Optimization for Approximate Submodularity

Avinatan Hassidim
Bar Ilan University and Google
avinatan@cs.biu.ac.il

Yaron Singer
Harvard University
yaron@seas.harvard.edu

Abstract

We consider the problem of maximizing a submodular function when given access to its approximate version. Submodular functions are heavily studied in a wide variety of disciplines since they are used to model many real world phenomena and are amenable to optimization. There are many cases however in which the phenomena we observe is only *approximately* submodular and the optimization guarantees cease to hold. In this paper we describe a technique that yields strong guarantees for maximization of monotone submodular functions from approximate surrogates under cardinality and intersection of matroid constraints. In particular, we show tight guarantees for maximization under a cardinality constraint and $1/(1+P)$ approximation under intersection of P matroids.

1 Introduction

In this paper we study maximization of *approximately* submodular functions. For nearly half a century submodular functions have been extensively studied since they are amenable to optimization and are recognized as an effective modeling instrument. In machine learning, submodular functions capture a variety of objectives that include entropy, diversity, image segmentation, and clustering.

Although submodular functions are used to model real world phenomena, in many cases the functions we encounter are only *approximately* submodular. This is either due to the fact that the objectives are not exactly submodular, or alternatively they are, but we only observe their approximate version.

In the literature, approximate utility functions are modeled as surrogates of the original function that have been corrupted by random variables drawn i.i.d from some distribution. Some examples include:

- **Revealed preference theory.** Luce's famous model assumes that an agent's revealed utility $\tilde{f} : 2^N \to \mathbb{R}$ can be approximated by a well-behaved utility function $f : 2^N \to \mathbb{R}$ s.t. $\tilde{f}(S) = f(S) + \xi_S$ for every $S \subseteq N$ where ξ_S is drawn i.i.d from a distribution [CE16]. $\tilde{f}$ is a utility function that approximates f, and multiple queries to $\tilde{f}$ return the same response;
- **Statistics and learning theory.** The assumption in learning is that the data we observe is generated by $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \xi_{\mathbf{x}}$ where f is in some well-behaved hypothesis class and $\xi_{\mathbf{x}}$ is drawn i.i.d from some distribution. The use of $\tilde{f}$ is not to model corruption by noise but rather the fact that data is not *exactly* manufactured by a function in the hypothesis class;
- **Active learning.** There is a long line of work on noise-robust learning where one has access to a noisy membership oracle $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \xi_{\mathbf{x}}$ and for every $\mathbf{x}$ we have that $\xi_{\mathbf{x}}$ is drawn i.i.d from a distribution [Ang88, GKS90, Jac94, SS95, BF02, Fel09]. In this model as well, the oracle is consistent and multiple queries return the same response. For set functions, one can consider active learning in experimental design applications where the objective function is often submodular and the goal would be to optimize $f : 2^N \to \mathbb{R}$ given $\tilde{f}$.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Similar to the above examples, we say that a function $\tilde{f} : 2^N \to \mathbb{R}_+$ is *approximately submodular* if there is a submodular function $f : 2^N \to \mathbb{R}_+$ and a distribution $\mathcal{D}$ s.t. for each set $S \subseteq N$ we have that $\tilde{f}(S) = \xi_S f(S)$ where ξ_S is drawn i.i.d from $\mathcal{D}$.[1] The modeling assumption that ξ_S is not adversarially chosen but drawn i.i.d is crucial. Without this assumption for $\tilde{f}(S) = \xi_S f(S)$ where $\xi_S \in [1-\epsilon, 1+\epsilon]$ even for subconstant $\epsilon > 0$ no algorithm can obtain an approximation strictly better than $n^{-1/2}$ to maximizing either $\tilde{f}$ or f under a cardinality constraint when $n = |N|$ [HS17]. This hardness stems from the fact that approximate submodularity implies that the function is close to submodular, but its marginals (or gradients) are not well approximated by those of the submodular function. When given an α approximation to the marginals, the greedy algorithm produces a $1 - 1/e^{\alpha}$ approximation. Furthermore, even for *continuous* submodular functions, a recent line of work shows that gradient methods produce strong guarantees with approximate gradient information of the function [HSK17, CHHK18, MHK18].

In contrast to the vast literature on submodular optimization, optimization of approximate submodularity is nascent. For distributions bounded in $[1 - \epsilon/k, 1 + \epsilon/k]$ the function is sufficiently close to submodular for the approximation guarantees of the greedy algorithm to go through [HS, LCSZ17, QSY$^+$17]. Without this assumptions, the greedy algorithm performs arbitrarily poorly. In [HS17] the authors give an algorithm that obtains an approximation arbitrarily close to $1 - 1/e$ under a cardinality constraint that is sufficiently large. For arbitrary cardinality and general matroid constraints there are no known approximation guarantees.

1.1 Our Contribution

In this paper we consider the problem $\max_{S \in \mathcal{F}} f(S)$ when $f : 2^N \to \mathbb{R}$ is non-negative monotone submodular defined of a ground set N of size n and the algorithm is only given access to an approximate surrogate $\tilde{f}$ and $\mathcal{F}$ is a uniform matroid (cardinality) or intersection of matroids constraint. We introduce a powerful technique which we call the *sampled mean approximation* and show:

- **Optimal guarantees for maximization under a cardinality constraint.** In [HS17] the result gives an approximation arbitrarily close to $1 - 1/e$ for $k \in \Omega(\log\log n)$. This is a fundamental limitation in their technique that initializes the solution with a set of size $\Omega(\log\log n)$ used for "smoothing" the approximately submodular function (see Appendix G.1 for more details). The technique in this paper is novel and yields an approximation of $1 - 1/e$ for any $k \geq 2$, and $1/2$ for $k = 1$, which is information theoretically tight, as we later show;
- $1/(1+P)$ **approximation for intersection of** P **matroids.** We utilize the sampled mean approximation method to produce the first results for the more challenging case of maximization under general matroid constraints. Our approximation guarantees are comparable with those achievable with a greedy algorithm for monotone submodular functions;
- **Information theoretic lower bounds.** We show that no randomized algorithm can obtain an approximation strictly better than $1/2 + \mathcal{O}(n^{-1/2})$ for $\max_{a \in N} f(a)$ given an approximately submodular oracle $\tilde{f}$, and that no randomized algorithm can obtain an approximation strictly better than $(2k-1)/2k + \mathcal{O}(n^{-1/2})$ for maximization under cardinality constraint k;
- **Bounds in Extreme Value Theory.** As we later discuss, some of our results may be of independent interest to Extreme Value Theory (EVT) which studies the bounds on the maximum sample (or the top samples) from some distribution. To achieve our main result we prove subtle properties about extreme values of random variables where not all samples are created equal and the distributions generalize those typically studied in EVT.

The results above are for the problem $\max_{S \in \mathcal{F}} f(S)$ when the algorithm is given access to $\tilde{f}$. In some applications, however, $\tilde{f}$ is the function that we actually wish to optimize, i.e. our goal is to solve $\max_{S \in \mathcal{F}} \tilde{f}(S)$. If $\tilde{f}(S)$ approximates $f(S)$ well on all sets S, we can use the solution for $\max_{S \in \mathcal{F}} f(S)$ as a solution for $\max_{S \in \mathcal{F}} \tilde{f}(S)$. In general, however, a solution that is good for $\max_{S \in \mathcal{F}} f(S)$ can be arbitrarily bad for $\max_{S \in \mathcal{F}} \tilde{f}(S)$. In Appendix E we give a black-box reduction showing that these problems are essentially equivalent. Specifically, we show that given a solution

[1] Describing $\tilde{f}$ as a multiplicative approximation of f is more convenient for analyzing multiplicative approximation guarantees. This is w.l.o.g as all our results apply to additive approximations as well.

to $\max_{S\in\mathcal{F}} f(S)$ one can produce a solution that is of arbitrarily close quality to $\max_{S\in\mathcal{F}} \tilde{f}(S)$ when $\mathcal{F}$ is any uniform matroid, an intersection of matroids of rank $\Omega(\log n)$, and an intersection of matroids of any rank when the distribution $\mathcal{D}$ has bounded support.

1.2 Technical Overview

The approximately submodular functions we consider approximate a submodular function f using samples from a distribution $\mathcal{D}$ of the class of *generalized exponential tail* distributions defined as:

Definition. *A noise distribution $\mathcal{D}$ has a* ***generalized exponential tail*** *if there exists some x_0 such that for $x > x_0$ the probability density function $\rho(x) = e^{-g(x)}$, where $g(x) = \sum_i a_i x^{\alpha_i}$ for some (not necessarily integers) $\alpha_0 \geq \alpha_1 \geq \ldots$, s.t. $\alpha_0 \geq 1$. If $\mathcal{D}$ has bounded support we only require that either it has an atom at its supremum, or that ρ is continuous and non-zero at the supremum.*

This class of distributions is carefully defined. On the one hand it is general enough to contain Gaussian and Exponential distributions, as well as any distribution with bounded support. On the other hand it has enough structure one can leverage. Note that optimization in this setting always requires that the support is independent of n and that n is sufficiently large [2]. Throughout the paper we assume that $\mathcal{D}$ has a generalized exponential tail and that n is sufficiently large.

Theorem. *For any non-negative monotone submodular function there is a deterministic polynomial-time algorithm which optimizes the function under a cardinality constraint $k \geq 3$ and obtains an approximation ratio that is arbitrarily close to $1 - 1/e$ with probability $1 - o(1)$ using access to an approximate oracle. For $k \geq 2$ there is a randomized algorithm whose approximation ratio is arbitrarily close to $1 - 1/e$, in expectation over the randomization of the algorithm. For $k = 1$ the algorithm achieves a $1/2$ approximation in expectation, and no randomized algorithm can achieve an approximation better than $1/2 + o(1)$, in expectation.*

The main part of the proof involves analysis of the following greedy algorithm. The algorithm iteratively chooses *bundles* of elements of size $\mathcal{O}(1/\epsilon)$. In each iteration,the algorithm first identifies a bundle $\mathbf{x}$ whose addition to the current solution approximately maximizes the *approximate mean value* $\tilde{F}$. Informally, $\tilde{F}(\mathbf{x})$ is the average value of $\tilde{f}$ evaluated on all bundles at Hamming distance one from $\mathbf{x}$. Then, the algorithm does not choose $\mathbf{x}$ but rather the bundle at Hamming distance one from $\mathbf{x}$ whose addition to the current solution maximizes the approximate submodular value $\tilde{f}$.

The major technical challenge is in analyzing the regime in which $k \in \Omega(1/\epsilon^2) \cap \mathcal{O}(\sqrt{\log n})$. At a high level, in this regime the analysis relies on showing that the marginal contribution of the bundle of elements selected in every iteration is approximately largest. Doing so requires proving subtle properties about extreme values of random variables drawn from the generalized exponential tail distribution, and the analysis fully leverages the properties of the distribution and the fact that $k \in \mathcal{O}(\sqrt{\log n})$. This is of independent interest to Extreme Value Theory (EVT) which tries to bound the maximum sample (or the top samples) from some distribution. If we would consider the constant function $f(S) = 1$ for $S \neq \emptyset$, and try to maximize an approximate version of f with respect to some distribution, this would be a classical EVT setting. One can view the bounds we develop as bounds on a generalization of EVT, where not all samples are created equal.

For general matroid constraints we apply the sampled-mean technique and obtain an approximation comparable to that of applying the greedy algorithm on a monotone submodular function.

Theorem. *For any non-negative monotone submodular function there is a deterministic polynomial-time algorithm which optimizes the function under an intersection of P matroids constraint and obtains an approximation ratio arbitrarily close to $1/(1 + P)$ given an approximate oracle.*

Paper organization. The main contribution of the paper is the definition of sampled mean approximation in Section 2 and the subsequent analysis of the algorithm for cardinality constraint in Section 3 and matroid constraints in Section 4. The techniques are novel, and the major technical crux of the paper is in analyzing the algorithm in Section 3. Optimization for small rank and lower bounds are in Appendix B.1. In Appendix G we further discuss related work, and in Appendix F we discuss extensions of the algorithms to related models.

[2] For example, if for every S the noise is s.t. $\xi_S = 2^{100}$ w.p. $1/2^{100}$ and 0 otherwise, but $n = 50$, it is likely that the oracle will always return 0, in which case we cannot do better than selecting an element at random.

2 The Sampled Mean Approximation

We begin by defining the sampled mean approximation of a function. This approach considers *bundles* $\mathbf{x}$ of size $c \in \mathcal{O}(1/\epsilon)$. We can then run a variant of a greedy algorithm which adds a bundle of size c in every iteration. For a given bundle $\mathbf{x}$ and set S, we define a *ball* to be all bundles obtained by swapping a single element from $\mathbf{x}$ with another not in $S \cup \mathbf{x}$. We denote $\mathbf{x}_{ij} = (\mathbf{x} \setminus \{x_i\}) \cup \{x_j\}$.

Definition. *For $S \subset N$ and bundle $\mathbf{x} \subset N$, the ball around $\mathbf{x}$ is $\mathcal{B}_S(\mathbf{x}) := \{\mathbf{x}_{ij} : i \in \mathbf{x}, j \notin S \cup \mathbf{x}\}$.*

We illustrate a ball in Figure 1. Notice that as long as $|S| \leq (1-\delta)|N|$ for some fixed $\delta > 0$, we have that $|\mathcal{B}_S(\mathbf{x})| \in \Omega(n)$. This will allow us to derive weak (but sufficient) concentration bounds.

Definition. *Let $f : 2^N \to \mathbb{R}$. For a set $S \subseteq N$ and bundle $\mathbf{x} \subseteq N$, the **mean value**, **noisy mean value**, and **mean marginal contribution** of $\mathbf{x}$ given S are, respectively:*

$$\begin{array}{lll} (1) & F(S \cup \mathbf{x}) := & \mathbb{E}_{\mathbf{z} \sim \mathcal{B}_S(\mathbf{x})}\left[\; f(S \cup \mathbf{z}) \;\right] \\ (2) & \tilde{F}(S \cup \mathbf{x}) := & \mathbb{E}_{\mathbf{z} \sim \mathcal{B}_S(\mathbf{x})}\left[\; \tilde{f}(S \cup \mathbf{z}) \;\right] \\ (3) & F_S(\mathbf{x}) := & \mathbb{E}_{\mathbf{z} \sim \mathcal{B}_S(\mathbf{x})}\left[\; f_S(\mathbf{z}) \;\right] \end{array}$$

The following lemma implies that under the right conditions, the bundle that maximizes the noisy mean value is a good approximation of the bundle whose (non-noisy) mean value is largest.

Lemma 2.1. *Let $\mathbf{x} \in \operatorname{argmax}_{\mathbf{z}:|\mathbf{z}|=c} \tilde{F}(S \cup \mathbf{z})$ where $c = 2/\epsilon$. Then, w.p. $\geq 1 - \exp\left(-\Omega\left(n^{1/4}\right)\right)$:*

$$F_S(\mathbf{x}) \geq (1-\epsilon) \max_{\mathbf{z}:|\mathbf{z}|=c} F_S(\mathbf{z}).$$

The above lemma gives us a *weak* concentration bound in the following sense. While it is generally not true that $\tilde{F}(S \cup \mathbf{z}) \approx F(S \cup \mathbf{z})$, we can upper bound $\tilde{F}(S \cup \mathbf{z})$ in a meaningful way and show that $\tilde{F}(S \cup \mathbf{x}^\star) \approx F(S \cup \mathbf{x}^\star)$ for $\mathbf{x}^\star \in \operatorname{argmax}_{\mathbf{z}} f_S(\mathbf{z})$ by using submodularity. This allows us to show that the mean marginal contribution of $\mathbf{x}$ that maximizes the *noisy* mean value is an ϵ-approximation to the maximal (non-noisy) mean marginal contribution. Details and proofs are in Appendix A.

