Hardness of Agnostically Learning Halfspaces from Worst-Case Lattice Problems

Stefan Tiegel STEFAN.TIEGEL@INF.ETHZ.CH
Andreasstrasse 5, 8050 Zürich

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We show hardness of improperly learning halfspaces in the agnostic model, both in the distribution-independent and the distribution-specific setting, based on the assumption that worst-case lattice problems, e.g., approximating the shortest vector to within polynomial factors, are hard. In particular, we show that under this assumption there is no efficient algorithm that outputs any binary hypothesis, not necessarily a halfspace, achieving misclassification error better than $\frac{1}{2} - \gamma$ even if the optimal misclassification error is as small is as small as δ. Here, γ can be smaller than the inverse of any polynomial in the dimension and δ as small as $\exp\left(-\Omega\left(\log^{1-c}(d)\right)\right)$, where $0 < c < 1$ is an arbitrary constant and d is the dimension. For the distribution-specific setting, we show that if the marginal distribution is standard Gaussian, for any $\beta > 0$ learning halfspaces up to error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ takes time at least $d^{\tilde{\Omega}(1/\varepsilon^{2-\beta})}$ under the same hardness assumptions. Similarly, we show that learning degree-ℓ polynomial threshold functions up to error $\mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$ takes time at least $d^{\tilde{\Omega}(\ell^{2-\beta}/\varepsilon^{4-2\beta})}$. $\mathrm{OPT}_{\mathrm{LTF}}$ and $\mathrm{OPT}_{\mathrm{PTF}_\ell}$ denote the best error achievable by any halfspace or polynomial threshold function, respectively.

Our lower bounds qualitatively match algorithmic guarantees and (nearly) recover known lower bounds based on non-worst-case assumptions. Previously, such hardness results Daniely (2016); Diakonikolas et al. (2021) were based on average-case complexity assumptions, specifically, variants of Feige's random 3SAT hypothesis, or restricted to the statistical query model. Our work gives the first hardness results basing these fundamental learning problems on well-understood worst-case complexity assumption. It is inspired by a sequence of recent works showing hardness of learning well-separated Gaussian mixtures based on worst-case lattice problems.

Keywords: Agnostic Learning, Lower Bounds, Worst-Case Hardness

1. Introduction

An important question in theoretical computer science, and in learning theory in particular, is understanding the relation between average-case and worst-case problems (cf. Levin's work on distributional analogs of NP Levin (1986) and Impagliazzo's five worlds Impagliazzo (1995), and also the survey of Bogdanov and Trevisan Bogdanov et al. (2006)). In particular, to understand for what kind of average-case problems we can show hardness based on worst-case assumptions, thus unlocking the power of the machinery of classical worst-case reductions. In this work, we make progress on this question by evidencing a strong connection between fundamental and well-studied learning problems and worst-case assumptions with a plethora of other applications. Specifically, we will show that learning halfspaces and polynomial threshold functions, in either the distribution-independent or distribution-specific setting, are as hard as standard worst-case lattice problems frequently used as a basis of hardness in cryptography Peikert et al. (2016).

There are several barriers for basing the hardness of average-case problems on classical assumptions such as $\mathrm{P} \neq \mathrm{NP}$ Applebaum et al. (2008); Feigenbaum and Fortnow (1993); Bogdanov and Trevisan (2006) and results in the context of learning theory have either been restricted to the PAC learning setting, in which there is no noise, Kearns and Valiant (1994); Klivans and Sherstov (2009)[1], or restricted to hardness results for (semi-)proper learning, where, loosely speaking, the hypothesis output by the algorithm has to be of the same kind as the one which generated the samples Feldman (2006); Feldman et al. (2006); Guruswami and Raghavendra (2006); Gopalan et al. (2010). In fact, there is evidence that this might be inherent Applebaum et al. (2008). On the other hand, there is a plethora of strong hardness results ruling out even *improper* learning algorithms, i.e., that output an arbitrary hypothesis that well-approximates a certain function to be learned, often matching known algorithmic upper bounds. However, these results can be based only on average-case assumptions Kalai et al. (2008); Klivans and Kothari (2014); Daniely (2016); Daniely and Vardi (2021) or be shown for restricted models of computations Diakonikolas et al. (2021). So far it remained unclear if these results can also be based on well-understood worst-case assumptions.

In contrast to this, basing hardness of average-case problems on worst-case assumptions is ubiquitous in cryptography and a highly desirable feature. In particular, many problems are based on worst-case hardness of lattice problems such as the Shortest Independent Vectors Problem (SIVP) or the Gap Shortest Vector Problem (gapSVP) (cf. problems 7 and 8). Roughly speaking, in the second, the task is to decide whether a given lattice contains a non-zero point of small norm and in the first, to find a basis for the lattice in which each vector has small norm. We refer to section 3 for exact definitions. We do not attempt to survey the vast literature on the topic and instead refer to Peikert et al. (2016). Recent breakthrough results Bruna et al. (2021); Gupte et al. (2022) have provided a bridge between these lattice problems and learning problems by showing that a certain Gaussian Mixture Model is hard to learn assuming the worst-case hardness of either SIVP or gapSVP. In this work we extend this bridge by showing that hardness of other fundamental learning problems can also be based on these assumptions. Specifically, assuming worst-case hardness of either of the above lattice problems, we show that weak improper learning of halfspaces in the agnostic model is hard. Further, we extend our results to the setting in which the marginal distribution is fixed to be a standard Gaussian, evidencing that even average-case problems with very specific distributional requirements can be shown to be hard under worst-case assumptions. This second result also extends to learning polynomial threshold functions. Precise definitions will follow below.

The task of agnostically learning a class $\mathcal{C}$ of boolean functions, called a concept class, is defined as follows: Given samples $(\boldsymbol{x}, y) \in \mathbb{R}^M \times \{-1, +1\}$ from an arbitrary distribution D compute a binary hypothesis $h \colon \mathbb{R}^M \to \{-1, +1\}$ achieving small *misclassification error*:

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

In particular, we aim to achieve error close to the minimum misclassification error achieved by any function in $\mathcal{C}$, denoted by $\mathrm{OPT}_{\mathcal{C}}$. We say that h is a *weak learner*, if it achieves error better than $1/2 - 1/\operatorname{poly}(M)$. Concept classes relevant to this work are the ones of all halfspaces, also known as linear threshold functions (LTFs), defined as $\boldsymbol{x} \mapsto \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)$ for some unknown $\boldsymbol{w} \in \mathbb{S}^{M-1}$, and degree-$\ell$ polynomial threshold functions (PTFs), defined as $\boldsymbol{x} \mapsto \operatorname{sign}(p(\boldsymbol{x}))$ for some unknown degree-ℓ polynomial p. Note, that we do not restrict the output hypothesis h to belong to $\mathcal{C}$. This is called *improper* learning and stands in contrast to so-called proper learning for which most hardness results based on worst-case assumption are known. In this work we show strong

[1] We will talk about this a bit more below.

limitations for improperly learning both LTFs and PTFs agnostically under worst-case assumptions. We remark that if $\mathrm{OPT}_{\mathrm{LTF}} = 0$, we can efficiently find a halfspace which achieves arbitrarily small misclassification error Maass and Turán (1994). This can be extended to the case when $\mathrm{OPT}_{\mathrm{LTF}} = O\left(\frac{\log M}{M}\right)$. Our first result states that even if $\mathrm{OPT}_{\mathrm{LTF}}$ is just slightly larger, we cannot output any binary hypothesis which achieves error significantly better than a random guess:

Theorem 1 (Informal version of theorem 9) *Assuming hardness of either* SIVP *or* gapSVP, *there is no* $\mathrm{poly}(M)$*-time algorithm that learns* M*-dimensional halfspaces in the agnostic model up to error* $1/2 - 1/\mathrm{poly}(M)$*. This holds already if* $\mathrm{OPT}_{\mathrm{LTF}}$ *is as small as* $\exp\big(-\log^{1-c}(M)\big)$*, where* $0 < c < 1$ *is an absolute constant.*

Hence, weak improper learning of halfspaces in the agnostic model is likely to be computationally challenging. It is natural to ask whether the problem becomes easier by making stronger distributional assumptions. This turns out to indeed be the case. Specifically, if we restrict to the case that samples $(\boldsymbol{x}, y)$ come from a distribution whose $\boldsymbol{x}$-marginal $D_{\boldsymbol{x}}$ is standard Gaussian, the L_1-regression algorithm from Kalai et al. (2008) is known to learn LTFs up to error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ in time $M^{O(1/\varepsilon^2)}$ and degree-ℓ PTFs up to error $\mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$ in time $M^{O(\ell^2/\varepsilon^4)}$. Our second main result shows that under the same assumptions as in theorem 1, these results are qualitatively tight.

Theorem 2 (Informal version of theorem 12) *Let* $\beta > 0$ *be arbitrary and* $\varepsilon > 0$*. There exists a distribution* D *over* $\mathbb{R}^M \times \{-1, +1\}$ *such that* $D_{\boldsymbol{x}}$ *is standard Gaussian and assuming hardness of either* SIVP *or* gapSVP *there is no* $M^{O(\frac{1}{\log(1/\varepsilon)\cdot\varepsilon^{2-\beta}})}$*-time algorithm which achieves misclassification error* $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ *over* D*. Similarly, there is no* $M^{O(\frac{\ell^{2-\beta}}{\log(\ell/\varepsilon)\cdot\varepsilon^{2-\beta}})}$*-time algorithm which achieves misclassification error* $\mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$ *over* D*.*

Our result is inspired by recent hardness results for learning mixtures of well-separated Gaussians Bruna et al. (2021); Gupte et al. (2022) based on the same worst-case lattice problems. In particular, we show a simple reduction from the *Continuous Learning with Errors* (CLWE) problem introduced in Bruna et al. (2021), a continuous analogue of Regev's Learning with Errors problem (LWE) Regev (2009). Indeed, our hard instance in theorem 1 will correspond to a mixture of (a small modification of) two *homogenous* CLWE distributions. The construction for theorem 2 will be similar. See section 2 for more details.

1.1. Relation to Previous Hardness Results

Our main theorems (almost) match algorithmic upper bounds and (nearly) recover known lower bounds under either average-case hardness assumptions or in restricted models of computation. In essence, we show that for a class of fundamental learning problems there is no price to pay for basing hardness of learning problems on worst-case assumptions. Hardness of improperly weakly learning halfspaces in the agnostic model, quantitatively matching the above theorem exactly, was known under a variant of Feige's random 3SAT hypothesis and when assuming D is supported on the boolean hypercube Daniely (2016). Later a weaker result, that achieving error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ is hard, was shown under a different assumption on the existence of a certain kind of pseudo-random generators Daniely and Vardi (2021).

For the distribution-specific setting, when $D_{\boldsymbol{x}}$ is standard Gaussian, lower bounds were either far from algorithmic guarantees Klivans and Kothari (2014) or only known in the statistical query

(SQ) model Kearns (1998). In particular, it was known that any SQ algorithm achieving error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ needs at least $2^{M^{\Omega(1)}}$ queries or queries of accuracy at $M^{-\Omega(1/\varepsilon^2)}$. Similarly, any SQ algorithm achieving error $\mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$ needs at least $2^{M^{\Omega(1)}}$ queries or queries of accuracy at $M^{-\Omega(\ell/\varepsilon^4)}$ Diakonikolas et al. (2021). This can be seen as evidence that every algorithm solving the above problems needs time at least $2^{M^{\Omega(1)}}$ or $M^{\Omega(1/\varepsilon^2)}$, respectively, $M^{\Omega(\ell^2/\varepsilon^4)}$, samples. This (nearly) matches our lower bounds in theorem 2. We remark that, for learning PTFs, both lower bounds are a $1/\varepsilon^2$ factor away from known upper bounds and closing this gap is an interesting open question.

Hardness Based on Public-Key Cryptosystems. We would like to further highlight the connection of our work to two lines of work for proving lower bounds for learning problems. In a seminal work, Kearns and Valiant pushed forward the idea of basing hardness of learning a concept class $\mathcal{C}$, specifically when $\mathrm{OPT}_{\mathcal{C}} = 0$, on the conjectured security of cryptographic public-key encryption schemes by creating samples for the learning problem by encryption messages oneself Kearns and Valiant (1994). They use this to show that improperly learning boolean formulae and deterministic finite automata is hard assuming, e.g., that breaking the RSA cryptosystem is hard. Later, this approach was used in Klivans and Sherstov (2009) to show that learning the class of intersections of halfspaces is hard assuming cryptosystems based on LWE are hard Regev (2003, 2005), which in turn is implied by hardness of either SIVP or gapSVP. Again assuming $\mathrm{OPT}_{\mathcal{C}} = 0$. Hence, in the case where the public-key encryption scheme used is hard under worst-case assumptions, also the learning problem enjoys the same hardness guarantees. However, there are two shortcomings to this approach: First, we have to find a suitable encryption scheme for a learning problem and additionally, this scheme has to be hard under worst-case assumptions. Second, it is not clear how to extend this method to the agnostic setting studied in this paper, where $\mathrm{OPT}_{\mathcal{C}} > 0$. Our approach gives a more principled approach for establishing the desired hardness guarantees.

Hardness Based on Learning Parities with Noise. Secondly, in the past the Learning Parities with Noise (LPN) problem has played a central role in deriving lower bounds for learning problems. LPN is a special case of LWE whose continuous version we base our lower bounds on. Crucially however, known worst-case hardness results for LWE do not extend to LPN. The following hardness results based on LPN are known: First, Feldman et al. (2006) shows hardness of agnostically learning various boolean functions, not including halfspaces, based on the hardness of a sparse version of LPN - more precisely, that learning parities that depend on only k variables, takes time at least $M^{\Omega(k)}$. Under the same assumption, Klivans and Kothari (2014) shows that agnostically learning halfspaces under the Gaussian distribution up to error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ takes time at least $M^{\Omega(\log(1/\varepsilon))}$. Second, and more relevant to this work, Kalai et al. (2008) shows that for any $\beta > 0$ an algorithm for agnostically learning halfspaces under the uniform distribution over the hypercube that runs in time $M^{O(1/\varepsilon^{2-\beta})}$ implies an algorithm for LPN with constant noise rate running in time roughly $2^{O(M^{1-\beta/2})}$. While LPN certainly is a central problem in the field of learning theory and all of the above assumptions are widely believed to be true, its worst-case hardness remains poorly understood. To the best of our knowledge, there is no worst-case hardness result for the sparse version. The version used by Kalai et al. (2008), was recently shown to be hard under a non-standard version of some worst-case assumption[2] Brakerski et al. (2019); Yu and Zhang (2021). Hence, lower bounds based on LPN can only constitute a weak link between fundamental learning problems and

[2] More specifically, a promise version of the Nearest Codeword Problem with additional assumptions.

worst-case assumptions. It however is a very interesting question, if this link can be strengthened by basing LPN on more standard worst-case assumptions as it is possible for its cousin LWE Regev (2010).

