Improved Hardness Results for Learning Intersections of Halfspaces

Stefan Tiegel STEFAN.TIEGEL@INF.ETHZ.CH
Andreasstrasse 5, 8092 Zürich

Editors: Shipra Agrawal and Aaron Roth

Abstract

We show strong (and surprisingly simple) lower bounds for weakly learning intersections of halfspaces in the improper setting. Strikingly little is known about this problem. For instance, it is not even known if there is a polynomial-time algorithm for learning the intersection of only two halfspaces. On the other hand, lower bounds based on well-established assumptions (such as approximating worst-case lattice problems or variants of Feige's 3SAT hypothesis) are only known (or are implied by existing results) for the intersection of super-logarithmically many halfspaces Klivans and Sherstov (2009, 2006); Daniely and Shalev-Shwartz (2016). With intersections of fewer halfspaces being only ruled out under less standard assumptions Daniely and Vardi (2021) (such as the existence of local pseudo-random generators with large stretch). We significantly narrow this gap by showing that even learning $\omega(\log \log N)$ halfspaces in dimension N takes super-polynomial time under standard assumptions on worst-case lattice problems (namely that SVP and SIVP are hard to approximate within polynomial factors). Further, we give unconditional hardness results in the statistical query framework. Specifically, we show that for any k (even constant), learning k halfspaces in dimension N requires accuracy $N^{-\Omega(k)}$, or exponentially many queries – in particular ruling out SQ algorithms with polynomial accuracy for $\omega(1)$ halfspaces. To the best of our knowledge this is the first unconditional hardness result for learning a super-constant number of halfspaces.

Our lower bounds are obtained in a unified way via a novel connection we make between intersections of halfspaces and the so-called parallel pancakes distribution Diakonikolas et al. (2017b); Bubeck et al. (2019); Bruna et al. (2021) that has been at the heart of many lower bound constructions in (robust) high-dimensional statistics in the past few years.

Keywords: intersections of halfspaces, weak learning, cryptographic hardness

1. Introduction

This work studies the computational complexity of weakly learning intersections of halfspaces in the PAC model Valiant (1984). A halfspace $h_w \colon \mathbb{R}^N \to \{\pm 1\}$, or linear threshold function (short LTF), is a function $x \mapsto \operatorname{sign}(\langle x, w \rangle)$ for some unit vector $w \in \mathbb{R}^N$. A fundamental question is to what extent we can predict the output of h_w on a fresh example, when given random example-label-pairs $(x, h_w(x))$ (where x can follow an arbitrary distribution). This problem is very well-understood and known to be solvable in time polynomial in the dimension and the inverse of the desired accuracy Maass and Turán (1994). On the other hand, surprisingly little is known when considering only slightly more complex functions such as a function of a small number of halfspaces. This holds true, even if the functions are simple functions, such as the AND function. Note that the AND function of several halfspaces corresponds to their intersection since for (x, y) it holds that $y = 1$ if and only if x is classified as positive by all halfspaces.

This class is particularly appealing since, depending on the number of halfspaces, it interpolates naturally between very simple (a single halfspace) and very complex boolean functions (such as polytopes with many facets). Studying the performance of efficient algorithms in this setting, parametrized by the number of halfspaces, can thus serve as a benchmark of how complex functions we could hope to learn. Formally, the problem is defined as follows:

Definition 1 *Let $k, N \in \mathbb{N}$. A distribution D over $\mathbb{R}^N \times \{\pm 1\}$ is an* intersection of k halfspaces*, if it can be described as follows: Let $w_1, \dots, w_k \in \mathbb{R}^N$ be unit vectors and for $i \in [k]$ let $h_{w_i} = \operatorname{sign}(\langle w_i, x \rangle)$. Let D_x be an arbitrary distribution over $\mathbb{R}^N$. A sample (x, y) from D is produced by first drawing $x \sim D_x$ and then setting $y = 1$ if and only if $h_{w_i}(x) = 1$ for all i.*

For brevity, we will sometimes write ”learning k halfspaces” when we mean ”learning the intersection of k halfspaces”.

We measure the performance of an algorithm as follows: For any function $f \colon \mathbb{R}^N \to \{\pm 1\}$, we define the misclassification error with respect to a distribution D over $\mathbb{R}^N \times \{\pm 1\}$ as $\operatorname{err}_D(f) :=$ $\mathbb{P}_{(x,y)\sim D}(f(x) \neq y)$. We say an algorithm *weakly* learns D, if given i.i.d. samples from D, it outputs a function $\hat{f}$ such that $\operatorname{err}_D(\hat{f}) \leqslant \frac{1}{2} - \frac{1}{\operatorname{poly}(N)}$, for some polynomial. Intuitively, this means the algorithm does slightly better than randomly guessing the label y. This paper studies to what extent we can hope to weakly learn the intersection of few (with respect to the dimension) halfspaces.

We remark that we do not restrict our algorithm to output an intersection of k (or more) halfspaces, but allow that it returns an arbitrary boolean function. This setting is called *improper learning*. Whereas the setting in which the hypothesis needs to be of the same (or a slightly larger) family, is referred to as (semi-)proper learning. Proving lower bounds against improper learners has proven to be significantly more difficult than against proper learners. In particular, while it is known how to show NP-hardness (under randomized reductions) of properly learning many natural classes of functions Feldman (2006); Feldman et al. (2006); Guruswami and Raghavendra (2006); Gopalan et al. (2010), there are inherent barriers for showing such reductions in the improper setting Applebaum et al. (2008). In fact, improper learners are known to be strictly more powerful. For instance, there are concept classes for which it is known that it is NP-hard to find a proper learner, but efficient improper learners exist Valiant (1984); Pitt and Valiant (1988).[1]

Previous hardness results Indeed, in the proper setting it is known that it is NP-hard to learn the intersection of two halfspaces, even if the learner is allowed to output a function that is an intersection of any constant number of halfspaces Alekhnovich et al. (2004). Whereas in the improper setting, despite extensive work on this topic Klivans et al. (2004a,b, 2008, 2009); Vempala (2010); Sherstov (2010, 2021), it is not even known whether there are polynomial-time algorithms for (improperly) learning the intersection of two halfspaces unless we make additional assumption about the marginal distribution D_x. Nor is there any evidence of hardness[2].

Due to the dearth of algorithmic results, researchers have started to look for evidence of hardness. Most of these are reduction-based, while a few are unconditional but restricted to the statistical query (SQ) model. The first result being the seminal work of Klivans and Sherstov (2009) showing that for any $\varepsilon > 0$, weakly learning N^ε halfspaces[3] is not possible in polynomial time, assuming hardness of certain worst-case lattice problems that form the basis of a large branch of cryptography

[1] The class being 3-Term DNFs that are known to be efficiently learnable via 3-CNFs

[2] Except for some structural observations Sherstov (2010, 2021). We will come back to this later.

[3] This was later strengthened to $\log^C(N)$ for some constant $C > 2$ Klivans and Sherstov (2006). See section 1.1 for a more detailed discussion.

(specifically, approximating SVP and SIVP up to polynomial factors, see theorems 2 and 3 and the end of this section for precise definitions and a discussion, we also refer to Peikert et al. (2016)). This was slightly strengthened in Daniely and Shalev-Shwartz (2016) to showing that learning $\omega(\log N)$ halfspaces is hard assuming a widely believed variant of Feige's hypothesis about refuting random 3SAT instances Feige (2002).

Going beyond this, researchers had to resort to less standard assumptions to show reduction-based hardness of even fewer halfspaces. In particular, Daniely and Vardi (2021) showed that assuming the existence of so-called *local* pseudo-random generators with polynomial stretch, learning even $\omega(1)$ halfspaces is hard – assuming that a specific candidate function actually satisfies these properties, they are able to show that learning k halfspaces takes time at least $n^{\Omega(k)}$. While this indeed gives some evidence of hardness, we believe verifying these predictions based on more standard assumptions or via unconditional lower bounds in restricted model of computation is an important line of work. Yet, proving such strong, or even fine-grained results, under more standard assumptions, such as approximating worst-case lattice problems or (variants of) Feige's hypothesis, has remained elusive. In our work, we make significant progress in this direction, by showing that learning even $\omega(\log \log N)$ halfspaces is hard under standard assumptions about approximating SVP and SIVP similar to Klivans and Sherstov (2009).

In terms of unconditional lower bounds, Klivans and Sherstov (2007) showed that (roughly speaking), restricted to the SQ model, learning k halfspaces takes time at least $N^{\Omega(k/\log \log N)}$, ruling out efficient SQ algorithms learning intersections of $\omega(\log \log N)$ halfspaces. As a by-product of our results, we will give an improved SQ lower bound (via a different hard instance than Klivans and Sherstov (2007)), showing that learning k halfspaces needs precision at least $N^{-\Omega(k)}$. Note that this rules out efficient SQ algorithms for learning $\omega(1)$ halfspaces, but also gives a fine-grained hardness result for learning $k = O(1)$ halfspaces.

Hardness assumption, SQ model and main results We will next state the precise hardness assumption we make. We remark that we do not expect the reader to be familiar with lattices or these problems and such familiarity is not necessary in order to understand and appreciate the remainder of this paper. Our reductions will start from a different learning problem that can be stated in elementary terms (see section 2). For more background on lattices and these problems, we refer to Peikert et al. (2016). An n-dimensional *lattice* L is defined to be a discrete additive subgroup of $\mathbb{R}^n$. It can be fully specified by a basis $B \in \mathbb{R}^{n \times n}$ as $L = B\mathbb{Z}^n$. We will only consider the case in which B is full-rank. For $1 \leqslant i \leqslant n$, consider

$$\lambda_i(L) \coloneqq \inf\left\{r > 0 \mid \dim\left(\operatorname{Span}\left(L \cap B_r(0)\right) \geqslant i\right)\right\} .$$

We can now define GapSVP and SIVP.

