable from samples over $\mathcal{D}$ in $\mathcal{M} = \{S : |S| \leq k\}$.

Before discussing what is achievable in this framework, the following points are worth noting:

- Optimization from samples is defined per distribution. Note that if we demand optimization from samples to hold on all distributions, then trivially no function would be optimizable from samples (e.g. for the distribution which always returns the empty set);

- Optimization from samples seeks to approximate the global optimum. In learning, we evaluate a hypothesis on the same distribution we use to train it since it enables making a prediction about events that are similar to those observed. For optimization it is trivial to be competitive against a sample by simply selecting the feasible solution with maximal value from the set of samples observed. Since an optimization algorithm has the power to select any solution, the hope is that polynomially many samples contain enough information for optimizing the function. In influence maximization, for example, we are interested in selecting a set of influencers, even if we did not observe a set of highly influential individuals that initiate a cascade together.

As we later show, there are interesting classes of functions and distributions that indeed allow us to approximate the global optimum well, in polynomial-time using polynomially many samples. The question is therefore not whether optimization from samples is possible, but rather which function classes are optimizable from samples.

1.1 Optimizability and learnability

Optimization from samples is particularly interesting when functions are *learnable* and *optimizable*.

- **Optimizability.** We are interested in functions $f : 2^N \to \mathbb{R}$ and constraint $\mathcal{M}$ such that given access to a *value oracle* (given S the oracle returns $f(S)$), there exists a constant factor approximation algorithm for $\max_{S \in \mathcal{M}} f(S)$. For this purpose, monotone submodular functions are a convenient class to work with, where the canonical problem is $\max_{|S| \leq k} f(S)$. It is well known that there is a $1 - 1/e$ approximation algorithm for this problem [50] and that this is tight using polynomially many value queries [27]. Influence maximization is an example of maximizing a monotone submodular function under a cardinality constraint [41].

- **PMAC-learnability.** The standard framework in the literature for learning set functions is *Probably Mostly Approximately Correct* (α-PMAC) learnability due to Balcan and Harvey [4]. This framework nicely generalizes Valiant's notion of *Probably Approximately Correct* (PAC) learnability [58]. Informally, PMAC-learnability guarantees that after observing polynomially many samples of sets and their function values, one can construct a surrogate function that is likely to, α-approximately, mimic the behavior of the function observed from the samples (see Appendix D for formal definitions). Since the seminal paper of Balcan and Harvey, there has been a great deal of work on learnability of submodular functions [28, 5, 2, 29, 30, 3].

Are functions that are learnable and optimizable also optimizable from samples?

1.2 Main result

Our main result is an impossibility. We show that there is an interesting class of functions that is PMAC-learnable and optimizable but not optimizable from samples. This class is coverage functions.

Definition. *A function is called* **coverage** *if there exists a family of sets $T_1, \ldots, T_n$ that covers subsets of a universe U with weights $w(a_j)$ for $a_j \in U$ such that for all S, $f(S) = \sum_{a_j \in \cup_{i \in S} T_i} w(a_j)$. A coverage function is* **polynomial-sized** *if the universe is of polynomial size in n. Influence maximization is a generalization of maximizing coverage functions under a cardinality constraint.*

Coverage functions are a canonical example of monotone submodular functions and are hence optimizable. In terms of learnability, for any constant $\epsilon > 0$, coverage functions are $(1 - \epsilon)$-PMAC learnable over any distribution [2], unlike monotone submodular functions which are generally not PMAC learnable [4]. Somewhat surprisingly, coverage functions are not optimizable from samples.

Theorem. *No algorithm can obtain an approximation better than $2^{-\Omega(\sqrt{\log n})}$ for maximizing a polynomial-sized coverage function under a cardinality constraint, using polynomially many samples drawn from any distribution.*

Coverage functions are heavily used in machine learning [55, 60, 35, 43, 1, 45, 56], data-mining [13, 21, 52, 54, 16, 34], mechanism design [17, 44, 23, 24, 9, 22], privacy [36, 28], as well as influence maximization [41, 53, 8]. In many of these applications, the functions are learned from data and the goal is to optimize the function under a cardinality constraint. In addition to learnability and optimizability, coverage functions have many other desirable properties (see Section E). One important fact is that they are *parametric*: if the sets $T_1, \ldots, T_n$ are known, then the coverage function is completely defined by the weights $\{w(a) : a \in U\}$. Our impossibility result holds even in the case where the sets $T_1, \ldots, T_n$ are known.

Technical overview. In the value query model, information theoretic impossibility results use functions defined over a partition of the ground set [46, 59, 26]. The hardness then arises from hiding all the information about the partition from the algorithm. Although the constructions in the OPS model also rely on a partition, the techniques are different since the impossibility is quasi-polynomial and not constant. In particular, the algorithm may learn the entire partition, and the hardness arises from hiding which parts of the partition are "good" or "bad". We begin by describing a framework which reduces the problem of showing hardness results to constructing good and bad functions which satisfy certain properties. The desired good and bad functions must have equal value on small sets of equal sizes and a large gap in value on large sets. Interestingly, a main technical difficulty is to simultaneously satisfy these two simple properties, which we do with novel techniques for constructing coverage functions. Another technical part is the use of tools from pseudorandomness to obtain coverage functions of polynomial size.

1.3 Algorithms for OPS

There are classes of functions and distributions for which optimization from samples is possible. Most of the algorithms use a simple technique that consists of estimating the expected marginal contribution of an element to a random sample. For general submodular functions, we show an essentially tight bound using a non-trivial analysis of an algorithm that uses such estimates.

Theorem. *There exists an $\tilde{\Omega}(n^{-1/4})$-OPS algorithm over a distribution $\mathcal{D}$ for monotone submodular functions. Furthermore, this approximation ratio is essentially tight.*

For unit-demand and additive functions, we give near-optimal optimization from samples results. The result for unit-demand is particularly interesting as it shows one can easily optimize a function from samples even when recovering it is impossible (see Section 4). For monotone submodular functions with curvature c, we obtain a $((1 - c)^2 - o(1))$-OPS algorithm.

1.4 Paper Organization

We begin with the hardness result in Section 2. The OPS algorithms are presented in Section 3. We discuss the notion of recoverability in Section 4 and additional related work in Section 5. The proofs are deferred to the appendix.

2 Impossibility of Optimization from Samples

We show that optimization from samples is in general impossible, over any distribution $\mathcal{D}$, even when the function is learnable and optimizable. Specifically, we show that there exists no constant α and distribution $\mathcal{D}$ such that coverage functions are α-optimizable from samples, even though they are $(1-\epsilon)$-`PMAC` learnable over any distribution $\mathcal{D}$ and can be maximized under a cardinality constraint within a factor of $1 - 1/e$. In Section 2.1, we construct a framework which reduces the problem of proving information theoretic lower bounds to constructing functions that satisfy certain properties. We then construct coverage functions that satisfy these properties in Section 2.2.

2.1 A Framework for OPS Hardness

The framework we introduce partitions the ground set of elements into *good, bad,* and *masking* elements. We derive two conditions on the values of these elements so that samples do not contain enough information to distinguish good and bad elements with high probability. We then give two additional conditions so that if an algorithm cannot distinguish good and bad elements, the solution returned by this algorithm has low value compared to the optimal set consisting of the good elements. We begin by defining the partition.

Definition. *The collection of partitions $\mathcal{P}$ contains all partitions P of the ground set N in r parts $T_1, \dots, T_r$ of k elements and a part M of remaining $n - rk$ elements, where $n = |N|$.*

The elements in T_i are called the *good* elements, for some $i \in [r]$. The *bad* and *masking* elements are the elements in $T_{-i} := \cup_{j=1, j\neq i}^{r} T_j$ and M respectively. Next, we define a class of functions $\mathcal{F}(g, b, m, m^+)$ such that $f \in \mathcal{F}(g, b, m, m^+)$ is defined in terms of good, bad, and masking functions g, b, and m^+, and a masking fraction $m \in [0, 1]$.[1]

Definition. *Given functions g, b, m, m^+, the class of functions $\mathcal{F}(g, b, m, m^+)$ contains functions $f^{P,i}$, where $P \in \mathcal{P}$ and $i \in [r]$, defined as*

$$f^{P,i}(S) := (1 - m(S \cap M))\Big(g(S \cap T_i) + b(S \cap T_{-i})\Big) + m^+(S \cap M).$$

We use probabilistic arguments over the partition $P \in \mathcal{P}$ and the integer $i \in [r]$ chosen uniformly at random to show that for any distribution $\mathcal{D}$ and any algorithm, there exists a function in $\mathcal{F}(g, b, m, m^+)$ that the algorithm optimizes poorly given samples from $\mathcal{D}$. The functions g, b, m, m^+ have desired properties that are parametrized below. At a high level, the identical on small samples and masking on large samples properties imply that the samples do not contain enough information to learn i, i.e. distinguish good and bad elements, even though the partition P can be learned. The gap and curvature property imply that if an algorithm cannot distinguish good and bad elements, then the algorithm performs poorly.

Definition. *The class of functions $\mathcal{F}(g, b, m, m^+)$ has an (α, β)-**gap** if the following conditions are satisfied for some t, where $\mathcal{U}(\mathcal{P})$ is the uniform distribution over $\mathcal{P}$.*

1. ***Identical on small samples.*** *For a fixed $S : |S| \leq t$, with probability $1 - n^{-\omega(1)}$ over partition $P \sim \mathcal{U}(\mathcal{P})$, $g(S \cap T_i) + b(S \cap T_{-i})$ is independent of i;*
2. ***Masking on large samples.*** *For a fixed $S : |S| \geq t$, with probability $1 - n^{-\omega(1)}$ over partition $P \sim \mathcal{U}(\mathcal{P})$, the masking fraction is $m(S \cap M) = 1$;*

[1] The notation m^+ refers to the role of this function, which is to maintain monotonicity of masking elements. These four functions are assumed to be normalized such that $g(\emptyset) = b(\emptyset) = m(\emptyset) = m^+(\emptyset) = 0$.

3. α-***Gap.*** *Let $S : |S| = k$, then $g(S) \geq \max\{\alpha \cdot b(S), \alpha \cdot m^+(S)\}$;*

4. β-***Curvature.*** *Let $S_1 : |S_1| = k$ and $S_2 : |S_2| = k/r$, then $g(S_1) \geq (1 - \beta) \cdot r \cdot g(S_2)$.*

The following lemma reduces the problem of showing an impossibility result to constructing g, b, m, and m^+ which satisfy the above properties.

Theorem 2.1. *Assume the functions g, b, m, m^+ have an (α, β)-gap, then $\mathcal{F}(g, b, m, m^+)$ is not $2\max(1/(r(1-\beta)), 2/\alpha)$-optimizable from samples over any distribution $\mathcal{D}$.*

Consider a distribution $\mathcal{D}$. The proof of this result consists of three parts.

1. Fix a set S. With probability $1 - n^{-\omega(1)}$ over $P \sim \mathcal{U}(\mathcal{P})$, $f^{P,i}(S)$ is independent of i, by the identical on small samples and masking on large samples properties.

2. There exists a partition $P \in \mathcal{P}$ such that with probability $1 - n^{-\omega(1)}$ over polynomially many samples $\mathcal{S}$ drawn from $\mathcal{D}$, $f^{P,i}(S)$ is independent of i for all $S \in \mathcal{S}$. Thus, given samples $\{(S_j, f^{P,i}(S_j))\}_j$ for such a partition P, the decisions of the algorithm are independent of i.

3. There exists $f^{P,i}$ such that the algorithm does not obtain a $2\max(1/(r(1-\beta)), 2/\alpha)$ approximation for $f^{P,i}$ with samples from $\mathcal{D}$. This holds by a second probabilistic argument, this time over $i \in \mathcal{U}([r])$, and by the gap and curvature properties.

2.2 OPS Hardness of Coverage Functions

We use this framework to show that there exists no constant α and distribution $\mathcal{D}$ such that coverage functions are α-optimizable from samples over $\mathcal{D}$. We first state a definition of coverage functions that is equivalent to the traditional definition and that is used through this section.

Definition 1. *A function $f : 2^N \to \mathbb{R}$ is coverage if there exists a bipartite graph $G = (N \cup \{a_j\}_j, E)$ between elements N and children $\{a_j\}_j$ with weights $w(a_j)$ such that the value of a set S is the sum of the weights of the children covered by S, i.e., for all $S \subseteq N$, $f(S) = \sum_{a_j : \exists e a_j \in E, e \in S} w(a_j)$. A coverage function is polynomial-sized if the number of children is polynomial in $n = |N|$.*

The construction of good and bad coverage functions g and b that combine the identical on small samples property and a large α-gap on large sets as needed by the framework is a main technical challenge. The bad function b needs to increase slowly (or not at all) for large sets to obtain a large α-gap, which requires a non-trivial overlap in the children covered by bad elements (this is related to coverage functions being *second-order supermodular* [42]). The overlap in children covered by good elements then must be similar (identical on small samples) while the good function still needs to grow quickly for large sets (large gap), as illustrated in Figure 1. We consider the cardinality constraint $k = n^{2/5-\epsilon}$ and a number of parts $r = n^{1/5-\epsilon}$. At a high level, the proof follows three main steps.

1. **Constructing the good and bad functions.** In Section 2.2.1, we construct the good and bad functions whose values are identical on small samples for $t = n^{3/5+\epsilon}$, have gap $\alpha = n^{1/5-\epsilon}$, and curvature $\beta = o(1)$. These good and bad functions are affine combinations of primitives $\{C_p\}_{p \in \mathbb{N}}$ which are coverage functions with desirable properties;

2. **Constructing the masking function.** In Section 2.2.2, we construct m and m^+ that are masking on large samples for $t = n^{3/5+\epsilon}$ and that have a gap $\alpha = n^{1/5}$. In this construction, masking elements cover the children from functions g and b such that t masking elements cover all the children, but k masking elements only covers an $n^{-1/5}$ fraction of them.

