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

Ilias Diakonikolas [*1] **Daniel M. Kane** [*2] **Lisheng Ren** [*1]

Abstract

We study the task of agnostically learning halfspaces under the Gaussian distribution. Specifically, given labeled examples $(\mathbf{x}, y)$ from an unknown distribution on $\mathbb{R}^n \times \{\pm 1\}$, whose marginal distribution on $\mathbf{x}$ is the standard Gaussian and the labels y can be arbitrary, the goal is to output a hypothesis with 0-1 loss $\mathrm{OPT} + \epsilon$, 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.

1. Introduction

A halfspace or Linear Threshold Function (LTF) is any Boolean-valued function $f : \mathbb{R}^n \to \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - t)$, where $\mathbf{w} \in \mathbb{R}^n$ is called the weight vector and $t \in \mathbb{R}$ is called the threshold. Here the univariate function $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\mathrm{sign}(u) = 1$ if $u \geq 0$ and $\mathrm{sign}(u) = -1$ 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 (Freund & Schapire, 1997). In the realizable setting (Valiant, 1984), halfspaces are known to be efficiently learnable (see, e.g., (Maass & Turan, 1994)) without distributional assumptions. In contrast, in the distribution-free agnostic model (Haussler, 1992; Kearns et al., 1994), even *weak* learning is computationally hard (Guruswami & Raghavendra, 2006; Feldman et al., 2006; Daniely, 2016; Tiegel, 2022). Due to this computational intractability, a significant branch of research has focused on agnostically learning halfspaces in the *distribution-specific* setting. Intuitively, the underlying structure of the data distribution can potentially be leveraged to obtain non-trivial efficient algorithms robust to adversarial label noise.

Here we focus on the well-studied task of agnostically learning halfspaces *when the underlying distribution on examples is assumed to be Gaussian*. That is, we are given i.i.d. samples from a joint distribution D on labeled examples $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^n$ is the example and $y \in \mathbb{R}$ is the corresponding label, and the goal is to compute a hypothesis that is competitive with the best-fitting halfspace. Moreover, we assume that the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. As we will explain subsequently, the distributional assumption makes the learning problem computationally easier, as compared to the distribution-free setting. Interestingly, *even the Gaussian version of the problem exhibits information-computation tradeoffs that we explore — and essentially resolve — in this paper.*

For concreteness, we introduce some notation followed by the definition of the aforementioned problem. For a boolean-valued hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ and a distribution D supported on $\mathbb{R}^n \times \{\pm 1\}$, we use $R_{0-1}(h; D)$ to denote the 0-1 error of h with respect to D, i.e., $R_{0-1}(h; D) \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y]$. For a class $\mathcal{C}$ of boolean-valued functions on $\mathbb{R}^n$, we use $R_{0-1}(\mathcal{C}; D)$ to denote the minimum 0-1 error of any $h \in \mathcal{C}$, i.e., $R_{0-1}(\mathcal{C}; D) \stackrel{\mathrm{def}}{=} \min_{h \in \mathcal{C}} R_{0-1}(h; D)$.

Problem 1.1 (Agnostically Learning Halfspaces under Gaussian Marginals)**.** *Let* LTF *be the class of halfspaces on $\mathbb{R}^n$. Given an error parameter $0 < \epsilon < 1$ and i.i.d. samples $(\mathbf{x}, y)$ from a distribution D on $\mathbb{R}^n \times \{\pm 1\}$, where the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and no assumptions are made on the labels y, the goal of the learning algorithm $\mathcal{A}$ is to output a hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such that $R_{0-1}(h; D) \leq R_{0-1}(\mathrm{LTF}; D) + \epsilon$ with high probability. We will say that the algorithm $\mathcal{A}$ agnostically learns halfspaces (or LTFs) under Gaussian*

[*]Equal contribution [1]University of Wisconsin-Madison [2]University of California, San Diego. Correspondence to: Lisheng Ren <lren29@wisc.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

marginals to additive error ϵ.

Prior Work on Problem 1.1 By standard results (Haussler, 1992; Kearns et al., 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 (Kalai et al., 2008) solves Problem 1.1 with sample complexity and running time $n^{O(1/\epsilon^2)}$ (Diakonikolas et al., 2010a;b). While the L_1-regression algorithm is not proper, recent work developed a proper learner with qualitatively similar sample and time complexities (namely, $n^{\mathrm{poly}(1/\epsilon)}$) (Diakonikolas et al., 2021a). 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 (Goel et al., 2020; Diakonikolas et al., 2020b; 2021b) has established tight hardness in the Statistical Query (SQ) model. SQ algorithms (Kearns, 1998) are a class of algorithms that are only allowed to query expectations of bounded functions of the distribution rather than directly access samples. (Diakonikolas et al., 2021b) leveraged the framework of (Diakonikolas et al., 2017) to show that any SQ algorithm for the problem either requires $2^{n^{\Omega(1)}}$ 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., (Dachman-Soled et al., 2015)) 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 (Klivans & Kothari, 2014). 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 (Klivans & Kothari, 2014) implies a computational lower bound of $n^{\Omega(\log(1/\epsilon))}$ 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 $C \cdot \mathrm{OPT} + \epsilon$, for a sufficiently large universal constant $C > 1$, $\mathrm{poly}(d/\epsilon)$ time algorithms are known (Awasthi et al., 2017; Daniely, 2015; Diakonikolas et al., 2018).

In summary, the best known algorithm for Problem 1.1 has sample complexity and running time $n^{\mathrm{poly}(1/\epsilon)}$, while the best known computational hardness result gives an $n^{\Omega(\log(1/\epsilon))}$ 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 $n^{\mathrm{poly}(1/\epsilon)}$ (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 $\sigma((\langle \mathbf{w}, \mathbf{x} \rangle - t))$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a univariate activation. If the activation is better behaved than the sign function, specifically if σ 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 $f(\mathbf{x}) = \mathrm{ReLU}(\langle \mathbf{w}, \mathbf{x} \rangle - t)$, where $\mathbf{w} \in \mathbb{R}^n$ is called the weight vector and $t \in \mathbb{R}$ is called the threshold. The activation $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}_+$ is defined as $\mathrm{ReLU}(u) = \max\{0, u\}$. 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., (Soltanolkotabi, 2017; Goel et al., 2017; Manurangsi & Reichman, 2018; Goel et al., 2019; Frei et al., 2020; Diakonikolas et al., 2020a; 2022c; Awasthi et al., 2022). 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 (Manurangsi & Reichman, 2018; Diakonikolas et al., 2022a). 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., $R_2(h; D) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim D}[(h(\mathbf{x}) - y)^2]$. 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., $R_2(\mathcal{C}; D) \stackrel{\text{def}}{=} \min_{h\in\mathcal{C}} R_2(h; D)$.

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

Prior Work on Problem 1.2 While there is no black-box relation with Problem 1.1, the situation for both problems is analogous. (Diakonikolas et al., 2020a) gave an algorithm for Problem 1.2 with sample complexity and runtime $n^{\mathrm{poly}(1/\epsilon)}$. While $\mathrm{poly}(n/\epsilon)$ time algorithms are known with weaker guarantees (Goel et al., 2019; Diakonikolas et al., 2020a; 2022c), the fastest known algorithm with $\mathrm{OPT} + \epsilon$ error is the one of (Diakonikolas et al., 2020a). In terms of computational hardness, (Goel et al., 2019) 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 $n^{\mathrm{poly}(1/\epsilon)}$ have been shown (Goel et al., 2020; Diakonikolas et al., 2020b; 2021b).