Problem 2 (Gap Shortest Vector Problem (GapSVP)) *Let $\alpha = \operatorname{poly}(n)$ be arbitrary. Given an n-dimensional lattice L and $d > 0$ such that either (a) $\lambda_1(L) \leqslant d$ or (b) $\lambda_1(L) > \alpha \cdot d$, decide whether (a) or (b) holds.*

Problem 3 (Shortest Independent Vector Problem (SIVP)) *Let $\alpha = \operatorname{poly}(n)$ be arbitrary. Given an n-dimensional lattice L output a set of linearly independent lattice points of length at most $\alpha \cdot \lambda_n(L)$.*

We make the following assumption

Assumption 4 *There is no quantum algorithm that runs in time $2^{o(n)}$ and uses only $2^{o(n)}$ samples that solves either problem 2 or problem 3.*

All known (quantum) algorithms for problem 2 and problem 3 require time $2^{\Omega(n)}$. Further, a falsification of the above assumption would be considered a major breakthrough in cryptography (cf. Peikert et al. (2016) and references therein for more context).

Similarly, we give some necessary background on the SQ model. In particular, SQ algorithms only have access to the distributions via query functions $\phi\colon \mathbb{R}^N \times \{\pm 1\} \to [-1, 1]$. Upon making a query ϕ, they receive as an answer a value in $[\mathbb{E}_D\, \phi(x, y) - \tau, \mathbb{E}_D\, \phi(x, y) + \tau]$. τ is called the *accuracy* or *precision* of the query. The query function can be arbitrary and outside of making these queries, the algorithms can perform arbitrary computation. When comparing to sample-based algorithms, typically the number of queries is taken as a proxy for run-time and $1/\tau^2$ as a proxy for the number of samples – since this many samples are needed to estimate the expectation of a query from samples up to accuracy τ.

Our reduction-based hardness result is as follows

Theorem 5 (See theorem 12 for full version) *Let $N, k \in \mathbb{N}$ such that $k \leqslant O(\sqrt{N})$. Under assumption 4, there is no $T = N^{o(\frac{k}{\log k + \log\log N})}$-time algorithm using $O(T)$ samples that learns the intersection of k halfspaces up to error better than $\frac{1}{2} - \frac{1}{\Omega(T)}$.*

It is insightful to explicitly compute the time lower bound for specific values of k. First, note that this rules out polynomial-time algorithms for weakly learning $\omega(\log\log N)$ halfspaces. A few other examples are as follows: For any $0 < \varepsilon \leqslant \frac{1}{2}$, not necessarily constant, learning $k = N^\varepsilon$ halfspaces takes time at least $\exp(\Omega(N^\varepsilon \cdot \frac{\log N}{\varepsilon \log N + \log\log N}))$ In particular, taking ε to be an absolute constant, we obtain that learning N^ε halfspaces takes time at least $\exp(\Omega(N^\varepsilon))$. Taking $\varepsilon = \frac{\log\log N}{\log N}$, we obtain that learning $\log N$ halfspaces takes time $\exp(\Omega(\frac{\log^2 N}{\log\log N}))$. Finally, for $\varepsilon = \omega(\frac{\log\log\log N}{\log N})$, we recover that learning $k = \omega(\log\log N)$ halfspaces takes time at least $\exp(\omega(\log N)) = N^{\omega(1)}$. Finally, under the more conservative assumption that there is no algorithm for theorems 2 and 3 running in time $2^{\Omega(n^{1-\delta})}$ for any constant $\delta > 0$, we are still able to rule out weakly learning $\omega(\log^\delta(N))$ halfspaces. See section 1.1 to a more detailed comparison with prior work.

Our SQ hardness results is as follows:

Theorem 6 *Let $k, N \in \mathbb{N}$ such that $k \leqslant N^\gamma$ for a sufficiently small absolute constant γ. Any SQ algorithms using queries of accuracy $\tau = N^{-\Omega(k)}$ that learns the intersection of k halfspaces over $\mathbb{R}^N$ up to error better than $\frac{1}{2} - 4\tau$ must make at least $2^{N^{\Omega(1)}}$ queries.*

Note that this shows that even weakly learning $\omega(1)$ halfspaces requires super-polynomial precision in the SQ model or exponentially many queries. Similarly, it shows that the fine-grained complexity of learning k halfspaces scales as $N^{\Omega(k)}$. We remark that we prioritized clarity and did not attempt to optimize any constants, neither in the condition that $k \leqslant N^\gamma$ nor in the exponent of the accuracy or the number of queries.

Future work We remark that both our lower bound instance can be solved in time $N^{O(k)}$ since they can be represented as a degree-$O(k)$ polynomial threshold function (see section 2 for all details) – and thus can be learned in time $N^{O(k)}$ via linear programming Maass and Turán (1994). This suggests that we should look for instances that cannot be represented as low-degree polynomial threshold functions. This approach seems particularly motivated since it is known that, at least when the input comes from the boolean hypercube, there exists an intersection of even 2 halfspaces that cannot be represented by degree-$o(n)$ polynomial threshold functions Sherstov (2010, 2021).

1.1. More on Previous Results

We elaborate a bit more on the connection between our work and previous hardness results below.

The work closest to us is Klivans and Sherstov (2009) (and the companion work Klivans and Sherstov (2006)). Their hardness results are ultimately also based on the hardness of theorems 2 and 3. However, their hardness result follows by showing that intersections of halfspaces can encode a public-key encryption system due to Regev Regev (2009) known to be secure assuming hardness of these lattice problems. Thus, a learning algorithm could break the crypto-system and hence falsify theorems 2 and 3. While we start from the same assumptions, we give a more direct reduction, completely bypassing the need to introduce any public-key encryption schemes. This more direct reduction is what enables our improved SQ lower bounds.

On a quantitative level, Klivans and Sherstov (2009) shows that for any absolute constant $\varepsilon > 0$ a $\mathrm{poly}(N)$-time algorithm for learning $k = N^\varepsilon$ halfspaces in dimension N would yield a $\mathrm{poly}(n)$-time algorithm for theorems 2 and 3 in dimension n. In particular, their results are implied by a weaker version of theorem 4 in which we only assume that there is no $\mathrm{poly}(n)$-time algorithm for theorems 2 and 3[4]. In Klivans and Sherstov (2006), the same authors observed that their reduction implies stronger lower bounds under quantitatively stronger assumptions on theorems 2 and 3 (closer to our theorem 4). Pushed to the limit, their result yields that theorem 4 implies that learning $\omega(\log N)$ halfspaces in dimension N takes super-polynomial time, matching the result of Daniely and Shalev-Shwartz (2016) under a different assumption – we remark that this is not formally stated in Klivans and Sherstov (2006) but follows immediately from their techniques. In particular, allowing ε to be sub-constant, their techniques can be used to show that theorem 4 implies that learning N^ε halfspaces takes time $\exp(\Omega(N^\varepsilon))$ (see the discussion at the end of section 2 for a more detailed argument and technical comparison to our work). This should be compared with our lower bound $\exp(\Omega(N^\varepsilon \frac{\log N}{\log\log N}))$ for $\varepsilon \leqslant \frac{\log\log N}{\log N}$ (and similar for larger ε). The latter is significantly larger and in particular allows to obtain hardness results of exponentially fewer halfspaces ($\omega(\log\log N)$).

We strongly believe that our techniques also allow for a trade-off of the following form: To rule out polynomial-time algorithms for learning more halfspaces, but under quantitatively weaker assumptions. We choose not to make this explicit for clarity of exposition and since already a $2^{o(n)}$-time algorithm for either of theorems 2 and 3 would be a major breakthrough.

Along a different direction, Tiegel (2023); Diakonikolas et al. (2022, 2023) show lower bounds for learning a single halfspaces in various error models under assumption 4.

2. Technical Overview

Relation to parallel pancakes and SQ lower bound Our lower bounds are based on a novel connection we make between the so-called ”parallel pancakes” distribution Diakonikolas et al. (2017a); Bubeck et al. (2019); Bruna et al. (2021) and intersections of halfspaces. On a high level, the former

[4] More specifically, they show that is true even when setting $\alpha = \tilde{O}(n^{1.5})$. We strongly believe that this is also true for our reduction, but did not attempt to make this explicit for clarity. theorems 2 and 3 are believed to be hard for any $\alpha = \mathrm{poly}(n)$.

is a mixture of few Gaussians, that is hard to distinguish from the standard Gaussian distribution. It (or versions thereof) has played a pivotal role in obtaining computational hardness results for learning theory problems. Yet, the connection to intersections of halfspaces had not been observed before. Similar ideas, without any reference to parallel pancakes, were implicitly used in Klivans and Sherstov (2009). By making this connection explicit and expanding on it, we are able to obtain improved lower bounds in both the SQ model and under assumption 4. More specifically, our connection allows us to leverage that (variants of this) distribution are known to be hard to learn in the SQ model and based on assumption 4. Fleshing out all details and satisfying all distribution requirements exactly will take some additional work.

We start by describing one version of the parallel pancakes distribution and showing our SQ lower bound (theorem 6). Unfortunately, this connection alone is not enough to establish our reduction-based result (theorem 5) as well. The reason being that the known hardness results for parallel pancakes under assumption 4 are quantitatively weaker than those known under SQ – and in particular would by themselves only rule out efficient algorithms for learning $\omega(\log N)$ halfspaces. Towards the end of this section, we will show how to show a hardness result for $\omega(\log \log N)$ halfspaces using a modified construction.

It is known Bubeck et al. (2019) (see also Diakonikolas et al. (2017a)) that there are two one-dimensional distributions A, B satisfying the following properties (see appendix A for all details):

1. A and B are mixtures of k Gaussians.

2. There exists two unions of k disjoint intervals S_A and S_B, such that only a negligible fraction of the probability mass of A (resp. B) lies outside S_A (resp. S_B).

