Optimal SQ Lower Bounds for Learning Halfspaces with Massart Noise

Rajai Nasser RAJAI.NASSER@INF.ETHZ.CH

Stefan Tiegel STEFAN.TIEGEL@INF.ETHZ.CH
Universitätstrasse 6, 8006 Zürich, Switzerland

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We give tight statistical query (SQ) lower bounds for learning halfspaces in the presence of Massart noise. In particular, suppose that all labels are corrupted with probability at most η. We show that for arbitrary $\eta \in [0, 1/2]$ every SQ algorithm achieving misclassification error better than η requires queries of superpolynomial accuracy or at least a superpolynomial number of queries. Further, this continues to hold even if the information-theoretically optimal error OPT is as small as $\exp(-\log^c(d))$, where d is the dimension and $0 < c < 1$ is an arbitrary absolute constant, and an overwhelming fraction of examples are noiseless. Our lower bound matches known polynomial time algorithms, which are also implementable in the SQ framework. Previously, such lower bounds only ruled out algorithms achieving error $\mathrm{OPT} + \varepsilon$ or error better than $\Omega(\eta)$ or, if η is close to $1/2$, error $\eta - o_\eta(1)$, where the term $o_\eta(1)$ is constant in d but going to 0 for η approaching $1/2$.

As a consequence, we also show that achieving misclassification error better than $1/2$ in the (A, α)-Tsybakov model is SQ-hard for A constant and α bounded away from 1.

Keywords: Massart Noise, PAC learning, Statistical query lower bounds

1. Introduction

Arguably one of the most fundamental problems in the area of machine learning and learning theory, going back to the Perceptron Algorithm Rosenblatt (1958), is the problem of learning halfspaces, or Linear Threshold Functions (LTFs): Fix $w \in \mathbb{R}^M$ and $\theta \in \mathbb{R}$, an LTF is a function $f\colon \mathbb{R}^M \to \{-1, 1\}$ such that $f(x) = 1$ if $\langle w, x\rangle \geqslant \theta$ and -1 otherwise. The associated learning problem is as follows: We observe samples (x, y) where $x \in \mathbb{R}^M$ is drawn from a fixed but unknown distribution D_x and $y \in \{-1, +1\}$ is, a possibly noisy version of, $f(x)$. We call x the *example* and y the *label*. Let D denote the joint distribution of (x, y), the goal is to output a hypothesis h such that the *misclassification error*

$$\mathrm{err}(h) \coloneqq \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

is minimized. For the purpose of this paper we consider the case where $\theta = 0$. In this work we make progress on a central question in the field: Identifying under which types of noise achieving small misclassification error is possible. On a conceptual level, we show that already as soon as only very few of the labels are flipped with some probability η, it is likely to be computationally hard to achieve error better than η. Even if the optimal error is much smaller than this.

© 2022 R. Nasser & S. Tiegel.

Realizable Case, Random Classification Noise, and Agnostic Model In the noiseless case, also called *realizable case*, it holds that $y = f(x)$ for all x. In this setting it is well-known that linear programming can achieve misclassfication error at most ε efficiently, i.e., in time polynomial in M and $\frac{1}{\varepsilon}$, and reliably, i.e., with probability close to 1, corresponding to Valiant's PAC model Valiant (1984). When considering noisy labels, the two most well-studied models are *Random Classification Noise* (RCN) Angluin and Laird (1988) and the *agnostic model* Haussler (1992); Kearns et al. (1994). In the former each sample (x, y) is generated by first drawing $x \sim D_x$ and then setting $y = f(x)$ with probability $1 - \eta$ and setting $y = -f(x)$ with probability η for some $\eta \in (0, 1) \setminus \{1/2\}$. It can be shown that in this model the information-theoretic optimal misclassification error is η and it is known how to efficiently find an LTF achieving misclassification error arbitrarily close to this Blum et al. (1998). However, one clear drawback is that the assumption that the magnitude of the noise is uniform across all examples is unrealistic. On the other extreme, in the agnostic model, no assumption whatsoever is placed on the joint distribution D. It is now believed that it is computationally hard to output any hypothesis that achieves error even slightly better than $1/2$. This holds even when the information-theoretic misclassification error is a function going to zero when the ambient dimension goes to infinity Daniely (2016). This is based on a hardness reduction to a problem widely believed to be computationally intractable.

A More Realistic Yet Computationally Tractable Noise Model Given the above results a natural question to ask is whether there exists a more realistic noise model in which it is still computationally tractable to achieve non-trivial guarantees. A promising candidate is the so-called *Massart noise model* which is defined as follows

Definition 1 *Let D_x be a distribution over $\mathbb{R}^M$ and let $f\colon \mathbb{R}^M \to \{-1, 1\}$ be an LTF. For $\eta \in [0, 1/2]$, we say that a distribution D over $\mathbb{R}^M \times \{-1, 1\}$ satisfies the* η-Massart noise condition *with respect to the* hypothesis *f and to the* marginal distribution *D_x if there exists a function $\eta : \mathbb{R}^M \to [0, \eta]$ such that samples $(x, y) \sim D$ are generated as follows: First, $x \sim D_x$ is drawn and then we output (x, y) where $y = f(x)$ with probability $1 - \eta(x)$ and $y = -f(x)$ with probability $\eta(x)$, i.e., $\eta(x)$ is the* flipping probability.

In the problem of learning halfspaces in the Massart noise model*, we observe samples $(x, y) \sim D$ from an unknown distribution D satisfying the η-Massart noise condition for some known bound $\eta \in [0, 1/2]$, and the goal is to output a hypothesis $h\colon \mathbb{R}^M \to \{-1, 1\}$ minimizing the* misclassification error

$$\mathrm{err}(h) \coloneqq \mathbb{P}_{(x,y)\sim D}[h(x) \neq y].$$

Note that the marginal distribution D_x, the true hypothesis $f\colon \mathbb{R}^M \to \{-1, 1\}$, and the flipping probability function $\eta : \mathbb{R}^M \to [0, \eta]$ are all unknown.

The model was proposed in Massart and Nédélec (2006).[1] Note that if $\eta(x) = \eta$ for all x, we obtain the Random Classification Noise model. As previously mentioned, the information-theoretically optimal error in the RCN model is equal to η. However, in the more general case of η-Massart noise, the information-theoretically optimal error is equal to

$$\mathrm{OPT} \coloneqq \mathbb{P}_{(x,y)\sim D}[f(x) \neq y] = \mathbb{E}\,\eta(x)\,,$$

1. Note that Rivest and Sloan (1994); Sloan (1996) introduced an equivalent model called "malicious misclassification noise".

which can potentially be much smaller than η. Information-theoretically, it was shown in Massart and Nédélec (2006) that for η bounded away from $1/2$, a number $n = O\left(\frac{M\log(1/\varepsilon)}{(1-2\eta)^2\varepsilon}\right)$ of samples suffices to achieve misclassification error $\mathrm{OPT}+\varepsilon$ and that this is tight up to constants. More generally, if the target halfspace is replaced by an unknown boolean function in a class of VC-dimension d, a number $n = O\left(\frac{d\log(1/\varepsilon)}{(1-2\eta)^2\varepsilon}\right)$ of samples suffices to achieve error $\mathrm{OPT}+\varepsilon$. [2]

However, until recently, algorithmic results were only known when assuming that the marginal distribution of the examples D_x belongs to some known class, e.g., is uniform or log-concave Awasthi et al. (2015, 2016); Zhang et al. (2017) or even more general in Diakonikolas et al. (2020). Under no assumption on the marginal distribution, Diakonikolas et al. (2019) was the first work that provided an efficient (improper) learning algorithm outputting a hypothesis h (which is not a halfspace) such that $\mathrm{err}(h) \leqslant \eta + \varepsilon$. They use time and sample complexities which are polynomial in M and $\frac{1}{\varepsilon}$. Building on this, Chen et al. (2020) provided an efficient (proper) learning algorithm with the same error guarantees but whose output is itself a halfspace. We remark that the sample complexity of both of the above works depends on the bit complexity of points in the support of D_x although this is information-theoretically not necessary. This assumption was recently removed in Diakonikolas et al. (2021a). Further, the above works assume $\eta < 1/2$. See Diakonikolas et al. (2021b) for a quasi-polynomial-time algorithmic result without this assumption but under Gaussian marginals.

On the other hand, until very recently, no matching computational lower bounds were known and it remained an open question to determine whether it is possible to efficiently achieve error guarantees that are better than η, potentially going all the way to OPT. This question is especially intriguing since the above algorithmic results imply that non-trivial guarantees can be achieved in the Massart noise model, which is much more realistic than RCN. The question then becomes if there are any computational limits at all in this model. As we will see, such limits do indeed exist, at least when restricting to the class of Statistical Query algorithms.

Statistical Query Algorithms and Known Lower Bounds. Statistical Query (SQ) algorithms do not have access to actual samples from the (unknown) distribution D but rather are allowed to query expectations of bounded functions over the underlying distribution. These queries return the correct value up to some accuracy. Since every such query can be simulated by samples from the distribution this is a restriction of Valiant's PAC model. Note that a simple Chernoff bound shows that in order to simulate a query of accuracy τ, a number of $O(1/\tau^2)$ samples is sufficient. Hence, SQ algorithms using N queries of accuracy at most τ can be taken as a proxy for algorithms using $O(1/\tau^2)$ samples and running in time $\mathrm{poly}(N, 1/\tau)$. The SQ model was originally introduced by Kearns (1998). See Feldman (2016) for a survey. Note, that it has also found applications outside of PAC learning, see e.g., Kasiviswanathan et al. (2011); Feldman et al. (2021) for examples.

Intriguingly, Kearns (1998) shows that any concept class that is PAC learnable in the realizable case using an SQ algorithm can also be learned in the PAC model under Random Classification Noise. Further, almost all known learning algorithms are either SQ or SQ-implementable, except for those that are based on Gaussian elimination, e.g., learning parities with noise Kearns (1998); Blum et al. (2003). One clear advantage of this framework is that it is possible to prove unconditional

2. We remark that previous works on algorithmic aspects of the Massart model stated this sample complexity as $O(d/\varepsilon^2)$. While this is correct, from Massart and Nédélec (2006) it follows that this is only necessary when $\eta \geqslant \frac{1}{2}\cdot(1-\sqrt{d/n})$. For η smaller than this the bound of $O\left(\frac{d\log(1/\varepsilon)}{(1-2\eta)^2\varepsilon}\right)$ holds.

lower bounds. This proceeds via the so-called *SQ dimension* first introduced in Blum et al. (1994) and later refined in Feldman et al. (2017); Feldman (2017). Although we will not see it explicitly, the lower bounds in this paper are also based on this parameter. See Diakonikolas and Kane (2021) and the references therein for more detail.

For learning halfspaces under Massart noise, Chen et al. (2020) initiated the study of computational lower bounds. The authors proved that when OPT is within a factor of 2 of η, achieving error $\mathrm{OPT}+\varepsilon$ requires superpolynomially many queries. While this shows that obtaining optimal error is hard, it does not rule out the possibility of an efficient (SQ) algorithm achieving constant factor approximations. More recently Diakonikolas and Kane (2021) proved that for $\tau = M^{-\omega(1)}$, achieving error better than $\Omega(\eta)$ requires queries of accuracy better than τ or at least $1/\tau$ queries. This holds even when η is a constant but OPT goes to zero as the ambient dimension M becomes large. This rules out any constant factor approximation algorithm, and also rules out efficient algorithms achieving error $O(\mathrm{OPT}^c)$ for any $c < 1$. Further, for η close to $1/2$ the authors show that achieving error that is better than $\eta - o_\eta(1)$ for some term $o_\eta(1)$ that is constant in M, but depends on η and goes to 0 as η goes to 1/2, also requires superpolynomial time in the SQ framework. For the special case of $\eta = 1/2$, Diakonikolas et al. (2021b) shows that achieving error $\mathrm{OPT}+\varepsilon$ requires queries of accuracy better than $d^{-\Omega(\log(1/\varepsilon))}$ or at least $2^{d^{\Omega(1)}}$ queries even under Gaussian marginals. However, as with Chen et al. (2020), this result only applies to exact learning.

