Agnostic Learning of Arbitrary ReLU Activation under Gaussian Marginals

Anxin Guo ANXINBGUO@GMAIL.COM
Northwestern University

Aravindan Vijayaraghavan ARAVINDV@NORTHWESTERN.EDU
Northwestern University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider the problem of learning an arbitrarily-biased ReLU activation (or neuron) over Gaussian marginals with the squared loss objective. Despite the ReLU neuron being the basic building block of modern neural networks, we still do not understand the basic algorithmic question of whether an arbitrary ReLU neuron is learnable in the non-realizable setting. In particular, all existing polynomial time algorithms only provide approximation guarantees for the better-behaved unbiased setting or restricted bias setting.

Our main result is a polynomial time statistical query (SQ) algorithm that gives the first constant factor approximation for arbitrary bias. It outputs a ReLU activation that achieves a loss of $O(\mathrm{OPT})+\varepsilon$ in time $\mathrm{poly}(d, 1/\varepsilon)$, where OPT is the loss obtained by the optimal ReLU activation. Our algorithm presents an interesting departure from existing algorithms, which are all based on gradient descent and thus fall within the class of correlational statistical query (CSQ) algorithms. We complement our algorithmic result by showing that no polynomial time CSQ algorithm can achieve a constant factor approximation. Together, these results shed light on the intrinsic limitation of gradient descent, while identifying arguably the simplest setting (a single neuron) where there is a separation between SQ and CSQ algorithms.

Keywords: agnostic learning, general ReLU activation, CSQ-SQ separation

1. Introduction

The Rectified Linear Unit (ReLU) is the predominant choice of activation function in machine learning. A ReLU neuron has two parameters – a vector $w \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$, and acts on an input $x \in \mathbb{R}^d$ as

$$\sigma(\langle x, w\rangle + b), \text{ where } \forall z \in \mathbb{R}, \sigma(z) = \max\{z, 0\}. \tag{1}$$

Given a distribution $\mathcal{D}$ over samples $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, the loss incurred by a ReLU function with parameters $w \in \mathbb{R}^d, b \in \mathbb{R}$ on distribution $\mathcal{D}$ is given by the least squares error

$$L(w, b) := \frac{1}{2} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\big(y - \sigma(\langle x, w\rangle + b)\big)^2\right], \text{ and } \mathrm{OPT} = \min_{\substack{w\in\mathbb{R}^d, b\in\mathbb{R}:\\ \|w\|_2 \le W}} L(w, b) \tag{2}$$

is the error of the best fit ReLU function for the given data distribution $\mathcal{D}$, and W is an upper bound on the norm of weight vector which one can think as being a large polynomial in d. The interesting setting when $\mathrm{OPT} > 0$ is called the agnostic, or non-realizable setting of problem (Kearns et al., 1994). The goal in agnostic learning is to design an algorithm that takes in as input i.i.d. samples from the distribution $\mathcal{D}$ and outputs a ReLU neuron whose loss is competitive with best fit error OPT, e.g., with loss $\mathrm{OPT} + \varepsilon$ or $O(\mathrm{OPT}) + \varepsilon$.

The problem of agnostic learning of a ReLU activation, also called ReLU regression, has been studied extensively over the past decade. This problem is computationally intractable without additional assumptions on the marginal distribution of x. Most algorithmic works study the setting where the marginal distribution on x is a standard Gaussian, or make similar distributional assumptions (Diakonikolas et al., 2020a; Frei et al., 2020; Awasthi et al., 2023).

Despite the ReLU neuron being the basic building block of modern neural networks, we still do not understand the basic algorithmic question of whether an arbitrary ReLU neuron is learnable in the non-realizable setting. In particular, we are not aware of any polynomial time algorithm that achieves a non-trivial multiplicative approximation for the best-fit ReLU of arbitrary bias.[1] The main question we address is:

Can we design a polynomial time algorithm that learns an arbitrary ReLU activation under Gaussian marginals that achieves approximately optimal loss of $O(\mathrm{OPT})$, where OPT is defined in (2)?

Over the past decade there have been several algorithmic results in the unbiased setting i.e., when the bias $b = 0$ (Goel and Klivans, 2019; Frei et al., 2020; Diakonikolas et al., 2020a, 2022c; Gollakota et al., 2023; Wang et al., 2023; Zarifis et al., 2024; Wang et al., 2024). In particular, the algorithm of Diakonikolas et al. (2020a) gives the first efficient $O(\mathrm{OPT})$ guarantee by minimizing a convex surrogate loss. On the other hand, there is evidence that $\mathrm{OPT} + \varepsilon$ may be computationally intractable (Diakonikolas et al., 2021b, 2023c). To the best of our knowledge, the only prior algorithmic result on learning negatively-biased ReLU neuron is the recent work of Awasthi et al. (2023), which handles the moderate bias setting when $|b| = O(1)$. In fact, we are aware of few algorithmic results for agnostic learning *any linear model* in the arbitrary bias setting. (One notable exception is the line of work initiated by Diakonikolas et al. (2018) that gives an $O(\mathrm{OPT})$ agnostic learning guarantee for general linear halfspaces.) The arbitrary bias setting seems to represent a common challenge across many problems in computational learning theory. We refer to Appendix A for other related works, including those related to single-index models.

Our main result gives an affirmative answer to the above question.

Theorem 1 (SQ algorithm that gets $O(\mathrm{OPT})$, informal) *There exists a constant α, such that for all $W > 0$ the following holds. Let $\mathcal{D}$ be a the joint distribution of $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, where the x-marginal is $\mathcal{N}(0, I_d)$. Algorithm 1 uses $\mathrm{poly}(d, \frac{1}{\varepsilon}, \frac{1}{\delta}, W)$ time and samples from $\mathcal{D}$, and with probability at least $1 - \delta$, outputs parameters $\hat{w} \in \mathbb{R}^n, \hat{b} \in \mathbb{R}$, such that:*

$$L(\hat{w}, \hat{b}) \leq \alpha \cdot \inf_{\substack{w \in \mathbb{R}^d, b \in \mathbb{R}: \\ \|w\|_2 \leq W}} L(w, b) + \varepsilon.$$

[1] The existing state-of-the-art algorithms incur approximation factors that depend polynomially on $1/\mathrm{OPT}$ when bias is very negative (Awasthi et al., 2023).

The above theorem gives the desired agnostic learning guarantee in polynomial time for an arbitrary ReLU activation. In addition, it is a proper learner[2] i.e., it outputs a ReLU activation function that achieves this loss. The algorithm fits into the Statistical Query (SQ) framework (Kearns, 1993), which gives oracle access to statistics $\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x,y)]$ for some user-specified function f, up to an error tolerance of τ (typically, $\tau = 1/\mathrm{poly}(d)$ for polynomial sample complexity). Our algorithm consists of two main steps – an initial phase that finds a coarse initalizer w_0 to the true vector v, and then an iterative procedure based on a reweighted version of projected gradient descent to get the desired error of $O(\mathrm{OPT})$. We further note that the time/sample complexity dependence on W is standard, since the upper bound on the weight's norm controls the "scale" of our problem.

Our algorithm presents an interesting departure from existing approaches for agnostic learning with zero or restricted bias, which uses gradient descent (Frei et al., 2020; Awasthi et al., 2023; Diakonikolas et al., 2022c) and other algorithms that fit in the framework of Correlation SQ (CSQ), where the algorithm is only allowed to query values of the form $\mathbb{E}_{(x,y)\sim\mathcal{D}}[yf(x)]$, for some function f, up to a tolerance τ. In fact, we can prove the following strong lower bound for all CSQ algorithms.

Theorem 2 (CSQ lower bound of $\omega(\mathrm{OPT})$) *There exists a function $F(\varepsilon)$ that goes to infinity as $\varepsilon \to 0$, such that for any $\varepsilon > 0$ and any constant $\alpha \geq 1$, there exists a family of distributions with $\mathrm{OPT} \leq \varepsilon/\alpha$, under which any CSQ algorithm that can agnostically learn an arbitrary ReLU neuron with loss at most $\alpha \cdot \mathrm{OPT} + \varepsilon$ (as defined in Eq. (2)) must use either $2^{d^{\Omega(1)}}$ queries or queries of tolerance $d^{-F(\varepsilon)}$.*

In other words, no efficient CSQ algorithm can achieve error of $O(\mathrm{OPT})$, since doing so requires either $\exp(d^{\Omega(1)})$ queries (exponential time) or $d^{-F(\varepsilon)} = d^{\omega(1)}$ tolerance (superpolynomial samples) in the CSQ framework. In particular, since many variants of gradient descent (GD) under the L^2 loss are captured by the CSQ model, this points to the sub-optimality of GD for learning even a single neuron, and motivates the design of new hybrid algorithms for learning neural networks.

Theorems 1 and 2 together identify a new problem on which there is a separation between SQ and CSQ algorithms. Chen et al. (2021) gave the first such separation by designing an SQ algorithm for PAC learning (the realizable setting) that is fixed parameter tractable[3] for learning a neural network with k neurons under Gaussian marginals, where superpolynomial $d^{\Omega(k)}$ lower bounds for CSQ algorithms were known (Diakonikolas et al., 2020c). Such a separation has also been identified for learning sparse polynomials (Kiani et al., 2024; Andoni et al., 2014) and planted multi or single-index models (Damian et al., 2022, 2024), both under Gaussian marginals. Our two-phase algorithm is also inspired by Chen and Meka (2020) on learning certain multi-index models using a two-step non-CSQ algorithms. However, these works are all in the realizable setting ($\mathrm{OPT} = 0$). The agnostic setting that we study in this work introduces different challenges that require new algorithmic ideas that we describe in more detail in Section 3. Our problem of agnostic learning of a single ReLU neuron is arguably the simplest setting, and the first agnostic setting where such separation has been identified.

Subsequent work: In very recent independent work, Zarifis et al. (2025) gave a different algorithm that applies a smoothing operator for learning neurons, and extended the $O(OPT)$ approximation guarantee to more general monotone Lipschitz activations. However, they do not prove the CSQ lower bound and SQ-CSQ separation that we establish in the biased ReLU setting.

[2] We allow our algorithm to output $\mathbf{0}$, by the simple argument that $\mathbf{0}$ is the L^2-limit of $\sigma(\langle x, w\rangle + b)$, as $b \to -\infty$.

[3] i.e. in time $f(k)\mathrm{poly}(d)$, where $f(k)$ can grow very fast with k, but $\mathrm{poly}(d)$ is independent of k.

2. Warm-up and preliminaries

In this section, we first introduce some auxiliary results that simplify our problem. Then, we present intuitions for the unique challenges in our regime, and why previous methods fail.

2.1. Preliminaries

Let $L(w, b) = \frac{1}{2}\mathbb{E}[(y - \sigma(\langle x, w\rangle + b))^2]$ be the loss of the ReLU neuron with weight $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. In our proofs, we often compare our loss $L(w, b)$ to $L(v, b)$, where v is the unit vector that minimizes this loss for fixed b. This is a proxy of OPT in the "normalized" problem defined below.

We use φ and Φ to denote the standard Gaussian pdf and cdf respectively, and $\int f(z)\, d\Phi(z)$ denotes the integral of f with respect to (1-dimensional) standard Gaussian measure. For any vector $w \in \mathbb{R}^d$ we use $\|w\|$ to denote its L^2 norm.

Normalizing target vector v. Suppose the best-fit ReLU for $\mathcal{D}$ is some $\sigma(\langle x, v\rangle + b)$ with $\|v\| \neq 1$. Then, we can instead work with the normalized problem $(x, \hat{y}) \sim \hat{\mathcal{D}}$, where $\hat{y} = y/\|v\|$. Let $\hat{L}$ be the loss function under $\hat{\mathcal{D}}$. We prove in Appendix C the following claim: if we *know* the value $\|v\|$, then scaling by $1/\|v\|$ reduces the problem to unit vector case, with at most polynomial runtime overhead.