3. The intervals in $S_A \cup S_B$.

4. Both A and B match k moments with $N(0, 1)$.

Consider now the following distribution $D_{A,B}$ over $\mathbb{R}^N \times \{\pm 1\}$: First, pick a uniformly random unit vector w, and let D_A (resp. D_B) be the distribution over $\mathbb{R}^N$ that is A (resp. B) along w and a standard Gaussian in the complement. Then, set $D_{A,B} = \frac{1}{2}(D_A, +1) + \frac{1}{2}(D_B, -1)$. Using results from Bubeck et al. (2019); Diakonikolas and Kane (2022) it is not hard to deduce that $D_{A,B}$ is hard to distinguish from $N(0, \mathrm{Id}_N) \times \mathrm{Be}(\frac{1}{2})$ in the SQ model. In particular, this task either requires queries of accuracy better than $N^{-\Omega(k)}$ (suggesting that we need at least $N^{\Omega(k)}$ samples) or $2^{N^{\Omega(1)}}$ queries. (We give a full argument for our variant of this distribution in appendix A.) Reviewer 2 asked for an explanation in words what the distribution $D_{A,B}$ (cf. the first paragraph on Page 6, also see Figure 1) in our lower bound instance is. We offer some intuition on $D_{A,B}$: This distribution corresponds to a mixture of two related instances of a labelled version of the parallel pancakes distribution alluded to earlier. In particular, for $(x, y) \sim D_{A,B}$, if $y = +1$, x follows the standard parallel pancakes distribution and if $y = -1$, x follows a "shifted" parallel pancakes distribution in which the pancakes are shifted along the hidden direction such that they are (mostly) disjoint from the pancakes for $y = +1$. See figure 1 for an illustration.

Modifying the instance to obtain an intersection of degree-2 PTFs Our SQ lower bound follows from the simple but powerful observation that a slight variant of this distribution can be realized as an intersection of k degree-2 polynomial threshold functions (short PTFs). Note that this is enough to show our hardness result. Indeed, recall that we aim to show that learning k halfspaces

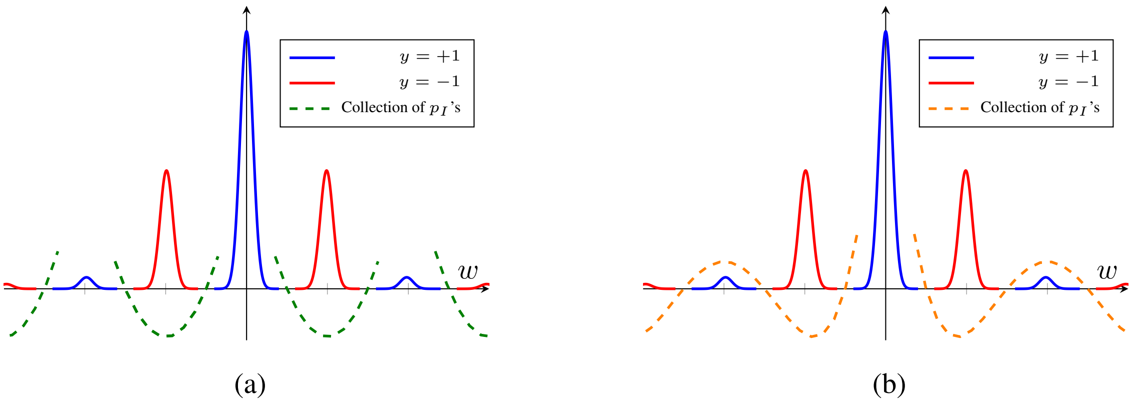

Figure 1: (a) shows how to capture the parallel pancakes distribution using degree-2 PTFs and (b) shows how to do the same using higher-degree PTFs (degree-4 in this case).

in dimension N takes time $N^{\Omega(k)}$. For this it is sufficient to show that learning the intersection of k degree-2 PTFs in dimension N takes time at least $N^{\Omega(k)}$ since we can represent these as intersections of k halfspaces over an $O(N^2)$-dimensional space. We can absorb the quadratic blow-up in the dimension in the $\Omega(\cdot)$-notation.

Note that a priori $D_{A,B}$ cannot be realized as such an intersection: Since the density of both A and B are positive on all of $\mathbb{R}$, there exists a region in which the label can be both + and -1 with some small probability. Since our model is noiseless, this should not be possible. Fortunately, these regions only make up a small fraction of the total probability mass and we can get rid of them by truncating the mixture components. Indeed, let $\tilde{A}, \tilde{B}$ be obtained by conditioning A (resp. B) to lie in S_A (resp. S_B) and let $D_{\tilde{A},\tilde{B}}$ be obtained analogously as before (replacing A and B by $\tilde{A}$ and $\tilde{B}$). In appendix A we show that $D_{\tilde{A},\tilde{B}}$ enjoys the same hardness guarantees in the SQ model as $D_{A,B}$, i.e, that this distribution is still hard to distinguish from $N(0, \mathrm{Id}_N) \times \mathrm{Be}(\frac{1}{2})$ in the relevant parameter regime. This follows by showing that the first k moments of both $\tilde{A}$ and $\tilde{B}$ still match those of $N(0, 1)$ up to small error ($N^{-\Omega(k)}$) and their χ^2-divergence with $N(0, 1)$ is not too large ($2^{O(k)} \log N$) – this uses results based on Diakonikolas and Kane (2022).

To see that $D_{\tilde{A},\tilde{B}}$ is an intersection of k degree-2 PTFs, note the following: By construction, for a sample $(x, y) \sim D_{\tilde{A},\tilde{B}}$, $y = 1$ if and only if $\langle x, w\rangle \in S_A$. Further $y = -1$ if and only if $\langle x, w\rangle \in S_B$. Thus, for every interval $I \subseteq S_B$, consider the polynomial $p_I : \mathbb{R} \to \mathbb{R}$ that is symmetric around the mid-point of B, is negative on I, and has its roots at half the distance between the end of I and the next interval in S_A. Note that p_I is negative on I and positive on all other intervals in both S_A and S_B. The final choice of degree-2 PTFs is then $\tilde{p}_I : \mathbb{R}^N \to \mathbb{R}$ such that $\tilde{p}_I(x) = p_I(\langle x, w\rangle)$. By construction, if $\langle x, w\rangle \in S_A$, $\tilde{p}_I(x) \geqslant 0$ for all I and if $\langle x, w\rangle \in S_B$ there exists $\tilde{p}_I$ such that $\tilde{p}_I(x) < 0$. It follows that D corresponds to the intersection of the $\tilde{p}_I$. Since S_B contains k intervals, this yields the claim. See Figure 1 (a) for an illustration.

To solve the distinguishing problem, we can run our weak learner on our input distribution and with one additional query compute the misclassification error of the produced hypothesis. Since in the null case the label y is independent of x, this should be 1/2. While it should be bounded away from $1/2$ under planted by assumption on our weak learner. We can thus solve the distinguishing problem.

Lower bound based on assumption 4 ”Parallel Pancakes”-type distribution are also known to be hard to distinguish from a standard Gaussian under assumption 4. In particular, using results from Bruna et al. (2021); Gupte et al. (2022) one could show that a similar distribution, that also has k components, takes time roughly at least $2^{\Omega(k)}$ to distinguish from a standard Gaussian. Unfortunately, using this, we could only hope to rule out learning intersections of $\omega(\log N)$ halfspaces, which is exponentially worse than $\omega(\log \log N)$. In order to obtain our improved lower bound, we make use of the following observation: Instead of showing that intersections of degree-2 PTFs are hard to learn, we can also show that degree-d PTFs are hard to learn for $d > 2$. Note that this introduces a fundamental tradeoff: The larger we choose d, the smaller the number of halfspaces becomes but the blow-up in the dimension is exponential in d. Luckily for us, there is still a choice of d that rules out learning $\omega(\log \log N)$ halfspaces.

Tiegel (2023) (building on Bruna et al. (2021)) showed the following (see section 4 for all details[5]): There are two one-dimensional distributions A, B satisfying

1. A, B are mixtures of infinitely many (truncated) Gaussians.
2. There exists two unions of infinitely many disjoint intervals S_A, S_B, such that A (resp. B) is supported on S_A (resp. S_B).
3. The intervals in $S_A \cup S_B$ are disjoint and ”interlacing” in the sense that they alternate.
4. If there is an algorithm distinguishing $D_{A,B}$ from $N(0, \mathrm{Id}_n) \times \mathrm{Be}(\frac{1}{2})$ using $2^{o(n)}$ samples and running in time $2^{o(n)}$, then assumption 4 is false.

In what follows we will denote the dimension of $D_{A,B}$ by n. We will denote the dimension of the halfspaces by N (which will roughly be n^d). Our first observation is that we can restrict to the $2n+1$ most central intervals in A and B respectively. It is not hard to show that the resulting $D_{A,B}$ is $2^{-\Omega(n)}$-close to the original one in total variation distance. Thus, even when seeing $2^{o(n)}$ samples from this distribution, the respective product distributions are still $2^{-\Omega(n)}$ close in total variation, and hence, the associated distinguishing problem is just as hard. We can hence assume that S_A, S_B contain only $2n+1$ intervals.

Let $d = \frac{2n+1}{k} + 1$ and for simplicity assume this is an even integer. By a similar construction as for the SQ lower bound, $D_{A,B}$ can be realized as an intersection of k degree-d PTFs – this time each PTF traces out $d-1$ intervals in S_B, instead of just 1. See Figure 1 (b) for an illustration. These can be realized as an intersection of k halfspaces in dimension $N = \Theta(n^d)$. Recall that we want to rule out algorithms weakly learning the intersections of halfspaces that run in time $N^{o(\frac{k}{\log k + \log \log N})}$. We claim that such an algorithm can distinguish $D_{A,B}$ from $N(0, \mathrm{Id}_n) \times \mathrm{Be}(()\frac{1}{2})$. In particular, since $\log N = \Theta(d \cdot \log n) = \Theta(\frac{n}{k} \cdot \log n)$ an algorithm running in time $N^{o(\frac{k}{\log k + \log \log N})}$ runs in time $2^{o(n)}$. Indeed,

$$
\begin{aligned}
N^{o\left(\frac{k}{(\log k + \log \log N)}\right)} &= \exp\left(o\left(\frac{k \cdot \log N}{\log k + \log \log N}\right)\right) \\
&= \exp\left(o\left(\frac{n \cdot \log n}{\log k + \log n - \log k + \log \log n}\right)\right)
\end{aligned}
$$

[5] Tiegel (2023) used a construction based on these distributions to show that learning a single halfspace in the agnostic model is hard under assumption 4. Note that this is different from our setting as we do not allow noise in the labels.

$$= 2^{o(n)} .$$

Thus, to solve the distinguishing problem we can use a similar reduction as in the SQ model: Run the learner on the first half of the input samples and compute the empirical misclassification error on the second. Again, under null this should be very close to 1/2 whereas under planted it should be bounded away from 1/2.