In summary, the best known algorithm for Problem 1.2 has sample complexity and running time $n^{\mathrm{poly}(1/\epsilon)}$, 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 $n^{\mathrm{poly}(1/\epsilon)}$ can be established. Similarly to the case of LTFs, we prove such a statement (Theorem 1.4) 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 $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ and the goal is to distinguish between the following two cases:

- Each $\mathbf{x}_i$ is drawn uniformly at random (u.a.r.) from $\mathbb{Z}_q^n$, and there is a hidden secret vector $\mathbf{s} \in \mathbb{Z}_q^n$ such that $y_i = \langle \mathbf{x}_i, \mathbf{s}\rangle + z_i$, where $z_i \in \mathbb{Z}_q$ is discrete Gaussian noise (independent of $\mathbf{x}_i$).
- Each $\mathbf{x}_i$ 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 $2^{n^{1-\Omega(1)}}$ 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 ϵ requires running time at least $\min\{n^{\Omega(1/(\epsilon\sqrt{\log n})^{\alpha})}, 2^{n^{0.99}}\}$.*

Some comments are in order to interpret this statement. The minimum of the two terms is necessary to handle the case where ϵ is very small, specifically $\epsilon = \tilde{O}(1/\sqrt{n})$. (Since the problem can always be solved in time $2^{\tilde{O}(n)}$ via brute-force, the first term cannot be a time lower bound for this range of ϵ.) On the other hand, for $\tilde{\Omega}(1/\sqrt{n}) = \epsilon \leq 1/\log^{1/2+c}(n)$, Theorem 1.3 gives a time lower bound of $n^{\Omega(1/(\epsilon\sqrt{\log n})^{\alpha})}$, for any constant $\alpha < 2$. This bound nearly matches the upper bound of $n^{O(1/\epsilon^2)}$ (Kalai et al., 2008), up to the $\sqrt{\log n}$ factor in the exponent. Note that the extraneous factor of $\sqrt{\log n}$ is negligible if ϵ 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 $\alpha < 2$.

Independent Work In independent and concurrent work, (Tiegel, 2022) showed a quantitatively similar lower bound as our Theorem 1.3. Their result requires that $\epsilon \leq 1/\sqrt{n}$ in Problem 1.1, while our result allows a wider range of ϵ, roughly as long as $\epsilon \leq 1/\sqrt{\log(n)}$.

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 ϵ 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 ϵ 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 $n^{\mathrm{poly}(1/\epsilon)}$ (Diakonikolas et al., 2020a) and exponentially improves on the best known computational lower bound of $n^{\Omega(\log(1/\epsilon))}$ (Goel et al., 2019).

1.2. Techniques

Our computational hardness reductions build on two main ideas. The first idea is inspired by the approach of (Di-

akonikolas et al., 2022b). We note that (Diakonikolas et al., 2022b) 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 (Diakonikolas et al., 2022b) have no implications for the Gaussian setting. Yet one of their ideas is useful in our context.

The key idea of (Diakonikolas et al., 2022b) 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 (Gupte et al., 2022),) 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, (Diakonikolas et al., 2022b) manage to produce a joint distribution on $(\mathbf{x}, y)$ over $\mathbb{R}^n \times \{\pm 1\}$ such that:

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

(ii) in the alternative hypothesis case[1], y is given by a Polynomial Threshold Function (PTF) applied to $\mathbf{x}$ with Massart noise.

Given the above, (Diakonikolas et al., 2022b) 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 leaning halfspaces and ReLUs under Gaussian marginals. A key difference in our setting is that we require the distribution of $\mathbf{x}$ be the standard Gaussian — a property inherently not satisfied by the aforementioned construction. Roughly speaking, (Diakonikolas et al., 2022b) 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 (Diakonikolas et al., 2022b) 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.

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

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 $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[f_{\mathbf{s}}(\mathbf{x}) g(\mathbf{x})] = \epsilon = \Omega(1/\|\mathbf{s}\|_2)$. This implies that any algorithm that agnostically learns LTFs to error $\mathrm{OPT} + \epsilon/3$, where $\mathrm{OPT} = R_{0-1}(\mathrm{LTF}; D)$, can be used to distinguish between the case that y is independent of $\mathbf{x}$ (in which case $\mathrm{OPT} = 1/2$) and the case described above (i.e., $y = f_{\mathbf{s}}(\mathbf{x}) = (-1)^{\lfloor \langle \mathbf{x}, \mathbf{s} \rangle \rfloor}$), where $\mathrm{OPT} = 1/2 - \epsilon$. 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 $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[f_{\mathbf{s}}(\mathbf{x}) g(\mathbf{x})] = \epsilon = \Omega(1/\|\mathbf{s}\|_2^2)$. 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 $y = f_{\mathbf{s}}(\mathbf{x})$ (and thus the minimal L_2^2-error for ReLUs is at most $1 - \epsilon^2$) and the case where y is independent of $\mathbf{x}$ (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 (Bruna et al., 2021) which has secret vector $\mathbf{s}$ with $\|\mathbf{s}\|_2 = \sqrt{n}$, we can obtain a time lower bound for our agnostic learning problems *only if the additive error ϵ is tiny*, namely $\epsilon = \tilde{O}(1/\sqrt{n})$. In order to prove lower bounds for a wider range of ϵ, we will need to instead start from a *small norm version* of the continuous LWE problem, where the secret vector $\mathbf{s}$ roughly satisfies $\|\mathbf{s}\|_2 \approx 1/\epsilon$. We accomplish this via a non-trivial modification of a reduction in (Gupte et al., 2022), which we view as an additional technical contribution of this work. Specifically, (Gupte et al., 2022) gave a reduction of the standard discrete LWE problem to a discrete LWE problem with a sparse secret (namely, secret vector $\mathbf{s} \in \{0, \pm 1\}^n$ with $\|\mathbf{s}\|_1 = k$). (This itself leverages an idea in (Micciancio, 2018).) After that, (Gupte et al., 2022) further reduces the sparse secret discrete LWE problem to a continuous LWE problem whose secret vector has small ℓ_2-norm. The limitation here is

[1] This leverages a construction of such a distribution from (Diakonikolas & Kane, 2022).

that their ℓ_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 B.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 $\langle \mathbf{x}, \mathbf{y} \rangle$ for the inner product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. For $p \geq 1$ and $\mathbf{x} \in \mathbb{R}^n$, we use $\|\mathbf{x}\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^n |\mathbf{x}_i|^p\right)^{1/p}$ to denote the ℓ_p-norm of $\mathbf{x}$. We use $\mathbb{S}^{n-1}$ to denote the unit sphere in $\mathbb{R}^n$, i.e., the set $\mathbb{S}^{n-1} \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$. For $q \in \mathbb{N}$, we denote $\mathbb{Z}_q \stackrel{\text{def}}{=} \{0, 1, \cdots, q-1\}$ and $\mathbb{R}_q \stackrel{\text{def}}{=} [0, q)$. We use $\mathrm{mod}_q : \mathbb{R}^n \to \mathbb{R}_q^n$ to denote the function that applies the mod_q operation on each coordinate of the vector $\mathbf{x}$. For a set $S \subset \mathbb{R}^n$, we use $U(S)$ to denote the uniform distribution over S. We use $\mathbf{x} \sim D$ to denote a random variable $\mathbf{x}$ with distribution D. For a random variable $\mathbf{x}$ (resp. a distribution D), we use $P_{\mathbf{x}}$ (resp. P_D) to denote the probability density function or probability mass function of the random variable $\mathbf{x}$ (resp. distribution D). We will require the following notion of partially supported Gaussian.