Distributions That Are Hard to Distinguish From a Gaussian. At the core of our results, and more specifically, the CLWE problem (see section 2 for a definition), lies the fact that a certain distribution is hard to distinguish from the standard Gaussian. We remark that this idea is also present in previous lower bound constructions. In particular, the ”parallel pancakes” construction in Diakonikolas et al. (2017b) is the starting point for many lower bounds in the statistical query model Diakonikolas et al. (2019, 2022b); Diakonikolas and Kane (2022); Nasser and Tiegel (2022). A similar construction was used in Bubeck et al. (2019) to show hardness of a certain binary classification problem in the statistical query model. Further, CLWE was used in Song et al. (2021) to show hardness of learning a single periodic neuron.

Lastly, concurrent and independent work Diakonikolas et al. (2022a) shows lower bounds for learning in the so-called Massart model Massart and Nédélec (2006) based on LWE and hence also provides a link between learning and worst-case lattice problems. Previously, such lower bounds were only known in the statistical query model Chen et al. (2020); Diakonikolas and Kane (2022); Nasser and Tiegel (2022).

2. Technical Overview

Continuous Learning with Errors. Before we start describing our lower bound constructions, we introduce the continuous learning with errors (CLWE) problem. Let $\boldsymbol{w}$ be uniform over the unit sphere, $\boldsymbol{y} \sim N(0, I)$, and $\gamma, \beta > 0$ be some parameters. We are given samples $(\boldsymbol{y}, z)$, where

$$z = \gamma \langle \boldsymbol{w}, \boldsymbol{y} \rangle + e \mod 1\,,$$

for $e \sim N(0, \beta^2)$[3]. The task is to distinguish these samples from samples $(\boldsymbol{y}, z)$, where $\boldsymbol{y} \sim N(0, I)$ as well, but z is independently and uniformly at random drawn from $[0, 1)$,[4] For convenience, we call this second distribution $\mathrm{CLWE}^{\mathrm{null}}$. Bruna et al. (2021) gave a (quantum) reduction from approximating the Gap Shortest Vector Problem (GapSVP) or the Shortest Independent Vectors Problem (SIVP) within polynomial factors to CLWE. In Gupte et al. (2022) this was strengthened, for some set of parameters, to a reduction directly from standard LWE implying hardness also when only assuming the classical hardness of the above lattice problems. Both works use the CLWE problem to obtain hardness results for density estimation of well-separated mixtures of Gaussians. As remarked earlier, the idea of designing a distribution that is hard to distinguish earlier also lies at the heart of many statistical query lower bounds. See e.g. the influential work Diakonikolas et al. (2017a) and subsequent works.

Distribution-Independent Setting. We next give a sketch of the proof of theorem 1. First, it is clear that in order to show lower bounds for learning halfspaces, it is enough to show lower bounds for learning polynomial threshold functions over a lower-dimensional space. More specifically, let $M, n, \ell \in \mathbb{N}$ be such that $M = \binom{n+\ell}{n} \leqslant n^\ell$, then any degree-$\ell$ PTF over $\mathbb{R}^n$ can be viewed as a halfspace over $\mathbb{R}^M$ by using an embedding that maps $\boldsymbol{x}$ to the vector containing all monomials of

[3] For ease of notation we have slightly rescaled the problem. See definition 3 for the exact definition we use.

[4] This is called the *decision version*. In the *search* version one asks instead to recover the hidden direction $\boldsymbol{w}$.

degree at most ℓ.[5] In what follows we will choose parameters such that $n \approx \log(M)^{1+c}$ for some constant $c > 0$. Hence, to rule out polynomial-time algorithms, in M, for learning halfspaces over $\mathbb{R}^M$ it is enough to show an exponential lower bound, in n, for learning degree-ℓ PTFs over $\mathbb{R}^n$.

There are two parts to showing theorem 1. We aim to find a distribution D such that: First, in sub-exponential time we cannot compute a binary hypothesis that has misclassification error significantly better than $1/2$ on D and second, there exists a degree-ℓ PTF such that $\mathrm{OPT}_{\mathrm{PTF}_\ell}$ is vanishing. By the discussion above this implies that $\mathrm{OPT}_{\mathrm{LTF}}$ is vanishing as well. We will choose D to correspond to a mixture of variants of the CLWE distribution. In what follows we set $\gamma \geqslant 2\sqrt{n}$ and $\beta = 1/\operatorname{poly}(n)$. Bruna et al. (2021); Gupte et al. (2022) show that for this choice of parameters there is no sub-exponential time algorithm for distinguishing such samples from $N(0, I_n) \times \mathcal{U}([0,1))$ assuming that there is no, quantum or classical, respectively, sub-exponential time algorithm for GapSVP and SIVP.[6]

Moreover, they introduced a variant of the CLWE distribution, which intuitively can be thought of as the CLWE distribution conditioned on $z \approx 0$. This is called the *homogeneous* CLWE (short hCLWE) distribution (cf. definition 4) and will be the basis of our hardness result. They show that it is equal to an infinite mixture of Gaussians and has density roughly proportional to

$$\sum_{k \in \mathbb{Z}} N\left(0, \gamma^2\right)(k) \cdot N\left(0, I_n - \boldsymbol{w}\boldsymbol{w}^\top\right)(\pi_{\boldsymbol{w}^\perp}(\boldsymbol{y})) \cdot N\left(\tfrac{k}{\gamma}, \tfrac{\beta^2}{\gamma^2}\right)(\langle \boldsymbol{w}, \boldsymbol{y} \rangle)\,,$$

where $N(\mu, \Sigma)(\boldsymbol{x})$ denotes the density of $N(\mu, \Sigma)$ evaluated at $\boldsymbol{x}$ and $\pi_{\boldsymbol{w}^\perp}(\boldsymbol{y})$ the projection of $\boldsymbol{y}$ onto the space orthogonal to $\boldsymbol{w}$. Note that the components are equally spaced along direction $\boldsymbol{w}$ with spacing $1/\gamma$ and the k-th component has weight roughly $\exp\left(-k^2/\gamma^2\right)$. Second, along the direction of $\boldsymbol{w}$ they have variance $\approx \beta/\gamma \ll 1/\gamma$, i.e., they are almost non-overlapping, and in all other directions have variance 1. The authors show that under the same hardness assumption, there is no sub-exponential time algorithm that can distinguish the hCLWE distribution from the standard Gaussian.

In particular, let H_0 be the hCLWE distribution. Additionally, let $H_{1/2}$ be obtained in the same way but instead of conditioning on $z \approx 0$ we condition on $z \approx 1/2$. The resulting distribution will be the same as H_0 but the components are shifted along the direction $\boldsymbol{w}$ by $1/(2\gamma)$. Further, it enjoys the same hardness guarantees as H_0. Since $\beta \ll \gamma$ the two distributions will only overlap in a region of exponentially small probability mass. In fact, if we consider the distributions H_0' and $H_{1/2}'$ in which each component of the mixture is truncated such that they are completely disjoint (by some small margin) this only introduces a negligible change in total variation distance. It follows by a standard argument (cf. lemma 20), that H_0' and $H_{1/2}'$ will still be hard to distinguish from a standard Gaussian. Bruna et al. (2021) showed how to obtain samples from H_0 using CLWE samples and their argument straightforwardly extends to obtaining samples from H_1. Hence, we can also obtain samples from the mixture distribution over $\mathbb{R}^n \times \{-1, +1\}$ defined as

$$D = \frac{1}{2} \cdot (H_0, +1) + \frac{1}{2} \cdot \left(H_{1/2}, -1\right)$$

by deciding for each sample whether it should be generated from H_0 or H_1 with probability $1/2$ and setting the label accordingly. Applying this same procedure to samples from $\mathrm{CLWE}^{\mathrm{null}}$, we

[5] This is sometimes referred to as the Veronese mapping, or a feature map.

[6] In the hardness result of Gupte et al. (2022), $\boldsymbol{w}$ is not a random unit vector but rather a random sparse unit vector.

can see that D is hard to distinguish from $D_n^{\text{null}} := N(0, I_n) \times \text{Be}\big(\frac{1}{2}\big)$, where $\text{Be}\big(\frac{1}{2}\big)$ denotes the distribution that is -1 with probability $1/2$ and $+1$ with probability $1/2$. Again, we can instead consider the distribution

$$D' = \frac{1}{2} \cdot \big(H'_0, +1\big) + \frac{1}{2} \cdot \Big(H'_{1/2}, -1\Big) .$$

First, notice that since any learning algorithm has error $1/2$ on D_n^{null} it follows that we cannot compute, in sub-exponential time, a hypothesis with misclassification error significantly better on D' either since otherwise we could distinguish the two distributions. Now that we have established that D' is hard to learn, to show our hardness result, we need to show that there is indeed a PTF which achieves vanishing error. First, note that we can restrict our attention to the direction $\boldsymbol{w}$ by considering a one-dimensional polynomial $p_{\boldsymbol{w}} \colon \mathbb{R} \to \mathbb{R}$ and then obtaining the final polynomial $p \colon \mathbb{R}^n \to \mathbb{R}$ as $p(\boldsymbol{x}) = p_{\boldsymbol{w}}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$. Consider the union of intervals

$$S_+ = \bigcup_{k \in \mathbb{Z}} \left[\tfrac{k}{\gamma} - \alpha, \tfrac{k}{\gamma} + \alpha\right] , \qquad S_- = \bigcup_{k \in \mathbb{Z}} \left[\tfrac{k}{\gamma} + \tfrac{1}{2\gamma} - \alpha, \tfrac{k}{\gamma} + \tfrac{1}{2\gamma} + \alpha\right] ,$$

where $\alpha < 1/(2\gamma)$ is the radius around which we truncate the components. Note that by construction the supports of H'_0 and $H'_{1/2}$ are equal to

$$\text{supp}\big(H'_0\big) = \{\boldsymbol{x} \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_+\} , \qquad \text{supp}\Big(H'_{1/2}\Big) = \{\boldsymbol{x} \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_-\} .$$

Further, let

$$S_+^{(\ell)} = \bigcup_{k=-\ell}^{\ell} \left[\tfrac{k}{\gamma} - \alpha, \tfrac{k}{\gamma} + \alpha\right] , \qquad S_-^{(\ell)} = \bigcup_{k=-\ell}^{\ell-1} \left[\tfrac{k}{\gamma} + \tfrac{1}{2\gamma} - \alpha, \tfrac{k}{\gamma} + \tfrac{1}{2\gamma} + \alpha\right] .$$

Consider the degree-4ℓ polynomial $p_{\boldsymbol{w}}$ that has is positive on $S_+^{(\ell)}$ and negative on $S_-^{(\ell)}$ and positive for points of magnitude larger than those in $S_+^{(\ell)} \cup S_-^{(\ell)}$. By choosing it such that its roots are halfway between the intervals we will have some small margin. Clearly, for $(\boldsymbol{x}, y) \sim D'$ such that $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_+^{(\ell)} \cup S_-^{(\ell)}$ we have $y = \text{sign}(p(\boldsymbol{x}))$ always. The same holds for $(\boldsymbol{x}, y)$ such that $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_+ \setminus S_+^{(\ell)}$. On the flip side, we note that for $(\boldsymbol{x}, y)$ such that $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_- \setminus S_-^{(\ell)}$ we have

$$-1 = y \neq \text{sign}(p(\boldsymbol{x})) = 1$$

always. Hence, the total misclassification error is equal to the probability that $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_- \setminus S_-^{(\ell)}$, This happens if and only if $\boldsymbol{x}$ comes from $H'_{1/2}$ and in particular from a component that doesn't belong to the 2ℓ most central ones. Since the k-th component has weight $\approx \exp\big(-k^2/\gamma^2\big)$ it follows that this event happens with probability roughly $\exp\big(-\ell^2/\gamma^2\big)$. For our choice of parameters we have $\gamma = 2\sqrt{n} \approx \log^{(1+c)/2}(M)$ and $\ell \approx \log(M)$ and hence the error of p becomes

$$\exp\big(-\ell^2/\gamma^2\big) = \exp\Big(-\log^{(1-c)}(M)\Big)$$

as desired.

Distribution-Specific Setting. For the distribution-specific setting (cf. theorem 2), we have the additional requirement that the marginal distribution needs to be standard Gaussian. Note that this implies that the above lifting to PTFs no longer works: Indeed, it even is unclear how the distribution before the lifting should look like so that it is standard Gaussian afterwards. Hence, we work directly with the CLWE problem in dimension M. Recall that this means that $\gamma = 2\sqrt{M}$ and $\beta = 1/\operatorname{poly}(M)$. This time, to preserve the marginal distribution, let H_0 be obtained by conditioning the CLWE distribution on $z \in [0, 1/2)$ and H_1 by conditioning on $z \in [1/2, 1)$. Our hard distribution will be

$$D = \frac{1}{2} \cdot (H_0, +1) + \frac{1}{2} \cdot (H_1, -1) \,.$$

Note that since $[0, 1/2)$ and $[1/2, 1)$ partition $[0, 1)$, it follows that the marginal of D is the same as the marginal distribution of y in CLWE, i.e., standard Gaussian. Note that, given CLWE samples, we can obtain samples from D by rejection sampling. If we apply the same rejection sampling procedure to samples from $\text{CLWE}^{\text{null}}$ we obtain samples from $D_M^{\text{null}} := N(0, I_M) \times \text{Be}\big(\frac{1}{2}\big)$. Hence, a sub-exponential, in M, algorithm to distinguish D and D_M^{null} with non-negligible advantage can be used to distinguish samples from CLWE and $\text{CLWE}^{\text{null}}$.

It remains to show that if we could learn LTFs and PTFs over D up to error better than $\text{OPT} + \varepsilon$ we can distinguish D from D_M^{null}. For this, we first inspect D more closely. As for the distribution-independent setting, the label of samples from D only depends on the direction $\boldsymbol{w}$. Second, let $A_k = [\frac{k}{\gamma}, \frac{k+1/2}{\gamma})$, $B_k = [\frac{k+1/2}{\gamma}, \frac{k+1}{\gamma})$ and

$$S_+ = \bigcup_{k \in \mathbb{Z}} A_k \,, \qquad S_- = \bigcup_{k \in \mathbb{Z}} B_k \,.$$

It turns out that D is sufficiently well approximated (cf. lemma 18) by the distribution D' whose marginal is standard Gaussian and for which it holds that $y = 1$ if and only if $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in S_+$. More specifically, the total variation distance between D and D' is at most $1/\operatorname{poly}(M)$ and hence affects the misclassification error by at most this same additive factor. We hence continue to work with D' below. Regarding LTFs, consider the function f defined as $\boldsymbol{x} \mapsto \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$. For simplicity, denote $z = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. Clearly, this function only misclassifies samples for which either $z \geqslant 0$ and $z \in S_-$ or $z \leqslant 0$ and $z \in S_+$. Let $X \sim N(0, 1)$. By symetry it follows that

$$\operatorname{err}_{D'}(f) = 2\mathbb{P}(z \geqslant 0 \,, z \in S_-) = 2 \sum_{k \geqslant 0} \mathbb{P}(X \in B_k) \,.$$

Notice that for $k \geqslant 0$, $\mathbb{P}(X \in B_k) \leqslant \mathbb{P}(X \in A_k)$ always since the pdf of a one-dimensional Gaussian is decreasing for $z \geqslant 0$. Further, one can show (cf. lemma 19) that for $k \geqslant \gamma$ is it decreasing sufficiently fast such that

$$2\mathbb{P}(X \in B_k) \leqslant \left(1 - \frac{1}{\gamma}\right) \cdot \left[\mathbb{P}(X \in A_k) + \mathbb{P}(X \in B_k)\right] .$$

Hence, we obtain that there exists an absolute constant $c > 0$ such that

$$\operatorname{err}_{D'}(f) \leqslant \sum_{0 \leqslant k < \gamma} \mathbb{P}(X \in A_k) + \mathbb{P}(X \in B_k) + \left(1 - \frac{1}{\gamma}\right) \cdot \sum_{k \geqslant \gamma} \mathbb{P}(X \in A_k) + \mathbb{P}(X \in B_k)$$

$$= \frac{1}{2} - \frac{1}{\gamma} \cdot \mathbb{P}(X \geqslant 1) = \frac{1}{2} - \frac{c}{\sqrt{M}} \,.$$

Hence, for arbitrary $\beta > 0$, an algorithm achieving misclassification error $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$ for $\varepsilon \approx 1/\sqrt{M}$ necessarily needs time at least $2^{\Omega\left(M^{1-\beta}\right)} = M^{\Omega\left(\frac{1}{\varepsilon^{2-\beta} \cdot \log(1/\varepsilon)}\right)}$.