Proposition 3 *In the above setting, the optimal ReLU for $\hat{\mathcal{D}}$ is $\sigma(\langle x, \hat{v}\rangle + \hat{b})$, where $\|\hat{v}\| = v/\|v\|$ is a unit vector and $\hat{b} = b/\|v\|$, and it has loss $\widehat{\mathrm{OPT}} = \hat{L}(\hat{v}, \hat{b}) = \mathrm{OPT}/\|v\|^2$. Moreover, suppose parameters w, b_w incur loss $\hat{L}(w, b_w) \leq \alpha \cdot \widehat{\mathrm{OPT}} + \varepsilon$ on the normalized problem $\hat{\mathcal{D}}$. Then, the pair $(\|v\|w, \|v\|b_w)$ would incur loss $L(\|v\|w, \|v\|b_w) \leq \alpha \cdot \mathrm{OPT} + \|v\|^2\varepsilon$ on the original problem $\mathcal{D}$.*

"Guessing" $\|v\|$ and b. To apply appropriate scaling, we must have some knowledge about $\|v\|$. To this end, we apply grid search over various guesses (β, γ) of the values $(\|v\|, b)$, where v, b are the parameters of the optimal ReLU over $\mathcal{D}$. (See Algorithm 1.) This approach is effective as long as the approximation factor w.r.t. an "estimated optimal ReLU" is at most a constant multiple of that w.r.t. the true $\mathrm{OPT} = L(v, b)$. We formalize this claim in the next proposition, also proven in Appendix C:

Proposition 4 *Let $\hat{v}$ be the unit vector in v's direction. Let $\delta_v = \beta - \|v\|$ and $\delta_b = \gamma - b$ denote the additive errors of our parameter estimations. Then if $|\delta_v|, |\delta_b| \leq 0.1\sqrt{\varepsilon}$, we have:*

$$\alpha \cdot L(\beta\hat{v}, \gamma) \leq O(\alpha) \cdot L(v, b) + O(\varepsilon).$$

The time/sample complexity of this grid search will be analyzed in Theorem 32 in the appendix. For now, we will assume the problem is normalized and the optimal ReLU has $\|v\| = 1$, and b is a known value. This suffices to bound the loss $L(w, b)$, as long as $L(v, b)$ is a good proxy of OPT.

2.2. Challenges with significant negative bias

In all previous works where b is not too negative, the optimal ReLU $\sigma(\langle x, v\rangle + b)$ is *linear* on a considerable portion of inputs. Intuitively, this shape should be easier to learn than a small wedge at the far end of the number line, which happens when b is very negative. Our results corroborate this intuition: we can apply previous algorithms when b is bounded below, but as $b \to -\infty$, the CSQ hardness (Theorem 2) takes effect and a specialized algorithm for this limit is needed.

Structural observations. The $b \to -\infty$ regime comes with a benefit: we can now apply asymptotic analysis and make claims that hold *when the optimal ReLU has sufficiently negative* b. Most importantly, we have the following two lemmas, which are proven in Appendix C:

Lemma 5 *Suppose* $\alpha \geq 3$. *Let* $\sigma(\langle x, v\rangle + b)$ *be the optimal ReLU with loss* OPT, *where* $\|v\| = 1$ *and* b *sufficiently negative. If the zero function incurs loss at least* $\alpha \cdot \mathrm{OPT} + \varepsilon$, *then we have*

$$\mathrm{OPT} < \frac{3\Phi(b)}{\alpha b^2}, \textit{ and } \varepsilon < \frac{3\Phi(b)}{b^2}.$$

The intuition behind the two lemmas is the following: if the optimal ReLU has very negative b, then the zero function should be a reasonable approximation. Thus, any non-trivial setting (where $\mathbf{0}$ is not good enough) would have OPT and ε both be extremely small.

An Analysis of (projected) gradient descent. Our new algorithm is motivated by the need to overcome the failure of the following gradient descent (GD) algorithm, as $b \to -\infty$:

$$(w_{t+1}, b_{t+1}) \leftarrow (w_t, b_t) - \eta \nabla_{w,b} L(w_t, b_t), \text{ for } t = 1, \ldots, T.$$

Consider the change in *direction* of w_{t+1}, compared to that of w_t. At iteration t, w_t updates in the following direction:

$$-\nabla_w L(w_t, b_t) = \mathbb{E}[(y - \sigma(\langle x, w_t\rangle + b)) \cdot \mathbb{1}\{\langle x, w_t\rangle \geq -b_t\} \cdot x],$$

where the indicator is due to the derivative of the ReLU function: $\sigma'(z) = \mathbb{1}\{z \geq 0\}$.

Let $v^\perp$ be the component of v that is perpendicular to w_t.[4] Observe that, for w_t's direction to approach that of v, the update $-\nabla_w L(w_t, b_t)$ must have a *positive component* on $v^\perp$:

$$\begin{aligned} \big\langle -\nabla_w L(w_t, b_t), v^\perp \big\rangle &= \mathbb{E}\big[\big(y - \sigma(\langle x, w_t\rangle + b_t)\big) \cdot \mathbb{1}\{\langle x, w_t\rangle \geq |b_t|\} \cdot \langle x, v^\perp\rangle\big] \\ &= \mathbb{E}[y \cdot \mathbb{1}\{\langle x, w_t\rangle \geq |b_t|\} \cdot \langle x, v^\perp\rangle] > 0. \end{aligned} \tag{3}$$

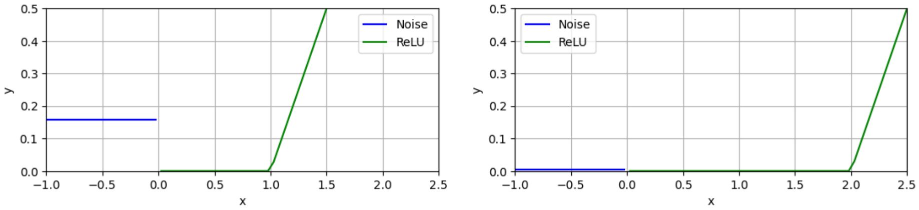
This can be viewed as a *conditioned* correlation between y and $\langle x, v^\perp\rangle$, where we call $\mathbb{1}\{\langle x, w_t\rangle \geq |b_t|\}$ the *condition function*. We can then write $y = \sigma(\langle x, v\rangle + b) - (\sigma(\langle x, v\rangle + b) - y)$, where the first term comes from ReLU and the second from noise. Through these lens, GD makes progress when ReLU (conditionally) correlates with $\langle x, v^\perp\rangle$ more than noise does.

One-dimensional adversarial example. We now give a simple one-dimensional example to show how the correlation between $\langle x, v^\perp\rangle$ and y can point to the *opposite direction*, when b is sufficiently negative. Fix an α, let $\mathrm{OPT} = \Phi(b)/100\alpha b^2$ such that $\mathbf{0}$ is not an α-approximation. Consider the following distribution which has loss exactly OPT, illustrated in Fig. 1:

$$y = \begin{cases} \sigma(x + b) & \text{when } x \geq 0, \\ \sqrt{2\mathrm{OPT}} & \text{when } x < 0. \end{cases} \tag{4}$$

Two seemingly contradictory facts are true about this distribution:

[4] When w_t is a unit vector, we have $v^\perp = \frac{v - \langle v, w_t\rangle w_t}{\|v - \langle v, w_t\rangle w_t\|}$.



(*a*) Moderate bias.

(*b*) Larger bias.

Figure 1: One-dimensional bad example in Eq. (4). The noise (blue) decreases exponentially in $|b|$.

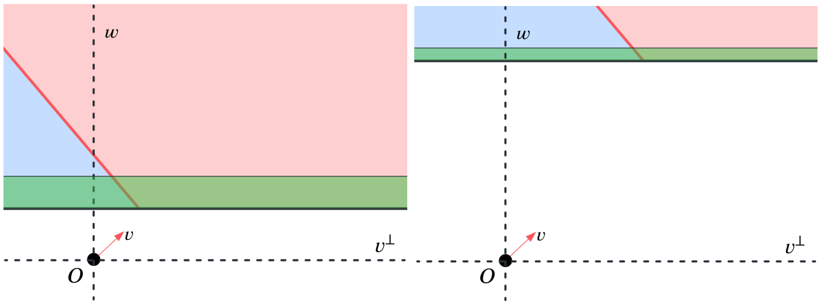
1. The ReLU produces more *loss* than noise: the zero function incurs about $\frac{\Phi(b)}{b^2} \approx 100\alpha \cdot \mathrm{OPT}$ loss when $x \geq 0$, by not fitting the ReLU.
2. The noise dominates the *correlation* with x, pulling $\mathbb{E}[xy]$ towards the negative direction:

$$\begin{cases} \mathbb{E}[\sigma(x+b) \cdot x] & = \Phi(b), \\ \mathbb{E}[(y - \sigma(x+b)) \cdot x] & = -\Omega(\sqrt{\mathrm{OPT}}) = -\tilde{\Omega}\left(\sqrt{\Phi(b)}\right). \end{cases}$$

An intuitive explanation is that the loss scales *quadratically* with y, yet the correlation scales (roughly) linearly. This issue appears when y is very close to zero, possible only as $b \to -\infty$.

Higher dimensions. Returning to high-dimensional ReLU regression, we demonstrate how the aforementioned adversarial example can occur in a way that harms GD, per the analysis in Eq. (3).

Suppose we have some estimated weight w_t and bias b_t. Consider the plane spanned by w_t and $v^\perp$ in Fig. 2. By removing the uncolored region through the condition function, the direction of correlation in Eq. (3) essentially depends on the thin green strip of width $\frac{1}{|b_t|^{0.99}}$, due to Gaussian decay.



(*a*) Moderate bias.

(*b*) Larger bias.

Figure 2: Plan view of the w_t-$v^\perp$ plane in high dimensional analysis of GD. The ReLU is positive on the red region. $1 - o(1)$ probability mass of the colored region falls in the green strip.

We thus identify a reason for the failure of GD : the red region is too biased in the green strip, so ReLU may fail to dominate the correlation (Eq. (3)), and GD may update in the wrong direction.[5] We note that this is strong evidence that *any* GD variant would fail at this task, even those not captured by the CSQ model (Theorem 33).

[5] Once exception is when $b_t \leq \frac{b}{\langle w_t, v \rangle}$. But then b_t is too negative and other problems start to occur.

3. Algorithmic results

```
Input: Accuracy parameter $\varepsilon$; sample access to $\mathcal{D}$; norm upper bound $W$.
  1. Let $h_0 = \mathbf{0}$ be the zero function, and let $L_0 = \mathbb{E}_{\mathcal{D}}[y^2]$ be its loss.
  2. Run a linear regression algorithm with bias, in the domain $b \geq O\big(\sqrt{\log(1/\varepsilon)}\big)$, and obtain parameters $w_{\text{lin}} \in \mathbb{R}^d, b_{\text{lin}}$. Let $h_1$ be the ReLU neuron $x \mapsto \sigma(\langle x, w_{\text{lin}} \rangle + b_{\text{lin}})$.
  3. Run the GD algorithm from Awasthi et al. (2023) with accuracy $\varepsilon$, and obtain ReLU neuron $h_2$ with loss $L_2$. This is competitive against the best moderately-biased ReLU.
  4. Run grid search over parameter space $[0, W] \times [-\Theta\big(\sqrt{\log(1/\varepsilon)}\big), \Theta\big(\sqrt{\log(1/\varepsilon)}\big)]$ with accuracy $0.1\sqrt{\varepsilon}$, where each pair of parameters $\{(\beta_i, \gamma_i) : i \in 3, \ldots, N\}$ indicate "guesses" for $(\|v\|, b)$.
     For each pair $(\beta_i, \gamma_i)$, we apply the following two subroutines on the normalized distribution $\hat{\mathcal{D}}$ (Theorem 3), and obtain ReLU $h_i$ with loss
         $L_i \leq 0.1\alpha \cdot \mathrm{OPT}_i + 0.1\frac{\varepsilon}{\beta_i^2},$
     where $\mathrm{OPT}_i$ is the minimum loss among ReLUs $\sigma(\langle x, v \rangle + b)$ with $\|v\| = \beta_i$ and $b = \gamma_i$.
     (a) Thresholded PCA: draw $m = \mathrm{poly}(\log d, \log 1/\delta, 1/\varepsilon)$ samples $S = \{(x_i, y_i)\}_{i=1}^m$. Output $w_0$, the top eigenvector of the following matrix:
             $\mathbb{E}_{(x,y)\sim S}\left[xx^\top \cdot \mathbb{1}\left\{|y| \geq \frac{1}{|b|}\right\}\right].$
     (b) Reweighted PGD: an iterative GD-like algorithm with initialization $w_0$ and step size $\eta$ being a fixed polynomial of $\varepsilon$, for $T = \mathrm{poly}(1/\varepsilon)$ steps. On each iteration $t \in [T]$, we draw $.m = \mathrm{poly}(d, 1/\delta, 1/\varepsilon)$ samples and truncate their $\|x\|$ and $y$ (see Appendix D.6) if necessary, resulting in training samples $S = \{(x_i, y_i)\}_{i=1}^m$.
         Then, we update $w_t$ by $w_{t+1} \leftarrow \frac{w_t + \eta v_{\text{update}}}{\|w_t + \eta v_{\text{update}}\|}$, where
             $v_{\text{update}} = \big(I_d - w_t w_t^\top\big)\mathbb{E}_{(x,y)\sim S}\left[f_t(x) \cdot y \cdot x \cdot \mathbb{1}\left\{\langle x, w_t \rangle \geq \tilde{b}\right\}\right].$
         Here, $f_t(x) = \exp\left(\rho|b| \langle x, w_t \rangle - \frac{\rho^2 b^2}{2}\right)$, and $\tilde{b} = (\rho + \lambda - \lambda^2\rho)|b|$. Both choices will be explained and the parameters $\lambda, \rho$ specified during analysis in Section 3.1.
         We will set the final output $h_i$ to be the ReLU neuron $x \mapsto \sigma) \langle x, w_t \rangle + b)$ with the smallest loss, for some $t \in [T]$.
Output: $h_{\text{ind}}$, where $\text{ind} = \arg\min_i\{L_i : i \in [N]\}$.
```

Algorithm 1: Complete algorithm, assuming population access to $\mathcal{D}$.