Definition 2.1 (Partially Supported Gaussian Distribution)**.** For $\sigma \in \mathbb{R}_+$ and $\mathbf{x} \in \mathbb{R}^n$, let $\rho_\sigma(\mathbf{x}) \stackrel{\text{def}}{=} \sigma^{-n} \exp\left(-\pi(\|\mathbf{x}\|_2/\sigma)^2\right)$. For any countable set[2] $S \subseteq \mathbb{R}^n$, we let $\rho_\sigma(S) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in S} \rho_\sigma(\mathbf{x})$, and let $D^{\mathcal{N}}_{S,\sigma}$ be the distribution supported on S with pmf $P_{D^{\mathcal{N}}_{S,\sigma}}(\mathbf{x}) = \rho_\sigma(\mathbf{x})/\rho_\sigma(S)$.

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

Learning with Errors The Learning with Errors (LWE) problem was introduced in (Regev, 2005). 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 $\mathbb{R}^n, \mathbb{R}^n, \mathbb{R}$ respectively. In the $\mathrm{LWE}(m, D_{\text{sample}}, D_{\text{secret}}, D_{\text{noise}}, \mathrm{mod}_q)$ problem, we are given m independent samples $(\mathbf{x}, y)$ and want to distinguish between the following two cases:

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

(ii) **Null hypothesis**: The random variables $\mathbf{x}$ and y are independent. Moreover, $\mathbf{x}$ 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_{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 $\mathrm{LWE}(2^{O(n^\beta)}, \mathbb{Z}_q^n, \mathbb{Z}_q^n, D^{\mathcal{N}}_{\mathbb{Z},\sigma}, \mathrm{mod}_q)$ 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. (Regev, 2005; Peikert, 2009) gave a polynomial-time *quantum* reduction from approximating (the decision version of) the Shortest Vector Problem (GapSVP) to LWE (with similar n, q, σ parameters). We note that the fastest known algorithm for GapSVP takes $2^{O(n)}$ time (Aggarwal et al., 2020). Thus, refuting the conjecture would be a major breakthrough. A similar assumption was also used in (Gupte et al., 2022) and (Diakonikolas et al., 2022b) 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 (Bruna et al., 2021)) 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 (Gupte et al., 2022) and gives the reduction from the standard LWE to the continuous LWE. The proof is deferred to Appendix B.

Proposition 2.4 (Hardness of continuous LWE (cLWE) with Small-Norm Secret)**.** *Under Assumption 2.3, for any $n \in \mathbb{N}$, any constants $\beta \in (0, 1)$, $\kappa \in \mathbb{N}$, $\gamma \in \mathbb{R}_+$ and any $\log^\gamma n \leq k \leq cn$ where $c > 0$ is a sufficiently small universal constant, the problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ 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.*

[2] We will take the sets S to be shifts of lattices, guaranteeing that $\rho_\sigma(S)$ is finite and the distribution is well-defined.

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 establish the desired cryptographic hardness of agnostically learning LTFs under the Gaussian distribution.

The high-level idea is the following. Given samples $(\mathbf{x}, y)$ from a distribution D on $\mathbb{R}^n \times \mathbb{R}_T$, which is an instance of the cLWE problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ (note that T is the "period" of the periodic signal on the hidden direction) from Proposition 2.4, we efficiently generate samples $(\mathbf{x}, y')$ (we leave $\mathbf{x}$ 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 $h : \mathbb{R}^n \to \{\pm 1\}$ such that $R_{0-1}(h; D') \leq 1/2 - \Omega(T)$.

(ii) If D is from the null hypothesis case, then for $(\mathbf{x}, y') \sim D'$, we have that $y' = +1$ with probability $1/2$ and $y' = -1$ with probability $1/2$ independent of $\mathbf{x}$; 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 $R_{0-1}(\mathrm{LTF}; D') + o(T)$, 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 $(\mathbf{x}, y)$ supported on $\mathbb{R}^n \times \{\pm 1\}$ with marginal $D_{\mathbf{x}} = D^{\mathcal{N}}_{\mathbb{R}^n,1}$, with $n^{-O(k^\beta)}$ advantage:*

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

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

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 $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ 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 $(\mathbf{x}, y')$ from a distribution D', which is an instance of the problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ from Proposition 2.4, we simply output $(\mathbf{x}, y) \sim D$, 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 $D_{\mathbf{x}}$ of D satisfies $D_{\mathbf{x}} = D^{\mathcal{N}}_{\mathbb{R}^n,1}$, therefore it suffices to verify that $R_{0-1}(\mathrm{LTF}; D) = 1/2 - \Omega\left(1/\sqrt{k \log n}\right)$ and $y = +1$ with probability $1/2$ independent of $\mathbf{x}$ for each case respectively.

For the alternative hypothesis case, let D' be from the alternative hypothesis case of the LWE. Let $\mathbf{s}$ be the secret vector in the LWE problem. We consider the following two LTFs: $h_1(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{s}, \mathbf{x} \rangle - T/6)$ and $h_2(\mathbf{x}) = \mathrm{sign}(-\langle \mathbf{s}, \mathbf{x} \rangle + T/3)$. 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 $R_{0-1}(\mathrm{LTF}; D) \leq R_{0-1}(h; D) \leq 1/2 - \Omega(1/\sqrt{k \log n})$ 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

$$\begin{aligned} B \stackrel{\mathrm{def}}{=} &\{\mathbf{t} \in \mathbb{R}^n \mid h_1(\mathbf{t}) = h_2(\mathbf{t})\} \\ = &\{\mathbf{t} \in \mathbb{R}^n \mid \langle \mathbf{s}, \mathbf{t} \rangle \in [T/6, T/3]\} \, . \end{aligned}$$

Since for any $\mathbf{t} \in B$, it is always the case that $h_1(\mathbf{t}) = h_2(\mathbf{t}) = +1$, we can write

$$\begin{aligned} &R_{0-1}(h_1; D) + R_{0-1}(h_2; D) \\ =&\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h_1(\mathbf{x})] + \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h_2(\mathbf{x})] \\ =&\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \notin B \wedge y \neq h_1(\mathbf{x})] \\ &+ \mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \notin B \wedge y \neq h_2(\mathbf{x})] \\ &+ 2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \in B \wedge y = -1] \, . \end{aligned}$$

Since for any $\mathbf{x} \notin B$ we have that $h_1(\mathbf{x}) \neq h_2(\mathbf{x})$, the first two terms sum to $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x} \notin B]$. Therefore, we have

that

$$\begin{aligned}&R_{0-1}(h_1;D)) + R_{0-1}(h_2;D)\\ =&\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x}\notin B] + 2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x}\in B\wedge y=-1]\\ =&1+\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x}\in B\wedge y=-1]\\ &-\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathbf{x}\in B\wedge y=+1]\\ =&1-\mathbf{Pr}[\mathbf{x}\in B](1-2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1\mid \mathbf{x}\in B])\ .\end{aligned}$$

From the definition of B and $\mathbf{x}\sim D^{\mathcal{N}}_{\mathbb{R}^n,1}$, we have $\mathbf{Pr}[\mathbf{x}\in B]=\Omega(T)$. Thus, we obtain

$$\begin{aligned}&R_{0-1}(h_1;D)+R_{0-1}(h_2;D)\\ =&1-\Omega(T)\left(1-2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1\mid \mathbf{x}\in B]\right)\ .\end{aligned}\tag{1}$$

If we can show that $\mathbf{Pr}[y=-1\mid \mathbf{x}\in B]\le 1/3$, then we are done since this implies that $R_{0-1}(h_1;D)+R_{0-1}(h_2;D)\le 1-\Omega(T)$.

