Near-Optimal Cryptographic Hardness of Agnostically Learning Halfspaces and ReLU Regression under Gaussian Marginals

Ilias Diakonikolas∗
UW Madison
ilias@cs.wisc.edu

Daniel M. Kane†
UC San-Diego
dakane@ucsd.edu

Lisheng Ren‡
UW Madison
lren29@wisc.edu

February 14, 2023

Abstract

We study the task of agnostically learning halfspaces under the Gaussian distribution. Specifically, given labeled examples \((x, y)\) from an unknown distribution on \(\mathbb{R}^n \times \{\pm 1\}\), whose marginal distribution on \(x\) is the standard Gaussian and the labels \(y\) can be arbitrary, the goal is to output a hypothesis with 0-1 loss \(OPT + \epsilon\), where \(OPT\) is the 0-1 loss of the best-fitting halfspace. We prove a near-optimal computational hardness result for this task, under the widely believed sub-exponential time hardness of the Learning with Errors (LWE) problem. Prior hardness results are either qualitatively suboptimal or apply to restricted families of algorithms. Our techniques extend to yield near-optimal lower bounds for related problems, including ReLU regression.

∗Supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant.
†Supported by NSF Medium Award CCF-2107547, NSF Award CCF-1553288 (CAREER), a Sloan Research Fellowship, and a grant from CasperLabs.
‡Supported by NSF Award CCF-1652862 (CAREER) and a DARPA Learning with Less Labels (LwLL) grant.
1 Introduction

A halfspace or Linear Threshold Function (LTF) is any Boolean-valued function $f : \mathbb{R}^n \to \{\pm 1\}$ of the form $f(x) = \text{sign}(\langle w, x \rangle - t)$, where $w \in \mathbb{R}^n$ is called the weight vector and $t \in \mathbb{R}$ is called the threshold. Here the univariate function $\text{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\text{sign}(u) = 1$ if $u \geq 0$ and $\text{sign}(u) = -1$ otherwise. The task of learning an unknown halfspace is a classical problem in machine learning that has been extensively studied since the 1950s, starting with the Perceptron algorithm [Rosenblatt 1958], and has lead to practically important techniques such as SVMs [Vapnik 1998] and AdaBoost [FS 1997]. In the realizable setting [Valiant 1984], halfspaces are known to be efficiently learnable (see, e.g., [MT 1994]) without distributional assumptions. In contrast, in the distribution-free agnostic AdaBoost [FS 1997]. In the realizable setting [Valiant 1984], halfspaces are known to be efficiently learnable (see, e.g., [MT 1994]) without distributional assumptions. In contrast, in the distribution-free agnostic AdaBoost [FS 1997].

Due to this computational intractability, a significant branch of research has focused on agnostically adversarial label noise. Weak model [Hausler 1992, KSS 1994], even (see, e.g., [MT 1994]) without distributional assumptions. In contrast, in the distribution-free agnostic AdaBoost [FS 1997]. In the realizable setting [Valiant 1984], halfspaces are known to be efficiently learnable (see, e.g., [MT 1994]) without distributional assumptions. In contrast, in the distribution-free agnostic AdaBoost [FS 1997].

Prior Work on Problem 1.1

By standard results [Hausler 1992, KSS 1994], it follows that the sample complexity of the agnostic learning problem for halfspaces is $O(n/\epsilon^2)$. The $L_1$-regression algorithm of [KKMS 2008] solves Problem 1.1 with sample complexity and running time $n^{O(1/\epsilon^2)}$ [DGJ 2010]. While the $L_1$-regression algorithm is not proper, recent work developed a proper learner with qualitatively similar sample and time complexities (namely, $n^{\text{poly}(1/\epsilon)}$) [DKK 2021]. Importantly, the $L_1$-regression algorithm remains the most efficient known algorithm for the problem.

Given the gap between the sample complexity of the problem and the complexity of known algorithms, it is natural to ask whether the limitations of known efficient algorithms are inherent. There are two general approaches to establish information-computation tradeoffs for statistical problems. One approach focuses on restricted families of algorithms (e.g., Statistical Query algorithms or low-degree polynomial tests). It should be noted that such results do not have any implications...
for the family of all polynomial-time algorithms. Another, arguably more convincing approach, is via efficient reductions from known (average-case) hard problems. This is the approach we adopt in this work.

Returning to Problem 1.1, a line of work \[GGK20, DKZ20, DKPZ21\] has established tight hardness in the Statistical Query (SQ) model. SQ algorithms [Kea98] are a class of algorithms that are only allowed to query expectations of bounded functions of the distribution rather than directly access samples. [DKPZ21] showed that any SQ algorithm for the problem either requires \(2^{\Omega(1)}\) queries or at least one query of very high accuracy (suggesting a sample complexity lower bound of \(n^{\Omega(1/\epsilon^2)}\)). Interestingly, it is known (see, e.g., [DFT+15]) that the \(L_1\)-regression algorithm can be efficiently implemented in the SQ model. However, since the SQ model is restricted, this SQ lower bound has no implications for general efficient algorithms.

Prior to the our work, the only known computational hardness for Problem 1.1 is due to Klivans and Kothari [KK14]. That work gave a reduction from the problem of learning sparse parities with noise to Problem 1.1. Under the plausible assumption that learning \(k\)-sparse parities with noise over \(\{0, 1\}^n\) requires time \(n^{\Omega(k)}\), the reduction of [KK13] implies a computational lower bound of \(n^{\Omega((1/\epsilon)^2)}\) for Problem 1.1. Interestingly, this lower bound cannot be improved in the sense that the corresponding hard instances can be solved in time \(n^{O(\log(1/\epsilon))}\).

Finally, we note that for the qualitatively weaker error guarantee of \(C \cdot \text{OPT} + \epsilon\), for a sufficiently large universal constant \(C > 1\), \(\text{poly}(d/\epsilon)\) time algorithms are known [ABL17, Dan15, DKS18].

In summary, the best known algorithm for Problem 1.1 has sample complexity and running time \(n^{\text{poly}(1/\epsilon)}\), while the best known computational hardness result gives an \(n^{\Omega((1/\epsilon)^2)}\) lower bound. Moreover, a tight lower bound is known for the restricted class of SQ algorithms. This raises the following natural question:

**Can we establish a near-optimal computational hardness result for Problem 1.1?**

In this paper, we answer this question in the affirmative by exhibiting a computational hardness reduction from a classical cryptographic problem, showing that current algorithms are essentially best possible. Specifically, we prove a complexity lower bound of \(n^{\text{poly}(1/\epsilon)}\) (Theorem 1.3), assuming the widely believed sub-exponential hardness of the Learning with Errors (LWE) problem (Definition 2.2).

The task of learning halfspaces is as a special case of the more general setting that the underlying function is of the form \(\sigma((\langle w, x \rangle - t))\), where \(\sigma : \mathbb{R} \to \mathbb{R}\) is a univariate activation. If the activation is better behaved than the sign function, specifically if \(\sigma\) is monotone and Lipschitz (aka the setting of Generalized Linear Models), then the learning problem can be easier computationally. Here we show that our techniques can be extended to prove near-optimal hardness for some of these cases as well. Specifically, we focus on the well-studied problem of ReLU regression.

A ReLU is any function \(f : \mathbb{R}^n \to \mathbb{R}_+\) of the form \(f(x) = \text{ReLU}((w, x) - t)\), where \(w \in \mathbb{R}^n\) is called the weight vector and \(t \in \mathbb{R}\) is called the threshold. The activation ReLU : \(\mathbb{R} \to \mathbb{R}_+\) is defined as \(\text{ReLU}(u) = \max\{0, u\}\). ReLUs are the most commonly used activations in modern deep neural networks. Moreover, finding the best-fitting ReLU with respect to square-loss is a fundamental primitive in the theory of neural networks. A line of work studied this problem from the perspectives of both algorithms and lower bounds, see, e.g., [Sol17, GKKKT17, MR18, GKK19, FCG20, DGK+20, DKTZ22, ATV22]. Similarly to the case of halfspaces, ReLU regression is efficiently solvable in the realizable setting and computationally hard (even for weak error guarantees) in the distribution-independent agnostic setting [MR13, DKMR22a]. Here we study the agnostic setting with Gaussian marginals.

Since ReLU regression is a real-valued task, we will require the analogous terminology. For a real-valued hypothesis \(h : \mathbb{R}^n \to \mathbb{R}\) and a distribution \(D\) supported on \(\mathbb{R}^n \times \{\pm 1\}\), we use \(R_2(h; D)\)
to denote the $L_2^2$-error of $h$ with respect to $D$, i.e., $R_2(h; D) \overset{\text{def}}{=} E_{(x,y) \sim D}[(h(x) - y)^2]$. For a class $\mathcal{C}$ of real-valued functions on $\mathbb{R}^n$, we use $R_2(\mathcal{C}; D)$ to denote the minimum $L_2^2$-error of any $h \in \mathcal{C}$, i.e., $R_2(\mathcal{C}; D) \overset{\text{def}}{=} \min_{h \in \mathcal{C}} R_2(h; D)$.

**Problem 1.2** (ReLU Regression under Gaussian Marginals). Let ReLU be the class of ReLUs on $\mathbb{R}^n$ with weight vectors in the set $\{w \in \mathbb{R}^n : \|w\|_2 \leq 1\}$. Given an additive error parameter $0 < \epsilon < 1$ and i.i.d. samples $(x,y)$ from a distribution $D$ on $\mathbb{R}^n \times \mathbb{R}$, where the marginal $D_x$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}(0,\mathbf{I})$ and the labels $y$ are bounded, the goal of the learning algorithm $A$ is to output a hypothesis $h : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $R_2(h; D) \leq R_2(\text{ReLU}; D) + \epsilon$ with high probability. We will say that the algorithm $A$ agnostically learns ReLUs under Gaussian marginals to additive error $\epsilon$.

**Prior Work on Problem 1.2** While there is no black-box relation with Problem 1.1, the situation for both problems is analogous. [DGK+20] gave an algorithm for Problem 1.2 with sample complexity and runtime $n^{\text{poly}(1/\epsilon)}$. While poly($n/\epsilon$) time algorithms are known with weaker guarantees, [GKK19, DGK+20, DKTZ22], the fastest known algorithm with OPT + $\epsilon$ error is the one of [DGK+20]. In terms of computational hardness, [GKK19] gave a reduction from sparse noisy parity implying a computational lower bound of $n^{\Omega(\log(1/\epsilon))}$ for Problem 1.2. In the restricted SQ model, (near-optimal) SQ lower bounds of $n^{\text{poly}(1/\epsilon)}$ have been shown [GGK20, DKTZ20, DKPZ21].

In summary, the best known algorithm for Problem 1.2 has sample complexity and running time $n^{\text{poly}(1/\epsilon)}$, while the best known computational hardness result gives an $n^{\Omega(\log(1/\epsilon))}$ lower bound. It is thus natural to ask whether computational hardness of $n^{\text{poly}(1/\epsilon)}$ can be established. Similarly to the case of LTFs, we prove such a statement (Theorem 1.3) under the sub-exponential hardness of LWE.