As can be seen, the best previously known lower bounds are a constant factor away from the best known algorithmic guarantees, but they do not match yet. In the present work, we close this gap by showing that the algorithmic guarantees are actually tight, at least in the SQ framework. More precisely, we will show that for arbitrary $\eta \in (0, 1/2]$ any SQ algorithms that achieves error better than η either requires a superpolynomial number of queries, or requires queries of superpolynomial accuracy. Further, as for Diakonikolas and Kane (2021) the result holds even when OPT goes to zero as a function of the ambient dimension M and η is a constant less or equal to 1/2.

1.1. Results

The following theorem is our main result (see Theorem 3 for a more detailed version):

Theorem 2 (Informal version) *Let $M \in \mathbb{R}$ be sufficiently large and $\eta \in (0, 1/2]$ be arbitrary. There exists no SQ algorithm that learns M-dimensional halfspaces in the η-Massart noise model to error better than η using at most $\mathrm{poly}(M)$ queries of accuracy no better than $1/\mathrm{poly}(M)$.*

This holds even if the optimal halfspace achieves error OPT that vanishes as fast as $2^{-(\log M)^c}$ for some $c < 1$, and even if we assume that all flipping probabilities are either 0 or η.

Some remarks are in order:

- As we mentioned earlier, this lower bound matches the guarantees that are achievable in polynomial time Diakonikolas et al. (2019); Chen et al. (2020). Moreover, since these algorithms can be implemented in the SQ learning model, this completely characterizes the error guarantees that are efficiently achievable in the SQ framework for the class of halfspaces under Massart noise. Further, this also suggests that improving over this guarantee with efficient non-SQ algorithms might be hard.

- For the special case $\eta = 1/2$, the result implies that handling $1/2$-Massart noise is as hard as the much more general agnostic model – again for the class of halfspaces and in the SQ

model. Namely, it is hard to achieve error better than a random hypothesis. Note that even though $\eta = 1/2$ means that there can be examples x with completely random labels, the fact that OPT can be made go to zero implies that there would be a vanishing fraction of such examples. We remark that Daniely gave a similar SQ lower bound for the agnostic model Daniely (2016).

- The fact that hardness still holds even if for all x we have $\eta(x) \in \{0, \eta\}$ and even if OPT is very small implies that achieving error better than η remains hard even if an overwhelming fraction of the samples have no noise in their labels. In light of the previous point this implies that even if the overwhelming majority of the points have no noise but the labels of just very few are random, outputting a hypothesis which does better than randomly classifying the points is SQ-hard.
- The case when $\mathrm{OPT} = O\left(\log(M)/M\right)$ is computationally easy. This follows since with high probability there is a subset of the observed samples in which no labels were flipped and which is sufficiently large to apply algorithms designed for the realizable case. Hence, for values of OPT only slightly smaller than allowed by Theorem 2 achieving optimal misclassfication error is possible in polynomial time.
- In previous works the probability p_+ of observing a $(+1)$-label was a function of η. While the same will be true for our construction, in fact, we will have $p_+ = 1 - \eta$, we show in Section 5 how to decouple these and show that for a fixed η the learning problem remains hard for all $p_+ \in [\eta, 1-\eta]$. Note that since one of the constant functions $+1$ or -1 trivially achieves error $\min\{p_+, 1-p_+\}$, this will be smaller than η for p_+ outside this range.

As a consequence of the above theorem, we immediately obtain strong hardness results for a more challenging noise model, namely the Tsybakov noise model Mammen and Tsybakov (1999); Tsybakov (2004) defined as follows: Let $A > 0$ and $\alpha \in [0, 1)$. Samples are generated as in the Massart model but the flipping probabilites $\eta(x)$ are not uniformly bounded by some constant but rather need to satisfy the following condition:

$$\forall 0 < t \leqslant 1/2 \colon \mathbb{P}[\eta(x) \geqslant 1/2 - t] \leqslant A \cdot t^{\alpha/(1-\alpha)} \,.$$

It is known that information-theoretically $O\left(\frac{A \cdot M}{\varepsilon^{2-\alpha}} \cdot \log(1/(A\varepsilon^{\alpha}))\right)$ samples suffice to learn halfspaces up to misclassification error $\mathrm{OPT} + \varepsilon$ in this model (Hanneke et al., 2014, Chapter 3). On the other hand, algorithmic results are only known when restricting the marginal distribution to belong to a fixed class of distributions (e.g., log-concave or even more general Diakonikolas et al. (2021c)). On the other hand, we claim that our hardness result about Massart noise implies that it is SQ-hard to achieve error even slightly better than $1/2$ in the Tsybakov model. Indeed, let $\zeta = 2^{-(\log M)^c}$ for some $0 < c < 1$. Further, let A be a constant, $\alpha \in (0, 1)$ be bounded away from 1, and

$$\eta = \frac{1}{2} - \left(\frac{\zeta}{A}\right)^{(1-\alpha)/\alpha} = \frac{1}{2} - \exp\left(-\Theta((\log M)^c)\right) \,.$$

Then the η-Massart condition together with the condition that $\eta(x) \in \{0, \eta\}$ and $\mathrm{OPT} = \zeta$ implies the (A, α)-Tsybakov condition. To see this note that for $t \geqslant 1/2 - \eta$ we obtain that

$$\mathbb{P}[\eta(x) \geqslant 1/2 - t] \leqslant \mathbb{P}[\eta(x) \geqslant 0] = \zeta = A \cdot \left(\frac{1}{2} - \eta\right)^{\alpha/(1-\alpha)} \leqslant A \cdot t^{\alpha/(1-\alpha)} \,,$$

and for $t < 1/2 - \eta$ we have

$$\mathbb{P}[\eta(x) \geqslant 1/2 - t] \leqslant \mathbb{P}[\eta(x) > \eta] = 0 \leqslant A \cdot t^{\alpha/(1-\alpha)} .$$

Hence, by Theorem 2, or Theorem 3, achieving error better than $1/2 - \exp\left(-\Theta((\log M^c)\right)$ requires queries of accuracy better than the inverse of any polymonial or at least superpolynomially many queries, even though $\mathrm{OPT} = 2^{-(\log M)^c}$. Similarly, for $\alpha = 1 - 1/\log(M)^{c'}$ where $0 < c' < c$ it is hard to achieve error better than $1/2 - \exp\left(-\Theta(((\log M)^{c-c'})\right)$ in the sense above. This stands in strong contrast to the fact that information-theoretically $\mathrm{poly}(M, 1/\varepsilon)$ samples and time suffice to achieve misclassification optimal error. That is, even if the fraciton of flipping probabilites decreases very fast as we approach 1/2 learning in the model remains hard.

Lastly, we would like to mention that we closely follow the techniques developed in Diakonikolas and Kane (2021) (and previous works cited therein). At the heart of their work one needs to design two distributions matching many moments of the standard Gaussian (and satisfying some additional properties). The main difference in our work lies in how exactly we construct these distributions, which eventually leads to the tight result.

2. Techniques

In this section, we will outline the techniques used to prove Theorem 2. On a high level, we will closely follow the approach of Diakonikolas and Kane (2021). First, note that for $M, m, d \in \mathbb{N}$ satisfying $M = \binom{m+d}{m} \leqslant m^d$, any degree-$d$ polynomial over $x \in \mathbb{R}^m$ can be viewed as a linear function[3] over $\mathbb{R}^M$. Hence, any lower bound against learning polynomial-threshold functions (PTFs) in $\mathbb{R}^m$ would yield a lower bound against learning halfspaces in $\mathbb{R}^M$. Further, if we choose m, d so that $m \approx \log(M)^{1+\alpha}$ for some constant $\alpha > 0$, then an exponential lower bound against learning PTFs in $\mathbb{R}^m$ would yield a superpolynomial lower bound against learning halfspaces in $\mathbb{R}^M$.

One key step of the SQ hardness result in Diakonikolas and Kane (2021) is to construct two specific distributions over $(x, y) \in \mathbb{R}^m \times \{-1, 1\}$ and show that a mixture of these two distributions is SQ-hard to distinguish from a certain null distribution[4]. The authors then argue that any algorithm that learns η-Massart PTFs up to error better than $\Omega(\eta)$ can be used to distinguish these distributions from the null distribution. We follow a similar proof strategy. The main difference lies in how we construct the two hard distributions (in a simpler way), allowing us to obtain the optimal lower bound η. In fact, we will show that two simple modifications of the standard Gaussian distribution will work.

Both distributions constructed in Diakonikolas and Kane (2021) as well as the ones that we will construct have the following common structure: Let $v \in \mathbb{R}^m$ be fixed but unknown, $q \in (0, 1)$, and let A, B be two one-dimensional distributions. Define D_+ (respectively, D_-) as the distribution over $\mathbb{R}^m$ that is equal to A (respectively, B) in the direction of v and equal to a standard Gaussian in the orthogonal complement. Then, define the distribution D over $\mathbb{R}^m \times \{-1, 1\}$ as follows: With probability q draw $x \sim D_+$ and return $(x, 1)$, and with probability $1 - q$ draw $x \sim D_-$ and return $(x, -1)$. The goal is to output a hypothesis h minimizing the misclassification error

3. We use an embedding $\mathbb{R}^m \to \mathbb{R}^M$ whose component functions are the monomials of degree $\leqslant d$.
4. The null distribution is the one where the example $x \in \mathbb{R}^m$ is standard Gaussian and the label $y \in \{-1, 1\}$ is independent from x.

$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$. It is easy to see that one of the constant functions 1 or -1 achieves error $\min\{q, 1-q\}$. The question is whether it is possible to achieve error better than $\min\{q, 1-q\}$.

Roughly speaking[5], the authors of Diakonikolas and Kane (2021) show the following hardness result: Suppose the first k moments of A and B match those of $N(0,1)$ upto additive error at most 2^{-k} and their χ^2-divergence with respect to $N(0,1)$ is not too large. Then every SQ algorithm outputting a hypothesis achieving misclassification error slightly smaller than $\min\{q, 1-q\}$ must either make queries of accuracy at least $2^{-k/2}$ or must make at least 2^{m-k} queries. Hence, if we can choose k to be a small constant multiple of m we get an exponential lower bound as desired. The authors then proceed to construct distributions satisfying the moment conditions with $\min\{q, 1-q\} = \Omega(\eta)$ and such that D corresponds to an η-Massart PTF. In this paper, we construct distributions satisfying the moment conditions with $\min\{q, 1-q\} = \eta$. However, the χ^2-divergence will be too large to apply the hardness result of Diakonikolas and Kane (2021) in a black-box way. To remedy this, we show that its proof can be adapted to also work in this regime. Further, by choosing the parameters slightly differently, the reduction still works. In the following, we briefly describe our construction. We will give a more detailed comparison with Diakonikolas and Kane (2021) in Section 2.1.