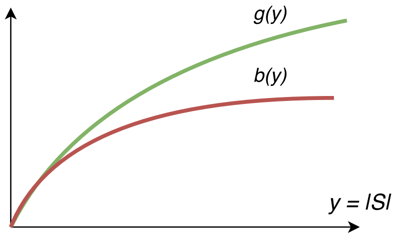

Figure 1: Sketches of the desired $g(S)$ and $b(S)$ in the simple case where they only depend on $y = |S|$.

3. **From exponential to polynomial-sized coverage functions.** Lastly in Section 2.2.3 we prove the hardness result for polynomial-sized coverage functions. This construction relies on constructions of ℓ-wise independent variables to reduce the number of children.

2.2.1 Constructing the Good and the Bad Coverage Functions

In this section we describe the construction of good and bad functions that are identical on small samples for $t = n^{3/5+\epsilon}$, with a gap $\alpha = n^{1/5-\epsilon}$ and curvature $\beta = o(1)$. To do so, we introduce a class of primitive functions C_p, through which we express the good and bad functions. For symmetric functions h (i.e. whose value only depends on the size of the set), we abuse notation and simply write $h(y)$ instead of $h(S)$ for a set S of size y.

The construction. We begin by describing the primitives we use for the good and bad functions. These primitives are the family $\{C_p\}_{p\in\mathbb{N}}$, which are symmetric, and defined as:

$$C_p(y) = p \cdot (1 - (1 - 1/p)^y) \,.$$

These are *coverage* functions defined over an *exponential* number of children.

Claim 1. *Consider the coverage function over ground set N where for each set S, there is a child a_S that is covered by exactly S, and child a_S has weight $w(a_S) = p \cdot \Pr(S \sim \mathrm{B}(N, 1/p))$ where the binomial distribution $\mathrm{B}(N, 1/p)$ picks each element in N independently with probability $1/p$, then this coverage function is C_p.*

For a given $\ell \in [n]$, we construct g and b as affine combinations of ℓ coverage functions $C_{p_j}(y)$ weighted by variables x_j for $j \in [\ell]$:

- **The good function** is defined as:

$$g(y) := y + \sum_{j\,:\,x_j<0} (-x_j) C_{p_j}(y)$$

- **The bad function** is defined as $b(S) = \sum_{j=1, j\neq i}^{r} b'(S \cap T_j)$, with

$$b'(y) := \sum_{j\,:\,x_j>0} x_j C_{p_j}(y)$$

Overview of the analysis of the good and bad functions. Observe that if $g(y) = b'(y)$ for all $y \leq \ell$ for some sufficiently large ℓ, then we obtain the identical on small samples property. The main idea is to express these ℓ constraints as a system of linear equations $A\mathbf{x} = \mathbf{y}$ where $A_{ij} := C_{p_j}(i)$ and $y_j := j$, with $i, j \in [\ell]$. We prove that this matrix has two crucial properties:

1. A **is invertible.** In Lemma A.1 we show that there exists $\{p_j\}_{j=1}^{\ell}$ such that the matrix A is invertible by interpreting its entries defined by C_{p_j} as non-zero polynomials of degree ℓ. This implies that the system of linear equations $A \cdot \mathbf{x} = \mathbf{y}$ can be solved and that there exists a coefficient $\mathbf{x}^*$ needed for our construction of the good and the bad functions;

2. $||\mathbf{x}^\star||_\infty$ **is bounded.** In Lemma A.4 we use Cramer's rule and Hadamard's inequality to prove that the entries of $\mathbf{x}^\star$ are bounded. This implies that the linear term y in $g(y)$ dominates $x_j^\star \cdot C_{p_j}(y)$ for large y and all j. This then allows us to prove the curvature and gap properties.

These properties of A imply the desired properties of g and b for $\ell = \log \log n$.

Lemma 2.1. *For every constant $\epsilon > 0$, there exists coverage functions g, b such that the identical on small samples property holds for $t = n^{3/5+\epsilon}$, with gap $\alpha = n^{1/5-\epsilon}$ and curvature $\beta = o(1)$.*

Lemma A.3 shows the identical on small samples property. It uses Lemma A.2 which shows that if $|S| \leq n^{3/5+\epsilon}$, then with probability $1 - n^{-\omega(1)}$, $|S \cap T_j| \leq \log \log n$ for all j. The property then follows from the system of linear equations. The gap and curvature properties are proven in Lemmas A.5 and A.6 using the fact that the term y in g dominates the other terms in g and b.

2.2.2 Constructing the Masking Function

Masking elements allow the indistinguishability of good and bad elements from large samples.

The masking elements. The construction of the coverage functions g and b defined in the previous section is generalized so that we can add masking elements with desirable properties. For each child a_i in the coverage function defined by $g+b$, we divide a_i into $n^{3/5}$ children $a_{i,1}, \ldots, a_{i,n^{3/5}}$ with equal weights $w(a_{i,j}) = \frac{w(a_i)}{n^{3/5}}$ for all j. Each element covering a_i according to g and b now covers children $a_{i,1}, \ldots, a_{i,n^{3/5}}$. Note that the value of $g(S)$ and $b(S)$ remains unchanged with this new construction and thus, the previous analysis still holds. Each masking elements in M is defined by drawing $j \sim \mathcal{U}[n^{3/5}]$ and having this element cover children $a_{i,j}$ for all i.

The masking function $m^+(S)$ is the total weight covered by masking elements S and the masking fraction $m(S)$ is the fraction of $j \in [n^{3/5}]$ such that j is drawn for at least one element in S.

Masking properties. Masking elements cover children that are already covered by good or bad elements. A large number of masking elements mask the good and bad elements, which implies that good and bad elements are indistinguishable.

- In Lemma A.8 we prove that the masking property holds for $t = n^{3/5+\epsilon}$.
- We show a gap $\alpha = n^{1/5}$ in Lemma A.9. For any $S : |S| \leq k$, we have $g(S) \geq n^{1/5} \cdot m^+(S)$.

An impossibility result for exponential size coverage functions. We have the four properties for a $(n^{1/5-\epsilon}, o(1))$-gap. The functions $f^{P,i}$ are coverage since g, b, m^+ are coverage and m^+ is the fraction of overlap between children from g, b, and m^+

Claim 2. *Coverage functions are not $n^{-1/5+\epsilon}$-optimizable from samples over any distribution $\mathcal{D}$, for any constant $\epsilon > 0$.*

2.2.3 From Exponential to Polynomial Size Coverage Functions

The construction above relies on the primitives C_p which are defined with exponentially many children. In this section we modify the construction to use primitives c_p which are coverage with *polynomially* many children. The function class $\mathcal{F}(g, b, m, m^+)$ obtained are then coverage functions with polynomially many children. The functions c_p we construct satisfy $c_p(y) = C_p(y)$ for all $y \le \ell$, and thus the matrix A for polynomial size coverage functions is identical to the general case. We lower the cardinality constraint to $k = 2^{\sqrt{\log n}} = |T_j|$ so that the functions $c_p(S \cap T_j)$ need to be defined over only $2^{\sqrt{\log n}}$ elements. We also lower the number of parts to $r = 2^{\sqrt{\log n}/2}$.

Maintaining symmetry via ℓ-wise independence. The technical challenge in defining a coverage function with polynomially many children is in maintaining the symmetry of non-trivial size sets. To do so, we construct coverage functions $\{\zeta^z\}_{z \in [k]}$ for which the elements that cover a random child are approximately ℓ-wise independent. The next lemma reduces the problem to that of constructing coverage functions ζ^z that satisfy certain properties.

Lemma 2.2. *Assume there exist symmetric (up to sets of size ℓ) coverage functions ζ^z with $\mathsf{poly}(n)$ children that are each covered by $z \in [k]$ parents. Then, there exists coverage functions c_p with $\mathsf{poly}(n)$ children that satisfy $c_p(S) = C_p(y)$ for all S such that $|S| = y \le \ell$, and $c_p(k) = C_p(k)$.*

The proof is constructive. We obtain c_p by replacing, for all $z \in [k]$, all children in C_p that are covered by z elements with children from ζ^z with weights normalized that sum up $C_p(k)$. Next, we construct such ζ^z. Assume without loss that k is prime (otherwise pick some prime close to k). Given $\mathbf{a} \in [k]^\ell$, and $x \in [z]$, let

$$h_{\mathbf{a}}(x) := \sum_{i \in [\ell]} a_i x^i \mod k$$

The children in ζ^z are $U = \{\mathbf{a} \in [k]^\ell : h_{\mathbf{a}}(x_1) \neq h_{\mathbf{a}}(x_2) \text{ for all distinct } x_1, x_2 \in [z]\}$. The k elements are $\{j : 0 \le j < k\}$. Child $\mathbf{a}$ is covered by elements $\{h_{\mathbf{a}}(x) : x \in [z]\}$. Note that $|U| \le k^\ell = 2^{\ell\sqrt{\log n}}$ and we pick $\ell = \log\log n$ as previously. The next lemma shows that we obtain the desired properties for ζ^z.

Lemma 2.3. *The coverage function ζ^z is symmetric for all sets of size at most ℓ.*

At a high level, the proof uses Lemma A.10 which shows that the parents of a random child $\mathbf{a}$ are approximately ℓ-wise independent. This follows from $h_{\mathbf{a}}(x)$ being a polynomial of degree $\ell - 1$, a standard construction for ℓ-wise independent random variables. Then, using inclusion-exclusion over subsets T of a set S of size at most ℓ, the probability that T is the parents of a child $\mathbf{a}$ only depends on $|T|$ by Lemma A.10. Thus, $\zeta^z(S)$ only depends on $|S|$. We are now ready to show the properties for g, b, m, m^+ with polynomially many children,

Lemma 2.4. *There exists polynomial-sized coverage functions g, b, m, and m^+ that satisfy an $(\alpha = 2^{\Omega(\sqrt{\log n})}, \beta = o(1))$-gap with $t = n^{3/5+\epsilon}$.*

We construct g, b, m, m^+ as in the general case but in terms of primitives c_p instead of C_p. By Lemmas 2.2 and 2.3, we obtain the same matrix A and coefficients $\mathbf{x}^\star$ as in the general case, so the identical on small samples property holds. The masking on large samples and curvature property hold almost identically as previously. Finally, since k is reduced, the gap α is reduced to $2^{\Omega(\sqrt{\log n})}$.

OPS Hardness for Coverage Functions. We get our main result by combining Theorem 2.1 with this $(\alpha = 2^{\Omega(\sqrt{\log n})}, \beta = o(1))$-gap.

Theorem 2.2. *For every constant $\epsilon > 0$, coverage functions are not $n^{-1/5+\epsilon}$-optimizable from samples over any distribution $\mathcal{D}$. In addition, polynomial-sized coverage functions are not $2^{-\Omega(\sqrt{\log n})}$-optimizable from samples over any distribution $\mathcal{D}$.*

3 Algorithms for OPS

In this section we describe `OPS`-algorithms. Our algorithmic approach is not to learn a surrogate function and to then optimize this surrogate function. Instead, the algorithms estimate the expected marginal contribution of elements to a random sample directly from the samples (Section 3.1) and solve the optimization problem using these estimates. The marginal contribution of an element e to a set S if $f_S(e) := f(S \cup \{e\}) - f(S)$. If these marginal contributions are decreasing, i.e., $f_S(e) \geq f_T(e)$ for all $e \in N$ and $S \subseteq T \subseteq N$, then f is submodular. If they are positive, i.e., $f_S(e) \geq 0$ for all $e \in N$ and $S \subseteq N$, then f is monotone.

This simple idea turns out to be quite powerful; we use these estimate to develop an $\tilde{\Omega}(n^{-1/4})$ `OPS`-algorithm for monotone submodular functions in Section 3.2. This approximation is essentially tight with a hardness result for general submodular functions shown in Appendix F that uses the framework from the previous section. In Section 3.3 we show that when samples are generated from a product distribution, there are interesting classes of functions that are amenable to optimization from samples.

3.1 OPS via Estimates of Expected Marginal Contributions

A simple case in which the expected marginal contribution $\mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e_i)]$ of an element e_i to a random set $S \sim \mathcal{D}$ can be estimated arbitrarily well is that of product distributions. We now show a simple algorithm we call EEMC which estimates the expected marginal contribution of an element when the distribution $\mathcal{D}$ is a product distribution. This estimate is simply the difference between the average value of a sample containing e_i and the average value of a sample not containing e_i.

Algorithm 1 EEMC Estimates the Expected Marginal Contribution $\mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e)]$.

```
Input: $\mathcal{S} = \{S_j : (S_j, f(S_j)) \text{ is a sample})\}$
  for $i \in [n]$ do
    $\mathcal{S}_i \leftarrow \{S : S \in \mathcal{S}, e_i \in S\}$
    $\mathcal{S}_{-i} \leftarrow \{S : S \in \mathcal{S}, e_i \not\in S\}$
    $\hat{v}_i \leftarrow \frac{1}{|\mathcal{S}_i|}\sum_{S\in\mathcal{S}_i} f(S) - \frac{1}{|\mathcal{S}_{-i}|}\sum_{S\in\mathcal{S}_{-i}} f(S)$
  end for
  return $(\hat{v}_1, \ldots, \hat{v}_n)$
```

Lemma 3.1. *Let $\mathcal{D}$ be a product distribution with bounded marginals.*[2] *Then, with probability at least $1 - O(e^{-n})$, the estimations $\hat{v}_i$ are ϵ accurate, for any $\epsilon \geq f(N)/\mathsf{poly}(n)$ and for all e_i, i.e.,*

$$|\hat{v}_i - \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e_i)]\,| \leq \epsilon.$$

[2] The marginals are bounded if for all e, $e \in S \sim \mathcal{D}$ and $e \not\in S \sim \mathcal{D}$ w.p. at least $1/\mathsf{poly}(n)$ and at most $1-1/\mathsf{poly}(n)$.