### 1.1 Our Results and Techniques

We start with an informal definition of the LWE problem. In the LWE problem, we are given samples $(x_1, y_1), \ldots, (x_m, y_m)$ and the goal is to distinguish between the following two cases:

- Each $x_i$ is drawn uniformly at random (u.a.r.) from $\mathbb{Z}_q^n$, and there is a hidden secret vector $s \in \mathbb{Z}_q^n$ such that $y_i = \langle x_i, s \rangle + z_i$, where $z_i \in \mathbb{Z}_q$ is discrete Gaussian noise (independent of $x_i$).
- Each $x_i$ and each $y_i$ are independent and are sampled u.a.r. from $\mathbb{Z}_q^n$ and $\mathbb{Z}_q$ respectively.

Formal definitions of LWE (Definition 2.2) together with the precise computational hardness assumption (Assumption 2.3) we rely on are given in Section 2.

For Problem 1.1 we prove:

**Theorem 1.3** (Hardness of Agnostically Learning Gaussian Halfspaces). Assume that LWE cannot be solved in $2^{n^{1-O(1)}}$ time. Then for any constants $c > 0$ and $\alpha < 2$ the following holds: If $\epsilon \leq 1/\log^{1/2+c}(n)$, any algorithm that agnostically learns LTFs on $\mathbb{R}^n$ with Gaussian marginals to additive error $\epsilon$ requires running time at least $\min\{n^{\Omega(1/(\sqrt{\log n})^\alpha)}, 2^{0.99}\}$.

Some comments are in order to interpret this statement. The minimum of the two terms is necessary to handle the case where $\epsilon$ is very small, specifically $\epsilon = \tilde{O}(1/\sqrt{n})$. (Since the problem can always be solved in time $2^{O(n)}$ via brute-force, the first term cannot be a time lower bound for this range of $\epsilon$.) On the other hand, for $\Omega(1/\sqrt{n}) = \epsilon \leq 1/\log^{1/2+c}(n)$, Theorem 1.3 gives a time lower bound of $n^{\Omega(1/(\sqrt{\log n})^\alpha)}$, for any constant $\alpha < 2$. This bound nearly matches the upper bound of $n^{O(1/\epsilon^2)}$ [KMS08], up to the $\sqrt{\log n}$ factor in the exponent. Note that the extraneous
factor of \(\sqrt{\log n}\) is negligible if \(\epsilon\) is sufficiently small. For example, if \(\epsilon \leq 1/\log n\), the implied lower bound is \(n^{\Omega(1/\epsilon^\alpha)}\) for any constant \(\alpha < 1\). For \(\epsilon = O(n^{-c})\), for a small constant \(c > 0\), we get a lower bound of \(n^{\tilde{\Omega}(1/\epsilon^\alpha)}\), for any constant \(\alpha < 2\).

For Problem 1.2, we prove:

**Theorem 1.4 (Hardness of Gaussian ReLU Regression).** Assume that LWE cannot be solved in \(2^{n^{1+\Omega(1)}}\) time. Then for any constants \(c > 0\) and \(\alpha < 1/2\) the following holds: If \(\epsilon \leq 1/\log^{2+c}(n)\), any algorithm for ReLU regression on \(\mathbb{R}^n\) under Gaussian marginals with additive error \(\epsilon\) requires running time at least \(\min\{n^{\Omega(1/(\epsilon \log^{2}(n)^{\alpha})}, 2^{n^{0.99}}\}\).

Intuitively, the above statement says that any algorithm for Problem 1.2 requires time at least \(n^{(1/\epsilon)^\Omega(1)}\), if \(\epsilon\) is sufficiently small (e.g., \(\epsilon = O(1/\log^3 n)\)) and not too small (in which case the latter term dominates the obvious brute-force algorithm). This runtime lower bound qualitatively matches the upper bound of \(n^{\text{Poly}(1/\epsilon)}\) \([\text{DGK}^{+}20]\) and exponentially improves on the best known computational lower bound of \(n^{\Omega(\log(1/\epsilon))}\) \([\text{GKK}^{19}\text{]}\).

### 1.2 Techniques

Our computational hardness reductions build on two main ideas. The first idea is inspired by the approach of \([\text{DKMR}22\text{b}]\). We note that \([\text{DKMR}22\text{b}]\) established a hardness reduction from LWE to distribution-free PAC learning halfspaces with Massart noise. While the Massart noise model is technically easier than the adversarial label noise model, here we are interested in the (much simpler) regime where the marginal distribution is Gaussian. Indeed, the results of \([\text{DKMR}22\text{b}]\) have no implications for the Gaussian setting. Yet one of their ideas is useful in our context.

The key idea of \([\text{DKMR}22\text{b}]\) is that by applying rejection sampling to a continuous variant of LWE supported on \(\mathbb{R}^n\) (this variant was shown to be as hard as the standard LWE problem supported on \(\mathbb{Z}_q^n\) in \([\text{GVV}22]\)) one obtains either (i) a standard Gaussian in the null hypothesis case or (ii) a distribution that is approximately a discrete Gaussian plus a little noise in a hidden direction and a standard Gaussian in the orthogonal directions in the alternative hypothesis case. By taking a mixture of such rejection sampling distributions, \([\text{DKMR}22\text{b}]\) manage to produce a joint distribution on \((x, y)\) over \(\mathbb{R}^n \times \{\pm 1\}\) such that:

(i) in the null hypothesis case, \(y\) is independent of \(x\), and

(ii) in the alternative hypothesis case, \(y\) is given by a Polynomial Threshold Function (PTF) applied to \(x\) with Massart noise.

Given the above, \([\text{DKMR}22\text{b}]\) conclude that any learner for Massart halfspaces LTFs can be used to distinguish between the alternative and null hypothesis cases, and thus solves the LWE problem.

In this paper, we apply a similar technique to the tasks of agnostically learning halfspaces and ReLUs under Gaussian marginals. A key difference in our setting is that we require the distribution of \(x\) be the standard Gaussian — a property inherently not satisfied by the aforementioned construction. Roughly speaking, \([\text{DKMR}22\text{b}]\) showed that it is LWE-hard to distinguish between a standard Gaussian and a distribution that is standard Gaussian in all directions except for a hidden direction in which it is approximately a specified mixture of discrete Gaussians plus a little noise. The learning application in \([\text{DKMR}22\text{b}]\) was obtained via the construction of a PTF with Massart noise such that both the conditional distributions on \(y = 1\) and on \(y = -1\) were such (noisy) mixtures of discrete Gaussians. In our context, we need to construct different pairs of such conditional distributions.

---

1This leverages a construction of such a distribution from \([\text{DK22}]\).
We do this as follows. Let $x$ be sampled from a standard Gaussian and consider the function $f_s(x) = (-1)^{(x,s)}$ for some unknown vector $s$ with relatively large norm. If we consider the distribution of $x$ conditioned on $f_s(x) = 1$, we obtain a distribution that is (i) Gaussian in the directions orthogonal to $s$, and (ii) a Gaussian conditioned on $[(x,s)]$ being in the $s$-direction. One can see that this is a mixture of discrete Gaussians. The same can be argued for the distribution of $s$ conditioned on $f_s(x) = -1$. Thus, using the techniques described above, we can show that given labeled samples $(x,y)$ with $x$ a standard Gaussian, it is LWE-hard to distinguish between the cases that (i) $y$ is independent of $x$, and (ii) $y = f_s(x)$ for some unknown vector $s$.

This result forms the basis for our two learning applications. Specifically, for the problem of agnostically learning Gaussian LTFs, it is not hard to show that there exists an LTF $g$ such that $E_{x \sim \mathcal{N}(0,1)}[f_s(x)g(x)] = \epsilon = \Omega(1/\|s\|_2^2)$. This implies that any algorithm that agnostically learns LTFs to error $\text{OPT} + \epsilon/3$, where $\text{OPT} = R_{0-1}(\text{LTF}; D)$, can be used to distinguish between the case that $y$ is independent of $x$ (in which case $\text{OPT} = 1/2$) and the case described above (i.e., $y = f_s(x) = (-1)^{(x,s)}$), where $\text{OPT} = 1/2 - \epsilon$. This implies that the agnostic learning of Gaussian LTFs is LWE-hard.

For ReLU regression, we show that there exists a ReLU $g$ such that $E_{x \sim \mathcal{N}(0,1)}[f_s(x)g(x)] = \epsilon = \Omega(1/\|s\|_2^2)$. In particular, this correlation means that the $L_2^2$-distance between $f$ and an appropriately scaled version of $g$ is bounded away from 1 in the negative direction. Thus, it is LWE-hard to distinguish between the case where $y = f_s(x)$ (and thus the minimal $L_2^2$-error for ReLUs is at most $1 - \epsilon^2$) and the case where $y$ is independent of $x$ (in which case the minimum $L_2^2$-error of any ReLU is at least 1).

The above sketch glossed over the following important technical point. By applying the aforementioned reduction directly to the standard version of the (continuous) LWE problem \cite{BGST21}, which has secret vector $s$ with $\|s\|_2 = \sqrt{n}$, we can obtain a time lower bound for our agnostic learning problems only if the additive error $\epsilon$ is tiny, namely $\epsilon = \tilde{O}(1/\sqrt{n})$. In order to prove lower bounds for a wider range of $\epsilon$, we will need to instead start from a small norm version of the continuous LWE problem, where the secret vector $s$ roughly satisfies $\|s\|_2 \approx 1/\epsilon$. We accomplish this via a non-trivial modification of a reduction in \cite{GVV22}, which we view as an additional technical contribution of this work. Specifically, \cite{GVV22} gave a reduction of the standard discrete LWE problem to a discrete LWE problem with a sparse secret (namely, secret vector $s \in \{0, \pm 1\}^n$ with $\|s\|_1 = k$). (This itself leverages an idea in \cite{Mic18}.) After that, \cite{GVV22} further reduces the sparse secret discrete LWE problem to a continuous LWE problem whose secret vector has small $\ell_2$-norm.

The limitation here is that their $\ell_2$-norm bound has a factor of $\sqrt{\log m}$, where $m$ is the number of samples. Unfortunately, this quantitative dependence prevents us from obtaining the near optimal lower bound for our learning LTFs tasks. To address this issue, we present a (slightly) improved reduction (see Lemma 3.5), removing the $\sqrt{\log m}$ factor on the secret vector norm. This allows us to apply our reduction technique to the small norm continuous LWE problem, giving nearly tight lower bounds for our learning problems.