Let $0 < \eta \leqslant 1/2$ be the bound of the Massart model and fix $q = 1-\eta$. We will show that we can choose A and B satisfying the moment conditions above, in such a way that D corresponds to an η-Massart PTF. Note that this will directly imply Theorem 2 via the previously outlined reduction. We partition $\mathbb{R}$ into three regions J_1, J_2 and $\mathbb{R} \setminus (J_1 \cup J_2)$ such that the following conditions hold:

1. $A(x) = 0$ for $x \in J_2$,
2. $B(x) = 0$ for $x \in J_1$,
3. $A(x) \geqslant B(x)$ for all $x \in \mathbb{R} \setminus (J_1 \cup J_2)$.

Suppose that J_2 can be written as the union of d intervals and hence there is a degree-$2d$ polynomial $p : \mathbb{R} \to \mathbb{R}$ which is non-negative on $\mathbb{R} \setminus J_2$ and non-positive on J_2. We claim, that D is an η-Massart PTF for the polynomial $p_v : \mathbb{R}^m \to \mathbb{R}$ defined as

$$p_v(x) = p(\langle v, x\rangle)\,.$$

Let $D_x(x) \coloneqq \sum_y D(x,y)$ be the marginal distribution of D on x. Then this means that for all $x \in \mathbb{R}^m$, such that $D_x(x) > 0$ it needs to hold that

$$\eta(x) \coloneqq \mathbb{P}_{(x,y)\sim D}\left[y \neq \operatorname{sign}(p_v(x)) \mid x\right] \leqslant \eta\,.$$

Indeed, consider x such that $\langle x, v\rangle \in J_1$. Since $p_v(x) \geqslant 0$ and $B(\langle x, v\rangle) = 0$ it follows that $\eta(x) = 0$. On a high level, this is because none of the samples with label -1 lie in this region. Similarly, the same holds for x such that $\langle x, v\rangle \in J_2$. Now consider $x \in \mathbb{R}^m$ such that $D_x(x) > 0$ and $\langle x, v\rangle \in \mathbb{R} \setminus (J_1 \cup J_2)$. Since $\operatorname{sign}(p_v(x)) = 1$ and $A(x) \geqslant B(x)$ it follows

$$\begin{aligned}\mathbb{P}_{(x,y)\sim D}\left[y \neq \operatorname{sign}(p(x)) \mid x\right] = \frac{\mathbb{P}_{(x,y)\sim D}[y \neq \operatorname{sign}(p(x)), x]}{D_x(x)} &= \frac{(1-q)\cdot B(x)}{q\cdot A(x) + (1-q)\cdot B(x)} \\ &\leqslant 1 - q = \eta\,.\end{aligned} \tag{2.1}$$

5. This sweeps under the rock some details, see Section 4 for all details.

Note that in our cosntruction it will actually hold that $A(x) = B(x)$ for all $x \in \mathbb{R} \setminus (J_1 \cup J_2)$. Hence, it even holds that $\eta(x) \in \{0, \eta\}$ for all x. Note that this is also why we needed the condition $A(x) \geqslant B(x)$ since otherwise the η-Massart condition would be violated.

Our work crucially departs from Diakonikolas and Kane (2021) in our choice of A and B to satisfy Conditions 1 to 3 and the moment-matching condition. In fact, giving a very clean construction will turn out to be essential for achieving the tightest possible lower bound. On a high level, A will be equal to an appropriate multiple of the standard Gaussian distribution on periodically spaced intervals of small size and 0 otherwise. B will be equal to A for x of large magnitude. For smaller x we will slightly displace the intervals.

Concretely, let $0 < \delta, \varepsilon < 1$ be such that $\varepsilon < \delta/8$ and consider the infinte union of intervals

$$J = \bigcup_{n \in \mathbb{Z}} [n\delta - \varepsilon, n\delta + \varepsilon] .$$

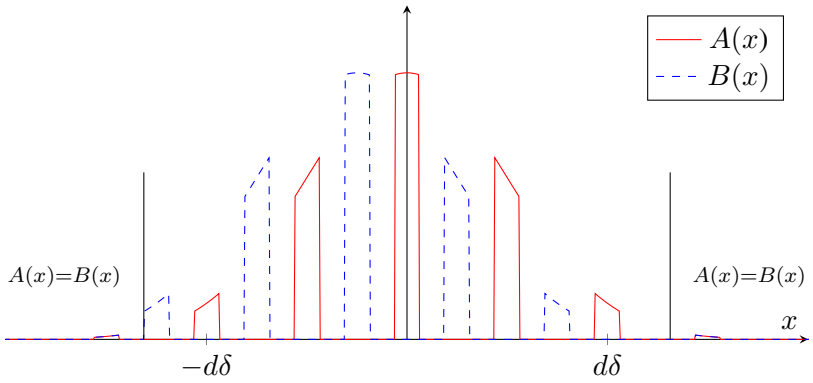
Denote by G the pdf of a standard Gaussian distribution. We define (the unnormalized measures)

$$A(x) = \begin{cases} \frac{\delta}{2\varepsilon} \cdot G(x) , & \text{if } x \in J , \\ 0 , & \text{otherwise.} \end{cases} \qquad B(x) = \begin{cases} A(x) , & \text{if } |x| > d\delta + 5\varepsilon , \\ A(x + 4\varepsilon) , & \text{otherwise.} \end{cases}$$

Clearly, the total probability mass of the two is the same. It can be shown that it is $1 \pm \exp(-\Omega(1/\delta)^2)$, so for the sake of this exposition assume that it is exactly one and that A and B are in fact probability distributions (see Appendix B for all details). Further, consider

$$J_1 = \bigcup_{n=-d}^{d} [n\delta - \varepsilon, n\delta + \varepsilon] , \qquad J_2 = \bigcup_{n=-d}^{d} [n\delta - 5\varepsilon, n\delta - 3\varepsilon] .$$

It is not hard to verify that A, B together with J_1, J_2 satisfy Conditions 1 to 3. Hence, our final distribution D will satisfy the Massart condition. Since $\eta(x) \neq 0$ only if $A(x) = B(x) > 0$ which only is the case when $|x| \gtrsim d\delta$ it follows that OPT is very small as well.[6]



The fact that the moments of A match those of a standard Gaussian will follow from the fact that it is obtained by only slightly modifying it. This part is similar to Diakonikolas and Kane (2021). Note that B is equal to A for x of magnitude larger than roughly $d\delta$ and for smaller x is

6. Note, that here J_2 is the union of $2d + 1$ intervals. It is straightforward to adapt the previous discussion to this case.

obtained by displacing A by ε. Hence, it will follow that its first k moments match those of A (and hence also those of a standard Gaussian) up to error $\varepsilon(d\delta)^k$. In Section 4, we will show that we can choose the parameters such that for k slightly smaller than m we can make the first k moments of A and B match those of a standard Gaussian up to error at most roughly $\exp(-\Omega(m))$ which will be sufficient.

2.1. Comparison with Diakonikolas and Kane (2021)

The key property that allowed us to achieve the sharp lower bound of η was that $A(x) \geqslant B(x)$ on $\mathbb{R} \setminus (J_1 \cup J_2)$. Indeed, if we only had $A(x) \geqslant c \cdot B(x)$ for some constant $0 < c < 1$, the resulting distribution D would no longer be η-Massart (cf. Eq. (2.1)), and the only way to still make it so is to increase q which in turn degrades the resulting lower bound. More precisely, if we only have $A(x) \geqslant c \cdot B(x)$, then the upper bound in Eq. (2.1) will now be $\frac{1-q}{c\cdot q+1-q}$ instead of $1-q$. Basic manipulations show that this is less than or equal to η if and only if $q \geqslant \frac{1}{1-\eta(1-c)} \cdot (1-\eta) > 1-\eta$, which means that the lower bound that we get from the distinguishing problem is at best $\min\{q, 1-q\} = \Omega(\eta)$.

While our construction can avoid this issue because we can ensure that $A(x) \geqslant B(x)$ for $x \notin J_1 \cup J_2$ (in fact, we will have $A(x) = B(x)$), it is unclear if the same can be achieved using the construction of Diakonikolas and Kane (2021), or a slight modification of it. In their work, the supports of A and B also consist of unions of intervals, but they increase in size as we move away from the origin. The intervals of A are disjoint from those of B for x of small magnitude, but they start to overlap when $|x|$ becomes large. On each interval the distribution is also a constant multiple of $G(x)$, however, their specific choice makes exact computations difficult and the authors only show that $A(x) \geqslant \Omega(B(x))$ where the constants in the Ω-notation can be smaller than 1. [7] We note, however, that the moment bounds the authors use for their distribution are very similar to the one we use for our distribution A.

On a more technical level, we cannot directly apply the hardness result (Diakonikolas and Kane, 2021, Proposition 3.8) the authors used. Suppose the first k moments of A and B match those of a standard Gaussian up to additve error ν and the χ^2-divergence of A and B with respect to the standard Gaussian is at most $\alpha/2$. Further, let

$$\tau = \nu^2 + 2^{-k}\alpha\,, \qquad\qquad N = 2^{\Omega(m)}\tau/\alpha\,.$$

Then this result says that for every SQ algorithm achieving misclassification error better than $\min\{q, 1-q\} - 4\sqrt{\tau}$ must either make queries of accuracy better than $2\sqrt{\tau}$ or must make at least N queries. Since in our construction we need to choose ε sufficiently small to match many moments — which in turn will increase the χ^2-divergence — we will have $\alpha \gg 2^k$ which is too large for the above. On the flip side, the proof of (Diakonikolas and Kane, 2021, Proposition 3.8) can readily be adapted (in fact, this is already implicit in the proof) so that the same conclusion also holds for

$$\tau = \nu^2 + c^k\alpha\,, \qquad\qquad N = 2^{c^2\cdot\Omega(m)}\tau/\alpha\,,$$

for some arbitrarily small c where the constant in $\Omega(m)$ is independent of c. It will turn out that we can choose c sufficiently small and in turn m slightly larger so that the above yields the desired bounds. See Section 4 and Appendix C for an in-depth discussion.

7. We remark that the authors do not work with distributions directly but with unnormalized measures. Normalizing them does not change the construction but makes the comparison easier.

3. Preliminaries

For two functions $f, g\colon \mathbb{R} \to \mathbb{R}$, we will write $f \ll g$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. Similarly, we will write $f \gg g$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \infty$.

All logarithms will be to the base e.

We will use $N(0,1)$ to denote the one-dimensional standard Gaussian distribution. We will denote its pdf by G and with a slight abuse of notation we will also refer to a standard Gaussian random variable by G.

For two probability distribution A and B we denote their χ^2-divergence by $\chi^2(A,B) = \int_{-\infty}^{\infty} \frac{A(x)^2}{B(x)}\,dx - 1$. For an unnormalized positive measure A we denote its total measure by $\|A\|_1$.

4. Hardness Result

In this section, we will prove the full version of Theorem 2. Concretely, we will show

Theorem 3 *Let $0 < \zeta \leqslant \eta \leqslant \frac{1}{2}$ and $M \in \mathbb{N}$ be such that $l := \frac{\log M}{(\log\log M)^3 \log(1/\zeta)}$ is at least a sufficiently large constant. There exists a parameter $\tau := M^{-\Theta(l)}$ for which there is no SQ algorithm that learns the class of halfspaces over $\mathbb{R}^M$ with η-Massart noise using at most $1/\tau$ queries of accuracy τ and which achieves misclassification error that is better than $\eta - \tau$. This holds even if the optimal halfspace has misclassification error that is as small as ζ and all flipping probabilites are either 0 or η.*

Note that $\zeta = 2^{-\log(M)^c}$ with $0 < c < 1$ satisfies the assumption of the theorem and we recover Theorem 2. As previously mentioned, the setting is the same as in Diakonikolas and Kane (2021) except that we achieve the tight lower bound of η.