The proof consists of the following two steps. First note that

$$\mathbf{E}_{S\sim\mathcal{D}|e_i\notin S}[f_S(e_i)] = \mathbf{E}_{S\sim\mathcal{D}|e_i\notin S}[f(S\cup e_i)] - \mathbf{E}_{S\sim\mathcal{D}|e_i\notin S}[f(S)] = \mathbf{E}_{S\sim\mathcal{D}|e_i\in S}[f(S)] - \mathbf{E}_{S\sim\mathcal{D}|e_i\notin S}[f(S)]$$

where the second equality is since $\mathcal{D}$ is a product distribution. Then, from standard concentration bounds, the average value $(\sum_{S\in\mathcal{S}_i} f(S))/|\mathcal{S}_i|$ of a set containing e_i estimates $\mathbf{E}_{S\sim\mathcal{D}|e_i\in S}[f(S)]$ well. Similarly, $(\sum_{S\in\mathcal{S}_{-i}} f(S))/|\mathcal{S}_{-i}|$ estimates $\mathbf{E}_{S\sim\mathcal{D}|e_i\notin S}[f(S)]$.

3.2 A Tight Approximation for Submodular Functions

We develop an $\tilde{\Omega}(n^{-1/4})$-OPS algorithm over $\mathcal{D}$ for monotone submodular functions, for some distribution $\mathcal{D}$. This bound is essentially tight since submodular functions are not $n^{-1/4+\epsilon}$-optimizable from samples over *any* distribution (Appendix F). We first describe the distribution for which the approximation holds. Then we describe the algorithm, which builds upon estimates of expected marginal contributions.

The distribution. Let $\mathcal{D}_i$ be the uniform distribution over all sets of size i. Define the distribution $\mathcal{D}^{sub}$ to be the distribution which draws from $\mathcal{D}_k$, $\mathcal{D}_{\sqrt{n}}$, and $\mathcal{D}_{\sqrt{n}+1}$ at random. In Lemma B.2 we generalize Lemma 3.1 to estimate $\hat{v}_i \approx \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\notin S}[f_S(e_i)]$ with samples from $\mathcal{D}_{\sqrt{n}}$ and $\mathcal{D}_{\sqrt{n}+1}$.

The algorithm. We begin by computing the expected marginal contributions of all elements. We then place the elements in $3\log n$ bins according to their estimated expected marginal contribution $\hat{v}_i$. The algorithm then simply returns either the best sample of size k or a random subset of size k of a random bin. Up to logarithmic factors, we can restrict our attention to just one bin. We give a formal description below.

Algorithm 2 An $\tilde{\Omega}(n^{-1/4})$-optimization from samples algorithm over $\mathcal{D}^{sub}$ for monotone submodular functions.

```
Input: $\mathcal{S} = \{S_i : (S_i, f(S_i)) \text{ is a sample})\}$
  With probability $\frac{1}{2}$:
      return  $\text{argmax}_{S\in\mathcal{S}\,:\,|S|=k} f(S)$                    best sample of size $k$
  With probability $\frac{1}{2}$:
      $(\hat{v}_1, \ldots, \hat{v}_n) \leftarrow \text{EEMC}(\mathcal{S})$
      $\hat{v}_{max} \leftarrow \max_i \hat{v}_i$
      for $j \in [3\log n]$ do
          $B_j \leftarrow \left\{i : \frac{\hat{v}_{max}}{2^{j1-1}} \leq \hat{v}_i < \frac{\hat{v}_{max}}{2^{j1}}\right\}$
      end for
      Pick $j \in [3\log n]$ u.a.r.
      return  $S$, a subset of $B_j$ of size $\min\{|B_j|, k\}$ u.a.r.      a random set from a random bin
```

Analysis of the algorithm. The main crux of this result is in the analysis of the algorithm.

Theorem 3.1. *Algorithm 2 is an $\tilde{\Omega}(n^{-1/4})$-OPS algorithm over $\mathcal{D}^{sub}$ for monotone submodular functions.*

The analysis is divided in two cases, depending if a random set $S \sim \mathcal{D}_{\sqrt{n}}$ of size $\sqrt{n}$ has low value or not. Let $S^\star$ be the optimal solution.

- Assume that $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S)] \leq f(S^\star)/4$. Thus, optimal elements have large estimated expected marginal contribution $\hat{v}_i$ by submodularity. Let $B^\star$ be the bin with the largest value among the bins with contributions $\hat{v} \geq f(S^\star)/(4k)$. We argue that a random subset of $B^\star$ of size k performs well. Lemma B.4 shows that a random subset of $B^\star$ is a $|B^\star|/(4k\sqrt{n})$-approximation. At a high level, a random subset S of size $\sqrt{n}$ contains $|B^\star|/\sqrt{n}$ elements from bin $B^\star$ in expectation, and these $|B^\star|/\sqrt{n}$ elements $S_{B^\star}$ have contributions at least $f(S^\star)/(4k)$ to $S_{B^\star}$. Lemma B.5 shows that a random subset of $B^\star$ is an $\tilde{\Omega}(k/|B^\star|)$-approximation to $f(S^\star)$. The proof first shows that $f(B^\star)$ has high value by the assumption that a random set $S \sim \mathcal{D}_{\sqrt{n}}$ has low value, and then uses the fact that a subset of $B^\star$ of size k is a $k/|B^\star|$ approximation to $B^\star$. Note that either $|B^\star|/(4k\sqrt{n})$ or $\tilde{\Omega}(k/|B^\star|)$ is at least $\tilde{\Omega}(n^{-1/4})$.

- Assume that $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S)] \geq f(S^\star)/4$. We argue that the best sample of size k performs well. Lemma B.6 shows that, by submodularity, a random set of size k is a $k/(4\sqrt{n})$ approximation since a random set of size k is a fraction $k/(\sqrt{n})$ smaller than a random set from $\mathcal{D}_{\sqrt{n}}$ in expectation. Lemma B.7 shows that the best sample of size k is a $1/k$-approximation since it contains the elements with the highest value with high probability. Note that either $k/(4\sqrt{n})$ or $1/k$ is at least $n^{-1/4}$.

3.3 Bounded Curvature and Additive Functions

A simple $((1-c)^2 - o(1))$-OPS algorithms for monotone submodular functions with curvature c over product distributions follows immediately from estimating expected marginal contributions. This result was recently improved to $(1-c)/(1+c-c^2)$, which was shown to be tight [6]. An immediate corollary is that additive (linear) functions, which have curvature 0, are $(1-o(1))$-OPS over product distributions. The curvature c of a function measures how far this function is to being additive.

Definition. *The* curvature c *of a submodular function* f *is* $c := 1 - \min_{e\in N, S\subseteq N} f_{S\setminus e}(e)/f(e)$.

This definition implies that $f_S(e) \geq (1-c)f(e) \geq (1-c)f_T(e)$ for all S, T and all $e \notin S \cup T$ since $f(e) \geq f(T \cup e) - f(T) = f_T(e)$ where the first inequality is by submodularity. The algorithm simply returns the k elements with the highest expected marginal contributions.

Algorithm 3 MaxMargCont: A $((1-c)^2 - o(1))$-optimization from samples algorithm for monotone submodular functions with curvature c.

```
Input: $\mathcal{S} = \{S_i : (S_i, f(S_i)) \text{ is a sample}\}$
  $(\hat{v}_1, \ldots, \hat{v}_n) \leftarrow$ EEMC($\mathcal{S}$)
  return  $S \leftarrow \text{argmax}_{|T|=k} \sum_{i\in T} \hat{v}_i$
```

Theorem 3.2. *Let f be a monotone submodular function with curvature c and $\mathcal{D}$ be a product distribution with bounded marginals. Then* MaxMargCont *is a $((1-c)^2 - o(1))$-OPS algorithm.*

The proof follows almost immediately from the definition of curvature. Let S be the set returned by the algorithm and $S^\star$ be the optimal solution, then $f(S)$ and $f(S^\star)$ are sums of marginal contributions of elements in S and $S^\star$ which are each at most a factor $1-c$ away from their estimated expected marginal contribution by curvature. A $1-o(1)$ approximation follows immediately for additive functions since they have curvature $c = 0$. A function f is additive if $f(S) = \sum_{e_i\in S} f(\{e_i\})$.

Corollary 1. *Let f be an additive function and $\mathcal{D}$ be a product distribution with bounded marginals. Then* MaxMargCont *is a $(1-o(1))$-OPS algorithm.*

4 Recoverability

The largely negative results from the above sections lead to the question of how well must a function be learned for it to be optimizable from samples? One extreme is a notion we refer to as recoverability (REC). A function is recoverable if it can be learned *everywhere* within an approximation of $1 \pm 1/n^2$ from samples. Does a function need to be learnable everywhere for it to be optimizable from samples?

Definition. *A function f is recoverable for distribution $\mathcal{D}$ if there exists an algorithm which, given a polynomial number of samples drawn from $\mathcal{D}$, outputs a function $\tilde{f}$ such that for all sets S,*

$$\left(1 - \frac{1}{n^2}\right) f(S) \leq \tilde{f}(S) \leq \left(1 + \frac{1}{n^2}\right) f(S)$$

with probability at least $1-\delta$ over the samples and the randomness of the algorithm, where $\delta \in [0,1)$ is a constant.

This notion of recoverability is similar to the problem of approximating a function everywhere from Goemans et al. [32]. The differences are that recoverability is from samples whereas their setting allows value queries and that recoverability requires being within an approximation of $1 \pm 1/n^2$. It is important for us to be within such bounds and not within some arbitrarily small constant because such perturbations can still lead to an $O(n^{-1/2+\delta})$ impossibility result for optimization [37]. We show that if a monotone submodular function f is recoverable then it is optimizable from samples by using the greedy algorithm on the recovered function $\tilde{f}$. The proof is similar to the classical analysis of the greedy algorithm.

Theorem 4.1. *If a monotone submodular function f is recoverable over $\mathcal{D}$, then it is $1-1/e-o(1)$-optimizable from samples over $\mathcal{D}$. For additive functions, it is $1-o(1)$-optimizable from samples.*

We show that additive functions are in REC under some mild condition. Combined with the previous result, we get an alternate proof from the previous section for additive functions being $1-o(1)$-optimizable from samples over product distributions.

Lemma 4.1. *Let f be an additive function with $v_{max} = \max_i f(\{e_i\})$, $v_{min} = \min_i f(\{e_i\})$ and let $\mathcal{D}$ be a product distribution with bounded marginals. If $v_{min} \geq v_{max}/poly(n)$, then f is recoverable for $\mathcal{D}$.*

We also note that submodular functions that are a *c-junta* for some constant c are recoverable. A function f is a c-junta [48, 29, 57] if it depends only on a set of elements T of size c. If c is constant, then with enough samples, T can be learned since each element not in T is in at least one sample which does not contain any element in T. For each subset of T, there is also at least one sample which intersects with T in exactly that subset, so f is exactly recoverable.

The previous results lead us to the following question: Does a function need to be recoverable to be optimizable from samples? We show that it is not the case since unit demand functions are optimizable from samples and not recoverable. A function f is a unit demand function if $f(S) = \max_{e_i \in S} f(\{e_i\})$.

Lemma 4.2. *Unit demand functions are not recoverable for $k \geq n^\epsilon$ but are 1-OPS.*

We conclude that functions do not need to be learnable everywhere from samples to be optimizable from samples.

5 Additional related work

Revenue maximization from samples. The discrepancy between the model on which algorithms optimize and the true state of nature has recently been studied in algorithmic mechanism design. Most closely related to our work are several recent papers (e.g. [14, 25, 11, 40, 47, 12]) that also consider models that bypass the learning algorithm and let the mechanism designer access samples from a distribution rather than an explicit Bayesian prior. In contrast to our negative conclusions, these papers achieve mostly positive results. In particular, Huang et al. [40] show that the obtainable revenue is much closer to the optimum than the information-theoretic bound on learning the valuation distribution.

Comparison to online learning and reinforcement learning. Another line of work which combines decision-making and learning is online learning (see survey [38]). In online learning, a player iteratively makes decisions. For each decision, the player incurs a cost and the cost function for the current iteration is immediately revealed. The objective is to minimize regret, which is the difference between the sum of the costs of the decisions of the player and the sum of the costs of the best fixed decision. The fundamental differences with our framework are that decisions are made online after each observation, instead of offline given a collection of observations. The benchmarks, regret in one case and the optimal solution in the other, are not comparable.

A similar comparison can be made with the problem of reinforcement learning, where at each iteration the player typically interacts with a Markov decision process (MDP) [51]. At each iteration, an action is chosen in an online manner and the player receives a reward based on the action and the state in the MDP she is in. Again, this differs from our setting where there is one offline decision to be made given a collection observations.

Additional learning results for submodular functions. In addition to the PMAC learning results mentioned in the introduction for coverage functions, there are multiple learning results for submodular functions. Monotone submodular functions are α-PMAC learnable over product distributions for some constant α under some assumptions [4]. Impossibility results arise for general distributions, in which case submodular functions are not $\tilde{\Omega}(n^{-1/3})$-PMAC learnable [4]. Finally, submodular functions can be $(1-\epsilon)$-PMAC learned for the uniform distribution over all sets with a running time and sample complexity exponential in ϵ and polynomial in n [29]. This exponential dependency is necessary since $2^{\Omega(\epsilon^{-2/3})}$ samples are needed to learn submodular functions with ℓ_1-error of ϵ over this distribution [31].

References

[1] Ioannis Antonellis, Anish Das Sarma, and Shaddin Dughmi. Dynamic covering for recommendation systems. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 26–34. ACM, 2012.

[2] Ashwinkumar Badanidiyuru, Shahar Dobzinski, Hu Fu, Robert Kleinberg, Noam Nisan, and Tim Roughgarden. Sketching valuation functions. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012.*

[3] Maria-Florina Balcan. Learning submodular functions with applications to multi-agent systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, 2015.

[4] Maria-Florina Balcan and Nicholas J. A. Harvey. Learning submodular functions. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, 2011.

[5] Maria-Florina Balcan, Florin Constantin, Satoru Iwata, and Lei Wang. Learning valuation functions. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, 2012.

[6] Eric Balkanski, Aviad Rubinstein, and Yaron Singer. The power of optimization from samples. In *NIPS*, 2016.