Comparison to Klivans and Sherstov (2009) The work Klivans and Sherstov (2009) shows that $O(n)$ degree-2 PTFs can encode the decryption function of a crypto-sytem by Regev Regev (2009). Under assumption 4 breaking this crypto-system requires time at least $2^{\Omega(n)}$. Using a similar argument as above, they deduce that learning $O(\sqrt{N})$ halfspaces in dimension N takes time at least $2^{\Omega(\sqrt{N})}$ – where the $\sqrt{N}$ comes from the quadratic blow-up in the dimension. Further, they argue the following: For any $\varepsilon > 0$, by padding all vectors with 0, we can artificially blow-up the dimension to $N = n^{\frac{1}{\varepsilon}}$. The number of halfspaces is then $k = N^{\varepsilon}$ (over the N-dimensional space) and the learning task requires time at least $2^{\Omega(n)} = \exp(N^{\varepsilon}) = \exp(k)$. It follows that learning $\omega(\log N)$ halfspaces in dimension N takes time at least $N^{\omega(1)}$.

Note that this simple padding argument cannot go beyond $\omega(\log N)$ halfspaces, intuitively, the padding strategy does not exploit the additional space available in higher dimensions. On the other hand, our argument based on higher-degree PTFs shows that exploiting this is indeed possible. Further, our arguments completely bypass the need to introduce any crypto-systems. In fact, it is not clear how the construction based on Regev's crypto-system would yield unconditional lower bounds in the SQ model.

3. Preliminaries

Notation We denote $\mathbb{R}_{\geqslant 0} = [0, \infty)$ and $\mathbb{R}_{>0} = (0, \infty)$. For a set S, we denote by $\mathcal{U}(S)$ the uniform distribution over S. We define the Total Variation Distance between two measures P and Q as

$$\mathrm{TVD}(P, Q) = \sup_A |P(A) - Q(A)| .$$

Let n be some parameter. For the problem of distinguishing two distributions D_n^0 and D_n^1 we define the advantage of an algorithm $\mathcal{A}$ as

$$\left|\mathbb{P}_{x\sim D_n^0}\left(\mathcal{A}(x) = 0\right) - \mathbb{P}_{x\sim D_n^1}\left(\mathcal{A}(x) = 0\right)\right| .$$

We say that an algorithm has non-negligible advantage if it has advantage $\Omega(n^{-c})$ for some constant $c > 0$.

Let $p \in [0, 1/2]$. We denote by $\mathrm{Be}(p)$ the distribution that is equal to +1 with probability p and equal to -1 with probability $1 - p$.

Let $\mathcal{X}$ be some set and D be a distribution over $\mathcal{X} \times \{-1, +1\}$. Further, let $h\colon \mathcal{X} \to \{-1, +1\}$ be a binary hypothesis. We denote the *misclassification error* of h as

$$\mathrm{err}_D\left(h\right) = \mathbb{P}_{(x,y)\sim D}\left(h(x) \neq y\right) .$$

Most of the time the distribution D will be clear from context and we will omit the subscript. We denote by D_x the marginal distribution of D over $\mathcal{X}$. If the domain of D_x is $\mathbb{R}^n$, we say an algorithm weakly learns D, if it outputs a binary hypothesis $\hat{h}$ such that $\mathrm{err}_D(\hat{h}) \leqslant \frac{1}{2} - \frac{1}{\mathrm{poly}(n)}$ for some choice of $\mathrm{poly}(n)$.

Gaussian distributions We denote the standard n-dimensional Gaussian distribution by $N(0, I_n)$. If the dimension is clear from context, we sometimes drop the subscript of the identity matrix. For $s > 0$, we denote by $\rho_s \colon \mathbb{R}^n \to \mathbb{R}_+$ the function

$$\rho_s(x) = \exp(-\pi\|x/s\|^2) \,.$$

If $s = 1$, we omit the subscript. Note that ρ_s/s^n is equal to the probability density function of the n-dimensional Gaussian distribution with mean 0 and covariance matrix $s^2/(2\pi) \cdot I_n$. In particular, it holds that

$$\int_{\mathbb{R}^n} \rho_s(x)\, dx = s^n \,.$$

We define $\rho_s(x\,;c) = \rho_s(x - c)$ and for $\alpha > 0$ we define

$$\rho_s^\alpha(x\,;c) = \begin{cases} \frac{1}{Z} \cdot \rho_s(x\,;c)\,, & \text{if } \|x-c\| \leqslant \alpha\,, \\ 0\,, & \text{otherwise,} \end{cases}$$

where

$$Z = \frac{\int_{\|x-c\|\leqslant\alpha} \rho_s(x\,;c)\, dx}{\int_{\mathbb{R}} \rho_s(x\,;c)\, dx} \,.$$

For a lattice $L \subseteq \mathbb{R}^n$ and $s > 0$ we define the discrete Gaussian distribution $D_{L,s}$ with width s as having support L and probability mass proportional to ρ_s. Further, for a discrete set S, we define $\rho_s\left(S\right) = \sum_{x\in S} \rho_s\left(x\right)$.

Various versions of Continuous LWE

Definition 7 (CLWE Distribution) *Let $w \in \mathbb{R}^n$ be a unit vector and $\beta, \gamma > 0$. Define the distribution $\mathrm{C}_{w,\beta,\gamma}$ over $\mathbb{R}^n \times [0,1)$ as follows. Draw $y \sim N(0, \frac{1}{2\pi} \cdot I_n)$, $e \sim N(0, \beta^2/(2\pi))$ and let*

$$z = \gamma \cdot \langle y, w\rangle + e \mod 1\,.$$

Note that the density of this distribution is given by

$$p(y,z) = \frac{1}{\beta} \cdot \rho\left(y\right) \cdot \sum_{k\in\mathbb{Z}} \rho_\beta\left(z + k - \gamma\langle w, y\rangle\right)\,.$$

Further, let $m \in \mathbb{N}$. We denote by $\mathrm{CLWE}\left(m, \gamma, \beta\right)$ the distribution obtained by first drawing $w \sim \mathcal{U}(\mathcal{S}^{n-1})$ and then drawing m independent samples from $\mathrm{C}_{w,\gamma,\beta}$.

Definition 8 (Homogeneous CLWE (hCLWE) Distribution) *Let $w \in \mathbb{R}^n$ be a unit vector, $c \in [0,1)$, and $\beta, \gamma > 0$. Let $\pi_{w^\perp}(y)$ be the projection of y onto the space orthogonal to w. Define the distribution $\mathrm{H}_{w,\beta,\gamma,c}$ over $\mathbb{R}^n$ as having density at y proportional to*

$$\sum_{k\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c) \cdot \rho\left(\pi_{w^\perp}(y)\right) \cdot \rho_{\beta/\sqrt{\beta^2+\gamma^2}}\left(\langle w, y\rangle\,; \frac{\gamma}{\beta^2+\gamma^2}(k-c)\right)\,. \tag{3.1}$$

Further, let $m \in \mathbb{N}$. We denote by $\mathrm{HCLWE}\left(m, \gamma, \beta, c\right)$ the distribution obtained by first drawing $w \sim \mathcal{U}(\mathcal{S}^{n-1})$ and then drawing m independent samples from $\mathrm{H}_{w,\gamma,\beta,c}$.

Intuitively, one can think of the $\mathrm{H}_{w,\gamma,\beta,c}$ distribution as $\mathrm{C}_{w,\gamma,\beta}$ conditioned on $z = c$.

Definition 9 (Truncated hCLWE Distribution) *Let $w \in \mathbb{R}^n$ be a unit vector, $c \in [0,1), \beta, \gamma > 0$ and $\alpha = \frac{1}{10} \cdot \frac{\gamma}{\gamma^2+\beta^2}$. Define the distribution $\mathrm{NH}^{(n)}_{w,\beta,\gamma,c}$ over $\mathbb{R}^n$ as having density proportional to*

$$\sum_{k=-n}^{n} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c) \cdot \rho\left(\pi_{w^\perp}(y)\right) \cdot \rho^{\alpha}_{\beta/\sqrt{\beta^2+\gamma^2}}\left(\langle w, y\rangle\,; \frac{\gamma}{\beta^2+\gamma^2}\left(k-c\right)\right) . \tag{3.2}$$

The superscript refers to the range of the summation.

Further, let $m \in \mathbb{N}$ and $\mathcal{S}$ be a distribution over unit vectors in $\mathbb{R}^n$. We denote by $\mathrm{NHCLWE}\left(m, \gamma, \beta, c\right)$ the distribution obtained by first drawing $w \sim \mathcal{U}(\mathcal{S}^{n-1})$ and then drawing m independent samples from $\mathrm{NH}_{w,\gamma,\beta,c}$.

Note that this is the same as the hCLWE distribution but with the individual components of the mixture truncated in the hidden direction and restricting to the middle $2n+1$ components. α is chosen such that the components become non-overlapping but the resulting distribution has small total variation distance to the corresponding non-truncated hCLWE distribution. Although this is strictly speaking not necessary to prove our result, we will see that having non-overlapping components will simplify our analysis.

We make the following hardness assumption about the CLWE distribution

Assumption 10 *Let $n, m \in \mathbb{N}$ and*

$$\gamma \geqslant 2\sqrt{n}\,, \qquad \beta = \frac{1}{\mathrm{poly}\left(n\right)}\,.$$

Further, let $\delta < 1$ be arbitrary and $m = 2^{n^\delta}$. There is no 2^{n^δ}-time distinguisher between

$$\mathrm{CLWE}\left(m, \gamma, \beta\right) \quad \textit{and} \quad N\left(0, \tfrac{1}{2\pi} \cdot I_n\right)^m \times U\left([0,1)\right)^m$$

with non-negligible advantage.

Note that by (Bruna et al., 2021, Corollary 3.2) this is implied by assuming theorem 4.

Hermite polynomials and moment-matching distributions We will also use the following facts about one-dimensional distributions matching moments with $N(0,1)$.

Fact 11 (Bubeck et al. (2019)) *For every $k \in \mathbb{N}$ greater or equal to 2, there exist two discrete A and B supported on at most k points such that*

- *A and B match at least $2k-1$ moments with $N(0,1)$,*
- *The points in the union of the supports of A and B are pairwise at distance at least $\Omega(1/\sqrt{k})$. Further, they are all contained in the interval $[-C\sqrt{k}, C\sqrt{k}]$ for some sufficiently large absolute constant $C > 0$.*

The support of A and B corresponds to the roots of the k-th and $(k-1)$-th normalized probabilist's Hermite polynomials.