In this section, we present our full algorithm (Algorithm 1), which in polynomial time outputs a ReLU neuron with at most $\alpha \cdot \mathrm{OPT} + \varepsilon$ loss, as defined in Eq. (2). We note that our algorithm fits in the SQ model, and for the sake of exposition, the description of Algorithm 1 assumes access to population expectations over $\mathcal{D}$. We account for finite-sample errors in the appendix.

Algorithm 1 is a two-stage process consisting of *thresholded principal component analysis (PCA)* and *reweighted projected gradient descent (PGD)*.[6] We will analyze reweighted PGD before thresholded PCA, as the former is intimately related to our observations above. Our algorithm is reminiscent of the two-stage algorithm used in Chen and Meka (2020) for polynomials in few relevant directions, but we note that we work under the more challenging agnostic setting, and we use a complete different iterative update algorithm.

All our formal lemmas and theorems are under the following assumption:

Assumption 6 *Let α be a sufficiently large constant. Let $\mathcal{D}$ be the distribution of (x, y) over $\mathbb{R}^d \times \mathbb{R}$, where the x-marginal is standard Gaussian. Let OPT be the loss of the best-fit ReLU $\sigma(\langle x, v\rangle + b)$ with $\|v\| = 1$, and we assume the zero function $\mathbf{0}$ has loss at least $\alpha \cdot \mathrm{OPT} + \varepsilon$.*

As described in Section 2, the assumption $\|v\| = 1$ is essentially without loss of generality. In the next two subsections we present main theorems for the two subroutines. The effectiveness of the complete Algorithm 1 is a natural consequence of the two and is proven in Appendix F.

3.1. Reweighted PGD

Recall the example in Fig. 2, that "correlation" between $\langle x, v^{\perp}\rangle$ and y can be dominated by noise instead of the optimal ReLU, hence GD updates in the wrong direction. To overcome this, when calculating the "gradient" on each iteration, we will slightly modify the *condition function* from $\mathbb{1}\{\langle x, w_t\rangle + b \geq 0\}$ to be $\mathbb{1}\{\langle x, w_t\rangle \geq \tilde{b}\}$ for some carefully chosen $\tilde{b}$. We also add a *weight function* $f_t(x)$ inside the expectation, defined as the change-of-measure function between two Gaussians. Specifically, it shifts the x-marginal to be centered around $\rho \cdot |b| \cdot w_t$, for some carefully chosen $\rho \in (0, 1)$. The exact expression of $f_t(x)$ and $\tilde{b}$ are specified in Algorithm 1. The parameters λ and ρ are constants to be determined during the analysis.

The insight behind reweighting[7] is the following: suppose w_t is close to v in direction, then shifting the Gaussian center along w_t would essentially *reduce the problem back to the moderate bias* case. This is clear from Fig. 2, where shifting the origin upwards in subfigure (b) would result in a similar landscape as subfigure (a). We note that this needs w_t to be mostly aligned with v in the first place, hence the necessity of thresholded PCA.

Below is our main theorem for reweighted PGD.

Theorem 7 *Suppose Assumption 6 holds with b sufficiently negative, and suppose the initialization w_0 has $\langle w_0, v\rangle \geq 0.9$. Then, reweighted PGD with $\rho = 0.5, \lambda = 0.9$, and η being a fixed polynomial of ε can with probability $1 - \delta$ output a unit vector w such that:*

$$L(\langle x, w\rangle + b) \leq \alpha \cdot \mathrm{OPT} + \varepsilon,$$

using $\mathrm{poly}(d, 1/\varepsilon, 1/\delta)$ time and samples per iteration, and in $\log(1/\varepsilon)$ iterations.

[6] This is also known as "Riemannian GD" in literature, as the unit sphere is a Riemannian manifold.

[7] Reweighting is conceptually similar to *rejection sampling*, a method used for learning general linear halfspaces in Diakonikolas et al. (2018, 2024). Compared to their approaches, we have to shift the Gaussian mean by a more carefully chosen amount as well as apply a conditioning indicator, due to the nature of our regression problem.

We prove that reweighting and conditioning allows the ReLU to overcome the noise and guide the algorithm to the correct direction. Contributions from ReLU and noise are measured by the vectors:

$$\begin{cases} v_{\text{relu}} &= \left(I_d - w_t w_t^\top\right)\mathbb{E}\left[f_t(x)\cdot x\cdot\sigma(\langle x,v\rangle+b)\cdot\mathbb{1}\{x_w\ge\tilde{b}\}\right], \text{ and} \\ v_{\text{noise}} &= \left(I_d - w_t w_t^\top\right)\mathbb{E}\left[f_t(x)\cdot x\cdot\big(y-\sigma(\langle x,v\rangle+b)\big)\cdot\mathbb{1}\{x_w\ge\tilde{b}\}\right]. \end{cases}$$

We also prove in Appendix D.3 the following, that closeness of direction implies small loss:

Lemma 8 *Suppose Assumption 6 holds. For any unit vector w, if $L(w,b) > \alpha\cdot\mathrm{OPT}+\varepsilon$, then for all sufficiently negative b:*

$$\|w-v\|^2 \ge \Omega\left(\frac{\alpha\cdot\mathrm{OPT}+\varepsilon}{\Phi(b)}\right).$$

3.1.1. CONTRIBUTION FROM THE OPTIMAL RELU

Let $x_w = \langle x, w\rangle$ and $x^\perp = \langle x, v^\perp\rangle$, which simplifies the notation in the following proof. We will focus on the plane spanned by x_w and $x_\perp$, since the other directions are all irrelevant to v_{relu}. In fact, by coordinate-wise independence of $\mathcal{N}(0, I_d)$, the direction of v_{relu} completely aligns with $v^\perp$.

Lemma 9 *Suppose Assumption 6 holds. Let $\rho,\lambda\in(0,1)$ be constants with $\rho\lambda\ge\frac{1}{2}$. If on iteration t we have $\langle w_t, v\rangle\ge\lambda$ and $L(w_t,b) > \alpha\cdot\mathrm{OPT}$, then for all sufficiently negative b, we have:*

$$\|v_{\text{relu}}\| = \langle v_{\text{relu}}, v_\perp\rangle = \Omega\big(\sqrt{\alpha\cdot\mathrm{OPT}+\varepsilon}\big)\cdot\frac{e^{-\frac{b^2}{2}\left((1-\lambda\rho)^2-\frac{1}{2}\right)}}{|b|^{3/2}}. \tag{5}$$

Proof The first equality follows from the coordinate-wise independence of standard Gaussian.

Note that $f_t(x)$ is really a function of x_w, so we write it as $f_t(x_w)$ in this proof. Using the coordinate system of w_t and $v^\perp$, we can write the desired expression as:

$$\begin{aligned} &\mathbb{E}\big[f_t(x)\cdot x_\perp\cdot\sigma(\langle x,v\rangle+b)\mathbb{1}\left\{x_w\ge(\rho+\lambda-\lambda^2\rho)|b|\right\}\big] \\ &= \int_{(\rho+\lambda-\lambda^2\rho)|b|}^{\infty} f_t(x_w)\int_{-\infty}^{\infty} x_\perp\cdot\sigma(\langle x,v\rangle+b)\,d\Phi(x_\perp)\,d\Phi(x_w) \\ &\ge \int_{(\rho+\lambda-\lambda^2\rho)|b|}^{\infty} f_t(x_w)\int_{0}^{\infty} x_\perp\cdot\sqrt{1-\langle w_t,v\rangle^2}\cdot\sigma\Big(x_\perp-\sqrt{1-\lambda^2}(1-\lambda\rho)|b|\Big)\,d\Phi(x_\perp)\,d\Phi(x_w) \\ &= \sqrt{1-\langle w_t,v\rangle^2}\int_{(\lambda-\lambda^2\rho)|b|}^{\infty}\Phi\Big(\sqrt{1-\lambda^2}(1-\lambda\rho)b\Big)\,d\Phi(x_w) \\ &= \sqrt{1-\langle w_t,v\rangle^2}\cdot\Phi\Big(\sqrt{1-\lambda^2}(1-\lambda\rho)b\Big)\cdot\Phi\big(\lambda(1-\lambda\rho)b\big) \\ &= \Omega(\|w-v\|)\cdot\frac{\varphi\Big(\sqrt{1-\lambda^2}(1-\lambda\rho)b\Big)\cdot\varphi\big(\lambda(1-\lambda\rho)b\big)}{\sqrt{1-\lambda^2}\lambda(1-\lambda\rho)^2b^2}. \end{aligned}$$

Here, the inequality is due to Theorem 22 in Appendix D, and the final estimation is by Mills' ratio in Appendix B. All other equalities are calculation of Gaussian integrals.

Using $\varphi(a) \cdot \varphi(b) = \Theta\big(e^{-\frac{a^2+b^2}{2}}\big)$, and plugging in Theorem 8, for any fixed λ, ρ we have:

$$\left\langle v_{\text{relu}}, v^{\perp} \right\rangle = \Omega(1) \cdot \sqrt{\frac{\alpha \cdot \mathrm{OPT} + \varepsilon}{\Phi(b)}} \cdot \frac{e^{-\frac{b^2}{2}(1-\lambda\rho)^2}}{b^2} = \Omega\big(\sqrt{\alpha \cdot \mathrm{OPT} + \varepsilon}\big) \cdot \frac{e^{-\frac{b^2}{2}\left((1-\lambda\rho)^2 - \frac{1}{2}\right)}}{|b|^{3/2}}$$

■

3.1.2. Bounding noise and showing progress

In Appendix D.4, we similarly upper bound the contribution from label noise $y - \sigma(\langle x, v \rangle + b)$:

Lemma 10 *Suppose Assumption 6 holds. Let $\lambda, \rho \in (0, 1)$ be constants. For all unit vector u such that $u \perp w$, and for all sufficiently negative b, we have:*

$$\langle v_{\text{noise}}, u \rangle = O(\sqrt{\mathrm{OPT}}) \cdot \frac{e^{-\frac{b^2}{2}\left(-\rho^2 + \frac{1}{2}(\lambda + \rho\lambda^2 - \rho)^2\right)}}{|b|^{1/2}}. \tag{6}$$

From these two lemmas, we can solve for the appropriate values of λ and ρ, which allows v_{relu} to dominate v_{update}, which indicates the direction of v. This proof is deferred to Appendix D.5.

Lemma 11 *Suppose Assumption 6 holds where b is sufficiently negative, and suppose $L(w_t, b) > \alpha \cdot \mathrm{OPT}$. Set $\lambda = 0.9$ and $\rho \in (0.3, 0.6)$. If $\langle w_t, v \rangle \geq \lambda$, the for all $u \perp w$ we have:*

$$\frac{\left|\left\langle v_{\text{noise}}, v^{\perp} \right\rangle\right|}{\left\langle v_{\text{relu}}, v^{\perp} \right\rangle} = e^{-\Omega(b^2)}.$$

Finally, we account for sampling error and finish the proof via the following lemma in Appendix D.6:

Lemma 12 *Suppose Assumption 6 holds where b is sufficiently negative, and suppose $L(w_t, b) > \alpha \cdot \mathrm{OPT} + \varepsilon$ at some iteration t. Then, after an iteration of reweighted PGD with $\lambda = 0.9, \rho = 0.5$, and $\eta = c_\eta \frac{\|w_t - v\|}{\|v_{\text{update}}\|}$ for some $c_\eta \leq 0.1$, then:*

$$\|w_{t+1} - v\|^2 \leq \big(1 - \Omega(c_\eta)\big)\|w_t - v\|^2.$$

A few extra comments about results in this subsection:

1. Eq. (9) does not have solution when $\lambda < \sqrt{\frac{1}{7} + \frac{2\sqrt{2}}{7}} \approx 0.74$, meaning that a warm start is necessary. This also agrees with the CSQ hardness, since reweighted PGD is a CSQ algorithm.

2. We also need $\rho > 0$ for the equation to be feasible, so reweighting is necessary.

3. It's not trivial how far we should shift the Gaussian mean along w. In particular, taking $\rho = 1$, then Eq. (9) is true only when $\lambda > 1$, an infeasible solution.