2 Preliminaries

**Notation** We use $\langle x, y \rangle$ for the inner product between vectors $x, y \in \mathbb{R}^n$. For $p \geq 1$ and $x \in \mathbb{R}^n$, we use $\|x\|_p \overset{\text{def}}{=} (\sum_{i=1}^{n} |x_i|^p)^{1/p}$ to denote the $\ell_p$-norm of $x$. We use $S^{n-1}$ to denote the unit sphere in $\mathbb{R}^n$, i.e., the set $S^{n-1} \overset{\text{def}}{=} \{ x \in \mathbb{R}^n : \|x\|_2 = 1 \}$. For $q \in \mathbb{N}$, we denote $\mathbb{Z}_q \overset{\text{def}}{=} \{ 0, 1, \cdots, q - 1 \}$ and $\mathbb{R}_q \overset{\text{def}}{=} [0, q)$. We use $\text{mod}_q : \mathbb{R}^n \rightarrow \mathbb{R}^n$ to denote the function that applies the $\text{mod}_q$ operation on each coordinate of the vector $x$. For a set $S \subset \mathbb{R}^n$, we use $U(S)$ to denote the uniform distribution over $S$. We use $x \sim D$ to denote a random variable $x$ with distribution $D$. For a random variable
Null hypothesis (resp. a distribution $D$), we use $P_x$ (resp. $P_D$) to denote the probability density function or probability mass function of the random variable $x$ (resp. distribution $D$). We will require the following notion of partially supported Gaussians.

**Definition 2.1** (Partially Supported Gaussian Distribution). For $\sigma \in \mathbb{R}_+$ and $x \in \mathbb{R}^n$, let $\rho_\sigma(x) \overset{\text{def}}{=} \sigma^{-n} \exp\left(-\pi(||x||^2/\sigma^2)\right)$. For any countable set $S \subseteq \mathbb{R}^n$, we let $\rho_\sigma(S) \overset{\text{def}}{=} \sum_{x \in S} \rho_\sigma(x)$, and let $D_{S,\sigma}$ be the distribution supported on $S$ with pmf $P_{D_{S,\sigma}}(x) = \rho_\sigma(x)/\rho_\sigma(S)$.

For consistency, we will use $D_{\mathbb{R}^n,\sqrt{2\pi}\sigma}$ to denote the $n$-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$.

**Learning with Errors** The Learning with Errors (LWE) problem was introduced in [Reg05]. Here we use a slightly more generic definition for the convenience of later reductions between different variants of LWE problems.

**Definition 2.2** (Generic LWE). Let $m,n \in \mathbb{N}$, $q \in \mathbb{R}_+$, and let $D_{\text{sample}}, D_{\text{secret}}, D_{\text{noise}}$ be distributions on $\mathbb{R}^m, \mathbb{R}^n, \mathbb{R}$ respectively. In the LWE($m,D_{\text{sample}}, D_{\text{secret}}, D_{\text{noise}}, \text{mod}_q$) problem, we are given $m$ independent samples $(x,y)$ and want to distinguish between the following two cases:

(i) **Alternative hypothesis**: A vector $s$ is drawn from $D_{\text{secret}}$ ($s$ is called “the secret vector”). Then each sample $(x,y)$ is generated by taking $x \sim D_{\text{sample}}$, $z \sim D_{\text{noise}}$, and letting $y = \text{mod}_q((x,s) + z)$.

(ii) **Null hypothesis**: The random variables $x$ and $y$ are independent. Moreover, $x$ has the same marginal distribution as in the alternative hypothesis, and $y$ has the marginal distribution as $U(S)$ where $S$ is the support of the marginal distribution of $y$ in the alternative hypothesis.

An algorithm $A$ solves the LWE problem with advantage $\alpha > 0$, if $p_{\text{alternative}} - p_{\text{null}} \geq \alpha$ where $p_{\text{alternative}}$ (resp. $p_{\text{null}}$) is the probability that $A$ outputs “alternative hypothesis” if the input distribution is from the alternative hypothesis (resp. null hypothesis). When a distribution in LWE is uniform over some set $S$, we may abbreviate $U(S)$ as $S$.

Our hardness assumption is the following:

**Assumption 2.3** (Sub-exponential LWE Assumption). Let $c > 0$ be a sufficiently large constant and $q \in \mathbb{N}$. For any constants $\beta \in (0,1)$, $\kappa \in \mathbb{N}$, the problem LWE($2^{O(n^\beta)}, \mathbb{Z}_q, \mathbb{Z}_q^n, D_{\mathbb{Z},\sigma}, \text{mod}_q$) with $q \leq n^\kappa$ and $\sigma = c\sqrt{n}$ cannot be solved in $2^{O(n^\beta)}$ time with $2^{-O(n^\beta)}$ advantage.

This is a widely-believed conjecture, supported by our current understanding of the field. [Reg05] [Pet09] gave a polynomial-time quantum reduction from approximating (the decision version of) the Shortest Vector Problem (GapSVP) to LWE (with similar $n,q,\sigma$ parameters). We note that the fastest known algorithm for GapSVP takes $2^{O(n)}$ time [ALNS20]. Thus, refuting the conjecture would be a major breakthrough. A similar assumption was also used in [GVV22] and [DKMR22] to establish computational hardness of learning Gaussian mixtures and distribution-independent learning of Massart halfspaces.

In addition to the standard LWE problem above, we will also consider a continuous variant of the LWE problem (introduced in [BRST21]) where supports of the distributions are continuous. In particular, the first part of our proof is the following proposition which slightly modifies the proof in [GVV22] and gives the reduction from the standard LWE to the continuous LWE. The proof is deferred to Appendix 13.

\footnote{We will take the sets $S$ to be shifts of lattices, guaranteeing that $\rho_\sigma(S)$ is finite and the distribution is well-defined.}
Proposition 2.4 (Hardness of continuous LWE (cLWE) with Small-Norm Secret). Under Assumption 2.3, for any \( n \in \mathbb{N} \), any constants \( \beta \in (0, 1) \), \( \gamma \in \mathbb{R}_+ \) and any \( \log \gamma n \leq k \leq cn \) where \( c > 0 \) is a sufficiently small universal constant, the problem \( \text{LWE}(n^{O(k^\beta)}, D_{\mathbb{R}^n,1}^N, S^{n-1}, D_{\mathbb{R},\sigma}, \text{mod}_T) \) with \( \sigma \geq k^{-\kappa} \) and \( T = 1/(c' \sqrt{k \log n}) \) where \( c' > 0 \) is a sufficiently large universal constant cannot be solved in \( n^{O(k^\beta)} \) time with \( n^{-O(k^\beta)} \) advantage.

3 Hardness of Agnostically Learning Gaussian LTFs

In this section, we continue from Proposition 2.4 (the proof of which is deferred to Appendix B) which is the first step of our reduction, and give the second and main part of the reduction. We thereby establishing the desired cryptographic hardness of agnostically learning LTFs under the Gaussian distribution.

The high-level idea is the following. Given samples \( (x, y) \) from a distribution \( D \) on \( \mathbb{R}^n \times \mathbb{R}_T \), which is an instance of the cLWE problem \( \text{LWE}(n^{O(k^\beta)}, D_{\mathbb{R}^n,1}^N, S^{n-1}, D_{\mathbb{R},\sigma}, \text{mod}_T) \) (note that \( T \) is the “period” of the periodic signal on the hidden direction) from Proposition 2.4, we efficiently generate samples \( (x, y') \) (we leave \( x \) unchanged) from a distribution \( D' \) on \( \mathbb{R}^n \times \{\pm 1\} \) such that:

(i) If \( D \) is from the alternative hypothesis case, then there exists an LTF \( h : \mathbb{R}^n \rightarrow \{\pm 1\} \) such that \( R_{0-1}(h; D') \leq 1/2 - \Omega(T) \).

(ii) If \( D \) is from the null hypothesis case, then for \( (x, y') \sim D' \), we have that \( y' = +1 \) with probability 1/2 and \( y' = -1 \) with probability 1/2 independent of \( x \); thus, no hypothesis can achieve error non-trivially better than 1/2.

Given the above properties, if an algorithm can agnostically learn LTFs with Gaussian marginals to error \( R_{0-1}(LTF; D') + o(T) \), then it can distinguish the two cases above and solve the LWE problem.

In the body of this section, we describe our reduction and formalize the above. The main theorem of this section, stated and proved below, establishes hardness for a natural decision version of agnostically learning LTFs.

Theorem 3.1 (Cryptographic Hardness of Agnostically Learning Gaussian LTFs). Under Assumption 2.3, for any \( n \in \mathbb{N} \), for any constants \( \beta \in (0, 1) \), \( \gamma \in \mathbb{R}_+ \) and any \( \log \gamma n \leq k \leq cn \) where \( c \) is a sufficiently small constant, there is no algorithm that runs in time \( n^{O(k^\beta)} \) and distinguishes between the following two cases of a joint distribution \( D \) of \( (x, y) \) supported on \( \mathbb{R}^n \times \{\pm 1\} \) with marginal \( D_x = D_{\mathbb{R}^n,1}^N \), with \( n^{-O(k^\beta)} \) advantage:

(i) **Alternative Hypothesis:** There exists an LTF with 0-1 error non-trivially smaller than 1/2, namely \( R_{0-1}(LTF; D) \leq 1/2 - \Omega(1/\sqrt{k \log n}) \).

(ii) **Null Hypothesis:** A sample \( (x, y) \sim D \) satisfies the following: \( y = +1 \) with probability 1/2 and \( y = -1 \) with probability 1/2 independent of \( x \).

Proof. We give an efficient method taking as input samples from a distribution \( D' \) — that is either from the alternative hypothesis or the null hypothesis of \( \text{LWE}(n^{O(k^\beta)}, D_{\mathbb{R}^n,1}^N, S^{n-1}, D_{\mathbb{R},\sigma}, \text{mod}_T) \) from Proposition 2.4 — and generates samples from another distribution \( D \) with the following properties: If \( D' \) is from the alternative (resp. null) hypothesis of the LWE problem, then the resulting distribution \( D \) will satisfy the alternative (resp. null) hypothesis requirement of the theorem for the agnostic LTF learning decision problem.
The reduction process is the following: For a sample \((x, y')\) from a distribution \(D'\), which is an instance of the problem \(\text{LWE}(n^O(k^8), D_{\mathbb{R}, 1}^N, S^{n-1}, D_{\mathbb{R}, \sigma}^N, \text{mod}_T)\) from Proposition 2.4, we simply output \((x, y) \sim D\), where \(y = +1\) if \(y' \leq T/2\) and \(y = -1\) otherwise. We argue that \(D\) satisfies the desired requirements stated above. We first note that the marginal \(D_x\) of \(D\) satisfies \(D_x = D_{\mathbb{R}, 1}^N\), therefore it suffices to verify that \(R_{0-1}(\text{LTF}; D) = 1/2 - \Omega\left(1/\sqrt{k \log n}\right)\) and \(y = +1\) with probability \(1/2\) independent of \(x\) for each case respectively.