We note that from the definition of the Alternative case distribution of the LWE problem, we have

$$y'=\mathrm{mod}_T(\langle\mathbf{s},\mathbf{x}\rangle+z)\ ,$$

and that $y=-1$ only if $y'>T/2$, which in turn happens only if

$$\langle\mathbf{s},\mathbf{x}\rangle+z>T/2 \text{ or } \langle\mathbf{s},\mathbf{x}\rangle+z<0\ .$$

For $\mathbf{x}\in B$, we have that $\langle\mathbf{s},\mathbf{x}\rangle\in[T/6,T/3]$, therefore $y=-1$ only if $|z|\ge T/6$. Notice that $z\sim D^{\mathcal{N}}_{\mathbb{R},\sigma}$ and Proposition 2.4 states that the LWE problem is hard for any fixed constant $\kappa\in\mathbb{N}$ and $\sigma\ge k^{-\kappa}$. 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, 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

$$\begin{aligned}\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1\mid\mathbf{x}\in B]&\le\mathbf{Pr}_{z\sim D^{\mathcal{N}}_{\mathbb{R},\sigma}}[|z|\ge T/6]\\ &=o(1)\ .\end{aligned}$$

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

$$\begin{aligned}&R_{0-1}(h_1;D)+R_{0-1}(h_2;D)\\ =&1-\Omega(T)\left(1-2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1\mid\mathbf{x}\in B]\right)\\ \le&1-\Omega(T)\ .\end{aligned}$$

Then, as argued above, if both $h=h_1$ and $h=h_2$ do not satisfy $R_{0-1}(h;D)\le 1/2-\Omega(T)$, 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 $R_{0-1}(h;D)\le 1/2-\Omega(T)\le 1/2-\Omega\left(1/\sqrt{k\log n}\right)$. 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 $\mathbf{x}$, since $y'\sim U([0,T))$ independent of $\mathbf{x}$ 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, for the problem $\mathrm{LWE}(n^{O(k^\beta)},D^{\mathcal{N}}_{\mathbb{R}^n,1},\mathbb{S}^{n-1},D^{\mathcal{N}}_{\mathbb{R},\sigma},\mathrm{mod}_T)$ with any $\sigma\ge k^{-\kappa}$ (where $\kappa\in\mathbb{N}$ is a constant) and $T=1/(c'\sqrt{k\log n})$, where $c'>0$ is a sufficiently large universal constant, the problem cannot be solved in $n^{O(k^\beta)}$ time with $n^{-O(k^\beta)}$ 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 $n^{O(k^\beta)}$ with $n^{-O(k^\beta)}$ 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 $c/(\sqrt{n\log n})\le\epsilon\le 1/\log^\gamma n$ 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 ϵ and runs in time $n^{O(1/(\epsilon\sqrt{\log n})^\alpha)}$.*

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 ϵ can solve the testing problem of Theorem 3.1 with probability $2/3$. Therefore, no such algorithm should run in time $n^{O(k^\beta)}$ for any $\beta\in(0,1)$. Since $\epsilon=c/\sqrt{k\log n}$, and if we chose $\beta=\alpha/2$, then the time lower bound can be rewritten as $n^{O(k^\beta)}=n^{O(1/(\epsilon\sqrt{\log n})^{2\beta})}=n^{O(1/(\epsilon\sqrt{\log n})^\alpha)}$. 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 $x\le t$. Roughly, our result says that any algorithm that solves this problem to error $\mathrm{OPT}+\epsilon$ with Gaussian marginals requires $n^{\mathrm{poly}(1/(\epsilon\log^2 n))}$ 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 $\log^\gamma n \le k \le cn$, 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 joint distribution D on $(\mathbf{x}, y)$ supported on $\mathbb{R}^n \times \{\pm 1\}$ with marginal $D_{\mathbf{x}} = D^{\mathcal{N}}_{\mathbb{R}^n,1}$, with $n^{-O(k^\beta)}$ advantage:

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

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

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 $D_u = \mathcal{D}^{\mathcal{N}}_{\mathbb{R},1}$ and $\mathbf{E}_{(u,y)\sim D}[y] = 0$. Suppose there is a ReLU of the form $f(u) = \mathrm{ReLU}(u - t)$ such that $t \ge 0$ and $\left|\mathbf{E}_{(u,y)\sim D}[yf(u)]\right| \ge \epsilon$. Then there exists $k \in (-1, 1)$ such that the ReLU $g(u) = \mathrm{ReLU}(ku - kt)$ satisfies $\mathbf{E}_{(u,y)\sim D}[(y - g(u))^2] \le 1 - \epsilon^2$.*

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

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

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

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

$$\begin{aligned}
\mathbf{E}_{(u,y)\sim D}[g(u)^2] &= k^2\mathbf{E}_{(u,y)\sim D}[f(u)^2] \\
&\le k^2\mathbf{E}_{(u,y)\sim D}[u^2] = k^2 \,.
\end{aligned}$$

In summary, we get that

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

We now choose the value of k. If $\mathbf{E}_{(u,y)\sim D}[yf(u)] > 0$, 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

$$\begin{aligned}
\mathbf{E}_{(u,y)\sim D}[(y - h(u))^2] &\le 1 + \epsilon^2 - 2\epsilon|\mathbf{E}_{(u,y)\sim D}[yf(u)]| \\
&\le 1 - \epsilon^2 \,.
\end{aligned}$$

□

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 $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ with any $\sigma \ge k^{-\kappa}$ ($\kappa \in \mathbb{N}$ is a constant) 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. We will give an efficient reduction of the LWE problem to the problem here.

For a sample $(\mathbf{x}, y')$ from a distribution D' which is an instance of the problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$, we will simply output $(\mathbf{x}, y)$ such that: (i) $y = +1$ if $y' \le T/2$, 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 $D_{\mathbf{x}}$ of D satisfies $D_{\mathbf{x}} = D^{\mathcal{N}}_{\mathbb{R}^n,1}$, it is enough to show that in the alternative hypothesis case, we have $R_2(\mathrm{ReLU}; D) = 1 - \Omega(1/(k \log n)^2)$, and in the null hypothesis case, we have $y = +1$ with probability $1/2$ independent of $\mathbf{x}$.

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

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

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

Proof. We let $h_t(\mathbf{x}) \stackrel{\mathrm{def}}{=} \mathrm{ReLU}(\langle \mathbf{s}, \mathbf{x}\rangle - t)$ and $r(t) = \mathbf{E}_{(\mathbf{x},y)\sim D}[yh_t(\mathbf{x})]$. 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

$$\begin{aligned}
r'(t) =& \frac{d\mathbf{E}_{(\mathbf{x},y)\sim D}[yh_t(\mathbf{x})]}{dt} \\
=& \frac{d\mathbf{E}_{(\mathbf{x},y)\sim D}[y(\langle \mathbf{s}, \mathbf{x}\rangle - t)\mathbf{1}(\langle \mathbf{s}, \mathbf{x}\rangle > t)]}{dt} \\
=& -\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = +1 \wedge \langle \mathbf{s}, \mathbf{x}\rangle > t] \\
&+ \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = -1 \wedge \langle \mathbf{s}, \mathbf{x}\rangle > t] \,,
\end{aligned}$$

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

$$\begin{aligned} r''(t) =& -\frac{d(\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=+1 \wedge \langle \mathbf{s},\mathbf{x}\rangle > t])}{dt} \\ &+ \frac{d(\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1 \wedge \langle \mathbf{s},\mathbf{x}\rangle > t])}{dt} \\ =& P_{\langle \mathbf{s},\mathbf{x}\rangle}(t)(\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1 \mid \langle \mathbf{s},\mathbf{x}\rangle = t] \\ &- \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=1 \mid \langle \mathbf{s},\mathbf{x}\rangle = t]) \\ =& P_{\langle \mathbf{s},\mathbf{x}\rangle}(t)\left(2\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y=-1 \mid \langle \mathbf{s},\mathbf{x}\rangle = t]-1\right) . \end{aligned}$$

Consider the interval $t \in [T/6, T/3]$. Note that $y=-1$ only if $\langle \mathbf{s},\mathbf{x}\rangle + z = t + z \notin [0,T/2]$. Thus, $y=-1$ only if $|z| \geq T/6$. Notice $z \sim D^{\mathcal{N}}_{\mathbb{R},\sigma}$ and $\sigma = o(T)$. Thus, for $t \in [T/6, T/3]$, we have that