In addition to approximating the mean marginal contribution given a noisy oracle, an important property of the sampled-mean approach is that it well-approximates its true marginal contribution.

Lemma 2.2. *For any $\epsilon > 0$ and any set $S \subset N$, let $\mathbf{x}$ be a bundle of size $1/\epsilon$, then:*

$$F_S(\mathbf{x}) \geq (1-\epsilon) f_S(\mathbf{x}),$$

The proof is in Appendix A and exploits a natural property of submodular functions: the removal of a random element from a sufficiently large set does not significantly affect its value, in expectation.

Let $\mathbf{x}^\star \in \operatorname{argmax}_{\mathbf{b}:|\mathbf{b}|=c} f_S(\mathbf{b})$. Lemma 2.1 and Lemma 2.2 together imply the following corollary:

Corollary 2.3. *For a fixed $\epsilon > 0$ let $c = 3/\epsilon$, and $\mathbf{x} \in \operatorname{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(S \cup \mathbf{b})$. Then, w.p. at least $1 - \exp(-\Omega(n^{1/4}))$ we have that: $F_S(\mathbf{x}) \geq (1-\epsilon) f_S(\mathbf{x}^\star)$.*

At a first glance, it may seem as if running the greedy algorithm with F instead of f suffices. The problem, however, is that the *mean* marginal contribution F_S may be an unbounded overestimate of the *true* marginal contribution f_S. Consider for example an instance where $S = \emptyset$ and there is a bundle $\mathbf{x}$ of size c s.t. for every $\emptyset \neq \mathbf{z} \subseteq \mathbf{x}$ we have $f(\mathbf{z}) = \delta$ for some arbitrarily small δ, while every other subset $T \subseteq N \setminus \mathbf{x}$ is complementary to $\mathbf{x}$ and has some arbitrarily large value M. In this case, $\mathbf{x} = \operatorname{argmax}_{\mathbf{z}:|\mathbf{z}|=c} F(\mathbf{z})$ and $F(\mathbf{x}) = M + \delta$ while $f(\mathbf{x}) = \delta$.

3 The Sampled Mean Greedy for Cardinality Constraints

The SM-GREEDY begins with the empty set S and at every iteration considers all bundles (sets) of size $c \in \mathcal{O}(1/\epsilon)$ to add to S. At every iteration, the algorithm first identifies the bundle $\mathbf{x}$ which maximizes the noisy mean value. After identifying $\mathbf{x}$, it then considers all possible bundles $\mathbf{z} \in \mathcal{B}_S(\mathbf{x})$ and takes the one whose noisy value is largest. We include a formal description below.

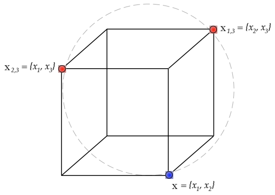

Figure 1: Illustration of a ball around $\mathbf{x} = \{x_1, x_2\}$ where $N = \{x_1, x_2, x_3\}$, and $S = \emptyset$. We think of $\mathbf{x}$ as a point in $[0,1]^3$ and $\mathcal{B}_S(\mathbf{x}) = \{\mathbf{x}_{12}, \mathbf{x}_{23}\} = \{\{x_2, x_3\}, \{x_1, x_3\}\} = \{(0,1,1), (1,0,1)\}$.

Algorithm 1 SM-GREEDY

```
Input: budget $k$, precision $\epsilon > 0$, $c \leftarrow \frac{56}{\epsilon}$
$S \leftarrow \emptyset$
while $|S| < c \cdot \lfloor \frac{k}{c} \rfloor$ do
    $\mathbf{x} \leftarrow \text{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(S \cup \mathbf{b})$
    $\hat{\mathbf{x}} \leftarrow \text{argmax}_{\mathbf{z} \in \mathcal{B}(\mathbf{x})} \tilde{f}(S \cup \mathbf{z})$
    $S \leftarrow S \cup \hat{\mathbf{x}}$
end while
return $S$
```

A key step in analyzing greedy algorithms like the one above is showing that in every iteration the marginal contribution of the element selected by the algorithm is arbitrarily close to maximal. This can then be used in a standard inductive argument to show that the algorithm obtains an approximation arbitrarily close to $1 - 1/e$. The main crux of the analysis is showing that this property indeed holds in SM-GREEDY w.h.p. when $k \in \Omega(1/\epsilon^2) \cap \mathcal{O}(\sqrt{\log n})$. For $k > \sqrt{\log n}$ the analysis become substantially simpler since it suffices to argue that the algorithm chooses an element whose marginal contribution is approximately optimal in expectation (details are in the proof of Theorem 3.5).

3.1 Analysis for $k \in \Omega(1/\epsilon^2) \cap \mathcal{O}(\sqrt{\log n})$

Throughout the rest of this section we will analyze a single iteration of SM-GREEDY in which a set $S \subset N$ was selected in previous iterations and the algorithm adds a bundle $\hat{\mathbf{x}}$ of size c. Specifically, $\hat{\mathbf{x}} \in \text{argmax}_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} \tilde{f}(S \cup \mathbf{z})$ where $\mathbf{x} \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(S \cup \mathbf{b})$.

As discussed above, for $\mathbf{x}^\star \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=c} f_S(\mathbf{b})$ we want to show that when $c \in \mathcal{O}(1/\epsilon)$:

$$f_S(\hat{\mathbf{x}}) \geq (1 - \epsilon) f_S(\mathbf{x}^\star).$$

To do so, we will define two kinds of bundles in $\mathcal{B}_S(\mathbf{x})$, called **good** and **bad**. A good bundle is a bundle $\mathbf{z}$ for which $f_S(\mathbf{z}) \geq (1 - \frac{2}{3}\epsilon) f_S(\mathbf{x}^\star)$ and a bad bundle $\mathbf{z}$ is one s.t. $f_S(\mathbf{z}) \leq (1 - \epsilon) f_S(\mathbf{x}^\star)$. Our goal is to prove that the bundle $\hat{\mathbf{x}}$ added by the algorithm is w.h.p. not bad. Since according to the definition of good and bad the true marginal contribution of good bundles has a fixed advantage over the true marginal contribution of bad bundles, and $\hat{\mathbf{x}}$ is the bundle with largest *noisy* value, essentially what we need to show is that the largest noise multiplier of a good bundle is sufficiently close to the largest noise multiplier of a bad bundle, with sufficiently high probability.

As a first step we quantify the fraction of good bundles in a ball, which will then allow us to bound the values of the noise multipliers of good and bad bundles. The following claim implies at least half of the bundles in $\mathcal{B}_S(\mathbf{x})$ are good and at most half are bad (the proof is in Appendix B).

Claim 3.1. *Suppose $F_S(\mathbf{x}) \geq \left(1 - \frac{\epsilon}{3}\right) f_S(\mathbf{x}^\star)$. Then at least half of the bundles in $\mathcal{B}_S(\mathbf{x})$ are good.*

We next define two thresholds ξ_{inf} and ξ_{sup}. The threshold ξ_{inf} is used as a lower bound on the largest value obtained when sampling *at least* $\frac{|B_S(\mathbf{x})|}{2}$ random variables from the noise distribution. Since at least half of the bundles in the ball are good, ξ_{inf} is a lower bound on the largest noise multiplier of a good bundle. The threshold ξ_{sup} is used as an upper bound on the largest value obtained when sampling *at most* $\frac{|B_S(\mathbf{x})|}{2}$ random variables from the noise distribution. Since at most half of the bundles in the ball are bad, ξ_{sup} upper bounds the value of a noise multiplier of a bad bundle. Throughout the rest of this section $\mathcal{D}$ will denote a generalized exponential tail distribution.

Definition. *Let $m = \frac{|\mathcal{B}_S(\mathbf{x})|}{2}$. For probability density function $\rho(x)$ of $\mathcal{D}$ we define:*

- *$\xi_{\sup}$ is the value for which: $\int_{\xi_{\sup}}^{\infty} \rho(x)dx = \frac{2}{m \log n}$;*
- *$\xi_{\inf}$ is the value for which: $\int_{\xi_{\inf}}^{\infty} \rho(x)dx = \frac{2 \log n}{m}$.*

Claim 3.2. *Let $m = \frac{|\mathcal{B}_S(\mathbf{x})|}{2}$, $\xi_1, \dots, \xi_m$ be i.i.d samples of $\mathcal{D}$ and $\xi^\star = \max\{\xi_1, \dots, \xi_m\}$. Then:*

$$\Pr\left[\xi_{\inf} \leq \xi^\star \leq \xi_{\sup}\right] \geq 1 - \frac{3}{\log n}$$

The proof is in Appendix B. Since at least half of the bundles in the ball are good and at most half are bad, the above claim implies that with probability $1 - 3/\log n$ the maximal value of noise multipliers of good and bad bundles fall within the range $[\xi_{\inf}, \xi_{\sup}]$. Since this holds in every iteration, when $k \in O(\sqrt{\log n})$ by a union bound we get that this holds in all iterations w.p. $1 - o(1)$.

At this point, our lower bound on the largest noisy value of a good bundle is:

$$\max_{\mathbf{z} \in \mathtt{good}} \tilde{f}(S \cup \mathbf{z}) \geq \xi_{\inf} \times \left(f(S) + \left(1 - \frac{2}{3}\epsilon\right) f_S(\mathbf{x}^\star) \right)$$

and the upper bound on the noisy value of any bad bundle is:

$$\max_{\mathbf{z} \in \mathtt{bad}} \tilde{f}(S \cup \mathbf{z}) \leq \xi_{\sup} \times \left(f(S) + (1 - \epsilon) f_S(\mathbf{x}^\star) \right).$$

Let $\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} \tilde{f}(S \cup \mathbf{z})$. We show that given the right bound on $\xi_{\inf}$ against $\xi_{\sup}$, then $\hat{\mathbf{x}}$ is not a bad bundle (it does not have to be a good bundle). Lemma 3.3 below gives us such a bound.

Lemma 3.3. *For any generalized exponential tail distribution, fixed $\gamma > 0$ and sufficiently large n:*

$$\xi_{\inf} \geq \left(1 - \frac{\gamma}{\sqrt{\log n}}\right) \xi_{\sup}$$

The proof is quite technical and the fully leverages the fact that $k \in \mathcal{O}(\sqrt{\log n})$ and the properties of generalized exponential tail distributions. The main challenge is that the tail of the distribution is not necessarily monotone (see Appendix B for further discussion). We defer the proof to Appendix B.

Proving the Main Lemma. We can now prove our main lemma which shows that for any $k \in \omega(1/\epsilon) \cap \mathcal{O}(\sqrt{\log n})$ taking a bundle according to the sample mean approach is guaranteed to be close to the optimal bundle. We state it for a cardinality constraint, but this fact holds more generally for any matroid of rank k. We include a short proof sketch and defer the full proof to Appendix B.

Lemma 3.4. *Let $\epsilon > 0$, and assume that $k \in \omega(1/\epsilon) \cap \mathcal{O}(\sqrt{\log n})$ and that $f_S(\mathbf{x}^\star) \in \Omega\left(\frac{f(S)}{k}\right)$. Then, in every iteration of* SM-Greedy *we have that with probability at least $1 - \frac{4}{\log n}$:*

$$f_S(\hat{\mathbf{x}}) \geq (1 - \epsilon) f_S(\mathbf{x}^\star).$$

Proof Sketch. From Corollary 2.3 we know that in every iteration with overwhelming high probability $F_S(\mathbf{x}) \geq (1 - \delta) f(\mathbf{x}^\star)$ for $\delta = \epsilon/3$. Since Lemma 3.3 applies for any fixed $\gamma > 0$ we know that for sufficiently large n we can lower bound the value of the maximal good bundle in the ball:

$$\max_{\mathbf{z} \in \mathtt{good}} \tilde{f}(S \cup \mathbf{z}) \geq \left(1 - \frac{\delta^2}{3\sqrt{\log n}}\right) \xi_{\sup} \times [f(S) + (1 - 2\delta) f_S(\mathbf{x}^\star)]$$

Let $\mathbf{b}$ be the bad bundle with the maximal noisy value for f_S. To bound this:

$$\tilde{f}(S \cup \mathbf{b}) = \max_{\mathbf{z} \in \mathtt{bad}} \xi_{S \cup \mathbf{z}} f(S \cup \mathbf{z}) \leq \xi_{\sup} \times [f(S) + (1 - 3\delta) f_S(\mathbf{x}^\star)]$$

The difference between the two bounds is positive, implying that a bad bundle is not selected. □

Theorem 3.5. *For any fixed $\epsilon > 0$ and $k \geq 4/\epsilon^2$* SM-Greedy *returns a set S s.t. w.p. $1 - o(1)$:*

$$f(S) \geq (1 - 1/e - \epsilon)\mathtt{OPT}.$$

Proof. Let $\kappa = \lfloor \frac{k}{c} \rfloor$ and use O to denote the set of κ bundles of size c whose total value is largest. To simplify notation we will treat sets of bundles as sets of elements. We will show that $f(S) \geq (1 - 1/e - \delta)f(O)$ where $\delta = \epsilon/2$. Notice that this implies $f(S) \geq (1 - 1/e - \epsilon)$ since $\kappa > \frac{k}{c} - c$ and by submodularity $f(O) \geq (1 - \delta)\texttt{OPT}$ when $k \geq \frac{1}{\delta^2} = \frac{4}{\epsilon^2}$.

