2015 IEEE 56th Annual Symposium on Foundations of Computer Science

Approximate Modularity

Flavio Chierichetti*, Abhimanyu Das†, Anirban Dasgupta‡, Ravi Kumar†
* *Sapienza University of Rome*
Rome, Italy
Email: flavio@di.uniroma1.it
† *Google*
Mountain View, CA
Email: abhi.das@gmail.com, ravi.k53@gmail.com
‡ *Indian Institute of Technology*
Gandhinagar, India
Email: anirban.dasgupta@gmail.com

Abstract

A set function on a ground set of size n is approximately modular if it satisfies every modularity requirement to within an additive error; approximate modularity is the set analog of approximate linearity. In this paper we study how close, in additive error, can approximately modular functions be to truly modular functions.

We first obtain a polynomial time algorithm that makes $O(n^2 \log n)$ queries to any approximately modular function to reconstruct a modular function that is $O(\sqrt{n})$-close. We also show an almost matching lower bound: any algorithm would need superpolynomially many queries to construct a modular function that is $o\left(\sqrt{n/\log n}\right)$-close.

In a striking contrast to these near-tight computational reconstruction bounds, we then show that for any approximately modular function, there exists a modular function that is $O(\log n)$-close.

I. Introduction

A real-valued set function f is said to be modular if it satisfies the following *modularity* property for all subsets X and Y of a ground set:

$$f(X) + f(Y) = f(X \cup Y) + f(X \cap Y).$$

Modular functions arise in a number of discrete optimization settings [10], [23], [25], [31]; a function is modular if it is both submodular and supermodular. Modularity is the set analog of linearity and the latter has been extensively studied in the context of property testing [4]; see also the survey [28].

In this paper we study the properties of approximately modular functions, i.e., functions that satisfy the modularity property to within an additive error. Our motivation for studying approximate modularity arises from the following. Firstly, properties similar to approximate modularity have been studied extensively in property testing and in the area of sublinear algorithms [5], [6], [9], [12], [16], [18], [26], [29]. They have also been studied in functional analysis, in the context of the stability of functional equations [7], [13]. Secondly, there has been recent interest in the machine learning community on approximately submodular functions [19], [20] and the so-called submodularity ratio [8], [21]. The goal there is to push the boundaries of greedy algorithms to settings where the submodularity property is only loosely satisfied. Unfortunately, the theory of approximate submodularity developed in these papers is rudimentary and the bounds obtained are pessimistic. Furthermore, there are diverging definitions of approximate submodularity that these papers have promoted—from additive relaxations of the submodularity conditions [20] to submodularity ratios [8], [21]—and it is not obvious which of the definitions is preferable. We believe that understanding

* Work done in part while visiting Google. Supported in part by a Google Focused Research Award.

0272-5428/15 $31.00 © 2015 IEEE
DOI 10.1109/FOCS.2015.74

approximate modularity is the first step towards developing a richer theory of approximate submodularity. The additive definition that we use is the discrete analog of the one used in functional analysis results on approximate linearity and convexity [7], [13], [15]. We also believe that our results indicate that this additive definition of approximate modularity (and by extension, that of approximate submodularity) could be the most theoretically appealing one[1]. Lastly, we think that approximate modularity is interesting in its own right as a basic and natural question; it was a bit surprising that this question had not been addressed before.

Our contributions: We address the following specific question in this paper: given an arbitrary approximately modular function f over a ground set of size n, how close (in the ℓ_∞-distance) is f to a modular function and how efficiently can we reconstruct a close modular function? Our main result is that for each approximately modular function f there exists a modular function g that is $O(\log n)$-close to f (Section IV). We use a probabilistic method to establish this result, and the reconstruction algorithm is not efficient. We also obtain a polynomial time algorithm (Section III) for reconstruction that is based on a careful self-correction applied to the approximately modular function. However, this algorithm only obtains a modular function that is $O(\sqrt{n})$-close. This exponential loss with respect to the existential result might seem disappointing, but it is inevitable: we show (Section V) that no polynomial time algorithm can obtain a function that is better than $o\left(\sqrt{n/\log n}\right)$-close to f.

We first give an overview of the technical steps in obtaining the $O(\sqrt{n})$-close algorithm. An easy observation is that a modular function is completely characterized by its values on the singletons and on the empty set. It is then easy to see that by induction one can get an $O(n)$-close modular approximation of f. Obtaining any improvement over this bound seems challenging. A natural attempt is to define a modular function by analyzing the marginals of each of the elements induced by the various sets, i.e., one obtains the weight of element i by aggregating the values of the multiset $\{f(X \cup \{i\}) - f(X) \mid X \subseteq [n] \setminus \{i\}\}$. We propose a weighted averaging of the elements of this multiset, which results in a modular function that is $O(\sqrt{n})$-close to f. To prove this bound, we define an appropriate linear program and use duality to obtain an upper bound on the gap between f and the proposed modular function by relating the objective to a hypergeometric distribution.

Our $O(\log n)$ existential result is obtained through a different approach. First, we observe that it is enough to concentrate on functions f that have the all-zeros function as their closest modular approximation, and hence their distance to modularity is the same as their maximum absolute value. Then, we characterize the subsets of the ground set where such an f attains its maximum and the minimum values: we show there exist probability distributions on these two classes of subsets that have the same marginals over the elements. We then create a system of union-intersection combinations (of logarithmic depth) and show that, with high probability, we can generate the same sets from each of the two extremal classes. This implies a bound on the number of times the approximate modular inequality has to be used in order to relate the smallest and largest values of f, which entails a bound on the gap between f and its modular approximation. The main technical novelty is the construction of the system of union-intersection combinations. This system produces, with high probability, approximately the same output sets from the two sides as long as the inputs are drawn from distributions with the same marginals over the elements. We conjecture that the $O(\log n)$ existential bound is optimal.

An additional implication of our results is to characterize the hardness of maximizing (or minimizing) an approximately modular function. Our algorithm immediately implies a polynomial time algorithm for finding out the maximum or minimum (possibly under some cardinality constraint), to an additive $\sqrt{n}$ approximation (which is also multiplicative if the optimal value is $\Omega(\sqrt{n})$), simply by constructing

[1]For some of the other variants, e.g., a multiplicative definition, it is easy to show that desirable mathematical properties such as the existence of a close modular function or a nontrivial approximation in polynomial time, do not exist. Furthermore, multiplicative definitions do not make sense if the function can take zero, positive, or negative values. However, this is not to say that there is no other *nice* definition for approximate modularity or submodularity.

and optimizing the corresponding modular function. More interestingly, our query lower bound for reconstruction shows that the maximum (or minimum) of an approximately modular function cannot be additively-approximated to better than $\Omega(\sqrt{n/\log n})$ in polynomial time. In other words, the algorithm in Section III is almost tight even for approximating the maximum (or minimum) of an approximately modular function.

Related work: As mentioned earlier, the stability of approximate functional properties has been studied in the classical Hyers–Ulam–Rassias theory. In the linearity case, the question addressed in these papers is the same as ours: given a function on Banach space that is approximately linear, is it close to a linear function? In a celebrated paper, Hyers [13] showed that the distance to the closest linear function is independent of the domain size. There have been several extensions of this basic result to other function families including approximate convexity [7], [15] and to convex domains [15], [22]; for a detailed survey on the known results, see the monographs [14], [17]. Our definition of approximate modularity follows the same additive notion that approximate linearity and convexity use. Unfortunately, the machinery in proving these results heavily rely of the niceness of the domain, i.e., the fact that it is a Banach space or that it is convex. This allows the usage of limiting arguments or of arguments that depend on an infinitesimal division of the domain; in other words, this allows the assumption that the domain has no "holes". Our domain, on the other hand, is a lattice and it is *not* a closed set (let alone convex). To circumvent this seemingly technical issue, one might try to extend the approximate modular functions to approximately linear functions (on a suitable convex domain), using their Lovász extensions or their multilinear extensions, and then apply known results on the stability of approximately linear functions. Unfortunately, the continuous extension of an approximately modular function with a unit additive error could be, in principle, very far from it. Existing results do not seem to give an upper bound on the distance better than $O(n)$; it is interesting to note that, in fact, our work entails that the closest continuous extension is no further than $O(\log n)$.

There has been a lot of work on testing and reconstructing important families of functions. A partial list of function families with efficient testers and correctors includes: linear functions [2]–[4], low-degree polynomials [29], functional equations [27], monotone functions [5], [6], [12], [30], convex and submodular functions [26], Lipschitz functions [6], [16], etc. Approximately linear functions and approximately low-degree polynomials were studied in [9], [18]. In these works, the question is if a given function is close (in Hamming distance) to a given family or far from it. Our question is different in that we want to understand the (ℓ_∞) distance between the families of approximately modular functions and modular functions. Furthermore, the techniques there are tailored more towards well-behaved convex domains with algebraic properties.

Our upper bound learns the closest modular function using polynomially many queries and obtaining an additive $\sqrt{n}$ approximation. Similar learning questions have been studied for submodular functions by Goemans et al. [11] (in the arbitrary query model) and Balcan et al. [1] (in the PMAC model) who show that using a polynomial number of queries, one can get multiplicative approximations of monotone-increasing submodular functions. As we will note, getting a multiplicative approximation in polynomial time is impossible in our setting.

II. Preliminaries

Let $[n] = \{1, \ldots, n\}$ and, for a set S, let $2^S = \{T \mid T \subseteq S\}$. Let $f : 2^{[n]} \to \mathbb{R}$ be a set function defined on the subsets of $[n]$. We recall the definition of modular functions, which are set analogs of linear functions:

Definition 1 (Modularity)**.** *A set function $f : 2^{[n]} \to \mathbb{R}$ is* modular *if for all subsets $X, Y \subseteq [n]$, it satisfies the* modularity *property:*

$$f(X) + f(Y) = f(X \cup Y) + f(X \cap Y). \tag{1}$$

An equivalent characterization of modular functions is the following: there exists weights $w_0, w_1, \ldots, w_n \in \mathbb{R}$ such that for any $X \subseteq [n]$,

$$f(X) = w_0 + \sum_{i \in X} w_i. \tag{2}$$

We will consider a setting where the modularity property is satisfied approximately.