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

where the last equality follows from $\sigma = o(T)$ and $P_{\langle \mathbf{s},\mathbf{x}\rangle}(t) = \Omega(1)$ since $\langle \mathbf{s},\mathbf{x}\rangle \sim D^{\mathcal{N}}_{\mathbb{R},1}$ 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

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

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

$$\begin{aligned} & r(T/6) - r(T/4) \\ =& r'(T/4)(-T/12) + \int_{T/4}^{T/6} r''(t)(T/6-t)dt \\ \leq& \int_{T/4}^{T/6} r''(t)(T/6-t)dt = -\Omega(T^2) . \end{aligned}$$

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. □

We will apply Lemma 4.3 on the joint distribution of $(\mathbf{x}, y)$ 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 $h(\mathbf{x}) = f(\langle \mathbf{s},\mathbf{x}\rangle) = \mathrm{ReLU}(\langle \mathbf{s},\mathbf{x}\rangle - t)$ such that $t \geq 0$ and $\left|\mathbf{E}_{(\mathbf{x},y)\sim D}[yh(\mathbf{x})]\right| = \left|\mathbf{E}_{(\mathbf{x},y)\sim D}[yf(\langle \mathbf{x},\mathbf{s}\rangle)]\right| = \Omega(T^2) = \Omega(1/(k\log n))$. If we apply Lemma 4.2 to the joint distribution of $(\langle \mathbf{x},\mathbf{s}\rangle, y)$ and the ReLU function f, we get that there must be a ReLU of the form $h'(\mathbf{x}) = kf(\langle \mathbf{x},\mathbf{s}\rangle) = \mathrm{ReLU}(\langle k\mathbf{s},\mathbf{x}\rangle - kt)$ such that $k < 1$ and

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

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

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

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

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 $c/(n\log n)^2 \leq \epsilon \leq 1/\log^{\gamma} n$ where c is a sufficiently large constant, there is no algorithm for ReLU regression on $\mathbb{R}^n$ under Gaussian marginals to error $R_2(\mathrm{ReLU}; D) + \epsilon$ 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 ϵ 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. □

Acknowledgements

Ilias Diakonikolas was 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. Daniel M. Kane was supported by NSF Medium Award CCF-2107547, NSF Award CCF-1553288 (CAREER), a Sloan Research Fellowship, and a grant from CasperLabs. Lisheng Ren was supported by NSF Award CCF-1652862 (CAREER) and a DARPA Learning with Less Labels (LwLL) grant.

References

Aggarwal, D., Li, J., Nguyen, P. Q., and Stephens-Davidowitz, N. Slide reduction, revisited - filling the gaps in SVP approximation. In *Advances in Cryptology - CRYPTO 2020 - 40th Annual International Cryptology Conference, CRYPTO 2020*, volume 12171 of *Lecture Notes in Computer Science*, pp. 274–295. Springer, 2020.

Awasthi, P., Balcan, M. F., and Long, P. M. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

Awasthi, P., Tang, A., and Vijayaraghavan, A. Agnostic learning of general relu activation using gradient descent. *CoRR*, abs/2208.02711, 2022.

Bruna, J., Regev, O., Song, M. J., and Tang, Y. Continuous LWE. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 694–707. ACM, 2021.

Dachman-Soled, D., Feldman, V., Tan, L., Wan, A., and Wimmer, K. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pp. 498–511. SIAM, 2015.

Daniely, A. A PTAS for agnostically learning halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pp. 484–502, 2015.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pp. 105–117, 2016.

Diakonikolas, I. and Kane, D. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4258–4282. PMLR, 2022.

Diakonikolas, I., Gopalan, P., Jaiswal, R., Servedio, R., and Viola, E. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010a.

Diakonikolas, I., Kane, D. M., and Nelson, J. Bounded independence fools degree-2 threshold functions. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010*, pp. 11–20. IEEE Computer Society, 2010b.

Diakonikolas, I., Kane, D. M., and Stewart, A. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pp. 73–84, 2017. Full version at http://arxiv.org/abs/1611.03473.

Diakonikolas, I., Kane, D., and Stewart, A. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pp. 1061–1073. ACM, 2018.

Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A. R., and Soltanolkotabi, M. Approximation schemes for ReLU regression. In *Conference on Learning Theory, COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1452–1485. PMLR, 2020a.

Diakonikolas, I., Kane, D., and Zarifis, N. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020b.

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. Agnostic proper learning of halfspaces under gaussian marginals. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1522–1551. PMLR, 2021a.

Diakonikolas, I., Kane, D. M., Pittas, T., and Zarifis, N. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1552–1584. PMLR, 2021b. URL http://proceedings.mlr.press/v134/diakonikolas21c.html.

Diakonikolas, I., Kane, D., Manurangsi, P., and Ren, L. Hardness of learning a single neuron with adversarial label noise. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8199–8213. PMLR, 2022a.

Diakonikolas, I., Kane, D. M., Manurangsi, P., and Ren, L. Cryptographic hardness of learning halfspaces with massart noise. *CoRR*, abs/2207.14266, 2022b. doi: 10.48550/arXiv.2207.14266. Conference version in NeurIPS'22.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning a single neuron with adversarial label noise via gradient descent. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4313–4361. PMLR, 2022c.

Feldman, V., Gopalan, P., Khot, S., and Ponnuswami, A. New results for learning noisy parities and halfspaces. In *Proc. FOCS*, pp. 563–576, 2006.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Freund, Y. and Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997.

Goel, S., Kanade, V., Klivans, A. R., and Thaler, J. Reliably learning the ReLU in polynomial time. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pp. 1004–1042, 2017.

Goel, S., Karmalkar, S., and Klivans, A. R. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

Goel, S., Gollakota, A., and Klivans, A. R. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Gupte, A., Vafa, N., and Vaikuntanathan, V. Continuous LWE is as hard as LWE & applications to learning Gaussian Mixtures. *arXiv preprint arXiv:2204.02550*, 2022. Conference version in FOCS'22.

Guruswami, V. and Raghavendra, P. Hardness of learning halfspaces with noise. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 543–552. IEEE Computer Society, 2006.

Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

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

Kearns, M., Schapire, R., and Sellie, L. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2/3):115–141, 1994.

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

Klivans, A. R. and Kothari, P. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014*, pp. 793–809, 2014.

Maass, W. and Turan, G. How fast can a threshold gate learn? In Hanson, S., Drastal, G., and Rivest, R. (eds.), *Computational Learning Theory and Natural Learning Systems*, pp. 381–414. MIT Press, 1994.

Manurangsi, P. and Reichman, D. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

Micciancio, D. On the hardness of learning with errors with binary secrets. *Theory Comput.*, 14(1):1–17, 2018.

Peikert, C. Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, 2009*, pp. 333–342. ACM, 2009.

Regev, O. On lattices, learning with errors, random linear codes, and cryptography. In *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 2005.

Rosenblatt, F. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

Soltanolkotabi, M. Learning ReLUs via gradient descent. In *Advances in neural information processing systems*, pp. 2007–2017, 2017.

Tiegel, S. Hardness of agnostically learning halfspaces from worst-case lattice problems. *CoRR*, abs/2207.14030, 2022.

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

Vapnik, V. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

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 $S_{n,k} \stackrel{\text{def}}{=} \{\mathbf{x} \in \{0, \pm 1\}^n : \|\mathbf{x}\|_1 = k\}$. We use $\mathrm{negl}(\lambda)$ 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 σ scale" to be the distribution of $D^{\mathcal{N}}_{T\mathbb{Z}+y,\sigma}$.