3.2. Thresholded PCA

The final piece of our algorithmic result is to give a warm start via *thresholding on y*. Our algorithm uses threshold $\tau = \frac{1}{|b|}$ and estimate direction v via the top eigenvector of the following matrix:

$$M = \mathbb{E}[xx^\top \mathbb{1}\{|y| \geq \tau\}].$$

The value $\tau = \Theta(1/|b|)$ is the smallest threshold that reduces the noise to the necessary level (per Theorem 14 and Theorem 16). We note that thresholding is related to *trimmed/filtered PCA* (Chen and Meka, 2020; Chen et al., 2021), which uses a different matrix and works under the realizable setting. In our agnostic setting, the top eigenvector does not necessarily align with v, hence thresholded PCA only gives a coarse estimation and requires a different analysis. In this section we will show:

Theorem 13 *Suppose Assumption 6 holds. If bias b satisfies $b \leq -\sqrt{\alpha/\log\alpha}$, then thresholded PCA with $\tau = \frac{1}{|b|}$ can with high probability find a unit vector w such that:*

$$\left|\langle w, v\rangle\right| \geq 1 - O\left(\frac{\log\alpha}{\alpha}\right),$$

using $\mathrm{poly}(\log d, 1/\varepsilon, \log 1/\delta)$ *time and samples.*

Now we will present the proof, with details deferred to Appendix E. Imagine an adversary trying to perturb the top eigenvector of M. They can only do this in two ways:

1. Generating noise in the (otherwise flat) region of the ReLU where $\langle x, v\rangle + b < 0$, increasing M's magnitude in some direction $u \perp v$;
2. Suppressing some of the y-value when x has a high v-component, so the magnitude of M along v decreases.

We use three lemmas to show that these actions have limited effects when $\tau = \Theta(1/|b|)$. For convenience, we define $M_0 = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[xx^\top \mathbb{1}\{|y| \geq \tau, \langle v, x\rangle + b < 0\}\right]$, and $M_1 = M - M_0$.

3.2.1. KEY LEMMAS

First, we show that the "flat" region (M_0) contributes very little magnitude in any direction:

Lemma 14 *Suppose Assumption 6 holds. For all sufficiently negative b, we have:*

$$\|M_0\|_{\mathrm{op}} = O\left(\frac{\log\alpha}{\alpha} b^2 \Phi(b)\right).$$

Proof Since the target ReLU is zero on the region $\{v^\top x + b < 0\}$, by Markov's inequality:

$$p := \mathbb{P}[|y| \geq \tau, \langle x, v\rangle + b < 0] \leq \frac{\mathbb{E}[y^2 \mathbb{1}\{|y| \geq \tau\}]}{\tau^2} \leq \frac{\mathrm{OPT}}{\tau^2}.$$

By Theorem 5, we have $\mathrm{OPT} = O\left(\frac{\Phi(b)}{\alpha b^2}\right)$. Plugging in the value of τ, this means $p = O\left(\frac{\Phi(b)}{\alpha}\right)$. For sufficiently negative b, p would be small enough for us to apply Fact 29. For all unit $u \in \mathbb{R}^d$,

$$\begin{aligned} u^\top M_0 u &= \mathbb{E}[\langle x, u\rangle^2 \mathbb{1}\{|y| \geq \tau, \langle x, w\rangle + b < 0\}] \\ &= O\Big(p \log\frac{1}{p}\Big) = O\left(\frac{\log\alpha}{\alpha} b^2 \Phi(b)\right). \end{aligned}$$

■

Then, by the coordinate-wise independence property of standard Gaussian, we prove the following.

Lemma 15 *For all sufficiently negative b, and for any unit vector $u \perp v$, we have $u^\top M_0 u \le \Phi(b)$.*

The last lemma shows that that the ReLU always has substantial contribution to M. The main observation is that, since OPT only has an $\frac{1}{\alpha}$-fraction of the ReLU's squared L^2 norm, with an OPT budget the adversary can only remove a small fraction of the ReLU's contribution from the calculation of M. The proofs of Theorem 15 and Theorem 16 can be found in Appendix E.

Lemma 16 *Suppose Assumption 6 holds. For all sufficiently negative b, $v^\top M_1 v = \Omega\big(b^2\Phi(b)\big)$.*

4. Lower Bound for CSQ algorithms

Our correlational statistical query (CSQ) hardness result follows an established template via a family $\mathcal{G}$ of functions with small *pairwise correlation* (Feldman et al., 2013). Our first step is to identify an appropriate familiy of functions $\mathcal{G}$. Essentially, identifying one function from $\mathcal{G}$ is hard under the CSQ model, and our goal is to reduce this task to agnostic learning a ReLU neuron. Notably, this proof only goes through for very negative b, as mentioned in Section 2. (More details in Appendix G.)

Let $\|\cdot\|_{\mathcal{N}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{N}}$ denote the L^2 norm and inner product with respect to the 1-dimensional standard Gaussian. Let H_k be the kth (unnormalized) Hermite polynomial. The key lemma for our proof is the following:

Lemma 17 *The following holds for all sufficiently small ε. Let $g_\varepsilon(x) = \sigma(x - b_\varepsilon)$, where b_ε is chosen so that $\|g_\varepsilon\|_{\mathcal{N}}^2 = 3\varepsilon$. Let integer $t_\varepsilon \in \mathbb{N}$ be:*

$$t_\varepsilon := \max\left\{ t \in \mathbb{N} : \sum_{k=0}^{t} \left\langle g_\varepsilon, \frac{H_k}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2 \le \frac{\varepsilon}{\alpha} \right\},$$

then, we have $t_\varepsilon \to \infty$ as $\varepsilon \to 0$.

In words, as $b \to -\infty$, we can remove more and more lower-order Hermite components from the function $\sigma(x-b)$, while still having a constant fraction of its L^2 norm preserved.

The fact that g_ε has *no correlation* with lower-order Hermite polynomials allows us to construct the following family of functions with small pairwise correlation:

Lemma 18 *Let $\tilde{g}_\varepsilon(x)$ be the functions $g_\varepsilon(x)$ with its first t_ε Hermite components removed. Let S be a set of vectors with $|\langle u, v\rangle| = O(d^{-\Omega(1)})$ for distinct $u, v \in S$, and $|S| = 2^{\Omega(d^c)}$. Consider the following family of functions:*

$$\mathcal{G} = \{ g_\varepsilon(\langle x, v\rangle) : v \in S \},$$

then $\mathcal{G}$ has low pairwise correlation. Specifically, for any distinct $u, v \in S$, we have:

$$\begin{cases} \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[g_\varepsilon(\langle x,u\rangle)\cdot g_\varepsilon(\langle x,v\rangle)] \le d^{-\Omega(t_\varepsilon)}\cdot 3\varepsilon, \\ \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[g_\varepsilon(\langle x,v\rangle)^2] = 3\varepsilon. \end{cases}$$

The low pairwise correlation and the large cardinality $|\mathcal{G}|$ translates into high CSQ dimension, which leads to our desired conclusion: any CSQ algorithm that learns $g_\varepsilon(\langle x, v\rangle) \in \mathcal{G}$ to a non-trivial squared loss (better than the zero function) would make $2^{d^{\Omega(1)}}$ queries[8], or queries of tolerance $d^{-\Omega(t_\varepsilon)}$.

Finally, an agnostic learner for ReLU with $\alpha \cdot \mathrm{OPT} + \varepsilon$ error can indeed learn any $g_\varepsilon(\langle x, v\rangle)$ up to nontrivial squared loss. Note that $\|g_\varepsilon\|_{\mathcal{N}}^2 = 3\varepsilon$: the zero function incurs squared loss of $\|\tilde{g}_{t_\varepsilon}\|_{\mathcal{N}}^2 \geq (3 - \frac{1}{\alpha})\varepsilon$, while the ReLU function $\sigma(\langle x, v\rangle + b_\varepsilon)$ has squared loss $\|g_\varepsilon - \tilde{g}_\varepsilon\|_{\mathcal{N}}^2 = \frac{\varepsilon}{\alpha}$.

Acknowledgments

We thank Alex Tang and Santosh Vempala for helpful discussions. Both authors were supported by the National Science Foundation under Grant Nos. ECCS-2216970 and CCF2154100, and also acknowledge the partial support of Adobe Research.

References

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In Chandra Chekuri, editor, *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 500–510. SIAM, 2014. doi: 10.1137/1.9781611973402.37. URL https://doi.org/10.1137/1.9781611973402.37.

Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general relu activations. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13485–13496, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/700fdb2ba62d4554dc268c65add4b16e-Abstract.html.

Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Agnostic learning of general relu activation using gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=EnrY5TOrbQ.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 1161–1227. PMLR, 2020. URL http://proceedings.mlr.press/v125/chen20a.html.

Sitan Chen, Adam R. Klivans, and Raghu Meka. Learning deep relu networks is fixed-parameter tractable. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 696–707. IEEE, 2021. doi: 10.1109/FOCS52979.2021.00073. URL https://doi.org/10.1109/FOCS52979.2021.00073.

[8] In this context, we can make correlational queries in the form $\mathbb{E}[yf(x)]$, where $x \sim \mathcal{N}(0, I_d)$ and y is labeled by some $g_\varepsilon(\langle x, v\rangle)$.

Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models (extended abstract). In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, page 1262. PMLR, 2024. URL https://proceedings.mlr.press/v247/damian24a.html.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 73–84. IEEE Computer Society, 2017. doi: 10.1109/FOCS.2017.16. URL https://doi.org/10.1109/FOCS.2017.16.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1061–1073. ACM, 2018. doi: 10.1145/3188745.3188754. URL https://doi.org/10.1145/3188745.3188754.

Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R. Klivans, and Mahdi Soltanolkotabi. Approximation schemes for relu regression. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020a. URL http://proceedings.mlr.press/v125/diakonikolas20b.html.

Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer relu networks. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 1514–1539. PMLR, 2020c. URL http://proceedings.mlr.press/v125/diakonikolas20d.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado,* *USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1522–1551. PMLR, 2021a. URL http://proceedings.mlr.press/v134/diakonikolas21b.html.

Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584. PMLR, 2021b. URL http://proceedings.mlr.press/v134/diakonikolas21c.html.

Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Hardness of learning a single neuron with adversarial label noise. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 8199–8213. PMLR, 2022a. URL https://proceedings.mlr.press/v151/diakonikolas22a.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 874–885. ACM, 2022b. doi: 10.1145/3519935.3519970. URL https://doi.org/10.1145/3519935.3519970.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning a single neuron with adversarial label noise via gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4313–4361. PMLR, 2022c. URL https://proceedings.mlr.press/v178/diakonikolas22c.html.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5118–5141. PMLR, 2022d. URL https://proceedings.mlr.press/v162/diakonikolas22b.html.

Ilias Diakonikolas, Jelena Diakonikolas, Daniel Kane, Puqian Wang, and Nikos Zarifis. Near-optimal bounds for learning gaussian halfspaces with random classification noise. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/44c150733f9c5b6f98cb0caad0c664c7-Abstract-Conference.html.

Ilias Diakonikolas, Jelena Diakonikolas, Daniel M. Kane, Puqian Wang, and Nikos Zarifis. Information-computation tradeoffs for learning margin halfspaces with random classification noise. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference* *on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 2211–2239. PMLR, 2023b. URL https://proceedings.mlr.press/v195/diakonikolas23a.html.

Ilias Diakonikolas, Daniel Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7922–7938. PMLR, 2023c. URL https://proceedings.mlr.press/v202/diakonikolas23b.html.

Ilias Diakonikolas, Daniel M. Kane, Sihan Liu, and Nikos Zarifis. Testable learning of general halfspaces with adversarial label noise. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 1308–1335. PMLR, 2024. URL https://proceedings.mlr.press/v247/diakonikolas24a.html.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 655–664. ACM, 2013. doi: 10.1145/2488608.2488692. URL https://doi.org/10.1145/2488608.2488692.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3a37abdeefe1dab1b30f7c5c7e581b93-Abstract.html.

Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499. PMLR, 2019.

Surbhi Goel, Sushrut Karmalkar, and Adam R. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8582–8591, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/067a26d87265ea39030f5bd82408ce7c-Abstract.html.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam R. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3587–3596. PMLR, 2020a. URL http://proceedings.mlr.press/v119/goel20a.html.

Surbhi Goel, Aravind Gollakota, and Adam R. Klivans. Statistical-query lower bounds via functional gradients. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/17257e81a344982579af1ae6415a7b8c-Abstract.html.

Aravind Gollakota, Parikshit Gopalan, Adam R. Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2f46ef5725a8eca24f7f24a17955ad1a-Abstract-Conference.html.

Sham M. Kakade, Adam Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 927–935, 2011a. URL https://proceedings.neurips.cc/paper/2011/hash/30bb3825e8f631cc6075c0f87bb4978c-Abstract.html.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011b.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*. Citeseer, 2009.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005), 23-25 October 2005, Pittsburgh, PA, USA, Proceedings*, pages 11–20. IEEE Computer Society, 2005. doi: 10.1109/SFCS.2005.13. URL https://doi.org/10.1109/SFCS.2005.13.

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and Amir Salman Avestimehr. Fitting relus via SGD and quantized SGD. In *IEEE International Symposium on Information Theory, ISIT 2019, Paris, France, July 7-12, 2019*, pages 2469–2473. IEEE, 2019. doi: 10.1109/ISIT.2019.8849667. URL https://doi.org/10.1109/ISIT.2019.8849667.

Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In S. Rao Kosaraju, David S. Johnson, and Alok Aggarwal, editors, *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 392–401. ACM, 1993. doi: 10.1145/167088.167200. URL https://doi.org/10.1145/167088.167200.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994. doi: 10.1007/BF00993468. URL https://doi.org/10.1007/BF00993468.

Bobak T. Kiani, Thien Le, Hannah Lawrence, Stefanie Jegelka, and Melanie Weber. On the hardness of learning under symmetries. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ARPrtuzAnQ.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5), October 2006. ISSN 0090-5364. doi: 10.1214/009053606000000786. URL http://dx.doi.org/10.1214/009053606000000786.

Christopher G Small. *Expansions and asymptotics for statistics*. Chapman and Hall/CRC, 2010.

Mahdi Soltanolkotabi. Learning relus via gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2007–2017, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3404–3413. PMLR, 2017. URL http://proceedings.mlr.press/v70/tian17a.html.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning a single neuron via sharpness. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 36541–36577. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23aq.html.

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, and Jelena Diakonikolas. Sample and computationally efficient robust learning of gaussian single-index models. *arXiv preprint arXiv:2411.05708*, 2024.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning single-index models via alignment sharpness. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AZ1tWCa9j3.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning monotone generalized linear models via data augmentation. *arXiv preprint arXiv:2502.08611*, 2025.

Appendix A. Related works

Here we give a more detailed survey of some of the most relevant results in algorithmic learning theory, relating to either *arbitrary bias* or *ReLU regression*.

Biased linear models. Several works explicitly considered learning halfspaces with arbitrary bias, or *general halfspaces*. Diakonikolas et al. (2018) considers learning general halfspaces under the more general *nasty noise* model. They achieve an optimal $O(\varepsilon)$ error rate, which translates into constant-factor approximation for agnostic learning. Later, Diakonikolas et al. (2022d) gave a faster constant-factor approximation in the agnostic setting. More recently, Diakonikolas et al. (2024) proposed an $\tilde{O}(\sqrt{\mathrm{OPT}}) + \varepsilon$ agnostic *tester-learner* for general halfspaces.

Going beyond the agnostic model, Diakonikolas et al. (2022b) gives algorithm and hardness results for learning general halfspace under *Massart noise* (Massart and Nédélec, 2006). Diakonikolas et al. (2023a) considers learning general halfspaces under the weaker *random classification noise (RCN)* model. Also on under RCN, Diakonikolas et al. (2023b) makes no anti-concentration assumption on x and therefore their results carry over to the biased case.

We note that the techniques for classification do not carry over to the regression case. For instance, in the agnostic setting of classification problems, the ground truth classifier is correct $1 - \mathrm{OPT}$ of the time. In regression problems, however, the y-value can be perturbed on all outcomes.

Single index models. Learning a ReLU neuron, in the realizable setting, is a special case of the *single index models* (SIMs) (Kakade et al., 2011a), where the y value depends on a single, unknown direction v. One key difference of SIM is that the joint distribution of $\langle x, v\rangle$ and y is often known, corresponding to the case where we know $\|v\|$ and b before-hand, and the task is to statistically estimate the direction v. The CSQ and SQ hardness of this problem are completely characterized by the information exponent (Damian et al., 2022) and the generative exponent (Damian et al., 2024), and a SQ-CSQ separation appears as the link function (activation function) becomes more involved.

Lately, several works (Gollakota et al., 2023; Zarifis et al., 2024; Wang et al., 2024) sought to extend the study of SIMs to agnostic case. Particularly, the recent independent work of Wang et al. (2024) also proposed a two-stage algorithm to give a constant approximation, for link functions that satisfy a set of assumptions. We note that the key difference between these regimes and ours is again due to our allowing for *arbitrary bias*, which breaks all usual assumptions for SIMs such as $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)^2] = 1$.

The task of realizably learning a single ReLU neuron, however, is indeed solved by earlier works on SIMs and isotonic regression (Kakade et al., 2011b; Kalai and Sastry, 2009). A number of other works (Tian, 2017; Soltanolkotabi, 2017; Kalan et al., 2019) also showed the effectiveness of various gradient methods in this relatively simple setting.

Agnostic learning of a single ReLU. While a single ReLU neuron is relatively easy to learn to error $\mathrm{OPT} + \varepsilon$ with the realizable assumption, one line of work (Goel et al., 2019; Diakonikolas et al., 2020b; Goel et al., 2020b; Diakonikolas et al., 2021b, 2023c) gave strong evidence that the same task takes quasi-polynomial time without this assumption. In other words, it is hard to substantially outperform polynomial regression (Kalai et al., 2005) on this goal, in the agnostic setting. On the other hand, there exists PTAS with error $(1 + \mu)\mathrm{OPT} + \varepsilon$ which runs in time $\mathrm{poly}(d, \varepsilon)\exp(1/\mu)$(Diakonikolas et al., 2020a, 2021a) for unbiased ReLUs.

Further relaxing the objective, a number of $\mathrm{poly}(d/\varepsilon)$-time algorithms are known. Goel et al. (2019) gave an $O(\mathrm{OPT}^{2/3}) + \varepsilon$ algorithm for unbiased ReLUs via reducing to learning *homogeneous halfspace*. Some work applied GD on the convex *matching loss* instead of the L^2 objective itself, achieving better time complexity under weaker distributional assumptions, for a class of activation functions including unbiased ReLU Diakonikolas et al. (2020a); Wang et al. (2023). Notably, we always need *some* distributional assumptions for these results, as Diakonikolas et al. (2022a) demonstrated a cryptographic hardness result for distribution-free constant-factor approximation.

Another line of work proposes to apply GD on the L^2 objective itself. For unbiased ReLU, Frei et al. (2020) proved an $O(\mathrm{OPT}^{2/3}) + \varepsilon$ loss guarantee. Later, Diakonikolas et al. (2022c) showed that GD actually achieves the best-possible $O(\mathrm{OPT}) + \varepsilon$ error, for unbiased ReLUs. Their method also extends for ReLUs with a known *positive* bias. The most relevant work to ours is that of Awasthi et al. (2023), where GD is shown to produce constant approximation even under moderately negative bias. However, as $b \to -\infty$, their approximation factor depends exponentially on b, which translates to a $(\mathrm{OPT}^{-\mathrm{poly}(\varepsilon)} + \varepsilon)$-approximation. All methods above are CSQ in nature, so they do not extend to our problem.

Appendix B. Mills ratio and Gaussian integrals

Let φ, Φ be the pdf and cdf of a one-dimensional Gaussian respectively. The *Mills ratio* at t, for $t > 0$, is defined as:

$$m(t) = \frac{1 - \Phi(t)}{\varphi(t)},$$

namely the ratio between Gaussian tail and Gaussian density. This ratio has the following asymptotic expansion “around infinity” (see e.g. Small (2010)):

$$m(t) \sim \frac{1}{t} - \frac{1}{t^3} + \frac{1 \cdot 3}{t^5} - \frac{1 \cdot 3 \cdot 5}{t^7} + \dots,$$

where the error of every partial sum is bounded by the absolute value of the next term.

We note that this series diverges for any fixed t, since the numerator grows like a factorial. On the other hand, at the limit $t \to \infty$, the more terms we have in a partial sum, the quicker the approximation error of this partial sum converges to zero.

Using Mills ratio, we now estimate the values of the following integrals:

- The following is often used in thresholded PCA: As $b \to -\infty$,

$$\begin{aligned}\int_{|b|}^{\infty} t^2 \, d\Phi(t) &= |b|\varphi(b) + \Phi(b) \\ &= \big(1 + o(1)\big) b^2 \Phi(b).\end{aligned}$$

- The next one is very useful in general, as it bounds the L^2 norm of a negatively biased ReLU nrueon: as $b \to -\infty$,

$$\begin{aligned}\int_{|b|}^{\infty} (t + b)^2 \, d\Phi(t) &= (b^2 + 1)\Phi(b) - |b|\varphi(b) \\ &= \frac{2 + o(1)}{b^2} \Phi(b).\end{aligned}$$

- The following does not need $b \to -\infty$ but is often used. For all b, we have:

  $$\int_{-b}^{\infty} t(t+b)\,d\Phi(t) = \Phi(b).$$

  We note that this is a special case of Theorem 39.

- This is used in proof for Theorem 16 with $b - \tau$ in place of the value b below. Let $h = \frac{c}{|b|}$ for some constant c, then:

  $$\int_{|b|+h}^{\infty} (t+b)^2 \varphi(t)\,dt = (1+o(1)) \cdot e^{-c} \cdot \left(1 + c + \frac{c^2}{2}\right) \cdot \int_{|b|}^{\infty} (t+b)^2 \varphi(t)\,dt.$$

  Using the asymptotic expansion of Mills ratio:

  $$\begin{aligned}
  \int_{|b|+h}^{\infty} (t+b)^2 \varphi(t)\,dt &= (b^2+1)\Phi\left(b-h\right) + (b+h)\varphi\left(b-h\right) \\
  &= \varphi\left(b-h\right)\left[(b^2+1)\left(\frac{1}{|b-h|} - \frac{1}{|b-h|^3} + \frac{3+o(1)}{|b-h|^5}\right) + \underbrace{(b+h)}_{\frac{b^2-h^2}{-|b-h|}}\right] \\
  &= \frac{\varphi(b-h)}{|b-h|}\left[(b^2+1)\left(1 - \frac{1}{(b-h)^2} + \frac{3+o(1)}{(b-h)^4}\right) - (b^2-h^2)\right] \\
  &= \Phi(b-h)\left[(b^2+1)\left(-\frac{1}{(b-h)^2} + \frac{3+o(1)}{(b-h)^4}\right) + 1 + h^2)\right] \\
  &= \Phi(b-h) \cdot \frac{-(b^2+1) + (1+h^2)(b-h)^2 + \frac{(3+o(1))(b^2+1)}{(b-h)^2}}{(b-h)^2} \\
  &= \Phi(b-h) \cdot \frac{-1 - 2bh + h^2 + b^2h^2 - 2bh^3 + h^4 + \frac{(3+o(1))(b^2+1)}{(b-h)^2}}{(b-h)^2}.
  \end{aligned}$$

  Plugging in $h = \frac{c}{|b|} = o(1)$, we have

  $$\begin{aligned}
  &\int_{|b|+h}^{\infty} (t+b)^2 \varphi(t)\,dt \\
  &\qquad = \Phi(b-h) \cdot \frac{-1 + 2c + h^2 + c^2 + 2ch^2 + h^4 + \frac{(3+o(1))(b^2+1)}{(b-h)^2}}{(b-h)^2} \\
  &\qquad = \Phi(b-h) \cdot \frac{c^2 + 2c + 2 + o(1)}{b^2} \\
  &\qquad = (1+o(1)) \cdot e^{-c} \cdot \left(1 + c + \frac{c^2}{2}\right) \cdot \int_{|b|}^{\infty} (t+b)^2 \varphi(t)\,dt.
  \end{aligned}$$

Appendix C. Omitted proofs from Section 2

Recall that $\hat{\mathcal{D}}$ is the distribution over $(x, \hat{y})$ where $\hat{y} = \frac{y}{\|v\|}$. Suppose the best-fitting ReLU for $\mathcal{D}$ is some $\sigma(\langle x, v\rangle + b)$ where v is not necessarily a unit vector.

Proposition 19 (Same as Theorem 3) *In the above setting, the optimal ReLU for $\hat{\mathcal{D}}$ is $\sigma(\langle x, \hat{v}\rangle + \hat{b})$ where $\|\hat{v}\| = v/\|v\|$ is a unit vector and $\hat{b} = b/\|v\|$, and it has loss $\widehat{\mathrm{OPT}} = \hat{L}(v, b) = \mathrm{OPT}/\|v\|^2$.*

Moreover, suppose parameters w, b_w incur loss $\hat{L}(w, b_w) \leq \alpha \cdot \widehat{\mathrm{OPT}} + \varepsilon$ on the normalized problem $\hat{\mathcal{D}}$. Then, the pair $(\|v\|w, \|v\|b_w)$ would incur loss $L(\|v\|w, \|v\|b_w) \leq \alpha \cdot \mathrm{OPT} + \|v\|^2\varepsilon$ on the original problem $\mathcal{D}$.