4. Hardness Under Assumption 4

In this section, we will prove a slightly more general version of theorem 5. We remark that we will not directly work with theorems 2 and 3 but rather with the continuous LWE problem introduced in Bruna et al. (2021).

Theorem 12 *Let $0 \leqslant \delta < 1$. Let $k, N \in \mathbb{N}$ such that $k \leqslant O(\sqrt{N})$. Assume there is an algorithm that learns the intersection of $k \leqslant O(N)$ halfspaces in dimension N in time $T = N^{o\left(\frac{k^{1-\delta}}{(\log k + \log\log N)^{1-\delta} \cdot \log^{\delta} N}\right)}$ up to error better than $\frac{1}{2} - \frac{1}{\Omega(T)}$, then there is an algorithm that solves CLWE in dimension n in time $2^{o\left(n^{1-\delta}\right)}$. Furthermore, every halfspace in the hard instance has a margin of $\Omega(\frac{1}{N \cdot \sqrt{k \cdot \log N}})$*

We will use the following two facts which are a straightforward extensions of facts in Tiegel (2023). We will prove them in appendix B.2.

Fact 13 (Adaptation of Theorem 15 in Tiegel (2023)) *Let $n, m \in \mathbb{N}$ with $2^{o(n)} = m > n$, and let $\gamma, \beta, \varepsilon \in \mathbb{R}_{>0}$ such that $0 \leqslant \beta \leqslant \gamma, \beta = \frac{1}{\mathrm{poly}(n)}$. Assume that there is no $(T + \mathrm{poly}(n, m))$-time distinguisher between*

$$\mathrm{CLWE}\left(m, \gamma, \beta\right) \quad \textit{and} \quad \left(N\left(0, \tfrac{1}{2\pi} \cdot I_n\right) \times U\left([0,1)\right)\right)^{\otimes m}$$

with advantage ε. Let $m' = \frac{m}{\mathrm{poly}(n)}$. Then there is no T-time distingiusher between

$$\frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}, +1\right) + \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}}, -1\right) \quad \textit{and} \quad N\left(0, \tfrac{1}{2\pi} \cdot I_n\right) \times \mathrm{Be}\left(\frac{1}{2}\right)$$

with advantage $\varepsilon - \mathrm{negl}(n)$ that uses at most m' samples.

Further, we will use the following fact about the supports of the mixture of homogeneous CLWE distributions. Its proof is contained in the proof of Lemma 11 in Tiegel (2023):

Fact 14 *Let $S^{(0)}, S^{(1)}$ be the support of $\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}$ and $\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}}$ respectively. Let $\alpha = \frac{1}{10} \cdot \frac{\gamma}{\gamma^2+\beta^2}$ and for $k \in \mathbb{N}$, let $\mu_k^{(0)} = \frac{\gamma}{\gamma^2+\beta^2} k$, $\mu_k^{(1/2)} = \frac{\gamma}{\gamma^2+\beta^2}(k - \frac{1}{2})$ then*

$$S^{(0)} = \bigcup_{k=-n}^{n} \left\{ x \in \mathbb{R}^n \;\middle|\; \langle x, w \rangle \in [\mu_k^{(0)} - \alpha, \mu_k^{(0)} + \alpha] \right\},$$

$$S^{(1)} = \bigcup_{k=-n}^{n} \left\{ x \in \mathbb{R}^n \;\middle|\; \langle x, w \rangle \in [\mu_k^{(1/2)} - \alpha, \mu_k^{(1/2)} + \alpha] \right\}.$$

Further, $S^{(0)}$ and $S^{(1)}$ are disjoint and at distance at least $\frac{1}{5} \cdot \frac{\gamma}{\gamma^2+\beta^2}$.

Proof [Proof of theorem 12] Let $d, k \in \mathbb{N}$ and $0 \leqslant \delta < 1$ (it might be instructive to think of $\delta = 0$ first). For simplicity assume that $2n+1$ is divisible by d and let $k = \frac{2n+1}{d}$. Let $m \leqslant N^{o\left(\frac{k^{1-\delta}}{(\log k + \log\log N)^{1-\delta} \log^{\delta} N}\right)}$ and $\tau \geqslant \frac{5}{\sqrt{m}}$. We will choose N such that $m \leqslant 2^{o(n^{1-\delta})}$. It follows by theorem 13, that if there is an algorithm that can distinguish between

$$D^{(p)} = \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}, +1\right) + \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,1/2}, -1\right) \qquad \text{and} \qquad D^{(n)} = N\left(0, \tfrac{1}{2\pi} I_n\right) \times \mathrm{Be}\left(\tfrac{1}{2}\right)$$

with probability at least $2/3$ in time $2^{o(n^{1-\delta})}$ and using at most $2^{o(n^{1-\delta})}$ samples, then there also is an algorithm that solves CLWE with probability at least, say, 0.6 in time $2^{o(n^{1-\delta})}$ and using at most $2^{o(n^{1-\delta})}$ samples. We will show that a learning algorithm would imply the former.

The reduction Suppose we are given m samples $((x_i, y_i))_{i=1}^m \in \mathbb{R}^n \times \{-1, +1\}$ from either of the two distributions. Our reduction does the following: Let $N = \sum_{j=0}^{2d}(n+1)^j = \Theta(n^{2d})$. We apply the Veronese mapping to the x_i, obtaining $((\tilde{x}_i, y_i))_{i=1}^m \in \mathbb{R}^N \times \{-1, +1\}$, where $\tilde{x}_i = ((1, x_i)^\alpha)_{|\alpha| \leqslant 2d}$. For simplicity, assume that m is even. We run our learning algorithm on the first $m/2$ samples to obtain a function $\hat{f} \colon \mathbb{R}^N \to \{+1, -1\}$. Let

$$\widehat{\operatorname{err}(f)} = \frac{2}{m} \sum_{i=m/2}^{m} \mathbf{1}\left(\hat{f}(x_i') \neq y_i\right) .$$

If $\left|\widehat{\operatorname{err}(f)} - \frac{1}{2}\right| > \frac{\tau}{2}$ we output planted and else we output null.

First assuming that the learner runs in time $N^{o\left(\frac{k}{\log k + \log\log N}\right)}$, notice that the procedure described above runs in the same time – the reduction only add an overhead of $N^{O(1)}$. We claim that this total time is equal to $2^{o(n^{1-\delta})}$. Indeed, using that $\log N = \Theta(d \cdot \log n) = \Theta(\frac{n}{k} \cdot \log n)$, we obtain

$$\begin{aligned} N^{o\left(\frac{k^{1-\delta}}{(\log k + \log\log N)^{1-\delta} \log^\delta N}\right)} &= \exp\left(o\left(\left(\frac{k \cdot \log N}{\log k + \log\log N}\right)^{1-\delta}\right)\right) \\ &= \exp\left(o\left(\left(\frac{n \cdot \log n}{\log k + \log n - \log k + \log\log n}\right)^{1-\delta}\right)\right) \\ &= 2^{o\left(n^{1-\delta}\right)} . \end{aligned}$$

To argue that it successfully distinguishes between $D^{(p)}$ and $D^{(n)}$, we proceed in two parts. If the input comes from $D^{(n)}$, $y_i \sim \operatorname{Be}(\frac{1}{2})$ and is independent of x_i', hence $\widehat{\operatorname{err}(f)}$ will be close to $\frac{1}{2}$. If the input comes from $D^{(p)}$, we will show that the samples input to our learning algorithm can be realized as the intersection of k halfspaces – we will assume this for now in the next paragraph. Hence, since we assume access to a weak learner, $\widehat{\operatorname{err}(f)}$ will be sufficiently smaller than $\frac{1}{2}$.

Indeed, under both null and planted the random variables $\mathbf{1}\left(f(x_i') \neq \tilde{y}_i\right)$ are i.i.d. Bernoulli with some expectation $p_n, p_p \in [0,1]$ respectively. Assume $(x_i, y_i) \sim D^{(n)} = N\left(0, \frac{1}{2\pi} I_n\right) \times \operatorname{Be}(\frac{1}{2})$. Since $(\tilde{x_i}, y_i) = (g(x_i), y_i)$ for a deterministic function g, it follows that y_i is independent from $\tilde{x}_i$. Since clearly, $y_i \sim \operatorname{Be}(\frac{1}{2})$ it follows that $p_n = \frac{1}{2}$. By assumption, $p_p \leqslant \frac{1}{2} - \tau$ is the success probability of our learning algorithm. It follows by Hoeffding's Inequality Hoeffding (1994) and since $\tau \geqslant \frac{5}{\sqrt{m}}$that in either case (i.e., for $p = p_n$ or $p = p_p$) it holds that

$$\mathbb{P}\left(\left|\widehat{\operatorname{err}(f)} - p\right| \geqslant \tfrac{\tau}{3}\right) \leqslant 2\exp\left(-\tfrac{m}{9}\tau^2\right) \leqslant 2\exp\left(-2.5\right) \leqslant \frac{1}{3} .$$

Hence, under the null distribution we correctly output null with probability at least $1/3$. Similarly, since under the planted distribution with probability at least $1/3$

$$\left|\widehat{\operatorname{err}(f)} - \frac{1}{2}\right| \geqslant \left(\frac{1}{2} - p_p\right) - \frac{\tau}{3} > \frac{\tau}{2}$$

we correctly output planted with the same probability.

The planted distribution is an intersection of k halfspaces Next, assume

$$(x_i, y_i) \sim D^{(p)} = \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}, +1\right) + \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,1/2}, -1\right) .$$

We argue that $D^{(p)}$ can be realized as an intersection of k degree-d polynomial threshold functions. That is, we show that there exists polynomials $p_1, \ldots, p_n \colon \mathbb{R}^n \to \mathbb{R}$ of degree at most d, such that for all $(x, y) \sim D^{(p)}$ it holds that $y = 1$ if and only if $p_j(x) \geqslant 0$ for all $j = 1, \ldots, k$. Note that this directly implies that the transformed samples $(\tilde{x}, y)$ we feed to our learning algorithm can be realized as an intersection of k halfspaces. In particular, the halfspaces correspond to the linearizations of $p_1, \ldots, p_k$.