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 $D^{\text{expand}}_{\mathbb{R}^n_1,\sigma}$ denote the distribution of $\mathbf{x}'$ drawn as follows: first sample $\mathbf{x} \sim U(\mathbb{R}^n_1)$, and then sample $\mathbf{x}' \sim D^{\mathcal{N}}_{\mathbb{Z}^n+\mathbf{x},\sigma}$.

Definition A.3 (Collapsed Gaussian Distribution on $\mathbb{R}^n_1$)**.** For $\sigma \in \mathbb{R}_+$, we will use $D^{\text{collapse}}_{\mathbb{R}^n_1,\sigma}$ to denote the distribution of $\mathrm{mod}_1(\mathbf{x})$ on $\mathbb{R}^n_1$, where $\mathbf{x} \sim D^{\mathcal{N}}_{\mathbb{R}^n,\sigma}$.

B. Hardness of cLWE with Small-Norm Secret

Here we give the proof of Proposition 2.4, 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_{sample} is the discrete set $\mathbb{Z}^n_q$ — into a continuous LWE (cLWE) problem — where the support of D_{sample} is $\mathbb{R}^n$. This kind of cLWE problem was first introduced in (Bruna et al., 2021), where the paper gives a quantum reduction from approximating (the decision version of) the Shortest Vector Problem (GapSVP) to cLWE. Subsequently, (Gupte et al., 2022) 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 *sparse* discrete LWE instance whose secret vector $\mathbf{s}$ is sampled uniformly from $S_{n,k}$; after the reduction, we get a cLWE instance whose dimension is n and the ℓ_2-norm of the secret is roughly $\sqrt{k}$ ($\sqrt{k} \approx \log^{0.01} n$, compared with the $\sqrt{n}$ ℓ_2-norm secret vector in (Bruna et al., 2021)).

To achieve this, we slightly modify an idea from (Gupte et al., 2022) to get rid of the $\log m$ (where m is the number of samples) blowup in the ℓ_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 $\mathbf{s} \in S_{n,k}$).

Lemma B.1 (Corollary 4 in (Gupte et al., 2022))**.** *For any $n, m, q, l, \lambda, k \in \mathbb{N}$, $\sigma \in \mathbb{R}_+$ suppose that $\log(q)/2^l = \mathrm{negl}(\lambda)$, $\sigma \geq 4\sqrt{\omega(\log\lambda) + \ln n + \ln m}$ and $k\log_2(n/k) \geq (l+1)\log_2(q) + \omega(\log\lambda)$. Then, if the testing problem $\mathrm{LWE}(n, \mathbb{Z}^l_q, \mathbb{Z}^l_q, D^{\mathcal{N}}_{\mathbb{Z},\sigma}, \mathrm{mod}_q)$ has no $T + \mathrm{poly}(n, m, q, \lambda)$ time distinguisher with advantage ϵ, then the problem $\mathrm{LWE}(m, \mathbb{Z}^n_q, S_{n,k}, D^{\mathcal{N}}_{\mathbb{Z},\sigma'}, \mathrm{mod}_q)$ has no T-time distinguisher with advantage $2\epsilon m + \mathrm{negl}(\lambda)$, where $\sigma' = 2\sigma\sqrt{k+1}$.*

The above lemma reduces $\mathrm{LWE}(n, \mathbb{Z}^l_q, \mathbb{Z}^l_q, D^{\mathcal{N}}_{\mathbb{Z},\sigma}, \mathrm{mod}_q)$ to $\mathrm{LWE}(m, \mathbb{Z}^n_q, S_{n,k}, D^{\mathcal{N}}_{\mathbb{Z},\sigma'}, \mathrm{mod}_q)$. The λ here acts as a security parameter. Notice that the original problem $\mathrm{LWE}(n, \mathbb{Z}^l_q, \mathbb{Z}^l_q, D^{\mathcal{N}}_{\mathbb{Z},\sigma}, \mathrm{mod}_q)$ has $2^{l\log q}$ possible choices of secret vector, while the new problem $\mathrm{LWE}(m, \mathbb{Z}^n_q, S_{n,k}, D^{\mathcal{N}}_{\mathbb{Z},\sigma'}, \mathrm{mod}_q)$ has roughly at least $2^{k\log_2(n/k)}$ 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 $D^{\mathcal{N}}_{\mathbb{Z},\sigma}$ to a continuous Gaussian $D^{\mathcal{N}}_{\mathbb{R},\sigma'}$ where σ' is going to be slightly larger than σ. This leads to the following lemma:

Lemma B.2 (Lemma 15 in (Gupte et al., 2022)). *Let $n, m, q, \lambda \in \mathbb{N}$, $\sigma \in \mathbb{R}_+$, $\epsilon \in (0, 1]$ and suppose $\sigma > \sqrt{4 \ln m + \omega(\log \lambda)}$. For any $S \subseteq \mathbb{R}^n$, suppose there is no $T + \mathrm{poly}(m, n, \log(q), \log(\sigma))$-time distinguisher for the problem $\mathrm{LWE}(m, \mathbb{Z}_q^n, S, D^{\mathcal{N}}_{\mathbb{Z},\sigma}, \mathrm{mod}_q)$ with advantage ϵ. Then there is no T-time distinguisher for the problem $\mathrm{LWE}(m, \mathbb{Z}_q^n, S, D^{\mathcal{N}}_{\mathbb{R},\sigma'}, \mathrm{mod}_q)$ with advantage $\epsilon + \mathrm{negl}(\lambda)$, 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 B.2, $\sigma > \sqrt{4 \ln m + \omega(\log \lambda)}$ 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_{sample} from $U(\mathbb{Z}_q^n)$ to $U(\mathbb{R}_q^n)$. We thus obtain the following:

Lemma B.3 (Lemma 16 in (Gupte et al., 2022)). *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 ℓ_2-norm r, and suppose that $\sigma \geq 3r\sqrt{\ln n + \ln m + \omega(\log \lambda)}$. Suppose there is no $T + \mathrm{poly}(m, n, \log(q), \log(\sigma))$-time distinguisher for $\mathrm{LWE}(m, \mathbb{Z}_q^n, S, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_q)$ with advantage ϵ, then there is no T-time distinguisher for the problem $\mathrm{LWE}(m, \mathbb{R}_q^n, S, D^{\mathcal{N}}_{\mathbb{R},\sigma'}, \mathrm{mod}_q)$ with advantage $\epsilon + \mathrm{negl}(\lambda)$, 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 (Gupte et al., 2022). We first need to introduce the following fact from (Diakonikolas et al., 2022b).

Fact B.4 (Fact A.4 in (Diakonikolas et al., 2022b)). *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*

$$\frac{P_{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\sigma}/\sigma}(\mathbf{t})}{P_{D^{\mathcal{N}}_{\mathbb{R}^n,1}}(\mathbf{t})} = \frac{P_{U(\mathbb{R}_1^n)}(\mathrm{mod}_1(\sigma\mathbf{t}))}{P_{D^{\mathrm{collapse}}_{\mathbb{R}_1^n,\sigma}}(\mathrm{mod}_1(\sigma\mathbf{t}))} = 1 \pm O(\epsilon) \ ,$$

for all $\mathbf{t} \in \mathbb{R}^n$, and

$$d_{\mathrm{TV}}\left(\frac{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\sigma}}{\sigma}, D^{\mathcal{N}}_{\mathbb{R}^n,1}\right), d_{\mathrm{TV}}\left(D^{\mathrm{collapse}}_{\mathbb{R}_1^n,\sigma}, U(\mathbb{R}_1^n)\right) = \exp\left(-\Omega(\sigma^2)\right) \ .$$