[7] Ian F Blake and Chris Studholme. Properties of random matrices and applications. *Unpublished report available at http://www. cs. toronto. edu/˜ cvs/coding*, 2006.

[8] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. Maximizing social influence in nearly optimal time. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 946–957. SIAM, 2014.

[9] Dave Buchfuhrer, Michael Schapira, and Yaron Singer. Computation and incentives in combinatorial public projects. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 33–42. ACM, 2010.

[10] Deeparnab Chakrabarty and Zhiyi Huang. Testing coverage functions. In *Automata, Languages, and Programming*, pages 170–181. 2012.

[11] Shuchi Chawla, Jason D. Hartline, and Denis Nekipelov. Mechanism design for data science. In *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*, pages 711–712, 2014.

[12] Yu Cheng, Ho Yee Cheung, Shaddin Dughmi, Ehsan Emamjomeh-Zadeh, Li Han, and Shang-Hua Teng. Mixture selection, mechanism design, and signaling. In *FOCS*, 2015. To appear.

[13] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Max-cover in map-reduce. In *Proceedings of the 19th international conference on World wide web*, pages 231–240. ACM, 2010.

[14] Richard Cole and Tim Roughgarden. The sample complexity of revenue maximization. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 243–252, 2014.

[15] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 793–801, 2014.

[16] Anirban Dasgupta, Arpita Ghosh, Ravi Kumar, Christopher Olston, Sandeep Pandey, and Andrew Tomkins. The discoverability of the web. In *Proceedings of the 16th international conference on World Wide Web*, pages 421–430. ACM, 2007.

[17] Shahar Dobzinski and Michael Schapira. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1064–1073. Society for Industrial and Applied Mathematics, 2006.

[18] Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3147–3155, 2013.

[19] Nan Du, Yingyu Liang, Maria-Florina Balcan, and Le Song. Learning time-varying coverage functions. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3374–3382, 2014.

[20] Nan Du, Yingyu Liang, Maria-Florina Balcan, and Le Song. Influence function learning in information diffusion networks. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 2016–2024, 2014.

[21] Nan Du, Yingyu Liang, Maria-Florina F Balcan, and Le Song. Learning time-varying coverage functions. In *Advances in neural information processing systems*, pages 3374–3382, 2014.

[22] Shaddin Dughmi. A truthful randomized mechanism for combinatorial public projects via convex optimization. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 263–272. ACM, 2011.

[23] Shaddin Dughmi and Jan Vondrák. Limitations of randomized mechanisms for combinatorial auctions. *Games and Economic Behavior*, 92:370–400, 2015.

[24] Shaddin Dughmi, Tim Roughgarden, and Qiqi Yan. From convex optimization to randomized mechanisms: toward optimal combinatorial auctions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 149–158. ACM, 2011.

[25] Shaddin Dughmi, Li Han, and Noam Nisan. Sampling and representation complexity of revenue maximization. In *Web and Internet Economics - 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings*, pages 277–291, 2014.

[26] Alina Ene, Jan Vondrák, and Yi Wu. Local distribution and the symmetry gap: Approximability of multiway partitioning problems. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 306–325, 2013.

[27] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45 (4):634–652, 1998.

[28] Vitaly Feldman and Pravesh Kothari. Learning coverage functions and private release of marginals. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014.

[29] Vitaly Feldman and Jan Vondrák. Optimal bounds on approximation of submodular and XOS functions by juntas. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, 2013.

[30] Vitaly Feldman and Jan Vondrák. Tight bounds on low-degree spectral concentration of submodular and xos functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 923–942. IEEE, 2015.

[31] Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. Representation, approximation and learning of submodular functions using low-rank decision trees. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, 2013.

[32] Michel X Goemans, Nicholas JA Harvey, Satoru Iwata, and Vahab Mirrokni. Approximating submodular functions everywhere. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009.

[33] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1019–1028, 2010.

[34] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[35] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 265–272, 2005.

[36] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. *SIAM Journal on Computing*, 42(4):1494–1520, 2013.

[37] Avinatan Hassidim and Yaron Singer. Submodular optimization under noise. 2015. Working paper.

[38] Elad Hazan. Draft: Introduction to online convex optimization. In *Foundations and Trends in Optimization, vol. XX, no. XX*, pages 1–172. 2016.

[39] Xinran He and David Kempe. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 885–894, 2016.

[40] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. Making the most of your samples. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15, Portland, OR, USA, June 15-19, 2015*, pages 45–60, 2015.

[41] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[42] Nitish Korula, Vahab S. Mirrokni, and Morteza Zadimoghaddam. Online submodular welfare maximization: Greedy beats 1/2 in random order. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, 2015.

[43] Andreas Krause and Carlos Guestrin. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pages 1650–1654, 2007.

[44] Benny Lehmann, Daniel Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 18–28. ACM, 2001.

[45] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics, 2011.

[46] Vahab Mirrokni, Michael Schapira, and Jan Vondrák. Tight information-theoretic lower bounds for welfare maximization in combinatorial auctions. In *Proceedings of the 9th ACM conference on Electronic commerce*, 2008.

[47] Jamie Morgenstern and Tim Roughgarden. The pseudo-dimension of nearly-optimal auctions. In *NIPS*, page Forthcoming, 12 2015. URL papers/auction-pseudo.pdf.

[48] Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 206–212. ACM, 2003.

[49] Harikrishna Narasimhan, David C. Parkes, and Yaron Singer. Learnability of influence in networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3186–3194, 2015.

[50] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions ii. *Math. Programming Study 8*, 1978.

[51] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[52] Barna Saha and Lise Getoor. On maximum coverage in the streaming model & application to multi-topic blog-watch. In *SDM*, volume 9, pages 697–708. SIAM, 2009.

[53] Lior Seeman and Yaron Singer. Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 459–468. IEEE, 2013.

[54] Yaron Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 733–742. ACM, 2012.

[55] Ashwin Swaminathan, Cherian V Mathew, and Darko Kirovski. Essential pages. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 173–182, 2009.

[56] Hiroya Takamura and Manabu Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781–789. Association for Computational Linguistics, 2009.

[57] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 11–20. IEEE, 2012.

[58] Leslie G. Valiant. A Theory of the Learnable. *Commun. ACM*, 1984.

[59] Jan Vondrák. Symmetry and approximability of submodular maximization problems. *SIAM J. Comput.*, 42(1):265–304, 2013.

[60] Yisong Yue and Thorsten Joachims. Predicting diverse subsets using structural svms. In *Proceedings of the 25th international conference on Machine learning*, pages 1224–1231, 2008.

Appendix

A Impossibility of OPS

A Framework for OPS Hardness

We reduce the problem of showing hardness results to the problem of constructing g, b, m, m^+ with an (α, β)-gap. Recall that a partition P has r parts $T_1, \dots, T_r$ of k elements and a part M of remaining $n - rk$ elements. The functions $f^{P,i}(S) \in \mathcal{F}(g, b, m, m^+)$ are defined as $f^{P,i}(S) := (1 - m(S \cap M))\,(g(S \cap T_i) + b(S \cap T_{-i})) + m^+(S \cap M)$ with $i \in [r]$.

Theorem 2.1. *Assume the functions g, b, m, m^+ have an (α, β)-gap, then $\mathcal{F}(g, b, m, m^+)$ is not $2\max(1/(r(1-\beta)), 2/\alpha)$-optimizable from samples over any distribution $\mathcal{D}$.*

Proof. Fix any distribution $\mathcal{D}$. We first claim that for a fixed set S, $f^{P,i}(S)$ is independent of i with probability $1 - n^{-\omega(1)}$ over a uniformly random partition $P \sim \mathcal{U}(\mathcal{P})$. If $|S| \leq t$, then the claim holds immediately by the identical on small samples property. If $|S| \geq t$, then $m(S \cap M) = 1$ with probability $1 - n^{-\omega(1)}$ over P by the masking on large samples property and $f^{P,i}(S) = m^+(S \cap M)$.

Next, we claim that there exists a partition $P \in \mathcal{P}$ such that $f^{P,i}(S)$ is independent of i with probability $1 - n^{-\omega(1)}$ over $S \sim \mathcal{D}$. Denote the event that $f^{P,i}(S)$ is independent of i by $I(S, P)$. By switching sums,

$$\begin{aligned}
&\sum_{P \in \mathcal{P}} \Pr(P \sim \mathcal{U}(\mathcal{P})) \sum_{S \in 2^N} \Pr(S \sim \mathcal{D})\, \mathbb{1}_{I(S,P)} \\
=&\sum_{S \in 2^N} \Pr(S \sim \mathcal{D}) \sum_{P \in \mathcal{P}} \Pr(P \sim \mathcal{U}(\mathcal{P}))\, \mathbb{1}_{I(S,P)} \\
\geq &\sum_{S \in 2^N} \Pr(S \sim \mathcal{D}) \left(1 - n^{-\omega(1)}\right) \\
=&1 - n^{-\omega(1)}
\end{aligned}$$

where the inequality is by the first claim. Thus there exists some P such that

$$\sum_{S \in 2^N} \Pr(S \sim \mathcal{D})\, \mathbb{1}_{I(S,P)} \geq 1 - n^{-\omega(1)},$$

which proves the desired claim.

Fix a partition P such that the previous claim holds, i.e., $f^{P,i}(S)$ is independent of i with probability $1 - n^{-\omega(1)}$ over a sample $S \sim \mathcal{D}$. Then, by a union bound over the polynomially many samples, $f^{P,i}(S)$ is independent of i for all samples S with probability $1 - n^{-\omega(1)}$, and we assume this is the case for the remaining of the proof. It follows that the choices of the algorithm given samples from $f \in \{f^{P,i}\}_{i=1}^r$ are independent of i. Pick $i \in [r]$ uniformly at random and consider the (possibly randomized) set S returned by the algorithm. Since S is independent of i, we get $\mathbf{E}_{i,S}[|S \cap T_i|] \leq k/r$. Let $S_{k/r} = \text{argmax}_{S:|S|=k/r}(g(S))$, we obtain

$$\begin{aligned}
\mathbf{E}_{i,S}\left[f^{P,i}(S)\right] &\leq \mathbf{E}_{i,S}\left[g(S \cap T_i) + b(S \cap T_{-i}) + m^+(S \cap M)\right] && (m(S \cap M) \leq 1) \\
&\leq g(S_{k/r}) + b(S) + m^+(S) && \text{(monotone and submodular)} \\
&\leq \frac{1}{r(1-\beta)} g(T_i) + \frac{2}{\alpha} g(T_i) && \text{(curvature and gap)} \\
&\leq 2\max\left(\frac{1}{r(1-\beta)}, \frac{2}{\alpha}\right) f^{P,i}(T_i)
\end{aligned}$$

Thus, there exists at least one i such that the algorithm does not obtain a $2\max(1/(r(1-\beta)), 2/\alpha)$-approximation to $f^{\mathcal{P},i}(T_i)$, and T_i is the optimal solution. □

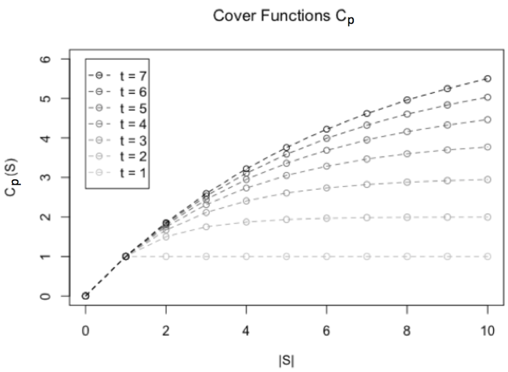

Figure 2: The value of coverage functions $C_p(y)$ for $1 \leq p \leq 7$ and sets $y \in [10]$.

OPS Hardness of Coverage Functions

We consider the cardinality constraint $k = n^{2/5-\epsilon}$ and the number of parts $r = n^{1/5-\epsilon}$.

Construction the Good and the Bad Coverage Functions

For symmetric functions h (i.e. whose value only depends on the size of the set), we abuse notation and simply write $h(y)$ instead of $h(S)$ for a set S of size y. We begin by showing that the primitives $C_p(y) = p \cdot (1 - (1 - 1/p)^y)$ (illustrated in Figure 2) are coverage functions. It then follows that the functions g and b are coverage.

Claim 1. *Consider the coverage function over ground set N where for each set S, there is a child a_S that is covered by exactly S, and child a_S has weight $w(a_S) = p \cdot \Pr(S \sim \mathrm{B}(N, 1/p))$ where the binomial distribution $\mathrm{B}(N, 1/p)$ picks each element in N independently with probability $1/p$, then this coverage function is C_p.*

Proof. Note that

$$\begin{aligned}
C_p(S) &= \sum_{T\,:\,|T\cap S|\geq 1} w(a_T) \\
&= t \cdot \sum_{T\,:\,|T\cap S|\geq 1} \Pr(T \sim \mathrm{B}(N, 1/p)) \\
&= t \cdot \left(1 - \sum_{T\,:\,|T\cap S|=0} \Pr(T \sim \mathrm{B}(N, 1/p))\right) \\
&= t \cdot \left(1 - \Pr_{T\sim \mathrm{B}(N,1/p)}[|T \cap S| = 0]\right) \\
&= t \cdot \left(1 - \left(1 - \frac{1}{p}\right)^{|S|}\right).
\end{aligned}$$

□

In the remaining of this section, we prove Lemma 2.1.

Lemma 2.1. *For every constant $\epsilon > 0$, there exists coverage functions g, b such that the identical on small samples property holds for $t = n^{3/5+\epsilon}$, with gap $\alpha = n^{1/5-\epsilon}$ and curvature $\beta = o(1)$.*

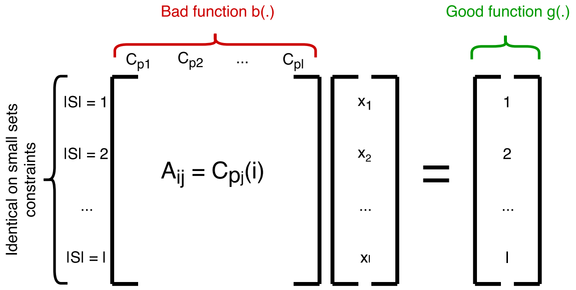

Figure 3: The matrix A.