For the alternative hypothesis case, let \(D'\) be from the alternative hypothesis case of the LWE. Let \(s\) be the secret vector in the LWE problem. We consider the following two LTFs: \(h_1(x) = \text{sign}(\langle x, s \rangle - T/6)\) and \(h_2(x) = \text{sign}(\langle x, s \rangle + T/3)\). If we can show that \(R_{0-1}(h_1; D) + R_{0-1}(h_2; D) \leq 1 - \Omega(T)\), then either \(h = h_1\) or \(h = h_2\) satisfies \(R_{0-1}(h; D) \leq 1/2 - \Omega(T)\), which implies that \(R_{0-1}(\text{LTF}; D) \leq R_{0-1}(h; D) \leq 1/2 - \Omega(1/\sqrt{k \log n})\) by the definition of \(T\).

To show that \(R_{0-1}(h_1; D) + R_{0-1}(h_2; D) \leq 1 - \Omega(T)\), we examine the subset of the domain where \(h_1\) and \(h_2\) agree, namely the region
\[
B \overset{\text{def}}{=} \{t \in \mathbb{R}^n \mid h_1(t) = h_2(t)\} = \{t \in \mathbb{R}^n \mid \langle s, t \rangle \in [T/6, T/3]\}.
\]
Since for any \(t \in B\), it is always the case that \(h_1(t) = h_2(t) = +1\), we can write
\[
R_{0-1}(h_1; D) + R_{0-1}(h_2; D) = \Pr_{(x, y) \sim D}[y \neq h_1(x)] + \Pr_{(x, y) \sim D}[y \neq h_2(x)]
= \Pr_{(x, y) \sim D}[x \notin B \land y \neq h_1(x)]
+ \Pr_{(x, y) \sim D}[x \notin B \land y \neq h_2(x)]
+ 2\Pr_{(x, y) \sim D}[x \in B \land y = -1].
\]
Since for any \(x \notin B\) we have that \(h_1(x) \neq h_2(x)\), the first two terms sum to \(\Pr_{(x, y) \sim D}[x \notin B]\). Therefore, we have that
\[
R_{0-1}(h_1; D)) + R_{0-1}(h_2; D)
= \Pr_{(x, y) \sim D}[x \notin B] + 2\Pr_{(x, y) \sim D}[x \in B \land y = -1]
= 1 + \Pr_{(x, y) \sim D}[x \in B \land y = -1]
- \Pr_{(x, y) \sim D}[x \in B \land y = +1]
= 1 - \Pr[x \in B]|(1 - 2\Pr_{(x, y) \sim D}[y = -1 \mid x \in B])).
\]
From the definition of \(B\) and \(x \sim D_{\mathbb{R}, 1}^N\), we have \(\Pr[x \in B] = \Omega(T)\). Thus, we obtain
\[
R_{0-1}(h_1; D) + R_{0-1}(h_2; D) = 1 - \Omega(T) \left(1 - 2\Pr_{(x, y) \sim D}[y = -1 \mid x \in B]\right). \tag{1}
\]
If we can show that \(\Pr[y = -1 \mid x \in B] \leq 1/3\), then we are done since this implies that \(R_{0-1}(h_1; D) + R_{0-1}(h_2; D) \leq 1 - \Omega(T)\).

We note that from the definition of the Alternative case distribution of the LWE problem, we have
\[
y' = \text{mod}_T(\langle s, x \rangle + z),
\]
and that \(y = -1\) only if \(y' > T/2\), which in turn happens only if
\[
\langle s, x \rangle + z > T/2 \text{ or } \langle s, x \rangle + z < 0.
\]

8
For $x \in B$, we have that $(s, x) \in [T/6, T/3]$, therefore $y = -1$ only if $|z| \geq T/6$. Notice that $z \sim D_{k, \sigma}^N$ and Proposition 2.4 states that the LWE problem is hard for any fixed constant $\kappa \in \mathbb{N}$ and $\sigma \geq k^{-\kappa}$. Given the constant $\gamma \in \mathbb{R}^+$ in this theorem, we will take $\kappa = \lceil 1/(2\gamma) \rceil + 1/2 + 1$ which is a fixed constant. Then, by Proposition 2.4, the LWE problem is hard for $\sigma = k^{-\kappa} = 1/(k^{3/2} \sqrt{\log n}) = o(T)$. Therefore, we have that

$$\Pr_{(x, y) \sim D}[y = -1 \mid x \in B] \leq \Pr_{z \sim D_{k, \sigma}^N}[|z| \geq T/6] = o(1).$$

Thus, plugging the above back to (1), we can conclude that

$$R_{0-1}(h_1; D) + R_{0-1}(h_2; D) - 1 \Omega(T) \left(1 - 2\Pr_{(x, y) \sim D}[y = -1 \mid x \in B]\right) \leq 1 - \Omega(T).$$

Then, as argued above, if both $h = h_1$ and $h = h_2$ do not satisfy $R_{0-1}(h; D) \leq 1/2 - \Omega(T)$, then $R_{0-1}(h_1; D) + R_{0-1}(h_2; D) > 1 - \Omega(T)$, a contradiction. Thus, either $h = h_1$ or $h = h_2$ satisfies $R_{0-1}(h; D) \leq 1/2 - \Omega(T) \leq 1/2 - \Omega(1/\sqrt{k \log n})$. This completes the proof for the alternative hypothesis case.

For the null hypothesis case, it is immediate that $y = +1$ with probability $1/2$ independent of $x$, since $y' \sim U([0, T])$ independent of $x$ in the null hypothesis case of the LWE problem. This completes the proof of correctness.

It remains to verify the time lower bound and the distinguishing advantage for agnostically learning LTFs. From Proposition 2.4, we know that under Assumption 2.3, LWE\((n^{O(k^3)}; D_{k, \sigma}^N, S_{k, \sigma}^n, D_{\mathbb{R}, k, \sigma}^N, \mod T)\) with any $\sigma \geq k^{-\kappa}$ (where $\kappa \in \mathbb{N}$ is a constant) and $T = 1/(c \sqrt{\log n})$, where $c > 0$ is a sufficiently large universal constant, the problem cannot be solved in $n^{O(k^3)}$ time with $n^{-O(k^3)}$ advantage. Therefore, under the same assumption, there is no algorithm that solves the decision version of the agnostic learning LTFs problem (defined in the theorem statement) in time $n^{O(k^3)}$ with $n^{-O(k^3)}$ advantage.

The following corollary immediately follows from Theorem 3.1.

**Corollary 3.2.** Under Assumption 2.3, for any constants $\alpha \in (0, 2)$, $\gamma > 1/2$ and any $c/(\sqrt{n \log n}) \leq \epsilon \leq 1/\log^\gamma n$ where $c$ is a sufficiently large constant, there is no algorithm that agnostically learns LTFs on $\mathbb{R}^n$ with Gaussian marginals to additive error $\epsilon$ and runs in time $n^{O(1/(\epsilon \sqrt{\log n}))}$.

**Proof.** We chose the parameter $k$ in Theorem 3.1 to be the value that $\epsilon = c/\sqrt{k \log n}$, where $c$ is a sufficiently small constant. Then any algorithm that agnostically learns LTFs to additive error $\epsilon$ can solve the testing problem of Theorem 3.1 with probability $2/3$. Therefore, no such algorithm should run in time $n^{O(k^3)}$ for any $\beta \in (0, 1)$. Since $\epsilon = c/\sqrt{k \log n}$, and if we chose $\beta = \alpha/2$, then the time lower bound can be rewritten as $n^{O(k^3)} = n^{O(1/(\epsilon \sqrt{\log n}))} = n^{O(1/(\alpha \sqrt{\log n}))}$. This completes the proof.

## 4 Hardness of ReLU Regression with Gaussian Marginals

In this section, we establish near-optimal computational hardness for ReLU regression under Gaussian marginals. It is worth pointing out that this hardness result would also apply to any $L$-Lipschitz activation function $f : \mathbb{R} \to \mathbb{R}$, for $L = O(1)$, such that there exists $t \in \mathbb{R}$ so that $f(x)$ is a constant for any $x \leq t$. Roughly, our result says that any algorithm that solves this problem to error $\text{OPT} + \epsilon$ with Gaussian marginals requires $n^{\text{poly}(1/\epsilon \sqrt{\log n})}$ time.

The idea is to show that the same hard instance as in Section 3 can be distinguished by a ReLU regression algorithm. The main theorem of this section, stated and proved below, establishes hardness for a natural decision version of agnostically learning ReLU.
Theorem 4.1. Under Assumption 2.3, for any constants $\beta \in (0, 1)$, $\gamma \in \mathbb{R}_+$ and any $\log^2 n \leq k \leq cn$, where $c$ is a sufficiently small constant, there is no algorithm that runs in time $n^{O(k^3)}$ and distinguishes between the following two cases of joint distribution $D$ on $(x, y)$ supported on $\mathbb{R}^n \times \{\pm 1\}$ with marginal $D_x = D_{\mathbb{R}^n, 1}$, with $n^{-O(k^3)}$ advantage:

(i) **Alternative Hypothesis:** There exists a ReLU with $L_2^2$-error non-trivially smaller than 1, namely $R_2(\text{ReLU}; D) \leq 1 - \Omega(1/(k \log n))$.

(ii) **Null Hypothesis:** A sample $(x, y) \sim D$ satisfies the following: $y = +1$ with probability $1/2$ and $y = -1$ with probability $1/2$ independent of $x$.

Proof. We start with the following intermediate lemma. The lemma roughly says that if there exists a ReLU nontrivially correlated with a distribution, then there must be another ReLU with nontrivial $L_2^2$-error.

Lemma 4.2. Let $\epsilon \in (0, 1)$ and $D$ be a joint distribution of $(u, y)$ supported on $\mathbb{R} \times \{\pm 1\}$ such that the marginal $D_u = D_{\mathbb{R}^n, 1}$ and $\mathbb{E}_{u \sim D}[y] = 0$. Suppose there is a ReLU of the form $f(u) = \text{ReLU}(u - t)$ such that $t \geq 0$ and $|\mathbb{E}_{u \sim D}[yf(u)]| \geq \epsilon$. Then there exists $k \in (-1, 1)$ such that the ReLU $g(u) = \text{ReLU}(ku - kt)$ satisfies $\mathbb{E}_{u \sim D}[(y - g(u))^2] \leq 1 - \epsilon^2$.

Proof. We first note that $g(u) = kf(u)$, thus

$$
\mathbb{E}_{(u, y) \sim D}[(y - g(u))^2] = \mathbb{E}_{(u, y) \sim D}[y^2] + \mathbb{E}_{(u, y) \sim D}[g(u)^2] - 2\mathbb{E}_{(u, y) \sim D}[yg(u)] = \mathbb{E}_{(u, y) \sim D}[y^2] + k^2\mathbb{E}_{(u, y) \sim D}[f(u)^2] - 2k\mathbb{E}_{(u, y) \sim D}[yf(u)].
$$

Since $y$ is supported on $\{\pm 1\}$, we have that the first term satisfies $\mathbb{E}_{(u, y) \sim D}[y^2] = 1$.

