Statistical Algorithms and a Lower Bound for Detecting Planted Cliques

Vitaly Feldman
Almaden Research Center
IBM
San Jose, CA 95120
vitaly@post.harvard.edu

Elena Grigorescu[*] [†]
Dept. of Computer Science
Purdue University
West Lafayette, IN 47907
elena-g@purdue.edu

Lev Reyzin[‡] [†]
Dept. of Mathematics
University of Illinois at Chicago
Chicago, IL 60607
lreyzin@math.uic.edu

Santosh S. Vempala[†]
School of Computer Science
Georgia Inst. of Technology
Atlanta, GA 30332
vempala@cc.gatech.edu

Ying Xiao[†]
School of Computer Science
Georgia Inst. of Technology
Atlanta, GA 30332
yxiao32@cc.gatech.edu

ABSTRACT

We introduce a framework for proving lower bounds on computational problems over distributions, based on a class of algorithms called *statistical algorithms*. For such algorithms, access to the input distribution is limited to obtaining an estimate of the expectation of any given function on a sample drawn randomly from the input distribution, rather than directly accessing samples. Most natural algorithms of interest in theory and in practice, e.g., moments-based methods, local search, standard iterative methods for convex optimization, MCMC and simulated annealing, are statistical algorithms or have statistical counterparts. Our framework is inspired by and generalize the statistical query model in learning theory [34].

Our main application is a nearly optimal lower bound on the complexity of *any* statistical algorithm for detecting planted bipartite clique distributions (or planted dense subgraph distributions) when the planted clique has size $O(n^{1/2-\delta})$ for any constant $\delta > 0$. Variants of these problems have been assumed to be hard to prove hardness for other problems and for cryptographic applications. Our lower bounds provide concrete evidence of hardness, thus supporting these assumptions.

[*]This material is based upon work supported by the National Science Foundation under Grant #1019343 to the Computing Research Association for the CIFellows Project.

[†]Research supported in part by NSF awards CCF-0915903 and CCF-0910584.

[‡]Research supported in part by a Simons Postdoctoral Fellowship.

STOC'13, June 1-4, 2013, Palo Alto, California, USA.

Categories and Subject Descriptors

F.2 [**Analysis of Algorithms and Problem Complexity**]: General

Keywords

Statistical algorithms; Lower bounds; Planted clique; Statistical query

1. INTRODUCTION

We study the complexity of problems where the input consists of independent samples from an unknown distribution. Such problems are at the heart of machine learning and statistics (and their numerous applications) and also occur in many other contexts such as compressed sensing and cryptography. Many methods exist to estimate the sample complexity of such problems (e.g. VC dimension [47]). Proving lower bounds on the computational complexity of these problems has been much more challenging. The traditional approach to this is via reductions and finding distributions that can generate instances of some problem conjectured to be intractable (e.g., assuming $\mathsf{NP} \neq \mathsf{RP}$).

Here we present a different approach, namely showing that a broad class of algorithms, which we refer to as *statistical algorithms*, have high complexity, unconditionally. Our definition encompasses most algorithmic approaches used in practice and in theory on a wide variety of problems, including Expectation Maximization (EM) [16], local search, MCMC optimization [45, 25], simulated annealing [36, 48], first and second order methods for linear/convex optimization, [17, 6], k-means, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Naïve Bayes, Neural Networks and many others (see [13] and [9] for proofs and many other examples). In fact, we are aware of only one algorithm that does not have a statistical counterpart: Gaussian elimination for solving linear equations over a field (e.g., mod 2).

Informally, statistical algorithms can access the input distribution only by asking for the value of any bounded real-valued function on a random sample from the distribution, or the average value of the function over a specified number

of independent random samples. As an example, suppose we are trying to solve $\min_{u \in U} \mathbf{E}_{x \sim D}[f(x,u)]$ by gradient descent. Then the gradient of the objective function is (by interchanging the derivative with the integral)

$$\nabla_u \mathop{\mathbf{E}}_x[f(x,u)] = \mathop{\mathbf{E}}_x[\nabla_u(f(x,u))]$$

and can be estimated from samples; thus the algorithm can proceed without ever examining samples directly. A more involved example is Linear Programming. One version is the feasibility problem: find a nonzero w s.t. $a \cdot w \geq 0$ for all a in some set A. We can formulate this as

$$\max_w \mathop{\mathbf{E}}_a[\mathsf{sign}(a \cdot w)]$$

and the distribution over a could be uniform over the set A if it is finite. This can be solved by a statistical algorithm [10, 17]. This is also the case for semidefinite programs and for conic optimization [6]. The key motivation for our definition of statistical algorithms is the empirical observation that almost all algorithms that work on explicit instances are already statistical in our sense or have natural statistical counterparts. Thus, proving lower bounds for statistical algorithms strongly indicates the need for new approaches even for explicit instances. We present the formal oracle-based definitions of statistical algorithms in Section 2.

The inspiration for our model is the *statistical query (SQ)* model in learning theory [34] defined as a restriction of Valiant's PAC learning model [46]. The primary goal of the restriction was to simplify the design of noise-tolerant learning algorithms. As was shown by Kearns and others in subsequent works, almost all classes of functions that can be learned efficiently can also be efficiently learned in the restricted SQ model. A notable and so far the only exception is the algorithm for learning parities, based on Gaussian elimination. As was already shown by Kearns [34], parities require exponential time to learn in the SQ model. Further, Blum et al. [11] proved that the number of SQs required for weak learning (that is, for obtaining a non-negligible advantage over the random guessing) of a class of functions C over a fixed distribution D is characterized by a combinatorial parameter of C and D, referred to as SQ-DIM(C,D), the statistical query dimension.

Our notion of statistical algorithms generalizes SQ learning algorithms to any computational problem over distributions. For any problem over distributions we define a parameter of the problem that lower bounds the complexity of solving the problem by any statistical algorithm in the same way as SQ-DIM lower bounds the complexity of learning in the SQ model. Our techniques for proving lower bounds on statistical algorithms are also based on methods developed for lower-bounding the complexity of SQ algorithms. However, as we will describe later, they depart from the known techniques in a number of significant ways that are necessary for our more general definition and our applications.

We apply our techniques to the problem of detecting planted bipartite cliques/dense subgraphs.

Detecting Planted Cliques. In the planted clique problem, we are given a graph G whose edges are generated by starting with a random graph $G_{n,1/2}$, then "planting," i.e., adding edges to form a clique on k randomly chosen vertices. Jerrum [30] and Kučera [37] introduced the planted clique problem as a potentially easier variant of the classical problem of finding the largest clique in a random graph [33]. A random graph $G_{n,1/2}$ contains a clique of size $2\log n$ with high probability, and a simple greedy algorithm can find one of size $\log n$. Finding cliques of size $(2-\epsilon)\log n$ is a hard problem for any $\epsilon > 0$. Planting a larger clique should make it easier to find one. The problem of finding the smallest k for which the planted clique can be detected in polynomial time has attracted significant attention. For $k \geq c\sqrt{n \log n}$, simply picking vertices of large degrees suffices [37]. Cliques of size $k = \Omega(\sqrt{n})$ can be found using spectral methods [2, 38, 14], via SDPs [19], nuclear norm minimization [3] or combinatorial methods [21, 15].

While there is no known polynomial-time algorithm that can detect cliques of size below the threshold of $\Omega(\sqrt{n})$, there is a quasipolynomial algorithm for any $k \geq 2\log n$: enumerate subsets of size $2\log n$; for each subset that forms a clique, take all common neighbors of the subset; one of these will be the planted clique. This is also the fastest known algorithm for any $k = O(n^{1/2-\delta})$, where $\delta > 0$.