In order to prove this, we will give a formal definition of the pair of hard distributions described in Section 2 and the properties which they need to satisfy, see Appendix B for all details. Further, we will show how these can be used to prove Theorem 3. Again, we refer the reader to Appendix A for all details. On a high level, this will be done via a reduction to the following classification problem, which was introduced in Diakonikolas and Kane (2021), and then applying a lower bound that was proved in the same reference.

Definition 4 (Hidden Direction Classification Problem) *Let A, B be two probability distributions over $\mathbb{R}$, let $p \in (0,1)$, and v be a unit vector in $\mathbb{R}^m$. Let D_+ (respectively D_-) be the distribution that is equal to A (respectively B) in the direction of v and equal to a standard Gaussian in its orthogonal complement. Consider the distribution $D_v^{A,B,p}$ on $\mathbb{R}^m \times \{-1,1\}$ defined as follows: With probability p draw $x \sim D_+$ and output $(x,1)$, and with probability $1-p$ draw $x \sim D_-$ and return $(x,-1)$. The* Hidden Direction Classification Problem *is the following: Given sample access to $D_v^{A,B,p}$ for a fixed but unknown v, output a hypothesis $h\colon \mathbb{R}^m \to \{-1,1\}$ (approximately) minimizing $\mathbb{P}_{(x,y)\sim D_v^{A,B,p}}[h(x) \neq y]$.*

Misclassification error $\min\{p, 1-p\}$ can trivially be achieved by one of the constant functions 1 or -1. The following lemma shows that in the SQ framework, one cannot do better if the distributions A and B (approximately) match many moments of the standard Gaussian distribution. Its proof is analogous to the one of Proposition 3.8 in Diakonikolas and Kane (2021). See Appendix C for a more detailed discussion.

Lemma 5 (Adaptation of Proposition 3.8 in Diakonikolas and Kane (2021)) *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 additive 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 $D_v^{A,B,p}$ for a fixed but unknown $v \in \mathbb{R}^m$, outputs a hypothesis $h \colon \mathbb{R}^m \to \{-1, 1\}$ such that*

$$\mathbb{P}_{(x,y)\sim D_v^{A,B,p}}[h(x) \neq y] < \min\{p, 1-p\} - 4\sqrt{\rho}\,,$$

must either make queries of accuracy better than $2\sqrt{\rho}$ or make at least $N = 2^{c^2 \cdot \Omega(m)} \cdot (\rho/\alpha)$, where the constant in the $\Omega(\cdot)$ is independent of c.

We will show that the distributions A, B introduced in Section 2 satisfy the conditions of Lemma 5 and are such that the distribution $D_v^{A,B,p}$ fits the Massart noise model. Let $0 < \delta, \varepsilon, d$ be to be fixed later. It is not hard to see that the construction described in Section 2 is equivalent to the following. Consider the positive measures

$$G_{\delta,\varepsilon}(x) = \sum_{n\in\mathbb{Z}} G(x) \cdot \left(\frac{\delta}{2\varepsilon}\right) \cdot \mathbf{1}_{[x\in[n\delta-\varepsilon, n\delta+\varepsilon]]}\,, \qquad G_{\delta,\varepsilon}^{(n)}(x) = \frac{G_{\delta,\varepsilon}(x)}{\|G_{\delta,\varepsilon}\|_1}\,,$$

note that $G_{\delta,\varepsilon}^{(n)}$ is in fact a probability measure, and define

$$A(x) = G_{\delta,\varepsilon}^{(n)}(x)\,, \qquad B(x) = \begin{cases} A(x)\,, & \text{if } |x| > d\delta + 5\varepsilon\,, \\ A(x+4\varepsilon)\,, & \text{otherwise.} \end{cases} \tag{4.1}$$

Further, as in Section 2, consider

$$J_1 = \bigcup_{-d\leqslant n\leqslant d} [n\delta - \varepsilon, n\delta + \varepsilon]\,, \qquad J_2 = \bigcup_{-d\leqslant n\leqslant d} [n\delta - 5\varepsilon, n\delta - 3\varepsilon]\,. \tag{4.2}$$

We will show that they satisfy the following set of properties

Proposition 6 *Let $0 < \zeta < 1/2$ and let $d \geqslant 2$ be an integer. Assume that $\delta = 4\sqrt{\log(1/\zeta)}/d$ and $\varepsilon < \delta/8$. If $\delta < 1$, the distributions A, B defined in Eq. (4.1) together with the intervals J_1, J_2 defined in Eq. (4.2) satisfy the following*

1. *$J_1 \cap J_2 = \emptyset$ and $J_1 \cup J_2 \subseteq [-d\delta - 5\varepsilon, d\delta + 5\varepsilon]$,*
2. *(a) $A = 0$ on J_2, $B = 0$ on J_1, and (b) for all $x \notin J_1 \cup J_2$ we have $A(x) = B(x)$,*
3. *for all $k \in \mathbb{N}$ the first k moments of A and B match those of a standard Gaussian within additive error $\nu = O(k!) \cdot \exp(-\Omega(1/\delta^2)) + 4\varepsilon\left(12\sqrt{\log(1/\zeta)}\right)^k$,*
4. *at most a ζ-fraction of the measure A (respectively B) lies outside J_1 (respectively J_2),*
5. *$\chi^2(A, N(0,1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$ and $\chi^2(B, N(0,1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$.*

The first two properties are straightforward. For distribution A the third will follow using a Fourier argument since the pdf of A corresponds to that of a standard Gaussian multiplied with a periodic rectangle function (cf. Fact 15 in Appendix D) and for B since it is only a small displacement of A. Further, properties 4 and 5 will follow by using standard properties of the Gaussian distribution, such as basic tail bounds. We refer the reader to Appendix B for detailed proofs.

Further, similar as outlined in Section 2, for a given paramter m, we can show that the distribution $D_v^{A,B,p}$ with $p = 1 - \eta$ is a degree-d η-Massart PTF with $\mathrm{OPT} \leqslant \zeta$ over $\mathbb{R}^m$. In addition, we can choose m and the parameters δ, ε, d in relation to M such that N and ρ in Lemma 5 are superpolynomial and the inverse of a superpolynomial function in M, respectively. This, together with Proposition 6, immediately yields Theorem 3. More precisely, in Appendix A we will show that the following choice of parameters works. Let $0 < \zeta < \eta$ be such that

$$\frac{\log M}{(\log\log M)^3} \geqslant C_\zeta \log(1/\zeta)$$

for a sufficiently large constant C_ζ. Further, let

$$\tau = M^{-\frac{\log M}{C_\tau (\log\log M)^3 \log(1/\zeta)}} .$$

In addition, for C_m and C_d sufficiently large constants, set

$$m = \left\lceil C_m \log(1/\tau) \log(1/\zeta)^4 \right\rceil , \qquad d = \left\lceil C_d \sqrt{\log(1/\zeta) \log(1/\tau) \log\log(1/\tau)} \right\rceil .$$

Lastly, let

$$\delta = \frac{4\sqrt{\log(1/\zeta)}}{d} = \Theta\left(\frac{1}{C_d \sqrt{\log(1/\tau) \log\log(1/\tau)}} \right)$$

and

$$k = \frac{4 \log(1/\tau)}{\log\log(1/\zeta)} , \qquad \varepsilon = \tau \cdot \left(12\sqrt{\log(1/\zeta)}\right)^{-k} , \qquad c = \frac{1}{144 \log(1/\zeta)^2} .$$

This choice implies that $\nu^2 + \alpha \cdot c^k \leqslant \tau$, where

$$\alpha = \chi^2(A, N(0,1)) + \chi^2(B, N(0,1)) = O\left(\frac{\delta}{\varepsilon}\right)^2 .$$

Hence, we can choose

$$\rho = \tau = M^{-\Theta\left(\frac{\log M}{C_\tau (\log\log M)^3 \log(1/\zeta)}\right)} .$$

Further, this implies

$$N = \frac{2^{c^2 \cdot \Omega(m)} \cdot \rho}{\alpha} \geqslant \Omega(\varepsilon^2) \cdot \exp\left(c^2 \cdot \Omega(m) - \log(1/\tau)\right) \geqslant 1/\tau^{\Theta(1)} .$$

We again refer the reader to Appendix A for all details.

5. Decoupling the Level of Massart Noise and the Probability of Observing a $(+1)$-label

Note that for every $\eta \in (0, 1/2]$, in the construction used to prove Theorem 3, the marginal probability p_+ of observing a $(+1)$-label is equal to $1 - \eta$. In this section, we show that the hardness result continues to hold for arbitrary $p_+ \in [\eta, 1 - \eta]$. Note that for p_+ outside this range it is indeed possible to achieve error better than η, since one of the constant functions $+1$ or -1 trivially achieves error $\min\{p_+, 1 - p_+\}$.

Theorem 7 *Let $\eta \in (0, 1/2]$ be fixed and $p_+ \in [\eta, 1-\eta]$ be arbitrary. The conclusion of Theorem 3 continues to hold if the marginal probability of observing a $(+1)$-label is equal to p_+.*

Proof Let D be the hard-to-learn distribution over $\mathbb{R}^M$ and assume that it corresponds to an η-Massart corruption of the halfspace corresponding to $w \in \mathbb{R}^M$. By our construction we know that

$$p_+(D) \coloneqq \mathbb{P}_{(x,y)\sim D}[y = +1] = 1 - \eta\,.$$

Consider now the following distribution D' over $\mathbb{R}^{2M}$: Let $q \in [0, 1]$ and $\mathbf{0} \in \mathbb{R}^M$ be the M-dimensional all-zeros vector. First, draw a sample (x, y) from D and then output

$$(x', y') = \begin{cases} ((x, \mathbf{0}), y)\,, & \text{with probability } q, \\ ((\mathbf{0}, x), -y)\,, & \text{with probability } 1 - q. \end{cases}$$

Clearly, it holds that

$$\begin{aligned} p_+(D') &\coloneqq \mathbb{P}_{(x',y')\sim D'}[y' = +1] = q \cdot \mathbb{P}_{(x,y)\sim D}[y = +1] + (1 - q) \cdot \left(1 - \mathbb{P}_{(x,y)\sim D}[y = +1]\right) \\ &= q \cdot (1 - \eta) + (1 - q) \cdot \eta \in [\eta, 1 - \eta]\,. \end{aligned}$$

Further, all of the values in the interval can be obtained by varying q. We claim that D' corresponds to an η-Massart halfspace with respect to $(w, -w) \in \mathbb{R}^{2M}$. Denote by $D'_{x'}$ the marginal of D' over $\mathbb{R}^{2M}$ and by D_x the marginal of D over $\mathbb{R}^M$. Indeed, each $x' \in \mathbb{R}^{2M}$ lying in the support of $D'_{x'}$ is of the form $(x, \mathbf{0})$ or $(\mathbf{0}, x)$ for $x \in \mathbb{R}^M$. Further, it holds that

$$\begin{aligned} \eta\left((x, \mathbf{0})\right) &= \mathbb{P}_{(x',y')\sim D'}\left[y' \neq \operatorname{sign}\left((w, -w)^\top x'\right) \,\middle|\, x' = (x, \mathbf{0})\right] \\ &= \mathbb{P}_{(x,y)\sim D}\left[y \neq \operatorname{sign}\left(w^\top x\right) \,\middle|\, x\right] = \eta(x)\,, \end{aligned}$$

and

$$\begin{aligned} \eta\left((\mathbf{0}, x)\right) &= \mathbb{P}_{(x',y')\sim D'}\left[y' \neq \operatorname{sign}\left((w, -w)^\top x'\right) \,\middle|\, x' = (\mathbf{0}, x)\right] \\ &= \mathbb{P}_{(x,y)\sim D}\left[-y \neq -\operatorname{sign}\left(w^\top x\right) \,\middle|\, x\right] = \eta(x)\,. \end{aligned}$$

Hence, it follows that $\eta(x') \leqslant \eta$ and further, that $\eta(x') \in \{0, \eta\}$. Moreover, we have

$$\mathrm{OPT}(D') = \mathop{\mathbb{E}}_{x'\sim D'_{x'}} \eta(x') = q \cdot \mathop{\mathbb{E}}_{x\sim D_x} \eta(x) + (1 - q) \cdot \mathop{\mathbb{E}}_{x\sim D_x} \eta(x) = \mathrm{OPT}(D)\,.$$

Lastly, suppose there is an SQ algorithm outputting a hypothesis h achieving misclassification error better than η with respect to D'. Then clearly, at least one of h and $-h$ restricted to the first M and last M coordinates, respectively, achieves misclassification error better than η with respect to D. Since we can estimate which one with one more statistical query, this completes the proof. ■

Acknowledgments

We thank David Steurer for helpful discussions regarding this project. 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

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 167–190, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Awasthi15b.html.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 152–192, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/awasthi16.html.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/5f8b73c0d4b1bf60dd7173b660b87c29-Abstract.html.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117, 2016. URL https://arxiv.org/pdf/1505.05800.pdf.