To bound the second term, we show that $f(u)^2 \leq u^2$ for any $u$. Notice that for $u \geq t$, since $t \geq 0$ by assumption, we have that $f(u)^2 = (u - t)^2 \leq u^2$. For $u < t$, we have that $f(u)^2 = 0 \leq u^2$. Therefore, combining with the fact that $u \sim D_{\mathbb{R}, 1}$, we can conclude that

$$
\mathbb{E}_{(u, y) \sim D}[g(u)^2] = k^2\mathbb{E}_{(u, y) \sim D}[f(u)^2] \leq k^2\mathbb{E}_{(u, y) \sim D}[u^2] = k^2.
$$

In summary, we get that

$$
\mathbb{E}_{(u, y) \sim D}[(y - g(u))^2] \leq 1 + k^2 - 2k\mathbb{E}_{(u, y) \sim D}[yf(u)].
$$

We now choose the value of $k$. If $\mathbb{E}_{(u, y) \sim D}[yf(u)] > 0$, then we take $k = \epsilon$; otherwise, we take $k = -\epsilon$, in which case we always have $k \in (-1, 1)$ (since $\epsilon \in (0, 1)$) and

$$
\mathbb{E}_{(u, y) \sim D}[(y - g(u))^2] \leq 1 + \epsilon^2 - 2\epsilon|\mathbb{E}_{(u, y) \sim D}[yf(u)]| \leq 1 - \epsilon^2.
$$

We now give a reduction similar to the proof of Theorem 3.1 using Proposition 2.4. We know that under Assumption 2.3 the following holds: the problem LWE($n^{O(k^3)}$, $D_{\mathbb{R}^n, 1}$, $\mathbb{S}^{n-1}$, $D_{\mathbb{R}^n, \sigma}$, mod-$T$) with any $\sigma \geq k^{-\kappa}$ ($\kappa \in \mathbb{N}$ is a constant) and $T = 1/(\epsilon' \sqrt{k \log n})$, where $\epsilon' > 0$ is a sufficiently large universal constant, cannot be solved in $n^{O(k^3)}$ time with $n^{-O(k^3)}$ advantage. We will give an efficient reduction of the LWE problem to the problem here.

For a sample $(x, y')$ from a distribution $D'$ which is an instance of the problem LWE($n^{O(k^3)}$, $D_{\mathbb{R}^n, 1}$, $\mathbb{S}^{n-1}$, $D_{\mathbb{R}^n, \sigma}$, mod-$T$), we will simply output $(x, y)$ such that: (i) $y = +1$ if $y' \leq \check{\gamma}$.
$T/2$, and (ii) $y = -1$ otherwise as samples from another distribution $D$. We argue that $D$ will satisfy the following property: if $D'$ is from the alternative (resp. null) hypothesis of the LWE problem, then the resulting distribution $D$ will satisfy the alternative (resp. null) hypothesis requirement of ReLU regression decision problem of Theorem 4.1.

Since the marginal $D_x$ of $D$ satisfies $D_x = D_{\mathbb{R}_+}^N$, it is enough to show that in the alternative hypothesis case, we have $R_2(ReLU; D) = 1 - \Omega(1/(k\log n)^2)$, and in the null hypothesis case, we have $y = +1$ with probability $1/2$ independent of $x$.

For the alternative hypothesis case, we first introduce the following lemma.

**Lemma 4.3.** For any $s \in S^{n-1}$, $\sigma, T \in \mathbb{R}_+$, let $D$ be the joint distribution of $(x, y)$ supported on $\mathbb{R}^n \times \{\pm 1\}$ such that each sample $(x, y)$ is generated in the following way. We take $x \sim D_{\mathbb{R}^n, 1}^N$, $z \sim D_{\mathbb{R}, \sigma}^N$, and letting $y = +1$ if $\text{mod}_T(\langle x, s \rangle + z) \leq T/2$ and $y = -1$ otherwise. Given $\sigma = o(T)$, then there is a ReLU of the form $h(x) = \text{ReLU}(\langle s, x \rangle - t)$ such that $t \geq 0$ and

$$|\mathbb{E}_{(x,y)\sim D}[yh(x)]| = \Omega(T^2).$$

**Proof.** We let $h_t(x) \overset{\text{def}}{=} \text{ReLU}(\langle s, x \rangle - t)$ and $r(t) = \mathbb{E}_{(x,y)\sim D}[yh_t(x)]$. Then we just need to show that there is a $t > 0$ such that $|r(t)| = \Omega(T)$. We observe that the derivative of $r(t)$ is

$$r'(t) = \frac{d\mathbb{E}_{(x,y)\sim D}[yh_t(x)]}{dt} = \frac{d\mathbb{E}_{(x,y)\sim D}[y(\langle s, x \rangle - t)1(\langle s, x \rangle > t)]}{dt} = -\mathbb{P}_{(x,y)\sim D}[y = +1 \land \langle s, x \rangle > t] + \mathbb{P}_{(x,y)\sim D}[y = -1 \land \langle s, x \rangle > t],$$

and the second derivative of $r(t)$ is

$$r''(t) = \frac{d(-\mathbb{P}_{(x,y)\sim D}[y = +1 \land \langle s, x \rangle > t] + \mathbb{P}_{(x,y)\sim D}[y = -1 \land \langle s, x \rangle > t])}{dt} = \mathbb{P}_{(s,x)}(t)(\mathbb{P}_{(x,y)\sim D}[y = -1 \mid \langle s, x \rangle = t] - \mathbb{P}_{(x,y)\sim D}[y = 1 \mid \langle s, x \rangle = t]) = \mathbb{P}_{(s,x)}(t)(2\mathbb{P}_{(x,y)\sim D}[y = -1 \mid \langle s, x \rangle = t] - 1).$$

Consider the interval $t \in [T/6, T/3]$. Note that $y = -1$ only if $\langle s, x \rangle + z = t + z \not\in [0, T/2]$. Thus, $y = -1$ only if $|z| \geq T/6$. Notice $z \sim D_{\mathbb{R}, \sigma}^N$ and $\sigma = o(T)$. Thus, for $t \in [T/6, T/3]$, we have that

$$r''(t) = \mathbb{P}_{(s,x)}(t)(2\mathbb{P}_{(x,y)\sim D}[y = -1 \mid \langle s, x \rangle = t] - 1) \leq \mathbb{P}_{(s,x)}(t)\left(2\mathbb{P}_{z\sim D_{\mathbb{R}, \sigma}^N}[|z| \geq T/6] - 1\right) = -\Omega(1),$$

where the last equality follows from $\sigma = o(T)$ and $\mathbb{P}_{(s,x)}(t) = \Omega(1)$ since $\langle s, x \rangle \sim D_{\mathbb{R}, 1}^N$ and $t \in [T/6, T/3]$ for $T < 1$.

We then prove that it holds either $r(T/3) - r(T/4) = \Omega(T^2)$ or $r(T/6) - r(T/4) = \Omega(T^2)$. First note that either $r'(T/4) \leq 0$ or $r'(T/4) > 0$. If $r'(T/4) \leq 0$, then

$$r(T/3) - r(T/4) = r'(T/4)(T/12) + \int_{T/4}^{T/3} r''(t)(T/3 - t) dt \leq \int_{T/4}^{T/3} r''(t)(T/3 - t) dt = -\Omega(T^2).$$

11
If $r'(T/4) > 0$, then

$$r(T/6) - r(T/4) = r'(T/4)(-T/12) + \int_{T/4}^{T/6} r''(t)(T/6 - t)dt$$

$$\leq \int_{T/4}^{T/6} r''(t)(T/6 - t)dt = -\Omega(T^2).$$

Since either $r(T/4) - r(T/3) = \Omega(T^2)$ or $r(T/4) - r(T/6) = \Omega(T^2)$, then one of $|r(T/6)|, |r(T/4)|, |r(T/3)|$ must be $\Omega(T^2)$. This completes the proof. \(\square\)

We will apply Lemma 4.3 on the joint distribution of $(x, y)$ here. Recall that Proposition 2.4 states that the LWE problem is hard for any fixed constant $\kappa \in \mathbb{N}$ and $\sigma \geq k^{-\kappa}$. Given the constant $\gamma \in \mathbb{R}^+$ in this theorem, we will take $\kappa = \lceil 1/(2\gamma) + 1/2 + 1 \rceil$ which is a fixed constant. Then from Proposition 2.4, the LWE problem is hard for $\sigma = k^{-\kappa} = 1/(k^{3/2}\sqrt{\log n}) = o(T)$. Therefore, by Lemma 4.3, there is a ReLU of the form $h(x) = f((s, x)) = \text{ReLU}((s, x) - t)$ such that $t \geq 0$ and $|E_{(x,y)\sim D}[yh(x)]| = |E_{(x,y)\sim D}[yf((x,s))]| = \Omega(T^2) = \Omega(1/(k \log n))$. If we apply Lemma 4.2 to the joint distribution of $((x,s), y)$ and the ReLU function $f$, we get that there must be a ReLU of the form $h'(x) = kf((x, s)) = \text{ReLU}((ks, x) - kt)$ such that $k < 1$ and

$$E_{(x,y)\sim D}[(y - h'(x))^2] \leq 1 - \Omega(1/(k \log n)^2).$$

Since $k < 1$, we have that $\|ks\|_2 \leq \|s\|_2 = 1$, thus $h' \in \text{ReLU}$. This implies that

$$R_2(\text{ReLU}; D) \leq 1 - \Omega(1/(k \log n)^2).$$

For the null hypothesis case, it is immediate that $y = +1$ with probability $1/2$ and $y = -1$ with probability $1/2$ independent of $x$, since $y' \sim U([0, T])$ independent of $x$ in the null hypothesis case of the LWE problem. This completes the proof. \(\square\)

The following corollary can be obtained directly from Theorem 4.1

**Corollary 4.4.** Under Assumption 2.3 for any constants $\alpha \in (0, 1/2)$, $\gamma > 2$ and any $c/(n \log n)^2 \leq \epsilon \leq 1/\log^7 n$ where $c$ is a sufficiently large constant, there is no algorithm for ReLU regression on $\mathbb{R}^n$ under Gaussian marginals to error $R_2(\text{ReLU}; D) + \epsilon$ and runs in time $n^{O(1/(\epsilon \log^2 n)^\alpha)}$.

**Proof.** We chose the parameter $k$ in Theorem 4.1 to be the value so that $\epsilon = c/(k \log n)^2$, where $c$ is a sufficiently small constant. Then any algorithm that agnostically learns a ReLU to additive error $\epsilon$ can solve the testing problem of Theorem 4.1 with probability 2/3. Therefore, no such algorithm should run in time $n^{O(k^\beta)}$ for any $\beta \in (0, 1)$. Since $\epsilon = c/(k \log n)^2$, and if we chose $\beta = 2\alpha$, then the time lower bound can be rewritten as $n^{O(k^\beta)} = n^{O(1/\epsilon \log^2 n)^{\beta/2})} = n^{O(1/\epsilon \log^2 n)^{\alpha}}$. This completes the proof. \(\square\)

**References**


APPENDIX

A Additional Technical Background

For \( n, k \in \mathbb{N} \) with \( k \leq n \), we use \( S_{n,k} \) to denote the \( k \)-sparse set \( S_{n,k} \overset{\text{def}}{=} \{ x \in \{0, \pm 1\}^n : \|x\|_1 = k \} \). We use \( \text{negl}(\lambda) \) to denote \( \lambda^{-\omega(1)} \).