Essentially, Fact B.4 says that, given $\mathbf{x} \sim D^{\mathcal{N}}_{\mathbb{R}^n,\sigma}$, the distribution of $\mathrm{mod}_1(\mathbf{x})$ is pointwise close (for its pdf function) to $U(\mathbb{R}_1^n)$ for sufficiently large σ. So if we consider the reverse of this process, given a $\mathbf{v} \sim U(\mathbb{R}_1^n)$, we sample $\mathbf{u} \sim D^{\mathcal{N}}_{\mathbb{R}^n+\mathbf{v},\sigma}$, then the distribution of $\mathbf{u}$ is sufficiently close to $D^{\mathcal{N}}_{\mathbb{R}^n,\sigma}$. We can leverage this fact to change the sample distribution in the LWE problem from $U(\mathbb{R}_q^n)$ to $D^{\mathcal{N}}_{\mathbb{R}^n,1}$ since $U(\mathbb{R}_q^n)$ is basically $U(\mathbb{R}_1^n)$ after rescaling. The difference here is that the original Lemma 18 takes a large σ so that $d_{\mathrm{TV}}\left(\frac{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\sigma}}{\sigma}, D^{\mathcal{N}}_{\mathbb{R}^n,1}\right) \approx 1/m$, thus m samples will not see the difference. However, since these two distributions are actually pointwise close, we can instead take a smaller σ 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 (Gupte et al., 2022).

Lemma B.5 (Modified Lemma 18 in (Gupte et al., 2022)). *Let $n, m, q \in \mathbb{N}, \sigma, r, \alpha \in \mathbb{R}_+$. Let $S \subseteq \mathbb{Z}^n$ where all elements in the support have fixed ℓ_2-norm r. Suppose there is no $T + \mathrm{poly}(n, m, \log(q))$-time distinguisher for the problem $\mathrm{LWE}(m, \mathbb{R}_q^n, S, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_q)$ with ϵ advantage. Then there is no T-time distinguisher for the problem $\mathrm{LWE}(m', D^{\mathcal{N}}_{\mathbb{R}^n,1}, S/r, D^{\mathcal{N}}_{\mathbb{R},\alpha\sigma/q}, \mathrm{mod}_\alpha)$ 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 $\mathrm{LWE}(m, \mathbb{R}_q^n, S, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_q)$, we can generate a sample $(\mathbf{x}', y')$ from the problem $\mathrm{LWE}(m', D^{\mathcal{N}}_{\mathbb{R}^n,1}, S/r, D^{\mathcal{N}}_{\mathbb{R},\alpha\sigma/q}, \mathrm{mod}_\alpha)$ with at least a constant success probability in the following manner.

We take a $\tilde{\sigma} = 1/r\alpha$ and sample $\tilde{\mathbf{x}} \sim D^{\mathcal{N}}_{\mathbb{Z}^n + \mathbf{x}/q, \tilde{\sigma}}/\tilde{\sigma}$. We define the function $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(\mathbf{t}) \stackrel{\mathrm{def}}{=} \frac{P_{D^{\mathcal{N}}_{\mathbb{R}^n,1}}(\mathbf{t})}{P_{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\tilde{\sigma}}/\tilde{\sigma}}(\mathbf{t})} \ .$$

With probability $f(\tilde{\mathbf{x}})/\max_{\mathbf{t}\in\mathbb{R}^n} f(\mathbf{t})$, we take $\mathbf{x}' = \tilde{\mathbf{x}}$ and $y' = y/(qr\tilde{\sigma})$ and output $(\mathbf{x}', y')$ as a sample for $\mathrm{LWE}(m', D^{\mathcal{N}}_{\mathbb{R}^n,1}, S/r, D^{\mathcal{N}}_{\mathbb{R},\alpha\sigma/q}, \mathrm{mod}_\alpha)$. Otherwise, we output failure.

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

$$\mathrm{mod}_q(\langle \mathbf{s}, \mathbf{x}\rangle) = \mathrm{mod}_q(\langle \mathbf{s}, q\tilde{\sigma}\tilde{\mathbf{x}}\rangle + \langle \mathbf{s}, \mathbf{x} - q\tilde{\sigma}\tilde{\mathbf{x}}\rangle) = \mathrm{mod}_q(\langle \mathbf{s}, q\tilde{\sigma}\tilde{\mathbf{x}}\rangle) \ .$$

Then we can write

$$\begin{aligned}
y' =& y/(qr\tilde{\sigma}) \\
=& \mathrm{mod}_q(\langle \mathbf{s}, \mathbf{x}\rangle + z)/(qr\tilde{\sigma}) \\
=& \mathrm{mod}_q(\langle \mathbf{s}, q\tilde{\sigma}\tilde{\mathbf{x}}\rangle + z)/(qr\tilde{\sigma}) \\
=& \mathrm{mod}_1(\langle \mathbf{s}, \tilde{\sigma}\tilde{\mathbf{x}}\rangle + z/q)/(r\tilde{\sigma}) \\
=& \mathrm{mod}_{1/(r\tilde{\sigma})}(\langle \mathbf{s}/r, \tilde{\mathbf{x}}\rangle + z/(qr\tilde{\sigma})) \\
=& \mathrm{mod}_\alpha(\langle \mathbf{s}/r, \mathbf{x}'\rangle + \alpha z/q) \ ,
\end{aligned}$$

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

It is immediate that $\mathbf{s}/r \sim U(S/r)$. 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 $\mathbf{x}'$, we first notice that $\mathbf{x}/q \sim \mathbb{R}_1^n$, and therefore $\tilde{\mathbf{x}} \sim D^{\mathrm{expand}}_{\mathbb{R}_1^n,\tilde{\sigma}}/\tilde{\sigma}$. Combining with the rejection sampling procedure we performed, we have that

$$P_{\mathbf{x}'}(\mathbf{u}) \propto \frac{f(\mathbf{u})}{\max_{\mathbf{t}\in\mathbb{R}^n} f(\mathbf{t})} P_{\tilde{\mathbf{x}}}(\mathbf{u}) = \frac{f(\mathbf{u})}{\max_{\mathbf{t}\in\mathbb{R}^n} f(\mathbf{t})} P_{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\tilde{\sigma}}/\tilde{\sigma}}(\mathbf{u}) = \frac{P_{D^{\mathcal{N}}_{\mathbb{R}^n,1}}(\mathbf{t})}{\max_{\mathbf{t}\in\mathbb{R}^n} f(\mathbf{t})} \propto P_{D^{\mathcal{N}}_{\mathbb{R}^n,1}}(\mathbf{u}) \ .$$

Thus, we conclude that $\mathbf{x}' \sim D^{\mathcal{N}}_{\mathbb{R}^n,1}$. For $\alpha z/q$, notice that $z \sim D^{\mathcal{N}}_{\mathbb{R},\sigma}$, and therefore $\alpha z/q \sim D^{\mathcal{N}}_{\mathbb{R},\alpha\sigma/q}$.

For the null hypothesis case, it is easy to see that the marginals satisfy $D_{\mathbf{x}'} = D^{\mathcal{N}}_{\mathbb{R}^n,1}$ and $D_{y'} = U(\mathbb{R}_\alpha)$, and $\mathbf{x}'$ and y' are independent — since $\mathbf{x}$ and y are independent and $\mathbf{x}'$ (resp. y') only depends on $\mathbf{x}$ (resp y).