Proof The first fact follows from the fact that the ReLU function σ is homogeneous:

$$\sigma(\langle x, w\rangle + b_w) - y = \|v\| \cdot \left(\sigma\Big(\Big\langle x, \frac{w}{\|v\|}\Big\rangle + \frac{b_w}{\|v\|}\Big) - \hat{y}\right), \text{ for all choices of } w \in \mathbb{R}, b_w \in \mathbb{R}.$$

The second is a result of scaling everything by $1/\|v\|$, except the additive ε:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_{\mathcal{D}}[(\|v\|h(x) - y)^2] &= \|v\|^2 \cdot \frac{1}{2}\mathbb{E}[(h(x) - \hat{y})^2] \\ &\leq \|v\|^2\big(\frac{\alpha}{2} \cdot \mathbb{E}_{\hat{\mathcal{D}}}[(\sigma(\langle x, \hat{v}\rangle + \hat{b}) - \hat{y})^2] + \varepsilon\big) \\ &= \alpha \cdot \frac{1}{2}\mathbb{E}[(\sigma(\langle x, v\rangle + b) - y)^2] + \|v\|^2\varepsilon \\ &= \alpha \cdot \mathrm{OPT} + \|v\|^2\varepsilon.\end{aligned}$$

■

Proposition 20 (Same as Theorem 4) *Let $\hat{v}$ be the unit vector in v's direction. Let $\delta_v = \beta - \|v\|$ and $\delta_b = \gamma - b$ denote the additive errors of our parameter estimations. Then if $|\delta_v|, |\delta_b| \leq 0.1\sqrt{\varepsilon}$, we have:*

$$\alpha \cdot L(\beta\hat{v}, \gamma) \leq O(\alpha) \cdot L(v, b) + O(\varepsilon).$$

Proof We will repeatedly apply the elementary inequality $(a + b)^2 \leq 2a^2 + 2b^2$. First, we compare the loss of the estimated optimal ReLU with OPT:

$$\begin{aligned}&\mathbb{E}\big[\big(\sigma(\langle x, v + \delta_v\hat{v}\rangle + b + \delta_b) - y\big)^2\big] \\ &\qquad \leq 2\mathbb{E}\big[\big(\sigma(\langle x, v + \delta_v\hat{v}\rangle + b + \delta_b) - \sigma(\langle x, v\rangle + b)\big)^2\big] + 2\mathrm{OPT} \\ &\qquad \leq 4\mathbb{E}\big[\big(\sigma(\langle x, v + \delta_v\hat{v}\rangle + b + \delta_b) - \sigma(\langle x, v\rangle + b + \delta_b)\big)^2\big] \\ &\qquad\quad + 4\mathbb{E}\big[\big(\sigma(\langle x, v\rangle + b + \delta_b) - \sigma(\langle x, v\rangle + b)\big)^2\big] + 2\mathrm{OPT}.\end{aligned}$$

Since σ is 1-Lipschitz, we can remove it, cancel the terms before squaring, and get an upper bound:

$$\mathbb{E}\big[\big(\sigma(\langle x, v + \delta_v\hat{v}\rangle + b + \delta_b) - y\big)^2\big] \leq 4\mathbb{E}[\big(\langle x, \delta_v\hat{v}\rangle\big)^2] + 4\delta_b^2 + 2\mathrm{OPT}.$$

Note that the first term on the right is just $4\delta_v^2$. Taking $|\delta_v| = |\delta_b| \leq 0.1\sqrt{\varepsilon}$, and the proof is finished. ■

Lemma 21 (Same as Lemma 5) *Suppose $\alpha \geq 3$. Let $\sigma(\langle x, v\rangle + b)$ be the optimal ReLU with loss* OPT*, where $\|v\| = 1$ and b sufficiently negative. If the zero function incurs loss at least $\alpha \cdot \mathrm{OPT} + \varepsilon$, then we have*

$$\mathrm{OPT} < \frac{3\Phi(b)}{\alpha b^2}, \text{ and } \varepsilon < \frac{3\Phi(b)}{b^2}.$$

Proof

By assumption we know that $\frac{1}{2}\mathbb{E}[y^2] > \alpha\mathrm{OPT}$. Using Gaussian integrals and Mills ratio, we have:

$$\begin{aligned}\alpha \cdot \mathrm{OPT} + \varepsilon < \frac{1}{2}\mathbb{E}[y^2] &\leq \mathbb{E}\left[\big(y - \sigma(\langle x, v\rangle + b)\big)^2\right] + \mathbb{E}[\sigma(\langle x, v\rangle + b)^2] \\ &= 2\mathrm{OPT} + \big(2 + o(1)\big)\frac{\Phi(b)}{b^2}.\end{aligned}$$

Rearranging the terms:

$$\begin{aligned}\mathrm{OPT} + \frac{\varepsilon}{\alpha - 2} &< \frac{\big(2 + o(1)\big)\Phi(b)}{(\alpha - 2)b^2} \\ &= \frac{3\Phi(b)}{\alpha b^2}.\end{aligned}$$

Since $\alpha \geq 3$, both claims follow. ■

Appendix D. Omitted proofs from reweighted PGD(Section 3.1)

In this section we provide the missing details from Section 3.1. First, we introduce some notations used in this section. We will focus heavily on the plane spanned by vectors w_t and v. As in Section 3.1, we use $x_w = \langle x, w_t\rangle$ and $x_\perp = \langle x, v^\perp\rangle$, where:

$$v^\perp = \frac{v - \langle v, w_t\rangle w_t}{\|v - \langle v, w_t\rangle w_t\|}.$$

We note that every x on this plane can be written as $x = x_w \cdot w_t + x_\perp \cdot v^\perp$ by orthogonality.

D.1. Re-centering the Gaussian covariates

As noted in Section 2, GD modifies the direction of w_i by gearing it towards the direction of the *conditioned correlation* between x and y, $\mathbb{E}[xy \cdot \mathbb{1}\{\langle x, w_t\rangle \geq -b\}]$, to learn about the direction $v^\perp$. One reason for its failure is that the condition function is too coarse to have the following ideal properties:

1. It should attempt to ignore the regions where $\langle x, v\rangle \leq -b$, since those are where the optimal ReLU neuron is zero and not informative. Conversely, it would put more weight on regions with large $\langle x, v\rangle$.
2. It should do so using the current estimated direction w_i, without knowing the true direction v.

In the rest of this subsection we will describe our approach, which applies both a *weight function* $f_t(x)$ and a *condition function* $\mathbb{1}\{\langle x, w_t\rangle \geq \tilde{b}\}$, in order to achieve the objectives above. Instead of $\mathbb{E}[xy \cdot \mathbb{1}\{x_w \geq |b|\}]$, on each iteration we make a "gradient" update using the vector

$$v_{\text{update}} = \left(I_d - w_t w_t^\top\right)\mathbb{E}[f_t(x) \cdot x \cdot y \cdot \mathbb{1}\{x_w \geq \tilde{b}\}],$$

where the two functions are defined via parameters $\rho, \lambda \in (0,1)$, to be determined in **??**:

$$\begin{cases} f_t(x) = \exp\left(\rho|b|x_w - \frac{1}{2}\rho^2 b^2\right), \text{ and} \\ \tilde{b} = (\rho + \lambda - \lambda^2\rho)|b|. \end{cases}$$

Now we explain the choice of these values. Weight function $f_t(x)$ is the relative density that shifts the x-marginal from standard Gaussian $\mathcal{N}(0, I_d)$ to $\mathcal{N}(\rho|b|w_t, I_d)$. Parameter ρ controls the amount of shift, relative to b.

Parameter λ can be understood as a lower bound for $\langle w_t, v\rangle$. The condition function $\tilde{b} = (\rho + \lambda + \lambda^2\rho)|b|$ is chosen to minimize the influence of noise while maintaining most of contribution from ReLU. Specifically, if $\langle w_t, v\rangle = \lambda$, then the value $\tilde{b}$ is precisely the x_w-coordinate of the point in region $\{x : \langle x, v\rangle \geq |b|\}$ that is the *closest to the new Gaussian center* $\rho|b|w_t$. In other words, it cuts through the region where ReLU makes a positive contribution, in a way that at least half of the contribution from ReLU are preserved.

D.2. Contribution from optimal ReLU

The following is the technical lemma used in the proof of Theorem 9. In short, we lower-bound the contribution from ReLU on every "horizontal slice" (as in Fig. 2) of fixed x_w. Let $\Lambda = \langle w_t, v\rangle$.

Lemma 22 *Suppose $\rho, \lambda \in (0,1)$ satisfy $\rho\lambda > \frac{1}{2}$, and assume $b < 0$. If $\Lambda \geq \lambda$, then on event $E = \{x_w \geq (\lambda - \lambda^2\rho + \rho)|b|\}$, for all $x_\perp \geq 0$ we have:*

$$\begin{aligned} \sigma\big(\langle x_w w + x_\perp v^\perp, v\rangle + b\big) - &\sigma\big(\langle x_w w - x_\perp v^\perp, v\rangle + b\big) \\ &\geq \sqrt{1-\Lambda^2} \cdot \sigma\Big(x_\perp - \sqrt{1-\lambda^2}(1-\lambda\rho)|b|\Big). \end{aligned} \tag{7}$$

Proof Since $v = \Lambda w + \sqrt{1-\Lambda^2}v^\perp$, we have:

$$\begin{aligned} \sigma\big(\langle x_w w + x_\perp v^\perp, v\rangle + b\big) - &\sigma\big(\langle x_w w - x_\perp v^\perp, v\rangle + b\big) \\ &= \sigma\Big(\Lambda x_w + \sqrt{1-\Lambda^2}x_\perp - |b|\Big) - \sigma\Big(\Lambda x_w - \sqrt{1-\Lambda^2}x_\perp - |b|\Big) \\ &\geq \sigma\Big(\Lambda x_w + \sqrt{1-\Lambda^2}x_\perp - |b|\Big) - \sigma(\Lambda x_w - |b|) \end{aligned}$$

Case 1, when $\Lambda x_w < |b|$. In this case the second term is zero, and the first term is:

$$\begin{aligned} \sigma\Big(\Lambda x_w + \sqrt{1-\Lambda^2}x_\perp - |b|\Big) &\geq \sigma\Big(\sqrt{1-\Lambda^2}x_\perp + \Lambda(\lambda - \lambda^2\rho + \rho)|b| - |b|\Big) \\ &\geq \sigma\Big(\sqrt{1-\Lambda^2}x_\perp + \Lambda(\Lambda - \Lambda^2\rho + \rho)|b| - |b|\Big), \end{aligned}$$

since $\lambda - \lambda^2\rho + \rho$ has derivative $1 - 2\lambda\rho < 0$. We can thus further bound it by:

$$\begin{aligned}
\sigma\Big(\sqrt{1-\Lambda^2}x_\perp + \Lambda(\Lambda - \Lambda^2\rho + \rho)|b| - |b|\Big) &= \sigma\Big(\sqrt{1-\Lambda^2}x_\perp + (\Lambda^2 - \Lambda^3\rho + \Lambda\rho - 1)|b|\Big) \\
&= \sigma\Big(\sqrt{1-\Lambda^2}x_\perp - (1-\Lambda^2)(1-\Lambda\rho)|b|\Big) \\
&\geq \sigma\Big(\sqrt{1-\Lambda^2}x_\perp - \sqrt{1-\Lambda^2}\sqrt{1-\lambda^2}(1-\lambda\rho)|b|\Big) \\
&= \sqrt{1-\Lambda^2}\cdot\sigma\Big(x_\perp - \sqrt{1-\lambda^2}(1-\lambda\rho)|b|\Big),
\end{aligned}$$

as desired.

Case 2, when $\Lambda x_w < |b|$. In this case we have:

$$\begin{aligned}
\sigma\Big(\Lambda x_w + \sqrt{1-\Lambda^2}x_\perp - |b|\Big) - \sigma(\Lambda x_w - |b|) &= \Lambda x_w + \sqrt{1-\Lambda^2}x_\perp - |b| - (\Lambda x_w - |b|) \\
&= \sqrt{1+\Lambda x_\perp} \\
&\geq \sqrt{1-\Lambda^2}\cdot\sigma\Big(x_\perp - \sqrt{1-\lambda^2}(1-\lambda\rho)|b|\Big).
\end{aligned}$$

■

D.3. Proof for Lemma 8

Lemma 23 (same as Lemma 8) *Suppose Assumption 6 holds. For any unit vector w, if $L(w,b) > \alpha \cdot \mathrm{OPT} + \varepsilon$, then for all sufficiently negative b:*

$$\|w - v\|^2 \geq \Omega\left(\frac{\alpha\cdot\mathrm{OPT} + \varepsilon}{\Phi(b)}\right).$$

Proof Let $F(w,b)$ be the realizable loss $\mathbb{E}[(\sigma(\langle x,w\rangle + b) - \sigma(\langle x,v\rangle + b))^2]$. By an elementary inequality we have $L(w,b) \leq 2\mathrm{OPT} + F(w,b)$, which implies $F(w) > 0.4\alpha\cdot\mathrm{OPT} + \varepsilon$ for large α. Meanwhile, $F(w)$ is upper bounded by $\|w - v\|$ by the following:

$$\begin{aligned}
F(w,b) &= \mathbb{E}[(\sigma(\langle x,w\rangle + b) - \sigma(\langle x,v\rangle + b))^2] \\
&\leq \mathbb{E}[(\langle x,w\rangle - \langle x,v\rangle)^2 \mathbb{1}\left\{\langle x,w\rangle + b \geq 0 \text{ or } \langle x,v\rangle + b \geq 0\right\}] \\
&\leq 2\mathbb{E}[\langle x, w - v\rangle^2 \,\mathbb{1}\left\{\langle x,w\rangle \geq |b|\right\}] \\
&= 2\int_{|b|}^{\infty}\int_{-\infty}^{\infty}\left\langle x_w w + x_\perp v^\perp, w - v\right\rangle^2 d\Phi(x_\perp)\,d\Phi(x_w) \\
&= 2\int_{|b|}^{\infty}\int_{-\infty}^{\infty}\left(x_w\frac{\|w-v\|^2}{2} - x_\perp\frac{\|w-v\|\|w+v\|}{2}\right)^2 d\Phi(x_\perp)\,d\Phi(x_w) \\
&\leq \int_{|b|}^{\infty}\int_{-\infty}^{\infty}\left(x_w^2\|w-v\|^4 - x_\perp^2\|w-v\|^2\right) d\Phi(x_\perp)\,d\Phi(x_w) \\
&= \int_{|b|}^{\infty}\left(x_w^2\|w-v\|^4 + \|w-v\|^2\right) d\Phi(x_w) \\
&\leq \big(1.1\|w-v\|^4 b^2 + \|w-v\|^2\big)\Phi(b),
\end{aligned}$$

for sufficiently negative b.