The definition of the discrete Gaussian distribution will also be useful here. Essentially, the discrete Gaussian is a univariate discrete distribution supported on equally spaced points on \( \mathbb{R} \) such that the probability mass on any point in its support is proportional to the probability density of a Gaussian on that point. Following Definition 2.1, the discrete Gaussian distribution can be written as the following.

Definition A.1 (Discrete Gaussian). For \( T \in \mathbb{R}_+, y \in \mathbb{R} \) and \( \sigma \in \mathbb{R}_+ \), we define the “\( T \)-spaced, \( y \)-offset discrete Gaussian distribution with \( \sigma \) scale” to be the distribution of \( D_{Tz+y,\sigma}^N \).

Throughout our proofs, we will need to manipulate Gaussian distributions that are taken modulo 1 and those with noise added to them. Due to this, it will be convenient to introduce the following definitions.

Definition A.2 (Expanded Gaussian Distribution from \( \mathbb{R}_n^1 \)). For \( \sigma \in \mathbb{R}_+ \), let \( D_{R_1^\sigma,\sigma}^{\text{expand}} \) denote the distribution of \( x' \) drawn as follows: first sample \( x \sim U(\mathbb{R}_n^1) \), and then sample \( x' \sim D_{Z^n+x,\sigma}^N \).

Definition A.3 (Collapsed Gaussian Distribution on \( \mathbb{R}_n^1 \)). For \( \sigma \in \mathbb{R}_+ \), we will use \( D_{R_1^\sigma,\sigma}^{\text{collapse}} \) to denote the distribution of \( \text{mod}_1(x) \) on \( \mathbb{R}_n^1 \), where \( x \sim D_{\mathbb{Z}_n^1,\sigma}^N \).

B Hardness of cLWE with Small-Norm Secret

Here we give the proof of Proposition 2.1 which is the first step of our hardness reduction. Specifically, we reduce the standard discrete LWE problem in Assumption 2.3 — where the support of \( D_{\text{sample}} \) is the discrete set \( \mathbb{Z}_q^n \) — into a continuous LWE (cLWE) problem — where the support of \( D_{\text{sample}} \) is \( \mathbb{R}_n^1 \). This kind of cLWE problem was first introduced in [BRST21], where the paper gives a quantum reduction from approximating (the decision version of) the Shortest Vector Problem (GapSVP) to cLWE. Subsequently, [GVV22] gave a classical reduction from the classic LWE problem to cLWE problem, indicating that cLWE problem is at least as hard as the LWE problem.

Notably, we will not directly use the cLWE hardness statement here. Instead, we reduce the standard discrete LWE to cLWE. The advantage of such a reduction is that we will be able to start from a \emph{sparse} \( \text{discrete \ LWE} \) instance whose secret vector \( s \) is sampled uniformly from \( S_{n,k} \); after the reduction, we get a cLWE instance whose dimension is \( n \) and the \( \ell_2 \)-norm of the secret is roughly \( \sqrt{k} \approx \log^{0.01} n \), compared with \( \sqrt{n} \ell_2 \)-norm secret vector in [BRST21]).

To achieve this, we slightly modify an idea from [GVV22] to get rid of the \( \log m \) (where \( m \) is the number of samples) blowup in the \( \ell_2 \)-norm of the secret vector.

To prove the proposition, we start with the following lemma which reduces the standard LWE to an LWE with a \( k \)-sparse secret vector (i.e., a secret vector \( s \in S_{n,k} \)).

Lemma B.1 (Corollary 4 in [GVV22]). For any \( n, m, q, l, \lambda, k \in \mathbb{N} \), \( \sigma \in \mathbb{R}_+ \), suppose that \( \log(q)/2^l = \text{negl}(\lambda), \sigma \geq 4\sqrt{\omega(\log \lambda)} + \ln n + \ln m \) and \( k \log_{2}(n/k) \geq (l+1)\log_{2}(q)+\omega(\log \lambda) \). Then, if the testing problem \( \text{LWE}(n, \mathbb{Z}_q^l, \mathbb{Z}_q^l, D_{\mathbb{Z}_n^1,\sigma}^N, \text{mod}_q) \) has no \( T + \text{poly}(n, m, q, \lambda) \) time distinguisher with advantage \( \epsilon \), then the problem \( \text{LWE}(m, \mathbb{Z}_q^l, S_{n,k}, D_{\mathbb{Z}_n^1,\sigma}^N, \text{mod}_q) \) has no \( T \)-time distinguisher with advantage \( 2\epsilon m + \text{negl}(\lambda) \), where \( \sigma' = 2\sigma \sqrt{k+1} \).
The above lemma reduces LWE\((n, \mathbb{Z}_q^n, \mathbb{Z}_q^l, D_{\mathbb{Z}_\sigma}^N)\) to LWE\((m, \mathbb{Z}_q^n, S_{n,k}, D_{\mathbb{Z}_\sigma'}^N, \text{mod}_q)\). The \(\lambda\) here acts as a security parameter. Notice that the original problem LWE\((n, \mathbb{Z}_q^n, \mathbb{Z}_q^l, D_{\mathbb{Z}_\sigma}^N, \text{mod}_q)\) has \(2^{l \log q}\) possible choices of secret vector, while the new problem LWE\((m, \mathbb{Z}_q^n, S_{n,k}, D_{\mathbb{Z}_\sigma'}^N, \text{mod}_q)\) has roughly at least \(2^{k \log q (n/k)}\) possible choices of secret vector. This intuitively explains why there is the requirement of \(k \log_2 (n/k) \geq (l + 1) \log_2 (q) + \omega (\log \lambda)\) in the lemma in terms of entropy of the secret vector.

We then use a bit of extra Gaussian noise to massage the noise distribution from a discrete Gaussian \(D_{\mathbb{Z}_\sigma}^N\) to a continuous Gaussian \(D_{\mathbb{R},\sigma'}^N\), where \(\sigma'\) is going to be slightly larger than \(\sigma\). This leads to the following lemma:

**Lemma B.2** (Lemma 15 in [GVV22]). Let \(n, m, q, \lambda \in \mathbb{N}, \epsilon \in (0, 1]\) and suppose \(\sigma > \sqrt{4 \log m + \omega (\log \lambda)}\). For any \(S \subseteq \mathbb{R}^n\), suppose there is no \(T + \text{poly}(m, n, \log(q), \log(\sigma))\)-time distinguisher for the problem LWE\((m, \mathbb{Z}_q^n, S, D_{\mathbb{Z}_\sigma}^N, \text{mod}_q)\) with advantage \(\epsilon\). Then there is no \(T\)-time distinguisher for the problem LWE\((m, \mathbb{Z}_q^n, S, D_{\mathbb{R},\sigma'}^N, \text{mod}_q)\) with advantage \(\epsilon + \text{negl}(\lambda)\), where we set

\[
\sigma' = \sqrt{\sigma^2 + 4 \ln(m) + \omega (\log \lambda)} = O(\sigma).
\]

We first note that the two requirements of parameters in Lemma 15.2 \(\sigma > \sqrt{4 \log m + \omega (\log \lambda)}\) and \(\sigma' = \sqrt{\sigma^2 + 4 \ln(m) + \omega (\log \lambda)}\) imply that \(\sigma' = \sqrt{\sigma^2 + 4 \ln(m) + \omega (\log \lambda)} = O(\sigma)\). This says that we are only blowing up the noise scale by at most a universal constant multiplicative factor. After this lemma, we again use a bit of extra Gaussian noise to massage the sample distribution \(D_{\text{sample}}\) from \(U(\mathbb{Z}_q^n)\) to \(U(\mathbb{R}_q^n)\). We thus obtain the following:

**Lemma B.3** (Lemma 16 in [GVV22]). Let \(n, m, q, \lambda \in \mathbb{N}, \sigma, r \in \mathbb{R}_+\) and \(\epsilon \in (0, 1]\). Let \(S \subseteq \mathbb{R}^n\) where all elements in the support have fixed \(\ell_2\)-norm \(r\), and suppose that \(\sigma \geq 3r \sqrt{\ln n + \ln m + \omega (\log \lambda)}\). Suppose there is no \(T + \text{poly}(m, n, \log(q), \log(\sigma))\)-time distinguisher for LWE\((m, \mathbb{Z}_q^n, S, D_{\mathbb{R},\sigma'}^N, \text{mod}_q)\) with advantage \(\epsilon\), then there is no \(T\)-time distinguisher for the problem LWE\((m, \mathbb{R}_q^n, S, D_{\mathbb{R},\sigma'}^N, \text{mod}_q)\) with advantage \(\epsilon + \text{negl}(\lambda)\), where we set

\[
\sigma' = \sqrt{\sigma^2 + 9r^2(\ln n + \ln m + \omega (\log \lambda))} = O(\sigma).
\]

Similarly, the statements \(\sigma \geq 3r \sqrt{\ln n + \ln m + \omega (\log \lambda)}\) and \(\sigma' = \sqrt{\sigma^2 + 9r^2(\ln n + \ln m + \omega (\log \lambda))}\) imply that \(\sigma' = O(\sigma)\). So to make the samples continuous, we are again blowing up the noise scale by at most a constant multiplicative factor. Then we give a modified version of Lemma 18 in [GVV22]. We first need to introduce the following fact from [DKMR22b].