Recall that w is the hidden direction in the planted distribution. All polynomials p_j will be of the form $p_j(x) = \tilde{p}_j(\langle x, w\rangle)$ for one-dimensional polynomials $\tilde{p}_1, \ldots, \tilde{p}_k$. On a high level, these will trace out the support of the positive and negative examples. Indeed, let $\alpha = \frac{1}{10} \cdot \frac{\gamma}{\gamma^2+\beta^2}$ and for $k = -n, \ldots, n$ let $\mu_k^{(0)} = \frac{\gamma}{\gamma^2+\beta^2} k, \mu_k^{(1/2)} = \frac{\gamma}{\gamma^2+\beta^2}(k - \frac{1}{2})$. Define

$$J_\ell^+ = [\mu_\ell^{(0)} - \alpha, \mu_\ell^{(0)} + \alpha] , \qquad J_\ell^- = [\mu_\ell^{(1/2)} - \alpha, \mu_\ell^{(1/2)} + \alpha] .$$

Recall that $k = \frac{2n+1}{d}$. Let $\tilde{p}_j$ be a degree-$2d$ polynomial that is negative on $J^-_{-n+(j-1)\cdot d}, \ldots, J^-_{-n+j\cdot d-1}$, positive on $J^+_{-n+(j-1)\cdot d}, \ldots, J^+_{-n+j\cdot d-1}$ and positive starting some distance away from the left and right-most ”negative” interval. Let its root be at the midpoints between consecutive intervals and the left-most root at the same distance to the left-most interval. Note that by construction, the following two properties hold (the first property also uses that all $\tilde{p}_j$ are positive after their last root)

1. If $z \in J_\ell^+$ for some ℓ, then $\tilde{p}_j(z) \geqslant 0$ for all $j \in [k]$,
2. If $z \in J_\ell^-$ for some ℓ, then there exists $j^* \in [k]$ such that $\tilde{p}_{j^*}(z) < 0$.

Recall from theorem 14 that $y = 1$ implies that $\langle x, w\rangle \in \cup_{\ell=-n}^n J_\ell^+$ and $y = -1$ implies that $\langle x, w\rangle \in \cup_{\ell=-n}^n J_\ell^-$. Hence, the two properties above imply that $y = 1$ if and only if $p_j(x) \geqslant 0$ for all $j \in [k]$.

■

Acknowledgments

The author would like to thank Kiril Bangachev, Guy Bresler, and Vinod Vaikuntanathan for helpful discussions. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 815464).

References

Michael Alekhnovich, Mark Braverman, Vitaly Feldman, Adam R. Klivans, and Toniann Pitassi. Learnability and automatizability. In *FOCS*, pages 621–630. IEEE Computer Society, 2004.

Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 211–220. IEEE, 2008.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous LWE. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 694–707. ACM, 2021. doi: 10.1145/3406325.3451000.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840. PMLR, 2019.

Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In *Conference on Learning Theory*, pages 815–830. PMLR, 2016.

Amit Daniely and Gal Vardi. From local pseudorandom generators to hardness of learning. In *Conference on Learning Theory*, pages 1358–1394. PMLR, 2021.

Ilias Diakonikolas and Daniel Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282. PMLR, 2022. URL https://proceedings.mlr.press/v178/diakonikolas22b.html.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS*, pages 73–84. IEEE Computer Society, 2017a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017b.

Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. personal communication, 2022.

Ilias Diakonikolas, Daniel M Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. *arXiv preprint arXiv:2302.06512*, 2023.

Uriel Feige. Relations between average case complexity and approximation complexity. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 534–543. ACM, New York, 2002. doi: 10.1145/509907.509985. URL http://dx.doi.org/10.1145/509907.509985.

Vitaly Feldman. Optimal hardness results for maximizing agreements with monomials. In *21st Annual IEEE Conference on Computational Complexity (CCC'06)*, pages 9–pp. IEEE, 2006.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 563–574. IEEE, 2006.

Parikshit Gopalan, Subhash Khot, and Rishi Saket. Hardness of reconstructing multivariate polynomials over finite fields. *SIAM J. Comput.*, 39(6):2598–2621, 2010. ISSN 0097-5397. doi: 10.1137/070705258. URL http://dx.doi.org/10.1137/070705258.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. April 2022.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *FOCS*, pages 543–552. IEEE Computer Society, 2006.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Adam R Klivans and Alexander A Sherstov. Cryptographic hardness results for learning intersections of halfspaces. *Available as ECCC report TR06-057*, 2006.

Adam R Klivans and Alexander A Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69:97–114, 2007.

Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. System Sci.*, 68(4):808–840, 2004a. ISSN 0022-0000. doi: 10.1016/j.jcss.2003.11.002. URL http://dx.doi.org/10.1016/j.jcss.2003.11.002.

Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004b.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *FOCS*, pages 541–550. IEEE Computer Society, 2008.

Adam R Klivans, Philip M Long, and Alex K Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 588–600. Springer, 2009.

Wolfgang Maass and György Turán. How fast can a threshold gate learn? In *Proceedings of a workshop on Computational learning theory and natural learning systems (vol. 1): constraints and prospects: constraints and prospects*, pages 381–414, 1994.

Rajai Nasser and Stefan Tiegel. Optimal SQ lower bounds for learning halfspaces with massart noise. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 1047–1074. PMLR, 2022. URL https://proceedings.mlr.press/v178/nasser22a.html.

Chris Peikert et al. A decade of lattice cryptography. *Foundations and Trends® in Theoretical Computer Science*, 10(4):283–424, 2016.

Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *J. ACM*, 35(4):965–984, 1988.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56 (6):34:1–34:40, 2009.

Alexander A Sherstov. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, pages 523–532, 2010.

Alexander A Sherstov. The hardest halfspace. *computational complexity*, 30(2):11, 2021.

Stefan Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3029–3064. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/tiegel23a.html.

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

Santosh Vempala. Learning convex concepts from gaussian distributions with PCA. In *FOCS*, pages 124–130. IEEE Computer Society, 2010.

Appendix A. SQ Hardness

In this section, we will prove our SQ lower bound (theorem 6).

Theorem 15 *Let $\beta \in (0, \frac{1}{2})$ be an absolute constant and $k, N \in \mathbb{N}$ be such that $2 \leqslant k \leqslant N^\gamma$ for a sufficiently small absolute constant γ. Every SQ algorithm that uses queries of accuracy $\rho = N^{-\Omega(k)}$ and learns intersections of k halfspaces in dimension N up to error better than $1/2 - 4\rho$ needs at least $2^{N^{\Omega(1)}}$ queries.*

To favor clarity of exposition and since in our eyes the ”small k” regime is the most interesting one, we have not tried to optimize constants, i.e., γ. We will show the theorem above by constructing a distribution over $(x, y) \in \mathbb{R}^N \times \{-1, +1\}$ that (a) is an intersection of k halfspaces and (b) the conditional distribution of x given $y = -1$ and $y = +1$ respectively (nearly) matches k moments with the standard Gaussian.

In particular, our hard instance will follow the NGCA framework and will be similar to the construction of Bubeck et al. (2019) – in their distribution however, the labels are not without noise. So we will need to slightly modify it. That is, the distribution conditioned on $y = +1$ and $y = -1$ will be equal to the standard Gaussian distribution except in one direction (the same direction in both cases), and equal to a distribution that nearly matches k moment with $N(0, 1)$ along said direction. We start by describing the distribution along this direction: From theorem 11 we know that there exists discrete distributions A, B supported on at most k points both matching $2k - 1$ moments with $N(0, 1)$ and such that all points in the union of their supports are at distance at least

$\Omega(1/\sqrt{k})$. Let $\tilde{A}, \tilde{B}$ be the distributions that are obtained from A, B via the following process – we only describe it for A. First, let A' be the distribution obtained as follows: Let $\delta > 0$. Draw $X \sim A$ and $Z \sim N(0,1)$ independently. Output $\sqrt{1-\delta} \cdot X + \delta \cdot Z$. Note that A' is a mixture of at most k Gaussians. Second, truncate each component of A' at distance τ from its mean. Later we will choose $\delta = \frac{1}{k^2 \log N}$ and $\tau = c \cdot \sqrt{\delta \cdot k \log N}$ (for a small enough absolute constant $c > 0$) Our family of hard instances D can be described as follows:

1. Draw $w \sim \mathbb{S}^{n-1}$ uniformly at random.

2. Let D_w^A be the product distribution that is A along w and a standard Gaussian in the complement (and the same for D_w^B).

3. Set $D = \frac{1}{2} \cdot (D_w^A, +1) + \frac{1}{2} \cdot (D_w^B, -1)$.

We will use the notation above throughout the rest of this section.

We will use the following theorem to show that D is in fact hard to learn in the SQ model: It is an instantiation of results from Diakonikolas and Kane (2022) (and a slight refinement of Nasser and Tiegel (2022) already implicit in the first work). See appendix B for full details how this follows from their theorems.