The good and bad functions are defined as $g(y) = y + \sum_{j\,:\,x_j<0}(-x_j)C_{p_j}(y)$ and $b(S) = \sum_{j=1,j\neq i}^{r} b'(S \cap T_j)$ with $b'(y) := \sum_{j\,:\,x_j>0} x_j C_{p_j}(y)$. We obtain the coefficients $\mathbf{x}$ by solving the system of linear equations $A\mathbf{x} = \mathbf{y}$ where $A_{ij} := C_{p_j}(i)$ and $y_j := j$ as illustrated in Figure 3, with $i, j \in [\ell]$.

To prove Lemma 2.1, we begin by showing that A is invertible in Lemma A.1, so that the coefficients $\mathbf{x}$ satisfying the system of linear equations exist. We then show the three desired properties. Lemma A.2 shows that a set S of size at most $n^{3/5+\epsilon}$ contains at most ℓ elements from any part T_j w.p. $1 - n^{-\omega(1)}$, thus the identical on small samples property holds by the system of linear equations (Lemma A.3). Lemma A.4 bounds the coefficients $\mathbf{x}$, thus the y term in the good function dominates and we obtain the gap (Lemma A.5) and curvature (Lemma A.6) properties.

Lemma A.1. *Matrix $A(\{p_j\}_{j=1}^{\ell})$ is invertible for some set of integers $\{p_j\}_{j=1}^{\ell}$ such that $j \leq p_j \leq j(j+1)$ for all $1 \leq j \leq \ell$.*

Proof. The proof goes by induction on ℓ and shows that it is possible to pick p_ℓ such that the rows of $A(\{p_j\}_{j=1}^{\ell})$ are linearly independent. The base case is trivial. In the inductive step, assume $p_1, \cdots, p_{\ell-1}$ have been picked so that the $(\ell-1) \times (\ell-1)$ matrix $A(\{p_j\}_{j=1}^{\ell-1})$ is invertible. We show that for some choice of integer $p_\ell \in [p_{\ell-1}, \ell(\ell+1)]$ there does not exist a vector $\mathbf{z}$ such that $\sum_{i\leq\ell} z_i A_{i,j} = 0$ for all $j \leq \ell$ where $A = A(\{p_j\}_{j=1}^{\ell})$. We write the first $\ell - 1$ entries of row A_ℓ as a linear combination of the other $\ell - 1$ rows:

$$\sum_{i<\ell} z_i A_{i,j} = A_{\ell,j} \quad \forall j < \ell.$$

Since $A(\{p_j\}_{j=1}^{\ell-1})$ is invertible by the inductive hypothesis, there exists a unique solution $\mathbf{z}^\star$ to the above system of linear equations. It remains to show that $\sum_{i<\ell} z_i^\star A_{i,\ell} \neq A_{\ell,\ell}$, which by the uniqueness of $\mathbf{z}^\star$ implies that there does not exist a vector z such that $\sum_{i\leq\ell} z_i A_{i,j} = 0$ for all $j \leq \ell$. Observe that $A_{\ell,\ell} + \sum_{i<\ell} z_i^\star A_{i,\ell} = (p_\ell^\ell - (p_\ell - 1)^\ell + \sum_{i<\ell} z_i^\star (p_\ell^i - (p_\ell - 1)^i) p_\ell^{\ell-i})/p_\ell^{\ell-1}$ and that

$$p_\ell^\ell - (p_\ell - 1)^\ell + \sum_{i<\ell} z_i^\star (p_\ell^i - (p_\ell - 1)^i) p_\ell^{\ell-i}$$

is a non-zero polynomial of degree ℓ that has at most ℓ roots. Therefore, there exists p_ℓ such that $p_{\ell-1} < p_\ell \leq p_{\ell-1} + \ell + 1$ and $\sum_{i<\ell} z_i^\star A_{i,\ell} \neq A_{\ell,\ell}$. So the rows of $A(\{p_j\}_{j=1}^{\ell})$ are linearly independent and the matrix is invertible. We get the bounds on p_ℓ by the induction hypothesis, $p_\ell \leq p_{\ell-1} + \ell + 1 \leq (\ell - 1)\ell + \ell + 1 \leq \ell(\ell + 1)$. □

We need the following lemma to show the identical on small samples property.

Lemma A.2. *Let T be a uniformly random set of size $|T|$ and consider a set S such that $|T|\cdot|S|/n \leq n^{-\epsilon}$ for some constant $\epsilon > 0$, then $\Pr(|S \cap T| \geq \ell) = n^{-\Omega(\ell)}$.*

Proof. We start by considering a subset L of S of size ℓ. We first bound the probability that L is a subset of T,

$$\Pr(L \subseteq T) \leq \prod_{e \in L} \Pr(e \in T) \leq \prod_{e \in L} \frac{|T|}{n} = \left(\frac{|T|}{n}\right)^{\ell}.$$

We then bound the probability that $|S \cap T| > \ell$ with a union bound over the events that a set L is a subset of T, for all subsets L of S of size ℓ:

$$\Pr(|S \cap T| > \ell) \leq \sum_{L \subseteq S \,:\, |L| = \ell} \Pr(L \subseteq S) \leq \binom{|S|}{\ell} \cdot \left(\frac{|T|}{n}\right)^{\ell} \leq \left(\frac{|T| \cdot |S|}{n}\right)^{\ell} \leq n^{-\epsilon\ell}$$

where the last inequality follows from the assumption that $|T| \cdot |S|/n \leq n^{-\epsilon}$. □

For coverage functions, we let $\ell = \log\log n$.

Lemma A.3. *The identical on small samples property holds for $t = n^{3/5+\epsilon/2}$.*

Proof. Lemma A.2 implies that $|S \cap T_j| \leq \ell = \log\log n$ w.p. $1 - \omega(1)$ over $P \sim \mathcal{U}(\mathcal{P})$ for all j for a set S of size at most $n^{3/5+\epsilon/2}$. Thus, $g(S \cap T_j) = b^T(S \cap T_j)$ for all j w.p. $1 - \omega(1)$ by the system of linear equations, which implies the identical on small samples property for $t = n^{3/5+\epsilon/2}$. □

The gap and curvature properties require bounding the coefficients $\mathbf{x}$ (Lemma A.4). We recall two basic results from linear algebra (Theorems A.1 and A.2) that are used to bound the coefficients.

Theorem A.1 (Cramer's rule)**.** *Let A be an invertible matrix. The solution to the linear system $Ax = y$ is given by $x_i = \frac{\det A_i}{\det A}$, where A_i is the matrix A with the i-th column replaced by the vector y.*

Theorem A.2 (Hadamard's inequality)**.** $\det A \leq \prod \|v_i\|$*, where $\|v_i\|$ denotes the Euclidean norm of the i-th column of A.*

Lemma A.4. *Let $\mathbf{x}^\star$ be the solution to the system of linear equations $\left(A(\{p_j\}_{j=1}^{\ell})\right)\mathbf{x} = \mathbf{y}$, then the entries of this solution are bounded: $|x_i^\star| \leq \ell^{O(\ell^4)}$.*

Proof. Denote $A := A(\{p_j\}_{j=1}^{\ell})$. By Lemma A.1, A is invertible, so let $\mathbf{x}^\star = (A)^{-1}\mathbf{y}$. By Cramer's rule (Theorem A.1), $x_i^\star = \frac{\det A_i}{\det A}$, where A_i is A with the i-th column replaced by the vector $\mathbf{y}$. Using the bound from Lemma A.1, every entry in A can be represented as a rational number, with numerator and denominator bounded by $\ell^{O(\ell)}$. We can multiply by all the denominators, and get an integer matrix with positive entries bounded by $\ell^{O(\ell^3)}$. Now, by Hadamard's inequality (Theorem A.2), the determinants of the integral A and all the A_i's are integers bounded by $\ell^{O(\ell^4)}$. Therefore every entry in $\mathbf{x}^\star$ can be written as a rational number with numerator and denominator bounded by $\ell^{O(\ell^4)}$. □

Using the bounds previously shown for $\mathbf{x}^\star$, the two following lemmas establish the gap α and curvature β of the good and bad functions $g(\cdot)$ and $b(\cdot)$.

Lemma A.5. *The gap between the good and the bad functions $g(\cdot)$ and $b(\cdot)$ is at least $\alpha = n^{1/5-\epsilon}$ for general coverage functions and at least $\alpha = 2^{\Omega(\sqrt{\log n})}$ for polynomial-size coverage functions.*

Proof. We show the gap between the good and the bad function on a set S of size k. Recall that $b(S) \leq r \cdot b^T(k) = r \cdot \sum_{j\,:\,x_j^\star > 0, j \leq \ell} x_j^\star C_{p_j}(k)$. We can bound each summand as:

$$\begin{aligned}
x_j^\star C_{p_j}(k) &\leq x_j^\star p_j && (C_p \text{ and } c_p \text{ upper bounded by } p)\\
&\leq x_j^\star \ell(\ell+1) && (\text{Lemma A.1})\\
&\leq \ell^{O(\ell^4)} && (\text{Lemma A.4}),
\end{aligned}$$

and therefore $b^T(k) \leq \ell^{O(\ell^4)}$. On the other hand, the good function is bounded from below by the cardinality: $g(k) \geq k$. Plugging in $k = n^{2/5-\epsilon}$, $r = n^{1/5-\epsilon}$ and $\ell = \log\log n$, we get the following gap α,

$$\frac{g(k)}{b(S)} \geq \frac{n^{2/5}}{n^{1/5}(\log\log n)^{\log^4 \log n}} \gg n^{1/5-\epsilon}.$$

With $k = 2^{\sqrt{\log n}}$, $r = 2^{\sqrt{\log n}/2}$, and $\ell = \log\log n$, we get

$$\frac{g(k)}{b(S)} \geq \frac{2^{\sqrt{\log n}}}{2^{\sqrt{\log n}/2}(\log\log n)^{\log^4 \log n}} = 2^{(1-o(1))\sqrt{\log n}/2}.$$

□

Lemma A.6. *The curvature for both the general and polynomial-size good function is $\beta = o(1)$.*

Proof. Note that $k/r \geq 2^{\sqrt{\log n}/2}$ for both the general and polynomial size cases. Thus, the curvature β is

$$\begin{aligned}
1 - \frac{g(k)}{r \cdot g(k/r)} &\leq 1 - \frac{k}{r \cdot k/r + r \cdot (\log\log n)^{\log^4 \log n}}\\
&\leq 1 - (1 + 2^{-\sqrt{\log n}/2} \cdot (\log\log n)^{\log^4 \log n})^{-1}\\
&= o(1)
\end{aligned}$$

where the first inequality follows a similar reasoning as the one used to upper bound $b(S)$ in Lemma A.5. □

Finally, combining Lemmas A.3, A.5, and A.6, we get Lemma 2.1.

Constructing the Masking Function

To obtain the desired properties of the masking functions m^+ and masking fraction m^+, each child a_i in the universe of $g + b$ is divided into $n^{3/5}$ children $a_{i,1}, \dots a_{i,n^{3/5}}$ of equal weights $w(a_i)/n^{3/5}$. For each masking element, draw $j \in \mathcal{U}([n^{3/5}])$, then this masking element covers $a_{i,j}$ for all i. The function $m^+(S)$ is then the total weight covered by masking elements S and the masking fraction $m(S)$ is the fraction of $j \in [n^{3/5}]$ such that j is drawn for at least one element in S. Lemmas A.8 and A.9 show the masking property on large samples and the α-gap for masking elements. We begin by stating the Chernoff bound, used in Lemma A.8.

Lemma A.7 (Chernoff Bound)**.** *Let $X_1, \dots, X_n$ be independent indicator random variables such that* $\Pr(X_i = 1) = 1$. *Let* $X = \sum_{i=1}^n X_i$ *and* $\mu = \mathbf{E}[X]$. *For* $0 < \delta < 1$,

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

Lemma A.8. *Consider the construction above for masking elements, then the masking on large samples property holds for* $t = n^{3/5+\epsilon}$.

Proof. First, we show that a large set S contains a large number of masking elements with exponentially high probability.[3] We then show that a large number of masking elements covers all the children with exponentially high probability, thus $m(S \cap M) = 1$.

Masking elements are a $1 - o(1)$ fraction of N since there are $n - rk = n - n^{1/5-\epsilon}n^{2/5-\epsilon}$ masking elements. By Chernoff bound (Lemma A.7), a set of size at least $n^{3/5+\epsilon}$ contains at least $n^{3/5+\epsilon/2}$ masking elements with exponentially high probability. By another Chernoff bound, with $n^{3/5+\epsilon/2}$ masking elements, at least one of these elements cover a fixed child $a_{i,j}$ with exponentially high probability. By a union bound, this holds for all $j \in [n^{3/5}]$. Finally, note that if a set of masking elements cover $a_{i,1}, \ldots, a_{i,n^{3/5}}$ for some i, this set covers $a_{i,1}, \ldots, a_{i,n^{3/5}}$ for all i. Thus w.p. at least $1 - n^{-\omega(1)}$, $m(S \cap M) = 1$. □

Lemma A.9. *The masking function* m *has a gap* $\alpha = n^{1/5+\epsilon}$ *with the good function* g.