Some evidence of the hardness of the problem was shown by Jerrum [30] who proved that a specific approach using a Markov chain cannot be efficient for $k = o(\sqrt{n})$. More evidence of hardness is given in [20], where it is shown that Lovász-Schrijver SDP relaxations, which include the SDP used in [19], cannot be used to efficiently find cliques of size $k = o(\sqrt{n})$. The problem has been used to generate cryptographic primitives [31], and as a hardness assumption [1, 28, 40].

We focus on the bipartite planted clique problem, where a bipartite $k \times k$ clique is planted in a random bipartite graph. A densest-subgraph version of the bipartite planted clique problem has been used as a hard problem for cryptographic applications [4]. All known bounds and algorithms for the k-clique problem can be easily adapted to the bipartite case. Therefore it is natural to suspect that new upper bounds on the planted k-clique problem would also yield new upper bounds for the bipartite case.

The starting point of our investigation for this problem is the property of the bipartite planted k-clique problem is that it has an equivalent formulation as a problem over distributions defined as follows.

PROBLEM 1. *Fix an integer k, $1 \leq k \leq n$, and a subset of k indices $S \subseteq \{1,2,\ldots,n\}$. The input distribution D_S on vectors $x \in \{0,1\}^n$ is defined as follows: with probability $1-(k/n)$, x is uniform over $\{0,1\}^n$; and with probability k/n the k coordinates of x in S are set to 1, and the rest are uniform over their support. For an integer t, the **distributional planted k-biclique** problem with t samples is the problem of finding the unknown subset S using t samples drawn randomly from D_S.*

One can view samples $x_1,\ldots,x_t$ as adjacency vectors of a bipartite graph as follows: the bipartite graph has n vertices on the right (with k marked as members of the clique) and t vertices on the left. Each of the t samples gives the adjacency vector of the corresponding vertex on the left. It is not hard to see that for $t = n$, conditioned on the event of getting exactly k samples with planted indices, we will get a random bipartite graph with a planted k-biclique (we prove the equivalence formally in the full version of this paper [23]).

Another interesting approach for planted clique was proposed by Frieze and Kannan [24]. They gave a reduction

from finding a planted clique in a random graph to finding a direction that maximizes a tensor norm; this was extended to general r in [12]. Specifically, they show that maximizing the r'th moment (or the 2-norm of an r'th order tensor) allows one to recover planted cliques of size $\tilde{\Omega}(n^{1/r})$. A related approach is to maximize the 3rd or higher moment of the distribution given by the distributional planted clique problem. For this approach it is natural to consider the following type of optimization algorithm: start with some unit vector u, then estimate the gradient at u (via samples), move along that direction and return to the sphere; repeat to reach an approximate local maximum. Unfortunately, over the unit sphere, the expected r'th moment function can have (exponentially) many local maxima even for simple distributions. A more sophisticated approach [32] is through Markov chains or simulated annealing; it attempts to sample unit vectors from a distribution on the sphere which is heavier on vectors that induce a higher moment, e.g., u is sampled with density proportional to $e^{f(u)}$ where $f(u)$ is the expected r'th moment along u. This could be implemented by a Markov chain with a Metropolis filter [39, 27] ensuring a proportional steady state distribution. If the Markov chain were to mix rapidly, that would give an efficient approximation algorithm because sampling from the steady state likely gives a vector of high moment. At each step, all one needs is to be able to estimate $f(u)$, which can be done by sampling from the input distribution.

As we will see presently, these statistical approaches will have provably high complexity. For the bipartite planted clique problem, statistical algorithms that can query expectations of arbitrary functions to within a small tolerance need $n^{\Omega(\log n)}$ queries to detect planted cliques of size $k < n^{\frac{1}{2}-\delta}$ for any $\delta > 0$. Even stronger exponential bounds apply for the more general problem of detecting planted dense subgraphs of the same size. These bounds match the current upper bounds.

PROBLEM 2. *Fix $0 < q \leq p \leq 1$. For $1 \leq k \leq n$, let $S \subseteq \{1, 2, \ldots, n\}$ be a set of k vertex indices and D_S be a distribution over $\{0,1\}^n$ such that when $x \sim D_S$, with probability $1 - (k/n)$ the entries of x are independently chosen according to a q biased Bernoulli variable, and with probability k/n the k coordinates in S are independently chosen according to a p biased Bernoulli variable, and the rest are independently chosen according to a q biased Bernoulli variable. The* ***generalized planted bipartite densest k-subgraph*** *problem is to find the unknown subset S given access to samples from D_S.*

To describe these results precisely and discuss exactly what they mean for the complexity of these problems, we will need to define the notion of statistical algorithms, the complexity measures we use, and our main tool for proving lower bounds, a notion of statistical dimension of a set of distributions. We do this in the next section. In Section 3 we prove our general lower bound theorems, and in Section 5 we estimate the statistical dimension of detecting planted cliques and dense subgraphs.

2. DEFINITIONS AND OVERVIEW

Here we formally define statistical algorithms and the key notion of statistical dimension, and then describe the resulting lower bounds in detail.

2.1 Problems over Distributions

We begin by formally defining the class of problems addressed by our framework.

DEFINITION 1 (SEARCH PROBLEMS OVER DISTRIBUTIONS). *For a domain X, let $\mathcal{D}$ be a set of distributions over X, let $\mathcal{F}$ be a set of solutions and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map from a distribution $D \in \mathcal{D}$ to a subset of solutions $\mathcal{Z}(D) \subseteq \mathcal{F}$ that are defined to be valid solutions for D. For $t > 0$ the* distributional search problem *$\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ using t samples is to find a valid solution $f \in \mathcal{Z}(D)$ given access to t random samples from an unknown $D \in \mathcal{D}$.*

We note that this definition captures decision problems by having $\mathcal{F} = \{0, 1\}$. With slight abuse of notation, for a solution $f \in \mathcal{F}$, we denote by $\mathcal{Z}^{-1}(f)$ the set of distributions in $\mathcal{D}$ for which f is a valid solution.

It is important to note that the number of available samples t can have a major influence on the complexity of the problem. First, for most problems there is a minimum t for which the problem is information-theoretically solvable. This value is often referred to as the *sample complexity* of the problem. But even for t which is larger than the sample complexity of the problem, having more samples can make the problem easier computationally. For example, in the context of attribute-efficient learning, there is a problem that is intractable with few samples (under cryptographic assumptions) but is easy to solve with a larger (but still polynomial) number of samples [43]. Our distributional planted biclique problem exhibits the same phenomenon.

2.2 Statistical Algorithms

The statistical query learning model of Kearns [34] is a restriction of the PAC model [46]. It captures algorithms that rely on empirical estimates of statistical properties of random examples of an unknown function instead of individual random examples (as in the PAC model of learning).

For general search, decision and optimization problems over a distribution, we define statistical algorithms as algorithms that do not see samples from the distribution but instead have access to estimates of the expectation of any bounded function of a sample from the distribution.

DEFINITION 2 (STAT ORACLE). *Let D be the input distribution over the domain X. For a* tolerance *parameter $\tau > 0$, STAT(τ) oracle is the oracle that for any query function $h : X \to [-1, 1]$, returns a value*

$$v \in [E_{x \sim D}[h(x)] - \tau, E_{x \sim D}[h(x)] + \tau] .$$

The general algorithmic techniques mentioned earlier can all be expressed as algorithms using STAT oracle instead of samples themselves, in most cases in a straightforward way. We would also like to note that in the PAC learning model some of the algorithms, such as the Perceptron algorithm, did not initially appear to fall into the SQ framework but SQ analogues were later found for all known learning techniques except Gaussian elimination (for specific examples, see [34] and [10]). We expect the situation to be similar even in the broader context of search problems over distributions.