If $\|w-v\| \leq \frac{1}{|b|}$, then $F(w,b) \leq 2.1\|w-v\|^2\Phi(b)$, and we have $\|w-v\| > \sqrt{\frac{0.4\alpha\cdot\mathrm{OPT}+\varepsilon}{2.1\Phi(b)}}$ as desired.

On the other hand, if $\|w-v\| > \frac{1}{|b|}$, then $\|w-v\|^2\Phi(b) \geq \frac{\Phi(b)}{b^2} \geq \frac{\alpha\cdot\mathrm{OPT}+\varepsilon}{6}$ by Theorem 5, and the proof is finished. ■

D.4. Proof of Lemma 10: contribution from noise

Lemma 24 (same as Lemma 10) *Suppose Assumption 6 holds. Let $\lambda, \rho \in (0,1)$ be constants. For all unit vector u such that $u \perp w$, and for all sufficiently negative b, we have:*

$$\langle v_{\text{noise}}, u\rangle = O(\sqrt{\mathrm{OPT}}) \cdot \frac{e^{-\frac{b^2}{2}\left(-\rho^2+\frac{1}{2}(\lambda+\rho\lambda^2-\rho)^2\right)}}{|b|^{1/2}}. \tag{8}$$

Proof Using Cauchy-Schwarz, we can bound $\langle v_{\text{noise}}, u\rangle$ by:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{D}}[f(x)\cdot\langle x,u\rangle\cdot(y-\sigma(\langle x,v\rangle+b))\mathbb{1}\{E\}] \\
&\qquad\leq \sqrt{\mathbb{E}[(y-\sigma(\langle x,v\rangle+b))^2]}\sqrt{\mathbb{E}[f(x)^2\mathbb{1}\{x_w \geq (\rho+\lambda-\lambda^2\rho)|b|\}]} \\
&\qquad= \sqrt{2\mathrm{OPT}}\cdot\sqrt{\mathbb{E}[f(x)^2\mathbb{1}\{x_w \geq (\rho+\lambda-\lambda^2\rho)|b|\}]}
\end{aligned}$$

The second term can be explicitly written as an integral:

$$\begin{aligned}
\mathbb{E}[f(x)^2\mathbb{1}\left\{x_w \geq (\rho+\lambda-\lambda^2\rho)|b|\right\}] &= \int_{(\rho+\lambda+\rho\lambda^2)|b|}^{\infty} \varphi(x_w)\cdot\exp(2\rho|b|x_w-\rho^2b^2)\,dx_w \\
&= \frac{1}{\sqrt{2\pi}}\int_{(\rho+\lambda+\rho\lambda^2)|b|}^{\infty}\exp\left(-\frac{1}{2}x_2^2+2\rho|b|x_w-2\rho^2b^2+\rho^2b^2\right)dx_w \\
&= \exp(\rho^2b^2)\cdot\mathop{\mathbb{P}}_{x_w\sim\mathcal{N}(2\rho|b|,I_d)}\left[x_w \geq (\lambda+\rho\lambda^2+\rho)|b|\right] \\
&= \exp(\rho^2b^2)\cdot\Phi\big((\lambda+\rho\lambda^2-\rho)b\big).
\end{aligned}$$

Plugging this back in, we have:

$$\begin{aligned}
\langle v_{\text{noise}}, u\rangle &\leq \sqrt{2\mathrm{OPT}}\cdot\sqrt{\exp(\rho^2b^2)\cdot\Phi\big((\lambda+\rho\lambda^2-\rho)b\big)} \\
&= O\big(\sqrt{\mathrm{OPT}}\big)\cdot\frac{1}{\varphi(\rho b)}\cdot\sqrt{\Phi\big((\lambda+\rho\lambda^2-\rho)b\big)} \\
&= O\big(\sqrt{\mathrm{OPT}}\big)\cdot\frac{1}{\varphi(\rho b)}\cdot\sqrt{\frac{\varphi\big((\lambda+\rho\lambda^2-\rho)b\big)}{(\lambda+\rho\lambda^2-\rho)|b|}} \\
&= O\big(\sqrt{\mathrm{OPT}}\big)\cdot\frac{e^{-\frac{b^2}{2}\left(-\rho^2+\frac{1}{2}(\lambda+\rho\lambda^2-\rho)^2\right)}}{|b|^{1/2}},
\end{aligned}$$

as desired. ■

D.5. Proof for Lemma 11

Now, we combine the previous lemmas to show that the direction of v_{update} is dominated by the contribution from ReLU, rather than noise. We will also determine the choice for parameters ρ and λ.

Lemma 25 (same as Lemma 11) *Suppose Assumption 6 holds where b is sufficiently negative, and suppose $L(w_t, b) > \alpha \cdot \mathrm{OPT}$. Set $\lambda = 0.9$ and $\rho \in (0.3, 0.6)$. If $\langle w_t, v\rangle \geq \lambda$, the for all $u \perp w$ we have:*

$$\frac{\left|\left\langle v_{\text{noise}}, v^{\perp}\right\rangle\right|}{\left\langle v_{\text{relu}}, v^{\perp}\right\rangle} = e^{-\Omega(b^2)}.$$

Proof To compare Eq. (5) and Eq. (8), we first consider the terms involving $e^{-\frac{1}{b^2}}$. To ensure $\left\langle v_{\text{noise}}, v^{\perp}\right\rangle$ dominates, we want it to have a *smaller* coefficient inside $\exp\left(-\frac{b^2}{2}\right)$, namely:

$$(1-\rho\lambda)^2 - \frac{1}{2} - \left(-\rho^2 + \frac{1}{2}(\lambda + \rho\lambda^2 - \rho)^2\right) \leq -\Omega(1). \tag{9}$$

This is true when $\lambda \geq 0.9$ and $\rho \in [0.3, 0.6]$, in which case the right hand side is less than -0.01. Consequently, for any unit vector u such that $u \perp w$,

$$\begin{aligned}
\frac{\left|\left\langle v_{\text{noise}}, u\right\rangle\right|}{\left\langle v_{\text{relu}}, v^{\perp}\right\rangle} &= \frac{O(1)\cdot e^{-\frac{b^2}{2}\left(-\rho^2+\frac{1}{2}(\lambda+\rho\lambda^2-\rho)^2\right)}/|b|^{1/2}}{\Omega(\sqrt{\alpha})\cdot e^{-\frac{b^2}{2}\left((1-\lambda\rho)^2-\frac{1}{2}\right)}/|b|^{3/2}} \\
&\leq \frac{O(1)\cdot e^{-\frac{0.16\cdot b^2}{2}}\cdot|b|}{\Omega(\sqrt{\alpha})} \\
&= e^{-\Omega(b^2)}.
\end{aligned}$$

■

D.6. Making progress on each update

Recall that our algorithm makes update by setting $w_{t+1} := \frac{w_t + \eta\hat{v}_{\text{update}}}{\|w_t + \eta\hat{v}_{\text{update}}\|}$, where $\hat{v}_{\text{update}}$ is the estimation of v_{update} from n new samples. In this subsection we will prove that, with appropriate values of η and n, w_t gets provably close to v on each iteration with high probability.

First we upper bound on magnitude $\|v_{\text{update}}\|$:

Lemma 26 *Suppose Assumption 6 holds with b sufficiently negative, and suppose $L(w_t, b) > \alpha \cdot \mathrm{OPT}$. For any $\rho, \lambda \in (0,1)$, if $\langle w_t, v\rangle \geq \lambda$, then:*

$$\|v_{\text{update}}\| = O(\|w_t - v\|)\cdot\Phi\big((1-\rho)b\big) \leq \mathrm{poly}(\varepsilon).$$

Proof

$$\begin{aligned}
\|v_{\text{update}}\| &= \max_{\text{unit } u: u\perp w_t} \langle v_{\text{update}}, u\rangle \\
&\leq \left\langle v_{\text{relu}}, v^{\perp}\right\rangle + \max_{\text{unit } u: u\perp w} \langle v_{\text{noise}}, u\rangle \\
&= \big(1 + o(1)\big)\left\langle v_{\text{relu}}, v^{\perp}\right\rangle,
\end{aligned}$$

where the small o is taken as $b \to -\infty$. It therefore suffices for us to upper bound $\left\langle v_{\text{relu}}, v^{\perp} \right\rangle$. We will take $f_t(x)$ into account by considering the shifted distribution: let $\mathcal{D}'$ be the modified distribution of (x, y), with the x-marginal being $\mathcal{N}(\rho|b|w_t, I_d)$. Then we have:

$$\begin{aligned}\left\langle v_{\text{relu}}, v^{\perp} \right\rangle &= \mathop{\mathbb{E}}_{\mathcal{D}}\Big[f_t(x) \cdot x_{\perp} \cdot \sigma(\langle x, v\rangle + b) \cdot \big(1 - \mathbb{1}\left\{\langle x, w_t\rangle < (\rho + \lambda - \lambda^2\rho)|b|\right\}\big)\Big] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}'}[x_{\perp} \cdot \sigma(\langle x, v\rangle + b)] - \mathop{\mathbb{E}}_{\mathcal{D}'}[x_{\perp} \cdot \sigma(\langle x, v\rangle + b)\mathbb{1}\left\{\langle x, w_t\rangle < (\rho + \lambda - \lambda^2\rho)|b|\right\}].\end{aligned}$$

Note that the second term is nonnegative: for any fixed x_w, the value of $\sigma(\langle x, v\rangle + b)$ always grows with $x_{\perp}$, hence they have positive correlation. It now suffices to upper bound the first term:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}'}[\langle x, v^{\perp}\rangle \cdot \sigma(\langle x, v\rangle + b)] &\leq \mathop{\mathbb{E}}_{\mathcal{D}'}[\langle x - \rho|b|w_t, v^{\perp}\rangle \cdot \sigma(\langle x - \rho|b|w_t, v\rangle - (1-\rho)|b|)] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\langle x, v^{\perp}\rangle \cdot \sigma(\langle x, v\rangle - (1-\rho)|b|)] \\ &= \langle v, v^{\perp}\rangle \Phi\big((1-\rho)|b|\big).\end{aligned}$$

The first inequality uses $w_t \perp v^{\perp}$ and $\langle w_t, v\rangle \leq 1$. The second inequality applies change of variable, taking x to be the previous $x - \rho|b|w_t$, since both random variables have distribution $\mathcal{N}(0, I_d)$. The proof is finished by noting $\langle v, v^{\perp}\rangle = \sqrt{1 - \langle w_t, v\rangle^2} = \Theta(\sqrt{1 - \langle w_t, v\rangle}) = \Theta(\|w_t - v\|)$. ■

Now we bound the number of fresh samples we need on each iteration of reweighted PGD. First, assuming y is bounded, we have the following sample complexity bound:

Lemma 27 *Suppose $|y| \leq B$ almost surely. For all sufficiently negative b, if $\hat{v}_{\text{update}}$ is calculated using $m = \text{poly}\left(d, \frac{1}{\delta}, \frac{1}{\varepsilon}, B\right)$ new samples, with probability $\geq 1 - \delta$ we have:*

$$\langle \hat{v}_{\text{update}}, v^{\perp}\rangle \geq 0.9\|\hat{v}_{\text{update}}\|,$$

on any iteration t.