Proof. We first bound the value of all good and bad elements, and then bound the fraction of that total value obtained by k masking elements. The value of all bad elements is

$$\begin{aligned}
\sum_{j=1, j\neq j}^{m} b^T(T_j) &= (r-1) \sum_{1\le j\le \ell \,:\, x_j^\star \ge 0} x_j^\star C_{p_j}(k) && \text{(Definition of } b^T\text{)}\\
&\le r \sum_{1\le j\le \ell} \ell^{O(\ell^4)} C_{p_j}(k) && \text{(Lemma A.4)}\\
&\le r \sum_{1\le j\le \ell} \ell^{O(\ell^4)} p_j && (C_p, c_p \le p)\\
&\le r \sum_{1\le j\le \ell} \ell^{O(\ell^4)} && \text{(Lemma A.1)}\\
&\le o(k) && (\ell = \log\log n, r = n^{1/5-\epsilon}, k = n^{2/5-\epsilon})\\
&\le o(g(k))
\end{aligned}$$

Now note that a masking element covers a $1/n^{3/5}$ fraction of the value of all good and bad elements by the above construction. Thus, $k = n^{2/5-\epsilon}$ masking elements cover at most a $1/n^{1/5+\epsilon}$ fraction of the total value of all good and bad elements, combining this with the total value of bad elements that is upper bounded by $o(g(k))$ concludes the proof. □

Combining Lemmas 2.1, A.8, and A.9, we obtain an $(n^{1/5-\epsilon}, o(1))$-gap. The main result for exponential size coverage functions then follows from Theorem 2.1.

Claim 2. *Coverage functions are not* $n^{-1/5+\epsilon}$*-optimizable from samples over any distribution* $\mathcal{D}$*, for any constant* $\epsilon > 0$.

From Exponential to Polynomial Size Coverage Functions

We modify C_p to use primitives c_p which are coverage with *polynomially* many children. The function class $\mathcal{F}(g, b, m, m^+)$ obtained are then coverage functions over a polynomial-size universe. The matrix A for polynomial size coverage functions is identical as in the general case. We lower

[3] Formally, with exponentially high probability means with probability at least $1 - e^{-\Omega(n^\epsilon)}$ for some constant $\epsilon > 0$.

the cardinality constraint to $k = 2^{\sqrt{\log n}} = |T_j|$ so that the functions $c_p(S \cap T_j)$ need to be defined over only $2^{\sqrt{\log n}}$ elements. We also lower the number of parts to $r = 2^{\sqrt{\log n}/2}$.

The main technical challenge is to obtain symmetric coverage functions for sets of size at most ℓ with polynomially many children. We start by reducing the problem to constructing such functions with certain properties in Lemma 2.2. We then construct such functions and prove they satisfy these properties in Lemma 2.3. Combining these Lemmas, we obtain a $(2^{\Omega(\sqrt{\log n})}, o(1))$-gap (Lemma 2.4).

Lemma 2.2. *Assume there exist symmetric (up to sets of size ℓ) coverage functions ζ^z with $\mathsf{poly}(n)$ children that are each covered by $z \in [k]$ parents. Then, there exists coverage functions c_p with $\mathsf{poly}(n)$ children that satisfy $c_p(S) = C_p(y)$ for all S such that $|S| = y \le \ell$, and $c_p(k) = C_p(k)$.*

Proof. The proof starts from the construction for C_p with exponentially many children over a ground set of size p and modifies it into a coverage function with polynomially many children while satisfying the desired conditions. For each $z \le k$, replace all children in C_p that are covered by exactly z elements with $\zeta^z(\cdot)$. Define C_p^z to be C_p but only with these children that are covered by exactly z elements. Let the new children from $\zeta^z(\cdot)$ be such that $\zeta^z(k) = C_p^z(k)$.

Clearly c_p has polynomially many children in n since each $\zeta^z(\cdot)$ has polynomially many children. Then, note that

$$c_p(k) = \sum_{z=1}^{k} \zeta^z(k) = \sum_{z=1}^{k} C_p^z(k) = C_p(k).$$

Finally, we show that $c_p(S) = C_p(y)$ for all S such that $|S| = y \le l$. Note that it suffices to show that $\zeta^z(S) = C_p^z(y)$ for all z, which we prove by induction on y. The base case $y = 0$ is trivial. If $y > 0$, then consider some set S such that $|S| = y$ and let $e \in S$. By the inductive hypothesis, $\zeta^z(S \setminus e) = C_p^z(y-1)$. Let T be the set of all children in $\zeta^z(\cdot)$ not covered by $S \setminus e$. Define $\zeta_T^z(\cdot)$ to be $\zeta^z(\cdot)$ but only with children in T. Since all children in T are covered by z distinct elements that are not in $S \setminus e$,

$$\sum_{e' \notin S \setminus e} \zeta_T^z(e') = z \cdot (\zeta^z(k) - \zeta^z(y-1)).$$

By the assumptions on ζ^z, $\zeta^z(S) = \zeta^z(S \setminus e \cup e')$. For any $e' \notin S \setminus e$, by combining with both $\zeta^z(S) = \zeta^z(S \setminus e) + \zeta_T^z(e)$ and $\zeta^z(S \setminus e \cup e') = \zeta^z(S \setminus e) + \zeta_T^z(e')$,

$$\zeta_T^z(e) = \zeta_T^z(e') = z \cdot (\zeta^z(k) - \zeta^z(y-1))/(k - y + 1).$$

So,

$$\begin{aligned}
\zeta^z(S) &= \zeta^z(S \setminus e) + \zeta_T^z(e) \\
&= \zeta^z(y-1) + z \cdot (\zeta^z(k) - \zeta^z(y-1))/(k-y+1) \\
&= C_p^z(y-1) + z \cdot (C_p^z(k) - C_p^z(y-1))/(k-y+1) \\
&= C_p^z(y)
\end{aligned}$$

where the last equality is obtained for C_p^z similarly as how it was obtained for ζ^z. □

We now construct such ζ^z. Assume without loss that k is prime (o.w. pick some prime close to k). Given $\mathbf{a} \in [k]^\ell$, and $x \in [z]$, let $h_{\mathbf{a}}(x) := \sum_{i \in [\ell]} a_i x^i \mod k$. The children in ζ^z are $U = \{\mathbf{a} \in [k]^\ell : h_{\mathbf{a}}(x_1) \neq h_{\mathbf{a}}(x_2)$ for all distinct $x_1, x_2 \in [z]\}$. The k elements are $\{j : 0 \le j < k\}$. Child $\mathbf{a}$ is covered by elements $\{h_{\mathbf{a}}(x) : x \in [z]\}$. Note that $|U| \le k^\ell = 2^{\ell\sqrt{\log n}}$ and we pick $\ell = \log\log n$ as previously. The following lemma is useful to show the symmetricity of ζ^z.

Lemma A.10. *Let* $\mathbf{a}$ *be a uniformly random child, then* $\Pr(h_{\mathbf{a}}(x_1) = j_1, \dots, h_{\mathbf{a}}(x_\ell) = j_\ell)$ *is independent of distinct* $x_1, \dots, x_\ell \in [z]$ *and* $j_1, \dots, j_\ell \in [p]$*. More precisely,*

$$\Pr(h_{\mathbf{a}}(x_1) = j_1, \dots, h_{\mathbf{a}}(x_\ell) = j_\ell) = \prod_{i=1}^{\ell} \frac{1}{p+1-i}.$$

Proof. It is well-known that the random variables $(h_{\mathbf{a}}(0), \dots, h_{\mathbf{a}}(z-1))$ where $\mathbf{a}$ is chosen uniformly at random from $[p]^\ell$ are ℓ-wise independent since $h_{\mathbf{a}}(\cdot)$ is a polynomial of degree $\ell - 1$, so

$$\Pr_{\mathbf{a} \sim \mathcal{U}([p]^l)}(h_{\mathbf{a}}(x_1) = j_1, \dots, h_{\mathbf{a}}(x_\ell) = j_\ell) = \prod_{i=1}^{\ell} \Pr(h_{\mathbf{a}}(x_i) = j_i)\,.$$

By throwing away all children $\mathbf{a}$ such that there exists distinct x_1, x_2 with $h_{\mathbf{a}}(x_1) = h_{\mathbf{a}}(x_2)$, we obtain the following by combining with the symmetry of the children $\mathbf{a}$ removed (there exists exactly one polynomial defined by some $\mathbf{a}$ passing through any collection of ℓ points):

$$\Pr(h_{\mathbf{a}}(x_1) = j_1, \dots, h_{\mathbf{a}}(x_\ell) = j_\ell) = \prod_{i=1}^{\ell} \Pr(h_{\mathbf{a}}(x_i) = j_i | h_{\mathbf{a}}(x_i) \notin \{j_1, \dots, j_{i-1}\})\,.$$

Finally, note that

$$\prod_{i=1}^{\ell} \Pr(h_{\mathbf{a}}(x_i) = j_i | h_{\mathbf{a}}(x_i) \notin \{j_1, \dots, j_{i-1}\}) = \prod_{i=1}^{\ell} \frac{1}{p+1-i}.$$

by the symmetry induced by a_0. □

We are now ready to show the main lemma for the coverage functions $\zeta^z(\cdot)$.

Lemma 2.3. *The coverage function* ζ^z *is symmetric for all sets of size at most* ℓ*.*

Proof. Let $\mathbf{a}$ be a child chosen uniformly at random and S be a set of size at most ℓ. Then, $\zeta^z(S) = k \cdot \Pr(\cup_{j \in S}(\exists x \in [z] \text{ s.t. } h_{\mathbf{a}}(x) = j))$ and

$$\Pr(\cup_{j \in S}(\exists x \in [z] \text{ s.t. } h_{\mathbf{a}}(x) = j)) = \sum_{T \subseteq S} (-1)^{|T|+1} \Pr(T \subseteq \{h_{\mathbf{a}}(x) : x \in [z]\})$$

by inclusion-exclusion. Note that $\Pr(T \subseteq \{h_{\mathbf{a}}(x) : x \in [z]\})$ only depends on the size of T by Lemma A.10. Therefore $\zeta^z(S)$ only depends on the size of S and $\zeta^z(\cdot)$ is symmetric for all sets of size at most ℓ. □

We obtain an $(\alpha = 2^{\Omega(\sqrt{\log n})}, \beta = o(1))$-gap for polynomial sized coverage functions by using the primitives c_p.

Lemma 2.4. *There exists polynomial-sized coverage functions* $g, b, m,$ *and* m^+ *that satisfy an* $(\alpha = 2^{\Omega(\sqrt{\log n})}, \beta = o(1))$*-gap with* $t = n^{3/5+\epsilon}$*.*

Proof. We construct g, b, m, m^+ as in the general case but in terms of primitives c_p instead of C_p. By Lemmas 2.2 and 2.3, we obtain the same matrix A and coefficients $\mathbf{x}^\star$ as in the general case, so the identical on small samples property holds. The masking on large samples holds identically as for general coverage functions. The gap and curvature properties are shown in Lemmas A.5 and A.6. □

We conclude with the main result for coverage functions by combining Claim 2, Lemma 2.4, and Theorem 2.1.

Theorem 2.2. *For every constant $\epsilon > 0$, coverage functions are not $n^{-1/5+\epsilon}$-optimizable from samples over any distribution $\mathcal{D}$. In addition, polynomial-sized coverage functions are not $2^{-\Omega(\sqrt{\log n})}$-optimizable from samples over any distribution $\mathcal{D}$.*

B Algorithms for OPS

OPS via Estimates of Expected Marginal Contributions

We denote by $\mathcal{S}_i$ and $\mathcal{S}_{-i}$ the collections of all samples containing and not containing element e_i respectively. The estimate $\hat{v}_i$ is then the difference in the average value of a sample in $\mathcal{S}_i$ and the average value of a sample in $\mathcal{S}_{-i}$. By standard concentration bounds (Hoeffding's inequality, Lemma B.1), these are good estimates of $\mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e_i)]$ for product distributions $\mathcal{D}$ (Lemma 3.1).

Lemma B.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with values in $[0, b]$. Let $X = \frac{1}{m}\sum_{i=1}^m X_i$ and $\mu = \mathbf{E}[X]$. Then for every $0 < \epsilon < 1$,*

$$\Pr\left(|\bar{X} - \mathbf{E}\left[\bar{X}\right] | \geq \epsilon\right) \leq 2e^{-2m\epsilon^2/b^2}.$$

Lemma 3.1. *Let $\mathcal{D}$ be a product distribution with bounded marginals.*[4] *Then, with probability at least $1 - O(e^{-n})$, the estimations $\hat{v}_i$ are ϵ accurate, for any $\epsilon \geq f(N)/\mathsf{poly}(n)$ and for all e_i, i.e.,*

$$|\hat{v}_i - \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e_i)] \,| \leq \epsilon.$$

Proof. Let $\epsilon \geq f(N)/n^c$ for some constant c. Since $\mathcal{D}$ is a product distribution with marginals bounded away from 0 and 1, there are at least $2n^{2c+1}$ samples containing element e_i and at least $2n^{2c+1}$ samples not containing e_i for all i, with exponentially high probability, with a sufficiently large polynomial number of samples. Then by Hoeffding's inequality (Lemma B.1 with $m = 2n^{2c+1}$ and $b = f(N)$),

$$\Pr\left(\left|\frac{1}{|\mathcal{S}_i|}\sum_{S\in\mathcal{S}_i} f(S) - \mathbf{E}_{S\sim\mathcal{D}|e_i\in S}[f(S)]\right| \geq \epsilon/2\right) \leq 2e^{-4n^{2c+1}(\epsilon/2)^2/f(N)^2} \leq 2e^{-n^{2c+1}/n^{2c}} \leq 2e^{-n}$$

and similarly,

$$\Pr\left(\left|\frac{1}{|\mathcal{S}_{-i}|}\sum_{S\in\mathcal{S}_{-i}} f(S) - \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f(S)]\right| \geq \epsilon/2\right) \leq 2e^{-n}.$$

Since

$$\hat{v}_i = \frac{1}{|\mathcal{S}_i|}\sum_{S\in\mathcal{S}_i} f(S) - \frac{1}{|\mathcal{S}_{-i}|}\sum_{S\in\mathcal{S}_{-i}} f(S)$$

and

$$\mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f_S(e_i)] = \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f(S\cup e_i)] - \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f(S)] = \mathbf{E}_{S\sim\mathcal{D}|e_i\in S}[f(S)] - \mathbf{E}_{S\sim\mathcal{D}|e_i\not\in S}[f(S)]$$

where the second equality is since $\mathcal{D}$ is a product distribution, the claim then holds with probability at least $1 - 4e^{-n}$. □

[4] The marginals are bounded if for all e, $e \in S \sim \mathcal{D}$ and $e \not\in S \sim \mathcal{D}$ w.p. at least $1/\mathsf{poly}(n)$ and at most $1-1/\mathsf{poly}(n)$.