We introduce some notation: we use $\hat{\mathbf{x}}_i$ to denote the bundle selected at iteration $i \in [\kappa]$, $S_i = \cup_{j \leq i} \hat{\mathbf{x}}_j$, $\mathbf{x}_i \in \text{argmax}_{\mathbf{z}} \tilde{F}(S_{i\text{-}1} \cup \mathbf{z})$, $\mathbf{x}_i^\star \in \text{argmax}_{\mathbf{z}} f_{S_{i\text{-}1}}(\mathbf{z})$, and O is the set of κ bundles with maximal value. Define $\Delta_i = f(O) - f(S_{i\text{-}1})$, and $S_0 = \emptyset$. From submodularity and monotonicity:

$$f_{S_{i\text{-}1}}(\mathbf{x}_i^\star) \geq \frac{1}{\kappa}\Delta_i$$

Consider now a set of bundles $\{\mathbf{z}_1, \ldots, \mathbf{z}_\kappa\}$ where for every $i \in [\kappa]$ we have that $\mathbf{z}_i$ is drawn u.a.r. from $\mathcal{B}_{S_{i\text{-}1}}(\mathbf{x}_i)$. For each such bundle we can assign a random variable ζ_i for which $f_{S_{i\text{-}1}}(\mathbf{z}_i) = \zeta_i f_{S_{i\text{-}1}}(\mathbf{x}_i^\star)$. Since in every iteration $i \in [\kappa]$ we choose the set whose value is maximal in $\mathcal{B}_{S_{i\text{-}1}}(\mathbf{x}_i)$, by stochastic dominance we know that $f_{S_{i\text{-}1}}(\hat{\mathbf{x}}_i) \geq f(\mathbf{z}_i)$ and therefore:

$$f(S_i) - f(S_{i\text{-}1}) = f_{S_{i\text{-}1}}(\hat{\mathbf{x}}_i) \geq \zeta_i f_{S_{i\text{-}1}}(\mathbf{x}_i^\star) \geq \frac{\zeta_i}{\kappa}\Delta_i$$

We will now show by induction that for all $i \in [\kappa]$ we have that $\Delta_i \leq \prod_{j=1}^{i}\left(1 - \frac{\zeta_j}{\kappa}\right) f(O)$. This is clearly the case for $i = 1$ when $S_0 = \emptyset$ and in general applying the inductive hypothesis we get:

$$\Delta_{i+1} = f(O) - f(S_i) \quad = \Delta_i - (f(S_i) - f(S_{i-1})) \leq \Delta_i \left(1 - \frac{\zeta_{i+1}}{\kappa}\right) \quad \leq \prod_{j=1}^{i+1}\left(1 - \frac{\zeta_j}{\kappa}\right) f(O)$$

We therefore have that:

$$\Delta_\kappa \leq \prod_{j=1}^{\kappa}\left(1 - \frac{\zeta_j}{\kappa}\right) f(O) \leq e^{-\frac{1}{\kappa}\sum_{j=1}^{\kappa}\zeta_j} f(O)$$

Observe that the solution of the algorithm S respects $f(S) = f(S_\kappa) = f(O) - \Delta_\kappa$, thus:

$$f(S) \geq \left(1 - e^{-\frac{1}{\kappa}\sum_{j=1}^{\kappa}\zeta_j}\right) f(O) \tag{1}$$

From Lemma 3.4, when $k \leq \sqrt{\log n}$ for every $i \in [\kappa - 1]$ we have that w.p. $1 - 4/\log(n)$: [3]

$$f_{S_{i\text{-}1}}(\hat{\mathbf{x}}_i) \geq (1 - \delta) f_{S_{i\text{-}1}}(\mathbf{x}_i^\star)$$

Therefore by a union bound, with probability $1 - o(1)$, we have that $\zeta_i \geq (1 - \delta)$ for all $i \in [\kappa]$. In particular, $\frac{1}{\kappa}\sum_{j=1}^{\kappa}\zeta_j \geq (1 - \delta)$. Otherwise, when $k > \sqrt{\log n}$ from Lemma 2.1:

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{B}(\mathbf{x}_i)}[f_{S_{i\text{-}1}}(\mathbf{z})] = F_{S_{i\text{-}1}}(\mathbf{x}_i) \geq \left(1 - \frac{\delta}{2}\right) f_{S_{i\text{-}1}}(\mathbf{x}_i^\star)$$

Thus, $\mathbb{E}[\zeta_{i+1}] \geq (1 - \delta/2)$, and by Chernoff when $\kappa > \sqrt{\log n}$ we get $\frac{1}{\kappa}\sum_{i=1}^{\kappa}\zeta_i \geq (1 - \delta)$ w.p. at least $1 - \exp(-\delta^2\kappa/8)$. Therefore, in both cases, when $k \leq \sqrt{\log n}$ and when $k \geq \sqrt{\log n}$ we have that $\frac{1}{\kappa}\sum_{j=1}^{\kappa}\zeta_j \geq (1 - \delta)$ w.p. $1 - o(1)$. With 1 this implies $f(S) \geq (1 - 1/e - \delta)f(O)$. □

Constant k and information theoretic lower bounds. For any constant, a single iteration of a minor modification of SM-GREEDY suffices. In Appendix B.1 we show an approximation arbitrarily close to $1 - 1/k$ w.h.p. and $1 - 1/(k+1)$ in expectation. For $k = 1$ this is arbitrarily close to $1/2$. In Appendix D we show nearly matching lower bounds, and in particular that no randomized algorithm can obtain an approximation ratio better than $1/2 + o(1)$ when $k = 1$, and that it is impossible to obtain an approximation better than $(2k - 1)/2k + O(1/\sqrt{n})$ for the optimal set of size k.

Theorem 3.6. *For any non-negative monotone submodular function there is a deterministic polynomial-time algorithm which optimizes the function under a cardinality constraint $k \geq 3$ and obtains an approximation ratio that is arbitrarily close to $1 - 1/e$ with probability $1 - o(1)$ using access to an approximate oracle. For $k \geq 2$ there is a randomized algorithm whose approximation ratio is arbitrarily close to $1 - 1/e$, in expectation over the randomization of the algorithm. For $k = 1$ the algorithm achieves a $1/2$ approximation in expectation, and no randomized algorithm can achieve an approximation better than $1/2 + o(1)$, in expectation.*

[3] W.l.o.g we assume that in every iteration $f_{S_{i\text{-}1}}(\mathbf{x}_i^\star) \in \Omega(\frac{f(S_{i\text{-}1})}{k})$ to apply Lemma 3.4. Since $\kappa \in \Omega(1/\epsilon)$, $k \in \Omega(1/\epsilon^2)$, and $f(S) \leq \texttt{OPT}$, ignoring iterations where this does not hold costs $\gamma\epsilon\texttt{OPT}$ for a small fixed γ.

4 Approximation Algorithm for Matroids

For intersection of matroids $\mathcal{F}$ the algorithm from the Section 3 is generalized as described below.

Algorithm 2 SM-MATROID-GREEDY

```
Input: intersection of matroids $\mathcal{F}$, precision $\epsilon > 0$, $c \leftarrow \frac{56}{\epsilon}$
$S \leftarrow \emptyset, X \leftarrow N$
while $X \neq S$ do
    $X \leftarrow X \setminus \{\mathbf{x} : S \cup \mathbf{x} \notin \mathcal{F}\}$
    $\mathbf{x} \leftarrow \text{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(S \cup \mathbf{b})$
    $\hat{\mathbf{x}} \leftarrow \text{argmax}_{\mathbf{z}\in\mathcal{B}(\mathbf{x})} \tilde{f}(S \cup \mathbf{z})$
    $S \leftarrow S \cup \hat{\mathbf{x}}$
end while
return $S$
```

The analysis of the algorithm uses the lemma below, which is a generalization of the classic result of [NWF78a]. The proof can be found in the Appendix.

Lemma 4.1. *Let O be the optimal solution, $k = |O|$, and for every iteration i of* SM-MATROID-GREEDY *let S_i be the set of elements selected and $\mathbf{x}_i^\star \in \text{argmax}_{|\mathbf{z}|=c} f_{S_{i-1}}(\mathbf{z})$. Then:*

$$f(O) \le (P+1)\sum_{i=1}^{\frac{k}{c}} f_{S_i}(\mathbf{x}_i^\star)$$

Theorem 4.2. *Let $\mathcal{F}$ denote the intersection of $P \ge 1$ matroids on the ground set N, and $f : 2^N \to \mathbb{R}$ be a non-negative monotone submodular function. Then with probability $1 - o(1)$ the* SM-MATROID-GREEDY *algorithm returns a set $S \in \mathcal{F}$ s.t.:*

$$f(S) \ge \frac{1-\epsilon}{P+1} OPT$$

Proof Sketch. If the rank of the matroid is $\mathcal{O}(1/\epsilon^2)$ we can simply apply the case of small k as in the previous section. Otherwise, assume the rank is at least $\Omega(1/\epsilon^2)$ Let $\kappa = \frac{k}{c}$, $S_i = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_i\}$ be current solutions of bundles of size c at iteration $i \in [\kappa]$ of the algorithm, and let $\mathbf{x}_i^\star$ be the optimal bundle at iteration i, i.e. $\mathbf{x}_i^\star = \text{argmax}_{\mathbf{b}:|\mathbf{b}|\le c} f_{S_{i-1}}(\mathbf{b})$. In every iteration $i \in [\kappa]$, similar to the proof of Theorem 3.5, since we choose the set whose value is maximal in $\mathcal{B}_{S_{i-1}}(\mathbf{x}_i)$ we have:

$$f_{S_{i-1}}(\hat{\mathbf{x}}_i) \ge \zeta_i f_{S_{i-1}}(\mathbf{x}_i^\star)$$

where ζ_i is a random variable with mean $(1 - \frac{\delta}{2})$. Therefore:

$$f(S) = \sum_{i=1}^{\kappa} f_{i-1}(\hat{\mathbf{x}}_i) \ge \sum_{i=1}^{\kappa} \zeta_i f_{S_{i-1}}(\mathbf{x}_i^\star)$$

From Lemma 3.4, when $k \le \sqrt{\log n}$ for every $i \in [\kappa - 1]$ we have that w.p. $1 - 4/\log(n)$:

$$f_{S_{i-1}}(\hat{\mathbf{x}}_i) \ge (1-\delta) f_{S_{i-1}}(\mathbf{x}_i^\star)$$

Therefore by a union bound, with probability $1 - o(1)$, we have that $\zeta_i \ge (1-\delta)$ for all $i \in [\kappa]$. Otherwise, when $k > \sqrt{\log n}$ we apply a Chernoff bound. We get that with probability $1 - o(1)$:

$$f(S) \ge \sum_{i=1}^{\kappa} \zeta_i f_{S_{i-1}}(\mathbf{x}_i^\star) \ge (1-\delta)\sum_{i=1}^{\kappa} f_{S_{i-1}}(\mathbf{x}_i^\star)$$

From Lemma 4.1 this implies the result. □

Acknowledgements. A.H. is supported by 1394/16 and by a BSF grant. Y.S. is supported by NSF grant CAREER CCF 1452961, NSF CCF 1301976, BSF grant 2014389, NSF USICCS proposal 1540428, a Google Research award, and a Facebook research award.

References

[Ang88] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[AS04] Alexander A. Ageev and Maxim Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *J. Comb. Optim.*, 8(3), 2004.

[BDF$^+$10] D. Buchfuhrer, S. Dughmi, H. Fu, R. Kleinberg, E. Mossel, C. H. Papadimitriou, M. Schapira, Y. Singer, and C. Umans. Inapproximability for VCG-based combinatorial auctions. In *SIAM-ACM Symposium on Discrete Algorithms (SODA)*, pages 518–536, 2010.

[BF02] Nader H. Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

[BFNS12] Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 649–658, 2012.

[BFNS14] Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1433–1452, 2014.

[BH11] 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*, pages 793–802, 2011.

[BMKK14] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: massive data summarization on the fly. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 671–680, 2014.

[BSS10] D. Buchfuhrer, M. Schapira, and Y. Singer. Computation and incentives in combinatorial public projects. In *EC*, pages 33–42, 2010.

[CCPV07] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *Integer programming and combinatorial optimization*, pages 182–196. Springer, 2007.

[CE11] Chandra Chekuri and Alina Ene. Approximation algorithms for submodular multiway partition. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 807–816, 2011.

[CE16] C.P. Chambers and F. Echenique. *Revealed Preference Theory*. Econometric Society Monographs. Cambridge University Press, 2016.

[CHHK18] Lin Chen, Christopher Harshaw, Hamed Hassani, and Amin Karbasi. Projection-free online optimization with stochastic gradient: From convexity to submodularity. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 813–822, 2018.

[CJV15] Chandra Chekuri, T. S. Jayram, and Jan Vondrák. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, pages 201–210, 2015.

[DFK11] Shahar Dobzinski, Hu Fu, and Robert D. Kleinberg. Optimal auctions with correlated bidders are easy. In *STOC*, pages 129–138, 2011.

[DK14] J. Djolonga and A. Krause. From MAP to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[DLN08] Shahar Dobzinski, Ron Lavi, and Noam Nisan. Multi-unit auctions with budget limits. In *FOCS*, 2008.

[DNS05] Shahar Dobzinski, Noam Nisan, and Michael Schapira. Approximation algorithms for combinatorial auctions with complement-free bidders. In *STOC*, pages 610–618, 2005.

[DRY11] Shaddin Dughmi, Tim Roughgarden, and Qiqi Yan. From convex optimization to randomized mechanisms: toward optimal combinatorial auctions. In *STOC*, pages 149–158, 2011.

[DS06] 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.

[DV12] Shahar Dobzinski and Jan Vondrák. The computational complexity of truthfulness in combinatorial auctions. In *EC*, pages 405–422, 2012.

[Fei98] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[Fel09] Vitaly Feldman. On the power of membership queries in agnostic learning. *Journal of Machine Learning Research*, 10:163–182, 2009.

[FFI$^+$15] Uriel Feige, Michal Feldman, Nicole Immorlica, Rani Izsak, Brendan Lucier, and Vasilis Syrgkanis. A unifying hierarchy of valuations with complements and substitutes. In *AAAI*, pages 872–878, 2015.

[FMV11] Uriel Feige, Vahab S Mirrokni, and Jan Vondrak. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

[FNW78] Marshall L Fisher, George L Nemhauser, and Laurence A Wolsey. *An analysis of approximations for maximizing submodular set functions—II.* Springer, 1978.

[FV06] Uriel Feige and Jan Vondrak. Approximation algorithms for allocation problems: Improving the factor of 1-1/e. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 667–676. IEEE, 2006.

[GFK10] D. Golovin, M. Faulkner, and A. Krause. Online distributed sensor selection. In *IPSN*, 2010.

[GK10] R. Gomes and A. Krause. Budgeted nonparametric learning from data streams. In *Int. Conference on Machine Learning (ICML)*, 2010.

[GK11] D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, 42:427–486, 2011.

[GKS90] Sally A. Goldman, Michael J. Kearns, and Robert E. Schapire. Exact identification of circuits using fixed points of amplification functions (abstract). In *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990.*, page 388, 1990.

[HS] Thibaut Horel and Yaron Singer. Maximizing approximately submodular functions. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*.

[HS17] Avinatan Hassidim and Yaron Singer. Submodular optimization under noise. 2017. COLT.

[HSK17] S. Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5843–5853, 2017.

[Jac94] Jeffrey C. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. In *35th Annual Symposium on Foundations of Computer Science, Santa Fe, New Mexico, USA, 20-22 November 1994*, pages 42–53, 1994.

[JB11a] S. Jegelka and J. Bilmes. Approximation bounds for inference using cooperative cuts. In *Int. Conference on Machine Learning (ICML)*, 2011.

[JB11b] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1897–1904, 2011.

[KG07] A. Krause and C. Guestrin. Nonmyopic active learning of gaussian processes. an exploration–exploitation approach. In *Int. Conference on Machine Learning (ICML)*, 2007.

[KKT03] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.

[KLMM05] Subhash Khot, Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Inapproximability results for combinatorial auctions with submodular utility functions. In *Internet and Network Economics*, pages 92–101. Springer, 2005.

[KMVV13] Ravi Kumar, Benjamin Moseley, Sergei Vassilvitskii, and Andrea Vattani. Fast greedy algorithms in mapreduce and streaming. In *SPAA*, 2013.

[KOJ13] P. Kohli, A. Osokin, and S. Jegelka. A principled deep random field for image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[LB11a] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *ACL/HLT*, 2011.

[LB11b] H. Lin and J. Bilmes. Optimal selection of limited vocabulary speech corpora. In *Proc. Interspeech*, 2011.