Our argument for degree-ℓ PTFs will be similar. For simplicity, assume that ℓ is even and consider the one-dimensional polynomial p defined as follows: It has roots $-\frac{\ell}{2\gamma}, -\frac{\ell-1}{2\gamma}, \ldots, 0, \ldots, \frac{\ell-1}{2\gamma}, \frac{\ell}{2\gamma}$ and its sign is positive between 0 and $\frac{1}{2\gamma}$ and alternates on the other intervals. For simplicity, also assume without loss of generality that it has positive sign for $z \geqslant \frac{\ell}{2\gamma}$. We define the polynomial threshold function h as $\boldsymbol{x} \mapsto \mathrm{sign}(p(\langle \boldsymbol{w}, \boldsymbol{x} \rangle))$. Let again $X \sim N(0,1)$, by symmetry and using the results above it follows that there exists an absolute constant $c > 0$ such that

$$\begin{aligned}
\mathrm{err}_{D'}(h) = 2 \sum_{k \geqslant \ell/2} \mathbb{P}(X \in B_k) = 2 \sum_{k \geqslant 0} \mathbb{P}(X \in B_k) - 2 \sum_{k < \ell/2} \mathbb{P}(X \in B_k) &\leqslant \frac{1}{2} - \frac{c}{\gamma} - \mathbb{P}\left(\frac{1}{\gamma} \leqslant X \leqslant \frac{\ell/2+1}{\gamma}\right) \\
&\leqslant \frac{1}{2} - \frac{c}{\gamma} - \frac{1}{2} \cdot \mathbb{P}\left(0 \leqslant X \leqslant \frac{\ell/2+1}{\gamma}\right) .
\end{aligned}$$

Since $\ell \ll \gamma$ the pdf of the standard Gaussian is roughly constant between 0 and $\frac{\ell/2+1}{\gamma}$. Hence, it follows that there exists an absolute constant $c' > 0$ such that $\mathrm{err}_{D'}(h) \leqslant \frac{1}{2} - \frac{c'\ell}{\gamma}$. It follows as for LTFs, that, for arbitrary $\beta > 0$, an algorithm achieving misclassification error $\mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$ for $\varepsilon \approx \ell/\sqrt{M}$ necessarily needs time at least $2^{\Omega\left(M^{1-\beta}\right)} = M^{\Omega\left(\frac{\ell^{2-\beta}}{\varepsilon^{2-\beta} \cdot \log(\ell/\varepsilon)}\right)}$.

We remark that in both the LTF and the PTF case, OPT is very close to 1/2. Indeed, this is a property is shared with previous lower bounds and in particular, quantitatively matches the result of Kalai et al. (2008) based on LPN, for the setting in which the marginal distribution is uniform over the boolean hypercube. It would be very desirable to show lower bounds where this is not the case, as for the distribution-independent setting.

3. Preliminaries, CLWE Distributions, and Hardness Assumptions

Notation

We use bold font for vectors and non-bold-font for scalars. 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

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

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

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

Most of the time the distribution D will be clear from context and we will omit the subscript. We denote by $D_{\boldsymbol{x}}$ the marginal distribution of D over $\mathcal{X}$.

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

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

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

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

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

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

where

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

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 $\rho_s(S) = \sum_{x\in S} \rho_s(x)$.

Various Other Distributions

Definition 3 (CLWE Distribution) *Let $\boldsymbol{w} \in \mathbb{R}^n$ be a unit vector and $\beta, \gamma > 0$. Define the distribution $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$ over $\mathbb{R}^n \times [0, 1)$ as follows. Draw $\boldsymbol{y} \sim N(0, \frac{1}{2\pi} \cdot I_n)$, $e \sim N(0, \beta^2/(2\pi))$ and let*

$$z = \gamma\langle \boldsymbol{w}, \boldsymbol{y}\rangle + e \mod 1 \,.$$

Note that the density of this distribution is given by

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

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

Definition 4 (Homogeneous CLWE (hCLWE) Distribution) *Let $\boldsymbol{w} \in \mathbb{R}^n$ be a unit vector, $c \in [0,1)$, and $\beta, \gamma > 0$. Let $\pi_{\boldsymbol{w}^\perp}(\boldsymbol{y})$ be the projection of $\boldsymbol{y}$ onto the space orthogonal to $\boldsymbol{w}$. Define the distribution $\mathrm{H}_{\boldsymbol{w},\beta,\gamma,c}$ over $\mathbb{R}^n$ as having density at $\boldsymbol{y}$ proportional to*

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

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

Note that eq. (3.1) integrates to $Z = \frac{\beta}{\sqrt{\beta^2+\gamma^2}}\cdot\rho_{\sqrt{\beta^2+\gamma^2}}(\mathbb{Z}\,;c)$. Further, eq. (3.1) is equivalent to (see fact 23)

$$\rho(\boldsymbol{y})\cdot\sum_{k\in\mathbb{Z}}\rho_\beta(\gamma\langle\boldsymbol{w},\boldsymbol{y}\rangle\,;k-c)\,. \tag{3.2}$$

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

Definition 5 (Non-Overlapping hCLWE Distribution) *Let $\boldsymbol{w} \in \mathbb{R}^n$ 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 $\mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c}$ over $\mathbb{R}^n$ as having density proportional to*

$$\sum_{k\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c)\cdot\rho(\pi_{\boldsymbol{w}^\perp}(\boldsymbol{y}))\cdot\rho^{\alpha}_{\beta/\sqrt{\beta^2+\gamma^2}}\left(\langle\boldsymbol{w},\boldsymbol{y}\rangle\,;\frac{\gamma}{\beta^2+\gamma^2}(k-c)\right). \tag{3.3}$$

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

Note that this is the same as the hCLWE distribution but with the individual components of the mixture truncated in the hidden direction. By definition of ρ^α eqs. (3.1) and (3.3) integrate to the same value. 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.

Hardness Assumption

We make the following hardness assumption

Assumption 6 *Let $n, m \in \mathbb{N}$ and*

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

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

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

with non-negligible advantage.

Note that Bruna et al. (2021) showed (cf. their Corollary 3.2) that there is a polynomial time (quantum reduction) from approximating either the Shortest Independent Vectors Problem or the Gap Shortest Vector Problem within polynomial factors to CLWE. We will define these problems explicitly below. For more background, we refer to Peikert et al. (2016). It is widely believed that both of these problems do not admit an algorithm, neither classical nor quantum, running in time faster than $2^{\Omega(n)}$ – we remark that there do exist algorithms running in time $2^{O(n)}$. We will invoke the above assumption with $1/2 < \delta < 1$. Bruna et al. (2021) showed that just above this threshold, i.e., when allowing $2^{O(n)}$ time, the CLWE distinguishing problem can be solved.

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

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

We can now define GapSVP and SIVP.

Problem 7 (Gap Shortest Vector Problem (GapSVP)) *Let $\alpha = \mathrm{poly}(n)$ 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 8 (Shortest Independent Vector Problem (SIVP)) *Let $\alpha = \mathrm{poly}(n)$ be arbitrary. Given an n-dimensional lattice L output a set of linearly independent lattice points of length at most $\alpha \cdot \lambda_n(L)$.*

Acknowledgments

The author thanks Rajai Nasser, David Steurer, and Jingqiu Ding for helpful discussions and Lucas Slot for feedback on a draft of this paper.

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

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.

Andrej Bogdanov and Luca Trevisan. On worst-case to average-case reductions for np problems. *SIAM Journal on Computing*, 36(4):1119–1159, 2006.

Andrej Bogdanov, Luca Trevisan, et al. Average-case complexity. *Foundations and Trends® in Theoretical Computer Science*, 2(1):1–106, 2006.

Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 575–584, 2013.

Zvika Brakerski, Vadim Lyubashevsky, Vinod Vaikuntanathan, and Daniel Wichs. Worst-case hardness for lpn and cryptographic hashing via code smoothing. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part III*, pages 619–635. Springer, 2019.

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.

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.

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, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021.

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

Ilias Diakonikolas, Daniel M. Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. Robust sparse mean estimation via sum of squares. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4703–4763. PMLR, 02–05 Jul 2022b. URL https://proceedings.mlr.press/v178/diakonikolas22e.html.

Joan Feigenbaum and Lance Fortnow. Random-self-reducibility of complete sets. *SIAM J. Comput.*, 22(5):994–1005, 1993.

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.

Russell Impagliazzo. A personal view of average-case complexity. In *Proceedings of Structure in Complexity Theory. Tenth Annual IEEE Conference*, pages 134–147. IEEE, 1995.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

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

Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67–95, 1994.

Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

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.

Leonid A Levin. Average case complete problems. *SIAM Journal on Computing*, 15(1):285–286, 1986.

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.

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.

Daniele Micciancio and Chris Peikert. Trapdoors for lattices: Simpler, tighter, faster, smaller. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 700–718. Springer, 2012.

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. An efficient and parallel gaussian sampler for lattices. In *Annual Cryptology Conference*, pages 80–97. Springer, 2010.

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

Oded Regev. New lattice based cryptographic constructions. In *STOC*, pages 407–416. ACM, 2003.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *STOC*, pages 84–93. ACM, 2005.

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

Oded Regev. The learning with errors problem (invited survey). In *IEEE Conference on Computational Complexity*, pages 191–204. IEEE Computer Society, 2010.

Min Jae Song, Ilias Zadik, and Joan Bruna. On the cryptographic hardness of learning single periodic neurons. *Advances in neural information processing systems*, 34:29602–29615, 2021.

Yu Yu and Jiang Zhang. Smoothing out binary linear codes and worst-case sub-exponential hardness for lpn. In *Advances in Cryptology–CRYPTO 2021: 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16–20, 2021, Proceedings, Part III 41*, pages 473–501. Springer, 2021.

Appendix A. Hardness of Distribution-Independent Learning

In this section we are going to prove a formal version of theorem 1. In particular, we will show the following theorem

Theorem 9 *Let $M \in \mathbb{N}$ and $0 < c < c' < 1$ be arbitrary. There exists a distribution D over $\mathbb{R}^M \times \{-1, +1\}$ such that under assumption 6 (with $\frac{1+c}{1+c''} < \delta < 1$) there is no algorithm using fewer than $\exp\big(\Omega\big(\log^{1+c}(M)\big)\big)$ samples and running in time $\exp\big(\Omega\big(\log^{1+c}(M)\big)\big)$ that outputs any binary hypothesis f such that*

$$\mathrm{err}_D(f) \leqslant \frac{1}{2} - \exp\big(-\Omega\big(\log^{1+c}(M)\big)\big)\,.$$

This holds even if there exists a linear threshold function f^* such that

$$\mathrm{err}_D(f^*) \leqslant \exp\Big(-\Omega\Big(\log^{1-c'}(M)\Big)\Big)$$

and for all $x \in \mathbb{R}^M$ in the support of D it holds that

$$\mathbb{P}_{(\boldsymbol{x},y)\sim D}(f^*(x) \neq y \mid \boldsymbol{x}) \in \{0, 1\}\,.$$

We will show hardness by showing that a certain low-degree polynomial threshold function is hard to learn. Hardness of learning halfspaces then follows by embedding this into a higher-dimensional space. Note that the last two properties of the distribution imply that an overwhelming fraction of the observed points is in fact noiseless. More concretely, we will use the following lemma. We provide in a proof in appendix D for completeness.

Lemma 10 *Let $n, d \in \mathbb{N}$ and $M \geqslant n^d$. Further, let D be a distribution over $\mathbb{R}^n \times \{-1, +1\}$. There exists a distribution D' over $\mathbb{R}^M \times \{-1, +1\}$ such that*

1. *For every degree-d polynomial threshold function $h\colon \mathbb{R}^n \to \{-1, +1\}$ there exists a linear threshold function $f\colon \mathbb{R}^M \to \{-1, +1\}$ such that*

$$\mathrm{err}_{D'}\big(f'\big) = \mathrm{err}_D(h)\,.$$

2. *For every binary function $f\colon \operatorname{supp}(D') \to \{-1, +1\}$ there exists a binary function $h\colon \mathbb{R}^n \to \{-1, +1\}$ such that*

$$\mathrm{err}_{D'}\big(f'\big) = \mathrm{err}_D(h)\,.$$

In both cases such a function can be computed in time $\mathrm{poly}(M)$. Moreover, there exists a one-to-one mapping $\phi\colon \operatorname{supp}(D) \to \operatorname{supp}(D')$ such that in both of the above cases for all $\tilde{\boldsymbol{x}}' \in \operatorname{supp}(D')$ it holds that

$$\mathbb{P}_{(\boldsymbol{x}',y')\sim D'}\big(f(\boldsymbol{x}') \neq y' \mid \boldsymbol{x}' = \tilde{\boldsymbol{x}}'\big) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(h(\boldsymbol{x}) \neq y \mid \boldsymbol{x} = \phi^{-1}(\tilde{\boldsymbol{x}}')\big)\,.$$

The hard distribution will correspond to a mixture of two non-overlapping hCLWE instances for an appropriate choice of parameters. More precisely, we will use the following lemma

Lemma 11 *Let $d, n \in \mathbb{N}, \beta, \gamma \in \mathbb{R}_{>0}$ such that*

$$\beta^2 \leqslant \gamma^2 \,, \qquad \frac{d}{\gamma} = \Omega(1) \,.$$

Further, let $c_+ = 0, c_- = 1/2$, and $\boldsymbol{w} \in \mathbb{S}^{n-1}$. Let

$$D_+ = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_+} \,, \qquad D_- = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_-} \,.$$