The most natural realization of STAT(τ) oracle is one that computes h on $O(1/\tau^2)$ random samples from D and returns their average. Chernoff's bound would then imply that the estimate is within the desired tolerance (with constant probability). However, if $h(x)$ is very biased (e.g. equal

to 0 with high probability), it can be estimated with fewer samples. Our primary application requires a tight bound on the number of samples necessary to solve a problem over distributions. Therefore we define a stronger version of STAT oracle in which the tolerance is adjusted for the variance of the query function, that is the oracle returns the expectation to within the same tolerance (up to constant factors) as one expects to get from t samples. More formally, for a Boolean query h, VSTAT(t) can return any value v for which the distribution $B(t, v)$ (sum of t independent Bernoulli variables with bias v) is statistically close to $B(t, E[h])$ (see Sec. 3.2 for more details on this correspondence).

DEFINITION 3 (VSTAT ORACLE). *Let D be the input distribution over the domain X. For a* sample size *parameter $t > 0$, VSTAT(t) oracle is the oracle that for any query function $h : X \to \{0, 1\}$, returns a value $v \in [p - \tau, p + \tau]$, where $p = E_{x \sim D}[h(x)]$ and $\tau = \max\left\{\frac{1}{t}, \sqrt{\frac{p(1-p)}{t}}\right\}$.*

Note that VSTAT(t) always returns the value of the expectation within $1/\sqrt{t}$. Therefore it is stronger than STAT($1/\sqrt{t}$) but weaker than STAT($1/t$). (STAT, unlike VSTAT, allows non-boolean functions but this is not a significant difference as any $[-1, 1]$-valued query can be converted to a logarithmic number of $\{0, 1\}$-valued queries).

The STAT and VSTAT oracles we defined can return any value within the given tolerance and therefore can make adversarial choices. We also aim to prove lower bounds against algorithms that use a potentially more benign, "unbiased" statistical oracle. The unbiased statistical oracle gives the algorithm the true value of a boolean query function on a randomly chosen sample. This model is based on the Honest SQ model in learning by Yang [49] (which itself is based on an earlier model of Jackson [29]).

DEFINITION 4 (1-STAT ORACLE). *Let D be the input distribution over the domain X. The 1-STAT oracle is the oracle that given any function $h : X \to \{0, 1\}$, takes an independent random sample x from D and returns $h(x)$.*

Note that the 1-STAT oracle draws a fresh sample upon each time it is called. Without re-sampling each time, the answers of the 1-STAT oracle could be easily used to recover any sample bit-by-bit, making it equivalent to having access to random samples. The **query complexity** of an algorithm using 1-STAT is defined to be the number of calls it makes to the 1-STAT oracle. Note that the 1-STAT oracle can be used to simulate VSTAT (with high probability) by taking the average of $O(t)$ replies of 1-STAT for the same function h. While it might seem that access to 1-STAT gives an algorithm more power than access to VSTAT we will show that t samples from 1-STAT can be simulated using access to VSTAT($O(t)$). This will allow us to translate our lower bounds on algorithms with access to VSTAT to query complexity lower bounds for unbiased statistical algorithms.

In the following discussion, we refer to algorithms using STAT or VSTAT oracles (instead of samples) as **statistical algorithms**. Algorithms using the 1-STAT oracle are henceforth called **unbiased statistical algorithms**.

2.3 Statistical Dimension

The main tool in our analysis is an information-theoretic bound on the complexity of statistical algorithms. Our definitions originate from the statistical query (SQ) dimension in learning theory [11] used to characterize SQ learning algorithms. Roughly speaking, the SQ dimension corresponds to the number of nearly uncorrelated labeling functions in a class (see the full version for the details of the definition and the relationship to our bounds [23]).

We introduce a natural generalization and strengthening of this approach to search problems over arbitrary sets of distributions and prove lower bounds on the complexity of statistical algorithms based on the generalized notion. Our definition departs from SQ dimension in three aspects. (1) Our notion applies to any set of distributions whereas in the learning setting all known dimensions require fixing the distribution over the domain and only allow varying the labeling function. (2) Instead of relying on a bound on pairwise correlations, our dimension relies on a bound on average correlations in a large set of distributions. This weaker condition allows us to derive the tight bounds on the complexity of statistical algorithms for the planted k-biclique problem. (3) We show that our dimension also gives lower bounds for the stronger VSTAT oracle (without incurring a quadratic loss in the sample size parameter).

We now define our dimension formally. For two functions $f, g : X \to \mathcal{R}$ and a distribution D with probability density function $D(x)$, the inner product of f and g over D is defined as $\langle f, g \rangle_D \doteq \mathbf{E}_{x \sim D}[f(x)g(x)]$. The norm of f over D is $\|f\|_D = \sqrt{\langle f, f \rangle_D}$. We remark that, by convention, the integral from the inner product is taken only over the support of D, i.e. for $x \in X$ such that $D(x) \neq 0$. Given a distribution D over X let $D(x)$ denote the probability density function of D relative to some fixed underlying measure over X (for example uniform distribution for discrete X or Lebesgue measure over $\mathcal{R}^n$). Our bound is based on the inner products between functions of the following form: $(D'(x) - D(x))/D(x)$ where D' and D are distributions over X. For this to be well-defined, we will only consider cases where $D(x) = 0$ implies $D'(x) = 0$, in which case $D'(x)/D(x)$ is treated as 1. To see why such functions are relevant to our discussion, note that for every real-valued function f over X,

$$\mathop{\mathbf{E}}_{x \sim D'}[f(x)] - \mathop{\mathbf{E}}_{x \sim D}[f(x)] = \mathop{\mathbf{E}}_{x \sim D}[D'(x)f(x)/D(x)] - \mathop{\mathbf{E}}_{x \sim D}[f(x)] = \left\langle \frac{D' - D}{D}, f \right\rangle_D .$$

This means that the inner product of any function f with $(D' - D)/D$ is equal to the difference of expectations of f under the two distributions. We also remark that the quantity $\langle \frac{D'}{D} - 1, \frac{D'}{D} - 1 \rangle_D$ is known as the $\chi^2(D', D)$ distance and is widely used for hypothesis testing in statistics [41]. A key notion for our statistical dimension is the *average correlation* of a set of distributions $\mathcal{D}'$ relative to a distribution D. We denote it by $\rho(\mathcal{D}', D)$ and define as follows:

$$\rho(\mathcal{D}', D) \doteq \frac{1}{m^2} \sum_{D_1, D_2 \in \mathcal{D}'} \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| .$$

Known lower bounds for SQ learning are based on bounding the pairwise correlations between functions. Bounds on pairwise correlations when strong enough easily imply bounds on the average correlation. Bounding average correlation is more involved technically but it appears to be necessary for the tight lower bounds that we give. In Section 3.3 we describe a pairwise-correlation version of our bounds. It is sufficient for some applications and generalizes

the statistical query dimension from learning theory (see full version for the details).

We now define the concept of statistical dimension.

DEFINITION 5. *For $\bar{\gamma} > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, let d be the largest integer such that there exists a* reference *distribution D over X and a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ with the following property: for any solution $f \in \mathcal{F}$ the set $\mathcal{D}_f = \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ is non-empty and for any subset $\mathcal{D}' \subseteq \mathcal{D}_f$, where $|\mathcal{D}'| \geq |\mathcal{D}_f|/d$, $\rho(\mathcal{D}', D) < \bar{\gamma}$. We define the* ***statistical dimension*** *with average correlation $\bar{\gamma}$ of $\mathcal{Z}$ to be d and denote it by* $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$.

The statistical dimension with average correlation $\bar{\gamma}$ of a search problem over distributions gives a lower bound on the complexity of any (deterministic) statistical algorithm for the problem that uses queries to VSTAT$(1/(3\bar{\gamma}))$.