Ilias Diakonikolas and Daniel M. Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. *CoRR*, 2021. https://arxiv.org/pdf/2012.09720v3.pdf.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with massart noise. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4751–4762, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/358aee4cc897452c00244351e4d91f69-Abstract.html.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1486–1513. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/diakonikolas20c.html.

Ilias Diakonikolas, Daniel Kane, and Christos Tzamos. Forster decomposition and learning halfspaces with noise. *Advances in Neural Information Processing Systems*, 34, 2021a.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. *arXiv preprint arXiv:2108.08767*, 2021b. URL https://cseweb.ucsd.edu//~dakane/ThresholdPhenomenaLearningLTFsWithNoise.pdf.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Efficiently learning halfspaces with tsybakov noise. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 88–101, 2021c.

Vitaly Feldman. *Statistical Query Learning*, pages 2090–2095. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4. doi: 10.1007/978-1-4939-2864-4_401. URL https://doi.org/10.1007/978-1-4939-2864-4_401.

Vitaly Feldman. A general characterization of the statistical query complexity. In *Conference on Learning Theory*, pages 785–830. PMLR, 2017.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2): 1–37, 2017.

Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. *Mathematics of Operations Research*, 2021.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

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

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. URL https://arxiv.org/pdf/math/0702683.pdf.

Ronald L Rivest and Robert Sloan. A formal model of hierarchical concept-learning. *Information and Computation*, 114(1):88–114, 1994.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Robert H Sloan. Pac learning, noise, and geometry. In *Learning and Geometry: Computational Approaches*, pages 21–41. Springer, 1996.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR, 2017.

Appendix A. Full Proof of Theorem 3

In this section, we will give the full proof of Theorem 3. Recall that by Proposition 6 there exists two distributions A, B and unions of intervals J_1, J_2 satisfying the following.

Proposition 8 (Restatement of Proposition 6) *Let $0 < \zeta < 1/2$ and let $d \geqslant 2$ be an integer. Define $\delta = 4\sqrt{\log(1/\zeta)}/d$ and let $\varepsilon < \delta/8$. If $\delta < 1$, there exist probability distributions A, B on $\mathbb{R}$ and two unions J_1, J_2 of $2d+1$ intervals such that*

1. *$J_1 \cap J_2 = \emptyset$ and $J_1 \cup J_2 \subseteq [-d\delta - 5\varepsilon, d\delta + 5\varepsilon]$,*
2. *(a) $A = 0$ on J_2, $B = 0$ on J_1, and (b) for all $x \notin J_1 \cup J_2$ we have $A(x) = B(x)$,*
3. *for all $k \in \mathbb{N}$ the first k moments of A and B match those of a standard Gaussian within additive error $\nu = O(k!) \cdot \exp(-\Omega(1/\delta^2)) + 4\varepsilon \left(12\sqrt{\log(1/\zeta)}\right)^k$,*
4. *at most a ζ-fraction of the measure A (respectively B) lies outside J_1 (respectively J_2),*

5. $\chi^2(A, N(0,1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$ *and* $\chi^2(B, N(0,1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$.

Consider the distribution $D_v^{A,B,p}$ for $p = 1 - \eta$ and the parameters in the definition of A, B and J_1, J_2 to be chosen later. Although $D_v^{A,B,p}$ will not correspond to a Massart distribution when considering only halfspaces, it will turn out to work when considering polynomial threshold functions, i.e., $y = \operatorname{sign}(p(x))$ for some polynomial p in x. Further, we will be able to choose the parameters such that Lemma 5 will correspond to a superpolynomial lower bound in terms of M.

Unless explicitly indicated by a subscript, in what follows the $O(\cdot), \Theta(\cdot), \Omega(\cdot)$-notation will only contain universal constants indpendent of the ones we define throughout the section. Fix a unit vector $v \in \mathbb{R}^m$ and let $0 < \zeta < \eta$ be such that

$$\frac{\log M}{(\log\log M)^3} \geqslant C_\zeta \log(1/\zeta)$$

for a sufficiently large constant C_ζ. Further, let

$$\tau = M^{-\frac{\log M}{C_\tau (\log\log M)^3 \log(1/\zeta)}}$$

for a sufficiently large constant C_τ, so that

$$\log(1/\tau) = \frac{(\log M)^2}{C_\tau (\log\log M)^3 \log(1/\zeta)}\,.$$

We would like to find m and d such that we can represent degree-$8d$ polynomials over $\mathbb{R}^m$ as halfspaces over $\mathbb{R}^M$. It is sufficient to have

$$\binom{8d+m}{8d} \leqslant m^{8d} \leqslant M.$$

To this end, for C_m and C_d sufficiently large constants, consider

$$m = \left\lceil C_m \log(1/\tau) \log(1/\zeta)^4 \right\rceil$$

and

$$d = \left\lceil C_d \sqrt{\log(1/\zeta) \log(1/\tau) \log\log(1/\tau)} \right\rceil\,.$$

Notice that since

$$\log(1/\tau) \geqslant \frac{C_\zeta^2 \cdot (\log(1/\zeta))^2 \cdot (\log\log M)^3}{C_\tau \log(1/\zeta)} \gg \log(1/\zeta)$$

it follows that

$$\log m = \log\log(1/\tau) + 4\log\log(1/\zeta) + \Theta_{C_m}(1) = \Theta_{C_m}(\log\log(1/\tau))\,.$$

Hence,

$$m^{8d} = \exp(8d \cdot \log m) = \exp\left(\Theta_{C_m, C_d}\left(\sqrt{\log(1/\zeta)\log(1/\tau)\big(\log\log(1/\tau)\big)^3}\right)\right)$$

$$= \exp\left(\frac{1}{\sqrt{C_\tau}} \cdot \log(M) \cdot \Theta_{C_m, C_d}\left(\frac{\log\log(1/\tau)}{\log\log M}\right)^{3/2}\right) \leqslant M\,,$$

where the last inequality follows since $\log\log(1/\tau) \leqslant 2\log\log(M)$ and by choosing C_τ to be large enough with respect to C_m and C_d. Let

$$\delta = \frac{4\sqrt{\log(1/\zeta)}}{d} = \Theta\left(\frac{1}{C_d\sqrt{\log(1/\tau)\log\log(1/\tau)}}\right).$$

Further, let

$$k = \frac{4\log(1/\tau)}{\log\log(1/\zeta)}\,, \qquad\qquad \varepsilon = \tau \cdot \left(12\sqrt{\log(1/\zeta)}\right)^{-k}.$$

and consider the probability distributions A, B defined by Proposition 6 for our settings of δ, ζ, and ε. Also, let J_1, J_2 be the corresponding unions of intervals. Let

$$D^{(m)} \coloneqq D_v^{A,B,p} \text{ with } p = 1 - \eta\,,$$

so that $\min\{p, 1-p\} = \eta$. As we will shortly see, $D^{(m)}$ is an η-Massart polynomial-threshold function. In order to obtain an η-Massart halfspace, we will embed $D^{(m)}$ into the higher dimensional space $\mathbb{R}^M$.

Let

$$M' \coloneqq \binom{m+8d}{8d} \leqslant m^{8d} \leqslant M\,,$$

and define

$$\begin{aligned} V_{8d} : \mathbb{R}^m &\to \mathbb{R}^{M'} \\ x &\mapsto (x^\alpha)_{|\alpha| \leqslant 8d}\,, \end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m$ is a multi-index and $|\alpha| = \sum_{i\in[m]} \alpha_i$. Furthermore, let

$$\begin{aligned} E_{M'\to M} : \mathbb{R}^{M'} &\to \mathbb{R}^M \\ x &\mapsto (x, 0)\,, \end{aligned}$$

be the linear embedding of $\mathbb{R}^{M'}$ into $\mathbb{R}^M$ that is obtained by appending by zeros. We will embed $D^{(m)}$ into $\mathbb{R}^M$ using the embedding $E : \mathbb{R}^m \to \mathbb{R}^M$ defined as

$$E = E_{M'\to M} \circ V_{8d}.$$

The hard distribution D is as follows: Draw $(x, y) \sim D^{(m)}$ and return $(E(x), y)$. The next lemma shows that this distribution satisfies the η-Massart property with respect to the class of halfspaces.

Lemma 9 *The probability distribution D is an η-Massart halfspace with* $\mathrm{OPT} \leqslant \zeta$.

Proof Let $v \in \mathbb{R}^m$ and consider the function $g_v \colon \mathbb{R}^m \to \{-1, 1\}$ such that

$$g_v(x) = \begin{cases} -1\,, & \text{if } \langle v, x\rangle \in J_2\,, \\ +1\,, & \text{otherwise.} \end{cases}$$

Since J_2 is a union of $2d + 1 \leqslant 4d$ intervals, g_v can be written as $\operatorname{sign}(p_v(x))$, where $p_v(x) = p(\langle v, x\rangle)$ for some degree-$8d$ polynomial p. Now since $M' = \binom{m+8d}{m}$, there is a linear function $f \colon \mathbb{R}^{M'} \to \{-1, 1\}$ such that for all $x \in \mathbb{R}^m$ it holds that $g_v(x) = \operatorname{sign}(f(V_{8d}(x)))$. This in turn implies that there is a linear function $h \colon \mathbb{R}^M \to \{-1, 1\}$ such that for all $x \in \mathbb{R}^m$ we have $g_v(x) = \operatorname{sign}(h(E(x)))$.