Definition 2 (Approximate modularity)**.** *A set function f is* ϵ-approximately modular *if for all subsets $X, Y \subseteq [n]$, we have*

$$|f(X) + f(Y) - f(X \cup Y) - f(X \cap Y)| \leq \epsilon. \tag{3}$$

If $\epsilon > 0$, then a simple scaling argument allows us to assume, without loss of generality, that $\epsilon = 1$. Henceforth, we will use approximate modularity to denote 1-approximate modularity. Let $\mathcal{M}_n$ be the set of all modular functions on $[n]$ and let $\widetilde{\mathcal{M}}_n$ be the set of all approximately modular functions on $[n]$. We next define the *distance* between two set functions on the ground set $[n]$: $|f - g| = \max_{X \subseteq [n]} |f(X) - g(X)|$.

Definition 3 (Closeness)**.** *Two set functions f and g are* γ-close *if $|f - g| \leq \gamma$.*

A set function f is γ-close to modularity if there is a $g \in \mathcal{M}_n$ such that $|f - g| \leq \gamma$. The distance of a set function f to modularity is $\inf_{g \in \mathcal{M}_n} |f - g|$.

Notation: Let $E_{x \in X}[Y(x)]$ denote the average value of $Y(x)$ when x is sampled uniformly at random from X, i.e., $E_{x \in X}[Y(x)] = \sum_{x \in X} Y(x)/|X|$. Let $[\xi] = 1$ if the predicate ξ is true and 0 otherwise.

III. An $O(\sqrt{n})$-approximation algorithm

In this section we show that if f is approximately modular, then one can construct, using a polynomial number of oracle accesses to f, a modular function g that is $O(\sqrt{n})$-close to f. We will also show that the analysis of our construction is tight.

Theorem 4. *(i) For every $f \in \widetilde{\mathcal{M}}_n$, there exists a $g \in \mathcal{M}_n$ such that $|f - g| \leq O(\sqrt{n})$. (ii) Furthermore, there exists a polynomial-time algorithm that obtains such a g with probability $1 - o(1)$; this algorithm performs $O(n^2 \log n)$ non-adaptive oracle queries to f.*

The rest of the section is devoted to proving Theorem 4 by explicitly constructing a function g.

A. Proof of Theorem 4(i)

We start by defining a set of weights (Section III-D contains a description of the reasoning that led us to these weights). First $w_0 = f(\emptyset)$. For $i \in [n]$:

$$w_i = E_{k \in [n]} \left[E_{T \in \binom{[n] \setminus \{i\}}{k-1}} \left[f(T \cup \{i\}) - f(T) \right] \right] = \sum_{T \subseteq [n] \setminus \{i\}} \frac{f(T \cup \{i\}) - f(T)}{n \cdot \binom{n-1}{|T|}}. \tag{4}$$

In other words, for $i \in [n]$, w_i is defined as a natural (weighted) average: uniformly pick a level $k \in [n]$, uniformly pick a set T of cardinality $k - 1$ that does not contain i, and retrieve a value for w_i using the incremental change due to i. Then, g is defined as

$$g(S) = w_0 + \sum_{i \in S} w_i. \tag{5}$$

From the characterization in (2), we know that g is modular. We will show that $|f - g| \leq O\left(\sqrt{n}\right)$.

By collecting terms in (4) and rearranging, we have

$$g(S) - f(\emptyset) = \sum_{i \in S} w_i = \sum_{T \subseteq [n]} \left(\frac{|S \cap T|}{n\binom{n-1}{|T|-1}} - \frac{|S \setminus T|}{n\binom{n-1}{|T|}} \right) \cdot f(T) = \sum_{T \subseteq [n]} \left(\frac{\frac{|S \cap T|}{|T|} - \frac{|S \setminus T|}{n - |T|}}{\binom{n}{|T|}} \cdot f(T) \right), \tag{6}$$

where, for simplicity of notation, we have adopted the convention that $\frac{0}{0} = 0$.

Now, let $S \subseteq [n]$ be an arbitrary fixed subset. The idea is to use the linear program (7) to find an approximately modular function f that maximizes $|g(S) - f(S)|$, where g is defined from f using (4). Note that it suffices to maximize $g(S) - f(S)$; indeed, if $g(S) - f(S) < 0$, then by setting $f'(S) = -f(S)$ and by letting g' be our modular approximation to f', we obtain that $g'(S) - f'(S) = -(g(S) - f(S)) > 0$.

Consider the following LP $\mathcal{L}_S$:

$$\mathcal{L}_S : \begin{cases} & \max g(S) - f(S) \text{ subject to} \\ \forall \{X, Y\} \in \binom{2^{[n]}}{2} : & -1 \leq f(X \cup Y) + f(X \cap Y) - f(X) - f(Y) \leq 1. \end{cases} \tag{7}$$

Using the definition of g in (5) and the expression in (6), the objective in (7) can be written as:

$$g(S) - f(S) = \sum_{T \subseteq [n]} \phi(T) \cdot f(T), \tag{8}$$

where $\phi(T) = \phi_S(T) = \frac{\frac{|S \cap T|}{|T|} - \frac{|S \setminus T|}{n - |T|}}{\binom{n}{|T|}} - [T = S] + [T = \emptyset]$. Now, the dual of $\mathcal{L}_S$ is:

$$\mathcal{D}_S : \begin{cases} & \min \sum_{\{X,Y\} \in \binom{2^{[n]}}{2}} \left(\pi^+_{\{X,Y\}} + \pi^-_{\{X,Y\}} \right) \\ \forall T \subseteq [n] : & \sum_{\substack{X \subseteq [n] \\ X \neq T}} \left(\pi^+_{\{T,X\}} - \pi^-_{\{T,X\}} \right) + \sum_{\substack{\{X,Y\} \in \binom{2^{[n]}}{2} \\ T \in \{X \cup Y, X \cap Y\}}} \left(\pi^-_{\{X,Y\}} - \pi^+_{\{X,Y\}} \right) = \phi(T), \\ \forall \{X, Y\} \in \binom{2^{[n]}}{2} : & \pi^+_{\{X,Y\}}, \pi^-_{\{X,Y\}} \geq 0. \end{cases} \tag{9}$$

Observe that the optimality of a solution of $\mathcal{D}_S$ guarantees that, for each $\{X, Y\} \in \binom{2^{[n]}}{2}$, at most one of $\pi^+_{\{X,Y\}}$ and $\pi^-_{\{X,Y\}}$ is positive. Therefore, if we let $\pi_{\{X,Y\}} = \pi^+_{\{X,Y\}} - \pi^-_{\{X,Y\}}$, we can rewrite $\mathcal{D}_S$ as:

$$\mathcal{D}'_S : \begin{cases} & \min \sum_{\{X,Y\} \in \binom{2^{[n]}}{2}} \left| \pi_{\{X,Y\}} \right| \\ \forall T \subseteq [n] : & \alpha(T) - \beta(T) = \phi(T), \end{cases} \tag{10}$$

where

$$\alpha(T) = \sum_{\substack{X \subseteq [n] \\ X \neq T}} \pi_{\{T,X\}}, \qquad \beta(T) = \sum_{\substack{\{X,Y\} \in \binom{2^{[n]}}{2} \\ T \in \{X \cup Y, X \cap Y\}}} \pi_{\{X,Y\}},$$

and the variables $\pi_{\{X,Y\}}$ are unrestricted.

We now provide a solution to $\mathcal{D}'_S$. Define

$$\pi_{\{X,Y\}} = \begin{cases} \binom{n-1}{|X|}^{-1} \cdot \left(\frac{|X \cap S|}{|X|} - \frac{|S|}{n} \right), & \text{if } Y = S,\ X \not\subseteq S, \text{ and } S \not\subseteq X, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We will now show that this assignment satisfies all the constraints (Lemma 5). Then, we will obtain an upper bound on the value of the objective function with this assignment (Lemma 6).

Lemma 5. *The assignment in (11) is a feasible solution to (10).*

Proof: We proceed by a case analysis. If $T = S$, then we have:

$$\alpha(S) = \sum_{i=0}^{|S|-1} \sum_{d=1}^{n-|S|} \left(\binom{|S|}{i} \binom{n - |S|}{d} \binom{n-1}{i+d}^{-1} \left(\frac{i}{i+d} - \frac{|S|}{n} \right) \right) = \binom{n}{|S|}^{-1} - 1,$$

and $\beta(S) = 0$. We also have that $\phi(S) = \binom{n}{|S|}^{-1} - 1$ and hence the statement holds in this case.

Now consider the T's that satisfy $T \not\subseteq S$ and $T \not\supseteq S$. We have:

$$\alpha(T) = \binom{n-1}{|T|}^{-1}\left(\frac{|T\cap S|}{|T|} - \frac{|S|}{n}\right) = \binom{n}{|T|}^{-1}\frac{n}{n-|T|}\left(\frac{|T\cap S|}{|T|} - \frac{|S|}{n}\right)$$
$$= \binom{n}{|T|}^{-1}\left(\frac{|T\cap S|}{|T|} + \frac{|T\cap S|}{n-|T|} - \frac{|S|}{n-|T|}\right) = \binom{n}{|T|}^{-1}\left(\frac{|T\cap S|}{|T|} - \frac{|S\setminus T|}{n-|T|}\right),$$

and $\beta(T) = 0$ (since the sum that makes up $\beta(T)$ can have non-zero values only in its terms indexed by the sets X such that $T \in \{X\cup S, X\cap S\}$, and $T \not\supseteq S, T \not\subseteq S$ implies that $T \notin \{X\cup S, X\cap S\}$.) Therefore, for each T such that $T \not\supseteq S, T \not\subseteq S$, the statement holds.