A Tight Approximation for Submodular Functions

Let $\mathcal{D}_i$ be the uniform distribution over all sets of size i. Define the distribution $\mathcal{D}^{sub}$ to be the distribution which draws from $\mathcal{D}_k$, $\mathcal{D}_{\sqrt{n}}$, and $\mathcal{D}_{\sqrt{n}+1}$ at random. This section is devoted to show that Algorithm 2 is an $\tilde{\Omega}(n^{-1/4})$-$\mathtt{OPS}$ algorithm over $\mathcal{D}^{sub}$ (Theorem 3.1). Define $\mathcal{S}_{i,j}$ and $\mathcal{S}_{-i,j}$ to be the collections of samples of size j containing and not containing e_i respectively. For Algorithm 2, we use a slight variation of EEMC($\mathcal{S}$) where the estimates are

$$\hat{v}_i = \frac{1}{|\mathcal{S}_{i,\sqrt{n}+1}|} \sum_{S \in \mathcal{S}_{i,\sqrt{n}+1}} f(S) - \frac{1}{|\mathcal{S}_{-i,\sqrt{n}}|} \sum_{S \in \mathcal{S}_{-i,\sqrt{n}}} f(S).$$

These are good estimates of $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\not\in S}[f_S(e_i)]$, as shown by the following lemma. The proof follows almost identically as the proof for Lemma 3.1.

Lemma B.2. *With probability at least $1 - O(e^{-n})$, the estimations $\hat{v}_i$ defined above are ϵ accurate, for any $\epsilon \geq f(N)/\mathsf{poly}(n)$ and for all e_i, i.e.,*

$$|\hat{v}_i - \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\not\in S}[f_S(e_i)]\,| \leq \epsilon.$$

Proof. Let $\epsilon \geq f(N)/n^c$ for some constant c. With a sufficiently large polynomial number of samples, $\mathcal{S}_{i,\sqrt{n}+1}$ and $\mathcal{S}_{-i,\sqrt{n}}$ are of size at least $2n^{2c+1}$ with exponentially high probability. Then by Hoeffding's inequality (Lemma B.1 with $m = 2n^{2c+1}$ and $b = f(N)$),

$$\Pr\left(\left|\frac{1}{|\mathcal{S}_{i,\sqrt{n}+1}|} \sum_{S\in\mathcal{S}_{i,\sqrt{n}+1}} f(S) - \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}+1}|e_i\in S}[f(S)]\right| \geq \epsilon/2\right) \leq 2e^{-4n^{2c+1}(\epsilon/2)^2/f(N)^2} \leq 2e^{-n}$$

and similarly,

$$\Pr\left(\left|\frac{1}{|\mathcal{S}_{-i,\sqrt{n}}|} \sum_{S\in\mathcal{S}_{-i,\sqrt{n}}} f(S) - \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\not\in S}[f(S)]\right| \geq \epsilon/2\right) \leq 2e^{-n}.$$

By the definition of $\hat{v}_i$ and since

$$\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\not\in S}[f_S(e_i)] = \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}+1}|e_i\in S}[f(S)] - \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}|e_i\not\in S}[f(S)]\,,$$

the claim then holds with probability at least $1 - 4e^{-n}$. □

Next, we show a simple lemma that is useful when returning random sets (Lemma B.3). The analysis is then divided into two cases, depending if a random set $S \sim \mathcal{D}_{\sqrt{n}}$ has low value or not. If a set has low value, then we obtain an $\tilde{\Omega}(n^{-1/4})$-approximation by Corollary 2. Corollary 2 combines Lemmas B.4 and B.5 which respectively obtain $t/(4k\sqrt{n})$ and k/t approximations. If a random set has high value, then we obtain an $n^{-1/4}$-approximation by Corollary 3. Corollary 3 combines Lemmas B.6 and B.7 which respectively obtain $k/(4\sqrt{n})$ and $1/k$ approximations.

Lemma B.3. *For any monotone submodular function $f(\cdot)$, the value of a uniform random set S of size k is a k/n-approximation to $f(N)$.*

Proof. Partition the ground set into sets of size k uniformly at random. A uniform random set of this partition is a k/n-approximation to $f(N)$ in expectation by submodularity. A uniform random set of this partition is also a uniform random set of size k. □

In the first case of the analysis, we assume that $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S)] \leq f(S^\star)/4$. Let j' be the largest j such that bin B_j contains at least one element e_i such that $\hat{v}_i \geq f(S^\star)/(2k)$. So any element $e_i \in B_j$, $j \leq j'$ is such that $\hat{v}_i \geq f(S^\star)/(4k)$. Define $B^\star = \text{argmax}_{B_j : j \leq j'} f(S^\star \cap B_j)$ to be the bin B with high marginal contributions that has the highest value from the optimal solution. Let t be the size of $B^\star$.

Lemma B.4. *If $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S)] \leq f(S^\star)/4$, then a uniformly random subset of bin $B^\star$ of size $\min\{k,t\}$ is a $(1-o(1))\cdot\min(1/4, t/(4k\sqrt{n}))$-approximation to $f(S^\star)$.*

Proof. Note that

$$\begin{aligned}
\mathbf{E}_{S\sim\mathcal{D}_{t/\sqrt{n}}|S\subseteq B^\star}[f(S)] &\geq \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S\cap B^\star)] && \text{(submodularity)}\\
&\geq \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}\left[\sum_{e_i\in S\cap B^\star} f_{(S\cap B^\star)\setminus e_i}(e_i)\right] && \text{(submodularity)}\\
&\geq \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}\left[\sum_{e_i\in S\cap B^\star} \mathbf{E}_{S'\sim\mathcal{D}_{\sqrt{n}}|e_i\notin S'}[f_{S'}(e_i)]\right] && \text{(submodularity)}\\
&\geq \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}\left[\sum_{e_i\in S\cap B^\star} (\hat{v}_i - f(N)/n^3)\right] && \text{(Lemma 3.1)}\\
&= \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[|S\cap B^\star|]\,(1-o(1))\frac{f(S^\star)}{4k} && (\hat{v}_i \geq f(S^\star)/(4k) \text{ for } e_i \in B^\star)\\
&= (1-o(1))\frac{tf(S^\star)}{4k\sqrt{n}}
\end{aligned}$$

If $t/\sqrt{n} \geq k$, then a uniformly random subset of bin $B^\star$ of size k is a $k\sqrt{n}/t$ approximation to $\mathbf{E}_{S\sim\mathcal{D}_{t/\sqrt{n}}|S\subseteq B^\star}[f(S)]$ by Lemma B.3, so a $(1-o(1))/4$ approximation to $f(S^\star)$ by the above inequalities. Otherwise, if $t/\sqrt{n} < k$, then a uniformly random subset of bin $B^\star$ of size $\min\{k,t\}$ has value at least $\mathbf{E}_{S\sim\mathcal{D}_{t/\sqrt{n}}|S\subseteq B^\star}[f(S)]$ by monotonicity, and is thus a $(1-o(1))t/(4k\sqrt{n})$ approximation to $f(S^\star)$.

□

Lemma B.5. *If $\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S)] \leq f(S^\star)/4$, a uniformly random subset of bin $B^\star$ of size $\min(k,t)$ is an $\tilde{\Omega}(\min(1,k/t))$-approximation to $f(S^\star)$.*

Proof. We start by bounding the value of optimal elements not in bin B_j with $j \leq j'$.

$$\begin{aligned}
&f(S^\star \setminus (\cup_{B_j : j\leq j'} B_j))\\
\leq&\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}[f(S \cup S^\star \setminus (\cup_{B_j : j\leq j'} B_j))] && \text{(monotonicity)}\\
\leq&\mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}\left[f(S) + \sum_{e_i\in(S^\star\setminus(\cup_{B_j:j\leq j'}B_j))\setminus S} \mathbf{E}_{S'\sim\mathcal{D}_{\sqrt{n}}|e_i\notin S'}[f_{S'}(e_i)]\right] && \text{(submodularity)}\\
\leq&f(S^\star)/4 + \mathbf{E}_{S\sim\mathcal{D}_{\sqrt{n}}}\left[\sum_{e_i\in(S^\star\setminus(\cup_{B_j:j\leq j'}B_j))\setminus S} (\hat{v}_i + f(N)/n^3)\right] && \text{(assumption and Lemma 3.1)}\\
\leq&f(S^\star)/4 + f(S^\star)/2 + f(S^\star)/n && \text{(definition of } j')
\end{aligned}$$

Since $f(S^\star) \leq f(S^\star \cap (\cup_{B_j : j \leq j'} B_j)) + f(S^\star \setminus (\cup_{B_j : j \leq j'} B_j))$ by submodularity, we get that $f(S^\star \cap (\cup_{B_j : j \leq j'} B_j)) \geq f(S^\star)/5$. Since there are $3 \log n$ bins, $f(S^\star \cap B^\star)$ is a $3 \log n$ to $f(S^\star)/5$ by submodularity and the definition of $B^\star$. By monotonicity, $f(B^\star)$ is a $15 \log n$ to $f(S^\star)$. Thus, by Lemma B.3, a random subset S of size k of bin $B^\star$ is an $\tilde{\Omega}(\min(1, k/t))$ approximation to $f(S^\star)$. □

Corollary 2. *If $\mathbf{E}_{S \sim \mathcal{D}_{\sqrt{n}}}[f(S)] \leq f(S^\star)/4$, a uniformly random subset of size $\min\{k, |B_j|\}$ of a random bin B_j is an $\tilde{\Omega}(n^{-1/4})$-approximation to $f(S^\star)$.*

Proof. With probability $1/(3 \log n)$, the random bin is $B^\star$ and we assume this is the case. By Lemma B.4 and B.5, a random subset of $B^\star$ of size $\min(k, t)$ is both an $\Omega(\min(1, t/(k\sqrt{n})))$ and an $\tilde{\Omega}(\min(1, k/t))$ approximation to $f(S^\star)$. Assume $t/(k\sqrt{n}) \leq 1$ and $k/t \leq 1$, otherwise we are done. Finally, note that if $t/k \geq n^{1/4}$, then $\Omega(t/(k\sqrt{n})) \geq \Omega(n^{-1/4})$, otherwise, $\tilde{\Omega}(k/t) \geq \tilde{\Omega}(n^{-1/4})$. □

In the second case of the proof, we assume that $\mathbf{E}_{S \sim \mathcal{D}_{\sqrt{n}}}[f(S)] \geq f(S^\star)/4$

Lemma B.6. *For any monotone submodular function f, if $\mathbf{E}_{S \sim \mathcal{D}_{\sqrt{n}}}[f(S)] \geq f(S^\star)/4$, then a uniformly random set of size k is a $\min(1/4, k/(4\sqrt{n}))$ approximation to $f(S^\star)$.*

Proof. If $k \geq \sqrt{n}$, then a uniformly random set of size k is a $1/4$-approximation to $f(S^\star)$ by monotonicity. Otherwise, a uniformly random subset of size k of N is a uniformly random subset of size k of a uniformly random subset of size $\sqrt{n}$ of N. So by Lemma B.3,

$$\mathbf{E}_{S \sim \mathcal{D}_k}[f(S)] \geq \frac{k}{\sqrt{n}} \cdot \mathbf{E}_{S \sim \mathcal{D}_{\sqrt{n}}}[f(S)] \geq \frac{k}{4\sqrt{n}} \cdot f(S^\star).$$

□

Lemma B.7. *For any monotone submodular function $f(\cdot)$, the sample S with the largest value among at least $n \log n$ samples of size k is a $1/k$-approximation to $f(S^\star)$ with high probability.*

Proof. By submodularity, there exists an element $e_i^\star$ such that $\{e_i^\star\}$ is a $1/k$-approximation to the optimal solution. By monotonicity, any set which contains $e_i^\star$ is a $1/k$-approximation to the optimal solution. After observing $n \log n$ samples, the probability of never observing a set that contains $e_i^\star$ is polynomially small. □

Corollary 3. *If $\mathbf{E}_{S \sim \mathcal{D}_{\sqrt{n}}}[f(S)] \geq f(S^\star)/4$, then the sample of size k with the largest value is a $\min(1/4, n^{-1/4}/4)$ approximation to $f(S^\star)$.*

Proof. By B.6 and Lemma B.7, the sample of size k with the largest value k is a $\min(1/4, k/(4\sqrt{n}))$ and a $1/k$ approximation to $f(S^\star)$. If $k \geq n^{1/4}$, then $\min(1/4, k/(4\sqrt{n})) \geq \min(1/4, n^{-1/4}/4)$, otherwise, $1/k \geq 1/n^{-1/4}$. □

By combining Corollaries 2 and 3, we obtain the main result for this section.

Theorem 3.1. *Algorithm 2 is an $\tilde{\Omega}(n^{-1/4})$-OPS algorithm over $\mathcal{D}^{sub}$ for monotone submodular functions.*

Bounded Curvature and Additive Functions

The algorithm MaxMargCont simply returns the k elements with the largest estimate $\hat{v}_i$.