[LCSZ17] Qiang Li, Wei Chen, Xiaoming Sun, and Jialin Zhang. Influence maximization with ϵ-almost submodular threshold functions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3804–3814, 2017.

[LKG$^+$07] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[LMNS09] Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 323–332, 2009.

[LSST13] Brendan Lucier, Yaron Singer, Vasilis Syrgkanis, and Éva Tardos. Equilibrium in combinatorial public projects. In *WINE*, pages 347–360, 2013.

[MHK18] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1886–1895, 2018.

[MSV08] Vahab S. Mirrokni, Michael Schapira, and Jan Vondrák. Tight information-theoretic lower bounds for welfare maximization in combinatorial auctions. In *Proceedings 9th ACM Conference on Electronic Commerce (EC-2008), Chicago, IL, USA, June 8-12, 2008*, pages 70–77, 2008.

[NW78] George L Nemhauser and Leonard A Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of operations research*, 3(3):177–188, 1978.

[NWF78a] 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.

[NWF78b] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

[PP11] Christos H. Papadimitriou and George Pierrakos. On optimal single-item auctions. In *STOC*, pages 119–128, 2011.

[PSS08] Christos H. Papadimitriou, Michael Schapira, and Yaron Singer. On the hardness of being truthful. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 250–259, 2008.

[QSY$^+$17] Chao Qian, Jing-Cheng Shi, Yang Yu, Ke Tang, and Zhi-Hua Zhou. Subset selection under noise. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3563–3573, 2017.

[RLK11] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM TKDD*, 5(4), 2011.

[SGK09] M. Streeter, D. Golovin, and A. Krause. Online learning of assignments. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[SS95] Eli Shamir and Clara Schwartzman. Learning by extended statistical queries and its relation to PAC learning. In *Computational Learning Theory: Eurocolt '95*, pages 357–366. Springer-Verlag, 1995.

[SS08] Michael Schapira and Yaron Singer. Inapproximability of combinatorial public projects. In *WINE*, pages 351–361, 2008.

[VCZ11] Jan Vondrák, Chandra Chekuri, and Rico Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 783–792, New York, NY, USA, 2011. ACM.

[Von08] Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *STOC*, pages 67–74, 2008.

A The Sampled Mean Method

Recall that for a bundle $\mathbf{x} = \{x_1, \dots, x_c\}$ we defined $\mathbf{x}_{ij} = \mathbf{x} \setminus \{x_i\} \cup \{x_j\}$ and the mean value is:

$$F(S \cup \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{B}_S(\mathbf{x})}[f(\mathbf{z})] = \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t} f(S \cup \mathbf{x}_{ij})\right)$$

Throughout this section we make some technical assumptions that essentially hold w.l.o.g:

- First, we assume that $f_S(\mathbf{x}^\star) \in \Omega\left(\frac{f(S)}{k}\right)$. Since we run the algorithm for $\kappa = \lfloor k/c \rfloor$ iterations, ignoring iterations for which this does not hold will cost at most $\gamma\epsilon\texttt{OPT}$ for some small fixed γ of our choice;

- Next, we can assume that $k < (1-\delta)n$ for some small fixed δ. Otherwise, by selecting k elements u.a.r we obtain an approximation that beats $1 - 1/e$. Since $c \in \mathcal{O}(1/\epsilon)$ we have that $n - |S| - c \in \Omega(n)$ and there are $\Omega(n)$ elements that we can use to swap with elements in $\mathbf{x}$;

- Another assumption we make is that $t \in \omega(n^{1/2}k^2)$ (the choice of $n^{1/2}$ is arbitrary, and we can use $n^{1-\alpha}$ for any fixed $\alpha > 0$). Note that even when $k \in \Omega(n)$ we can make $t \in \Omega(n^d)$ by defining the ball $\mathcal{B}_S(\mathbf{x})$ to be all bundles obtained by removing and swapping d elements from $\mathbf{x}$, instead of swapping a single element as in the current description. Using $d = 3$ for example, suffices. For readability we chose to describe the sampled mean approximation method with a single swap. Since in the worst case we may need to increase the size of the bundles by a factor of 3, we account for this blowup in the description of SM-GREEDY. That is, instead of setting $c = 18/\epsilon$ which suffices when $k \in \mathcal{O}(n^{1/4})$, we use $c = 56/\epsilon$;

- Throughout the entire paper we assume that the values of the distribution are independent of n. As discussed in the Introduction, without this assumption no algorithm can obtain any finite approximation. Therefore, since we do not necessarily assume the distribution is bounded, we assume that n is sufficiently large.

Weak concentration bound. We now turn to prove the weak concentration bound. This bound implies that choosing the element that maximizes the noisy sampled mean is an arbitrarily good approximation to choosing the element that maximizes the (non-noisy) sampled mean.

Lemma. 2.1 *Let* $\mathbf{x} \in \operatorname{argmax}_{\mathbf{z}:|\mathbf{z}|=c} \tilde{F}(S \cup \mathbf{z})$ *where* $c = 2/\epsilon$. *Then, w.p. at least* $1 - \exp\left(-\Omega\left(n^{1/4}\right)\right)$*:*

$$F_S(\mathbf{x}) \geq (1-\epsilon) \max_{\mathbf{z}:|\mathbf{z}|=c} F_S(\mathbf{z}).$$

The proof uses lemmas A.1 and A.2. The first lemma lower bounds the noisy mean value of $\mathbf{x}^\star$ against its (non-noisy) mean value, and the second upper bounds the noisy mean value of an arbitrary bundle against its (non-noisy) mean value. We use μ to denote the mean of the distribution.

Lemma A.1. *For a fixed* $\epsilon > 0$*, let* $\mathbf{x}^\star \in \operatorname{argmax}_{\mathbf{z}:|\mathbf{z}|=1/\epsilon} f_S(\mathbf{z})$ *and let* $\eta > 0$ *be a constant. For any* $\lambda > 0$ *we have that with probability at least* $1 - \exp\left(-\Omega(\lambda^2 t^{1/2-\eta})\right)$*:*

$$\tilde{F}(S \cup \mathbf{x}^\star) \geq (1-\lambda)\mu \cdot (f(S) + (1-\epsilon)F_S(\mathbf{x}^\star))$$

Proof. Let $\omega = \max_{\mathbf{x}_{ij} \in \mathcal{B}_S(\mathbf{x})} \xi_{ij}$ be the upper bound on values of noise multipliers in the ball, and $t = n - c - |S|$, where $c = 1/\epsilon$. We can break up the sampled mean value to two terms:

$$\begin{aligned}
\tilde{F}(S \cup \mathbf{x}^\star) &= \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\tilde{f}(S \cup \mathbf{x}^\star_{ij})\right)\\
&= \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f(S \cup \mathbf{x}^\star_{ij})\right)\\
&= \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f(S) + \xi_{ij} f_S(\mathbf{x}^\star_{ij})\right)\\
&= \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f(S)\right) + \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f_S(\mathbf{x}^\star_{ij})\right)
\end{aligned}$$

For the first term, by a straightforward application of the Chernoff bound we know that it is lower bounded by $(1-\lambda)\mu f(S)$ with probability at least $1 - \exp(\frac{-\lambda^2 c\dot{t}}{\omega})$.

To bound the second term, we make the following observations:

- There is at most one set $\mathbf{x}^\star_{\text{-}i}$ for which $f_S(\mathbf{x}^\star_{-i}) < \frac{f_S(\mathbf{x}^\star)}{2}$. To see this, assume for purpose of contradiction there are $\mathbf{x}^\star_{\text{-}i}$ and $\mathbf{x}^\star_{\text{-}j}$ s.t. $f_S(\mathbf{x}^\star_{\text{-}i}) \leq f_S(\mathbf{x}^\star_{\text{-}j}) < f_S(\mathbf{x}^\star)/2$, then since $\mathbf{x}^\star = \mathbf{x}^\star_{\text{-}i} \cup \mathbf{x}^\star_{\text{-}j}$, by subadditivity we get a contradiction:
$$f_S(\mathbf{x}^\star) = f_S(\mathbf{x}^\star_{\text{-}i} \cup \mathbf{x}^\star_{\text{-}j}) \leq f_S(\mathbf{x}^\star_{\text{-}i}) + f_S(\mathbf{x}^\star_{\text{-}j}) < 2 \cdot \frac{f_S(\mathbf{x}^\star)}{2} = f_S(\mathbf{x}^\star).$$
- Let l be in the index of the set with lowest value. Since there is at most one $i \in [c]$ for which $f_S(\mathbf{x}^\star_{\text{-}i}) < f_S(\mathbf{x}^\star)$, for all $i \neq l$ we know that the minimal value is at least $f_S(\mathbf{x}^\star)/2$. Due to the maximality of $\mathbf{x}^\star$ we also know that $f_S(\mathbf{x}^\star_{ij}) \leq f_S(\mathbf{x}^\star)$, for all $i \in [c]$. We can therefore apply a Chernoff bound on $\sum_{j=1}^{t}\xi_{ij} f_S(\mathbf{x}^\star_{ij})$ for every $i \neq l$ and get that w.p. at least $1 - \exp(\frac{\lambda^2 t}{\omega})$:
$$\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f_S(\mathbf{x}^\star_{ij}) \geq (1-\lambda)\mu\frac{1}{t}\sum_{j=1}^{t} f_S(\mathbf{x}^\star_{ij});$$
- By the minimality of l, and since $c = 1/\epsilon$, we know that:
$$\frac{1}{c \cdot t}\sum_{i \neq l}\sum_{i=1}^{t} f_S(\mathbf{x}^\star_{ij}) \geq (1-\epsilon)\frac{1}{c \cdot t}\sum_{i=1}^{c}\sum_{i=1}^{t} f_S(\mathbf{x}^\star_{ij}) = (1-\epsilon)F_S(\mathbf{x}^\star)$$

Together, these observations give us our desired bound:

$$\begin{aligned}
\tilde{F}(S \cup \mathbf{x}^\star) &= \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f(S)\right) + \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f_S(\mathbf{x}^\star_{ij})\right)\\
&\geq \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f(S)\right) + \frac{1}{c}\sum_{i \neq l}\left(\frac{1}{t}\sum_{j=1}^{t}\xi_{ij} f_S(\mathbf{x}^\star_{ij})\right)\\
&\geq (1-\lambda)\mu f(S) + (1-\lambda)\mu\frac{1}{c}\sum_{i \neq l}\sum_{j=1}^{t} f_S(\mathbf{x}^\star_{ij})\\
&\geq (1-\lambda)\mu f(S) + (1-\lambda)\mu(1-\epsilon)F_S(\mathbf{x}^\star)\\
&= (1-\lambda)\mu\left(f(S) + (1-\epsilon)F_S(\mathbf{x}^\star)\right)
\end{aligned}$$

Finally, to upper bound ω, since the noise distribution is a generalized exponential tail we have that for any $\delta > 0$ and sufficiently large m:

$$\Pr[\omega < m^{\delta}] > 1 - \exp\left(-\Omega\left(\frac{m^{\delta}}{\log m}\right)\right)$$

To see this, note that this is trivially true when m tends to infinity when the noise distribution is bounded, or has finite support. If the noise distribution is unbounded, since its tail is subexponential, at any given sample the probability of seeing the value m^{δ} is at most $\exp(-\mathcal{O}(m^{\delta}))$ and iterating this a polynomial number of times achieves the bound.

Therefore we know that with probability $1 - \exp\left(-\Omega(\sqrt{c \cdot t}\log^{-1}(c \cdot t))\right)$ all $c \cdot t = |\mathcal{B}_S(\mathbf{x})|$ noise multipliers are bounded from above by $\omega = \sqrt{ct}$. Thus by a union bound, the bound above holds with probability of at least $1 - \exp\left(-\Omega(\lambda^2 t^{1/2-\eta})\right)$ for some arbitrarily small $\eta > 0$. □

The following bound shows that for sufficiently large t (which is proportional to n), we have that $\widetilde{F}(S) \approx (1+\lambda)\mu(F(S) + 3t^{-1/4} f_S(\mathbf{x}^{\star}))$ for small $\lambda > 0$.

Lemma A.2. *For $\epsilon > 0$, let $\mathbf{x}$ be a bundle of size $c = 1/\epsilon$ and let $\eta > 0$ be a constant. For any $\lambda > 0$ we have that with probability at least $1 - \exp\left(-\Omega\left(\lambda^2 t^{1/2-\eta}\right)\right)$:*

$$\widetilde{F}(S \cup \mathbf{x}) < (1+\lambda)\mu \cdot \left(f(S) + F_S(\mathbf{x}) + 3t^{-\eta/3} f_S(\mathbf{x}^{\star})\right).$$

Proof. As in the proof of the previous lemma, let ω be the upper bound on the value of a noise multipliers. Since $\widetilde{F}(\mathbf{x})$ is an average of these values over all $i \in [c]$, a concentration bound that holds for every $i \in [c]$ will give us the desired result.

For a given $i \in [c]$, let $\mathbf{z}_1, \ldots, \mathbf{z}_t$ denote the bundles $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$. For each bundle $\mathbf{z}_i$ we will use α_i to denote the marginal value $f_S(\mathbf{z}_i)$ and ξ_i to denote the noise multiplier $\xi_{S \cup \mathbf{z}_i}$. We have two sums:

$$\frac{1}{t}\sum_{i=1}^{t} \tilde{f}(S \cup \mathbf{x}_{ij}) = \frac{1}{t}\sum_{i=1}^{t} \xi_i f(S) + \frac{1}{t}\sum_{i=1}^{t} \xi_i \alpha_i. \tag{2}$$

As before, we can immediately apply a Chernoff bound on the first term and what remains is to show concentration on the second term. Define $\alpha^{\star} = \max_i \alpha_i$. Note that due to the maximality of $\mathbf{x}^{\star}$ we have that $\alpha^{\star} \leq f_S(\mathbf{x}^{\star})$. To apply concentration bounds on the second term, we partition the values of $\{\alpha_i\}_{i \in [t]}$ to bins of width $\alpha^{\star} \cdot t^{-\nu}$ for some arbitrarily small constant $\nu > 0$. We call a bin *dense* if it has at least $t^{1-2\nu}$ values and *sparse* otherwise. Using this terminology:

$$\sum_{i=1}^{t} \xi_i \alpha_i = \sum_{i \in \text{dense}} \xi_i \alpha_i + \sum_{i \in \text{sparse}} \xi_i \alpha_i.$$

Let BIN_{ℓ} be the dense bin whose elements have the largest values. Consider the $t^{1-2\nu}/2$ largest values in BIN_{ℓ} and call the set of indices associated with these values L. We can rewrite the sum as:

$$\sum_{i=1}^{t} \xi_i \alpha_i = \sum_{i \in \text{dense} \setminus L} \xi_i \alpha_i + \sum_{i \in L \cup \text{sparse}} \xi_i \alpha_i$$