**Fact B.4** (Fact A.4 in [DKMR22b]). Let \(n \in \mathbb{N}, \sigma \in \mathbb{R}_+, \epsilon \in (0, 1/3]\) be such that \(\sigma \geq \sqrt{\ln (2n(1 + 1/\epsilon)))/\pi}\). Then, we have

\[
\frac{P_{D_{\mathbb{Z}_1,\sigma}^{\text{expand}}}(t)}{P_{D_{\mathbb{R}_1}^N}(t)} = \frac{P_{U(\mathbb{R}_1) (\text{mod}_1(\sigma t))}}{P_{D_{\mathbb{R}_1}^\text{collapse} (\text{mod}_1(\sigma t))}} = \exp(\log(\lambda)) = 1 \pm O(\epsilon),
\]

for all \(t \in \mathbb{R}_n\), and

\[
d_{TV}\left(D_{\mathbb{Z}_1,\sigma}^{\text{expand}}/\sigma, D_{\mathbb{R}_1,n}^N\right), d_{TV}\left(D_{\mathbb{R}_1,\sigma}^\text{collapse}, U(\mathbb{R}_1^n)\right) = \exp\left(-\Omega(\sigma^2)\right).
\]

Essentially, Fact B.4 says that, given \(x \sim D_{\mathbb{R}_n,\sigma}^N\), the distribution of \(\text{mod}_1(x)\) is pointwise close (for its pdf function) to \(U(\mathbb{R}_1^n)\) for sufficiently large \(\sigma\). So if we consider the reverse of this
process, given a \( \mathbf{v} \sim U(\mathbb{R}^n_1) \), we sample \( \mathbf{u} \sim D_{\mathbb{R}^n_1 + \mathbf{v}, \sigma}^N \), then the distribution of \( \mathbf{u} \) is sufficiently close to \( D_{\mathbb{R}^n_1, \sigma}^N \). We can leverage this fact to change the sample distribution in the LWE problem from \( U(\mathbb{R}^n_q) \) to \( D_{\mathbb{R}^n_1, 1}^N \) since \( U(\mathbb{R}^n_q) \) is basically \( U(\mathbb{R}^n_1) \) after rescaling. The difference here is that the original Lemma 18 takes a large \( \sigma \) so that \( d_{\text{TV}} \left( D_{\mathbb{R}^n_1, \sigma}^N, D_{\mathbb{R}^n_1, 1}^N \right) \approx 1/m \), thus \( m \) samples will not see the difference. However, since these two distributions are actually pointwise close, we can instead take a smaller \( \sigma \) and do an extra rejection sampling step on \( \mathbf{u} \) to make the distribution exactly a Gaussian. This allows us to give the nearly optimal lower bound on agnostic learning LTFs with Gaussian marginals. Now we give the modified version of Lemma 18 in [GVV22].

**Lemma B.5** (Modified Lemma 18 in [GVV22]). Let \( n, m, q \in \mathbb{N}, \sigma, \alpha \in \mathbb{R}_+ \). Let \( S \subseteq \mathbb{Z}^n \) where all elements in the support have fixed \( \ell_2 \)-norm \( r \). Suppose there is no \( T + \text{poly}(n, m, \log(q)) \)-time distinguisher for the problem \( \text{LWE}(m, \mathbb{R}^n_q, S, D_{\mathbb{R}^n_1, \sigma, \alpha, m}^N, \text{mod}_q) \) with \( \epsilon \) advantage. Then there is no \( T \)-time distinguisher for the problem \( \text{LWE}(m', D_{\mathbb{R}^n_1, 1}^N, S/r, D_{\mathbb{R}^n_1, \alpha, m}^N, \text{mod}_q) \) with \( \epsilon + 2^{-\Omega(m)} \) advantage, where

\[
\alpha = c/ \left( r \sqrt{\log n} \right),
\]

\[
m' = cm,
\]

and \( c > 0 \) is a sufficiently small universal constant.

**Proof.** We will give a reduction argument. Given a sample \( (\mathbf{x}, y) \) from \( \text{LWE}(m, \mathbb{R}^n_q, S, D_{\mathbb{R}^n_1, \sigma, \alpha, m}^N, \text{mod}_q) \), we can generate a sample \( (\mathbf{x}', y') \) from the problem \( \text{LWE}(m', D_{\mathbb{R}^n_1, 1}^N, S/r, D_{\mathbb{R}^n_1, \alpha, m}^N, \text{mod}_q) \) with at least a constant success probability in the following manner.

We take a \( \hat{\sigma} = 1/ra \) and sample \( \hat{\mathbf{x}} \sim D_{\mathbb{Z}^n/q, \hat{\sigma}}^N \). We define the function \( f : \mathbb{R}^n \to \mathbb{R} \) as

\[
f(t) \overset{\text{def}}{=} \frac{P_{D_{\mathbb{R}^n_1}^N}(t)}{P_{D_{\mathbb{R}^n_1, \hat{\sigma}}^N}(t)}.
\]

With probability \( f(\hat{\mathbf{x}})/\max_{t \in \mathbb{R}^n} f(t) \), we take \( \mathbf{x}' = \hat{\mathbf{x}} \) and \( y' = y/(q \hat{\sigma}) \) and output \( (\mathbf{x}', y') \) as a sample for \( \text{LWE}(m', D_{\mathbb{R}^n_1, 1}^N, S/r, D_{\mathbb{R}^n_1, \alpha, m}^N, \text{mod}_q) \). Otherwise, we output failure.

We will prove that if \( (\mathbf{x}, y) \) is from the alternative hypothesis case, then it must be \( \mathbf{x}' \sim D_{\mathbb{R}^n_1, 1}^N \) and \( y' = \text{mod}_q(\langle s', \mathbf{x}' \rangle + z') \), where \( s' \sim U(S/q) \) and \( z' \sim D_{\mathbb{R}^n_1, \alpha, m}^N \). Since \( (\mathbf{x}, y) \) is from the alternative hypothesis case, it must satisfy \( \mathbf{x} \sim U(\mathbb{R}^n_q) \) and \( y = \text{mod}_q(\langle \mathbf{s}, \mathbf{x} \rangle + z) \), where \( \mathbf{s} \sim U(S) \) and \( z \sim D_{\mathbb{R}^n_1, \sigma}^N \). Then, the fact \( \hat{\mathbf{x}} \sim D_{\mathbb{Z}^n/q, \hat{\sigma}}^N \) implies that \( \hat{\sigma} \mathbf{x} - \mathbf{x}/q \in \mathbb{Z}^n \) and \( q \sigma \hat{\mathbf{x}} = \mathbf{x} \in q\mathbb{Z}^n \); combined with \( \mathbf{s} \in \mathbb{Z}^n \), we have that

\[
\text{mod}_q(\langle \mathbf{s}, \mathbf{x} \rangle) = \text{mod}_q(\langle \mathbf{s}, q \hat{\sigma} \hat{\mathbf{x}} \rangle + \langle \mathbf{s}, \mathbf{x} - q \hat{\sigma} \hat{\mathbf{x}} \rangle) = \text{mod}_q(\langle \mathbf{s}, q \hat{\sigma} \hat{\mathbf{x}} \rangle).
\]

Then we can write

\[
y' = y/(q \hat{\sigma})
\]

\[
= \text{mod}_q(\langle \mathbf{s}, \mathbf{x} \rangle + z)/(q \hat{\sigma})
\]

\[
= \text{mod}_q(\langle \mathbf{s}, q \hat{\sigma} \hat{\mathbf{x}} \rangle + z)/(q \hat{\sigma})
\]

\[
= \text{mod}_1((\mathbf{s}, \hat{\sigma} \hat{\mathbf{x}} + z/q)/(r \hat{\sigma}))
\]

\[
= \text{mod}_1/(r \hat{\sigma})(\langle \mathbf{s}/r, \hat{\mathbf{x}} \rangle + z/(q r \hat{\sigma}))
\]

\[
= \text{mod}_a \left( \langle \mathbf{s}/r, \mathbf{x}' \rangle + \alpha z/q \right),
\]
where the last equality follows from the fact $\tilde{\sigma} = 1/(r\alpha)$. Note that the three terms in the above expression, $s/r$, $x'$, and $\alpha z/q$ are independent (since $x'$, $s$, $z$ are independent). It only remains to verify the distribution of each of them.

It is immediate that $s/r \sim U(S/r)$. For the other two, we first define the following notation. For functions $f, g : U \to \mathbb{R}$, we write $f(u) \propto g(u)$ if there is a constant $c \in \mathbb{R} \setminus \{0\}$ such that for all $u \in U$, it holds $f(u) = cg(u)$. For $x'$, we first notice that $x/q \sim \mathbb{R}^2_1$, and therefore $\tilde{x} \sim \mathbb{R}^{\frac{\log n}{\alpha}}_1/\tilde{\sigma}$. Combining with the rejection sampling procedure we performed, we have that

$$P_{x'}(u) \propto \frac{f(u)}{\max_{t \in \mathbb{R}^n} f(t)} P_{x}(u) = \frac{f(u)}{\max_{t \in \mathbb{R}^n} f(t)} P_{D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}}(u) = \frac{P_{D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}}(u)}{\max_{t \in \mathbb{R}^n} f(t)} \propto P_{D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}}(u).$$

Thus, we conclude that $x' \sim D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}$. For $\alpha z/q$, notice that $z \sim D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}$, and therefore $\alpha z/q \sim D_{\mathbb{R}, \alpha \tilde{\sigma}}$.

For the null hypothesis case, it is easy to see that the marginals satisfy $D_{x'} = D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1}$ and $D_{y'} = U(\mathbb{R}_n)$, and $x'$ and $y'$ are independent — since $x$ and $y$ are independent and $x'$ (resp. $y'$) only depends on $x$ (resp $y$).

It remains to verify that the sampling will produce at least $m'$ many samples with $1 - 2^{-O(m)}$ probability. We first show that each individual rejection sampling succeeds with at least a positive constant probability. From Fact B.4, given $\tilde{\sigma}$, we first notice that $x/q \sim \mathbb{R}^2_1$, and therefore $\tilde{x} \sim D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1, \tilde{\sigma}}$. For the null hypothesis case, it is easy to see that the marginals satisfy $D_{x'} = D_{\mathbb{R}^{\frac{\log n}{\alpha}}_1}$ and $D_{y'} = U(\mathbb{R}_n)$, and $x'$ and $y'$ are independent — since $x$ and $y$ are independent and $x'$ (resp $y'$) only depends on $x$ (resp $y$).

We note that Lemma B.5 is stronger than Lemma 18 in [GVV22] in the sense that the original Lemma 18 has $\alpha = c/\left(r \sqrt{\log n + \log m + \omega(\log \lambda)}\right)$, compared with $\alpha = c/(r \sqrt{\log n})$ here. For the task of learning LTFs, if one uses Lemma 18 instead of Lemma B.5 and follows the same argument for rest of the proof, one will still get an $n^{O(1/(c \sqrt{\log n})^{0.99})}$ lower bound — compared with the $n^{O(1/(c \sqrt{\log n})^{1.99})}$ near-optimal lower bound we establish here.

Combining the above lemmas and Assumption 2.3, we establish the proof of Proposition 2.4.

Proof of Proposition 2.4. We provide an efficient reduction from Assumption 2.3 via Lemma B.1, Lemma B.2, Lemma B.3 and Lemma B.5. More precisely, the reduction will follow the following steps:

1. Let the problem in Assumption 2.3 be solving LWE($2^{O(l^\beta)}, m,n,q$, $m_0$, $\tilde{s}$), with $2^{-O(l^\beta)}$ advantage, where $l$ is the dimension. We then use Lemma B.1 to reduce to solving the problem LWE($n^{O(k^\beta)}, m,n,q$), with $n^{-O(k^\beta)}$ advantage, where $c$ is a sufficient large positive universal constant, $n$ is the dimension and the secret vector is from the sparse set $S_{n,k}$.