THEOREM 1. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\bar{\gamma} > 0$ let $d = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$. Any statistical algorithm requires at least d calls to VSTAT$(1/(3\bar{\gamma}))$ oracle to solve $\mathcal{Z}$.*

At a high level our proof works as follows. It is not hard to see that for any D, an algorithm that solves $\mathcal{Z}$ needs to "distinguish" all distributions in $\mathcal{D}_f$ from D for some solution f. Here by "distinguishing" we mean that the algorithm needs to ask a query g such that $\mathbf{E}_D[g]$ cannot be used as a response of VSTAT$(1/(3\bar{\gamma}))$ for $D' \in \mathcal{D}_f$. In the key component of the proof we show that if a query function g to VSTAT$(1/(3\bar{\gamma}))$ distinguishes between a distribution D and any distribution $D' \in \mathcal{D}'$, then $\mathcal{D}'$ must have average correlation of at least $\bar{\gamma}$ relative to D. The condition that for any $|\mathcal{D}'| \geq |\mathcal{D}_f|/d$, $\rho(\mathcal{D}', D) < \bar{\gamma}$ then immediately implies that at least d queries are required to distinguish any distribution in $\mathcal{D}_f$ from D.

In Section 3.1 we give a refinement of SDA which additionally requires that the set $\mathcal{D}_f$ is not too small (and not just non-empty). This refined notion allows us to extend the lower bound to randomized statistical algorithms and then to unbiased statistical algorithms.

2.4 Lower Bounds

We will prove lower bounds for the bipartite planted clique under both statistical oracles (VSTAT and 1-STAT) .

THEOREM 2. *For any constant $\delta > 0$, any $k \leq n^{1/2-\delta}$ and $r > 0$, at least $n^{\Omega(\log r)}$ queries to VSTAT$(n^2/(rk^2))$ are required to solve the distributional planted bipartite k-clique. No polynomial-time statistical algorithm can solve the problem using queries to VSTAT$(o(n^2/k^2))$ and any statistical algorithm will require $n^{\Omega(\log n)}$ queries to VSTAT$(n^{2-\delta}/k^2)$.*

This bound also applies to any randomized algorithm with probability of success being at least a (positive) constant. Note that this bound is essentially tight. For every vertex in the clique, the probability that the corresponding bit of a randomly chosen point is set to 1 is $1/2 + k/(2n)$ whereas for every vertex not in the clique, this probability is $1/2$. Therefore using n queries to VSTAT$(16n^2/k^2)$ (i.e., of tolerance $k/4n$) it is easy to detect the planted bipartite clique. Indeed, this can be done by using the query functions $h : \{0, 1, \}^n \to \{0, 1\}$, $h(x) = x_i$, for each $i \in [n]$. So, the answers of the VSTAT oracle represent the expected value of the ith bit over the sample.

There is also a statistical algorithm that uses $n^{O(\log n)}$ queries to VSTAT$(25n/k)$ (significantly higher tolerance) to find the planted set for any $k \geq \log n$. In fact, the same algorithm can be used for the standard planted clique problem that achieves complexity $n^{O(\log n)}$. We enumerate over subsets $T \subseteq [n]$ of $\log n$ indices and query VSTAT$(25n/k)$ with the function $g_T : \{0, 1\}^n \to \{0, 1\}$ defined as 1 if and only if the point has ones in all coordinates in T. Therefore, if the set T is included in the true clique then

$$\mathop{\mathbf{E}}_{D}[g_T] = \frac{k}{n} \cdot 1 + \left(1 - \frac{k}{n}\right) 2^{-\log n} \in \left[\frac{k}{n}, \frac{k+1}{n}\right].$$

With this expectation, VSTAT$(25n/k)$ has tolerance at most $\sqrt{k(k+1)/25n^2} \leq (k+1)/5n$ and will return at least $k/n - (k+1)/(5n) > 3k/(4n)$. If, on the other hand, at least one element of T is not from the planted clique, then $\mathbf{E}_D[g_T] \leq k/(2n) + 1/n$ and VSTAT$(25n/k)$ will return at most $(k+2)/(2n) + (k+2)/(5n) < 3k/(4n)$. Thus, we will know all $\log n$-sized subsets of the planted clique and hence the entire clique.

Our bounds have direct implications for the average-case planted bipartite k-clique problem. An instance of this problem is a random $n \times n$ bipartite graph with a $k \times k$ bipartite clique planted randomly. In the full version of this paper, we show that the average-case planted k-biclique is equivalent to our distributional planted k-biclique with n samples. For a statistical algorithm, n samples directly correspond to having access to VSTAT$(O(n))$. Our bounds show that this problem can be solved in polynomial time when $k = \Omega(\sqrt{n})$. At the same time, for $k \leq n^{1/2-\delta}$, any statistical algorithm will require $n^{\Omega(\log n)}$ queries to VSTAT$(n^{1+\delta})$.

We also give a bound for unbiased statistical algorithms.

THEOREM 3. *For any constant $\delta > 0$ and any $k \leq n^{1/2-\delta}$, any unbiased statistical that succeeds with probability at least $2/3$ requires $\tilde{\Omega}(n^2/k^2)$ queries to solve the distributional planted bipartite k-clique problem.*

Each query of an unbiased statistical algorithm requires a new sample from D. Therefore this bound implies that any algorithm that does not reuse samples will require $\tilde{\Omega}(n^2/k^2)$ samples. To place this bound in context, we note that it is easy to detect whether a clique of size k has been planted using $\tilde{O}(n^2/k^2)$ samples (as before, to detect if a coordinate i is in the clique we can compute the average of x_i on $\tilde{O}(n^2/k^2)$ samples). Of course, finding all vertices in the clique would require reusing samples (which statistical algorithms cannot do). Note that $n^2/k^2 \leq n$ if and only if $k \geq \sqrt{n}$.

These bounds match the state of the art for the average-case planted bipartite k-clique and planted k-clique problems. One reason why we consider this natural is that some of the algorithms that solve the problem for $k = \Omega(\sqrt{n})$ are obviously statistical. For example, the key procedure of the algorithm of Feige and Ron [21] removes a vertex that has the lowest degree (in the current graph) and repeats until the remaining graph is a clique. The degree is the number of ones in a column (or row) of the adjacency matrix and can be viewed of as an estimate of the expectation of 1 appearing in the corresponding coordinate of a random sample. Even the more involved algorithms for the problem, like finding the eigenvector with the largest eigenvalue or solving an SDP, have statistical analogues.

A closely related problem is the *planted densest subgraph* problem, where edges in the planted subset appear with higher probability than in the remaining graph. This is a variant of the densest k-subgraph problem, which itself is a natural generalization of k-clique that asks to recover the densest k-vertex subgraph of a given n-vertex graph [18, 35, 7, 8]. The conjectured hardness of its average case variant, the planted densest subgraph problem, has been used in public key encryption schemes [4] and in analyzing parameters specific to financial markets [5]. Our lower bounds extend in a straightforward manner to this problem.

THEOREM 4. *For any constant $\delta > 0$, any $k \leq n^{1/2-\delta}$, $\ell \leq k$ and density parameter $p = 1/2 + \alpha$, at least $n^{\Omega(\ell)}$ queries to VSTAT($n^2/(\ell\alpha^2 k^2)$) are required to solve the distributional planted bipartite densest k-subgraph with density p. For constants $c, \delta > 0$, density $1/2 < p \leq 1/2 + 1/n^c$, and $k \leq n^{1/2-\delta}$, any unbiased statistical algorithm requires $\tilde{\Omega}((n^{2+2c})/k^2)$ queries to find a planted densest subgraph of size k.*

For example, taking $k = n^{1/3}$, $l = k$ and $\alpha = 1/n^c$ yields a lower bound of $n^{\Omega(n^{1/3})}$ for the VSTAT(n^{1+2c}) oracle.