The set $L \cup$ sparse is of size at least $t^{1-2\nu}/2$ and at most $t^{1-2\nu}/2 + t^{1-\nu}$. This is because L is of size exactly $t^{1-2\nu}/2$ and there are at most $t^{1-\nu}$ values in bins that are sparse since there are t^{ν} bins, and sparse bins have less than $t^{1-2\nu}$ values. Thus, when ω is an upper bound on the value of the noise multiplier, from Chernoff, for any $\lambda < 1$ with probability $1 - \exp(-\Omega(\lambda^2 t^{1-2\nu}\omega^{-1}))$:

$$\begin{aligned}
\sum_{i \in L \cup \text{sparse}} \xi_i \alpha_i &\leq \sum_{i \in L \cup \text{sparse}} \xi_i \alpha^{\star} \\
&< (1+\lambda)\mu \cdot |L \cup \text{sparse}| \cdot \alpha^{\star} \\
&\leq (1+\lambda)\mu \cdot \left(\frac{t^{1-2\nu}}{2} + t^{1-\nu}\right)\alpha^{\star} \\
&< (1+\lambda)\mu \cdot 2t^{1-\nu}\alpha^{\star}
\end{aligned}$$

We will now apply concentration bounds on the values in the dense bins. For a dense bin $\textsc{bin}_q$, let $\alpha_{\max(q)}$ and $\alpha_{\min(q)}$ be the maximal and minimal values in the bin, respectively. Since the width of the bins is $\alpha^\star \cdot t^{-\nu}$ we have that $\alpha_{\min(q)} \geq \alpha_{\max(q)} - \alpha^\star \cdot t^{-\nu}$. Recall that a dense bin has at least $t^{1-2\nu}$ values. We can therefore apply a Chernoff bound on a dense bin $\textsc{bin}_q$. We get that for any $\lambda < 1$ w.p. $1 - \exp(-\lambda^2 t^{1-2\nu}\omega^{-1})$:

$$\begin{aligned}\sum_{i \in \textsc{bin}_q} \xi_i \alpha_i &\leq \sum_{i \in \textsc{bin}_q} \xi_i \cdot \alpha_{\max(q)} \\ &\leq (1+\lambda)\mu \cdot \alpha_{\max(q)} \cdot |\textsc{bin}_q| \\ &\leq (1+\lambda)\mu \cdot \left(\alpha_{\min(q)} + \alpha^\star \cdot t^{-\nu}\right) \cdot |\textsc{bin}_q| \\ &< (1+\lambda)\mu \cdot \left(|\textsc{bin}_q| \cdot \alpha_{\min(q)} + |\textsc{bin}_q|\alpha^\star \cdot t^{-\nu}\right)\end{aligned}$$

Applying a union bound over all bins we get with probability $1 - t^{\nu} \cdot \exp(-\lambda^2 t^{1-2\nu}\omega^{-1})$:

$$\sum_{i \in \text{dense} \setminus L} \xi_i \alpha_i < \sum_q (1+\lambda)\mu \cdot \left(|\textsc{bin}_q|\alpha^\star \cdot t^{-\nu} + |\textsc{bin}_q| \cdot \alpha_{\min(q)}\right) < (1+\lambda)\mu \cdot \left(\alpha^\star t^{1-\nu} + \sum_{i=1}^{t} \alpha_i\right)$$

Conditioning on both events, together we have:

$$\frac{1}{t}\sum_{j=1}^{t} f_S(\mathbf{x}_{ij}) = \frac{1}{t}\sum_{i=1}^{t} \xi_i \alpha_i = \frac{1}{t}\left(\sum_{i \in L \cup \text{sparse}} \xi_i \alpha_i + \sum_{i \in \text{dense} \setminus L} \xi_i \alpha_i\right) < (1+\lambda)\mu \cdot \left(3t^{-\nu} f_S(\mathbf{x}^\star) + F_S(\mathbf{x})\right)$$

By a union bound over all $i \in [c]$ we get that with probability $1 - \exp(-\Omega(\lambda^2 t^{1-3\nu}\omega^{-1}))$:

$$\tilde{F}(S \cup \mathbf{x}) = \frac{1}{c}\sum_{i=1}^{c}\left(\frac{1}{t}\sum_{j=1}^{t} \xi_{ij} f(S) + \frac{1}{t}\sum_{j=1}^{t} \xi_{ij} f_S(\mathbf{x}_{ij})\right) \leq (1+\lambda)\mu\left(f(S) + F_S(\mathbf{x}) + 3t^{-\nu} f_S(\mathbf{x}^\star)\right)$$

As in the previous lemma, since the distribution is a generalized exponential tail, we know that $\omega \leq \sqrt{c \cdot t}$ w.p. at least $1 - \exp\left(-\Omega((c \cdot t)^{1/2} \log^{-1}(c \cdot t))\right)$. Taking a union bound, the bound above holds with probability of at least $1 - \exp\left(-\Omega(\lambda^2 t^{1/2 - 3\nu})\right)$. Setting $\eta = \nu/3$ concludes our proof. □

Proof of Lemma 2.1. Let $\mathbf{x}^\star = \arg\max_{\mathbf{x}:|\mathbf{x}|=c} f_S(\mathbf{x})$ and let $\mathbf{b}$ be a bundle of size c for which $F_S(\mathbf{b}) < (1-\epsilon)F_S(\mathbf{x}^\star)$. We will show that $\mathbf{b}$ cannot be selected as $\mathbf{x}^\star$ beats $\mathbf{b}$ with overwhelming high probability:

$$\tilde{F}(S \cup \mathbf{x}^\star) > \tilde{F}(S \cup \mathbf{b}).$$

By taking a union bound over all possible $\mathcal{O}(n^c)$ bundles we will then conclude that the bundle whose noisy mean contribution is largest must have mean contribution at least factor of $(1-\epsilon)$ from that of $\mathbf{x}^\star$, with overwhelming high probability.

From Lemma A.1 since $c = 2/\epsilon$ we know that for $\lambda \in [0,1)$ with probability $1 - \exp\left(-\Omega(\lambda^2 t^{1/2-\eta})\right)$:

$$F(S \cup \mathbf{x}^\star) \geq (1-\lambda)\mu\left(f(S) + \left(1 - \frac{\epsilon}{2}\right)F_S(\mathbf{x}^\star)\right)$$

Similarly, from Lemma A.2 we know that with probability $1 - \exp\left(-\Omega(\lambda^2 t^{1/2-\eta})\right)$:

$$\begin{aligned}F(S \cup \mathbf{b}) &\geq (1+\lambda)\mu\left(f(S) + F_S(\mathbf{b}) + 3t^{-\eta/3} f_S(\mathbf{x}^\star)\right) \\ &\geq (1+\lambda)\mu\left(f(S) + (1 - \epsilon + 6t^{-\eta/3})F_S(\mathbf{x}^\star)\right)\end{aligned}$$

where we used the assumption $F_S(\mathbf{x}^\star) \geq (1-\epsilon)F_S(\mathbf{b})$ and the fact that $2F_S(\mathbf{x}^\star) \geq f_S(\mathbf{x}^\star)$. Recall that we're assuming that $f_S(\mathbf{x}^\star) \in \Omega\left(\frac{f(S)}{k}\right)$. Thus, for some small fixed $\alpha > 0$

we have that $F_S(\mathbf{x}^\star) \geq \alpha f(S)/2k$. Given the inequalities above, with probability at least $1 - \exp\left(-\Omega(\lambda^2 t^{1/2-\eta})\right)$:

$$\begin{aligned}\tilde{F}(S \cup \mathbf{x}^\star) - \tilde{F}(S \cup \mathbf{b}) &\geq \mu\left(\left(\frac{\epsilon}{2} - 2\lambda - (1+\lambda)6t^{-\eta/3}\right) F_S(\mathbf{x}^\star) - 2\lambda f(S)\right) \\ &\geq \mu\left(\left(\frac{\epsilon}{2} - 2\lambda - 12t^{-\eta/3} - \lambda\frac{k}{\alpha}\right) F_S(\mathbf{x}^\star)\right)\end{aligned}$$

For any $\lambda \leq \epsilon^2/4k$ the difference above is strictly positive. Without loss of generality we're assuming $t \gg n^{1/2} \cdot k^2$ and therefore the difference is positive with probability $1 - \exp\left(-\Omega(t^{1/4})\right)$. Since $t \in \Omega(n)$ this concludes our proof. □

The sampled mean is (almost) an upper bound of the function. We now show that when the size of the bundle is sufficiently large, the marginal contributions of the sampled mean nearly upper bound the true marginal contributions of a monotone submodular function.

Lemma. 2.2 *For any $\epsilon > 0$ and any set $S \subset N$, let $\mathbf{x}$ be a bundle of size $1/\epsilon$, then:*

$$F_S(\mathbf{x}) \geq (1-\epsilon) f_S(\mathbf{x})$$

Proof. Let $c = 1/\epsilon$ and consider an arbitrary ordering on the elements in the bundle $x_1, \ldots, x_c \in \mathbf{x}$. Define $\mathbf{x}_{\text{-}i} = \mathbf{x} \setminus \{x_i\}$, and $\mathbf{x}_{ij} = \mathbf{x}_{\text{-}i} \cup \{x_j\}$. From submodulairty we get that for any $i \in [c]$:

$$f_{S \cup \mathbf{x}_{\text{-}i}}(x_i) = f(S \cup \mathbf{x}_{\text{-}i} \cup x_i) - f(S \cup \mathbf{x}_{\text{-}i}) \leq f(S \cup \{x_1 \ldots, x_i\}) - f(S \cup \{x_1, \ldots, x_{i-1}\})$$

Thus:

$$\sum_{i=1}^{c} f_{S \cup \mathbf{x}_{\text{-}i}}(x_i) \leq \sum_{i=1}^{c} \left(f(S \cup \{x_1 \ldots, x_i\}) - f(S \cup \{x_1, \ldots, x_{i-1}\})\right) = f_S(\mathbf{x}) \tag{3}$$

Let $t = n - c - |S|$. By summing over all $\mathbf{x}_{\text{-}i}$ we get the desired bound:

$$\begin{aligned}F_S(\mathbf{x}) &= \frac{1}{c \cdot t}\sum_{j=1}^{t}\sum_{i=1}^{c} f_S(\mathbf{x}_{ij}) \\ &\geq \frac{1}{c}\sum_{i=1}^{c} f_S(\mathbf{x}_{\text{-}i}) \\ &= \frac{1}{c}\sum_{i=1}^{c} \left(f_S(\mathbf{x}_{\text{-}i} \cup x_i) - f_{S \cup \mathbf{x}_{\text{-}i}}(x_i)\right) \\ &= \frac{1}{c}\sum_{i=1}^{c} f_S(\mathbf{x}) - \frac{1}{c}\sum_{i=1}^{c} f_{S \cup \mathbf{x}_{\text{-}i}}(x_i) \\ &\geq f_S(\mathbf{x}) - \frac{1}{c} f_S(\mathbf{x}) && \text{by (3)} \\ &= \left(1 - \frac{1}{c}\right) f_S(\mathbf{x}) \\ &= (1-\epsilon)\, f_S(\mathbf{x}).\end{aligned}$$

□

B The Sampled-Mean Greedy

Claim. 3.1. *Suppose $F(\mathbf{x}) \geq \left(1 - \frac{\epsilon}{3}\right) f_S(\mathbf{x}^\star)$. Then at least half of the bundles in $\mathcal{B}_S(\mathbf{x})$ are good.*

Proof. For convenience we will use $\delta = \epsilon/3$. Let $\mathcal{B}_S^+(\mathbf{x})$ be the set of good bundles in $\mathcal{B}_S(\mathbf{x})$. Due to the maximality of $\mathbf{x}^\star$ we have that $f_S(\mathbf{z}) \leq f_S(\mathbf{x}^\star)$ for every $\mathbf{z} \in \mathcal{B}_S(\mathbf{x})$. Therefore:

$$\sum_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} f_S(\mathbf{z}) \leq |\, \mathcal{B}_{\mathrm{S}}^+(\mathbf{x}) \,| \cdot f_S(\mathbf{x}^\star) + \left(|\mathcal{B}_S(\mathbf{x})| - |\, \mathcal{B}_{\mathrm{S}}^+(\mathbf{x}) \,|\right) \cdot (1 - 2\delta) f_S(\mathbf{x}^\star) \tag{4}$$

By the definition of $F(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})}[f(\mathbf{z})]$ and our assumption $F(\mathbf{x}) \geq (1 - \delta) f(\mathbf{x})$:

$$\frac{1}{|\mathcal{B}_S(\mathbf{x})|} \sum_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} f_S(\mathbf{z}) \geq (1 - \delta) f_S(\mathbf{x}^\star) \tag{5}$$

Putting (4) and (5) together we get:

$$|\mathcal{B}_S(\mathbf{x})|(1 - \delta) f_S(\mathbf{x}^\star) \leq \left(|\, \mathcal{B}_{\mathrm{S}}^+(\mathbf{x}) \,| + (|\mathcal{B}_S(\mathbf{x})| - |\, \mathcal{B}_{\mathrm{S}}^+(\mathbf{x}) \,|)(1 - 2\delta)\right) f_S(\mathbf{x}^\star)$$

Rearranging we get that $|\, \mathcal{B}_{\mathrm{S}}^+(\mathbf{x}) \,| \geq |\mathcal{B}_S(\mathbf{x})|/2$, as required. □

Claim. 3.2 *Let $m = \frac{|\mathcal{B}_S(\mathbf{x})|}{2}$, $\xi_1, \ldots, \xi_m$ be i.i.d samples from $\mathcal{D}$ and $\xi^\star = \max\{\xi_1, \ldots, \xi_m\}$. Then:*

$$\Pr\left[\, \xi_{\text{inf}} \leq \xi^\star \leq \xi_{\text{sup}} \,\right] \geq 1 - \frac{3}{\log n}$$

Proof. For a single sample ξ from $\mathcal{D}$, we have that $\Pr[\xi \leq \xi_{\text{sup}}] = 1 - \frac{2}{m \log n}$. If we take m independent samples $\xi_1, \ldots \xi_m$, the probability they are all bounded by ξ_{sup} is:

$$\left(1 - \frac{2}{m \log n}\right)^m > \left(1 - \frac{2}{\log n}\right)$$

For ξ_{inf}, the probability a single sample ξ taken from $\mathcal{D}$ is at most ξ_{inf} is $\Pr\left[\xi \leq \xi_{\text{inf}}\right] = 1 - \frac{2 \log n}{m}$. If we take independent samples $\xi_1, \ldots \xi_m$, the probability they are all bounded by ξ_{inf} is:

$$\left(1 - \frac{2 \log n}{m}\right)^m < \frac{2}{2^{\log n}} = \frac{2}{n}$$

Therefore, by a union bound the likelihood that the maximum of m samples is bounded between ξ_{inf} and ξ_{sup} is at least $1 - \frac{3}{\log n}$. □

Bounding extreme values of the noise multipliers. Before proving Lemma 3.3, we illustrate the main challenge. First, consider a distribution which returns 0 with probability 0.99 and 1 with probability 0.01. If $m = 50$, clearly the lemma doesn't hold, but for $n > 1000$ the lemma would follow through. It is easy to generalize the problem to any distribution with an atom at its supremum.

One class of distributions for which the lemma may not hold, is one with an infinite number of atoms. For example, consider the distribution for which $\Pr[2^d] = 1/2^d$. In this case, the lemma is incorrect regardless of the value of m. The problem is not with the atoms, as it is easy to construct a density function which is non zero only around 2^d, and its integral around 2^d is exactly 2^{-d}. Note however that such a density function would be far from monotone. We do not want to require monotone noise distributions, as to not rule out bimodal distributions, and to allow for small fluctuations in the density function. Instead, we require that except for a finite number of modalities, the function's tail has a lower bound and an upper bound, which are somehow related. This requirement is rather weak, and encompasses in particular Exponential distributions, Gaussians (which are monotone), bounded distributions and distributions with a finite support.