Let $\mathcal{C}_{4d}$ be the class of all degree-$4d$ polynomial threshold functions (PTFs). Consider the distribution over $\mathbb{R}^n \times \{-1,+1\}$ given by

$$D = \frac{1}{2} \cdot (D_+, +1) + \frac{1}{2} \cdot (D_-, -1) \,.$$

There exists a degree-$4d$ PTF h^* such that

$$\mathrm{err}(h^*) \leqslant \exp\left(-\Omega\left(\frac{d^2}{\gamma^2}\right)\right) .$$

Moreover, it holds that

$$\forall \boldsymbol{x} \in \mathbb{R}^n \colon \quad \mathbb{P}_{(\boldsymbol{x},y)\sim D}(h^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \in \{0,1\} \,.$$

With this in hand, we continue with the proof of theorem 9

Proof [Proof of theorem 9] Let $1 > c' > c'' > c > 0$, $\frac{1+c}{1+c''} < \delta < 1$ and

$$d = \left\lceil \frac{1}{4} \cdot \frac{\delta}{1+c} \cdot \frac{\log M}{\log\log M} \right\rceil ,$$

where C is a large enough universal constant. Further, let n be the largest natural number such that $n^{4d} \leqslant M$. In what follows, we will for simplicity assume that $n^{4d} = M$, all arguments can readily be adapted to the general case. We will show that there exists a distribution D over $\mathbb{R}^n \times \{-1,+1\}$ such that under assumption 6 there is no algorithm using fewer than $\exp\left(n^\delta\right)$ samples and running in time at most $\exp\left(n^\delta\right)$ that outputs any binary hypothesis achieving misclassification error better than $1/2 - \tau$, for

$$\tau = \exp\left(-c_\tau \cdot n^\delta\right) ,$$

for a small enough absolute constant c_τ. Note that this implies the first part of the theorem since

$$n^\delta = \exp\left(\frac{\delta}{4d} \cdot \log M\right) = \exp((1+c) \cdot \log\log M + \Theta(1)) = \Theta\left(\log^{1+c} M\right) \tag{A.1}$$

and hence

$$\exp\left(n^\delta\right) = \exp\left(O\left(\log^{1+c}(M)\right)\right) \qquad \text{and} \qquad \tau = \exp\left(-\Omega\left(\log^{1+c}(M)\right)\right) .$$

For this choice of parameters it also holds that

$$d = \Theta\left(\frac{n^{\delta/(1+c)}}{\log n}\right) \qquad \text{and} \qquad M = \exp\left(n^{\delta/(1+c)}\right) .$$

Let $c_+ = 0, c_- = 1/2, \boldsymbol{w} \in \mathcal{U}\big(\mathbb{S}^{n-1}\big)$ and

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

First, consider

$$D_+ = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_+}\,, \qquad D_- = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_-}\,.$$

We then set

$$D = \frac{1}{2}\cdot(D_+,+1) + \frac{1}{2}\cdot(D_-,-1)\,.$$

Combining assumption 6 and theorem 15 it follows that there is no $O\big(\exp\big(n^\delta\big)\big)$-time distinguisher between D and $N\big(0,\frac{1}{2\pi}\cdot I_n\big)\times \mathrm{Be}\big(\frac{1}{2}\big)$ which uses at most $m = O\big(\exp\big(n^\delta\big)\big)$ samples and has non-negligible advantage. Let D' be the distribution obtained when applying lemma 10 to D. Assume towards a contraction that there is a learning algorithm that using time and samples (from D')

$$\exp\big(O\big(\log^{1+c}(M)\big)\big) = \exp\Big(n^\delta\Big)$$

outputs a binary function $f\colon\ \mathrm{supp}(D') \to \{-1,+1\}$ such that

$$\mathrm{err}_{D'}(f) \leqslant \frac{1}{2} - \exp\big(-\Omega\big(\log^{1+c}(M)\big)\big) = \frac{1}{2} - \tau\,.$$

We claim that we can use this to correctly determine the distribution of the above distinguishing problem in time $O\big(\exp\big(n^\delta\big)\big)$ and with probability at least $2/3$. Indeed, suppose we are given m samples from one of the two distributions. Note that in case they came from $N\big(0,\frac{1}{2\pi}\cdot I_n\big)\times \mathrm{Be}\big(\frac{1}{2}\big)$ the label of the resulting distribution will still be distributed as $\mathrm{Be}\big(\frac{1}{2}\big)$ independently of the example. We first transform the samples using the mapping of lemma 10 and then run our learning algorith on the first $m/2$ samples to obtain a hypothesis f with the guarantees above - for simplicity, assume that m is even. Next, we compute

$$\widehat{\mathrm{err}(f)} = \frac{2}{m}\sum_{i=m/2}^{m} \mathbf{1}(f(x_i) \neq y_i)\,.$$

If

$$\left|\widehat{\mathrm{err}(f)} - \frac{1}{2}\right| > \frac{\tau}{2}$$

we output D and else we output $N\big(0,\frac{1}{2\pi}\cdot I_n\big)\times \mathrm{Be}\big(\frac{1}{2}\big)$. Suppose for now, that the samples come from the distribution D. Then by assumption our learning algorithm outputs a hypothesis h such that

$$\mathrm{err}_{D'}(f) \leqslant \frac{1}{2} - \tau\,.$$

Note that $\widehat{\mathrm{err}(f)}$ is a sum of independent random variables bounded between 0 and 1 and with mean $\mathrm{err}(f)$. Hence, by Hoeffding's Inequality Hoeffding (1994) it follows that

$$\mathbb{P}\Big(\Big|\widehat{\mathrm{err}(f)} - \mathrm{err}(f)\Big| > \frac{\tau}{3}\Big) \leqslant 2\exp\big(-\tfrac{2m}{9}\cdot\tau^2\big) \leqslant \frac{1}{3}\,,$$

where we used that c_τ is a small enough absolute constant. Hence, with probability at least $2/3$ we have that

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

Similary, if the samples come from $N\big(0, \frac{1}{2\pi}\cdot I_n\big) \times \operatorname{Be}\big(\frac{1}{2}\big)$ it follows that $\mathbb{P}_{(\boldsymbol{x}',y)\sim D'}(f(\boldsymbol{x}') \neq y) = 1/2$ and hence

$$\mathbb{P}\Big(\Big|\widehat{\operatorname{err}(f)} - 1/2\Big| > \frac{\tau}{3}\Big) \leqslant \frac{1}{3}\,.$$

Together this yields that

$$\left|\widehat{\operatorname{err}(f)} - \frac{1}{2}\right| \leqslant \frac{\tau}{3} < \frac{\tau}{2}$$

with probability at least $2/3$.

Next, we will show the second part of the theorem. To this end, note that from eq. (A.1) it follows that

$$n = \log^{\frac{1+c}{\delta}}(M)$$

and hence

$$\frac{d}{\gamma} = \Omega\left(\frac{\log^{\left(1-\frac{1+c}{2\delta}\right)} M}{\log\log M}\right) = \Omega\left(\frac{\log^{\frac{1}{2}\cdot(1-c'')} M}{\log\log M}\right) = \Omega(1)\,,$$

where we used that $\frac{1+c}{1+c''} < \delta$ implies that

$$1 - \tfrac{1+c}{2\delta} > \tfrac{1}{2} - \tfrac{1}{2}\cdot c'' > 0\,.$$

Hence, from lemma 11 it follows that there exists a degree-$4d$ PTF h^* satisfying

$$\operatorname{err}_D(h^*) \leqslant \exp\left(-\Omega\left(\frac{d^2}{\gamma^2}\right)\right) = \exp\left(-\frac{\log^{1-c''} M}{\log\log M}\right) = \exp\Big(-\Omega\Big(\log^{1-c'} M\Big)\Big)\,,$$

for c' slightly larger than c''. Further, it holds that

$$\forall \boldsymbol{x} \in \mathbb{R}^n : \quad \mathbb{P}_{(\boldsymbol{x},y)\sim D}(h^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \in \{0,1\}\,.$$

By lemma 10 it follows that for the same distribution D' there exists a linear threshold function $f^* \colon \mathbb{R}^M \to \{-1,+1\}$ which has the same misclassification error and conditional error probabilites (with respect to D') which finishes the proof. ■

It remains to prove lemma 11

Proof [Proof of lemma 11] Let $d, n \in \mathbb{N}, \beta, \gamma \in \mathbb{R}_{>0}$ such that

$$\beta^2 \leqslant \gamma^2\,, \qquad \frac{d}{\gamma} = \Omega(1)\,.$$

Further, let $c_+ = 0, c_- = 1/2$ and $\boldsymbol{w} \in \mathbb{S}^{n-1}$. Recall that

$$D_+ = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_+}\,, \qquad D_- = \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c_-}\,.$$

We will first show that for our choice of parameters the supports of D_+ and D_- are disjoint. To this end, recall that for $c \in [0,1)$ the distribution $\mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c}$ has density proportional to

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

where $\alpha = \frac{1}{10}\cdot\frac{\gamma}{\gamma^2+\beta^2}$ and $\pi_{\boldsymbol{w}^\perp}(\boldsymbol{y})$ denotes the projection of $\boldsymbol{y}$ onto the orthogonal complement of $\boldsymbol{w}$. For $k \in \mathbb{Z}$ let

$$\mu_k^+ = \frac{\gamma}{\beta^2+\gamma^2}(k-c_+) = \frac{\gamma}{\beta^2+\gamma^2}k \qquad \text{and} \qquad \mu_k^- = \frac{\gamma}{\beta^2+\gamma^2}(k-c_-) = \frac{\gamma}{\beta^2+\gamma^2}\left(k-\frac{1}{2}\right).$$

Consider the intervals

$$J_k^+ = \left[\mu_k^+ - \alpha, \mu_k^+ + \alpha\right],$$
$$J_k^- = \left[\mu_k^- - \alpha, \mu_k^- + \alpha\right].$$

Then it follows that

$$\mathrm{supp}(D_+) = \bigcup_{k\in\mathbb{Z}}\left\{\boldsymbol{x}\in\mathbb{R}^n \;\middle|\; \langle\boldsymbol{w},\boldsymbol{x}\rangle \in J_k^+\right\},$$
$$\mathrm{supp}(D_-) = \bigcup_{k\in\mathbb{Z}}\left\{\boldsymbol{x}\in\mathbb{R}^n \;\middle|\; \langle\boldsymbol{w},\boldsymbol{x}\rangle \in J_k^-\right\}.$$

Since the intervals J_k^+, J_k^- are symmetric around μ_k^+ and μ_k^- respectively and

$$\min\{|\mu_k^+ - \mu_k^-|, |\mu_k^+ - \mu_{k+1}^-|\} = \frac{1}{2}\cdot\frac{\gamma}{\beta^2+\gamma^2},$$

it follows that the supports of D_+ and D_- are disjoint if and only if

$$\frac{1}{2}\cdot\frac{\gamma}{\beta^2+\gamma^2} > 2\alpha = \frac{1}{5}\cdot\frac{\gamma}{\beta^2+\gamma^2},$$

which always is the case. Hence, the supports of D_+ and D_- are indeed disjoint.

Consider next the $2d$ intervals $J_{-d+1}^-, \dots, J_d^-$ and the minimum-degree polynomial $p_{\boldsymbol{w}}\colon \mathbb{R}\to\mathbb{R}$ that is zero on exactly the points halfway between one of these intervals and the closest J_k^+ intervals. Further, choose this in such a way that it is non-positive on $J_{-d+1}^-, \dots, J_d^-$. Note by construction it has degree $4d$. Further, consider the degree-$4d$ PTF

$$p\colon \mathbb{R}^n \to \mathbb{R},$$
$$\boldsymbol{x} \mapsto \mathrm{sign}(p_{\boldsymbol{w}}(\langle\boldsymbol{w},\boldsymbol{x}\rangle)).$$

Let

$$S^- = \bigcup_{k=-d+1}^{d}\left\{\boldsymbol{x}\in\mathbb{R}^n \;\middle|\; \langle\boldsymbol{w},\boldsymbol{x}\rangle \in J_k^-\right\}.$$

Note that for all $\boldsymbol{x}$ such that $D_+(\boldsymbol{x}) \neq 0$ it holds that

$$\mathbb{P}_{(\boldsymbol{x},y)\sim D}(p(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) = 0$$

since for such $\boldsymbol{x}$ the label y is always equal to $+1$ and so is the value of p. For the same reason the same holds for all $\boldsymbol{x} \in S^-$. Hence, we obtain that

$$\mathbb{P}_{(\boldsymbol{x},y)\sim D}(p(\boldsymbol{x}) \neq y) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(\boldsymbol{x} \in \operatorname{supp}(D_-) \setminus S^-\big)\,.$$

Let $Z = \frac{\beta}{\sqrt{\beta^2+\gamma^2}} \cdot \rho_{\sqrt{\beta^2+\gamma^2}}(\mathbb{Z}\,;c_-)$ then by definition of D_- and using that for $s > 0$

$$\int_{|z-c|\leqslant\alpha} \rho_s^\alpha(z\,;c)\,dz = \int_{\mathbb{R}} \rho_s(z\,;c)\,dz = s\,.$$

it follows that

$$\begin{aligned}
\mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(\boldsymbol{x} \in \operatorname{supp}(D_-) \setminus S^-\big) &= \frac{1}{Z} \sum_{\substack{k\leqslant -d\,,\\ k\geqslant d+1}} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c_-) \cdot \int_{J_k^-} \rho^\alpha_{\beta/\sqrt{\beta^2+\gamma^2}}\left(z\,; \frac{\gamma}{\beta^2+\gamma^2}(k-c_-)\right) dz \\
&\leqslant \frac{1}{Z} \sum_{|k|\geqslant d} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c_-) \cdot \int_{J_k^-} \rho^\alpha_{\beta/\sqrt{\beta^2+\gamma^2}}\left(z\,; \frac{\gamma}{\beta^2+\gamma^2}(k-c_-)\right) dz \\
&= \frac{1}{Z} \sum_{|k|\geqslant d} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c_-) \cdot \int_{\mathbb{R}} \rho_{\beta/\sqrt{\beta^2+\gamma^2}}\left(z\,; \frac{\gamma}{\beta^2+\gamma^2}(k-c_-)\right) dz \\
&= \frac{\beta}{\sqrt{\beta^2+\gamma^2}\cdot Z} \sum_{|k|\geqslant d} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c_-) \\
&= \frac{1}{\rho_{\sqrt{\beta^2+\gamma^2}}(\mathbb{Z}\,;c_-)} \cdot \sum_{|k|\geqslant d} \rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c_-)\,.
\end{aligned}$$

It follows that

$$\mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(\boldsymbol{x} \in \operatorname{supp}(D_-) \setminus S^-\big) \leqslant \mathbb{P}(|U| \geqslant d)\,,$$

where $U \sim D_{\mathbb{Z}-c_-,\sqrt{\beta^2+\gamma^2}}$. By standard tailbounds for the discrete Gaussian distribution (Micciancio and Peikert, 2012, Lemma 2.8) we conclude that

$$\mathbb{P}(|U| \geqslant d) \leqslant \Theta(1)\cdot\exp\left(-\pi\cdot\frac{d^2}{\beta^2+\gamma^2}\right) = \exp\left(-\Omega\left(\frac{d^2}{\gamma^2}\right)\right),$$

where in the last equality we used that $\beta^2 \leqslant \gamma^2$ and $d/\gamma = \Omega(1)$.