It remains to verify that the sampling will produce at least m' many samples with $1 - 2^{-\Omega(m)}$ probability. We first show that each individual rejection sampling succeeds with at least a positive constant probability. From Fact B.4, given $\tilde{\sigma} = 1/r\alpha = \sqrt{\log n}/c$ for sufficiently small constant $c > 0$, we have

$$f(\mathbf{t}) = \frac{P_{D^{\mathcal{N}}_{\mathbb{R}^n,1}}(\mathbf{t})}{P_{D^{\mathrm{expand}}_{\mathbb{R}_1^n,\tilde{\sigma}}/\tilde{\sigma}}(\mathbf{t})} \in (1/2, 3/2) \ .$$

Notice that for any $\mathbf{x}$, we accept the sample with $f(\tilde{\mathbf{x}})/\max_{\mathbf{t}\in\mathbb{R}^n} f(\mathbf{t})$ probability, which is at least $1/3$ probability given the bound above. Then, by an application of the Chernoff bound, we have that the rejection sampling succeeds at least $m' = cm$ times with probability at least $1 - 2^{-\Omega(m)}$, where $c > 0$ is a sufficiently small constant. This completes the proof. □

We note that Lemma B.5 is stronger than Lemma 18 in (Gupte et al., 2022) in the sense that the original Lemma 18 has $\alpha = c/\left(r\sqrt{\log n + \log m + \omega(\log \lambda)}\right)$, compared with $\alpha = c/\left(r\sqrt{\log n}\right)$ 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 am $n^{\Omega(1/(\epsilon\sqrt{\log n})^{0.99})}$ lower bound — compared with the $n^{\Omega(1/(\epsilon\sqrt{\log n})^{1.99})}$ *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 $\mathrm{LWE}(2^{O(l^{\beta'})}, \mathbb{Z}_q^l, \mathbb{Z}_q^l, D^{\mathcal{N}}_{\mathbb{Z},\sigma'}, \mathrm{mod}_q)$ with $2^{-O(l^{\beta'})}$ advantage, where l is the dimension.
2. We then use Lemma B.1 to reduce to solving the problem $\mathrm{LWE}(n^{O(k^\beta)}, \mathbb{Z}_q^n, S_{n,k}, D^{\mathcal{N}}_{\mathbb{Z},c\sqrt{k}\sigma'}, \mathrm{mod}_q)$ 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 $S_{n,k}$.
3. The we apply 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 $\mathrm{LWE}(n^{O(k^\beta)}, \mathbb{R}_q^n, S_{n,k}, D^{\mathcal{N}}_{\mathbb{R},c\sqrt{k}\sigma'}, \mathrm{mod}_q)$ with $n^{-O(k^\beta)}$ advantage, where c is a sufficiently large positive universal constant.
4. To finish the reduction, we apply Lemma B.5 which mainly changes the sample distribution from $U(\mathbb{R}_q^n)$ to $D^{\mathcal{N}}_{\mathbb{R}^n,1}$ and reduce to solving the problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_\alpha)$ with $n^{-O(k^\beta)}$ advantage.

To start the reduction, we need to chose the values for parameters l, β', q, σ' 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 ψ be the value such that $k = \log^\psi n$ (ψ has dependence on n and k). We will chose the following values:

- $l = \log^t n$, 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
$$q = k^{\kappa+1} = \log^{\psi(\kappa+1)} n = l^{\psi(\kappa+1)/t} \le l^{\psi(\kappa+1)/(\psi(1-\delta))} = l^{(\kappa+1)/(1-\delta)} = l^{O(1)} .$$
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 $2^{l^{\beta'}} = n^{\omega(k^\beta)}$. Notice that
$$2^{l^{\beta'}} = 2^{\log^{t\beta'} n} = n^{\log^{t\beta'-1} n} = n^{k^{\frac{t\beta'-1}{\psi}}} .$$
Since $k = \log^\psi n$ and $k \ge \log^\gamma n$, it follows that $\psi \ge \gamma$; therefore, $\beta' = \frac{1+\gamma(1-2\delta)}{1+\gamma(1-\delta)} \ge \frac{1+\psi(1-2\delta)}{1+\psi(1-\delta)}$. Plugging this into the above, we get that
$$2^{l^{\beta'}} \ge n^{k^{\frac{t\frac{1+\psi(1-2\delta)}{1+\psi(1-\delta)}-1}{\psi}}} = n^{k^{1-2\delta}} = n^{\omega(k^\beta)} ,$$
where the last equality follows from the fact $\beta = 1 - 3\delta$. For the requirements, we have:

(a) It is immediate that $\log(q)/2^l = O(\log l)/2^l = \mathrm{negl}(\lambda)$ (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 $q = \mathrm{poly}(l)$ from step 1, it therefore suffices to show that $k\log n - k\log k = \omega(l\log l)$, which is $k\log n \geq cl\log l + k\log k$ for any constant c. We prove this by analyzing two cases, namely $cl\log l \leq k\log k$ and $cl\log l > k\log k$.

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 $l\log l = t\log^t n\log\log n$. Thus,

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

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

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

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

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

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 $2^{-O(l^{\beta'})}$ and advantage $2\epsilon m + \mathrm{negl}(\lambda)$ in Lemma B.1, where ϵ is the advantage before the reduction. Notice that since $2^{l^{\beta'}} = n^{\omega(k^\beta)}$, the time lower bound is at least any $n^{O(k^\beta)}$. For the advantage, by taking $\epsilon = 2^{-3(l^{\beta'})}$, we have that

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

where the last inequality and equality follows from the statement $2^{l^{\beta'}} = n^{\omega(k^\beta)}$ shown above. Thus, there is no $n^{O(k^\beta)}$-time distinguisher for solving $\mathrm{LWE}(n^{O(k^\beta)}, \mathbb{Z}_q^n, S_{n,k}, D^{\mathcal{N}}_{\mathbb{Z}, c\sqrt{k}\sigma'}, \mathrm{mod}_q)$ with $n^{-O(k^\beta)}$ advantage.

3. We then check the parameter requirements in Lemma B.2 and Lemma B.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 $n^{ck^\beta} - \mathrm{poly}(m, n, \log(q), \log(c\sqrt{k}\sigma'))$. Since $m = n^{O(k^\beta)}$, $\log(q) = \log(k^{\kappa+1}) = O(\log k)$, 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 $\mathrm{LWE}(n^{O(k^\beta)}, \mathbb{R}_q^n, S_{n,k}, D^{\mathcal{N}}_{\mathbb{R}, c\sqrt{k}\sigma'}, \mathrm{mod}_q)$ with $n^{-O(k^\beta)}$ advantage.

4. After applying Lemma B.5, we get that there is no $n^{O(k^\beta)}$-time distinguisher for solving the problem $\mathrm{LWE}(m', D^{\mathcal{N}}_{\mathbb{R}^n,1}, S_{n,k}/\sqrt{k}, D^{\mathcal{N}}_{\mathbb{R},c\alpha\sqrt{k}\sigma'/q}, \mathrm{mod}_\alpha)$ with $n^{-O(k^\beta)}$ advantage, where $\alpha = c/\left(\sqrt{k\log n}\right)$, $m' = cn^{O(k^\beta)}$, and $c > 0$ is a sufficiently small universal constant. We just need to check that it matches the values of σ, m, T in the body of Proposition 2.4. For the noise scale σ, we have

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

where the last inequality follows from $k = \log^\psi n$. For the number of samples, we have that $m' = cn^{c'(k^\beta)}$ which is any $n^{O(k^\beta)}$ by choosing c' to be sufficiently large. For the parameter T, we have that $\alpha = c/\left(\sqrt{k\log n}\right) = T$. Then, the only remaining difference is that the secret vector distribution is $U(S_{n,k}/\sqrt{k})$ instead of $U(\mathbb{S}^{n-1})$. 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 $U(\mathbb{S}^{n-1})$ distribution we want. Therefore, there is no $n^{O(k^\beta)}$-time distinguisher for solving the problem $\mathrm{LWE}(n^{O(k^\beta)}, D^{\mathcal{N}}_{\mathbb{R}^n,1}, \mathbb{S}^{n-1}, D^{\mathcal{N}}_{\mathbb{R},\sigma}, \mathrm{mod}_T)$ with $n^{-O(k^\beta)}$ advantage.

This proves Proposition 2.4.

□