Lemma. 3.3 *For any generalized exponential tail distribution, fixed $\gamma > 0$ and sufficiently large n:*

$$\xi_{\text{inf}} \geq \left(1 - \frac{\gamma}{\sqrt{\log n}}\right) \xi_{\text{sup}}$$

Proof. We will use $\varphi = \frac{\gamma}{\sqrt{\log n}}$ and first we give a proof for distributions with bounded support. Let M be an upper bound on $\mathcal{D}$. If there is an atom at M with some probability $p > 0$, then we are done, as $\xi_{\inf} = \xi_{\sup} = M$. Otherwise, since $\mathcal{D}$ has a generalized exponential tail we know that $\rho(M) = p$ for some $p > 0$, and that ρ is continuous at M. But then there is some $\delta > 0$ such that for any $M - \delta \leq x \leq M$ we have that $\rho(x) \geq p/2$. Choosing n to be large enough that $(1-\epsilon)p > p - \delta$:

$$\int_{(1-\epsilon)M}^{M} \rho(x) \geq p/2\epsilon$$

Choosing n large enough s.t. $2 \log n/m < \gamma/2\epsilon$ gives that $\xi_{\inf} \geq (1-\epsilon)M$. As $\xi_{\sup} \leq M$ we are done.

When the distribution does not have bounded support, recall that by definition of $\mathcal{D}$ for $x \geq x_0$, we have that $\rho(x) = e^{-g(x)}$, where $g(x) = \sum_i a_i x_i^{\alpha}$ and that we do not assume that all the α_i's are integers, but only that $\alpha_0 \geq \alpha_1 \geq \ldots$, and that $\alpha_0 \geq 1$. We do not assume anything on the other α_i values. In this case, the proof follows three stages:

1. We use properties of $\mathcal{D}$ to argue upper and lower bounds for $\rho(x)$;
2. We show an upper bound M on $\xi_{\sup}$;
3. We show that integrating a lower bound of $\rho(x)$ from $(1-\varphi)M$ to ∞, yields a probability mass at least $\frac{\log n}{\varphi\, m}$. Now suppose for contradiction that $\xi_{\inf} < (1-\varphi)\,\xi_{\sup}$, we would get that $\int_{\xi_{\inf}}^{\infty} \rho(x)$ is strictly greater than $\frac{\log n}{\varphi\, m}$, which contradicts the definition of $\xi_{\inf}$.

We now elaborate each on stage.

First stage. For the first stage we will show that for every $g(x)$, there exists n_0 such that for any $n > n_0$ and $x \geq \left(\frac{\log n}{2a_0}\right)^{1/\alpha_0}$ we have that for $\beta = \varphi/100 < 1/100$:

$$(1+\beta)a_0 x^{\alpha_0 - 1} e^{-(1+\beta)a_0 x^{\alpha_0}} \leq \rho(x) \leq (1-\beta)a_0 x^{\alpha_0-1} e^{-(1-\beta)a_0 x^{\alpha_0}}$$

We explain both directions of the inequality. To see $a_0 x^{\alpha_0 - 1}(1+\beta)e^{-(1+\beta)a_0 x^{\alpha_0}} \leq \rho(x)$ we first show:

$$e^{-(1+\beta/2)a_0 x^{\alpha_0}} \leq \rho(x)$$

This holds since for sufficiently large n, we have that:

$$x \geq \frac{(\log n)^{1/\alpha_0}}{2a_0} \geq \left(\frac{2\sum_{i=1} |a_i|}{\beta a_0}\right)^{\alpha_0 - \alpha_1}$$

So the term $\frac{\beta}{2} x^{\alpha_0}$ dominates the rest of the terms. We now show that:

$$e^{-(1+\beta/2)a_0 x^{\alpha_0}} \geq a_0 x^{\alpha_0 - 1}(1+\beta)e^{-(1+\beta)a_0 x^{\alpha_0}}$$

This is equivalent to:

$$e^{\beta a_0/2x^{\alpha_0}} \geq a_0 x^{\alpha_0-1}(1+\beta)$$

Which hold for $x = \log\log^3 n$ and large enough n.

The other side of the inequality is proved in a similar way. We want to show that:

$$\rho(x) \leq (1-\beta)a_0 x^{\alpha_0 - 1} e^{-(1-\beta)a_0 x^{\alpha_0}}$$

Clearly for $x > \log\log^3 n$ we have that $(1-\beta)a_0 x^{\alpha_0-1} > 1$. Hence we just need to show that:

$$\rho(x) \leq e^{-(1-\beta)a_0 x^{\alpha_0}}$$

But this holds for sufficiently large n s.t.:

$$x \geq \frac{(\log n)^{1/\alpha_0}}{2a_0} \geq \left(\frac{\sum_{i=1} |a_i|}{\beta a_0}\right)^{\alpha_0 - \alpha_1}$$

Second stage. We now proceed to the second stage, and compute an upper bound on ξ_{sup}. Note that if for every $x \geq M$ we have $\rho(x) \leq g(x)$ and

$$\int_{\xi_{\text{sup}}}^{\infty} \rho(x) = \int_{M}^{\infty} g(x)$$

then it must be that $M \geq \xi_{\text{sup}}$. Applying this to our setting, and using $m = \mathcal{B}_S(\mathbf{x}) = c \cdot (n - |S| - c)$ we bound $\rho(x) \leq (1-\beta)a_0 x^{\alpha_0 - 1} e^{-(1-\beta)a_0 x^{\alpha_0}}$ to get:

$$\frac{1}{m \log n} = \int_{M}^{\infty} (1-\beta)a_0 x^{\alpha_0 - 1} e^{-(1-\beta)a_0 x^{\alpha_0}} = -e^{-(1-\beta)a_0 x^{\alpha_0}}|_{M}^{\infty} = e^{-(1-\beta)a_0 M^{\alpha_0}}$$

Taking the logarithm of both sides, we get:

$$-(1-\beta)a_0 M^{\alpha_0} = \log \frac{1}{m \log n} = -\log(m \log n)$$

Multiplying by -1, dividing by $(1-\beta)a_0$ and taking the $1/\alpha_0$ root we get:

$$M = \left(\frac{\log m \log n)}{(1-\beta)a_0}\right)^{\alpha_0}$$

Note that $(1-\varphi)M > \left(\frac{\log n}{2a_0}\right)^{1/\alpha_0}$ and hence our bounds on $\rho(x)$ hold for this regime.

Third stage. We move to the third stage, and bound $\int_{(1-\varphi)M}^{\infty} \rho(x)$ from below. If we show that: $\int_{(1-\varphi)M}^{\infty} \rho(x)$ is greater than $\frac{\log n}{\varphi m}$, this implies that $\xi_{\text{inf}} \geq (1-\varphi)M$, as ξ_{inf} is defined as the value such that when we integrate $\rho(x)$ from ξ_{inf} to ∞ we get exactly $\frac{\log n}{\varphi m}$. We show:

$$\begin{aligned}
\int_{(1-\varphi)M}^{\infty} \rho(x) &\geq (1+\beta)a_0\alpha_0 x^{\alpha_0 - 1} e^{-(1+\beta)a_0 x^{\alpha_0}} \\
&= -e^{-(1+\beta)a_0 x^{\alpha_0}}|_{(1-\varphi)M}^{\infty} \\
&= e^{-(1+\beta)a_0((1-\varphi)M)^{\alpha_0}} \\
&= e^{-(1+\beta)a_0 M^{\alpha_0}(1-\varphi)^{\alpha_0}} \\
&\geq e^{-(1+\beta)a_0 M^{\alpha_0}(1-\varphi)}
\end{aligned}$$

However $a_0 M^{\alpha_0} = \left(\frac{\log m \log n)}{(1-\beta)}\right)$. Since $\beta < 0.1$ we have that $\frac{1+\beta}{1-\beta} < 1 + 3\beta$. Substituting both expressions we get:

$$\begin{aligned}
e^{-(1+\beta)a_0 M^{\alpha_0}(1-\varphi)} &\geq e^{-(1+3\beta)(1-\varphi)\log m \log n)} \\
&= \left(\frac{1}{m \log n}\right)^{(1-\varphi)(1+3\beta)} \\
&\geq \left(\frac{1}{m \log n}\right)^{(1-\varphi/2)}
\end{aligned}$$

where we used that $\beta = \varphi/100$ and hence $(1-\varphi)(1+3\beta) < 1 - \varphi/2$. We now need to compare this to $\frac{\sqrt{\log n}}{\varphi m}$. To do this, note that:

$$\left(\frac{1}{m \log n}\right)^{(1-\varphi/2)} \geq \frac{1}{m^{1-\varphi/2} \log n} \geq \frac{2^{\sqrt{\log n}}}{m \log n} \geq \frac{\log n}{\varphi m}$$

where n is large enough that $\frac{\varphi}{2} \log m > \sqrt{\log n}$. This completes the proof, since $\xi_{\text{inf}} \geq (1-\varphi)M \geq (1-\varphi)\,\xi_{\text{sup}}$ as required. □

Lemma. 3.4 *Let $\epsilon > 0$, and assume that $k \in \omega(1/\epsilon) \cap \mathcal{O}(\sqrt{\log n})$ and that $f_S(\mathbf{x}^\star) \in \Omega\left(\frac{f(S)}{k}\right)$. Then, in every iteration of* SM-GREEDY *we have that with probability at least $1 - \frac{4}{\log n}$:*

$$f_S(\hat{\mathbf{x}}) \geq (1-\epsilon) f_S(\mathbf{x}^\star).$$

Proof. From Corollary 2.3 we know that when $c = 9/\epsilon$, then in every iteration with overwhelming high probability we have that for $\delta = \epsilon/3$:

$$F_S(\mathbf{x}) \geq (1-\delta) f(\mathbf{x}^\star) \tag{6}$$

In every iteration the algorithm selects $\hat{\mathbf{x}} \in \operatorname{argmax}_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} \tilde{f}(S \cup \mathbf{z})$. Recall that a good bundle is $\mathbf{z} \in \mathcal{B}_S(\mathbf{x})$ for which $f(\mathbf{z}) \geq (1-2\epsilon/3) f(\mathbf{x}^\star)$ and a bad bundle is $\mathbf{z} \in \mathcal{B}_S(\mathbf{x})$ s.t. $f(\mathbf{z}) \leq (1-\epsilon) f(\mathbf{x}^\star)$. Conditioning on the assumption (6), from Claim 3.2 we know that with probability at least $1 - \frac{3}{\log n}$ the noise multipliers of both good and bad bundles in $\mathcal{B}_S(\mathbf{x})$ are in $[\xi_{\inf}, \xi_{\sup}]$. Since Lemma 3.3 applies for any fixed $\gamma > 0$ we know that for sufficiently large n we have that:

$$\xi_{\inf} \geq \left(1 - \frac{\delta^2}{3\sqrt{\log n}}\right) \xi_{\sup}$$

Thus, a lower bound on the maximal noisy value of a *good* bundle is:

$$\begin{aligned}
\max_{\mathbf{z} \in \mathtt{good}} \tilde{f}(S \cup \mathbf{z}) &= \max_{\mathbf{z} \in \mathtt{good}} \xi_{\mathbf{z}} \times [f(S) + f_S(\mathbf{z})] \\
&\geq \xi_{\inf} \times [f(S) + (1-2\delta) f_S(\mathbf{x}^\star)] \\
&\geq \left(1 - \frac{\delta^2}{3\sqrt{\log n}}\right) \xi_{\sup} \times [f(S) + (1-2\delta) f_S(\mathbf{x}^\star)]
\end{aligned}$$

An upper bound on the maximal noisy value of a *bad* bundle is:

$$\tilde{f}(S \cup \mathbf{b}) = \max_{\mathbf{z} \in \mathtt{bad}} \tilde{f}(S \cup \mathbf{z}) = \max_{\mathbf{z} \in \mathtt{bad}} \xi_{\mathbf{z}} f(S \cup \mathbf{z}) \leq \xi_{\sup}[f(S) + (1-3\delta) f_S(\mathbf{x}^\star)]$$

Since $f_S(\mathbf{x}) \in \Omega\left(\frac{f(S)}{k}\right)$, and importantly $k \leq \sqrt{\log n}$ we know that for sufficiently large n:

$$\frac{\sqrt{\log n}}{\delta} f_S(\mathbf{x}^\star) \geq f(S).$$

Putting it all together and conditioning on all events we have with probability at least $1 - \frac{4}{\log n}$:

$$\begin{aligned}
&\tilde{f}(S \cup \hat{\mathbf{x}}) - \tilde{f}(S \cup \mathbf{b}) \\
&\geq \left((1 - \frac{\delta^2}{3\sqrt{\log n}})\, \xi_{\sup}[f(S) + (1-2\delta) f_S(\mathbf{x}^\star)]\right) - \left(\xi_{\sup}[f(S) + (1-3\delta) f_S(\mathbf{x}^\star)]\right) \\
&\geq \xi_{\sup}\left(\delta f_S(\mathbf{x}^\star) - \frac{\delta^2}{3\sqrt{\log n}} \times [(1-2\delta) f_S(\mathbf{x}^\star) + f(S)]\right) \\
&\geq \xi_{\sup}\left(\delta f_S(\mathbf{x}^\star) - \frac{\delta^2}{3\sqrt{\log n}} \times \left[(1-2\delta) f_S(\mathbf{x}^\star) + \frac{\sqrt{\log n}}{\delta} f_S(\mathbf{x}^\star)\right]\right) \\
&> \xi_{\sup} \cdot \frac{\delta}{3} f_S(\mathbf{x}^\star)
\end{aligned}$$

Since the difference is strictly positive this implies that with probability at least $1 - \frac{4}{\log n}$ a bad bundle will not be selected as $\hat{\mathbf{x}}$, which concludes our proof. □

B.1 Optimization for Constant k

Redefining the ball. For every bundle $\mathbf{x}$ of size c we define the ball $\mathcal{B}(\mathbf{x}) = \{\mathbf{x} \cup z : z \notin \mathbf{x}\}$, and the mean value and noisy mean values are $F(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \in \mathcal{B}(\mathbf{x})}[f(\mathbf{z})]$ and $\tilde{F}(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \in \mathcal{B}(\mathbf{x})}[\tilde{f}(\mathbf{z})]$, respectively. Using the same reasoning as in Lemma 2.1 we get a concentration bound on $\operatorname{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(\mathbf{b})$.

Lemma B.1. *Let* $\mathbf{x} \in \operatorname{argmax}_{\mathbf{b}:|\mathbf{b}|=c} \tilde{F}(\mathbf{b})$. *Then, for any* $\lambda > 0$ *w.p.* $1 - \exp(-\Omega(\lambda^2 \sqrt{n-c}))$*:*

$$F(\mathbf{x}) \geq (1-\lambda) \max_{\mathbf{b}:|\mathbf{b}|=c} F(\mathbf{b}).$$

Algorithm 3 EXP-SM-GREEDY

```
Input: budget $k$
$\mathbf{x} \leftarrow \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k} \tilde{F}(\mathbf{b})$
$z \leftarrow$ select random element from $N \setminus \mathbf{x}$
$\hat{\mathbf{x}} \leftarrow$ random set of size $k$ from $\mathbf{x} \cup z$
return $\hat{\mathbf{x}}$
```

Approximation guarantee in expectation. We first present the algorithm whose approximation guarantee is arbitrarily close to $1 - \frac{1}{k+1}$, in expectation. The algorithm will simply select the set $\hat{\mathbf{x}}$ to be a random subset of k elements from a random set of $\mathcal{B}(\mathbf{x})$ where $\mathbf{x} \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k} \tilde{F}(\mathbf{b})$.