Theorem 16 *Let $\beta \in (0, \frac{1}{2})$ be an absolute constant and $\beta' < \beta$. Let $k, N \in \mathbb{N}$ be such that $k \leqslant N^\gamma$ for a sufficiently small absolute constant γ. Let A, B be two one-dimensional distributions that match k moments with $N(0,1)$ up to error $N^{-\Omega(k)}$ and such that $\chi^2(A, N(0,1)), \chi^2(A, N(0,1)) \leqslant 2^{O(k)} \log N$. Let the family of distributions $\mathcal{D}$ be as above. Then any SQ algorithm with accuracy $\rho = N^{-\Omega(k)}$ that learns $\mathcal{D}$ up to error $\frac{1}{2} - 4\rho$ needs at least $2^{N^{\Omega(1)}}$ queries.*

We can now proceed to prove theorem 15:

Proof [Proof of theorem 15] Let $A', B', \tilde{A}, \tilde{B}$ be as above. As mentioned before, our proof proceeds in two steps: First, we show that D corresponds to an intersection of k degree-2 polynomial threshold functions and second, we will appeal to theorem 16 to show that D is hard to learn. Just as in the proof of theorem 12 this will imply the claim by applying the Veronese mapping. Note that the blow-up in the dimension is only quadratic and thus can be absorbed in the $\Omega(\cdot)$- and $O(\cdot)$-notation in our theorem statement. We first set parameters, let $c > 0$ be a sufficiently small absolute constant, we set

$$\delta = \frac{1}{k^2 \log N} \qquad \text{and} \qquad \tau = c \cdot \sqrt{\delta k \log N} = \frac{c}{\sqrt{k}} \,.$$

The hard instance is an intersection of k degree-2 PTFs Recall that in A' (resp. B') the mixture components have variance δ and in $\tilde{A}$ (resp. $\tilde{B}$) we truncate them at distance τ from their means. In particular, let $S_A, S_B \subseteq \mathbb{R}$ be the collection of intervals of length 2τ around the means of the components of A' and B'. Since by theorem 11 the means of the components (of both A' and B' together) are at least $\Omega(\frac{1}{\sqrt{k}})$ apart, we can choose c in the definition of τ small enough such that the intervals in $S_A \cup S_B$ are disjoint and at distance $\Omega(\frac{1}{\sqrt{k}})$. Note that by construction, S_A and S_B contain at most k intervals

The proof is analogous to theorem 12 with the only difference that we will only use degree-2 polynomials. Indeed, by construction, for a sample $(x, y) \sim D$, $y = 1$ if and only if $\langle x, w \rangle \in S_A$. Further $y = -1$ if and only if $\langle x, w \rangle \in S_B$. Thus, for every interval $I \subseteq S_B$, consider the polynomial $p_I \colon \mathbb{R} \to \mathbb{R}$ that is symmetric around the mid-point of B, is negative on I, and has its roots at half the distance between the end of I and the next interval in S_A. Note that p_I is negative on I and positive on all other intervals. The final choice of degree-2 PTFs is then $\tilde{p}_I \colon \mathbb{R}^N \to \mathbb{R}$ such that $\tilde{p}_I(x) = p_I(\langle x, w\rangle)$. By construction, if $\langle x, w\rangle \in S_A$, $\tilde{p}_I(x) \geqslant 0$ for all I and if $\langle x, w\rangle \in S_B$ there exists $\tilde{p}_I$ such that $\tilde{p}_I(x) < 0$. It follows that D corresponds to the intersection of the $\tilde{p}_I$.

Set-up for SQ lower bound and χ^2-divergence Note that in order to prove theorem 15 it is now enough to verify that $\mathcal{D}$ satisfies the conditions of theorem 16. Since the conditions on N and k are assumed to be true, it only remains to verify the following

1. $\tilde{A}$ and $\tilde{B}$ match k moments with $N(0,1)$ up to error $N^{-\Omega(k)}$,
2. $\chi^2(A, N(0,1))$ and $\chi^2(B, N(0,1))$ are at most $2^{O(k)} \log N$.

We will verify the properties above only for $\tilde{A}$, $\tilde{B}$ is completely analogous. Then theorem 15 is implied by theorem 16.

We start with the χ^2-divergence. Let S_A be as in the previous paragraphs. Note that $\tilde{A}$ is the distribution A' conditioned on lying in S_A. In particular, it follows that $p_{\tilde{A}}(x) = 1\,(x \in S_A) \cdot \frac{p_{A'}(x)}{\mathbb{P}_{X\sim A'}(X \in S_A)}$. By standard concentration bounds for the Gaussian distribution, it follows that $\mathbb{P}_{X\sim A'}(X \notin S_A) \leqslant \exp(-\Omega(\frac{\tau^2}{\delta}))$ and hence also that $\mathbb{P}_{X\sim A'}(X \in S_A) \geqslant 1 - \exp(-\Omega(\frac{\tau^2}{\delta})) \geqslant \frac{1}{2}$. Denote the pdf of $N(0,1)$ by G. From (Diakonikolas et al., 2017b, Lemma 4.6), we now that $\chi^2(A', N(0,1)) \leqslant 2^{O(k)}/\sqrt{\delta}$. It follows that

$$\begin{aligned}
\chi^2(\tilde{A}, N(0,1)) + 1 &= \int_{-\infty}^{\infty} \frac{p_{\tilde{A}}(x)^2}{G(x)}\, dx = \frac{1}{\mathbb{P}_{X\sim A'}\left(X \in S_A\right)^2} \cdot \int_{S_A} \frac{p_{A'}(x)^2}{G(x)}\, dx \\
&\leqslant 4 \cdot \int_{-\infty}^{\infty} \frac{p_{A'}(x)^2}{G(x)}\, dx \leqslant 4\chi^2\left(A', N(0,1)\right) + 4 \\
&\leqslant \frac{2^{O(k)}}{\sqrt{\delta}}\,.
\end{aligned}$$

Recalling that $\delta = \frac{1}{k^2 \log N}$ we obtain that $\chi^2(\tilde{A}, N(0,1)) \leqslant 2^{O(k)} \log N$.

Moment matching By theorem 11 A matches $2k-1$ moments exactly with $N(0,1)$. We claim A' does too: Indeed, for every integer $0 \leqslant \ell \leqslant 2k-1$ we have (in the following X, Z, Z' are all independent)

$$\begin{aligned}
\mathop{\mathbb{E}}_{X'\sim A'} (X')^\ell &= \mathop{\mathbb{E}}_{X\sim A, Z\sim N(0,1)} \left(\sqrt{1-\delta} \cdot X + \delta \cdot Z\right)^\ell \\
&= \sum_{r=0}^{\ell} \binom{\ell}{r} \mathop{\mathbb{E}}_{X\sim A} (1-\delta)^{r/2} \cdot X^r \mathop{\mathbb{E}}_{Z\sim N(0,1)} \delta^{\ell-r} \cdot Z^{\ell-r} \\
&= \sum_{r=0}^{\ell} \binom{\ell}{r} \mathop{\mathbb{E}}_{Z'\sim N(0,1)} (1-\delta)^{r/2} \cdot X^r \mathop{\mathbb{E}}_{Z\sim N(0,1)} \delta^{\ell-r} \cdot Z^{\ell-r} \\
&= \mathop{\mathbb{E}}_{Z'\sim N(0,1), Z\sim N(0,1)} \left(\sqrt{1-\delta} \cdot Z' + \delta \cdot Z\right)^\ell \\
&= E_{Z\sim N(0,1)} Z^\ell\,,
\end{aligned}$$

We next show that the moments of $\tilde{A}$ are close to the moments of A'. We start with some observations: First, note that by construction $\mathbb{P}_{\tilde{A}}(X \notin S) = 0$. Second, let $C' > 0$ be a large enough constant, such that all means are at least 2τ away from the boundary of the interval $[-C'\sqrt{k}, C'\sqrt{k}]$. Note that the density of $\tilde{A}$ is 0 outside this interval by construction. Let μ_k be the mean of the rightmost component, by theorem 11 $\mu_k = O(\sqrt{k})$. Choose C' such that $C'\sqrt{k} - \mu_k \geqslant \tau$. Since $\tau = \frac{c}{\sqrt{k}}$ for some constant c, such a choice of $C' > 0$ exists. Note that

$$\begin{aligned}\mathbb{P}_{X\sim A'}\left(|X| \geqslant C'\sqrt{k}\right) &\leqslant O(k) \cdot \mathbb{P}_{X\sim N(\mu_k,\delta)}\left(X \geqslant C'\sqrt{k}\right) \\ &\leqslant O(k) \cdot \exp\left(-\frac{\left(\mu_k - C'\sqrt{k}\right)^2}{2\delta}\right) = \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right) ,\end{aligned}$$

where we used that $\frac{\tau^2}{\delta} = \Omega(k \log N) \gg \log k$. Lastly, we note that the total variation distance between A' and $\tilde{A}$ is at most $\exp(-\Omega(\frac{\tau^2}{\delta}))$:

$$\begin{aligned}\left\|p_{A'} - p_{\tilde{A}}\right\|_1 &= \int_{-\infty}^{\infty} \left|p_{A'}(x) - p_{\tilde{A}}(x)\right| \, dx \\ &= \int_S \left(\frac{1}{\mathbb{P}_{X\sim A'}(X \in S)} - 1\right) \cdot p_{A'}(x)\, dx + \int_{S^c} p_{A'}(x)\, dx \\ &= \int_S \frac{\mathbb{P}_{X\sim A'}(X \notin S)}{\mathbb{P}_{X\sim A'}(X \in S)} \cdot p_{A'}(x)\, dx + \mathbb{P}_{X\sim A'}(X \notin S) \leqslant 3 \cdot \mathbb{P}_{X\sim A'}(X \notin S) \\ &= \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right) .\end{aligned}$$

Using the above observations, we start our moment calculations. Let $0 \leqslant \ell \leqslant k$, then

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{N(0,1)} X^\ell - \mathop{\mathbb{E}}_{\tilde{A}} X^\ell\right| = \left|\mathop{\mathbb{E}}_{A'} X^\ell - \mathop{\mathbb{E}}_{\tilde{A}} X^\ell\right| = \left|\int_{-\infty}^{\infty} x^\ell (p_{A'}(x) - p_{\tilde{A}}(x))\, dx\right| \\ &\leqslant \left|\int_{C'\sqrt{k}}^{\infty} x^\ell p_{A'}(x)\, dx + \int_{-\infty}^{-C'\sqrt{k}} x^\ell p_{A'}(x)\, dx\right| + \left|\int_{-C'\sqrt{k}}^{C'\sqrt{k}} x^\ell (p_{A'}(x) - p_{\tilde{A}}(x))\, dx\right|\end{aligned}$$