Note that $D(x', y) \neq 0$ only if $x' = E(x)$ for some $x \in \mathbb{R}^m$. Furthermore,

- For $x \in \mathbb{R}^m$ satisfying $\langle x, v\rangle \in (J_1 \cup J_2)$, we have $y = \operatorname{sign}(h(E(x)))$ with probability 1.
- For $x \in \mathbb{R}^m$ satisfying $\langle x, v\rangle \notin (J_1 \cup J_2)$ and $D(E(x), y) \neq 0$, we have $\operatorname{sign}(h(E(x))) = 1$, and
$$y = \begin{cases} 1 = \operatorname{sign}(h(E(x))) & \text{with probability } p = 1 - \eta, \\ -1 = -\operatorname{sign}(h(E(x))) & \text{with probability } 1 - p = \eta. \end{cases}$$

Hence, D corresponds to an η-Massart distribution corresponding to the halfspace f. Furthermore, the flipping probability function $\eta : \mathbb{R}^M \to [0, \eta]$ satisfies

$$\eta(x') = \begin{cases} \eta\,, & \text{if } \exists x \in R^m,\ x' = E(x) \text{ and } \langle x, v\rangle \notin (J_1 \cup J_2)\,. \\ 0\,, & \text{otherwise.} \end{cases}$$

Therefore, $\eta(x') \in \{0, \eta\}$ for all $x' \in \mathbb{R}^M$ and

$$\mathrm{OPT} = \mathop{\mathbb{E}}_{(x', y)\sim D} \eta(x') = \eta \cdot \mathbb{P}_{x\sim D^{(m)}}[\langle v, x\rangle \notin (J_1 \cup J_2)] \leqslant \eta \cdot \zeta \leqslant \zeta.$$

The second last inequality follows from Item 2 and Item 4 of Proposition 6. Indeed, we have

$$\mathbb{P}_{x\sim D^{(m)}}[\langle v, x\rangle \notin (J_1 \cup J_2)] = p \cdot \mathbb{P}_{\langle v,x\rangle\sim A}[\langle v, x\rangle \notin (J_1 \cup J_2)] + (1-p) \cdot \mathbb{P}_{\langle v,x\rangle\sim B}[\langle v, x\rangle \notin (J_1 \cup J_2)] \leqslant \zeta\,.$$

■

Second, any hypothesis for predicting y from x can be turned into one predicting y from $E(x)$ and vice-versa. Hence, it is enough to show that the former is SQ-hard. Consider the setting of Lemma 5 with A and B given by Proposition 6. First, by Item 5 we know that $\chi^2(A, N(0, 1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$ and $\chi^2(B, N(0, 1)) = O\left(\frac{\delta}{\varepsilon}\right)^2$. Hence,

$$\alpha = O\left(\frac{\delta}{\varepsilon}\right)^2.$$

Further, let

$$\gamma = O(k!) \cdot \exp\left(-\Omega(1/\delta^2)\right)\,.$$

By Proposition 6 we know that the first k moments of A and B match those of a standard Gaussian up to additive error

$$\gamma + 4\varepsilon \left(12\sqrt{\log(1/\zeta)}\right)^k = \gamma + 4\tau\,,$$

where the equality follows from the fact that $\varepsilon = \tau \cdot \left(12\sqrt{\log(1/\zeta)}\right)^{-k}$.

We claim that by choosing C_d large enough we get $\gamma \ll \tau$. Indeed, since $k = \dfrac{4\log(1/\tau)}{\log\log(1/\zeta)}$ and $\delta = \Theta\left(\dfrac{1}{C_d\sqrt{\log(1/\tau)\log\log(1/\tau)}}\right)$, we have

$$\log(O(k!)) = O(k\log k) = O\left(\log(1/\tau)\frac{\log\log(1/\tau)}{\log\log(1/\zeta)}\right) \leqslant O\left(\frac{1}{C_d^2\delta^2}\right)\,.$$

Hence, by choosing C_d large enough, we get

$$\gamma = \exp\left(-\Omega(1/\delta^2)\right) = \exp\left(-\Omega(C_d^2\log(1/\tau)\log\log(1/\tau))\right) \ll \tau\,.$$

It follows that both A and B match the moments of a standard Gaussian up to additive error at most $\nu = 5\tau$. This, in addition to the fact that $\varepsilon = \tau \cdot \left(12\sqrt{\log(1/\zeta)}\right)^{-k}$, imply that the parameter $\nu^2 + \alpha\cdot c^k$ in Lemma 5 is equal to

$$\nu^2 + \alpha c^k = 25\tau^2 + O\left(\frac{\delta}{\varepsilon}\right)^2 \cdot c^k \leqslant 25\tau^2 + \frac{O(1)}{\varepsilon^2}\cdot c^k \leqslant 25\tau^2 + \frac{O(1)}{\tau^2}\cdot\left(144c\cdot\log(1/\zeta)\right)^k\,.$$

By choosing $c = \dfrac{1}{144\log(1/\zeta)^2}$ and recalling that $k = \dfrac{4\log(1/\tau)}{\log\log(1/\zeta)}$, we get

$$\left(144c\cdot\log(1/\zeta)\right)^k = \log(1/\zeta)^{-k} = \exp\left(-\log\log(1/\zeta)\cdot\frac{4\log(1/\tau)}{\log\log(1/\zeta)}\right) = \tau^4\,,$$

which implies that

$$\nu^2 + \alpha\cdot c^k \leqslant O(\tau^2) \leqslant \tau$$

for sufficiently large M (and hence sufficiently small τ). Therefore, we can choose the parameter ρ in Lemma 5 to be equal to τ.

Next, we claim that the parameter

$$N = \frac{2^{c^2\cdot\Omega(m)}\cdot\rho}{\alpha} \geqslant \Omega\left(\frac{\varepsilon}{\delta}\right)^2\cdot\exp(c^2\cdot\Omega(m))\cdot\tau \geqslant \Omega(\varepsilon^2)\cdot\exp\left(c^2\cdot\Omega(m) - \log(1/\tau)\right)$$

of Lemma 5 is at least $1/\tau^{\Theta(1)}$. In fact, recalling that $c = \dfrac{1}{144\log(1/\zeta)^2}$, $\varepsilon = \tau\cdot\left(12\sqrt{\log(1/\zeta)}\right)^{-k}$, $k = \dfrac{4\log(1/\tau)}{\log\log(1/\zeta)}$ and $m = \left\lceil C_m\log(1/\tau)\log(1/\zeta)^4\right\rceil$, we obtain

$$\Omega(\varepsilon^2)\exp\left(c^2\cdot\Omega(m) - \log(1/\tau)\right)$$

$$
\begin{aligned}
&= \exp\left(c^2 \cdot \Omega(m) - 3\log(1/\tau) - 2k\log\left(12\sqrt{\log(1/\zeta)}\right) - O(1)\right) \\
&\geqslant \exp\left(C_m \cdot \Omega(1) \cdot \log(1/\tau) - 3\log(1/\tau) - \Theta(k\log\log(1/\zeta))\right) \\
&= \exp\left((C_m \cdot \Omega(1) - \Theta(1)) \cdot \log(1/\tau)\right) .
\end{aligned}
$$

By choosing the constant C_m in the definition of m large enough, we conclude that $N \geqslant 1/\tau^{\Theta(1)}$.

Hence, by Lemma 5 any SQ algorithm that outputs a hypothesis h such that

$$\mathbb{P}_{(x,y)\sim D}[h(x) \neq y] \leqslant \eta - 4\sqrt{\tau}$$

must either make queries of accuracy better than $2\sqrt{\rho} = \tau^{\Theta(1)}$ or make at least $N = 1/\tau^{\Theta(1)}$ queries. Since

$$\tau^{\Theta(1)} = M^{-\Theta\left(\log M/\left[\log(1/\zeta)\cdot(\log\log M)^3\right]\right)} ,$$

Theorem 3 follows.

Appendix B. Full Proof of Hard Distributions

In this section, we will give a full proof of Proposition 6. For convenience, we restate here the definition of the distributions A, B (cf. Eq. (4.1)) and the union of intervals J_1, J_2 (cf. Eq. (4.2)). Let $0 < \delta, \varepsilon, d$ be to be fixed later and consider the positive measures

$$G_{\delta,\varepsilon}(x) = \sum_{n\in\mathbb{Z}} G(x) \cdot \left(\frac{\delta}{2\varepsilon}\right) \cdot \mathbf{1}_{[x\in[n\delta-\varepsilon, n\delta+\varepsilon]]} , \qquad G^{(n)}_{\delta,\varepsilon}(x) = \frac{G_{\delta,\varepsilon}(x)}{\|G_{\delta,\varepsilon}\|_1} .$$

Note, that $G^{(n)}_{\delta,\varepsilon}$ corresponds to a probability measure. Further, define

$$A(x) = G^{(n)}_{\delta,\varepsilon}(x) , \qquad B(x) = \begin{cases} A(x) , & \text{if } |x| > d\delta + 5\varepsilon , \\ A(x + 4\varepsilon) , & \text{otherwise.} \end{cases}$$

and

$$J_1 = \bigcup_{-d\leqslant n\leqslant d} [n\delta - \varepsilon, n\delta + \varepsilon] , \qquad J_2 = \bigcup_{-d\leqslant n\leqslant d} [n\delta - 5\varepsilon, n\delta - 3\varepsilon] .$$

Our goal is to show

Proposition 10 (Restatement of Proposition 6) *Let $0 < \zeta < 1/2$ and let $d \geqslant 2$ be an integer. Assume that $\delta = 4\sqrt{\log(1/\zeta)}/d$ and $\varepsilon < \delta/8$. If $\delta < 1$, the distributions A, B defined in Eq. (4.1) (or equivalently defined above) together with the intervals J_1, J_2 defined in Eq. (4.2) (or equivalently defined above) satisfy the following*

1. *$J_1 \cap J_2 = \emptyset$ and $J_1 \cup J_2 \subseteq [-d\delta - 5\varepsilon, d\delta + 5\varepsilon]$,*
2. *(a) $A = 0$ on J_2, $B = 0$ on J_1, and (b) for all $x \notin J_1 \cup J_2$ we have $A(x) = B(x)$,*
3. *for all $k \in \mathbb{N}$ the first k moments of A and B match those of a standard Gaussian within additive error $\nu = O(k!) \cdot \exp(-\Omega(1/\delta^2)) + 4\varepsilon\left(12\sqrt{\log(1/\zeta)}\right)^k$,*

4. *at most a ζ-fraction of the measure A (respectively B) lies outside J_1 (respectively J_2)*,
5. $\chi^2\left(A, N(0,1)\right) = O\left(\frac{\delta}{\varepsilon}\right)^2$ *and* $\chi^2\left(B, N(0,1)\right) = O\left(\frac{\delta}{\varepsilon}\right)^2$.

Since $\varepsilon < \delta/8$, Item 1 and Item 2 of Proposition 6 clearly hold.