2. We use Lemma B.2 and Lemma B.3. The two lemmas make the sample and noise distributions continuous. As we argued before, these two lemmas will only blow up the noise scale by a universal constant factor, so we reduce to solving LWE($n^{O(k^\beta)}, m,n,q$), with $n^{-O(k^\beta)}$ advantage, where $c$ is a sufficiently large positive universal constant.
4. To finish the reduction, we apply Lemma [B.3] which mainly changes the sample distribution from \( U(\mathbb{R}^n_0) \) to \( D^{N}_{\mathbb{R}^n,1} \), and reduce to solving the problem \( \text{LWE}(n^{O(k^\beta)}, D^{N}_{\mathbb{R}^n,1}, \mathbb{G}^{n-1}, D^{N}_{\mathbb{R}^{\sigma'}, \text{mod}_\alpha}) \) with \( n^{-O(k^\beta)} \) advantage.

To start the reduction, we need to chose the values for parameters \( l, \beta', q, \sigma' \) in the first step. Let \( n, k, \beta, \gamma, \kappa \) be the parameters in the body of Proposition [2.4] which are the target parameters we want to get after the reduction. For convenience, we let \( \delta > 0 \) be the constant such that \( 1 - 3\delta = \beta \).

Let \( \psi \) be the value such that \( k = \log^\psi n \) (\( \psi \) has dependence on \( n \) and \( k \)). We will chose the following values:

- \( l = \log^l n \), where \( t = 1 + \psi(1 - \delta) \);
- \( \beta' = \frac{1 + \gamma(1 - 2\delta)}{1 + \gamma(1 - \delta)} \), which is a constant, and \( \beta' \in (0, 1) \);
- \( q = k^{\kappa+1} \);
- \( \sigma' = c\sqrt{l} \), where \( c \) is a sufficiently large constant.

We now check validity of the parameters for each step of the reduction:

1. We first check that the parameters satisfy the requirements in Assumption [2.3] Notice that
   \[
   q = k^{\kappa+1} = \log^{(\kappa+1)} n = \log^t n 
   \leq \frac{n}{l} \leq \frac{n}{l} \leq \frac{l}{\log l} = l^{O(1)}.
   \]

2. We then check the requirements in Lemma [B.1] We chose the additional parameters as \( \lambda = 2^{l^{\beta'}} \) and \( m = n^{O(k^\beta)} \). For convenience, we first show that \( 2^{l^{\beta'}} = n^{O(k^\beta)} \). Notice that
   \[
   2^{l^{\beta'}} = 2^{\log^t n} = n^{\log^t n} = n^{\frac{t-1}{\log l}}.
   \]
   Since \( k = \log^\psi n \) and \( k \geq \log^\gamma n \), it follows that \( \psi \geq \gamma \); therefore, \( \beta' = \frac{1 + \gamma(1 - 2\delta)}{1 + \gamma(1 - \delta)} \geq \frac{1 + \psi(1 - 2\delta)}{1 + \psi(1 - \delta)} \).
   Plugging this into the above, we get that
   \[
   2^{l^{\beta'}} \geq n^{k^{\beta'}} = n^{\omega(k^\beta)}.
   \]

   where the last equality follows from the fact \( \beta = 1 - 3\delta \). For the requirements, we have:

   (a) It is immediate that \( \log(q)/2^l = O(\log l)/2^l = \text{negl}(l) \) (since \( q = l^{O(1)} \) from the last step).

   (b) For the requirement \( \sigma' \geq 4\sqrt{\omega(\log \lambda) + \ln n + \ln m} \), since \( \sigma' = c\sqrt{l} \), taking squares on both side, it can be rewritten as
   \[
   l = \omega(\log \lambda + \ln n + \ln m).
   \]

   Notice that \( \log \lambda = O(l^{\beta'}) \), where \( \beta' < 1 \); thus, \( l = \omega(\log \lambda) \). Since \( l = \log^t n \), where \( t = 1 + \psi(1 - \delta) \geq 1 + \gamma(1 - \delta) \) and \( \gamma(1 - \delta) \) is a positive constant, we have that \( l = \omega(\ln n) \).

   Then, since \( 2^{l^{\beta'}} = n^{\omega(k^\beta)} \) as shown above, and \( m = n^{O(k^\beta)} \), we get that \( 2^{l^{\beta'}} = \omega(m) \); thus, we get \( l = \omega(l^{\beta'}) = \omega(\log m) \). Combining the above gives us that \( l = \omega(\log \lambda + \ln n + \ln m) \).

   (c) For the requirement \( k \log_2(n/k) \geq (l + 1) \log_2(q) + \omega(\log \lambda) \), since \( l = \omega(\log \lambda) \) as shown above, we can rewrite it as \( k \log_2(n) - k \log_2(k) \geq 2l \log_2(q) \). Since \( q = \text{poly}(l) \) from step 1, it therefore suffices to show that \( k \log n - k \log k = \omega(l \log l) \), which is \( k \log n \geq c \log l + k \log k \) for any constant \( c \). We prove this by analyzing two cases, namely \( c \log l \leq k \log k \) and \( c \log l > k \log k \).
If $cl\log l \leq k\log k$, then since $k < c' n$, where $c'$ is a sufficiently small universal constant, we get that $k\log n \geq 2k\log k \geq cl\log l + k\log k$.

If $cl\log l > k\log k$, then it suffices to show that $k\log n = \omega(l\log l)$. Notice that $k\log n = \log^{1+\psi} n$ and $l\log l = t\log t n\log n$. Thus,

$$\frac{k\log n}{l\log l} = \frac{\log^{1+\psi-t} n}{t\log t}.$$

Notice that $1+\psi-t = \delta\psi \geq \delta\gamma$ (since $k \geq \log^2 n$ and $k = \log^\psi n$ implies $\psi \geq \gamma$) is at least a constant; thus,

$$\frac{k\log n}{l\log l} = \frac{\log^{1+\psi-t} n}{t\log t} = \frac{\log^\delta n}{t\log t} = \omega\left(\frac{\log^{\delta/2} n}{t}\right) = \omega\left(\frac{\log^{\delta/2} n}{1+\psi}\right),$$

where the last equality comes from the fact that $t = 1+\psi(1-\delta) \leq 1+\psi$. Therefore, we just need to show that $\log^{\delta/2} n$ is at least a constant. Notice that for any sufficiently large $n$ such that $\log^{\delta/2} n \geq e$, we have that

$$\log^{\delta/2} n = (\log^{\delta/2} n)^\psi \geq e^\psi \geq 1+\psi.$$

Thus, we have that

$$\frac{k\log n}{l\log l} = \omega(1),$$

which is $k\log n = \omega(l\log l)$.

Therefore, the requirement $k\log_2(n/k) \geq (l+1)\log_2(q) + \omega(\log \lambda)$ is satisfied in both cases.

(d) It only remains to verify the time lower bound of $2^{-O(l^{1/\psi})}$ and advantage $2\epsilon m + \negl(\lambda)$ in Lemma 13B.1, where $\epsilon$ is the advantage before the reduction. Notice that since $2^{l^{1/\psi}} \leq n^{\omega(k^\beta)}$, the time lower bound is at least any $n^{O(k^{\beta})}$. For the advantage, by taking $\epsilon = 2^{-3(l^{1/\psi})}$, we have that

$$2\epsilon m + \negl(\lambda) = 2^{-3(l^{1/\psi})} n^{O(k^{\beta})} + \negl(2^{l^{1/\psi}}) \leq 2^{-2(l^{1/\psi})} + \negl(2^{l^{1/\psi}}) = n^{-\omega(k^{\beta})},$$

where the last inequality and equality follows from the statement $2^{l^{1/\psi}} \leq n^{\omega(k^{\beta})}$ shown above. Thus, there is no $n^{O(k^{\beta})}$-time distinguisher for solving LWE($\lambda^{O(k^{\beta})}$, $Z_q^n$, $S_{n,k}$, $D_{\mathbb{R},c\sqrt{k^\sigma'}}$, $\mod_q$) with $n^{-O(k^{\beta})}$ advantage.

3. We then check the parameter requirements in Lemma 13B.2 and Lemma 13B.3. Note that it suffices to check that $c\sqrt{k^{\sigma'}} \geq 3r\sqrt{\ln n + \ln m + \omega(\log \lambda)}$ for sufficiently large constant $c$. Since $r = \sqrt{k}$ from its definition and we have already shown that $\sigma' \geq 4\sqrt{\omega(\log \lambda)} + \ln n + \ln m$ in Step 2b, this inequality holds.

Then it only remains to verify the time lower bound and advantage. The time lower bound is $n^{c\sqrt{k^{\sigma'}}} - \poly(m, n, \log(q), \log(c\sqrt{k^{\sigma'}}))$. Since $m = n^{O(k^{\beta})}$, $\log(q) = \log(k^{c+1}) = O(\log k)$, and $\log(c\sqrt{k^{\sigma'}}) = O(\log k + \log l) = O(\log^{1+\psi} n) = O(k\log n)$, by choosing $c$ to be a sufficiently large constant, the above lower bound is any $n^{O(k^{\beta})}$. Similarly, the advantage is any $n^{-O(k^{\beta})}$. Thus, there is no $n^{O(k^{\beta})}$-time distinguisher for solving the problem LWE($\lambda^{O(k^{\beta})}$, $\mathbb{R}_q^n$, $S_{n,k}$, $D_{\mathbb{R},c\sqrt{k^\sigma'}}$, $\mod_q$) with $n^{-O(k^{\beta})}$ advantage.
4. After applying Lemma B.5, we get that there is no $n^{O(k^3)}$-time distinguisher for solving the problem LWE($m', D_{R_{n,1}}^N, S_{n,k}/\sqrt{k}, D_{R,ca\sqrt{k}\sigma'/q}$, mod$_a$) with $n^{-O(k^3)}$ advantage, where $\alpha = c/(\sqrt{k \log n}$), $m' = cn^{O(k^3)}$, and $c > 0$ is a sufficiently small universal constant. We just need to check that it matches the values of $\sigma, m, T$ in the body of Proposition 2.4. For the noise scale $\sigma$, we have

$$ca\sqrt{k}\sigma'/q = c' \sqrt{l/(\sqrt{\log n q})} = c' \log n^{\psi(1-\delta)/2}/q \leq c' k^{1/2}/k^{\kappa+1} = o(k^{-\kappa}) = o(\sigma) ,$$

where the last inequality follows from $k = \log^\psi n$. For the number of samples, we have that $m' = cn^{\epsilon(k^3)}$ which is any $n^{O(k^3)}$ by choosing $c'$ to be sufficiently large. For the parameter $T$, we have that $\alpha = c/(\sqrt{k \log n}) = T$. Then, the only remaining difference is that the secret vector distribution is $U(S_{n,k}/\sqrt{k})$ instead of $U(S^{n-1})$. The catch here is that we can do a random rotation on all the samples and this makes the secret vector also randomly rotated and gives the $U(S^{n-1})$ distribution we want. Therefore, there is no $n^{O(k^3)}$-time distinguisher for solving the problem LWE($n^{O(k^3)}, D_{R_{n,1}}^N, S^{n-1}, D_{R,\sigma}$, mod$_T$) with $n^{-O(k^3)}$ advantage.

This proves Proposition 2.4. \qed