3. LOWER BOUNDS FROM STATISTICAL DIMENSION

We refer the reader to the full version for the technical details [23] omitted here.

3.1 Lower Bounds for Statistical Algorithms

We now prove Theorem 1 which is the basis of all our lower bounds. We will prove a stronger version of this theorem which also applies to randomized algorithms. For this version we need an additional parameter in the definition of SDA.

DEFINITION 6. *For $\bar{\gamma} > 0$, $\eta > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, let d be the largest integer such that there exists a* reference *distribution D over X and a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ with the following property: for any solution $f \in \mathcal{F}$ the set $\mathcal{D}_f = \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ has size at least $(1-\eta) \cdot |\mathcal{D}_D|$ and for any subset $\mathcal{D}' \subseteq \mathcal{D}_f$, where $|\mathcal{D}'| \geq |\mathcal{D}_f|/d$, $\rho(\mathcal{D}', D) < \bar{\gamma}$. We define the **statistical dimension** with average correlation $\bar{\gamma}$ and solution set bound η of $\mathcal{Z}$ to be d and denote it by* $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$.

Note that for any $\eta < 1$, $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}) \geq \mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$ and for $1 > \eta \geq 1 - 1/|\mathcal{D}_D|$, we get $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}) = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$, where $|\mathcal{D}_D|$ is the set of distributions that maximizes $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$.

THEOREM 5. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\bar{\gamma} > 0$ and $\eta \in (0,1)$ let $d = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$. Any randomized statistical algorithm that solves $\mathcal{Z}$ with probability $\delta > \eta$ requires at least $\frac{\delta-\eta}{1-\eta} d$ calls to VSTAT($1/(3\bar{\gamma})$).*

Theorem 1 is obtained from Theorem 5 by setting $\delta = 1$ and using any $1 - 1/|\mathcal{D}_D| \leq \eta < 1$. Further, for any $\eta < 1$, $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}) \geq \mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$ and therefore for any $\eta < 1$, a bound on $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$ can be used in Theorem 1 in place of bound on $\mathrm{SDA}(\mathcal{Z}, \bar{\gamma})$.

We will need the following simple lemma that bounds the distance between any $p \in [0,1]$ and p' which is returned by VSTAT(t) on a query with expectation p in terms of p'.

LEMMA 1. *For an integer t and any $p \in [0,1]$, let $p' \in [0,1]$ be such that $|p' - p| \geq \tau = \max\left\{\frac{1}{t}, \sqrt{\frac{p(1-p)}{t}}\right\}$. Then $|p' - p| \geq \sqrt{\frac{\min\{p', 1-p'\}}{3t}}$.*

We are now ready to prove Theorem 5.

PROOF OF THEOREM 5. We prove our lower bound by exhibiting a distribution over inputs (which are distributions over X) for which every deterministic statistical algorithm that solves $\mathcal{Z}$ with probability δ (over the choice of input) requires at least $(\delta - \eta) \cdot d/(1-\eta)$ calls to VSTAT($1/(3\bar{\gamma})$). The claim of the theorem will then follow by Yao's minimax principle [51].

Let D be the reference distribution and $\mathcal{D}_D$ be a set of distributions for which the value d is achieved. Let $\mathcal{A}$ be a deterministic statistical algorithm that uses q queries to VSTAT($1/(3\bar{\gamma})$) to solve $\mathcal{Z}$ with probability δ over a random choice of a distribution from $\mathcal{D}_D$. Following an approach from [22], we simulate $\mathcal{A}$ by answering any query $h : X \to \{0,1\}$ of $\mathcal{A}$ with value $\mathbf{E}_D[h(x)]$. Let $h_1, h_2, \ldots, h_q$ be the queries asked by $\mathcal{A}$ in this simulation and let f be the output of $\mathcal{A}$.

By the definition of SDA, for $\mathcal{D}_f = \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ it holds that $|\mathcal{D}_f| \geq (1-\eta)|\mathcal{D}_D|$ and for every $\mathcal{D}' \subseteq \mathcal{D}_f$, either $\rho(\mathcal{D}', D) < \bar{\gamma}$ or $|\mathcal{D}'| \leq |\mathcal{D}_f|/d$. Let the set $\mathcal{D}^+ \subseteq \mathcal{D}_D$ be the set of distributions on which $\mathcal{A}$ is successful. Let $\mathcal{D}_f^+ = \mathcal{D}_f \cap \mathcal{D}^+$ and we denote these distributions by $\{D_1, D_2, \ldots, D_m\}$. We note that $\mathcal{D}_f^+ = \mathcal{D}^+ \setminus (\mathcal{D}_D \setminus \mathcal{D}_f)$ and therefore

$$m = |\mathcal{D}_f^+| \geq |\mathcal{D}^+| - |\mathcal{D}_D \setminus \mathcal{D}_f| \geq \delta|\mathcal{D}_D| - |\mathcal{D}_D \setminus \mathcal{D}_f|$$
$$= \frac{\delta|\mathcal{D}_D| - |\mathcal{D}_D \setminus \mathcal{D}_f|}{|\mathcal{D}_D| - |\mathcal{D}_D \setminus \mathcal{D}_f|}|D_f| \geq \frac{\delta-\eta}{1-\eta}|\mathcal{D}_f|. \quad (1)$$

To lower bound q, we use a generalization of an elegant argument of Szörényi [44]. For every $k \leq q$, let A_k be the set of all distributions D_i such that

$$\left|\mathop{\mathbf{E}}_{D}[h_k(x)] - \mathop{\mathbf{E}}_{D_i}[h_k(x)]\right| > \tau_{i,k} \doteq \max\left\{\frac{1}{t}, \sqrt{\frac{p_{i,k}(1-p_{i,k})}{t}}\right\},$$

where we use t to denote $1/(3\bar{\gamma})$ and $p_{i,k}$ to denote $\mathbf{E}_{D_i}[h_k(x)]$. To prove the desired bound we first prove the following two claims:

1. $\sum_{k \leq q} |A_k| \geq m$;
2. for every k, $|A_k| \leq |\mathcal{D}_f|/d$.

Combining these two implies that $q \geq d \cdot m/|\mathcal{D}_f|$. By inequality (1), $q \geq \frac{\delta-\eta}{1-\eta} \cdot d$ giving the desired lower bound. In the rest of the proof for conciseness we drop the subscript D for inner products and norms.

To prove the first claim we assume, for the sake of contradiction, that there exists $D_i \notin \cup_{k \leq q} A_k$. Then for every $k \leq q$, $|\mathbf{E}_D[h_k(x)] - \mathbf{E}_{D_i}[h_k(x)]| \leq \tau_{i,k}$. This implies that the replies of our simulation $\mathbf{E}_D[h_k(x)]$ are within $\tau_{i,k}$ of $\mathbf{E}_{D_i}[h_k(x)]$. By the definition of $\mathcal{A}$ and VSTAT(t), this implies that f is a valid solution for $\mathcal{Z}$ on D_i, contradicting the condition that $D_i \in \mathcal{D}_f^+ \subseteq \mathcal{D}_D \setminus Z^{-1}(f)$.