Theorem B.2. *For any submodular $f : 2^N \to \mathbb{R}$,* EXP-SM-GREEDY *returns a $(1 - \frac{1}{k+1} - o(1))$ approximation for $\max_{S:|S|\leq k} f(S)$, in expectation, using a generalized exponential tail noisy oracle.*

Proof. Similar to Lemma 2.2, by submodularity we know that in expectation $f(\hat{\mathbf{x}}) \geq \frac{k}{k+1} F(\mathbf{x})$. Let $\mathbf{x}^\star = \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k} f(\mathbf{b})$. From monotonicity we know that $f(\mathbf{x}^\star) \leq F(\mathbf{x}^\star)$. Applying Lemma B.1 we get that for the set $F(\mathbf{x}) \geq (1 - o(1))F(\mathbf{x}^\star)$. Hence:

$$\mathbb{E}[f(\hat{\mathbf{x}})] \geq \left(\frac{k}{k+1}\right) F(\mathbf{x}) \geq \left(\frac{k}{k+1} - o(1)\right) F(\mathbf{x}^\star) \geq \left(\frac{k}{k+1} - o(1)\right) f(\mathbf{x}^\star) = \left(\frac{k}{k+1} - o(1)\right) \texttt{OPT}.$$

□

Approximation Guarantee with high probability. To obtain a result that holds w.h.p. we will consider a modest variant of the algorithm above. The algorithm enumerates all possible subsets of size $k - 1$, identifies the bundle $\mathbf{x} \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k-1} \tilde{F}(\mathbf{b})$ and then returns $\hat{\mathbf{x}} \in \text{argmax}_{\mathbf{z}\in\mathcal{B}(\mathbf{x})} \tilde{f}(\mathbf{z})$.

Algorithm 4 WHP-SM-GREEDY

```
Input: budget $k$
$\mathbf{x} \leftarrow \arg\max_{\mathbf{b}:|\mathbf{b}|=k-1} \tilde{F}(\mathbf{b})$
$\hat{\mathbf{x}} \leftarrow \text{argmax}_{\mathbf{z}\in\mathcal{B}(\mathbf{x})} \tilde{f}(\mathbf{z})$
return $\hat{\mathbf{x}}$
```

Theorem B.3. *For any submodular function $f : 2^N \to \mathbb{R}$ and any fixed $\epsilon > 0$ and constant k, there is a $(1 - 1/k - \epsilon)$-approximation algorithm for $\max_{S:|S|\leq k} f(S)$ which only uses a generalized exponential tail noisy oracle, and succeeds with probability at least $1 - 6/\log n$.*

Proof. Let $\mathbf{x} \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k-1} \tilde{F}(\mathbf{b})$, and let $\mathbf{x}^\star \in \text{argmax}_{\mathbf{b}:|\mathbf{b}|=k-1} f(\mathbf{b})$. Since $\mathbf{x}^\star$ is the optimal solution over $k - 1$ elements, from submodularity we know that $f(\mathbf{x}^\star) \geq (1 - 1/k)\texttt{OPT}$.

What now remains to show is that $\hat{\mathbf{x}} \in \text{argmax}_{z\in N\setminus\mathbf{x}} \tilde{f}(\mathbf{x} \cup z)$ is a $(1 - \epsilon)$ approximation to $F(\mathbf{x})$. To do so, recall the definitions of good and bad bundles from the previous section: let $\delta = \epsilon/3$, and say a bundle $\mathbf{z}$ is *good* if $f(\mathbf{z}) \geq (1 - 2\delta)f(\mathbf{x}^\star)$ and *bad* if $f(\mathbf{z}) \leq (1 - 3\delta)f(\mathbf{x}^\star)$. We show that with high probability the bundle $\hat{\mathbf{x}}$ selected by the algorithm has value at least as high as that of a bad bundle, i.e. $f(\hat{\mathbf{x}}) \geq (1 - 3\delta)f(\mathbf{x}^\star)$ which will complete the proof.

We first show that with probability at least $1 - 6/\log n$ the maximal noise multiplier of a good bundle is at least ξ_{inf} and of a bad bundle is at most ξ_{sup}, where we use the same definition of ξ_{inf} and ξ_{sup} as in Section 3. To do so we will first argue about the number of good bundles in the ball. From Lemma B.1 and the maximality of $\mathbf{x}$ we know that with overwhelming high probability $F(\mathbf{x}) \geq (1 - o(1))F(\mathbf{x}^\star)$. Therefore for $m = n - k$ and fixed δ:

$$F(\mathbf{x}) = \frac{1}{m} \sum_{z \notin \mathbf{x}} f(\mathbf{x} \cup z) \geq (1 - \delta) \frac{1}{m} \sum_{z \notin \mathbf{x}^\star} f(\mathbf{x}^\star \cup z) \geq (1 - \delta) f(\mathbf{x}^\star)$$

Let $\mathcal{B}^+(\mathbf{x})$ be the bundle of all good bundles in $\mathcal{B}(\mathbf{x})$. Due to the maximality of $\mathbf{x}^\star$ and submodularity we know that $f(\mathbf{x} \cup z) \leq 2f(\mathbf{x}^\star)$ for all $z \notin \mathbf{x}$:

$$\sum_{z \notin \mathbf{x}} f(\mathbf{x} \cup z) \leq |\mathcal{B}^+(\mathbf{x})| 2 f(\mathbf{x}^\star) + (m - |\mathcal{B}^+(\mathbf{x})|)(1 - 2\delta) f(\mathbf{x}^\star)$$

Putting the these bounds on $F(\mathbf{x})$ together and rearranging we get that:

$$|\mathcal{B}^+(\mathbf{x})| \geq \frac{\delta \cdot m}{1 + 2\epsilon} \geq \frac{\delta m}{3}$$

Since there are at least $\delta m/3$ good bundles we can bound the likelihood of at least one noise multiplier of a good bundle achieving value $\xi_{\inf}$:

$$\Pr\left[\max\{\xi_1, \ldots, \xi_{\delta \cdot m/3}\} \geq \xi_{\inf}\right] \geq 1 - \left(1 - \frac{2\log n}{m}\right)^{\frac{\delta m}{3}} \geq 1 - \frac{2}{n^{\delta/3}} \geq 1 - \frac{1}{\log n}$$

Since there are $m = n - k$ bundles in the ball, the likelihood that all noise multipliers of bad bunldes are bounded from above by $\xi_{\sup}$ is:

$$\Pr\left[\max\{\xi_1, \ldots \xi_m\} \leq \xi_{\sup}\right] \geq \left(1 - \frac{2}{m \log n}\right)^m > \left(1 - \frac{4}{\log n}\right)$$

Thus, by a union bound and conditioning on the event in Lemma B.1 we get that $\xi_{\sup}$ is an upper bound on the value of the noise multiplier of bad bundles and $\xi_{\inf}$ is with lower bound on the value of the noise multiplier of a good stem all with probability at least $1 - 6/\log n$.

We therefore know that with probability at least $1 - 6/\log n$ the maximal noise multiplier of a good bundle is at least $\xi_{\inf}$ and the noise multiplier of a bad bundle is at most $\xi_{\sup}$. From Lemma 3.3 we know that $\xi_{\inf} \geq (1 - o(1))\, \xi_{\sup}$. Thus:

$$\max_{\mathbf{z} \in \mathcal{B}^+(\mathbf{x})} \tilde{f}(\mathbf{z}) = \max_{\mathbf{z} \in \in \mathcal{B}^+(\mathbf{x})} \xi_{\mathbf{z}} f(\mathbf{z}) \geq \xi_{\inf} \cdot (1 - 2\delta) f(\mathbf{x}^\star) \geq \xi_{\sup} \cdot (1 - 2\delta - o(1)) f(\mathbf{x}^\star)$$

Let $\mathbf{b} \in \operatorname{argmax}_{\mathbf{z} \in \text{bad bundles}} \tilde{f}(\mathbf{x} \cup \mathbf{z})$:

$$\tilde{f}(\mathbf{b}) \geq \max_{\mathbf{z} \in \text{bad bundles}} \tilde{f}(\mathbf{z}) = \max_{\mathbf{z} \in \text{bad bundles}} \xi_{\mathbf{z}} f(\mathbf{z}) \leq \xi_{\sup} \cdot (1 - 3\delta) f(\mathbf{z})$$

Putting it all together we have with probability at least $1 - 6/\log n$:

$$\tilde{f}(\hat{\mathbf{x}}) - \tilde{f}(\mathbf{b}) \geq \xi_{\sup} f(\mathbf{x}^\star) \cdot \Big((1 - 2\delta - o(1)) - (1 - 3\delta)\Big) > \xi_{\sup} f(\mathbf{x}^\star) (\delta - o(1))$$

The difference is strictly positive, and since $\delta = \epsilon/3$ is fixed and this completes our proof. □

C Approximation Algorithm for Matroids

We begin with basic facts and definitions about Matroids and properties of submodular functions.

Claim C.1. *Let $f : 2^N \to \mathbb{R}$ be a submodular function and $S_k, O \subseteq N$. Then we have that:*

$$f(O) \leq f(O \cup S_k) \leq f(S_k) + \sum_{x \in O \setminus S_k} f_{S_k}(x) \tag{7}$$

Proof. This is a direct consequence of submodularity. □

Definition C.2 (rank and span of a matroid)**.** *For a set S and a matroid $\mathcal{M}_j$ in the family $\mathcal{F}$, we define $\mathtt{rank}_j(S)$, called the rank of S in $\mathcal{M}_j$ to be the cardinality of the largest subset of S which is independent in $\mathcal{M}_j$, and define $\mathtt{span}_j(S)$, called the* span *of S in $\mathcal{M}_j$ by:*

$$\mathtt{span}_j(S) = \{a \in N : \mathtt{rank}_j(S \cup a) = \mathtt{rank}_j(S)\}$$

Claim C.3. *Let $S_i, O \in \mathcal{M}_j$ be independent sets where $S_i = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_i\}$ is a set of bundles of size c each. Then:*

$$|\mathtt{span}_j(S_i) \cap O| \leq c \cdot i$$

Proof. Since S_i is independent in $\mathcal{M}_j$, we know that $\texttt{rank}_j(\texttt{span}_j(S_i)) = \texttt{rank}_j(S_i) = |S_i|$. In particular, we have that $\texttt{rank}_j(\texttt{span}_j(S_i)) = c \cdot i$. Since O is an independent set in $\mathcal{M}_j$ we have:

$$\texttt{rank}_j(\texttt{span}_j(S_i) \cap O) = |\texttt{span}_j(S_i) \cap O| \leq |\texttt{span}_j(S_i)| = |S_i| \leq c \cdot i$$

where the above inequality is due to the fact that for any independent set T in $\mathcal{M}_j$ we have that $\texttt{rank}_j(T) = |\texttt{span}_j(T)| = |T|$. This implies that $|\texttt{span}_j(S_i) \cap O| \leq c \cdot i$. □

Claim C.4 (Prop. 2.2 in [NWF78a]). *If for* $\forall t \in [k]$ $\sum_{i=0}^{t-1} \sigma_i \leq t$ *and* $p_{i-1} \geq p_i$, *with* $\sigma_i, p_i \geq 0$ *then:*

$$\sum_{i=0}^{k-1} p_i \sigma_i \leq \sum_{i=0}^{k-1} p_i.$$

Lemma. *4.1 Let O be the optimal solution, $k = |O|$, and for every iteration i of* SM-MATROID-GREEDY *let S_i be the set of elements selected and $\mathbf{x}_i^\star \in \operatorname{argmax}_{|\mathbf{z}|=c} f_{S_{i\text{-}1}}(\mathbf{z})$ be the optimal bundle at stage i of the algorithm. Then:*

$$f(O) \leq (P+1) \sum_{i=1}^{\frac{k}{c}} f_{S_i}(\mathbf{x}_i^\star)$$

Proof. Since S_i is independent in $\mathcal{M}_j$, we know that $\texttt{rank}_j(\texttt{span}_j(S_i)) = \texttt{rank}_j(S_i) = |S_i|$. In particular, we have that $\texttt{rank}_j(\texttt{span}_j(S_i)) = c \cdot i$. Now in each $1 \leq j \leq P$, since O is an independent set in M_j we have:

$$\texttt{rank}_j(\texttt{span}_j(S_i) \cap O) = |\texttt{span}_j(S_i) \cap O|$$

which by Claim C.3 implies that $|\texttt{span}_j(S_i) \cap O| \leq c \cdot i$.

Define $U_i = \cup_{j=1}^{P} \texttt{span}_j(S_i)$, to be the set of elements which are not part of the maximization in step $i+1$ of the procedure, and hence cannot give value at that stage. We have:

$$|U_i \cap O| = |(\cup_{j=1}^{P} \texttt{span}_j(S_i)) \cap O| \leq \sum_{j=1}^{P} |\texttt{span}_j(S_i) \cap O| \leq P \cdot c \cdot i$$

Let $V_i = (U_i \setminus U_{i-1}) \cap O$ be the elements of O which are not part of the maximization in step i, but were part of the maximization in step $i-1$. If $x \in V_i$ then it must be that

$$f_k(x) \leq f_{S_i}(x) \leq \frac{f_{S_i}(\mathbf{x}_i^\star)}{c}$$

where the first inequality is due to submodularity of f and the second is since x was not chosen in step i. Hence, we can upper bound:

$$\sum_{x \in O \setminus S_k} f_{S_k}(x) \leq \sum_{i=1}^{\frac{k}{c}} \sum_{x \in V_i} \frac{f_{S_i}(\mathbf{x}_i^\star)}{c} = \sum_{i=1}^{\frac{k}{c}} |V_i| f_{S_i}(\mathbf{x}_i^\star) \leq P \sum_{i=1}^{\frac{k}{c}} f_{S_i}(\mathbf{x}_i^\star)$$

where the last inequality uses $\sum_{t=1}^{i} |V_t| = |U_i \cap O| \leq Pi$ and the following arithmetic claim due to C.4. Together with (7), we get:

$$f(O) \leq (P+1) \sum_{i=1}^{\frac{k}{c}} f_{S_i}(\mathbf{x}_i^\star)$$

as required. □

D Information Theoretic Lower Bounds

Claim D.1. *There exists a submodular function and noise distribution s.t. no randomized algorithm can obtain an approximation better than $1/2 + O(1/\sqrt{n})$ for $\max_{a \in N} f(a)$ w.h.p. using a noisy oracle.*

Proof. We will construct two functions that are identical except that one function attributes a value of 2 for a special element $x^\star$ and 1 for all other elements, whereas the other is assigns a value of 1 for each element. In addition, these functions will be bounded from above by 2 so that the only queries that give any information are those of singletons. More formally, consider the functions $f_1(S) = \min\{|S|, 2\}$ and $f_2(S) = \min\{g(S), 2\}$ where $g : 2^N \to \mathbb{R}$ is defined for some $x^\star \in N$ as:

$$g(S) = \begin{cases} 2, & \text{if } S = x^\star \\ |S|, & \text{otherwise} \end{cases}$$

The noise distribution will return 2 with probability $1/\sqrt{n}$ and 1 otherwise.