Moreover, all points for which $\mathbb{P}_{(\boldsymbol{x},y)\sim D}(h^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \neq 0$ have their projection onto $\boldsymbol{w}$ in $\operatorname{supp}(D_-) \setminus S^-$. However, since for such $\boldsymbol{x}$ the distribution D always outputs a -1 label, whereas $p(\boldsymbol{x}) = +1$, it follows that for such $\boldsymbol{x}$

$$\mathbb{P}_{(\boldsymbol{x},y)\sim D}(h^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) = 1\,.$$

■

Appendix B. Hardness of Distribution-Specific Learning

In this section, we show hardness results for agnostic learning even when the $\boldsymbol{x}$ marginal distribution is Gaussian based on assumption 6. For consistency with the rest of the paper, we show the result where the marginal distribution is equal to $N(0, \frac{1}{2\pi} \cdot I_M)$ instead of standard Gaussian. Recall that for a distribution D over $\mathbb{R}^M \times \{-1, +1\}$, we denote by $D_{\boldsymbol{x}}$ its marginal distribution over $\mathbb{R}^M$. More specifically, we will show:

Theorem 12 *Let $M \in \mathbb{N}$ and $\varepsilon > 0$ be small enough. There exists a distribution D over $\mathbb{R}^M \times \{-1, +1\}$ such that $D_{\boldsymbol{x}} = N(0, \frac{1}{2\pi} \cdot I_M)$ and under assumption 6 for all $\delta < 1$, there is no algorithm using fewer than*

$$M^{\Omega\left(\frac{1}{\log(1/\varepsilon)} \cdot \left(\frac{1}{\varepsilon^2}\right)^{\delta}\right)}$$

time and samples that outputs any binary hypothesis f such that

$$\mathrm{err}_D(f) \leqslant \mathrm{OPT}_{\mathrm{LTF}} + \varepsilon\,.$$

Further, under the same assumption, there is no algorithm using fewer than

$$M^{\Omega\left(\frac{1}{\log(\ell/\varepsilon)} \cdot \left(\frac{\ell}{\varepsilon^2}\right)^{\delta}\right)}$$

time and samples that outputs any binary hypothesis f such that

$$\mathrm{err}_D(f) \leqslant \mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon\,.$$

The hard distribution D is defined as follows: Let $\gamma = 2\sqrt{M}$, $\beta = \frac{1}{\mathrm{poly}}(M)$ and $\boldsymbol{w}$ be uniform over $\mathbb{S}^{M-1}$.

- Draw a sample $(\boldsymbol{x}, z) \sim \mathrm{C}_{\boldsymbol{w},\beta,\gamma}$.
- If $z \in [0, 1/2)$ output $(\boldsymbol{x}, +1)$, else output $(\boldsymbol{x}, -1)$.

theorem 12 will follow directly by the following two lemmas.

Lemma 13 *Let D be as defined above and $\delta < 1$. Then $D_{\boldsymbol{x}} = N(0, \frac{1}{2\pi} \cdot I_M)$ and under assumption 6 there is no algorithm that uses fewer than 2^{M^δ} time and samples and can distinguish D from $\mathrm{N}(0, \frac{1}{2\pi} \cdot I_M) \times \mathrm{Be}(\frac{1}{2})$ with non-negligible advantage in M.*

Lemma 14 *Let D again be as above and $\varepsilon > 0$ be small enough. Suppose there is an algorithm using fewer than*

$$M^{\Omega\left(\frac{1}{\log(1/\varepsilon)} \cdot \left(\frac{1}{\varepsilon^2}\right)^{\delta}\right)}$$

time and samples that outputs a binary hypothesis f such that $\mathrm{err}_D(f) \leqslant \mathrm{OPT}_{\mathrm{LTF}} + \varepsilon$. Then there is an algorithm that uses the same amount of time and samples that can distinguish D from $N(0, \frac{1}{2\pi} \cdot I_M) \times \mathrm{Be}(\frac{1}{2})$ with non-negligible advantage. Similarly, if there is an algorithm using fewer than

$$M^{\Omega\left(\frac{1}{\log(\ell/\varepsilon)} \cdot \left(\frac{\ell}{\varepsilon^2}\right)^{\delta}\right)}$$

time and samples that outputs a binary hypothesis f such that $\mathrm{err}_D(f) \leqslant \mathrm{OPT}_{\mathrm{PTF}_\ell} + \varepsilon$. Then there is an algorithm that uses the same amount of time and samples that can distinguish D from $N(0, \frac{1}{2\pi} \cdot I_M) \times \mathrm{Be}(\frac{1}{2})$ with non-negligible advantage.

We start by proving lemma 13.

Proof [Proof of lemma 13] We first show that $D_{\boldsymbol{x}} = N(0, \frac{1}{2\pi} \cdot I_M)$. Let p be the density of $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$ and p_D the density of D. For $(\boldsymbol{x}, y) \in \mathbb{R}^M \times \{-1, +1\}$ it holds that

$$p_D(\boldsymbol{x} \mid y) = \begin{cases} p(\boldsymbol{x} \mid z \in [0, 1/2))\,, & \text{if } y = +1, \\ p(\boldsymbol{x} \mid z \in [1/2, 1))\,, & \text{if } y = -1. \end{cases}$$

Let $\boldsymbol{x} \in \mathbb{R}^M$, we compute the density $p_{\boldsymbol{x}}(\boldsymbol{x})$ of $D_{\boldsymbol{x}}$ at point $\boldsymbol{x}$.

$$\begin{aligned} p_{\boldsymbol{x}}(\boldsymbol{x}) &= \frac{1}{2} \cdot p(\boldsymbol{x} \mid z \in [0, 1/2)) + \frac{1}{2} \cdot p(\boldsymbol{x} \mid z \in [1/2, 1)) = \int_0^1 p(\boldsymbol{x}, c)\, dc \\ &= \rho(\boldsymbol{x}) = N(0, \tfrac{1}{2\pi} \cdot I_M)(\boldsymbol{x})\,. \end{aligned}$$

Further, let $m = 2^{M^\delta}$. Given a T-time distinguisher $\mathcal{A}$ between

$$D^m \quad \text{and} \quad N(0, \tfrac{1}{2\pi} \cdot I_M)^m \times \mathrm{Be}\big(\tfrac{1}{2}\big)^m$$

we construct a $O(T)$-time distinguisher between

$$\mathrm{CLWE}(\boldsymbol{w}, \beta, \gamma) \quad \text{and} \quad N(0, \tfrac{1}{2\pi} \cdot I_M)^m \times U([0, 1))^m\,.$$

Given samples $(\boldsymbol{x}, z)$ from either $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$ or $N(0, \frac{1}{2\pi} \cdot I_M) \times U([0, 1))$, we construct new samples $(\boldsymbol{x}', y')$ as follows.

- If $z \in [0, 1/2)$ output $(\boldsymbol{x}, +1)$,
- else output $(\boldsymbol{x}, -1)$.

In case $(\boldsymbol{x}, z)$ came from $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$, $(\boldsymbol{x}', y')$ will be distributed according to D by definition. In case $(\boldsymbol{x}, z)$ came from $N(0, \frac{1}{2\pi} \cdot I_M) \times U([0, 1))$, $\boldsymbol{x}'$ and y' will be independent and with marginals $N(0, \frac{1}{2\pi} \cdot I_M)$ and $\mathrm{Be}\big(\frac{1}{2}\big)$, respectively, as desired. Hence, we can directly use our distinguisher $\mathcal{A}$ to distinguish the two cases. ■

Next, we will proof lemma 14

Proof [Proof of lemma 14] We start by proving the result about LTFs, the result about PTFs will follow in the same way. Let $\delta > 0$ and $\tau = \frac{1}{\mathrm{poly}(M)}$. Suppose $\mathrm{OPT}_{\mathrm{LTF}}$ and ε are such that

$$\mathrm{err}_D(f) \leqslant \mathrm{OPT}_{\mathrm{LTF}} + \varepsilon \leqslant \tfrac{1}{2} - \tau\,.$$

We proceed similarly to the proof of theorem 9. Given $m = 2^{M^\delta}$[7] samples $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_m, y_m)$ from either D or $N(0, \frac{1}{2\pi} \cdot I_M) \times \mathrm{Be}\big(\frac{1}{2}\big)$ we run our algorithm on the first $m/2$ samples to obtain a binary hypothesis f. Let

$$\widehat{\mathrm{err}(f)} = \frac{2}{m} \sum_{i=m/2+1}^{m} \mathbf{1}(f(\boldsymbol{x}_i) \neq y_i)\,.$$

[7] For simplicity assume that m is even.

If $|\widehat{\mathrm{err}(f)} - \frac{1}{2}| > \frac{\tau}{2}$, output D, else output $N(0, \frac{1}{2\pi} \cdot I_M) \times \mathrm{Be}\big(\frac{1}{2}\big)$. By an application of Hoeffding's Inequality, it follows as in the proof of theorem 9, that this test successfully distinguishes between the two distributions with probability at least $2/3$.

Assume, that for an absolute constant $c > 0$, it holds that

$$\mathrm{OPT}_{\mathrm{LTF}} \leqslant \frac{1}{2} - \frac{c}{\gamma} = \frac{1}{2} - \frac{c}{2\sqrt{M}}\,.$$

We will verify this shortly. This implies, that we can choose $\varepsilon = \Omega\Big(1/\sqrt{M}\Big)$ and it still holds that $\mathrm{OPT}_{\mathrm{LTF}} + \varepsilon \leqslant \frac{1}{2} - \tau$. Since

$$2^{M^\delta} = M^{\Omega\left(\frac{1}{\log(1/\varepsilon)} \cdot \left(\frac{1}{\varepsilon^2}\right)^\delta\right)}\,,$$

the result will follow.

We next turn to bounding $\mathrm{OPT}_{\mathrm{LTF}}$. First, note that the density of D is equal to

$$p_D(\boldsymbol{x}, y) = \frac{1}{\beta}\rho(\boldsymbol{x}) \cdot \begin{cases} \sum_{k\in\mathbb{Z}} \int_0^{1/2} \rho_\beta(c + k - \gamma\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, dc\,, & \text{if } y = +1, \\ \sum_{k\in\mathbb{Z}} \int_{1/2}^{1} \rho_\beta(c + k - \gamma\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, dc\,, & \text{if } y = -1. \end{cases}$$

To simplify the analysis we will work with the distribution D' whose density is equal to

$$p_{D'}(\boldsymbol{x}, y) = \rho(\boldsymbol{x}) \cdot \begin{cases} \sum_{k\in\mathbb{Z}} \mathbf{1}(\gamma\langle \boldsymbol{w}, \boldsymbol{y}\rangle \in [k, k+1/2))\,, & \text{if } y = +1, \\ \sum_{k\in\mathbb{Z}} \mathbf{1}(\gamma\langle \boldsymbol{w}, \boldsymbol{y}\rangle \in [k+1/2, k+1))\,, & \text{if } y = -1. \end{cases}$$

By lemma 18 it holds that $\mathrm{TVD}(D, D') \leqslant \frac{1}{\mathrm{poly}(M)}$ and hence if $\tilde{f}$ is any linear threshold function it holds that

$$\mathrm{err}_D\Big(\tilde{f}\Big) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}\Big(\tilde{f}(\boldsymbol{x}) \neq y\Big) \leqslant \mathbb{P}_{(\boldsymbol{x},y)\sim D'}\Big(\tilde{f}(\boldsymbol{x}) \neq y\Big) + \frac{1}{\mathrm{poly}(M)} = \mathrm{err}_{D'}\Big(\tilde{f}\Big) + \frac{1}{\mathrm{poly}(M)}\,.$$

Finally, in lemma 19 we show that there exists a halfspace f^* and an absolute constant $c' > 0$ such that $\mathrm{err}_{D'}(f^*) \leqslant \frac{1}{2} - \frac{c'}{\gamma}$ implying that there is a second absolute constant $c > 0$ such that

$$\mathrm{OPT}_{\mathrm{LTF}} \leqslant \mathrm{err}_D(f^*) \leqslant \frac{1}{2} - \frac{c}{\gamma}\,.$$

Next, we turn to the result about PTFs. Analogously as above, it follows from lemma 19 that there exists a degree-ℓ PTF, such that $\mathrm{err}_{D'}(f^*) \leqslant \frac{1}{2} - \frac{c'\ell}{\gamma}$. Hence, in this case, there is an absolute constant $c > 0$, such that $\mathrm{OPT}_{\mathrm{PTF}_\ell} \leqslant \frac{1}{2} - \frac{c\ell}{\gamma}$. Hence, we can choose $\varepsilon = \Omega\Big(\ell/\sqrt{M}\Big)$, implying that

$$2^{M^\delta} = M^{\Omega\left(\frac{1}{\log(\ell/\varepsilon)} \cdot \left(\frac{\ell}{\varepsilon}\right)^{2\delta}\right)}$$

■

Appendix C. Reductions

The goal of this section is to show that the mixture distributions which we used to prove hardness of learning in the agnostic model are hard to learn under assumption 6. In particular, our goal will be to prove the following theorem

Theorem 15 *Let $n, m \in \mathbb{N}$ with $2^n > m > n$, and let $\gamma, \beta, \varepsilon \in \mathbb{R}_{>0}, c \in [0,1)$ such that*

$$0 \leqslant \beta \leqslant \gamma\,, \\ \beta = \tfrac{1}{\mathrm{poly}(n)}\,.$$

Assume that there is no $(T + \mathrm{poly}(n,m))$-time distinguisher between

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

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

$$D_c \coloneqq \mathrm{NHCLWE}\big(m', \gamma, 2\beta, c\big) \quad \textit{and} \quad N\big(0, \tfrac{1}{2\pi} \cdot I_n\big)^m$$

with advantage $\varepsilon - \mathrm{negl}(n)$. Moreover, let $c_+, c_- \in [0,1)$. Then there is no T-time distinguisher between

$$\frac{1}{2} \cdot \big(D_{c_+}, +1\big) + \frac{1}{2} \cdot \big(D_{c_-}, -1\big) \quad \textit{and} \quad N\big(0, \tfrac{1}{2\pi} \cdot I_n\big) \times \mathrm{Be}\left(\frac{1}{2}\right)$$

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

One key ingredient is the following straightforward adaptation of (Bruna et al., 2021, Lemma 4.1). We will include its proof for completeness at the end of this section.