Finally, consider $T \subset S$ or $T \supset S$. In either case, we have $\alpha(T) = 0$. If $T \subset S$, then

$$\beta(T) = \sum_{d=1}^{n-|S|}\left(\binom{n-|S|}{d}\cdot\frac{\frac{|T|}{|T|+d} - \frac{|S|}{n}}{\binom{n-1}{|T|+d}}\right) = -\frac{1-\frac{|S|-|T|}{n-|T|}}{\binom{n}{|T|}} = -\phi(T).$$

If $T \supset S$, then we have:

$$\beta(T) = \sum_{i=0}^{|S|-1}\left(\binom{|S|}{i}\cdot\frac{\frac{i}{|T|-|S|+i} - \frac{|S|}{n}}{\binom{n-1}{|T|-|S|+i}}\right) = -|S|\cdot\frac{\frac{1}{|T|} - \frac{1}{n}}{\binom{n-1}{|T|}} = -\frac{|S|}{|T|}\cdot\binom{n}{|T|}^{-1} = -\phi(T).$$

Hence, the statement holds in this case as well. ∎

Lemma 6. *Under the assignment in (11), the value of (10) is at most* $4\sqrt{\min(|S|, n-|S|)}$.

Proof: Let $s = |S|$. Then,

$$\sum\left|\pi_{\{X,Y\}}\right| = \sum_{i=0}^{s-1}\sum_{d=1}^{n-s}\left(\binom{s}{i}\binom{n-s}{d}\frac{\left|\frac{i}{i+d} - \frac{s}{n}\right|}{\binom{n-1}{i+d}}\right) \;\le\; 2\sum_{i=0}^{s-1}\sum_{d=1}^{n-s}\left(\binom{s}{i}\binom{n-s}{d}\frac{\left|\frac{i}{i+d} - \frac{s}{n}\right|}{\binom{n}{i+d}}\right)$$
$$\le 2\sum_{x=1}^{n}\sum_{i=0}^{\min(x,s)}\left(\frac{\binom{s}{i}\binom{n-s}{x-i}}{\binom{n}{x}}\left|\frac{i}{x} - \frac{s}{n}\right|\right) \;=\; 2\sum_{x=1}^{n}E\left[\left|\frac{H}{x} - \frac{s}{n}\right|\right],$$

where H denotes the following hypergeometric random variable: the number of red balls in a uniform at random sample (without replacement) of x balls from an urn containing n balls, s of which are red.

It is easy to see that $E[H] = xs/n$ and hence $E\left[\frac{H}{x} - \frac{s}{n}\right] = 0$. Now,

$$\mathrm{var}\left(\frac{H}{x}\right) = \frac{1}{x^2}\cdot x\cdot\frac{s}{n}\cdot\frac{n-s}{n}\cdot\frac{n-x}{n-1} \le \frac{\min(s,n-s)}{xn}\cdot\frac{\max(s,n-s)}{n}\cdot\frac{n-x}{n-1} \le \frac{\min(s,n-s)}{xn}.$$

Applying Jensen's inequality, we get

$$E\left[\left|\frac{H}{x} - \frac{s}{n}\right|\right] \le \sqrt{\mathrm{var}\left(\frac{H}{x}\right)} \le \sqrt{\frac{\min(s,n-s)}{xn}}.$$

We can then conclude

$$\sum\left|\pi_{\{X,Y\}}\right| \le 2\sum_{x=1}^{n}E\left[\left|\frac{H}{x} - \frac{s}{n}\right|\right] \le 2\sqrt{\frac{\min(s,n-s)}{n}}\sum_{x=1}^{n}\frac{1}{\sqrt{x}}.$$

The claim follows from $\sum_{x=1}^{n}\frac{1}{\sqrt{x}} \le 2\sqrt{n}$. ∎

Thus, Lemma 5 and Lemma 6 imply Theorem 4(i). In fact, Lemma 6 shows the stronger claim that for a subset S, we will have $|g(S) - f(S)| \le O\left(\sqrt{\min(|S|, n-|S|)}\right)$, i.e., the values of subsets of cardinality $o(n)$ and those of subsets of cardinality $n - o(n)$ will be approximated to $o(\sqrt{n})$.

B. Proof of Theorem 4(ii)

We now show that we can approximate the weights of g to obtain a function $\tilde{g} \in \mathcal{M}_n$ such that $|\tilde{g} - g| \leq O(\sqrt{n})$ (and thus $|\tilde{g} - f| \leq O(\sqrt{n})$) using $O\left(n^2 \log n\right)$ non-adaptive oracle queries to f.

Lemma 7. *There exists a randomized algorithm that for any $i \in [n]$ produces $\tilde{w}_i$ such that $\Pr[|w_i - \tilde{w}_i| \leq n^{-1/2}] \geq 1 - n^{-2}$ using $O(n \log n)$ non-adaptive oracle queries to f.*

Proof: For $i \in [n]$, let $\delta_i(X) = f(X \cup \{i\}) - f(X)$. Observe that for each $i \in [n]$ and for each $X \subseteq [n] \setminus \{i\}$, it holds by (3) that:

$$f(\{i\}) - f(\emptyset) - 1 \leq \delta_i(X) \leq f(\{i\}) - f(\emptyset) + 1.$$

Recall that w_i is the expectation of $\delta_i(X)$ under a suitable probability distribution over the subsets $X \subseteq [n] \setminus \{i\}$; the distribution is the one used in (4): we first pick $k \in [n]$ independently and uniformly at random and then $X \in \binom{[n]\setminus\{i\}}{k-1}$ also independently and uniformly at random.

If $X_1, \ldots, X_t$ are sets independently picked according to the above distribution, then we have that the random variables $\delta_i(X_j)$ are also iid, and for each $1 \leq j \leq t$, with probability 1 it holds that $\delta_i(X_j) \in [f(\{i\}) - f(\emptyset) - 1, f(\{i\}) - f(\emptyset) + 1]$. Therefore, $\mathrm{var}(\delta_i(X_j)) \leq 1$. Let $\tilde{w}_i = \sum_{j=1}^{t} \delta_i(X_j)/t$. By applying Bernstein's inequality, we get that:

$$\Pr\left[|w_i - \tilde{w}_i| \geq \frac{1}{\sqrt{n}}\right] \leq 2e^{-\frac{2t^2(1/\sqrt{n})^2}{4t}} = 2e^{-\frac{t}{2n}}.$$

Choosing $t = \lceil 4n \ln(2n) \rceil$ gives our claim. ∎

Since one has to learn n different w_i's, the maximum additional additive error to g is at most $n \cdot (1/\sqrt{n})$ with probability at least $1 - n^{-1}$, and we obtain Theorem 4(ii).

C. A tight lower bound on the error of g

We now give a family of functions that prove that our analysis of the level-average function g is tight. Fix an even $n \geq 2$, and let f be the function given by

$$f(S) = \frac{1}{2} \cdot \left[\left|S \cap \left\lceil \frac{n}{2} \right\rceil\right| > \left|S \setminus \left\lceil \frac{n}{2} \right\rceil\right|\right]. \tag{12}$$

Since, for each $S \subseteq [n]$, it holds that $f(S) \in \left\{0, \frac{1}{2}\right\}$, we have that f is 1-approximately modular.

We now bound the values of the weights $w_1, \ldots, w_{n/2}$ of our approximation g of the above f.

Lemma 8. *Using (4), $w_i = \Omega\left(n^{-1/2}\right)$, for $i \in \left\lceil \frac{n}{2} \right\rceil$.*

Proof: Fix $i \in \left\lceil \frac{n}{2} \right\rceil$ and let $X \subseteq [n] \setminus \{i\}$. By construction, $f(X \cup \{i\}) - f(X)$ is $\frac{1}{2}$ if $\left|X \cap \left\lceil \frac{n}{2} \right\rceil\right| = \left|X \setminus \left\lceil \frac{n}{2} \right\rceil\right|$ and 0 otherwise. Now, for an even $0 \leq \ell < n$, let $\mathcal{B}_\ell$ be the class of sets $Y \in \binom{[n]\setminus\{i\}}{\ell}$ for which $\left|Y \cap \left\lceil \frac{n}{2} \right\rceil\right| = \left|Y \setminus \left\lceil \frac{n}{2} \right\rceil\right|$. It is easy to see that $|\mathcal{B}_\ell| = \binom{n/2-1}{\ell/2} \cdot \binom{n/2}{\ell/2}$. Then, the weight of the i^{th} element is

$$w_i = \frac{1}{2n} \sum_{k=0}^{n/2-1} \frac{|\mathcal{B}_{2k}|}{\binom{n-1}{2k}} = \frac{1}{2n} \sum_{k=0}^{n/2-1} \frac{\binom{n/2-1}{k}\binom{n/2}{k}}{\binom{n-1}{2k}} \geq \frac{1}{2n}\, \Omega\left(\sum_{k=1}^{n/4} \frac{1}{\sqrt{k}}\right) \geq \Omega\left(\frac{1}{\sqrt{n}}\right).$$

∎

From this and using the definition of g in (5), we have $g\left(\left\lceil \frac{n}{2} \right\rceil\right) \geq \frac{n}{2} \cdot \Omega(n^{-1/2}) = \Omega(\sqrt{n})$. On the other hand, $f\left(\left\lceil \frac{n}{2} \right\rceil\right) = \frac{1}{2}$. Hence, $|f - g| \geq \Omega(\sqrt{n})$.

D. Motivating g as limit of an averaging process

We describe an interesting relationship between the w_i's and the limit of an averaging process that we now describe. First we note that the definition of w_i's is reminiscent of Shapley value computation [32],

Given f, we will define an infinite sequence of functions from $2^{[n]}$ to $\mathbb{R}$. The first function $f^{(0)}(S)$ is defined to be equal to $f(S)$ for each $S \subseteq [n]$. Given a function $f^{(t)}$, we define the function $f^{(t+1)}$ through a linear operator: for each $S \subseteq [n]$,

$$f^{(t+1)}(S) = 2^{-n} \cdot \sum_{T \subseteq [n]} \left(f^{(t)}(S \cup T) + f^{(t)}(S \cap T) - f^{(t)}(T) \right).$$

That is, we compute the value $f^{(t+1)}(S)$ by averaging the results of the applications of the $f^{(t)}$-approximately modular property on S and T, for every set $T \subseteq [n]$.

Observe that, since g is modular, we have $g(S \cup T) + g(S \cap T) - g(T) = g(S)$. Therefore, the function g is a fixed point of the the above linear operator. The following claim, stated as a conjecture in our submitted manuscript, was proved by one of the reviewers.