For simplicity, assume that $1.5k$ is an integer. Recall that A' matches $2k - 1 \geqslant 1.5k$ moments with $N(0,1)$. For the first absolute value, we can deduce using Hölder's Inequality with $q = \frac{3}{2}$ and $p = 3$, that

$$\begin{aligned}\int_{C'\sqrt{k}}^{\infty} x^\ell p_{A'}(x)\, dx + \int_{-\infty}^{-C'\sqrt{k}} x^\ell p_{A'}(x)\, dx &= \mathop{\mathbb{E}}_{A'} X^\ell\, \mathbf{1}\left\{|X| \geqslant C'\sqrt{k}\right\} \\ &\leqslant \left(\mathop{\mathbb{E}}_{A'} X^{1.5\ell}\right)^{\frac{2}{3}} \left(\mathbb{P}_{X\sim A'}\left(|X| \geqslant C'\sqrt{k}\right)\right)^{\frac{1}{3}} \\ &\leqslant \left(\mathop{\mathbb{E}}_{N(0,1)} X^{1.5k}\right)^{\frac{2}{3}} \left(\mathbb{P}_{X\sim A'}\left(|X| \geqslant C'\sqrt{k}\right)\right)^{\frac{1}{3}} \\ &\leqslant (2k)^k \cdot \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right) ,\end{aligned}$$

where we also used that the k-th moment of $N(0,1)$ can be upper bounded as $k^{k/2}$. Since $\frac{\tau^2}{\delta} = \Omega(k \log N)$ and $k\log(2k) \leqslant 2\gamma \cdot k \log N$ for a sufficiently small constant γ, it follows that this integral is at most $\exp(-\Omega(\frac{\tau^2}{\delta}))$. Using the total variation bound, we can bound the second absolute value:

$$\left|\int_{-C'\sqrt{k}}^{C'\sqrt{k}} x^\ell (p_{A'}(x) - p_{\tilde{A}}(x))\, dx\right| \leqslant \left(C'\sqrt{k}\right)^\ell \|p_{A'} - p_{\tilde{A}}\| \leqslant (C'\sqrt{k})^k \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right)$$

$$= \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right).$$

Combining the two above displays and using that $\frac{\tau^2}{\delta} = \Omega(k \log N)$, we obtain that

$$\left|\mathop{\mathbb{E}}_{N(0,1)} X^\ell - \mathop{\mathbb{E}}_{\tilde{A}} X^\ell\right| \leqslant \exp\left(-\Omega\left(\frac{\tau^2}{\delta}\right)\right) \leqslant N^{-\Omega(k)}.$$

■

Appendix B. Missing Lemmas

B.1. Missing Lemmas for SQ-Hardness

We will formally argue how theorem 16 follows from the results in Diakonikolas and Kane (2022); Nasser and Tiegel (2022). We start by restating Lemma 4.3 of Nasser and Tiegel (2022). We remark that this proof follows almost verbatim the proof of Diakonikolas and Kane (2022), but makes certain things more explicit which will be useful for us. The distribution $D_v^{A,B,p}$ with $p = \frac{1}{2}$ in their lemma corresponds to our $\mathcal{D}$. They denote the dimension by m instead of N. We use our notation in the restatement below.

Lemma 17 (Lemma 4.3 of Nasser and Tiegel (2022)) *Let $k \in \mathbb{N}$ and $\nu, \rho, c > 0$. Let A, B be probability distributions on $\mathbb{R}$ such that their first k moments agree with the first k moments of $N(0,1)$ up to error at most ν and such that $\chi^2(A, N(0,1))$ and $\chi^2(B, N(0,1))$ are finite. Denote $\alpha := \chi^2(A, N(0,1)) + \chi^2(B, N(0,1))$ and assume that $\nu^2 + \alpha \cdot c^k \leqslant \rho$. Then, any SQ algorithm which, given access to samples from $\mathcal{D}$, outputs a hypothesis $h \colon \mathbb{R}^N \to \{-1,+1\}$ such that*

$$\mathrm{err}_{\mathcal{D}}(h) < \frac{1}{2} - 4\sqrt{\rho},$$

must either make queries of accuracy better than $2\sqrt{\rho}$ or make at least $2^{c^2 \cdot \Omega(N)} \cdot (\rho/\alpha)$ queries.

The proof of theorem 16 follows mostly by setting parameters:

Proof By assumption, we have $\nu = N^{-\Omega(k)}$ and $\alpha = 2^{O(k)} \log N$. Let $0 < \beta < \frac{1}{2}$ be a small enough absolute constant and $c = N^{-\beta}$ such that

$$\alpha \cdot c^k = 2^{O(k)} \log(N) \cdot N^{-\beta k} \leqslant \tfrac{1}{2}\rho.$$

Then,

$$\nu^2 + \alpha \cdot c^k \leqslant N^{-\beta' k} = \rho.$$

Thus, by Lemma 4.3 any SQ algorithm that learns to up to error $\frac{1}{2} - 4\tau$ for $\tau = \sqrt{\rho}$ must either make queries of accuracy 2τ or must make at least

$$\exp\left(N^{-2\beta} \cdot \Omega(N) - \Omega(k \log N)\right) \cdot \frac{2^{-O(k)}}{\log N} = \exp\left(\Omega(N^{1-2\beta}) - \Omega(k \log N)\right)$$

queries. Since $k \leqslant N^{\gamma}$ for a sufficiently small γ, the above is at least $\exp\left(\Omega(N^{1-2\beta})\right) = \exp(N^{\Omega(1)})$.

Since we assumed that our SQ algorithm can make queries of accuracy $N^{\beta' k)} > 2\tau$, it follows that it needs at least $2^{\Omega(\sqrt{N})}$ queries. ■

We remark that we make the assumption that our SQ algorithm can make queries of accuracy $N^{-\Omega(k)}$ for the following reason: Lemma 4.3 of Nasser and Tiegel (2022) uses a reduction from an associated testing problem to learning, we believe this reduction needs at least one query of this high accuracy to work (the same applies to Diakonikolas and Kane (2022)). Such an assumption is not necessary to show hardness for the associated testing problem – which we believe still captures the essence of the learning problem.

Lemma 18 *Let $n \in \mathbb{N}, \varepsilon > 0$ and distributions D_n^0 and D_n^1 be such that there exists no T-time distinguisher with advatage at least ε between D_n^0 and D_n^1. Further, let $D_n^{1'}$ be a third distribution such that $\mathrm{TVD}(D_n^1, D_n^{1'}) = \mathrm{negl}(n)$. Then there exists no T-time distingiusher with advantage at least $\varepsilon - \mathrm{negl}(n)$ between D_n^0 and $D_n^{1'}$.*

Proof Suppose there exists a distinguisher $\mathcal{A}$ between D_n^0 and $D_n^{1'}$ with advantage at least $\varepsilon - \mathrm{negl}(n)$. Using this distinguisher to distinguish between D_n^0 and D_n^1 gives advantage

$$\left|\mathbb{P}_{x \sim D_n^0}\left(\mathcal{A}(x) = 0\right) - \mathbb{P}_{x \sim D_n^1}\left(\mathcal{A}(x) = 0\right)\right| \geqslant \left|\mathbb{P}_{x \sim D_n^0}\left(\mathcal{A}(x) = 0\right) - \mathbb{P}_{x \sim D_n^{1'}}\left(\mathcal{A}(x) = 0\right)\right| + \mathrm{negl}(n) \geqslant \varepsilon$$

which is a contradiction. ■

B.2. Missing Lemmas for Cryptographic Hardness

In this section, we will prove fact 13 restated below.

Fact 19 (Restatement of fact 13) *Let $n, m \in \mathbb{N}$ with $2^{o(n)} = m > n$, and let $\gamma, \beta, \varepsilon \in \mathbb{R}_{>0}$ such that $0 \leqslant \beta \leqslant \gamma, \beta = \frac{1}{\mathrm{poly}(n)}$. Assume that there is no $(T + \mathrm{poly}(n, m))$-time distinguisher between*

$$\mathrm{CLWE}\left(m, \gamma, \beta\right) \quad \textit{and} \quad \left(N\left(0, \tfrac{1}{2\pi} \cdot I_n\right) \times U\left([0, 1)\right)\right)^{\otimes m}$$

with advantage ε. Let $m' = \frac{m}{\mathrm{poly}(n)}$. Then there is no T-time distingiusher between

$$\frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}, +1\right) + \frac{1}{2} \cdot \left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}}, -1\right) \quad \textit{and} \quad N\left(0, \tfrac{1}{2\pi} \cdot I_n\right) \times \mathrm{Be}\left(\frac{1}{2}\right)$$

with advantage $\varepsilon - \mathrm{negl}(n)$ that uses at most m' samples.

Proof From (Tiegel, 2023, Theorem 15) we know that the conclusion is true for

$$\frac{1}{2}\cdot\left(\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0},+1\right)+\frac{1}{2}\cdot\left(\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}},-1\right)\quad\text{and}\quad N\left(0,\tfrac{1}{2\pi}\cdot I_n\right)\times\mathrm{Be}\left(\frac{1}{2}\right).$$

Our lemma follows by noting that the total variation distance between $\frac{1}{2}\cdot\left(\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0},+1\right)+\frac{1}{2}\cdot\left(\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}},-1\right)$ and $\frac{1}{2}\cdot\left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0},+1\right)+\frac{1}{2}\cdot\left(\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}},-1\right)$ is at most $2^{\Theta(-n)}$, even when considering their m-fold product for $m = 2^{o(n)}$. We can then imply theorem 18. By triangle inequality, it is enough to show that $\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0}$ and $\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}$ and $\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}}$ and $\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,\frac{1}{2}}$ satisfy this. Without loss of generality consider $\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0}$ and $\mathrm{NH}^{(n)}_{\boldsymbol{w},\beta,\gamma,0}$. Note that we can couple these two distributions as follows: We first draw a sample X from $\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0}$ if X comes from the central $2n+1$ components we set $X' = X$, else, we resample X' independently from $\mathrm{NH}^{(\infty)}_{\boldsymbol{w},\beta,\gamma,0}$ until it does. We output (X, X'). The marginals are correct by construction. Thus, the TVD is at most the probability that the first draw of X does not come from the central $2n+1$ components. This probability is at most

$$\frac{\sum_{|\ell|>n}\rho_{\sqrt{\beta^2+\gamma^2}}(\ell)}{\sum_{\ell=-\infty}^{\infty}\rho_{\sqrt{\beta^2+\gamma^2}}(\ell)}=\frac{\sum_{|\ell|>n}\exp\left(-\pi\ell^2/(\beta^2+\gamma^2)\right)}{\sum_{\ell=-\infty}^{\infty}\exp\left(-\pi\ell^2/(\beta^2+\gamma^2)\right)}\leqslant\sum_{|\ell|>n}\exp\left(-\pi\ell^2/(\beta^2+\gamma^2)\right)$$

$$\leqslant\sum_{|\ell|>n}\exp\left(-\frac{2\ell^2}{n}\right)\leqslant\exp\left(-\frac{n}{10}\right).$$

■