To prove the second claim, suppose that for some $k \in [d]$, $|A_k| > |\mathcal{D}_f|/d$. We will denote $p_k = \mathbf{E}_D[h_k(x)]$. We assume that $p_k \leq 1/2$ (when $p_k > 1/2$, we simply replace h_k by

$1 - h_k$). We note first that:

$$\begin{aligned}\mathop{\mathbf{E}}_{D_i}[h_k(x)] - \mathop{\mathbf{E}}_{D}[h_k(x)] &= \mathop{\mathbf{E}}_{D}\left[\frac{D_i(x)}{D(x)}\, h_k(x)\right] - \mathop{\mathbf{E}}_{D}[h_k(x)] \\ &= \left\langle h_k, \frac{D_i}{D} - 1 \right\rangle \\ &= p_{i,k} - p_k.\end{aligned}$$

Let $\hat{D}_i(x) = \frac{D_i(x)}{D(x)} - 1$, (where the convention is that $\hat{D}_i(x) = 0$ if $D(x) = 0$). We will next show upper and lower bounds on the following quantity

$$\Phi = \left\langle h_k, \sum_{D_i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right\rangle.$$

By Cauchy-Schwartz we have that

$$\begin{aligned}\Phi^2 &= \left\langle h_k, \sum_{D_i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right\rangle^2 \\ &\le \|h_k\|^2 \cdot \left\| \sum_{D_i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right\|^2 \\ &\le \|h_k\|^2 \cdot \left(\sum_{D_i, D_j \in A_k} \left|\langle \hat{D}_i, \hat{D}_j\rangle\right| \right) \\ &\le \|h_k\|^2 \cdot \rho(A_k, D) \cdot |A_k|^2. \end{aligned} \tag{2}$$

As before, we also have that

$$\begin{aligned}\Phi^2 &= \left\langle h_k, \sum_{D_i \in A_k} \hat{D}_i \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right\rangle^2 \\ &= \left(\sum_{D_i \in A_k} \langle h_k, \hat{D}_i\rangle \cdot \mathsf{sign}\langle h_k, \hat{D}_i\rangle \right)^2 \\ &\ge \left(\sum_{D_i \in A_k} |p_{i,k} - p_k| \right)^2. \end{aligned} \tag{3}$$

To evaluate the last term of this inequality we use the fact that $|p_{i,k} - p_k| \ge \tau_{i,k} = \max\{1/t, \sqrt{p_{i,k}(1-p_{i,k})/t}\}$ and Lemma 1 to obtain that for every $D_i \in A_k$,

$$|p_k - p_{i,k}| \ge \sqrt{\frac{\min\{p_k, 1-p_k\}}{3t}} = \sqrt{\frac{p_k}{3t}}. \tag{4}$$

By substituting equation (4) into (3) we get that $\Phi^2 \ge \frac{p_k}{3t} \cdot |A_k|^2$.

We note that, h_k is a $\{0,1\}$-valued function and therefore $\|h_k\|^2 = p_k$. Substituting this into equation (2) we get that $\Phi^2 \le p_k \cdot \rho(A_k, D) \cdot |A_k|^2$. By combining these two bounds on Φ^2 we obtain that $\rho(A_k, D) \ge 1/(3t) = \bar{\gamma}$ which contradicts the definition of SDA. □

3.2 Lower Bounds for Unbiased Statistical Algorithms

Next we address lower bounds on algorithms that use the 1-STAT oracle. We recall that the 1-STAT oracle returns the value of a function on a single randomly chosen point. To estimate the expectation of a function, an algorithm can simply query this oracle multiple times with the same function and average the results. A lower bound for this oracle directly translates to a lower bound on the number of samples that any statistical algorithm must use.

We note that responses of 1-STAT do not have the room for the possibly adversarial deviation afforded by the tolerance of the STAT and VSTAT oracles. The ability to use these slight deviations in a coordinated way is used crucially in our lower bounds against VSTAT and in known lower bounds for SQ learning algorithms. This makes proving lower bounds against unbiased statistical algorithms harder and indeed lower bounds for Honest SQ learning (which is a special case of unbiased statistical algorithms) required a substantially more involved argument than lower bounds for the regular SQ model [50].

Our lower bounds for unbiased statistical algorithms use a different approach. We show a direct simulation of 1-STAT oracle using VSTAT oracle.

THEOREM 6. *Let $\mathcal{Z}$ be a search problem and let $\mathcal{A}$ be a (possibly randomized) unbiased statistical algorithm that solves $\mathcal{Z}$ with probability at least δ using m samples from 1-STAT. For any $\delta' \in (0, 1/2]$, there exists a statistical algorithm $\mathcal{A}'$ that uses at most m queries to VSTAT(m/δ'^2) and solves $\mathcal{Z}$ with probability at least $\delta - \delta'$.*

Our proof relies on a simple simulation. which implies that success probability of the simulated algorithm is not much worse than that of the unbiased statistical algorithm.

THEOREM 7. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\bar{\gamma} > 0$ and $\eta \in (0,1)$, let $d = \mathrm{SDA}(\mathcal{Z}, \bar{\gamma}, \eta)$. Any (possibly randomized) unbiased statistical algorithm that solves $\mathcal{Z}$ with probability δ requires at least m calls to 1-STAT for*

$$m = \min\left\{ \frac{d(\delta - \eta)}{2(1-\eta)}, \frac{(\delta-\eta)^2}{12\bar{\gamma}} \right\}.$$

In particular, if $\eta \le 1/6$ then any algorithm with success probability of at least $2/3$ requires at least $\min\{d/4, 1/48\bar{\gamma}\}$ samples from 1-STAT.

To conclude we note that the reduction in the other direction is trivial, namely that VSTAT(t) oracle can be simulated using 1-STAT oracle. We provide details in the full version of this paper.

3.3 Statistical Dimension from Pairwise Correlations

In addition to SDA which is based on average, correlation, we introduce a simpler notion based on pairwise correlatons.

DEFINITION 7. *For $\gamma, \beta > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, let m be the maximum integer such that there exists a* reference *distribution D over X and a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that for any solution $f \in \mathcal{F}$, there exists a set of m distributions $\mathcal{D}_f = \{D_1, \ldots, D_m\} \subseteq \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ with the following property:*

$$\left| \left\langle \frac{D_i}{D} - 1, \frac{D_j}{D} - 1 \right\rangle_D \right| \le \begin{cases} \beta \text{ for } i = j \in [m] \\ \gamma \text{ for } i \ne j \in [m]. \end{cases}$$

*We define the **statistical dimension** with pairwise correlations (γ, β) of $\mathcal{Z}$ to be m and denote it by $\mathrm{SD}(\mathcal{Z}, \gamma, \beta)$.*

Using this notion, we can prove the following lower bound (which is a corollary of Theorem 1):

THEOREM 8. *Let X be a domain and $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X. For $\gamma, \beta > 0$, let $m = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$. For any $\tau > 0$, any statistical algorithm requires at least $m(\tau^2 - \gamma)/(\beta - \gamma)$ calls to the STAT(τ) oracle to solve $\mathcal{Z}$. In particular, if for $m > 0$, $\mathrm{SD}(\mathcal{Z}, \gamma = \frac{m^{-2/3}}{2}, \beta = 1) \geq m$ then at least $m^{1/3}/2$ calls of tolerance $m^{-1/3}$ to the STAT oracle are required to solve $\mathcal{Z}$.*

4. WARM-UP: MAX-XOR-SAT

In this section, we demonstrate our techniques on a warm-up problem, MAX-XOR-SAT. For this problem, it is sufficient to use pairwise correlations, rather than average correlations.

The MAX-XOR-SAT problem over a distribution is defined as follows.