Theorem 9. $\lim_{t \to \infty} f^{(t)} = g$.

E. Median

Another natural approach for approximating the function f is to take the median of the multiset $M_i = \{f(X \cup \{i\}) - f(X) \mid X \subseteq [n] \setminus \{i\}\}$ as the weight of element w_i. Unfortunately, this approach, for some f's, returns modular functions that are no better than $\Omega(n)$-close to f.

For instance, let $f(X) = \lfloor |X|/3 \rfloor$. This function is clearly $O(1)$-close to modularity (the modular function $|X|/3$ is $2/3$-close to f), and it is therefore $O(1)$-approximately modular. On the other hand, it is easy to check that the multiset M_i contains $\frac{2}{3} \cdot 2^{n-1} \pm O(1)$ many 0's, and $\frac{1}{3} \cdot 2^{n-1} \pm O(1)$ many 1's. Therefore its median is 0, and the median approximation of f would be the 0 function, which is $\Omega(n)$-far from f.

IV. An existential $O(\log n)$ upper bound

In this section we show our main result: if f is approximately modular, then there is a modular function h that is $O(\log n)$-close to f. However, unlike in Section III, our proof will not actually give an explicit construction for such a function h. Instead, we will show that h exists by showing that if one searches for the function f that is furthest from modularity, then

(i) it suffices to consider the f's whose best modular approximation is the all-zeros function,
(ii) the maximum value of such an f is its distance to modularity, and most importantly,
(iii) this maximum value is $O(\log n)$.

In Section IV-A we will study (i) through an LP that will translate the property of being best approximated by the all-zeros function into two probability distributions, which we will use to construct the heart of our argument: a random system of inequalities to prove (iii). From an algorithmic perspective, our upper bound proves that the natural LP whose output is the best modular approximation to f (see (13)) produces a modular function that is $O(\log n)$-close to f. However, this LP has size $2^{\Theta(n)}$ and therefore naively solving it will take exponential time.

A. Functions approximable by the all-zeros function

For a generic approximately modular function $f \in \widetilde{\mathcal{M}}_n$, consider its closest modular approximation $c(S) = w_0 + \sum_{i \in S} w_i$, parametrized by $w_0, w_1, \ldots, w_n$. For the function f, let $U(f)$ be its largest value and let $L(f)$ be its smallest value, i.e., $U(f) = \max_{S \subseteq [n]} f(S)$ and $L(f) = \min_{S \subseteq [n]} f(S)$. Let $M(f) = \max_{S \subseteq [n]} |f(S)| = \max(|U(f)|, |L(f)|)$ be its maximum absolute value. Moreover, let $\mathcal{U}(f) = \{S \mid f(S) = U(f)\}$ and $\mathcal{L}(f) = \{S \mid f(S) = L(f)\}$ be the sets on which the maximum and the minimum values are attained.

Note that, if we define a function f' as $f' = f - c$, then $\inf_{h\in\mathcal{M}} |f' - h| = \inf_{h\in\mathcal{M}} |f - (c + h)| = \inf_{h\in\mathcal{M}} |f - h|$, since $c \in \mathcal{M}$. Hence, the distance of f' to modularity is the same as that of f, and the all-zeros function is the closest modular approximation of f'. Therefore, if we show that every approximately modular function whose best modular approximation is the all-zeros function ($z \equiv 0$) is γ-close to modularity, then we actually show that every approximately modular function is also γ-close to modularity.

We now show a crucial characterization of approximately modular functions whose best approximation is the all-zeros function z.

Lemma 10. *Let $n \geq 2$ and $f : 2^{[n]} \to \mathbb{R}$ be a set function. The following are equivalent:*
(1) z is (one of) the best modular approximations of f.
(2) It holds $U(f) = -L(f)$, and there exist probability distributions P^+ and P^- such that:
(i) *the support of P^+ (resp., P^-) is a subset of $\mathcal{U}(f)$ (resp., $\mathcal{L}(f)$);*
(ii) *for each $i \in [n]$, we have that if S^+ is sampled from P^+, and S^- is sampled from P^-, then $\Pr[i \in S^+] = \Pr[i \in S^-]$.*

Proof: Observe that wlog, $U(f) > L(f)$, since otherwise $U(f) = L(f)$ and f is a constant function.

The best modular approximation of f is given by the following LP (with variables $w_0, w_1, \ldots, w_n, t$):

$$\begin{cases} & \min t \\ \forall S \subseteq [n]: & w_0 + \sum_{i\in S} w_i + t \geq f(S), \\ \forall S \subseteq [n]: & -w_0 - \sum_{i\in S} w_i + t \geq -f(S). \end{cases} \tag{13}$$

The dual of the above program is equal to:

$$\begin{cases} & \max \sum_{S\subseteq[n]} \left(f(S) \cdot \left(y_S^+ - y_S^-\right)\right) \\ \forall i \in [n]: & \sum_{S\ni i} \left(y_S^+ - y_S^-\right) = 0 \\ & \sum_{S\subseteq[n]} \left(y_S^+ - y_S^-\right) = 0 \\ & \sum_{S\subseteq[n]} \left(y_S^+ + y_S^-\right) = 1, \\ & y_S^+, y_S^- \geq 0. \end{cases}$$

Note that z is one of the best modular approximations f if and only if the optimal value of the primal (and dual) program is equal to $M(f)$, i.e., if and only if the following system is feasible:

$$\begin{cases} & \sum_{S\subseteq[n]} \left(f(S) \cdot \left(y_S^+ - y_S^-\right)\right) = M(f) \\ \forall i \in [n]: & \sum_{S\ni i} \left(y_S^+ - y_S^-\right) = 0 \\ & \sum_{S\subseteq[n]} \left(y_S^+ - y_S^-\right) = 0 \\ & \sum_{S\subseteq[n]} \left(y_S^+ + y_S^-\right) = 1, \\ & y_S^+, y_S^- \geq 0. \end{cases} \tag{14}$$

We first show (2) $\implies$ (1). If there exists probability distributions P^+ and P^-, then note that by setting $y^+(S) = \frac{1}{2}P^+(S)$ and $y^-(S) = \frac{1}{2}P^-(S)$, we satisfy the last three conditions on y^+ and y^- in (14). Furthermore, since y^+ and y^- have disjoint support (in fact they have support only on $\mathcal{U}(f)$ and $\mathcal{L}(f)$ respectively) and since $U(f) = -L(f)$, we also satisfy $\sum_{S\subseteq[n]} \left(f(S) \cdot \left(y_S^+ - y_S^-\right)\right) = \frac{M(f)}{2} \sum_{S\subseteq[n]} \left(P^+(S) + P^-(S)\right) = M(f)$.

We next show (1) $\implies$ (2). Observe that (1) implies that system (14) is feasible. We will show that the probability distributions P^+ and P^- can be constructed from the solutions y^+ and y^- of this system. Note that, since (i) $\sum_S \left(f(S) \cdot \left(y_S^+ - y_S^-\right)\right) = M(f)$, (ii) $|f(S)| \leq M(f)$ for each S, and (iii) $\sum_S \left(y_S^+ + y_S^-\right) = 1$, it must be that $y_S^+ > 0$ only if $f(S) = U(f) = M(f)$, and $y_S^- > 0$ only if $f(S) = L(f) = -M(f)$. Furthermore, by (ii) and (iii), $\sum_S y^+(S) = \sum_S y^-(S) = \frac{1}{2}$.

Define $P^+(S) = 2y^+(S)$ and $P^-(S) = 2y^-(S)$. Then, $P^+(S) \geq 0$ for all S. Moreover $P^+(S)$ and $P^-(S)$ are both probability distributions, since the values of each of them sum up to 1. The support

of P^+ (resp., P^-) are sets S where $f(S) = U(f)$ (resp., $f(S) = -L(f)$). Hence, the support of P^+ is a subset of $\mathcal{U}(f)$ and the support of P^- is a subset of $\mathcal{L}(f)$. Finally, the marginal conditions follow directly from the second set of equalities in (14). ■

B. Outline of the method

We will use the probabilistic method to show that no approximately modular function that is best approximated by the all-zeros function can have a maximum value larger than $O(\log n)$. At a very high level, we will show that it is possible to find random combinations (unions and intersections) of sets that will end up producing the same outputs, regardless of whether we start from sets in $\mathcal{U}(f)$ or from sets in $\mathcal{L}(f)$. These combinations can be organized in layers, or levels, and each layer will only "add" a unit error to the function. Since the number of layers will be bounded by $O(\log n)$, we will be able to conclude that the total distance to modularity is $O(\log n)$.

We first introduce a basic structure that will be a building block for the final proof. A *striped* tree of height h is a full binary tree with the following properties: every node in the tree has an associated set and each non-leaf level $i \in [h-1]$ is assigned a label $s_i \in \{\cap, \cup\}$. The leaves will be initialized with sets of elements. The set associated with an internal node labeled $\cup$ (resp., $\cap$) will be the union (resp., intersection) of the two sets associated with its two children.

Let P be a probability distribution over subsets of $[n]$. *Seeding* T with P means assigning sets sampled iid from P to the leaves of T. Given a distribution P and a striped tree T, we use $T(P)$ to denote the random set in the root of T seeded by P.

We will show that if one samples a striped tree T of height $h = \Theta(\log n)$ uniformly at random, makes two copies of T, and seeds one with P^+, and the other with P^-, then (since both distributions give the same marginals $p_1, \ldots, p_n$ to the elements) with probability $1 - o(1)$, we will have $T(P^-) = T(P^+)$; this is the first part of the proof. The second part will embed all the 2^h striped trees of height h in a network of depth h. We will make two copies of that network. We will select the sets in the leaves of one of them using the distribution P^+ iid, and the leaves of the other using the distribution P^-, again, iid. We will then use the two networks to prove that the maximum value of any approximately modular function f that is best approximated by the all-zeros function z, can be upper bounded by $\Theta(h) = \Theta(\log n)$. This will allow us to conclude that each approximately modular function on a ground set of n elements is at distance $O(\log n)$ from modularity.