In order to show Item 4, we bound the measure of $G^{(n)}_{\delta,\varepsilon}$ outside $J_1 \cup J_2$ by ζ. Indeed, it then follows that

$$\int_{x\notin J_1} A(x)\,dx = \int_{x\notin J_2} B(x)\,dx = \int_{x\notin J_1\cup J_2} G^{(n)}_{\delta,\varepsilon}(x)\,dx \leqslant \zeta\,.$$

In order to do so, we will upper bound the measure of $G_{\delta,\varepsilon}$ outside $J_1 \cup J_2$ and will lower bound the total measure $\|G_{\delta,\varepsilon}\|_1$. We have:

$$\begin{aligned}
\int_{x\notin J_1\cup J_2} G_{\delta,\varepsilon}(x)\,dx &\leqslant 2\int_{d\delta-5\varepsilon}^{\infty} G_{\delta,\varepsilon}(x)\,dx = 2\left(\frac{\delta}{2\varepsilon}\right)\sum_{n>d}\int_{n\delta-\varepsilon}^{n\delta+\varepsilon} G(x)\,dx \\
&\leqslant 2\sum_{n>d}\int_{(n-1)\delta+\varepsilon}^{n\delta+\varepsilon} G(x)\,dx \leqslant 2\mathbb{P}\big[N(0,1)\geqslant d\delta\big] \\
&\leqslant 2\exp\big(-(d\delta)^2/2\big) \leqslant 2\exp\left(-\left(4\sqrt{\log(1/\zeta)}\right)^2/2\right) \\
&\leqslant 2\exp\left(-8\log(1/\zeta)\right) = 2\zeta^8\,,
\end{aligned}$$

where we used the fact that $G(x)$ is decreasing for $x \geqslant 0$ and that $d = 4\sqrt{\log(1/\zeta)}/\delta \geqslant 2$. As for $\|G_{\delta,\varepsilon}\|_1$, we have

$$\begin{aligned}
\|G_{\delta,\varepsilon}\|_1 &= \int_{\mathbb{R}} G_{\delta,\varepsilon}(x)\,dx \geqslant 2\int_{\delta-\varepsilon}^{\infty} G_{\delta,\varepsilon}(x)\,dx \\
&= 2\left(\frac{\delta}{2\varepsilon}\right)\sum_{n\geqslant 1}\int_{n\delta-\varepsilon}^{n\delta+\varepsilon} G(x)\,dx \geqslant 2\sum_{n\geqslant 1}\int_{n\delta-\varepsilon}^{(n+1)\delta-\varepsilon} G(x)\,dx \qquad \text{(B.1)} \\
&= 2\mathbb{P}\big[N(0,1)\geqslant \delta-\varepsilon\big] \geqslant 2\mathbb{P}\big[N(0,1)\geqslant 1\big] \geqslant 2\cdot\frac{1}{10} = \frac{1}{5}\,,
\end{aligned}$$

where we also used the fact that $G(x)$ is decreasing for $x \geqslant 0$, and that $\delta < 1$. Now since $\zeta < \frac{1}{2}$, we deduce that

$$\int_{x\notin J_1\cup J_2} G^{(n)}_{\delta,\varepsilon}(x)\,dx = \frac{1}{\|G_{\delta,\varepsilon}\|_1}\int_{x\notin J_1\cup J_2} G_{\delta,\varepsilon}(x)\,dx \leqslant 10\zeta^8 \leqslant \zeta\,.$$

Next, we will bound the chi-square divergence

Lemma 11 *Let A, B be defined as above, then $\chi^2\left(A, N(0,1)\right) = O\left(\frac{\delta}{\varepsilon}\right)^2$ and $\chi^2\left(B, N(0,1)\right) = O\left(\frac{\delta}{\varepsilon}\right)^2$.*

Proof We will start with A:

$$\chi^2\left(A, N(0,1)\right) = \frac{1}{\|G_{\delta,\varepsilon}\|_1^2}\left(\frac{\delta}{2\varepsilon}\right)^2\sum_{n\in\mathbb{Z}}\int_{n\delta-\varepsilon}^{n\delta+\varepsilon}\frac{G(x)^2}{G(x)}\,dx - 1$$

$$\leqslant 25 \cdot \left(\frac{\delta}{2\varepsilon}\right)^2 \int_{-\infty}^{\infty} G(x)\,dx = O\left(\frac{\delta}{\varepsilon}\right)^2 ,$$

where we used $\|G_{\delta,\varepsilon}\|_1 \geqslant \frac{1}{5}$ from (B.1).

For B we get:

$$\chi^2\left(B, N(0,1)\right) = \frac{1}{\|G_{\delta,\varepsilon}\|_1^2}\left(\frac{\delta}{2\varepsilon}\right)^2 \left[\sum_{|n|>d}\int_{n\delta-\varepsilon}^{n\delta+\varepsilon} \frac{G(x)^2}{G(x)}\,dx + \sum_{|n|\leqslant d}\int_{n\delta-5\varepsilon}^{n\delta-3\varepsilon} \frac{G(x+4\varepsilon)^2}{G(x)}\,dx\right] - 1\,.$$

The term corresponding to the first sum is less than or equal to $\chi^2\left(A, N(0,1)\right) = O\left(\frac{\delta}{\varepsilon}\right)^2$, and for the second sum we notice that

$$\begin{aligned}\frac{G(x+4\varepsilon)^2}{G(x)} &= \frac{1}{\sqrt{2\pi}}\exp\left(-(x+4\varepsilon)^2 + \frac{x^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2} - 8x\varepsilon - 16\varepsilon^2\right)\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(16\varepsilon^2\right)\exp\left(-\frac{(x+8\varepsilon)^2}{2}\right)\end{aligned}$$

implying that

$$\sum_{|n|\leqslant d}\int_{n\delta-5\varepsilon}^{n\delta-3\varepsilon} \frac{G(x+4\varepsilon)^2}{G(x)}\,dx \leqslant \int_{-\infty}^{\infty} \frac{\exp\left(16\varepsilon^2\right)}{\sqrt{2\pi}}\exp\left(-\frac{(x+8\varepsilon)^2}{2}\right)\,dx = O(1)\,.$$

Putting everything together yields the claim. ■

Lastly, we show that the moments match up to the desired error. We start with A.

Lemma 12 *Let $k \in \mathbb{N}$. For the distribution A defined as above and all $t \leqslant k$ it holds that*

$$\left|\mathbb{E}\,A^t - \mathbb{E}\,G^t\right| \leqslant O(t!)\cdot\exp(-\Omega(1/\delta^2))\,.$$

Proof Since $G_{\delta,\varepsilon} = \sum_{n\in\mathbb{Z}} G(x)\cdot f(\frac{x-n\delta}{\delta})$ for $f(y) = \left(\frac{\delta}{2\varepsilon}\right)\cdot \mathbf{1}_{[y\in[-\varepsilon/\delta,\varepsilon/\delta]]}$ it follows from Fact 15 that

$$|\mathbb{E}\,G^t - \mathbb{E}\,G^t_{\delta,\varepsilon}| \leqslant t!\cdot\delta^t\cdot\exp(-\Omega(1/\delta^2))\,.$$

In particular, for $t = 0$, we get

$$|1 - \|G_{\delta,\varepsilon}\|_1| \leqslant \exp(-\Omega(1/\delta^2))\,,$$

which implies that

$$\left|\frac{1}{\|G_{\delta,\varepsilon}\|_1} - 1\right| \leqslant \frac{\exp(-\Omega(1/\delta^2))}{1-\exp(-\Omega(1/\delta^2))} \leqslant O(1)\cdot\exp(-\Omega(1/\delta^2))\,.$$

Now using the fact that $\mathbb{E}\,G^t = 0 \leqslant t!$ for odd t and that

$$\mathbb{E}\,G^t = (t-1)!! = (t-1)(t-3)(t-5)\cdots 1 \leqslant t!$$

for even t, we get

$$\left|\mathbb{E}\,G^t_{\delta,\varepsilon}\right| \leqslant \mathbb{E}\,G^t + t!\cdot\delta^t\cdot\exp(-\Omega(1/\delta^2)) \leqslant t!\left(1+\delta^t\cdot\exp(-\Omega(1/\delta^2))\right) \leqslant O(t!)\,.$$

It follows that

$$\left|\mathbb{E}\, A^t - \mathbb{E}\, G_{\delta,\varepsilon}^t\right| = \left|\mathbb{E}\left(G_{\delta,\varepsilon}^{(n)}\right)^t - \mathbb{E}\, G_{\delta,\varepsilon}^t\right| = \left|\mathbb{E}\, G_{\delta,\varepsilon}^t\right| \cdot \left|\frac{1}{\|G_{\delta,\varepsilon}\|_1} - 1\right| \leqslant O(t!) \cdot \exp(-\Omega(1/\delta^2))\,.$$

We conclude that

$$\begin{aligned}\left|\mathbb{E}\, A^t - \mathbb{E}\, G^t\right| &\leqslant \left|\mathbb{E}\, A^t - \mathbb{E}\, G_{\delta,\varepsilon}^t\right| + \left|\mathbb{E}\, G^t - \mathbb{E}\, G_{\delta,\varepsilon}^t\right| \\ &\leqslant O(t!) \cdot \exp(-\Omega(1/\delta^2)) + t! \cdot \delta^t \cdot \exp(-\Omega(1/\delta^2)) \\ &\leqslant O(t!) \cdot \exp(-\Omega(1/\delta^2))\,.\end{aligned}$$

■

Next, we will prove the bound for B:

Lemma 13 *Let $k \in \mathbb{N}$. For the distribution B defined as above and all $t \leqslant k$ it holds that*

$$\left|\mathbb{E}\, B^t - \mathbb{E}\, G^t\right| \leqslant O(t!) \cdot \exp(-\Omega(1/\delta^2)) + 4\varepsilon\left(12\sqrt{\log(1/\zeta)}\right)^t.$$

Proof Due to Theorem 12, it is sufficient to show that $\left|\mathbb{E}\, B^t - EA^t\right| \leqslant 4\varepsilon\left(12\sqrt{\log(1/\zeta)}\right)^t$. To this end, notice that the two distributions agree for x larger in magnitude than $d\delta - 5\varepsilon$. Thus

$$\begin{aligned}\mathbb{E}\, B^t - \mathbb{E}\, A^t &= \int_{-d\delta-5\varepsilon}^{d\delta+5\varepsilon} x^t dB(x) - \int_{-d\delta+5\varepsilon}^{d\delta-5\varepsilon} x^t dA(x) \\ &= \int_{-d\delta-5\varepsilon}^{d\delta-3\varepsilon} (x-4\varepsilon)^t dA(x) - \int_{-d\delta+\varepsilon}^{d\delta-\varepsilon} x^t dA(x) \\ &= \int_{-d\delta-\varepsilon}^{d\delta+\varepsilon} (x-4\varepsilon)^t dA(x) - \int_{-d\delta+\varepsilon}^{d\delta-\varepsilon} x^t dA(x) \\ &= \int_{-d\delta-\varepsilon}^{d\delta+\varepsilon} \left((x-4\varepsilon)^t - x^t\right) \cdot dA(x)\,.\end{aligned}$$

Therefore,

$$\left|\mathbb{E}\, B^t - \mathbb{E}\, A^t\right| \leqslant \sup_{-d\delta-\varepsilon \leqslant x \leqslant d\delta+\varepsilon} \left|(x-4\varepsilon)^t - x^t\right| \leqslant \sup_{-2d\delta \leqslant x \leqslant 2d\delta} \left|(x-4\varepsilon)^t - x^t\right|\,.$$

Now since

$$(x-4\varepsilon)^t - x^t = \sum_{l=1}^{t} \binom{t}{l} (-4\varepsilon)^l x^{t-l}\,,$$

we get

$$\left|(x-4\varepsilon)^t - x^t\right| \leqslant \sum_{l=1}^{t} \binom{t}{l} (4\varepsilon)^l \left|x\right|^{t-l} \leqslant \sum_{l=1}^{t} \binom{t}{l} 4\varepsilon \left(1+|x|\right)^t$$

$$\leqslant 2^t \cdot 4\varepsilon \left(1 + |x|\right)^t = 4\varepsilon \left(2 + 2\,|x|\right)^t .$$

Now since $d\delta = 4\sqrt{\log(1/\zeta)}$, we conclude that

$$\left|\mathbb{E}\, B^t - \mathbb{E}\, A^t\right| \leqslant 4\varepsilon(2 + 4d\delta)^t = 4\varepsilon \left(2 + 8\sqrt{\log(1/\zeta)}\right)^t .$$

■

Since $\zeta < 1/2$ and $\log(2) > 1/4$ it holds that

$$\sqrt{\log(1/\zeta)} > \sqrt{\log(2)} > \sqrt{1/4} = 1/2\,.$$

Hence, we obtain

$$\left|\mathbb{E}\, B^t - \mathbb{E}\, A^t\right| \leqslant 4\varepsilon \left(4 \cdot \frac{1}{2} + 8\sqrt{\log(1/\zeta)}\right)^t \leqslant 4\varepsilon \left(12\sqrt{\log(1/\zeta)}\right)^t .$$

Appendix C. Discussion of Lemma 5

As we already mentioned, the proof of Lemma 5 is verbatim the same as the one of Proposition 3.8 in Diakonikolas and Kane (2021). The only difference is that we apply their Lemma 3.5 with an arbitrary $c > 0$ instead of $c = 1/2$. Further, we need the following more precise version of their Fact 3.6. Also here, the proof is the same (and straightforward) but for us it is important to know the explicity dependence of the size of the set on the parameter c. In Diakonikolas and Kane (2021) it was only stated as $2^{\Omega_c(m)}$.