Lemma 16 (straightforward extension of Lemma 4.1 in Bruna et al. (2021)) *For every $\boldsymbol{w} \in \mathbb{R}^n$ there is a $\mathrm{poly}(n, 1/\delta)$-time probabilistic algorithm that takes as input parameters $\delta \in (0,1), c \in [0,1)$, and samples from $\mathrm{C}_{\boldsymbol{w},\gamma,\beta}$ and outputs samples from $\mathrm{H}_{\boldsymbol{w},\sqrt{\beta^2+\delta^2},\gamma,c}$. More specifically, given $\mathrm{poly}(n, 1/\delta)$ CLWE samples the algorithm runs in time $\mathrm{poly}(n, 1/\delta)$ and with probability at least $1 - \exp(-\mathrm{poly}(n, 1/\delta))$ outputs at least one HCLWE sample. Further, if given samples from $N\big(0, \frac{1}{2\pi} \cdot I_n\big) \times \mathcal{U}([0,1))$ the procedure will output samples from $N\big(0, \frac{1}{2\pi} \cdot I_n\big)$.*

With this in had, we can prove theorem 15

Proof [Proof of theorem 15] Let $\gamma, \beta, \varepsilon \in \mathbb{R}_{>0}, c \in [0,1)$. Assume that there is no $(T + \mathrm{poly}(n,m))$-time distinguisher between

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

with advantage ε. Let $m' = \frac{m}{\mathrm{poly}(n)}$. We claim that this implies that there is no T-time distinguisher between

$$\mathrm{HCLWE}\big(m', \gamma, 2\beta, c\big) \quad \text{and} \quad N\big(0, \tfrac{1}{2\pi} \cdot I_n\big)^{m'}$$

with advantage $\varepsilon - \mathrm{negl}(n)$. Note that this implies the conclusion of the theorem since by the second part of lemma 17 the total variation distance between $\mathrm{H}_{\boldsymbol{w},\beta,\gamma,c}$ and $\mathrm{NH}_{\boldsymbol{w},\gamma,\beta,c}$ is at most

$$4 \cdot \exp\left(-\frac{1}{100\beta^2}\right)$$

for every $\boldsymbol{w}$ that is unit. Hence, since $m < 2^n$ the total variation distance between the respective m'-fold product distributions is at most

$$4m \cdot \exp\left(-\frac{1}{100\beta^2}\right) = \operatorname{negl}(n)\,.$$

Note that we can apply this since in our case $0 \leqslant \beta \leqslant \gamma$. See lemma 20 for a formal proof of the fact that a small change in total variation distance results in only a small change in the distinguishing advantage.

To show the claim, we will use lemma 16. Concretely, assume that there is a T-time distinguisher between

$$\mathrm{HCLWE}\big(m', \gamma, 2\beta, c\big) \quad \text{and} \quad N\big(0, \tfrac{1}{2\pi} \cdot I_n\big)^{m'}\,.$$

We will use this to build a $(T + \operatorname{poly}(n, m))$-time distinguisher between

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

as follows: Let $\boldsymbol{w}$ denote the secret vector of the CLWE distribution. Given m samples from either $\mathrm{C}_{\boldsymbol{w},\gamma,\beta}$ or $N\big(0, \frac{1}{2\pi} \cdot I_n\big) \times U([0,1))$ we invoke the algorithm of lemma 16 with

$$\delta = \sqrt{3}\beta = \Omega(1/\operatorname{poly}(n))\,.$$

In case the samples came from $\mathrm{C}_{\boldsymbol{w},\gamma,\beta}$ with probability at least

$$1 - \exp(-\operatorname{poly}(n, 1/\delta)) = 1 - \operatorname{negl}(n)$$

we obtain in time $\operatorname{poly}(n)$ at least $m' = \frac{m}{\operatorname{poly}(n)}$ samples from $\mathrm{H}_{\boldsymbol{w},2\beta,\gamma,c}$. In case the samples came from $N\big(0, \frac{1}{2\pi} \cdot I_n\big)^n \times U([0,1))$ with at least the same probability we obtain in time $\operatorname{poly}(n)$ at least $m' = \frac{m}{\operatorname{poly}(n)}$ samples from D_1^n. Hence, if we had a T-time distinguisher between

$$\mathrm{HCLWE}\big(m', \gamma, 2\beta, c\big) \quad \text{and} \quad N\big(0, \tfrac{1}{2\pi} \cdot I_n\big)^{m'}$$

with advantage $\varepsilon - \operatorname{negl}(n)$, this would directly yield a $(T + \operatorname{poly}(n, m))$-time distinguisher between

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

with advantage ε. The shift of $\operatorname{negl}(n)$ in the advatage is due to the fact that the sample conversion algortihm can fail with probability $\operatorname{negl}(n)$.

For the second part of the theorem, we first note, that we can again replace the truncated mixture distributions by the non-truncated ones by invoking lemma 20. By construction, the respective mixture distributions have total variation distance at most

$$p \cdot \operatorname{negl}(n) + (1 - p) \cdot \operatorname{negl}(n) = \operatorname{negl}(n)\,.$$

The result follows since we can generate samples from this mixture from samples from the CLWE distribution as follows: With probabily p we invoke the procedure of lemma 16 with $c = c_+$ and with probability $1 - p$ with $c = c_-$. ■

Finally, we give the proof of lemma 16

Proof Let $\delta \in (0,1), c \in [0,1)$ and $\beta, \gamma > 0$. Without loss of generality assume that $\boldsymbol{w} = \boldsymbol{e}_1$. Given samples from $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$ the idea is to perform rejection sampling to obtain samples from $\mathrm{H}_{\boldsymbol{w},\sqrt{\beta^2+\delta^2},\gamma,c}$. Concretely, let $g\colon [0,1) \to [0,1]$ be given by $g(z) = g_0(z)/M$, where

$$g_0(z) = \sum_{k\in\mathbb{Z}} \rho_\delta(z+k+c)\,, \qquad M = \sup_{z\in[0,1)} g_0(z)\,.$$

For a CLWE sample $(\boldsymbol{y}, z)$, output $\boldsymbol{y}$ with probability $g(z)$.[8] Recall that the density of $\mathrm{C}_{\boldsymbol{w},\beta,\gamma}$ is given by

$$p(\boldsymbol{y}, z) = \frac{1}{\beta} \cdot \rho(\boldsymbol{y}) \cdot \sum_{k\in\mathbb{Z}} \rho_\beta(z+k-\gamma\boldsymbol{y}_1)\,.$$

Using fact 22 (in the third equality) and that for all $c \in \mathbb{R}$ it holds that $\int_{\mathbb{R}} \rho_s(x-c)\,dx = s$ we obtain that the density p' of the distribution given by the rejection sampling, i.e., of outputting $\boldsymbol{y}$ and accept, is given by

$$\begin{aligned}
p'(\boldsymbol{y}) &= \int_{[0,1)} p(\boldsymbol{y},z)g(z)\,dz = \frac{\rho(\boldsymbol{y})}{\beta\cdot M} \cdot \int_{[0,1)} \sum_{k_1,k_2\in\mathbb{Z}} \rho_\beta(z+k_1-\gamma\boldsymbol{y}_1)\cdot\rho_\delta(z+k_2+c)\,dz \\
&= \frac{\rho(\boldsymbol{y})}{\beta\cdot M} \cdot \int_{[0,1)} \sum_{k_1,k_2\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\delta^2}}(\gamma\boldsymbol{y}_1-k_1+k_2+c)\rho_{\beta\delta/\sqrt{\beta^2+\delta^2}}\Big(z+\tfrac{\beta^2}{\beta^2+\gamma^2}(k_2+c)+\tfrac{\gamma^2}{\beta^2+\gamma^2}(k_1-\gamma\boldsymbol{y}_1)\Big)\,dz \\
&= \frac{\rho(\boldsymbol{y})}{\beta\cdot M} \cdot \int_{[0,1)} \sum_{k,k_2\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\delta^2}}(\gamma\boldsymbol{y}_1-k+c)\cdot\rho_{\beta\delta/\sqrt{\beta^2+\delta^2}}\Big(z+k_2+\tfrac{\beta^2}{\beta^2+\delta^2}c+\tfrac{\delta^2}{\beta^2+\delta^2}(k-\gamma\boldsymbol{y}_1)\Big)\,dz \\
&= \frac{\rho(\boldsymbol{y})}{\beta\cdot M} \cdot \sum_{k\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\delta^2}}(\gamma\boldsymbol{y}_1-k+c)\cdot\int_{\mathbb{R}} \rho_{\beta\delta/\sqrt{\beta^2+\delta^2}}\Big(x+\tfrac{\beta^2}{\beta^2+\delta^2}c+\tfrac{\delta^2}{\beta^2+\delta^2}(k-\gamma\boldsymbol{y}_1)\Big)\,dx \\
&= \frac{\delta\cdot\rho(\boldsymbol{y})}{\sqrt{\beta^2+\delta^2}\cdot M} \sum_{k\in\mathbb{Z}} \rho_{\sqrt{\beta^2+\delta^2}}(\gamma\boldsymbol{y}_1\,;k-c)\,.
\end{aligned}$$

Hence, the distribution is indeed equal to $\mathrm{H}_{\boldsymbol{w},\sqrt{\beta^2+\delta^2},\gamma,c}$. It also follows, that the probability that we accept a given CLWE sample is equal to

$$\begin{aligned}
\int_{\mathbb{R}^n} p'(\boldsymbol{y})\,d\boldsymbol{y} &= \frac{\delta}{\sqrt{\beta^2+\delta^2}\cdot M} \cdot \frac{\sqrt{\beta^2+\delta^2}}{\sqrt{\beta^2+\delta^2+\gamma^2}} \cdot \rho\Big(\tfrac{1}{\sqrt{\beta^2+\delta^2+\gamma^2}}\mathbb{Z}\Big) \\
&= \frac{\delta}{\sqrt{\beta^2+\delta^2+\gamma^2}\cdot M} \cdot \rho\Big(\tfrac{1}{\sqrt{\beta^2+\delta^2+\gamma^2}}\mathbb{Z}\Big) \\
&= \frac{\delta}{\sqrt{\beta^2+\delta^2+\gamma^2}\cdot M} \cdot \rho_{\sqrt{\beta^2+\delta^2+\gamma^2}}(\mathbb{Z})
\end{aligned}$$

Note that using fact 21 it follows that

$$\rho_{\sqrt{\beta^2+\delta^2+\gamma^2}}(\mathbb{Z}) = \sqrt{\beta^2+\delta^2+\gamma^2}\cdot\rho_{1/\sqrt{\beta^2+\delta^2+\gamma^2}}(\mathbb{Z}) \geqslant \sqrt{\beta^2+\delta^2+\gamma^2}\,.$$

[8] Note that by (Brakerski et al., 2013, Section 5.2) the function $g(z)$ is efficiently computable.

Hence, the probability that we accept is at least δ/M. Further, for each $z \in [0,1)$ we have that

$$\begin{aligned} g_0(z) &= \sum_{k\in\mathbb{Z}} \rho_\delta(z+k+c) \\ &\leqslant 2 \cdot \sum_{k=0}^{\infty} \rho_\delta(k) \\ &< 2 \cdot \sum_{k=0}^{\infty} \exp(-\pi k) < 4\,. \end{aligned}$$

Hence, $M \leqslant 4$ and it follows that we accept with probability at least $\delta/4$. Thus, after $\mathrm{poly}(n, 1/\delta)$ we output at least one HCLWE sample with probability at least $1 - \exp(-\mathrm{poly}(n, 1/\delta))$.

Lastly, when given samples from $N\big(0, \frac{1}{2\pi} \cdot I_n\big) \times \mathcal{U}([0,1))$ the procedure will output samples from $N\big(0, \frac{1}{2\pi} \cdot I_n\big)$ since in this case $\boldsymbol{y}$ and z are independent. ■

Appendix D. Missing Lemmas

D.1. TVD Closeness of Supporting Distributions

Lemma 17 *Let $\boldsymbol{w} \in \mathbb{R}^d$ and $0 \leqslant \beta \leqslant \gamma, c \in [0,1)$. Then*

$$\mathrm{TVD}(\mathrm{H}_{\boldsymbol{w},\beta,\gamma,c}, \mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c}) \leqslant 4 \cdot \exp\left(-\frac{1}{100\beta^2}\right).$$

Proof Let P_H denote the density of $\mathrm{H}_{\boldsymbol{w},\beta,\gamma,c}$ and P_N the density of $\mathrm{NH}_{\boldsymbol{w},\beta,\gamma,c}$. Abusing notation slightly, we also use P_H and P_N to refer to the marginal of P_H and P_N on the span of $\boldsymbol{w}$. Since the densities agree in the space orthogonal to $\boldsymbol{w}$, and since they factorize over these two spaces, we obtain

$$\mathrm{TVD}(P_H, P_N) = \sup_{A \subseteq \mathrm{Span}(\boldsymbol{w})} |P_H(A) - P_N(A)|\,.$$

For ease of notation, we identify the span of w with the real line. Further, for $k \in \mathbb{Z}$, let $P_{H,k}$ and $P_{N,k}$ denote the density of the k-th component of P_H and P_N respectively. Further, let w_k denote the weight of the k-th component - which is the same in both cases. It follows that

$$\begin{aligned} \mathrm{TVD}(P_H, P_N) = \sup_{A\subseteq\mathbb{R}} |P_H(A) - P_N(A)| &= \sup_{A\subseteq\mathbb{R}} \left|\sum_{k\in\mathbb{Z}} w_k (P_{H,k}(A) - P_{N,k}(A))\right| \\ &\leqslant \sum_{k\in\mathbb{Z}} w_k \cdot \sup_{A\subseteq\mathbb{R}} |P_{H,k}(A) - P_{N,k}(A)| \end{aligned}$$

Next, fix $k \in \mathbb{Z}$ and let I_k denote the support of $P_{N,k}$. Let $Z = P_{H,k}(I_k)$ and recall that for $C \subseteq I_k$ it holds that $P_{N,k}(C) = \frac{1}{Z} \cdot P_{H,k}(C)$ and for C disjoint from I_k that $P_{N,k}(C) = 0$. We can then bound

$$\sup_{A\subseteq\mathbb{R}} |P_{H,k}(A) - P_{N,k}(A)| = \sup_{A\subseteq\mathbb{R}} |P_{H,k}(A \cap I_k) + P_{H,k}(A \cap I_k^c) - P_{N,k}(A \cap I_k) - P_{N,k}(A \cap I_k^c)|$$

$$\begin{aligned}&\leqslant \sup_{A\subseteq\mathbb{R}} |P_{H,k}(A\cap I_k) - P_{N,k}(A\cap I_k)| + \sup_{A\subseteq\mathbb{R}} |P_{H,k}(A\cap I_k^c) - P_{H,k}(A\cap I_k^c)|\\ &= \frac{1-Z}{Z}\cdot P_{H,k}(I_k) + P_{H,k}(I_k^c) = 2\cdot(1-Z)\,.\end{aligned}$$

Let $\mu_k = \frac{\gamma}{\beta^2+\gamma^2}(k-c)$ and $\sigma_k^2 = \frac{1}{2\pi}\cdot\frac{\beta^2}{\beta^2+\gamma^2}$ and denote by X_k the random variable distributed according to $P_{H,k}$. Note that $X_k \sim N(\mu_k, \sigma_k^2)$ and $I_k = [\mu_k - \alpha, \mu_k - \alpha]$, where $\alpha = \frac{1}{10}\cdot\frac{\gamma}{\gamma^2+\beta^2}$. It follows that

$$\begin{aligned}1 - Z = \mathbb{P}(|X_k| \geqslant \alpha) \leqslant 2\cdot\exp\left(-\frac{\alpha^2}{2\sigma_k^2}\right) &= 2\cdot\exp\left(-\frac{2\pi\cdot\gamma^2\cdot\left(\beta^2+\gamma^2\right)}{200\beta^2\cdot(\beta^2+\gamma^2)^2}\right) \leqslant 2\cdot\exp\left(-\frac{\gamma^2}{50\beta^2\cdot(\beta^2+\gamma^2)}\right)\\ &\leqslant 2\cdot\exp\left(-\frac{1}{100\beta^2}\right).\end{aligned}$$