C. Striped trees

For a given striped tree T and the probability distribution P that seeds it, we are interested in the random set $T(P)$. Rather than characterize $T(P)$ for a specific tree T, though, we characterize the set $T(P)$ when the stripes of the tree are picked uniformly at random. In order to do so, we analyze the function $F_h(\cdot)$, where h is the height of T. Intuitively, if the stripes are chosen uniformly at random, then $F_h(p)$ is going to be the marginal of an element in a leaf if its marginal in $T(P)$ is p. We present the formal definition below.

Consider a striped tree and let p be the marginal of an element in an internal node and let q be the marginal of the same element at its two children. If the node is labeled "$\cup$", then $p = 2q - q^2$, and if the node is labeled with "$\cap$", then $p = q^2$. Hence, given p to be the marginal at a node, the marginal at the children is given by either $1 - \sqrt{1-p}$ for a "$\cup$" node or by $\sqrt{p}$ for a "$\cap$" node.

We let $f_1(x) = \sqrt{x}$ and $f_2(x) = 1 - \sqrt{1-x}$ be defined over $\mathcal{X} = (0, 1)$. We also let $f(x)$ be a random function defined over $\mathcal{X}$ as follows:

$$f(x) = \begin{cases} f_1(x) & \text{with probability } 1/2, \\ f_2(x) & \text{with probability } 1/2. \end{cases}$$

Moreover, we let $F_0(x) = x$ and, for each $h \geq 0$, let $F_{h+1}(x) = F_h(f(x))$, where each f is chosen iid. The random function $F_h(x)$ is the *iterated function system defined by* f.

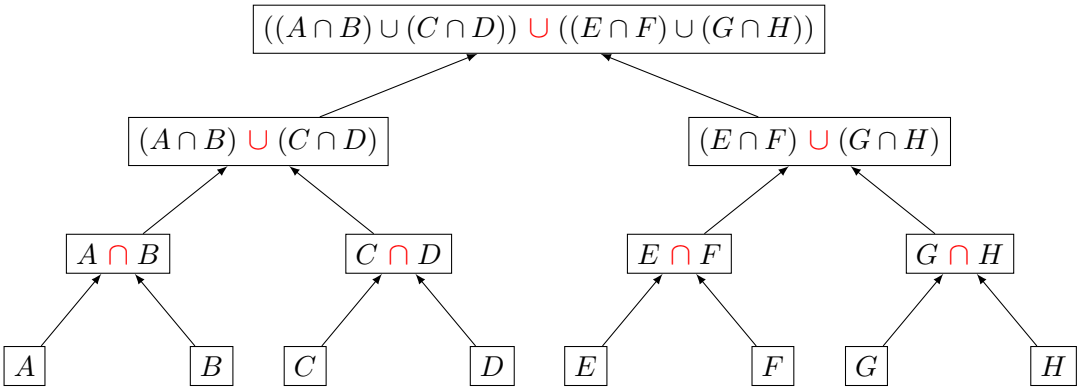

Figure 1: A striped tree of height $h = 3$, with stripes $\cup(\cup(\cap))$.

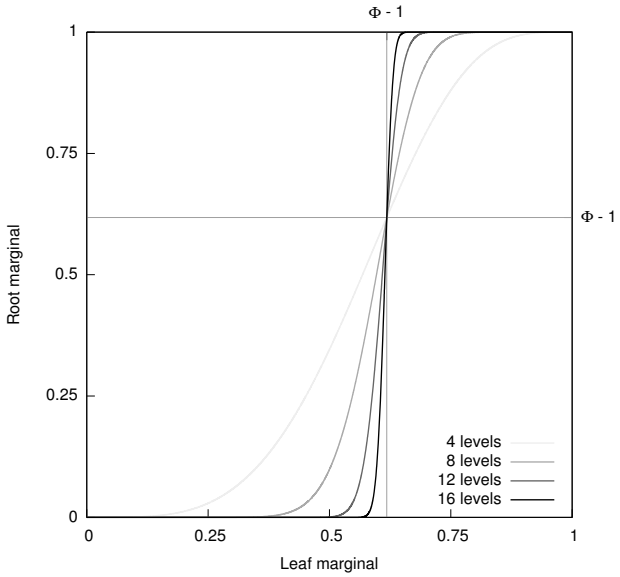

Figure 2: The threshold functions induced by a striped tree that alternates, level by level, $\cup$ and $\cap$ (with $\cup$ being the operator at the root node) for $4, 8, 12, 16$ levels — the plotted functions are the inverses of the random functions F_h described in the text (with the stripes described in the previous sentence). Alternating $\cup$'s and $\cap$'s in this fashion positions the threshold one unit to the left of the golden ratio, i.e., at $\phi - 1 = \frac{\sqrt{5}-1}{2}$. Indeed, if we let $p_\cup(x) = 2x - x^2$ and $p_\cap(x) = x^2$, and if we let $p(x) = p_\cup(p_\cap(x)) = 2x^2 - x^4$, we have that $p(x)$ is increasing in $[0, 1]$, that $p(0) = 0, p(1) = 1$ and that $\phi - 1$ is the (unique) fixed point of $p(x)$ in $(0, 1)$. The plot gives a visual indication that $p(p(\cdots p(x) \cdots)) = p_\cup(p_\cap(p_\cup(p_\cap(\cdots p_\cup(p_\cap(x)) \cdots))))$ converges quickly to a threshold function centered on its fixed point.

In other words, if we pick a striped tree of height h uniformly at random, then the system $F_h(p)$ gives us the marginal in the leaves of an element that has marginal p at the root. Our aim is to show that for a large enough h, for a random tree, $F_h(\cdot)$ is "close" to a uniformly chosen threshold in $[0, 1]$, and with high probability, each element i, depending on its marginal p_i, either has probability $o\left(n^{-1}\right)$ or $1 - o\left(n^{-1}\right)$ of being in the final set $T(P)$. Note that since the intended seeding distributions P^+ and P^- have the same marginals over elements, a simple union bound argument will ensure that $T(P^+) = T(P^-)$ w.h.p.

To formalize these arguments, we first need to show that $F_h(\cdot)$ is close to its stationary distribution (which, we will show, is the uniform distribution in $(0, 1)$). We will employ machinery from random iterated systems [33] for this purpose. We say that a process $F_h(x)$ is *attractive* if $F_\infty(x) = \lim_{h\to\infty} F_h(x)$ exists and if for each $x \in (0, 1)$, we have that $F_\infty(x)$ equals the same distribution F_∞, i.e., if the limit

distribution is independent of the starting point in $(0,1)$.

Observation 11. *The uniform distribution on $(0,1)$ is invariant for f.*

Proof: Take any set $[a,b] \subseteq (0,1)$. Let X be a random variable uniform in $(0,1)$. Then,

$$\Pr\left[f(X) \in [a,b]\right] = \frac{1}{2}\int_{a^2}^{b^2} 1\,dx + \frac{1}{2}\int_{2a-a^2}^{2b-b^2} 1\,dx = b - a = \Pr\left[X \in [a,b]\right].$$

■

The following corollary follows from the chain of theorems given in Section IV-E. In the following corollary F_h and F_∞ are related: the outermost h functions of F_h, and the outermost h functions of F_∞ are identical. The value r that we use in the the following Corollary 12 is equal to $\frac{2+\sqrt{2}}{4} < 0.86$; this value emerges from the analysis of our iterated function system (Lemma 16).

Corollary 12. *Let $0 < \delta < \frac{1}{2}$ be given. We have that*

$$E\left[\left|F_h\left(\frac{1}{2}\right) - F_\infty\left(\frac{1}{2}\right)\right|\right] \;\leq\; O\left(r^h\right) \quad \textit{and} \quad E\left[\left|F_h\left(1-\delta\right) - F_h\left(\delta\right)\right|\right] \;\leq\; O\left(\delta^{-1}r^h\right);$$

hence F_h is attractive. Moreover, the function $F_h(x)$ is increasing with probability 1 and the stationary distribution $F_\infty(\cdot)$ is unique.

Proof: By applying Theorem 17 and Theorem 18, and the bound we obtained for C_x in Lemma 19 for our iterated function system, we obtain:

$$E\left[\left|F_h\left(\frac{1}{2}\right) - F_\infty\left(\frac{1}{2}\right)\right|\right] \leq \frac{C_{1/2}}{1-r}r^h \leq \frac{32}{2-\sqrt{2}}r^h.$$

Moreover,

$$E\left[\left|F_h(1-\delta) - F_h(\delta)\right|\right| \leq \left|1-2\delta\right| \cdot \Phi\left(1-\delta,\delta\right) \cdot r^h = O\left(\delta^{-1}r^h\right).$$

The claim about F_h being increasing follows from the fact that both f_1 and f_2 are monotonically increasing, and that $F_h(x)$ is obtained by repeated applications of copies of f_1 and f_2 to themselves. The claim about the uniqueness of the stationary distribution follows from the attractivity of F_i and a result of Letac [24] (see Section IV-E.) ■

We then have our main theorem of the section. Again, $r = (2+\sqrt{2})/4$.

Theorem 13. *Let T be a uniform at random striped tree of height $h = \left\lceil 2\log_{1/r} n + 4\log_{1/r}\log n \right\rceil$. Then, the probability that T seeded with P^+ will produce the same set as T seeded with P^- is at least $1 - O\left(\log^{-1} n\right)$.*

Proof: Let $p_1 \leq \cdots \leq p_n$ be the marginals of the n elements under (each of) the two distributions P^+ and P^-. Let $t(n) = n\log n$. For each p_i, let $B_i = \left[p_i - \frac{1}{t(n)}, p_i + \frac{1}{t(n)}\right]$. Let $B = \bigcup_{i=1}^n B_i$.

Let X be a random variable uniform in $[0,1]$. Then, $P\left[X \in B\right] \leq O\left(\frac{n}{t(n)}\right) = O\left(\log^{-1} n\right)$. Observe that if $X \notin B$, then X is at distance at least $\frac{1}{t(n)}$ from each of the p_i's. Therefore, if $-\frac{1}{2t(n)} \leq \xi \leq \frac{1}{2t(n)}$, $X \notin B$ implies that $X + \xi$ is at distance at least $\frac{1}{2t(n)}$ from each p_i.