PROBLEM 3. *For $\epsilon \geq 0$, the ϵ-approximate **MAX-XOR-SAT** problem is defined as follows. Given samples from some unknown distribution D over XOR clauses on n variables, find an assignment x that maximizes up to additive ϵ the probability a random clause drawn from D is satisfied.*

In the worst case, it is known that MAX-XOR-SAT is NP-hard to approximate to within $1/2 - \delta$ for any constant δ [26]. In practice, local search algorithms such as WalkSat [42] are commonly applied as heuristics for maximum satisfiability problems. We give strong evidence that the distributional version of MAX-XOR-SAT is hard for algorithms that locally seek to improve an assignment by flipping variables as to satisfy more clauses, giving some theoretical justification for the observations of [42]. Moreover, our proof even applies to the case when there exists an assignment that satisfies all the clauses generated by the target distribution.

The bound we obtain can be viewed as a restatement of the known lower bound for learning parities using statistical query algorithms (indeed, the problem of learning parities is a special case of our distributional MAX-XOR-SAT).

THEOREM 9. *For any $\delta > 0$, any statistical algorithm requires at least $\tau^2(2^n - 1)$ queries to STAT(τ) to solve $\left(\frac{1}{2} - \delta\right)$-approximate MAX-XOR-SAT. In particular, at least $2^{n/3}$ queries of tolerance $2^{-n/3}$ are required.*

For a vector $c \in \{0,1\}^n$ we define the parity function $\chi_c(x)$ as usual $\chi_c(x) \doteq -(-1)^{c \cdot x}$. Further, let D_c to be the uniform distribution over the set $S_c = \{x \mid \chi_c(x) = 1\}$.

LEMMA 2. *For $c \in \{0,1\}^n$, $c \neq \bar{0}$ and the uniform distribution U over $\{-1,1\}^n$, the following hold:*

1. $\mathbf{E}_{x \sim D_c}[\chi_{c'}(x)] = \begin{cases} 1 \text{ if } c = c' \\ 0 \text{ otherwise.} \end{cases}$
2. $\mathbf{E}_{x \sim U}[\chi_c(x)\chi_{c'}(x)] = \begin{cases} 1 \text{ if } c = c' \\ 0 \text{ otherwise.} \end{cases}$

These two facts will imply that when $D = U$ (the uniform distribution) and the D_i's consist of the D_c's, we can set $\gamma = 0$ and $\beta = 1$ in our simpler variant of statistical dimension from which Theorem 9 follows.

THEOREM 10. *For the MAX-XOR-SAT problem, let $\mathcal{F} = \{\chi_x\}_{x \in \{0,1\}^n}$, let $\mathcal{D}$ be the set of all distributions over clauses $c \in \{0,1\}^n$, and for any $\delta > 0$, let $\mathcal{Z}$ be the $(\frac{1}{2} - \delta)$-approximate MAX-XOR-SAT problem defined over $\mathcal{F}$ and $\mathcal{D}$. Then $\mathrm{SD}(\mathcal{Z}, 0, 1) \geq 2^n - 1$.*

5. PLANTED CLIQUE

We now prove the lower bound claimed in Theorem 2 on the problem of detecting a planted k-clique in the given distribution on vectors from $\{0,1\}^n$ as defined above.

For a subset $S \subseteq [n]$, let D_S be the distribution over $\{0,1\}^n$ with a planted clique on the subset S. Let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k. For $i \in [m]$ we use D_i to denote D_{S_i}. The reference distribution in our lower bounds will be the uniform distribution over $\{0,1\}^n$ and let $\hat{D}_S$ denote $D_S/D - 1$. In order to apply our lower bounds based on statistical dimension with average correlation we now prove that for the planted clique problem average correlations of large sets must be small. We start with a lemma that bounds the correlation of two planted clique distributions relative to the reference distribution D as a function of the overlap between the cliques:

LEMMA 3. *For $i, j \in [m]$,*

$$\left\langle \hat{D}_i, \hat{D}_j \right\rangle_D \leq \frac{2^\lambda k^2}{n^2},$$

where $\lambda = |S_i \cap S_j|$.

PROOF. For the distribution D_i, we consider the probability $D_i(x)$ of generating the vector x. Then,

$$D_i(x) = \begin{cases} (\frac{n-k}{n})\frac{1}{2^n} + (\frac{k}{n})\frac{1}{2^{n-k}} & \text{if } \forall s \in S_i, x_s = 1 \\ (\frac{n-k}{n})\frac{1}{2^n} & \text{otherwise.} \end{cases}$$

We can now compute $\hat{D}_i$ and the inner product:

$$\begin{aligned} \left\langle \hat{D}_i, \hat{D}_j \right\rangle_D &\leq \frac{2^{n-2k+\lambda}}{2^n} \left(\frac{k2^k}{n} - \frac{k}{n}\right)^2 \\ &\quad + 2\left(\frac{2^{n-k}}{2^n}\right)\left(\frac{k2^k}{n} - \frac{k}{n}\right)\left(-\frac{k}{n}\right) + \left(-\frac{k}{n}\right)^2 \\ &\leq \frac{2^\lambda k^2}{n^2} \end{aligned}$$

□

We now give a bound on the average correlation of any $\hat{D}_S$ with a large number of distinct clique distributions.

LEMMA 4. *For $\delta > 0$ and $k \leq n^{1/2-\delta}$, let $\{S_1, \ldots, S_m\}$ be the set of all $\binom{n}{k}$ subsets of $[n]$ of size k and $\{D_1, \ldots, D_m\}$ be the corresponding distributions on $\{0,1\}^n$. Then for any integer $\ell \leq k$, set S of size k and subset $A \subseteq \{S_1, \ldots, S_m\}$ where $|A| \geq 4(m-1)/n^{2\ell\delta}$,*

$$\frac{1}{|A|} \sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle < 2^{\ell+2} \frac{k^2}{n^2}.$$

PROOF. In this proof we first show that if the total number of sets in A is large then most of sets in A have a small overlap with S. We then use the bound on the overlap of most sets to obtain a bound on the average correlation of D_S with distributions for sets in A.

Formally, we let $\alpha = \frac{k^2}{n^2}$ and using Lemma 3 get the bound $\langle \hat{D}_i, \hat{D}_j \rangle \leq 2^{|S_i \cap S_j|}\alpha$. Summing over $S_i \in A$,

$$\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle \leq \sum_{S_i \in A} 2^{|S \cap S_i|}\alpha.$$

For any set $A \subseteq \{S_1, \ldots, S_m\}$ of size t this bound is maximized when the sets of A include S, then all sets that intersect S in $k-1$ indices, then all sets that intersect S in $k-2$ indices and so on until the size bound t is exhausted. We can therefore assume without loss of generality that A is defined in precisely this way.

Let $T_\lambda = \{S_i \mid |S \cap S_i| = \lambda\}$ denote the subset of all k-subsets that intersect with S in exactly λ indices. Let λ_0 be the smallest λ for which $A \cap T_\lambda$ is non-empty. We first observe that for any $1 \leq j \leq k-1$,

$$\begin{aligned}\frac{|T_j|}{|T_{j+1}|} &= \frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{k}{j+1}\binom{n-k}{k-j-1}} \\ &= \frac{(j+1)(n-2k+j+1)}{(k-j-1)(k-j)} \\ &\geq \frac{(j+1)(n-2k)}{k(k+1)} \\ &\geq \frac{(j+1)n^{2\delta}}{2}.\end{aligned}$$

By applying this equation inductively we obtain,

$$|T_j| \leq \frac{2^j \cdot |T_0|}{j! \cdot n^{2\delta j}} < \frac{2^j \cdot (m-1)}{j! \cdot n^{2\delta j}}$$

and

$$\sum_{k \geq \lambda \geq j} |T_\lambda| < \sum_{k \geq \lambda \geq j} \frac{2^\lambda \cdot (m-1)}{\lambda! \cdot n^{2\delta\lambda}} \leq \frac{4(m-1)}{n^{\delta j}} \ .$$