Hence, we finally obtain that

$$\mathrm{TVD}(P_H, P_N) \leqslant \left(\sum_{k\in\mathbb{Z}} w_k\right)\cdot 4\cdot\exp\left(-\frac{1}{100\beta^2}\right) = 4\cdot\exp\left(-\frac{1}{100\beta^2}\right).$$

■

Lemma 18 *Let D, D' be distributions over $\mathbb{R}^M \times \{-1,+1\}$ with densities defined below, then $\mathrm{TVD}(D, D') \leqslant \frac{1}{\mathrm{poly}(M)}$. The densities are equal to*

$$p_D(\boldsymbol{x}, y) = \frac{1}{\beta}\rho(\boldsymbol{x})\cdot\begin{cases}\sum_{k\in\mathbb{Z}}\int_0^{1/2}\rho_\beta(c+k-\gamma\langle\boldsymbol{w},\boldsymbol{x}\rangle)\,dc\,, & \textit{if } y=+1,\\ \sum_{k\in\mathbb{Z}}\int_{1/2}^{1}\rho_\beta(c+k-\gamma\langle\boldsymbol{w},\boldsymbol{x}\rangle)\,dc\,, & \textit{if } y=-1.\end{cases}$$

and

$$p_{D'}(\boldsymbol{x}, y) = \rho(\boldsymbol{x})\cdot\begin{cases}\sum_{k\in\mathbb{Z}}\mathbf{1}(\gamma\langle\boldsymbol{w},\boldsymbol{x}\rangle\in[k,k+1/2))\,, & \textit{if } y=+1,\\ \sum_{k\in\mathbb{Z}}\mathbf{1}(\gamma\langle\boldsymbol{w},\boldsymbol{x}\rangle\in[k+1/2,k+1))\,, & \textit{if } y=-1.\end{cases}$$

Proof The proof proceeds similary to lemma 17. First, note that by symmetry

$$\begin{aligned}\mathrm{TVD}(D, D') &= \max\left\{\sup_{A\subseteq\mathbb{R}^M}|\mathbb{P}_D(\boldsymbol{x}\in A, y=\ell) - \mathbb{P}_{D'}(\boldsymbol{x}\in A, y=\ell)|\;\middle|\;\ell\in\{-1,+1\}\right\}\\ &= \sup_{A\subseteq\mathbb{R}^M}|\mathbb{P}_D(\boldsymbol{x}\in A, y=+1) - \mathbb{P}_{D'}(\boldsymbol{x}\in A, y=+1)|\\ &= \sup_{A\subseteq\mathrm{Span}(\boldsymbol{w})}|\mathbb{P}_D(\langle\boldsymbol{w},\boldsymbol{x}\rangle\in A, y=+1) - \mathbb{P}_{D'}(\langle\boldsymbol{w},\boldsymbol{x}\rangle\in A, y=+1)|\,.\end{aligned}$$

Without loss of generality, identfiy the span of $\boldsymbol{w}$ with the real line. Abusing notation, we denote the one-dimensional densities by p_D and $p_{D'}$ as well. Define $I_k(z) := \int_0^{1/2}\rho_\beta(c+k-\gamma z)\,dc$ and observe that

$$p_D(z) = \frac{1}{\beta}\rho(z)\cdot\sum_{k\in\mathbb{Z}} I_k(z)\,, \qquad\qquad p_{D'}(z) = \rho(z)\cdot\sum_{k\in\mathbb{Z}}\mathbf{1}(\gamma z\in[k,k+1/2])\,.$$

It follows that $\mathrm{TVD}(D, D') = \mathrm{TVD}(p_D, p_{D'})$ (referring to the one-dimensional densities). To bound this quantity, we introduce the following intermediate distribution $\tilde{D}$ defined as follows: For $z \in \mathbb{R}$ let $k^*(z) = \operatorname{argmin}_{k \in \mathbb{Z}} \min\{|\gamma z - k|, |\gamma z - k - 1/2|\}$, then we set

$$p_{\tilde{D}}(z) \propto \frac{1}{\beta} \rho(z) \cdot I_{k^*(z)}(z) \,.$$

We will first show that $\mathrm{TVD}(p_D, p_{\tilde{D}}) \leqslant \exp(-\operatorname{poly}(M))$ and second that $\mathrm{TVD}(p_{\tilde{D}}, p_{D'}) \leqslant \frac{1}{\operatorname{poly}(M)}$ which together imply the desired result. We will use that for measure P, Q it holds that $\mathrm{TVD}(P, Q) \leqslant \sqrt{2}\mathrm{H}(P, Q)$, where $\mathrm{H}(P, Q)$ is the Hellinger distance defined as $\mathrm{H}(P, Q) = \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$ for p, q the densitites of the measures P and Q respectively.

Let Z be the normalization constant in the density $p_{\tilde{D}}$, it follows that

$$Z = \int_{\mathbb{R}} \frac{1}{\beta} \rho(z) \cdot I_{k^*(z)}(z) \, dz = 1 - \int_{\mathbb{R}} \frac{1}{\beta} \rho(z) \cdot \sum_{k \neq k^*(z)} I_k(z) \, dz \,.$$

Note that for a given z and $k \neq k^*(z)$ it holds that γz is at distance at least $\frac{|k^*(z) - k|}{4}$ from the interval $[k, k + 1/2]$. Hence, we can bound $I_k(z)$ as follows:

$$I_k(z) = \int_0^{1/2} \rho_\beta(c + k - \gamma z) \, dc \leqslant \frac{1}{2} \rho_\beta\left(\tfrac{|k^*(z) - k|}{4}\right) = \exp\Big(-(k^*(z) - k)^2 \operatorname{poly}(n)\Big) \,.$$

It follows that

$$\sum_{k \neq k^*(z)} I_k(z) \leqslant \sum_{k \geqslant 1} \exp(-k \cdot \operatorname{poly}(n)) \leqslant \exp(-\operatorname{poly}(n)) \,,$$

implying that $Z \geqslant 1 - \exp(-\operatorname{poly}(n))$. Using this we can bound the Hellinger distance

$$\begin{aligned}
\mathrm{H}(D, \tilde{D}) &= \frac{1}{2\beta} \int_{\mathbb{R}} \rho(z) \left(\sqrt{\frac{I_{k^*(z)}(z)}{Z}} - \sqrt{\sum_{k \in \mathbb{Z}} I_k(z)}\right)^2 dz \\
&\leqslant \frac{1}{\beta} \left[\int_{\mathbb{R}} \rho(z) \left(\sqrt{\frac{I_{k^*(z)}(z)}{Z}} - \sqrt{I_{k^*(z)}(z)}\right)^2 dz + \int_{\mathbb{R}} \rho(z) \left(\sqrt{I_{k^*(z)}(z)} - \sqrt{\sum_{k \in z} I_k(z)}\right)^2 dz\right] \\
&\leqslant \exp(-\operatorname{poly}(n)) \,,
\end{aligned}$$

where in the last inequality we used that $\int_{\mathbb{R}} \rho(z) \, dz = 1$ and $I_{k^*(z)}(z) \leqslant 1$ for all z.

Next, we turn to bound $\mathrm{H}(\tilde{D}, D')$. To this end, let $\tau > 0$ be some parameter to be chosen later and note that $\sum_{k \in \mathbb{Z}} \mathbf{1}(\gamma z \in [k, k + 1/2]) = \mathbf{1}(\gamma z \in [k^*(z), k^*(z) + 1/2])$. We obtain

$$\mathrm{H}(\tilde{D}, D') = \frac{1}{2} \int_{\mathbb{R}} \rho(z) \left(\sqrt{\frac{1}{\beta} I_{k^*(z)}(z)} - \mathbf{1}(\gamma z \in [k^*(z), k^*(z) + 1/2])\right)^2 dc$$

To begin with, note that

$$\frac{1}{\beta} I_{k^*(z)}(z) \leqslant \int_{\mathbb{R}} \frac{1}{\beta} \rho_\beta(c + k - \gamma z) \, dc \leqslant 1 \,.$$

We proceed by making a case distinction. First, consider z such that $\gamma z \in [k^*(z)+\tau, k^*(z)+1/2-\tau]$ and let $X \sim N(0, \frac{\beta}{2\pi})$. For such z it holds that

$$\begin{aligned}\frac{1}{\beta} I_{k^*(z)}(z) &= \int_0^{1/2} \frac{1}{\beta}\rho_\beta(c + k^*(z) - \gamma z)\, dc = \mathbb{P}\big(k^*(z) - \gamma z \leqslant X \leqslant k^*(z) + \tfrac{1}{2} - \gamma z\big) \\ &\geqslant 1 - \mathbb{P}(|X| \geqslant \tau) \geqslant 1 - 2\exp\Big(-\tfrac{\pi\tau^2}{\beta^2}\Big)\,.\end{aligned}$$

Next, consider z such that $\min\big\{|\gamma z - k^*(z)|, |\gamma z - k^*(z) - \frac{1}{2}|\big\} \geqslant \tau$. We obtain that

$$\frac{1}{\beta} I_{k^*(z)}(z) \leqslant \mathbb{P}(|X| \geqslant \tau) \leqslant 2\exp\Big(-\tfrac{\pi\tau^2}{\beta^2}\Big)\,.$$

Let $S = \bigcup_{k\in\mathbb{Z}} [k-\tau, k+\tau] \cup [k + \frac{1}{2} - \tau, k + \frac{1}{2} + \tau]$. Using the above, we can bound

$$\frac{1}{2}\int_{\mathbb{R}\setminus S} \rho(z)\left(\sqrt{\frac{1}{\beta} I_{k^*(z)}(z)} - \mathbf{1}(\gamma z \in [k^*(z), k^*(z) + 1/2])\right)^2 dc \leqslant \exp\left(-\frac{\pi\tau^2}{\beta^2}\right).$$

It remains to bound the integral on S. For this, note that ρ is symmetric around $z = 0$ and monotone for $z \geqslant 0$ and $z \leqslant 0$. This yields

$$\begin{aligned}&\frac{1}{2}\int_{S} \rho(z)\left(\sqrt{\frac{1}{\beta} I_{k^*(z)}(z)} - \mathbf{1}(\gamma z \in [k^*(z), k^*(z) + 1/2])\right)^2 dc \\ &\leqslant 4\left[\int_0^\tau \rho(z)\, dz + \int_{1/2-\tau}^{1/2+\tau} \rho(z)\, dz + \sum_{k\geqslant 1}\int_{k-\tau}^{k+\tau} \rho(z)\, dz + \int_{k+1/2-\tau}^{k+1/2+\tau} \rho(z)\, dz\right] \\ &\leqslant 4\left[\int_0^\tau \rho(z)\, dz + \frac{1}{\frac{1}{4\tau} - 1}\int_{\tau}^{1/2+\tau} \rho(z)\, dz + \frac{1}{\frac{1}{4\tau} - 1}\sum_{k\geqslant 1}\int_{k-1/2+\tau}^{k+\tau} \rho(z)\, dz + \int_{k+\tau}^{k+1/2+\tau} \rho(z)\, dz\right] \\ &\leqslant 12\tau \int_0^\infty \rho(z)\, dz = 6\tau\,.\end{aligned}$$

Hence, combining the above bounds and choosing $\tau = \sqrt{\beta} = \frac{1}{\mathrm{poly}(M)}$, we obtain

$$\mathrm{H}(\tilde{D}, D') \leqslant \exp\left(-\frac{\pi\tau^2}{\beta^2}\right) + 6\tau = \exp(-\operatorname{poly}(M)) + \frac{1}{\mathrm{poly}(M)}$$

as desired.

■

D.2. Supporting Lemmas about Optimal Halfspaces

Lemma 19 *Consider the distribution D' over $\mathbb{R}^M \times \{-1, 1\}$ with density given by*

$$p_{D'}(\boldsymbol{x}, y) = \rho(\boldsymbol{x}) \cdot \begin{cases} \sum_{k\in\mathbb{Z}} \mathbf{1}(\gamma\langle \boldsymbol{w}, \boldsymbol{x}\rangle \in [k, k+1/2))\,, & \textit{if } y = +1, \\ \sum_{k\in\mathbb{Z}} \mathbf{1}(\gamma\langle \boldsymbol{w}, \boldsymbol{x}\rangle \in [k+1/2, k+1))\,, & \textit{if } y = -1. \end{cases}$$

Let $f^*(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)$, then there exists an absolute constant $c > 0$ such that $\operatorname{err}_{D'}(f^*) \leqslant \frac{1}{2} - \frac{c}{\gamma}$.

Further, for each $\ell \geqslant 2$ there exists a degree-ℓ PTF h^* such that $\operatorname{err}_{D'}(h^*) \leqslant \frac{1}{2} - \frac{c'\ell}{\gamma}$, for some absolute constant $c' > 0$.

Proof First note, that for $(\boldsymbol{x}, y) \sim D'$ only depends on $\langle \boldsymbol{w}, \boldsymbol{x}\rangle$. Let $z = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$ and $A_k = [\frac{k}{\gamma}, \frac{k+1/2}{\gamma}]$, $B_k = [\frac{k+1/2}{\gamma}, \frac{k+1}{\gamma}]$ for $k \in \mathbb{Z}$. Further, let $X \sim N(0, \frac{1}{2\pi})$.