Fact 14 *Let $c > 0$. There exists a set S of unit vectors over $\mathbb{R}^m$ of size $\exp(c^2 \cdot \Omega(m))$ such that for all $u, v \in S$ it holds that $|\langle u, v\rangle| \leqslant c$. Further, the constant inside the $\Omega(\cdot)$ notation is independent of c.*

Proof Let the elements of S be picked independently and uniformly at random from the unit sphere. Let $u, v \in S$ and w.l.o.g. assume that $u = e_1$. Since v has the same distribution as $X/\|X\|$ where $X \sim N(0, \mathrm{Id}_m)$ it holds that

$$\mathbb{P}[|\langle u, v\rangle| > c] = \mathbb{P}[|v_1| > c] = \mathbb{P}[|X_1| > c \cdot \|X\|] \leqslant \mathbb{P}[|X_1| > c\sqrt{m/2}] + \mathbb{P}[\|X\| < \sqrt{m/2}]\,.$$

By standard Gaussian tail bounds the first probability is at most $2\exp(-c^2 m/4)$ and by standard chi-squared tail bounds (e.g., Wainwright (2019)) the second probability is at most $2\exp(-m/32)$. Hence, the claim follows by a union bound over all pairs of distinct elements in S. ■

Appendix D. Moment Bounds

In the following, we prove a statement that is similar to a lemma that was previously shown in Diakonikolas and Kane (2021).

Fact 15 *Let $f : \mathbb{R} \to \mathbb{R}$ be a non-negative function such that*

- $f(x) = 0$ *for* $x \notin [0,1]$, *and*
- $\int_0^1 f(x)\,dx = 1$.

Let $G \sim N(0,1)$ *and with a little abuse of notation, we will denote the pdf of* G *also by* G. *For every* $\delta > 0$, *define*

$$G_\delta(x) = \sum_{n\in\mathbb{Z}} f\left(\frac{x+n\delta}{\delta}\right)\cdot G(x)\,.$$

We have

$$|\mathbb{E}\,G^t - \mathbb{E}\,G_\delta^t| \leqslant t!\cdot\delta^t\cdot\exp(-\Omega(1/\delta^2))\,.$$

The proof will make use of Fourier analysis. We will introduce here the necessary background. For a function $g\colon \mathbb{R}\to\mathbb{R}$ we define its Fourier transform to be

$$\hat{g}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x)\cdot e^{-i\omega x}\,dx\,.$$

It is well-known that for $a,b\in\mathbb{R}$ and $h\colon\mathbb{R}\to\mathbb{R}$ we have

$$\widehat{(a\cdot g + b\cdot h)} = a\cdot\hat{g} + b\cdot\hat{h} \quad\text{and}\quad \widehat{(g\cdot h)} = \frac{1}{\sqrt{2\pi}}\left(\hat{g}*\hat{h}\right)\,,$$

where $*$ denotes convolution. Further, if G denotes the pdf of a standard Gaussian then $\hat{G} = G$.

For a random variable X with pdf g, let

$$\varphi_X(t) = \int_{-\infty}^{\infty} g(x)\cdot e^{itx}\,dx$$

denote its characteristic function. Notice that $\varphi_X(t) = \sqrt{2\pi}\cdot\hat{g}(-t)$. For the t-th moment of X it follows

$$\begin{aligned}\mathbb{E}\,X^t = \frac{1}{i^t}\cdot\varphi_X^{(t)}(0) &= \sqrt{2\pi}\cdot(-i)^t\cdot(-1)^t\cdot\hat{g}^{(t)}(0)\\ &= \sqrt{2\pi}\cdot i^t\cdot\hat{g}^{(t)}(0)\,.\end{aligned}$$

Proof [Proof of Fact 15] By the above discussion it is enough to show that

$$|\hat{G}_\delta^{(t)}(0) - \hat{G}^{(t)}(0)| \leqslant t!\cdot\frac{\delta^t}{\sqrt{2\pi}}\cdot\exp^{-\Omega((1/\delta)^2)}\,.$$

We know that $\hat{G} = G$. Next, we will compute $\hat{G}_\delta(\omega)$. Let

$$F(x) = \sum_{n\in\mathbb{Z}} f\left(\frac{x+n\delta}{\delta}\right)\,,$$

so that $G_\delta = G(x)\cdot F(x)$ and hence

$$\hat{G}_\delta = \frac{1}{\sqrt{2\pi}}\hat{G}(\omega) * \hat{F}(\omega)\,.$$

For F we obtain the following: Since F is periodic with period δ, we can decompose it using the Fourier basis over $[0, \delta]$. We get that

$$F(x) = \sum_{n \in \mathbb{Z}} \hat{F}_n \cdot e^{2i\pi \frac{n}{\delta} x} \,,$$

where

$$\begin{aligned} \hat{F}_n &= \frac{1}{\delta} \int_0^\delta F(x) \cdot e^{-2i\pi \frac{n}{\delta} x} \, dx = \frac{1}{\delta} \int_0^\delta \sum_{l \in \mathbb{Z}} f\left(\frac{x + l\delta}{\delta}\right) \cdot e^{-2i\pi \frac{n}{\delta} x} \, dx \\ &= \frac{1}{\delta} \int_0^\delta f\left(\frac{x}{\delta}\right) \cdot e^{-2i\pi \frac{n}{\delta} x} \, dx = \int_0^1 f(y) \cdot e^{-2i\pi n y} \, dy \,. \end{aligned}$$

For $n = 0$, we clearly have $\hat{F}_0 = 1$ since f integrates to 1 over $[0, 1]$. Now for $n \neq 0$, we can write

$$|\hat{F}_n| = \left| \int_0^1 f(y) \cdot e^{-2i\pi n y} \, dy \right| \leqslant \int_0^1 |f(y) \cdot e^{-2i\pi n y}| \, dy = \int_0^1 f(y) \, dy = 1 \,,$$

where we used the fact that f is non-negative and that the complex exponential has magnitude 1.

Now using the fact that the Fourier transform of $e^{2i\pi \frac{n}{\delta} x}$ is equal to $\sqrt{2\pi} \boldsymbol{\delta}_{\mathrm{D}} \left(\omega - \frac{2\pi n}{\delta}\right)$, where $\boldsymbol{\delta}_{\mathrm{D}}$ is the Dirac delta-distribution[8], we get that

$$\hat{F}(\omega) = \sqrt{2\pi} \cdot \sum_{n \in \mathbb{Z}} \hat{F}_n \cdot \boldsymbol{\delta}_{\mathrm{D}} \left(\omega - \frac{2\pi n}{\delta}\right) \,,$$

and hence

$$\hat{G}_\delta(\omega) = \frac{1}{\sqrt{2\pi}} \hat{G}(\omega) * \hat{F}(\omega) = \sum_{n \in \mathbb{Z}} \hat{F}_n \cdot \hat{G}(\omega) * \boldsymbol{\delta}_{\mathrm{D}} \left(\omega - \frac{2\pi n}{\delta}\right) = \sum_{n \in Z} \hat{F}_n \cdot \hat{G} \left(\omega - \frac{2\pi n}{\delta}\right) \,.$$

Now since $\hat{F}_0 = 1$ and $|\hat{F}_n| \leqslant 1$ for $n \neq 1$, we get

$$\hat{G}_\delta^{(t)}(0) = \hat{G}^{(t)}(0) + \sum_{n \neq 0} \hat{F}_n \cdot \hat{G}^{(t)} \left(-\frac{2\pi n}{\delta}\right) \,,$$

and

$$|\hat{G}_\delta^{(t)}(0) - \hat{G}^{(t)}(0)| \leqslant \sum_{n \neq 0} \left| \hat{G}^{(t)} \left(\frac{2\pi n}{\delta}\right) \right| \,.$$

Now using Cauchy's integral formula, we have

$$\hat{G}^{(t)} \left(\frac{2\pi n}{\delta}\right) = \frac{t!}{2\pi i} \oint_{\gamma_n} \frac{\hat{G}(z)}{\left(z - \frac{2\pi n}{\delta}\right)^{t+1}} dz \,,$$

8. Note that we use bold font and the subscript D, i.e., $\boldsymbol{\delta}_{\mathrm{D}}$, to denote the Dirac delta-distribution and non-bold font for the paramter $\delta \in \mathbb{R}$.

where the complex (contour) integral is over the circle γ_n of center $\frac{2\pi n}{\delta}$ and of radius $\frac{\pi}{2\delta}$ in the complex plane. Now since the circle γ_n has length $\frac{\pi^2}{\delta}$, we get

$$\begin{aligned}\left|\hat{G}^{(t)}\left(\frac{2\pi n}{\delta}\right)\right| &= \frac{t!}{2\pi}\left|\oint_{\gamma_n}\frac{\hat{G}(z)}{\left(z-\frac{2\pi n}{\delta}\right)^{t+1}}dz\right| \leqslant \frac{t!}{2\pi}\cdot\frac{\pi^2}{\delta}\cdot\max_{z\in\gamma}\left|\frac{\hat{G}(z)}{\left(z-\frac{2\pi n}{\delta}\right)^{t+1}}\right| \\ &= t!\cdot\frac{\pi}{2\delta}\cdot\frac{\max_{z\in\gamma}|\hat{G}(z)|}{\left(\frac{\pi}{2\delta}\right)^{t+1}} \leqslant t!\cdot\frac{\delta^t}{\sqrt{2\pi}}\cdot e^{-\Omega((n/\delta)^2)},\end{aligned}$$

where in the last inequality we used the fact that

$$\begin{aligned}\max_{z\in\gamma_n}|\hat{G}(z)| = \max_{z\in\gamma_n}|G(z)| &= \frac{1}{\sqrt{2\pi}}\max_{z\in\gamma_n}\left|e^{-\frac{z^2}{2}}\right| = \frac{1}{\sqrt{2\pi}}\max_{x+iy\in\gamma_n}e^{-\frac{x^2-y^2}{2}} \\ &\leqslant \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\left(\frac{2\pi|n|}{\delta}-\frac{\pi}{2\delta}\right)^2-\left(\frac{\pi}{2\delta}\right)^2\right)} \leqslant \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{2\pi|n|}{\delta}-\frac{\pi}{\delta}\right)^2} \\ &\leqslant \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\pi|n|}{\delta}\right)^2} = \frac{1}{\sqrt{2\pi}}e^{-\Omega((n/\delta)^2)}.\end{aligned}$$

We conclude that

$$|\hat{G}_\delta^{(t)}(0)-\hat{G}^{(t)}(0)| \leqslant t!\cdot\frac{\delta^t}{\sqrt{2\pi}}\cdot\sum_{n\neq 0}e^{-\Omega((n/\delta)^2)} = t!\cdot\frac{\delta^t}{\sqrt{2\pi}}\cdot e^{-\Omega((1/\delta)^2)}.$$

■