We now consider the random variable $F_h\left(\frac{1}{2}\right)$. Using Corollary 12,

$$E\left[\left|F_h(1/2) - F_\infty(1/2)\right|\right] \leq O(r^h) = O\left(r^{2\log_{1/r} n + 4\log_{1/r}\log n}\right) = O\left(n^{-2}\log^{-4} n\right).$$

By applying Markov's inequality, we then get that:

$$\Pr\left[\left|F_h\left(\frac{1}{2}\right) - F_\infty\left(\frac{1}{2}\right)\right| \geq \Omega\left(n^{-1}\log^{-3} n\right)\right] \leq O\left(n^{-1}\cdot\log^{-1} n\right).$$

By Observation 11 and by Corollary 12, we have that $F_\infty\left(\frac{1}{2}\right)$ is uniformly distributed in $(0,1)$. Hence:

$$\Pr\left[\exists i \left| F_h\left(\frac{1}{2}\right) - p_i \right| \leq \frac{1}{2t(n)}\right] < O\left(\log^{-1} n\right).$$

We now set $\delta = \frac{1}{n \log n}$, and again apply Corollary 12 to get

$$E\left[\left| F_h\left(1 - \frac{1}{n\log n}\right) - F_h\left(\frac{1}{n \log n}\right)\right|\right] \leq O\left(n \log n \cdot r^h\right) = O\left(n^{-1}\log^{-3} n\right).$$

Therefore,

$$\Pr\left[\left| F_h\left(1 - \frac{1}{n\log n}\right) - F_h\left(\frac{1}{n \log n}\right)\right| > \Omega\left(n^{-1}\log^{-2} n\right)\right] \leq O\left(\log^{-1} n\right).$$

We define the following two events: $\xi_1 =$ "$\left| F_h\left(1 - \frac{1}{n\log n}\right) - F_h\left(\frac{1}{n \log n}\right)\right| > \Omega\left(n^{-1}\log^{-2} n\right)$" and $\xi_2 =$ "$\exists i \left| F_h\left(\frac{1}{2}\right) - p_i\right| \leq \frac{1}{2t(n)}$". By a union bound, $\Pr[\neg\xi_1 \text{ and } \neg\xi_2] \geq 1 - O\left(\log^{-1} n\right)$. We condition on the tree having been created, on $\neg\xi_1$ and $\neg\xi_2$, and start the other random process: use either P^+ or P^- to place sets iid in the leaves of the tree.

Consider an arbitrary element i. By $\neg\xi_1$ and $\neg\xi_2$, we have that the probability that element i will be in the root of the tree is either at most δ or at least $1-\delta$. Let $S^\star$ be the set of elements having probability at least $1-\delta$ of being in the root set. We compute the probability that the root set S is different from $S^\star$:

$$\begin{aligned}\Pr\left[S \neq S^\star\right] &= \Pr\left[\exists i \in (S^\star \setminus S) \text{ or } \exists j \in (S \setminus S^\star)\right] \\ &\leq \sum_{i \in S^\star} \Pr[i \notin S] + \sum_{j \in [n]\setminus S^\star} \Pr[j \in S] \\ &\leq |S^\star| \cdot \delta + (n - |S^\star|) \cdot \delta = n\delta = \frac{1}{\log n}.\end{aligned}$$

A union bound concludes the proof. ∎

D. Striped networks

In order to finish up the proof of bounding $M(f)$, we need yet another step, since a single striped tree is not enough to create an "efficient" system of inequalities for comparing $U(f)$ with $L(f)$. This is because each rule application creates two outputs whereas a striped tree considers only a single output (either intersection or union) per pair of input sets. This loss is too large for an upper bound proof to go through. To avoid this issue we consider *striped networks*; striped networks' rules have two outputs (intersection *and* union) per input pair. For each striped tree of a given height, there exists a subset of the nodes of a striped network of the same height that acts like that striped tree. This will allow us to use the probabilistic guarantees given by Theorem 13 in the analysis of striped networks, in order to finally prove our upper bound on $M(f)$.

A striped network is built inductively. Each striped network N_h of height h will have 2^h outputs and 2^h input nodes (which will be the ones seeded with random sets from P^+ or P^-). Each node will have one label, and each output node will have a unique label. The striped network N_0, having height $h = 0$, contains a single input (which is also an output), having the empty label. To produce the striped network N_{h+1}, we first create two copies of the striped network N_h. We pair up the outputs of the two copies having the same labels; for each of the 2^h pairs, each with its label ℓ, we apply the approximate modularity rule on the two elements of the pair and label the outputs of this rule application with $\cap(\ell)$ and $\cup(\ell)$ (resp., for the intersection output and the union output of the rule application). Observe that the total number of rule applications in N_h is $h2^{h-1}$. See Figure 3 for an illustration of N_3.

The value of r in the following theorem is $r = (2+\sqrt{2})/4$.

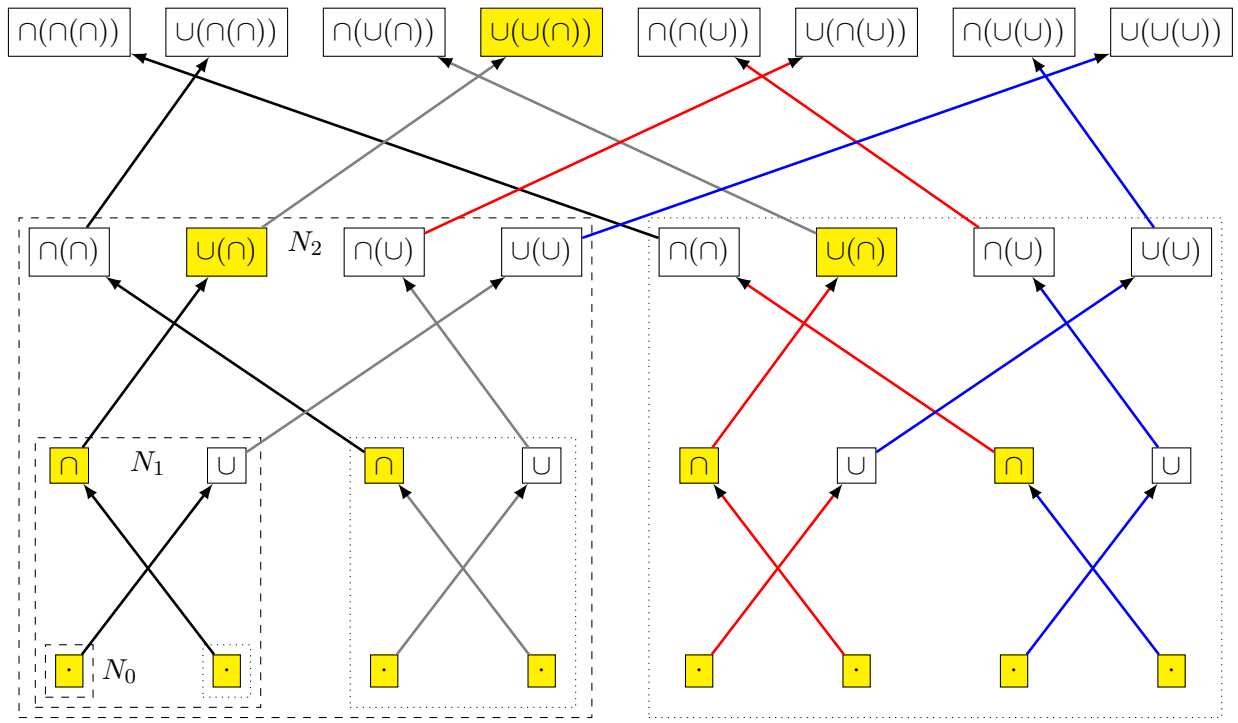

Figure 3: The striped network N_3 of height $h = 3$. Some of the copies of N_0, N_1 and N_2 that make up N_3 are enclosed in dashed rectangles. The highlighted nodes make up the striped tree in Figure 1. The bottom level's nodes will contain the random sets. In each level, each edge color corresponds to a distinct rule application. The number of rule applications shown in this figure is equal to $2^{h-1} \cdot h = 4 \cdot 3 = 12$.

Theorem 14. *For each $f \in \widetilde{\mathcal{M}}_n$ there exists a $g \in \mathcal{M}_n$ such that*

$$|f - g| \leq \log_{1/r} n + O(\log\log n) < 4.3774 \cdot \lg n + O(\log\log n).$$

Proof: Recall that if $g(S) = w_0 + \sum_{i \in S} w_i$ is an optimal modular approximation to f, then the all-zeros function z is an optimal modular approximation to $F(S) = f(S) - g(S)$, and the distance to modularity of f equals the distance to modularity of F. Since F has z as its best modular approximation, by Lemma 10 its maximum value M is the opposite of its minimum value $-M$. Therefore M is the distance to modularity of both F and f. Moreover, by Lemma 10, there exist a probability distribution P^+ over the sets where F achieves value M, and a probability distribution P^- over the sets where F achieves value $-M$, such that for each $i \in [n]$, if $X^+ \sim P^+$ and $X^- \sim P^-$ are independent, then $\Pr[i \in X^+] = \Pr[i \in X^-]$.

We now create two striped networks N^+ and N^- of height $h = \left\lceil 2\log_{1/r} n + 4\log_{1/r}\log n \right\rceil$ each. We seed the leaves of N^+ with iid samples of P^+ and we seed the leaves of N^- with an iid samples of P^-. Suppose that $X^+_{0,i}, \ldots, X^+_{2^h-1,i}$ are the sets that end up in the i^{th} level of N^+.

Then, for each $0 \leq i \leq h-1$, there exist a matching $\mathcal{M}_i$ between the indices in $\{0, \ldots, 2^h - 1\}$ and a matching $\mathcal{M}'_i$ between the indices in $\{0, \ldots, 2^h - 1\}$, such that for each $(j, j') \in \mathcal{M}_i$ there exists a unique $(j'', j''') \in \mathcal{M}'_i$ such that $X^+_{j'',i+1} = X^+_{j,i} \cap X^+_{j',i}$ and $X^+_{j''',i+1} = X^+_{j,i} \cup X^+_{j',i}$.