Theorem 3.2. *Let f be a monotone submodular function with curvature c and $\mathcal{D}$ be a product distribution with bounded marginals. Then* MaxMargCont *is a $((1-c)^2 - o(1))$-OPS algorithm.*

Proof. Let $S^\star = \{e_1^\star, \dots, e_k^\star\}$ be the optimal solution and $S = \{e_1, \dots, e_k\}$ be the set returned by Algorithm 3. Let $S_i^\star := \{e_1^\star, \dots, e_i^\star\}$ and $S_i := \{e_1, \dots, e_i\}$. If $e_j \not\in S$, then

$$\begin{aligned}
f_{S_{i-1}}(e_i) &\geq (1-c) \cdot \mathbf{E}_{S \sim \mathcal{D} | e_i \not\in S}[f_S(e_i)] && \text{(curvature)} \\
&\geq (1-c)\hat{v}_i - \frac{f(N)}{n^2} && \text{Lemma 3.1 with } \epsilon = \frac{f(N)}{(1-c)n^2} \\
&\geq (1-c)\hat{v}_j - \frac{f(N)}{n^2} && (e_i \in S, e_j \not\in S \text{ and by Algorithm 3}) \\
&\geq (1-c)^2 f_T(e_j) - \frac{f(N)}{n^2} && \text{(curvature)}
\end{aligned}$$

for any set T which does not contain e_j and we conclude that

$$f(S) = \sum_{i=1}^{k} f_{S_{i-1}}(e_i) \geq (1-c)^2 \left(\sum_{i=1}^{k} f_{S_{i-1}^\star}(e_i^\star) \right) - \frac{kf(N)}{n^2} = ((1-c)^2 - o(1)) f(S^\star).$$

□

C Recoverability

A function f is recoverable for distribution $\mathcal{D}$ if given samples drawn from $\mathcal{D}$, it is possible to output a function $\tilde{f}(\cdot)$ such that for all sets S, $\left(1-1/n^2\right) f(S) \leq \tilde{f}(S) \leq \left(1+1/n^2\right) f(S)$ with high probability over the samples and the randomness of the algorithm.

Theorem 4.1. *If a monotone submodular function f is recoverable over $\mathcal{D}$, then it is $1-1/e-o(1)$-optimizable from samples over $\mathcal{D}$. For additive functions, it is $1-o(1)$-optimizable from samples.*

Proof. We show that the greedy algorithm with $\tilde{f}(\cdot)$ for a recoverable function performs well. The proof follows similarly as the classical analysis of the greedy algorithm. We start with submodular functions and denote by $S_i = \{e_1, \cdots, e_i\}$ the set obtained after the ith iteration. Let $S^\star$ be the optimal solution, then by submodularity,

$$\begin{aligned} f(S^\star) &\leq f(S_{i-1}) + \sum_{e \in S^\star \setminus S_{i-1}} f_{S_{i-1}}(e) \\ &\leq f(S_{i-1}) + \sum_{e \in S^\star \setminus S_{i-1}} \left(\left(\frac{1+1/n^2}{1-1/n^2} \right) f(S_i) - f(S_{i-1}) \right) \end{aligned}$$

where the second inequality follows from $\tilde{f}(S_i) \geq \tilde{f}(S_{i-1} \cup \{e\})$ for all $e \in S^\star \setminus S_{i-1}$ by the greedy algorithm, so $(1+1/n^2) f(S_i) \geq (1-1/n^2) f(S_{i-1} \cup \{e\})$. We therefore get that

$$f(S^\star) \leq (1-k) f(S_{i-1}) + k \left(\frac{1+1/n^2}{1-1/n^2} \right) f(S_i).$$

By induction and similarly as in the analysis of the greedy algorithm, we then get that

$$f(S_k) \geq \left(\frac{1-1/n^2}{1+1/n^2} \right)^k \left(1-(1-1/k)^k\right) f(S^\star).$$

Since

$$\left(\frac{1-1/n^2}{1+1/n^2} \right)^k \geq \left(1 - \frac{2}{n^2}\right)^k \geq 1 - 2k/n^2 \geq 1 - 2/n$$

and $\left(1-(1-1/k)^k\right) \geq 1-1/e$, the greedy algorithm achieves an $(1-1/e-o(1))$-approximation for submodular functions.

For additive functions, let S be the set returned by the greedy algorithm and $\hat{v}_i = \tilde{f}(\{e_i\})$, then

$$f(S) = \sum_{e_i \in S} f(\{e_i\}) \geq \left(\frac{1}{1+1/n^2} \right) \sum_{e_i \in S} \hat{v}_i \geq \left(\frac{1}{1+1/n^2} \right) \sum_{e_i \in S^\star} \hat{v}_i \geq \left(\frac{1-1/n^2}{1+1/n^2} \right) f(S^\star).$$

We therefore get a $(1-o(1))$-approximation for additive functions.

□

Lemma 4.1. *Let f be an additive function with $v_{max} = \max_i f(\{e_i\})$, $v_{min} = \min_i f(\{e_i\})$ and let $\mathcal{D}$ be a product distribution with bounded marginals. If $v_{min} \geq v_{max}/poly(n)$, then f is recoverable for $\mathcal{D}$.*

Proof. We have already shown that the expected marginal contribution of an element to a random set of size $k-1$ can be estimated from samples for submodular functions[5]. In the case of additive functions, this marginal contribution of an element is its value $f(\{e_i\})$.

We apply Lemma 3.1 with $\epsilon = f(\{e_i\})/n^2$ to compute $\hat{v}_i$ such that $|\hat{v}_i - f(\{e_i\})| \leq f(\{e_i\})/n^2$. Note that $\epsilon = f(\{e_i\})/n^2$ satisfies $\epsilon \geq f(S^\star)/\mathsf{poly}(n)$ since $v_{min} \geq v_{max}/\mathsf{poly}(n)$. Let $\tilde{f}(S) = \sum_{e_i \in S} \hat{v}_i$, then

$$\tilde{f}(S) \leq \sum_{e_i \in S} (1 + 1/n^2) f(\{e_i\}) = (1 + 1/n^2) f(S)$$

and

$$\tilde{f}(S) \geq \sum_{e_i \in S} (1 - 1/n^2) f(\{e_i\}) = (1 - 1/n^2) f(S).$$

□

Lemma 4.2. *Unit demand functions are not recoverable for $k \geq n^\epsilon$ but are 1-OPS.*

Proof. We first show that unit demand functions are not recoverable. Define a hypothesis class of functions $\mathcal{F}$ which contains n unit demand functions $f_j(\cdot)$ with $f(\{e_1\}) = j/n$ and $f(\{e_i\}) = 1$ for $i \geq 2$, for all integers $1 \leq j \leq n$. We wish to recover function $f_j(\cdot)$ with j picked uniformly at random. With high probability, the sample $\{e_1\}$ is not observed when $k \geq n^\epsilon$, so the values of all observed samples are independent of j. Unit demand functions are therefore not recoverable.

Unit demand functions, on the other hand, are 1-optimizable from samples. With at least $n \log n$ samples, at least one sample contains, with high probability, the best element $e^\star := \operatorname{argmax}_{e_i} f(\{e_i\})$. Any set containing the best element is an optimal solution. Therefore, an algorithm which returns the sample with the largest value obtains an optimal solution with high probability. □

[5] For simplicity, this proof uses estimations that we know how to compute. However, The values $f(\{e_i\})$ can be recovered exactly by solving a system of linear equations where each row corresponds to a sample, provided that the matrix for this system is invertible, which is the case with a sufficiently large number of samples by using results from random matrix theory such as in the survey by Blake and Studholme [7].

D Learning Models

As a model for statistical learnability we use the notion of PAC learnability due to Valiant [58] and its generalization to real-valued functions PMAC learnability, due to Balcan and Harvey [4]. Let $\mathcal{F}$ be a hypothesis class of functions $\{f_1, f_2, \ldots\}$ where $f_i : 2^N \to \mathbb{R}$. Given precision parameters $\epsilon > 0$ and $\delta > 0$, the input to a learning algorithm is samples $\{S_i, f(S_i)\}_{i=1}^m$ where the S_i's are drawn i.i.d. from from some distribution $\mathcal{D}$, and the number of samples m is polynomial in $1/\epsilon, 1/\delta$ and n. The learning algorithm outputs a function $\widetilde{f} : 2^N \to \mathbb{R}$ that should approximate f in the following sense.

- $\mathcal{F}$ is PAC-learnable on distribution $\mathcal{D}$ if there exists a (not necessarily polynomial time) learning algorithm such that for every $\epsilon, \delta > 0$:

$$\Pr_{S_1,\ldots,S_m \sim \mathcal{D}} \left[\Pr_{S \sim \mathcal{D}} \left[\widetilde{f}(S) \neq f(S) \right] \geq 1 - \epsilon \right] \geq 1 - \delta$$

- $\mathcal{F}$ is α-PMAC-learnable on distribution $\mathcal{D}$ if there exists a (not necessarily polynomial time) learning algorithm such that for every $\epsilon, \delta > 0$:

$$\Pr_{S_1,\ldots,S_m \sim \mathcal{D}} \left[\Pr_{S \sim \mathcal{D}} \left[\alpha \cdot \widetilde{f}(S) \leq f(S) \leq \widetilde{f}(S) \right] \geq 1 - \epsilon \right] \geq 1 - \delta$$

A class $\mathcal{F}$ is PAC (or α-PMAC) learnable if it is PAC- (α-PMAC)-learnable on every distribution $\mathcal{D}$.

E Discussion

Beyond set functions. Thinking about models as set functions is a useful abstraction, but optimization from samples can be considered for general optimization problems. Instead of the max-k-cover problem, one may ask whether samples of spanning trees can be used for finding an approximately minimum spanning tree. Similarly, one may ask whether shortest paths, matching, maximal likelihood in phylogenetic trees, or any other problem where crucial aspects of the objective functions are learned from data, is optimizable from samples.

Coverage functions. In addition to their stronger learning guarantees, coverage functions have additional guarantees that distinguish them from general monotone submodular functions.

- Any polynomial-sized coverage function can be exactly *recovered*, i.e., learned exactly for *all* sets, using polynomially many (adaptive) queries to a value oracle [10]. In contrast, there are monotone submodular functions for which no algorithm can recover the function using fewer than exponentially many value queries [10]. It is thus interesting that despite being a distinguished class within submodular functions with enough structure to be exactly recovered via adaptive queries, polynomial-sized coverage functions are inapproximable from samples.
- In mechanism design, one seeks to design polynomial-time mechanisms which have desirable properties in equilibrium (e.g. truthful-in-expectation). Although there is an impossibility result for general submodular functions [23], one can show that for coverage functions there is a mechanism which is truthful-in-expectation [22, 24].

F Hardness of Submodular Functions

Using the hardness framework from Section 2.1, it is relatively easy to show that submodular functions are not $n^{-1/4+\epsilon}$-$\texttt{OPS}$ over any distribution $\mathcal{D}$. The good, bad, and masking functions g, b, m, m^+ we use are:

$$\begin{aligned} g(S) &= |S|, \\ b(S) &= \min(|S|, \log n), \\ m(S) &= \min(1, |S|/n^{1/2}), \\ m^+(S) &= n^{-1/4} \cdot \min(n^{1/2}, |S|). \end{aligned}$$

It is easy to show that $\mathcal{F}(g, b, m, m^+)$ is a class of monotone submodular functions (Lemma F.2). To derive the optimal $n^{-1/4+\epsilon}$ impossibility we consider the cardinality constraint $k = n^{1/4-\epsilon/2}$ and the size of the partition to be $r = n^{1/4}$. We show that $\mathcal{F}(g, b, m, m^+)$ has an $(n^{1/4-\epsilon}, 0)$-gap.

Lemma F.1. *The class $\mathcal{F}(g, b, m, m^+)$ as defined above has an $(n^{1/4-\epsilon}, 0)$-gap with $t = n^{1/2+\epsilon/4}$.*

Proof. We show that these functions satisfy the properties to have an $(n^{1/4-\epsilon}, 0)$-gap.

- **Identical on small samples.** Assume $|S| \leq n^{1/2+\epsilon/4}$. Then $|T_{-i}| \cdot |S|/n \leq n^{1/2-\epsilon/2} \cdot n^{1/2+\epsilon/4}/n \leq n^{-\epsilon/4}$, so by Lemma A.2, $|S \cap T_{-i}| \leq \log n$ w.p. $1 - \omega(1)$ over $P \sim \mathcal{U}(\mathcal{P})$. Thus

  $$g(S \cap T_i) + b(S \cap T_{-i}) = |S \cap (\cup_{j=1}^r T_j)|$$

  with probability $1 - \omega(1)$ over P.

- **Identical on large samples.** Assume $|S| \geq n^{1/2+\epsilon/4}$. Then $|S \cap M| \geq n^{1/2}$ with exponentially high probability over $P \sim \mathcal{U}(\mathcal{P})$ by Chernoff bound (Lemma A.7), and $m(S \cap M) = 1$ w.p. at least $1 - \omega(1)$.

- **Gap $n^{1/4-\epsilon}$.** Note that $g(S) = k = n^{1/4-\epsilon/2}$, $b(S) = \log n$, $m^+(S) = n^{-\epsilon/2}$ for $|S| = k$, so $g(S) \geq n^{1/4-\epsilon} b(S)$ for n large enough and $g(S) = n^{1/4} m^+(S)$.

- Curvature $\beta = 0$. The curvature $\beta = 0$ follows from g being linear.

□

We show that that we obtain monotone submodular functions.

Lemma F.2. *The class of functions $\mathcal{F}(g, b, m, m^+)$ is a class of monotone submodular functions.*

Proof. We show that the marginal contributions $f_S(e)$ of an element $e \in N$ to a set $S \subseteq N$ are such that $f_S(e) \geq f_T(e)$ for $S \subseteq T$ (submodular) and $f_S(e) \geq 0$ for all S (monotone) for all elements e. For $e \in T_j$, for all j, this follows immediately from g and b being monotone submodular. For $e \in M$, note that

$$f_S(e) = \begin{cases} -\frac{1}{n^{1/2}}(|S \cap T_i| + \min(|S \cap T_{-i}|, \log n)) + n^{-1/4} & \text{if } |S \cap M| < n^{1/2} \\ 0 & \text{otherwise} \end{cases}$$

Since $|S \cap T_i| + \min(|S \cap T_{-i}|, \log n) \leq n^{1/4}$, $f_S(e) \geq f_T(e)$ for $S \subseteq T$ and $f_S(e) \geq 0$ for all S. □

Together with Theorem 2.1, these two lemmas imply the hardness result.

Theorem F.1. *For every constant $\epsilon > 0$, monotone submodular functions are not $n^{-1/4+\epsilon}$-$\texttt{OPS}$ over any distribution $\mathcal{D}$.*