We claim that no algorithm can distinguish between the two functions with success probability greater than $1/2 + O(1/\sqrt{n})$. For all sets with two or more elements, both functions return 2, and so no information is gained when querying such sets. Hence, the only information the algorithm has to work with is the number of 1, 2, and 4 values observed on singletons. If it sees the value 4 on such a set, it concludes that the underlying function is f_2. This happens with probability $1/\sqrt{n}$.

Conditioned on the event that the value 4 is not realized, the only input that the algorithm has is the number of 1s and 2s it sees. The optimal policy is to choose a threshold, such if a number of 2s observed is or above this threshold, the algorithm returns f_2 and otherwise it reruns f_1. In this case, the optimal threshold is $\sqrt{n} + 1$.

The probability that f_2 has at most $\sqrt{n}$ twos is $1/2 - 1/\sqrt{n}$, and so is the probability that f_1 has at least $\sqrt{n} + 1$ twos, and hence the advantage over a random guess is $O(1/\sqrt{n})$ again.

An algorithm which approximates the maximal set on f_2 with ratio better than $1/2 + \omega(1/\sqrt{n})$ can be used to distinguish the two functions with advantage $\omega(1/\sqrt{n})$. Having ruled this out, the best approximation one can get is $1/2 + O(1/\sqrt{n})$ as required. □

We now construct a lower bound for general k of $(2k - 1)/2k$, where our upper bound is $(k - 1)/k$.

Claim D.2. *There exists a submodular function and noise distribution for which w.h.p. no randomized algorithm with a noisy oracle can obtain an approximation better than $(2k - 1)/2k + O(1/\sqrt{n})$ for the optimal set of size k.*

Proof. Consider the function:

$$f_1(S) = \begin{cases} 2|S|, & \text{if } |S| < k \\ 2k - 1, & \text{if } |S| = k \\ 2k, & \text{if } |S| > k \end{cases}$$

and the function f_2, which is dependent on the identity of some random set of size k, denoted $S^\star$:

$$f_2(S; S^\star) = \begin{cases} 2|S|, & \text{if } |S| < k \\ 2k - 1, & \text{if } |S| = k, S \neq S^\star \\ 2k, & \text{if } S = S^\star \\ 2k, & \text{if } |S| > k \end{cases}$$

Note that both functions are submodular.

The noise distribution will return $2k/(2k - 1)$ with probability $n^{-1/2}$ and 1 otherwise. Again we claim that no algorithm can distinguish between the functions with probability greater than $1/2$. Indeed, since f_1, f_2 are identical on sets of size different than k, and their value only depends on the set size, querying these sets doesn't help the algorithm (the oracle calls on these sets can be simulated). As for sets of size k, the algorithm will see a mix of $2k - 1$, $2k$, and at most one value of $4k^2/(k - 1)$. If the algorithm sees the value $4k^2/(k - 1)$ then it was given access to f_2. However, the algorithm will see this value only with probability $1/\sqrt{n}$. Conditioning on not seeing this value, the best policy the algorithm can adopt is to guess f_2 if the number of $2k$ values is at least $1 + \frac{\binom{n}{k}}{\sqrt{n}}$, and guess f_1 otherwise. The probability of success with this test is $1/2 + O(1/\sqrt{n})$ (regardless of whether the underlying function is f_1 or $f - 2$). Any algorithm which would approximate the best set of size k to an expected ratio better than $(2k - 1)/2k + \omega(1/\sqrt{n})$ could be used to distinguish between the function with an advantage greater than $1/\sqrt{n}$, and this puts a bound of $(2k - 1)/2k + O(1/\sqrt{n})$ on the expected approximation ratio. □

We note that if the algorithm is not allowed to query the oracle on sets of size greater than k, Claim D.1 can be extended to show a $\Omega(1/n)$ inapproximability, so choosing a random element is almost the best possible course of action.

E From maximizing f to maximizing $\tilde{f}$

Similar to the previous section, let f be a submodular function, let g be a function which is derived from f by sampling for each $x \subset [n]$ a function $h \in_{\mathcal{D}} \mathcal{H}$ and setting $g(x) = h(f(x))$. In this section we assume that the family $\mathcal{H}$ consists of monotone concave functions. We are trying to maximize g under an intersection of matroids $\mathcal{F}$. Suppose that we are allowed unlimited oracle access to f, but only n^c oracle invocation of g for some $c > 0$. Let $\mathcal{ALG}(n^c)$ be the following algorithm:

1. Find sets $S_1, S_2, \dots S_{n^c}$ such that $S_i = \arg\max_{S \in [n], S \in P, S \neq S_1, S_2 \dots S_{i-1}}$.
2. Output $\arg\max_{S_i} g(S_i)$

Lemma E.1. *Algorithm $\mathcal{ALG}(n^c)$ is optimal if we are only allowed n^c oracle invocations of g.*

Note that we are not necessarily finding the optimal set, but this is the best one can do in this setting.

To set a more realistic model, let $S* = \text{argmax}_{S \subset [n],\ S \in \mathcal{F}} f(S)$, and suppose that we are given a set $\tilde{S} \in \mathcal{F}$, $|\tilde{S}| \geq 1.01 c \log / (\epsilon \log(1/\epsilon))$ such that $f(\tilde{S}) \geq \alpha f(S*)$ for some $\alpha > 0$.

We are still allowed n^c oracle calls to g, but we are not allowed any oracle calls to f. Let $\mathcal{ALG}$ be the following algorithm:

1. Find sets $S_1, S_2, \dots S_{n^c} \subset \tilde{S}$ with no repetitions, such that S_i is chosen at random between all sets with maximal intersection size with $\tilde{S}$.
2. Output $\arg\max_{S_i} g(S_i)$

Lemma E.2. *Algorithm $\mathcal{ALG}$ gives an $\alpha(1-\epsilon)$ approximation to $\mathcal{ALG}(n^c)$.*

Proof. Let $\tilde{S}^{-j}$ be all the subsets of $\tilde{S}$ of size $k - j$. We have $|S^{-j}| = \binom{k}{j}$. Let $\alpha = 1.01clogn/log(1/\epsilon)$. We claim that in step 1 of $\mathcal{ALG}$ the set S_{n^c} has at least $k - \alpha$ elements. To see this, we let

$$\begin{aligned}\log(\sum_{j=1}^{\alpha}\binom{k}{j}) &\geq \log(\binom{k}{\alpha}) \\ &\geq 0.999kH(\alpha/k) \geq 0.999k(\alpha/k)\log(k/\alpha) \\ &\geq 0.999\alpha\log(k/\alpha) \geq 0.999\alpha\log(1/\epsilon) \geq c\log n\end{aligned}$$

So there are at least n^c different sets created in the first step. As $\alpha \leq |\tilde{S}|/\epsilon$, the expected value of $f(S_i)$ is at least $(1-\epsilon)f(S)$. The expected value of running $\mathcal{ALG}$ is at least the maximum of $h_1((1-c/k)f(S)), \dots h_{n^c}((1-c/k)f(S))$ where h_i is sampled independently from $\mathcal{H}$ according to $\mathcal{D}$. Since each h_i is convex, in expectation this is at least the maximum of n^c samples of the form $h((1-\epsilon)f(S))$ where h is sampled independently each time. □

If $\mathcal{H}$ is bounded and independent of n, then we get

Lemma E.3. *Algorithm $\mathcal{ALG}$ gives an $(1-2\epsilon)$ approximation to the optimal value.*

The proof relies on the fact that if one samples $\mathcal{H}$ enough times then one gets a value which is a $1-\epsilon$ approximation to the optimal value. We note that in this case it is enough to sample $\mathcal{H}$ any super constant number of times (when n is large enough), so we no longer need any requirement on the size of $\tilde{S}$.

F Extensions

F.1 Additive approximations

Throughout this paper we assumed the approximation is multiplicative, i.e. we defined the oracle to return $\tilde{f}(S) = \xi_S \cdot f(S)$. An alternative model is one where the approximation is *additive*, i.e. $\tilde{f}(S) = f(S) + \xi_S$, where $\xi_S \sim \mathcal{D}$. We note that the impossibility results for functions that are ϵ-close to submodular also apply to additive approximations.

From a technical perspective, the problem remains non-trivial. Fortunately, all the algorithms described above apply to the additive noise model, modulo the weak concentration bounds which become straightforward:

$$\begin{aligned}
\tilde{F}(S \cup \mathbf{x}) &= \frac{1}{|\mathcal{B}_S(\mathbf{x})|} \sum_{\mathbf{z} \in \mathcal{B}_S(x)} \tilde{f}(S \cup \mathbf{z}) \\
&= \frac{1}{|\mathcal{B}_S(\mathbf{x})|} \sum_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} (f(S \cup \mathbf{z}) + \xi_{S \cup \mathbf{z}}) \\
&= \frac{1}{|\mathcal{B}_S(\mathbf{x})|} \left(\sum_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} f(S \cup \mathbf{z}) + \sum_{\mathbf{z} \in \mathcal{B}_S(\mathbf{x})} \xi_{S \cup \mathbf{z}} \right)
\end{aligned}$$

Thus, by applying a concentration bound we can show that a bundle $\mathbf{x}$ whose noisy mean value is maximal implies that its non-noisy mean marginal contribution $F_S(\mathbf{x})$ is approximately maximal.

F.2 Correlated Approximations

No algorithm can optimize a monotone submodular function under a cardinality constraint given access to a approximate oracle whose multipliers are arbitrarily correlated across sets, even when the support of the distribution is arbitrarily small (this is implied from [HS17]). In light of this, one may wish to consider special cases of correlated distributions.

Guarantees for d-correlated distributions. Our algorithms can be extended to a model in which querying similar sets may return results that are arbitrarily correlated, as long as querying sets which are sufficiently far from each other gives independent answers.

Definition. *We say that the samples are d-**correlated** if for any two sets S and T, such that $|S \setminus T| + |T \setminus S| > d$ we have that the approximation is applied independently to S and to T.*

For this model we show that we can obtain an approximation arbitrarily close to $1 - 1/e$ for $O(1)$-correlated noise when $k \geq 2$.

Modification of algorithms for small k for $O(1)$-correlated noise. We add $c \gg d/\epsilon$ elements at each phase of the algorithm. We modify the definition of $\tilde{F}$ in the following way. First we take a an arbitrary partition $P_1, \ldots P_{(n-|S|)/d}$ on the elements not in S, in which each P_i is of size d, and a partition $Q_1 \ldots Q_{(|S|+|A|)/d}$ of the elements in $S \cup A$. We estimate the value of a set A given S using:

$$\tilde{F}(S \cup A) = \frac{d^2}{(|S| + |A|)(|N| - |S| - |A|)} \sum_{Q_i \in A} \sum_{P_j} \tilde{f}(((S \cup A) \setminus Q_i) \cup P_j)$$

and modify the rest of the algorithm accordingly.

Correctness relies on three steps:

1. First, when we are in iteration i of the algorithm (after we already added $(i-1)c$ elements to S), all the sets we apply the oracle on are of size $c \cdot i$, and hence they are independent of any set of size $c(i-1)$ or less which were used in previous phases;

2. Second, when we evaluate $\tilde{F}(S \cup A)$ for a specific set A, we only use sets which are independent in the comparison. Here we rely on changing d elements in A each time, and replacing them by another set of d elements;
3. Finally, we treat each set A separately, and show that if its marginal contribution is negligible then w.h.p its mean smooth value is not too large, and if its marginal contribution is not negligible, then w.h.p. $\tilde{F}(S \cup A)$ approximates $F(S \cup A)$ well. Taking a union bound over all the bad events we get that the set A chosen has large (non-noisy) smooth mean value.

G More related work

The seminal works of [NWF78b, FNW78] show that the greedy algorithm gives a factor of $1-1/e$ for maximizing a submodular function under a cardinality constraint and a factor $1/(P+1)$ approximation for intersection of P matroids. Thirty years later Vondrak [Von08] proposed the continuous greedy algorithm, which achieves a $1-1/e$ approximation for Matroid constraints. This guarantee is optimal in the value oracle model, even for simple problems like max-cover [Fei98, MSV08, KLMM05, NW78].

Other cases have different approximation constants. One can do better than $1-1/f$ with demand queries [FV06]. Non monotone functions have also been studied [FMV11, LMNS09, BFNS12, BFNS14] but we leave approximate versions of such functions to future research. As for techniques for submodular optimization which resemble ours, we note rounding frameworks, optimization of multilinear relaxations and convex relaxations [AS04, CE11, CCPV07, Von08, CJV15, VCZ11].

On the applications front, Submodular utilities have been studied extensively in game theory (e.g. [DNS05, DS06, DLN08, MSV08, BDF$^+$10, DFK11, PP11, DRY11, DV12, PSS08, SS08, BSS10, LSST13]. However, lately it is more and more common to assume that valuations are not exactly submodular [FFI$^+$15]. Similarly, many problems in ML are modeled via submodular functions (e.g. identifying influencers in social networks [KKT03, RLK11] sensor placement [LKG$^+$07, GFK10], learning in data streams [SGK09, GK10, KMVV13, BMKK14], information summarization [LB11a, LB11b], adaptive learning [GK11], vision [JB11b, JB11a, KOJ13], and general inference methods [KG07, JB11a, DK14]). We note that these functions are often inferred from data, and may not be exactly submodular.

We especially mention the notion of *Probably Mostly Approximately Correct* (PMAC) learnability due to Balcan and Harvey [BH11]. Informally, PMAC-learnability guarantees that after observing polynomially-many samples of sets and their function values, one can construct a surrogate function that is with constant probability over the distributions generating the samples, likely to be an approximation of the submodular function generating the data. If we assume that this function is generated in an i.i.d manner we can optimize it.

G.1 Comparison to with previous work

The algorithms we develop here are very different from the one in [HS17]. At a high level, the idea in [HS17] is to choose an arbitrarily set of elements H of size $\Omega(\frac{\log\log n}{\epsilon^2})$ for a technique called *smoothing*. Smoothing with the set H refers to the process of greedily selecting the element $a \in N \notin H$ that maximizes $\sum_{H' \subseteq H} \tilde{f}(S \cup H' \cup a)$ in every iteration, when S is the solution used in previous iterations. The size of H is determined to be $\Omega(\frac{\log\log n}{\epsilon}^2)$ for concentration bounds to hold. For the approximation guarantees to go through, it is necessary to include the set H in the solution. For this reason, this technique fundamentally requires a cardinality $\Omega(\frac{\log\log n}{\epsilon^2})$ and does not hold for matroids (consider a matroid in which committing to H prevents adding elements with high value).

The algorithm described here adds a constant number of elements at once, and does not commit to selecting any particular set a priori. This lets us prove far stronger results. The new algorithm works for any k, and not just for $k \gg \log\log n$. As our lower bounds show, somewhat counter-intuitively, the the smaller cardinality constraint is, the more difficult the problem becomes as there is less room for error. This technique is applicable for intersections of matroids. This relies on not committing to any arbitrary set. We note that randomness would not salvage the algorithm presented in [HS17].