We first prove the result about linear threshold functions. By symmetry it holds that

$$\operatorname{err}_{D'}(f^*) = 2\mathbb{P}(f^*(\boldsymbol{x}) \neq y, y = -1) = 2\sum_{k\in\mathbb{Z}} \int_{B_k} \mathbf{1}(f^*(z) \neq -1)\rho(z)\, dz = 2\sum_{k\geqslant 0} \mathbb{P}(X \in B_k)$$

Note that by construction, for $k \geqslant 0$, it holds that $2\mathbb{P}(X \in B_k) \leqslant \mathbb{P}(X \in B_k) + \mathbb{P}(X \in A_k)$. Let k^* be a non-negative integer to be chosen later and assume that there exists $\varepsilon = \varepsilon(k^*)$ such that $2\mathbb{P}(X \in B_k) \leqslant (1-\varepsilon)[\mathbb{P}(X \in B_k) + \mathbb{P}(X \in A_k)]$, then for $k^* \geqslant \gamma$

$$\begin{aligned}
2\sum_{k\geqslant 0} \mathbb{P}(X \in B_k) &\leqslant \sum_{0\leqslant k<k^*} \mathbb{P}(X \in B_k) + \mathbb{P}(X \in A_k) + (1-\varepsilon)\cdot \sum_{k\geqslant k^*} \mathbb{P}(X \in B_k) + \mathbb{P}(X \in A_k) \\
&= \frac{1}{2} - \mathbb{P}\left(X \geqslant \frac{k^*}{\gamma}\right) + (1-\varepsilon)\cdot \mathbb{P}\left(X \geqslant \frac{k^*}{\gamma}\right) = \frac{1}{2} - \varepsilon \cdot \mathbb{P}\left(X \geqslant \frac{k^*}{\gamma}\right) \\
&\leqslant \frac{1}{2} - \varepsilon \cdot \frac{k^*/\gamma}{2\pi(k^*/\gamma)^2 + 1} \exp\left(-\frac{\pi\cdot(k^*)^2}{\gamma^2}\right) \\
&\leqslant \frac{1}{2} - \varepsilon \cdot \frac{\gamma}{4\pi k^*} \exp\left(-\frac{\pi\cdot(k^*)^2}{\gamma^2}\right),
\end{aligned}$$

where we also used standard bounds for the pdf of the standard Gaussian distribution. Next, we aim to find ε and calculate

$$\begin{aligned}
\mathbb{P}(X \in A_k) - \mathbb{P}(X \in B_k) &= \int_{k/\gamma}^{(k+1/2)/\gamma} \exp\left(-\pi z^2\right) dz - \int_{(k+1/2)/\gamma}^{(k+1)/\gamma} \exp\left(-\pi z^2\right) dz \\
&= \int_{(k+1/2)/\gamma}^{(k+1)/\gamma} \left[\exp\left(-\pi(z - 1/(2\gamma))^2\right) - \exp\left(-\pi z^2\right)\right] dz \\
&= \int_{(k+1/2)/\gamma}^{(k+1)/\gamma} \exp\left(-\pi z^2\right) \cdot \left[\exp\left(\pi\frac{z}{\gamma}\right)\exp\left(-\frac{\pi}{4\gamma^2}\right) - 1\right] dz
\end{aligned}$$

For $k \geqslant 1$ we can bound

$$\begin{aligned}
\exp\left(\pi\frac{z}{\gamma}\right)\exp\left(-\frac{\pi}{4\gamma^2}\right) \geqslant \left(1 + \pi\frac{z}{\gamma}\right)\left(1 - \frac{\pi}{4\gamma^2}\right) &= 1 + \pi\frac{z}{\gamma} - \frac{\pi}{4\gamma^2} - \pi^2\frac{z}{4\gamma^3} \\
&\geqslant 1 + \pi\frac{z}{2\gamma} - \frac{\pi}{4\gamma^2} \geqslant 1 + \pi\frac{k}{4\gamma^2}\,.
\end{aligned}$$

Which implies

$$\mathbb{P}(X \in A_k) - \mathbb{P}(X \in B_k) \geqslant \pi\frac{k}{4\gamma^2}\mathbb{P}(X \in B_k)\,.$$

Rearringing implies that

$$\mathbb{P}(X \in B_k) \leqslant \frac{2}{2+\pi\frac{k}{4\gamma^2}} \cdot (\mathbb{P}(X \in B_k) + \mathbb{P}(X \in A_k))$$

which yields $\varepsilon(k) \geqslant \pi\frac{k}{12\gamma^2} \geqslant \frac{k}{4\gamma^2}$. Hence, for $k^* = \lceil\gamma\rceil$ we have

$$\varepsilon(k^*) \cdot \frac{\gamma}{4\pi k^*} \exp\left(-\frac{\pi \cdot (k^*)^2}{\gamma^2}\right) \geqslant \frac{1}{4\gamma} \cdot \frac{1}{4\pi} \cdot \exp(-4\pi)\,.$$

As desired, this implies that for an absolute constant $c > 0$ it holds that

$$\mathrm{err}_{D'}(f^*) \leqslant \frac{1}{2} - \frac{c}{\gamma}\,.$$

Next, we prove the result about degree-ℓ PTFs. Again, since the labels of D' only depend on the direction $\boldsymbol{w}$ it suffices to define a one-dimensional degree-ℓ polynomial $p_z\colon \mathbb{R} \to \mathbb{R}$. The final PTF will be defined as $h^*(\boldsymbol{x}) = \operatorname{sign} p_z(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)$. Note that p_z can be fully specified by ℓ roots and the sign it takes between any two roots. For simplicity, we assume that ℓ is odd and consider degree-$2\ell+1$ PTFs, the even case works analogously. Let p_z be the polynomial that has roots $-\frac{\ell}{2\gamma}, -\frac{\ell-1}{2\gamma}, \ldots, 0, \ldots, \frac{\ell-1}{2\gamma}, \frac{\ell}{2\gamma}$. Further, let its sign be positive between 0 and $\frac{1}{2\gamma}$ and alternate on the other intervals. Again, for simplicity and without loss of generality, also assume that its sign after the greatest positive root is positive. Observe that this implies that ℓ is even. Note that for $\ell = 0$ we recover the LTF from above. Let $c > 0$ be some absolute constant. By symmetry and the results above it follows that, note that we use that h^* agrees with the label of all samples $(\boldsymbol{x}, y)$ such that $\langle \boldsymbol{w}, \boldsymbol{x}\rangle \leqslant 0$ and $y = -1$.

$$\begin{aligned}\mathrm{err}_{D'}(f^*) &= 2\mathbb{P}(f^*(\boldsymbol{x}) \neq y, y=-1) = 2\sum_{k\in\mathbb{Z}} \int_{B_k} \mathbf{1}(f^*(z) \neq -1)\rho(z)\,dz = 2\sum_{k\geqslant \ell/2} \mathbb{P}(X \in B_k) \\ &= 2\sum_{k\geqslant 0} \mathbb{P}(X \in B_k) - 2\sum_{k<\ell/2} \mathbb{P}(X \in B_k) \leqslant \frac{1}{2} - \frac{c}{\gamma} - \mathbb{P}\left(\frac{1}{\gamma} \leqslant X \leqslant \frac{\ell/2+1}{\gamma}\right) \\ &\leqslant \frac{1}{2} - \frac{c}{\gamma} - \frac{1}{2} \cdot \mathbb{P}\left(X \leqslant \frac{\ell+1}{2\gamma}\right).\end{aligned}$$

Using that $\ell/\gamma \leqslant 1$ we bound as before

$$\begin{aligned}\tfrac{1}{2} \cdot \mathbb{P}\left(X \leqslant \frac{\ell+1}{2\gamma}\right) &\geqslant \frac{(\ell+1)/\gamma}{4\pi((\ell+1)/\gamma)^2+1} \exp\left(-\frac{\pi \cdot (\ell+1)^2}{\gamma^2}\right) \\ &\geqslant \frac{\ell+1}{36\gamma} \exp(-\pi)\,.\end{aligned}$$

Hence, there exists an absolute constant $c' > 0$ such that

$$\mathrm{err}_{D'}(h^*) \leqslant \frac{1}{2} - \frac{c'\ell}{\gamma}\,.$$

■

Next, we prove lemma 10.

Lemma *[Restatement of lemma 10] Let $n, d \in \mathbb{N}$ and $M \geqslant n^d$. Further, let D be a distribution over $\mathbb{R}^n \times \{-1,+1\}$. There exists a distribution D' over $\mathbb{R}^M \times \{-1,+1\}$ such that*

1. *For every degree-d polynomial threshold function $h\colon \mathbb{R}^n \to \{-1,+1\}$ there exists a linear threshold function $f\colon \mathbb{R}^M \to \{-1,+1\}$ such that*
$$\mathrm{err}_{D'}\big(f'\big) = \mathrm{err}_D(h)\,.$$
2. *For every binary function $f\colon\ \mathrm{supp}(D') \to \{-1,+1\}$ there exists a binary function $h\colon \mathbb{R}^n \to \{-1,+1\}$ such that*
$$\mathrm{err}_{D'}\big(f'\big) = \mathrm{err}_D(h)\,.$$

In both cases such a function can be computed in time $\mathrm{poly}(M)$. Moreover, there exists a one-to-one mapping $\phi\colon\ \mathrm{supp}(D) \to \mathrm{supp}(D')$ such that in both of the above cases for all $\tilde{\boldsymbol{x}}' \in \mathrm{supp}(D')$ it holds that

$$\mathbb{P}_{(\boldsymbol{x}',y')\sim D'}\big(f(\boldsymbol{x}') \neq y' \;\big|\; \boldsymbol{x}' = \tilde{\boldsymbol{x}}'\big) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(h(\boldsymbol{x}) \neq y \;\big|\; \boldsymbol{x} = \phi^{-1}(\tilde{\boldsymbol{x}}')\big)\,.$$

Proof We start by describing the mapping ϕ. Denote by $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ a multi-index and by $|\alpha| = \sum_{i=1}^n \alpha_i$ its size. Let $M' = \binom{n+d}{n}$ and let

$$\begin{aligned}\phi\colon \mathbb{R}^n &\to \mathbb{R}^M\,,\\ \boldsymbol{x} &\mapsto \Big(\big(\boldsymbol{x}^\alpha\big)_{|\alpha|\leqslant d}, \boldsymbol{0}\Big)\,,\end{aligned}$$

where by $\boldsymbol{0}$ we mean the vector containing $M - M'$ zeros. Define the distribution D' over $\mathbb{R}^M \times \{-1,+1\}$ as first drawing $(\boldsymbol{x}, y) \sim D$ and then outputting $(\phi(\boldsymbol{x}), y)$. Clearly, restricted to the support of D, the map ϕ is a bijection between $\mathrm{supp}(D)$ and $\mathrm{supp}(D')$.

Next, consider any degree-d polynomial threshold function $h\colon \mathbb{R}^n \to \{-1,+1\}$. Since $M' = \binom{n+d}{n} \leqslant M$ there exists a linear threshold function f such that for all $\boldsymbol{x} \in \mathbb{R}^n$ it holds that $h(\boldsymbol{x}) = f(\phi(\boldsymbol{x}))$. It follows that

$$\mathrm{err}_{D'}\big(f'\big) = \mathbb{P}_{(\boldsymbol{x}',y')\sim D'}\big(f(\boldsymbol{x}') \neq y'\big) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}(f(\phi(\boldsymbol{x})) \neq y) = \mathbb{P}_{(\boldsymbol{x},y)\sim D}(h(\boldsymbol{x}) \neq y) = \mathrm{err}_D(f)\,.$$

Similarly, for every binary function $f\colon\ \mathrm{supp}(D') \to \{-1,+1\}$ we can define the binary function $h\colon \mathbb{R}^n \to \{-1,+1\}$ such that $h(\boldsymbol{x}) = f(\phi(\boldsymbol{x}))$. Hence, we have

$$\mathrm{err}_{D'}\big(f'\big) = \mathrm{err}_D(f)\,.$$

Since in both cases we have to consider at most M coefficients we can compute the linear/polynomial threshold function in time $\mathrm{poly}(M)$. Moreover, in both cases, for $\tilde{\boldsymbol{x}}' \in \mathrm{supp}(D')$ it holds that

$$\begin{aligned}\mathbb{P}_{(\boldsymbol{x}',y')\sim D'}\big(f(\boldsymbol{x}') \neq y' \;\big|\; \boldsymbol{x}' = \tilde{\boldsymbol{x}}'\big) &= \mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(f(\phi(\boldsymbol{x})) \neq y \;\big|\; \phi(\boldsymbol{x}) = \tilde{\boldsymbol{x}}'\big)\\ &= \mathbb{P}_{(\boldsymbol{x},y)\sim D}\big(h(\boldsymbol{x}) \neq y \;\big|\; \boldsymbol{x} = \phi^{-1}(\tilde{\boldsymbol{x}}')\big)\,,\end{aligned}$$

as desired.

■

D.3. Small Facts

Lemma 20 *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

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

which is a contradiction. ■

Fact 21 (Poisson Summation Formula) *For any lattice L and any function f it holds that*

$$f(L) = \det(L^*) \cdot \hat{f}(L^*)$$

where $L^* = \{\boldsymbol{y} \in \mathbb{R}^n \mid \langle \boldsymbol{x}, \boldsymbol{y}\rangle \in \mathbb{Z}\, \text{for all}\ \boldsymbol{x} \in \mathbb{Z}\}$ is the dual lattice of L and $\hat{f}$ the Fourier transform of f.

Fact 22 (Peikert (2010)) *For any $r_1, r_2 > 0$ and vectors $\boldsymbol{x}, \boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathbb{R}^n$, let $r_0 = \sqrt{r_1^2 + r_2^2}, r_3 = \frac{r_1 r_2}{r_0}$, and $\boldsymbol{c}_3 = \frac{r_3^2}{r_1^2}\boldsymbol{c}_1 + \frac{r_3^2}{r_2^2}\boldsymbol{c}_2$. Then*

$$\rho_{r_1}(\boldsymbol{x} - \boldsymbol{c}_1) \cdot \rho_{r_2}(\boldsymbol{x} - \boldsymbol{c}_2) = \rho_{r_0}(c_1 - c_2) \cdot \rho_{r_3}(\boldsymbol{x} - \boldsymbol{c}_3)\,.$$

Fact 23 *Let $\gamma, \beta \geqslant 0$ and $\boldsymbol{w} \in \mathbb{R}^d$, then*

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

Proof Clearly, for $\boldsymbol{y}$ orthogonal to $\boldsymbol{w}$ the equality holds. Consider any $\boldsymbol{y}$ in the span of $\boldsymbol{w}$ and for convenience write $z = \langle \boldsymbol{y}, \boldsymbol{w}\rangle$. Fix $k \in \mathbb{Z}$ then we have that

$$\rho_{\sqrt{\beta^2+\gamma^2}}(k\,; c) \cdot \rho_{\beta/\sqrt{\beta^2+\gamma^2}}\left(z\,; \frac{\gamma}{\beta^2+\gamma^2}(k-c)\right) = \exp\left(-\pi\left[\frac{(k-c)^2}{\beta^2+\gamma^2} + \frac{(\beta^2+\gamma^2)\cdot\left(z - \frac{\gamma}{\beta^2+\gamma^2}(k-c)\right)^2}{\beta^2}\right]\right).$$

Focusing only on the expression inside the exponential function (and ignoring the π) we obtain

$$\begin{aligned}\frac{(k-c)^2}{\beta^2+\gamma^2} + \frac{(\beta^2+\gamma^2)\cdot\left(z - \frac{\gamma}{\beta^2+\gamma^2}(k-c)\right)^2}{\beta^2} &= \frac{(k-c)^2\cdot\beta^2 + \left[(\beta^2+\gamma^2)\cdot z - \gamma\cdot(k-c)\right]^2}{(\beta^2+\gamma^2)\cdot\beta^2} \\ &= \frac{(\beta^2+\gamma^2)\cdot z^2 + (k-c)^2 - 2\cdot(k-c)\cdot\gamma\cdot z}{\beta^2} \\ &= \frac{((k-c) - \gamma\cdot z)^2}{\beta^2} + z^2\,.\end{aligned}$$

Hence, it follows that

$$\rho_{\sqrt{\beta^2+\gamma^2}}(k\,;c)\cdot\rho_{\beta/\sqrt{\beta^2+\gamma^2}}\left(z\,;\frac{\gamma}{\beta^2+\gamma^2}(k-c)\right)=\rho_\beta(\gamma\cdot z\,;k-c)\cdot\rho(z)$$

which implies the claim. ■