By definition of λ_0, $|A| \leq \sum_{j \geq \lambda_0} |T_j| < 4(m-1)/n^{2\delta\lambda_0}$. In particular, if $|A| \geq 4(m-1)/n^{2\ell\delta}$ then $n^{2\ell\delta}/4 < n^{2\delta\lambda_0}/4$ or $\lambda_0 < \ell$. Now we can conclude that

$$\begin{aligned}\sum_{S_i \in A} \langle \hat{D}_S, \hat{D}_i \rangle &\leq \sum_{j=\lambda_0}^{k} 2^j |T_j \cap A|\alpha \\ &\leq \left(2^{\lambda_0}|T_{\lambda_0} \cap A| + \sum_{j=\lambda_0+1}^{k} 2^j |T_j|\right)\alpha \\ &\leq \left(2^{\lambda_0}|T_{\lambda_0} \cap A| + 2 \cdot 2^{\lambda_0+1}|T_{\lambda_0+1}|\right)\alpha \\ &\leq 2^{\lambda_0+2}|A|\alpha < 2^{\ell+2}|A|\alpha.\end{aligned}$$

To derive the last inequality we need to note that for every $j \geq 0$, $2^j|T_j| > 2(2^{j+1}|T_{j+1}|)$ we can therefore telescope the sum. □

Lemma 4 gives a simple way to bound the SDA.

THEOREM 11. *For $\delta > 0$ and $k \leq n^{1/2-\delta}$ let $\mathcal{Z}$ the planted bipartite k-clique problem. Then for any $\ell \leq k$,*

$$\mathrm{SDA}\left(\mathcal{Z}, 2^{\ell+2}k^2/n^2, 1/\binom{n}{k}\right) \geq n^{2\ell\delta}/4.$$

Combining Theorems 5 and 11 gives Theorem 2 by taking $\ell = \log(r)$. Theorems 7 and 11 imply the sample complexity lower bound stated in Theorem 3 by taking $\ell \geq 4/\delta$.

Analogously, for the generalized planted densest subgraph case, we can give a bound on statistical dimension SDA:

THEOREM 12. *Fix $0 < q \leq p \leq 1$. For $\delta > 0$ and $k \leq n^{1/2-\delta}$ let $\mathcal{Z}$ be the generalized planted bipartite densest subgraph problem. Then for any $\ell \leq k$,*

$$\mathrm{SDA}\left(\mathcal{Z}, \frac{2k^2}{n^2}\left(\left(1 + \frac{(p-q)^2}{q(1-q)}\right)^{\ell+1} - 1\right), \frac{1}{\binom{n}{k}}\right) \geq n^{2\ell\delta}/4.$$

provided $n^{2\delta} \geq 1 + (p-q)^2/q(1-q)$.

Acknowledgments

We thank Benny Applebaum, Avrim Blum, Uri Feige, Ravi Kannan, Michael Kearns, Robi Krauthgamer, Moni Naor, Jan Vondrak, and Avi Wigderson for insightful comments and helpful discussions.

6. REFERENCES

[1] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise independence. In *Proceedings of STOC*, pages 496–505, 2007.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. In *SODA*, pages 594–598, 1998.

[3] B. P. W. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, 129(1):69–89, 2011.

[4] B. Applebaum, B. Barak, and A. Wigderson. Public-key cryptography from different assumptions. In *STOC*, pages 171–180, 2010.

[5] S. Arora, B. Barak, M. Brunnermeier, and R. Ge. Computational complexity and information asymmetry in financial products (extended abstract). In *ICS*, pages 49–65, 2010.

[6] A. Belloni, R. M. Freund, and S. Vempala. An efficient rescaled perceptron algorithm for conic systems. *Math. Oper. Res.*, 34(3):621–641, 2009.

[7] A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan. Detecting high log-densities: an $o(n^{1/4})$ approximation for densest k-subgraph. In *STOC*, pages 201–210, 2010.

[8] A. Bhaskara, M. Charikar, A. Vijayaraghavan, V. Guruswami, and Y. Zhou. Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph. In *SODA*, pages 388–405, 2012.

[9] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proceedings of PODS*, pages 128–138, 2005.

[10] A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[11] A. Blum, M. L. Furst, J. C. Jackson, M. J. Kearns, Y. Mansour, and S. Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *STOC*, pages 253–262, 1994.

[12] S. Brubaker and S. Vempala. Random tensors and planted cliques. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 5687, pages 406–419. 2009.

[13] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of NIPS*, pages 281–288, 2006.

[14] A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability & Computing*, 19(2):227–284, 2010.

[15] Y. Dekel, O. Gurel-Gurevich, and Y. Peres. Finding hidden cliques in linear time with high probability. In *Proceedings of ANALCO*, pages 67–75, 2011.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[17] J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Math. Program.*, 114(1):101–114, 2008.

[18] U. Feige. Relations between average case complexity and approximation complexity. In *IEEE Conference on Computational Complexity*, page 5, 2002.

[19] U. Feige and R. Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Struct. Algorithms*, 16(2):195–208, 2000.

[20] U. Feige and R. Krauthgamer. The probable value of the Lovász–Schrijver relaxations for maximum independent set. *SICOMP*, 32(2):345–370, 2003.

[21] U. Feige and D. Ron. Finding hidden cliques in linear time. In *Proceedings of AofA*, pages 189–204, 2010.

[22] V. Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer System Sciences*, 78(5):1444–1459, 2012.

[23] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *CoRR*, abs/1201.1214, 2012.

[24] A. M. Frieze and R. Kannan. A new approach to the planted clique problem. In *FSTTCS*, pages 187–198, 2008.

[25] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[26] J. Håstad. Some optimal inapproximability results. *J. ACM*, 48:798–859, July 2001.

[27] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[28] E. Hazan and R. Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, 40(1):79–91, 2011.

[29] J. Jackson. On the efficiency of noise-tolerant PAC algorithms derived from statistical queries. *Annals of Mathematics and Artificial Intelligence*, 39(3):291–313, Nov. 2003.

[30] M. Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 3(4):347–360, 1992.

[31] A. Juels and M. Peinado. Hiding cliques for cryptographic security. *Des. Codes Cryptography*, 20(3):269–280, 2000.

[32] R. Kannan. personal communication.

[33] R. Karp. Probabilistic analysis of graph-theoretic algorithms. In *Proceedings of Computer Science and Statistics 12th Annual Symposium on the Interface*, page 173, 1979.

[34] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[35] S. Khot. Ruling out ptas for graph min-bisection, densest subgraph and bipartite clique. In *FOCS*, pages 136–145, 2004.

[36] S. Kirkpatrick, D. G. Jr., and M. P. Vecchi. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

[37] L. Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2-3):193–212, 1995.

[38] F. McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.

[39] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[40] L. Minder and D. Vilenchik. Small clique detection and approximate nash equilibria. 5687:673–685, 2009.

[41] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*, 50(302):157–175, 1900.

[42] B. Selman, H. Kautz, and B. Cohen. Local search strategies for satisfiability testing. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 521–532, 1995.

[43] R. Servedio. Computational sample complexity and attribute-efficient learning. *Journal of Computer and System Sciences*, 60(1):161–178, 2000.

[44] B. Szörényi. Characterizing statistical query learning: Simplified notions and proofs. In *ALT*, pages 186–200, 2009.

[45] M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

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

[47] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

[48] V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, Jan. 1985.

[49] K. Yang. On learning correlated boolean functions using statistical queries. In *Proceedings of ALT*, pages 59–76, 2001.

[50] K. Yang. New lower bounds for statistical query learning. *J. Comput. Syst. Sci.*, 70(4):485–509, 2005.

[51] A. Yao. Probabilistic computations: Toward a unified measure of complexity. In *FOCS*, pages 222–227, 1977.