Let $\Sigma^+_i = \sum_{j=0}^{2^h-1} F\left(X^+_{j,i}\right)$. By the upper-approximately modular inequality (i.e., $F(S) + F(T) \leq F(S \cap T) + F(S \cup T) + 1$), we have for each $i = 0, \ldots, h-1$ that

$$\Sigma^+_i \leq \Sigma^+_{i+1} + 2^{h-1}.$$

We unwind the inequalities to get:

$$\Sigma_0^+ \leq h2^{h-1} + \Sigma_h^+ = h2^{h-1} + \sum_{j=0}^{2^h-1} F\left(X_{j,h}^+\right).$$

By the definition of P^+, we have for each $j = 0, \ldots, 2^h - 1$ that $F\left(X_{j,0}^+\right) = M$. Therefore:

$$M2^h \leq h2^{h-1} + \sum_{j=0}^{2^h-1} F\left(X_{j,h}^+\right). \tag{15}$$

We now mirror our analysis of the N^+ network to the case of the N^- network. As before, we have that for each i, there exist two matchings $\mathcal{M}_i, \mathcal{M}_i'$ such that for each $\{j, j'\} \in \mathcal{M}_i$ there exists a unique $\{j'', j'''\} \in \mathcal{M}_i'$ such that $X_{j'',i+1}^- = X_{j,i}^- \cap X_{j',i}^-$ and $X_{j''',i+1}^- = X_{j,i}^- \cup X_{j',i}^-$.

We define $\Sigma_i^- = \sum_{j=0}^{2^h-1} F\left(X_{j,i}^-\right)$. This time, we use the lower-approximately modular inequality (i.e., $F(S) + F(T) \geq F(S \cap T) + F(S \cup T) - 1$) to get that for each $i = 0, \ldots, h-1$:

$$\Sigma_i^- \geq -2^{h-1} + \Sigma_{i+1}^-.$$

We use this to obtain

$$\Sigma_0^- \geq -h2^{h-1} + \sum_{j=0}^{2^h-1} F\left(X_{j,h}^-\right).$$

Again by definition, we get that all the $F\left(X_{j,0}^-\right)$ equal $-M$. Therefore:

$$-M2^h \geq -h2^{h-1} + \sum_{j=0}^{2^h-1} F\left(X_{j,h}^-\right). \tag{16}$$

We now subtract inequality (16) from inequality (15), obtaining:

$$M2^{h+1} \leq h2^h + \sum_{j=0}^{2^h-1} \left(F\left(X_{j,h}^+\right) - F\left(X_{j,h}^-\right)\right) \implies$$

$$M \leq \frac{h}{2} + 2^{-h-1} \sum_{j=0}^{2^h-1} \left(F\left(X_{j,h}^+\right) - F\left(X_{j,h}^-\right)\right).$$

Observe that the striped tree that made up $X_{j,h}^+$ is equal to the striped tree that made up $X_{j,h}^-$, for each $j = 0, \ldots, 2^h - 1$. Therefore, for a uniform at random j, Theorem 13 guarantees that with probability at least $1 - O\left(\log^{-1} n\right)$, it will hold $X_{j,h}^+ = X_{j,h}^-$. If this event happens, then the jth difference $F\left(X_{j,h}^+\right) - F\left(X_{j,h}^-\right)$ will equal 0. Moreover, since $-M \leq F(X) \leq M$ for each $X \subseteq [n]$, no difference can be larger than $2M$. Hence, if $T = \left|\left\{j \mid 0 \leq j \leq 2^h - 1 \text{ and } X_{j,h}^+ \neq X_{j,h}^-\right\}\right|$, then

$$M \leq \frac{h}{2} + 2^{-h-1} \cdot T \cdot 2M.$$

Since the expected number $E[T]$ of j's for which we will have $X_{j,h}^+ \neq X_{j,h}^-$ is upper bounded by $2^h \cdot O\left(\log^{-1} n\right)$, there exists an assignment of sets to the two 0-levels of the two striped networks that guarantees that $T \leq E[T] \leq 2^h O\left(\log^{-1} n\right)$. (In fact, by Markov inequality, a random assignment will guarantee that $T \leq 2^h O\left(\log^{-1} n\right)$ with probability $\Omega(1)$). With such an assignment, we get:

$$M \leq \frac{h}{2} + O\left(M \log^{-1} n\right).$$

Since $h = \Theta(\log n)$, we finally obtain $M \leq \frac{h}{2} + O(1)$, completing the proof. ∎

E. Showing that F_h is attractive

We first start with an observation about f_1' and f_2'.

Observation 15. *The functions f_1 and f_2 are differentiable in $\mathcal{X}$ and their derivatives are:*

$$f_1'(x) = \frac{1}{2\sqrt{x}} \quad \text{and} \quad f_2'(x) = \frac{1}{2\sqrt{1-x}}.$$

Let $\phi(x) = \frac{1}{(1-x)\cdot x}$ be defined over $\mathcal{X}$. Also, let $\Phi(x,y) = \sup_{0\le t\le 1} \phi(tx+(1-t)y)$ be defined over $\mathcal{X}^2$. Since $\phi(x)$ is convex, we have that $\Phi(x,y) = \max(\phi(x),\phi(y))$.

The growth rate of f (with respect to the one-dimensional linear metric $\rho(x,y) = |x-y|$) is defined to be equal to:

$$r(x) = \frac{E_f\left[\phi(f(x))\cdot f'(x)\right]}{\phi(x)}.$$

We let $r = \sup_{x\in\mathcal{X}} r(x)$ be the supremum of the growth rate of f over its domain. The following lemma shows that with this choice of ϕ, the growth rate of f is bounded.

Lemma 16. *For each $x \in \mathcal{X}$, we have that*

$$r(x) = \frac{1}{2} + \frac{\sqrt{x}+\sqrt{1-x}}{4},$$

Moreover, $r(x) \le r(1/2) = (2+\sqrt{2})/4 = r$.

Proof: We have that:

$$\begin{aligned}
r(x) = \frac{E_f\left[\phi(f(x))\cdot f'(x)\right]}{\phi(x)} &= \frac{1}{2}\left(\frac{\frac{1}{2\sqrt{x}}}{(1-\sqrt{x})\sqrt{x}} + \frac{\frac{1}{2\sqrt{1-x}}}{\sqrt{1-x}\cdot(1-\sqrt{1-x})}\right)(1-x)x \\
&= \frac{1}{4}\left(\frac{1}{x}\cdot\frac{1}{1-\sqrt{x}} + \frac{1}{1-x}\cdot\frac{1}{1-\sqrt{1-x}}\right)(1-x)x \\
&= \frac{1}{4}\left(\frac{1-x}{1-\sqrt{x}} + \frac{x}{1-\sqrt{1-x}}\right) \\
&= \frac{1}{4}\left(\frac{(1-\sqrt{x})(1+\sqrt{x})}{1-\sqrt{x}} + \frac{x\left(1+\sqrt{1-x}\right)}{\left(1-\sqrt{1-x}\right)\left(1+\sqrt{1-x}\right)}\right) \\
&= \frac{1}{4}\left(1+\sqrt{x}+\frac{x\left(1+\sqrt{1-x}\right)}{x}\right) = \frac{1}{2}+\frac{\sqrt{x}+\sqrt{1-x}}{4}.
\end{aligned}$$

The first derivative of $\sqrt{x}+\sqrt{1-x}$ is equal to $\frac{1}{2}\left(\frac{1}{\sqrt{x}} - \frac{1}{\sqrt{1-x}}\right)$; the latter is positive for $x < \frac{1}{2}$, and negative for $x > \frac{1}{2}$. Therefore the maximum of $\sqrt{x}+\sqrt{1-x}$ is achieved at $x = \frac{1}{2}$, and its value there is $\sqrt{2}$. The claim follows. ∎

We then use a result from [33] that helps us conclude that the iterated function system defined by repeated independent applications of f on the interval $\mathcal{X} = (0,1)$ is locally contractive.

Theorem 17 (Theorem 2 in [33])**.** *If $\phi : \mathcal{X} \to [1,\infty)$ is a continuous function whose growth rate r with respect to a random function f is strictly smaller than 1, then the iterated function system defined by f on the interval $\mathcal{X}$ is locally contractive with drift function ϕ and rate r.*

The final aim is to show that $F_h(\cdot)$ converges exponentially fast to a stationary distribution. We will do this by applying the following theorem of [33].

Theorem 18 (Theorem 1 in [33]). *Let $C_x = E_f\left[|f(x) - x| \cdot \Phi(x, f(x))\right]$. If the iterated function system F_h defined by f is locally contractive with drift function ϕ and rate $r < 1$, and if C_x is finite for all $x \in \mathcal{X}$:*

$$E\left[|F_h(x) - F_\infty(x)|\right] \leq \frac{C_x}{1-r} r^h.$$

Moreover, for all $x, y \in \mathcal{X}$:

$$E\left[|F_h(x) - F_h(y)|\right] \leq |x - y| \cdot \Phi(x, y) r^h.$$

Therefore, F_h is attractive.

We bound C_x in our case.

Lemma 19. *We have that*

$$C_x \leq \frac{2}{(1-x)x}.$$

Proof: Observe that, for $x \in \mathcal{X}$, we have $f(x) \in \mathcal{X}$. Therefore $|f(x) - x| \leq 1$. The function $\phi(x) : \mathcal{X} \to \mathbb{R}^{\geq 0}$ is decreasing for $x < \frac{1}{2}$, and it is increasing for $x > \frac{1}{2}$. Moreover, $\phi(x) = \phi(1-x)$. Therefore, since $\sqrt{x} > x$, for $x \in \mathcal{X}$, so long as $x \geq \frac{1}{2}$, we have $\phi(x) < \phi(\sqrt{x}) = \phi(f_1(x))$. Analogously, if $x < \frac{1}{2}$, we have that $\phi(x) = \phi(1-x) < \phi(\sqrt{1-x}) = \phi(1 - \sqrt{1-x}) = \phi(f_2(x))$.

Therefore, since $f \in \{f_1, f_2\}$, we have that

$$\Phi(x, f(x)) \leq \max(\phi(x), \phi(f_1(x)), \phi(f_2(x))) = \max(\phi(f_1(x)), \phi(f_2(x))) = \max\left(\phi\left(\sqrt{x}\right), \phi\left(\sqrt{1-x}\right)\right).$$

We now bound $\phi(\sqrt{y})$, for $y \in \mathcal{X}$:

$$\phi(\sqrt{y}) = \frac{1}{\sqrt{y} - y} = \frac{1}{\sqrt{y} - y} \frac{\sqrt{y} + y}{\sqrt{y} + y} = \frac{\sqrt{y} + y}{(1-y)y} \leq \frac{2}{(1-y)y}.$$

Clearly the same upper bounds holds for $\phi(\sqrt{1-y})$. Thus, we have that:

$$\Phi(x, f(x)) \leq \frac{2}{(1-x)x}.$$

The claim follows. ∎

V. Lower bounds

Although we showed that every approximately modular function f admits a modular approximation that is $O(\log n)$-close, the only algorithm that we have discussed for producing such a modular approximation solves an exponential sized-LP (see (13)) and requires knowing all the 2^n values of f. The algorithm in Section III, however, makes $O(n^2 \log n)$ queries to f and returns a modular approximation that is $O(\sqrt{n})$-close. We now show that if only a polynomial number of queries to f is allowed, then the algorithm of Section III is near-optimal in terms of the distance between the function in its input, and the one it outputs.

The lower bound will make use of the following distribution over approximately modular set functions.

Definition 20. *Given $n, q \in \mathbb{Z}^+$, let $x_1, \ldots, x_n$ be chosen independently and uniformly at random in $\{-1, 1\}$. The random set function $r_{n,q}$ is defined as $r_{n,q}(S) = \frac{1}{4} \cdot b\left(\sum_{i \in S} \frac{x_i}{\sqrt{2n \ln(4q)}}\right)$, $\forall S \subseteq [n]$, where $b(y) = \lfloor y \rfloor$ if $y \geq 0$, and $b(y) = -\lfloor -y \rfloor$ otherwise.*

Let the function $m_{n,q}(S)$ be defined as $m_{n,q}(S) = \frac{1}{4} \sum_{i \in S} \frac{x_i}{\sqrt{2n \ln(4q)}}$. Observe that $r_{n,q}(S) - m_{n,q}(S) \in \left(-\frac{1}{4}, \frac{1}{4}\right)$. Since $m_{n,q}$ is modular, we have that $r_{n,q}$ is 1-approximately modular. Moreover, the maximum

absolute value of $r_{n,q}$ is at least $\frac{1}{4}\left(\frac{n/2}{\sqrt{2n\ln(4q)}}-1\right) = \Omega\left(\sqrt{\frac{n}{\log q}}\right)$, since for any choice of the x_i's, there are either at least $n/2$ positive x_i's, or more than $n/2$ negative x_i's.

We will now show that using only q queries we cannot distinguish between a random $r_{n,q}$ and the all-zeros function z with probability $1-o(1)$. Since the maximum absolute value of z is 0, this implies the additive error is at least $\Omega\left(\sqrt{\frac{n}{\log q}}\right)$ with at least constant probability.

Lemma 21. *Let f be a random 1-approximately modular function defined as follows: $f \sim r_{n,q}$ with probability 1/2 and is z otherwise. There is a constant $c<1$ such that, if a randomized adaptive algorithm queries f at most q times, then the probability that the algorithm is able to guess correctly if $f=z$ is at most c.*

Proof: Let $S \subseteq [n]$. By the Chernoff bound, we have that

$$\Pr[r_{n,q}(S) \neq 0] = \Pr\left[\left|\sum_{i\in S} x_i\right| \geq \sqrt{2n\ln(4q)}\right] \leq 2e^{-\frac{2n\ln(4q)}{2n}} = \frac{1}{2q}.$$

By definition, $z(S)=0$ for each set S. Let $S_1,\ldots,S_q$ be the random variables that represent the (possibly random and adaptive) probes performed by the algorithm under the assumption that the function oracle always returns 0 as an answer. By the union bound, we have $\Pr\left[f(S_1)=\cdots=f(S_q)=0 \mid f=z\right]=1$, and $\Pr\left[f(S_1)=\cdots=f(S_q)=0 \mid f\sim r_{n,q}\right] \geq 1-q\cdot\frac{1}{2q}=\frac{1}{2}$. Thus, regardless of whether $f=z$ or $f\sim r_{n,q}$, with probability at least 1/2, the algorithm will get an answer of 0 to each of its q queries. Therefore, from the algorithm's perspective, f is in the support of $r_{n,q}$ with probability $\Omega(1)$, and $f=z$ with probability $\Omega(1)$. ∎

Theorem 22. *There exists a distribution over 1-approximately modular functions f over $[n]$ such that no algorithm performing at most q queries to f can distinguish, with probability $1-o(1)$, whether the maximum absolute value of f is 0, or whether it is at least $\Omega\left(\sqrt{\frac{n}{\log q}}\right)$.*

Proof: Let f be defined as in Lemma 21 with $q=n^{O(1)}$. By Lemma 21, no algorithm is able to tell with probability $1-o(1)$ whether $f=z$ or $f\sim r_{n,q}$. In the former case, the maximum absolute value of f is 0 whereas in the latter case, it is $\Omega\left(\sqrt{\frac{n}{\log n}}\right)$. ∎

The above theorem directly entails that no finite multiplicative approximation of f is possible with sub-exponentially many queries. It also entails the following additive inapproximability:

Corollary 23. *There exists a distribution over 1-approximately modular functions f over $[n]$ such that no algorithm performing at most $n^{O(1)}$ queries to f can return, with probability $1-o(1)$, a modular function that is $o\left(\sqrt{\frac{n}{\log n}}\right)$-close to f.*

ACKNOWLEDGMENTS

We thank the reviewers for their many valuable comments. We also thank one of the reviewers for the proof of Theorem 9.

REFERENCES

[1] M.-F. Balcan and N. J. Harvey. Learning submodular functions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 793–802. ACM, 2011.

[2] M. Bellare, D. Coppersmith, J. Håstad, M. A. Kiwi, and M. Sudan. Linearity testing in characteristic two. *IEEE Transactions on Information Theory*, 42(6):1781–1795, 1996.

[3] M. Ben-Or, D. Coppersmith, M. Luby, and R. Rubinfeld. Non-abelian homomorphism testing, and distributions close to their self-convolutions. In *RANDOM*, pages 273–285, 2004.

[4] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. *JCSS*, 47(3):549–595, 1993.

[5] D. Chakrabarty and C. Seshadhri. A $o(n)$ monotonicity tester for Boolean functions over the hypercube. In *STOC*, pages 411–418, 2013.

[6] D. Chakrabarty and C. Seshadhri. Optimal bounds for monotonicity and Lipschitz testing over hypercubes and hypergrids. In *STOC*, pages 419–428, 2013.

[7] P. W. Cholewa. Remarks on the stability of functional equations. *Aequationes Math.*, 27:76–86, 1984.

[8] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *ICML*, pages 1057–1064, 2011.

[9] F. Ergün, R. Kumar, and R. Rubinfeld. Checking approximate computations of polynomials and functional equations. *SICOMP*, 31(2):550–576, 2001.

[10] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

[11] M. X. Goemans, N. J. Harvey, S. Iwata, and V. Mirrokni. Approximating submodular functions everywhere. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 535–544. Society for Industrial and Applied Mathematics, 2009.

[12] O. Goldreich, S. Goldwasser, E. Lehman, D. Ron, and A. Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[13] D. H. Hyers. On the stability of the linear functional equation. *PNAS*, 27:222–224, 1941.

[14] D. H. Hyers, G. Isaac, and T. M. Rassias. *Stability of Functional Equations in Several Variables*. Birkhauser, 1998.

[15] D. H. Hyers and S. Ulam. Approximately convex functions. *Proc. Amer. Math. Soc.*, 3:821–828, 1952.

[16] M. Jha and S. Raskhodnikova. Testing and reconstruction of Lipschitz functions with applications to data privacy. *SICOMP*, 42(2):700–731, 2013.

[17] S. M. Jung. *Hyers–Ulam–Rassias Stability of Functional Equations in Nonlinear Analysis*. Springer, 2011.

[18] M. A. Kiwi, F. Magniez, and M. Santha. Approximate testing with error relative to input size. *JCSS*, 66(2):371–392, 2003.

[19] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *ICML*, pages 567–574, 2010.

[20] A. Krause, A. P. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008.

[21] M. J. Kusner, W. Chen, Q. Zhou, Z. E. Xu, K. Q. Weinberger, and Y. Chen. Feature-cost sensitive learning with submodular trees of classifiers. In *AAAI*, pages 1939–1945, 2014.

[22] M. Laczkovich. The local stability of convexity, affinity and of the Jensen equation. *Aequationes Mathematicae*, 58(1-2):135–142, 1999.

[23] J. Lee. *A First Course in Combinatorial Optimization*. Cambridge University Press, 2004.

[24] G. Letac. A contraction principle for certain Markov chains and its applications. *Contemp. Math.*, 50:263–273, 1986.

[25] H. Narayanan. *Submodular Functions and Electrical Networks*. North-Holland, 1997.

[26] M. Parnas, D. Ron, and R. Rubinfeld. On testing convexity and submodularity. *SICOMP*, 32(5):1158–1184, 2003.

[27] R. Rubinfeld. On the robustness of functional equations. *SICOMP*, 28(6):1972–1997, 1999.

[28] R. Rubinfeld. Linearity testing/testing Hadamard codes. In *Encyclopedia of Algorithms*. Springer, 2008.

[29] R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. In *STOC*, pages 147–156, 2005.

[30] R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.

[31] A. Schrijver. *Combinatorial Optimization*. Springer, 2003.

[32] L. S. Shapley. A Value for n-Person Games. In H. Kuhn and A. Tucker, editors, *Contributions to the Theory of Games*, volume 2 of *Annals of Mathematics Studies*. Princeton University Press, 1953.

[33] D. Steinsaltz. Locally contractive iterated function systems. *Annals of Probability*, 27(4):1952–1979, 1999.