Proof We can bound the LHS and RHS by:

$$\langle \hat{v}_{\text{update}}, v^{\perp}\rangle \geq \langle v_{\text{update}}, v^{\perp}\rangle - \|\hat{v}_{\text{update}} - v_{\text{update}}\|, \text{ and}$$

$$0.9\|\hat{v}_{\text{update}}\| \leq 0.9(\|v_{\text{update}}\| + \|\hat{v}_{\text{update}} - v_{\text{update}}\|).$$

Hence, it suffices to show that:

$$\langle v_{\text{update}}, v^{\perp}\rangle - \|\hat{v}_{\text{update}} - v_{\text{update}}\| \geq 0.9(\|v_{\text{update}}\| + \|\hat{v}_{\text{update}} - v_{\text{update}}\|).$$

By Lemma 11, we know that $\langle v_{\text{update}}, v^{\perp}\rangle \geq 0.95\|v_{\text{update}}\|$ as $b \to -\infty$, since v_{relu} dominates v_{noise}. Now we want to show that, for all sufficiently negative b:

$$0.05\|v_{\text{update}}, v^{\perp}\| \geq 1.9\|\hat{v}_{\text{update}} - v_{\text{update}}\|.$$

To lower bound the LHS we use Lemma 9, which states:

$$\begin{aligned}\|v_{\text{update}}\| &= (1+o(1))\|v_{\text{relu}}\| \\ &\geq \Omega\big(\sqrt{\alpha \cdot \text{OPT} + \varepsilon}\big) \cdot e^{-O(b^2)} \\ &= \text{poly}(\varepsilon) \text{ by Lemma 5.}\end{aligned}$$

Meanwhile, to upper bound the RHS, we will use the assumption that y is almost surely bounded, and apply multidimensional Chebyshev's inequality.

Let μ and Σ denote the mean and covariance matrix of the vector $f_t(x)\cdot x \cdot y \cdot \mathbb{1}\left\{x_w \geq (\rho+\lambda-\lambda^2\rho)|b|\right\}$. For any unit vector u, we have:

$$\begin{aligned}u^\top \Sigma u &\leq \mathbb{E}[f_t(x)^2 \langle x,u\rangle^2 y^2] \\ &\leq \big(\mathbb{E}[y^4]\big)^{1/2}\big(\mathbb{E}[f_t(x)^8]\big)^{1/4}\big(\mathbb{E}[\langle x,u\rangle^8]\big)^{1/4} \\ &= O(B^2)\cdot\left(\int_{-\infty}^{\infty}\exp(8\rho|b|x_w - 4\rho^2b^2)\varphi(x_w)\,dx\right)^{1/4} \\ &= O\big(B^2e^{3\rho^2b^2}\big),\end{aligned}$$

where $\rho \in (0.3, 0.6)$ as before. Suppose the empirical estimate $\hat{v}_{\text{update}}$ is calculated using m samples, then the random variable $\hat{v}_{\text{update}}$ has covariance $\frac{1}{m}\Sigma$. By multidimensional Chebyshev's inequality, for all $s > 0$ we have:

$$\begin{aligned}&\mathbb{P}\left[\sqrt{(\hat{v}_{\text{update}} - v_{\text{update}})^\top(\Sigma/m)^{-1}(\hat{v}_{\text{update}} - v_{\text{update}})} \geq s\right] \\ &\quad= \mathbb{P}\left[\|\hat{v}_{\text{update}} - v_{\text{update}}\| \geq \frac{sBe^{1.5\rho^2b^2}}{\sqrt{m}}\right] \leq \frac{d}{s^2},\end{aligned}$$

To make this at most δ, we take $s = \sqrt{\frac{d}{\delta}}$. The inequality now becomes:

$$\mathbb{P}\left[\|\hat{v}_{\text{update}} - v_{\text{update}}\| \geq \frac{\sqrt{d}Be^{1.5\rho^2b^2}}{\sqrt{m\delta}}\right] \leq \delta.$$

Therefore, to conclude, we must set m such that $\frac{\sqrt{d}sBe^{1.5\rho^2b^2}}{\sqrt{m\delta}}$ is at most some $\text{poly}(\varepsilon)$. A polynomial number of samples is sufficient:

$$m = \frac{O\left(\sqrt{d}s^2B^2e^{3\rho^2b^2}\right)}{\sqrt{\delta}\cdot\text{poly}(\varepsilon)} = \text{poly}\left(d, \frac{1}{\delta}, \frac{1}{\varepsilon}, B\right).$$

■

Now, if $\mathcal{D}$ has unbounded y, we can slightly modify the samples so that we can more efficiently sample the same desired direction. Let $B_x(d, m, \varepsilon, \delta)$ be the value with the following two properties:

1. With probability at least $1-\delta$, all m fresh samples $\{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{D}$ will have $\|x_i\| \le B_x(d, m, \varepsilon, \delta)$.
2. On the event $\|x_i\| \le B_x(d, m, \varepsilon, \delta)$, most of the value $\langle v_{\text{relu}}, v^\perp \rangle$ should be kept (c.f. Section 3.1):

$$\left|\mathbb{E}[f_t(x) \cdot x_\perp \cdot \sigma(\langle x, v\rangle + b) \cdot \mathbb{1}\{x_w \ge \tilde{b}, \|x\| > B_x(d, m, \varepsilon, \delta)\}]\right| \le 0.1\langle v_{\text{relu}}, v^\perp \rangle$$

To satisfy the first property, it suffices to take $B_x = O\left(\sqrt{d \log(m/\delta)}\right)$ by Gaussian concentration. The second property is satisfied by $B_x = \sqrt{d} \cdot \operatorname{polylog}(1/\varepsilon)$, again by Gaussian concentration, and the facts that $|b| = O\left(\sqrt{\log 1/\varepsilon}\right)$ and $\langle v_{\text{relu}}, v^\perp \rangle \le \operatorname{poly}(\varepsilon)$.

Let $B_y = |B_x| + O\left(\sqrt{\log(1/\varepsilon)}\right)$ be the maximum value of $\sigma(\langle x, v\rangle + b)$ on the event that $\|x\| \le B_x$. It follows that:

$$B_y(d, m, \varepsilon, \delta) = \operatorname{poly}(d, \log m, \log(1/\delta), 1/\varepsilon).$$

On the other hand, by Lemma 27, we know that assuming $|y| \le B_y$, the sample complexity is bounded by some $\operatorname{poly}(d, 1/\delta, 1/\varepsilon, B_y)$. Since B_y has at most poly-log growth in m, and m has at most polynomial growth in B_y, we can solve for an upper bound on B_y and m which are (at most) polynomial in the values $d, 1/\varepsilon$, and $1/\delta$.

Finally, it's easy to check that all the lemmas for $\mathcal{D}$ also hold for the new, truncated distribution, where we condition on the event that $\|x\| \le B_x$ and $|y| \le B_y$. Note that most of v_{relu} is kept, and v_{noise} becomes smaller as y' is now closer to the target ReLU.

Finally, we make provable progress on each iteration when η is set properly:

Lemma 28 (same as Lemma 12) *Suppose Assumption 6 holds where b is sufficiently negative, and suppose $L(w_t, b) > \alpha \cdot \mathrm{OPT} + \varepsilon$ at some iteration t. Then, after an iteration of reweighted PGD with $\lambda = 0.9, \rho = 0.5$, and $\eta = c_\eta \frac{\|w_t - v\|}{\|v_{\text{update}}\|}$ for some $c_\eta \le 0.1$, then:*

Proof

$$\begin{aligned}
\|w_t - v\|^2 - \|w_{t+1} - v\|^2 &= \frac{\langle w_{t+1}, v\rangle - \langle w_t, v\rangle}{2} \\
&= \frac{1}{2}\left[\left\langle \frac{w_t + \eta \hat{v}_{\text{update}}}{\|w_t + \eta \hat{v}_{\text{update}}\|}, v \right\rangle - \langle w_t, v\rangle\right] \\
&= \frac{\eta\langle \hat{v}_{\text{update}}, v\rangle + (1 - \|w_t + \eta \hat{v}_{\text{update}}\|)\langle w_t, v\rangle}{2\|w_t + \eta \hat{v}_{\text{update}}\|} \\
&\ge \frac{\eta\langle \hat{v}_{\text{update}}, v\rangle - \frac{\eta^2 \|\hat{v}_{\text{update}}\|^2}{2}\langle w_t, v\rangle}{2\|w_t + \eta \hat{v}_{\text{update}}\|}.
\end{aligned}$$

When b is sufficiently negative, by Theorem 27, with probability $\ge 1 - \delta$ we have $\langle \hat{v}_{\text{update}}, v^\perp \rangle \ge 0.9\|\hat{v}_{\text{update}}\|$. Therefore:

$$\langle \hat{v}_{\text{update}}, v\rangle = \sqrt{1 - \langle w_t, v\rangle^2}\langle \hat{v}_{\text{update}}, v^\perp\rangle \ge \|w_t - v\| \cdot 0.9\|\hat{v}_{\text{update}}\|$$

Suppose we take $\eta = c_\eta \frac{\|w_t - v\|}{\|\hat{v}_{\text{update}}\|}$ for any $c_\eta \leq 0.1$, then:

$$\begin{aligned}
\|w_t - v\|^2 - \|w_{t+1} - v\|^2 &\geq \eta \|\hat{v}_{\text{update}}\| \cdot \frac{0.9\|w_t - v\| - \frac{1}{2}\eta\|\hat{v}_{\text{update}}\|}{2\|w_t + \eta\hat{v}_{\text{update}}\|} \\
&= c_\eta \|w_t - v\| \cdot \frac{(0.9 - 0.5c_\eta)\|w_t - v\|}{O(1)} \\
&= \Omega(c_\eta \|w_t - v\|^2).
\end{aligned}$$

■

Note that $\|\hat{v}_{\text{update}}\| = \Theta(\|v_{\text{update}}\|)$, and the latter is bounded by:

$$\Omega\big(\sqrt{\alpha \cdot \mathrm{OPT} + \varepsilon}\big) \cdot e^{-O(b^2)} \leq \|v_{\text{update}}\| \leq O\Big(\|w_t - v\| \Phi\big((1-\rho)b\big)\Big).$$

Moreover, we have $\|w_t - v\| \geq \Omega\big(\sqrt{\alpha \cdot \mathrm{OPT} + \varepsilon}/\sqrt{\Phi(b)}\big)$. Therefore, we have $\frac{\|w_t - v\|}{\|\hat{v}_{\text{update}}\|} = e^{-\Theta(b^2)} = \mathrm{poly}(\varepsilon)$. Hence by setgin η to be some fixed polynomial in ε, we have:

$$c_\eta = \eta \frac{\|\hat{v}_{\text{update}}\|}{\|w_t - v\|} = \mathrm{poly}(\varepsilon) \leq -0.1.$$

We hence conclude that $T = \mathrm{poly}(\varepsilon)$ iterations suffice to obtain the desired $\|w_t - v\| \leq O\big((\alpha \cdot \mathrm{OPT} + \varepsilon)/\Phi(b)\big)$ for some $t \in [T]$, which produces $L(w_t, b) \leq \alpha \cdot \mathrm{OPT} + \varepsilon$ by Theorem 8. Plugging this T back to the sample complexity, and our main theorem for PGD (Theorem 7) follows.

Appendix E. Omitted proofs from thresholded PCA (Section 3.2)

Recall that thresholded PCA outputs the top eigenvector of matrix

$$M = \mathbb{E}\big[xx^\top \mathbb{1}\left\{|y| \geq \tau\right\}\big],$$

where $\tau = \frac{1}{|b|}$ is the threshold. We use $\hat{M}$ to denote the estimation of M from n samples:

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \mathbb{1}\left\{|y_i| \geq \tau\right\}.$$

During analysis we will also partition $\mathbb{R}^d$ into two regions based on the sign of $\langle x, v\rangle + b$, and identify the their contribution to M separately:

$$\begin{cases}
M_0 = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[xx^\top \mathbb{1}\left\{|y| \geq \tau, \langle v, x\rangle + b < 0\right\}\big], \\
M_1 = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[xx^\top \mathbb{1}\left\{|y| \geq \tau, \langle v, x\rangle + b \geq 0\right\}\big].
\end{cases}$$

The rest of this section is dedicated to proving this theorem. In Appendix E.1, we state and prove three lemmas in the following order:

1. For any unit vector u, $u^\top M_0 u$ is small. (Theorem 14)
2. For any unit vector u perpendicular to v, $u^\top M_1 u$ is small. (Theorem 15)
3. $v^\top M_1 v$ is large, which helps us identify the true direction. (Theorem 16)

In Appendix E.2, we give a simple proof of the main theorem using these lemmas.

E.1. Technical lemmas.

The following is used in the proof of Lemma 14:

Fact 29 *For all sufficiently small p and any event E with $\mathbb{P}[E] = p$, we have:*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\langle x, u\rangle^2 \mathbb{1}\{E\}] = O\Big(p \log \frac{1}{p}\Big).$$

Proof Consider event $E_u = \{|\langle x, u\rangle| \geq |\Phi^{-1}(p/2)|\}$. Clearly we have $\mathbb{P}[E_u] = p$, and

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\langle x, u\rangle^2 \mathbb{1}\{E\}] \leq \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\langle x, u\rangle^2 \mathbb{1}\{E_u\}].$$

It now suffices to upper bound the right hand side. Consider threshold $t = \Phi^{-1}(p/2)$. Then, because

$$\mathbb{P}[E_u] = 2\Phi(t) = \Theta\Big(\frac{e^{-t^2/2}}{t}\Big) = p,$$

we have $t = O\big(\sqrt{\log(1/p)}\big)$ for all sufficiently small p. Therefore:

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\langle x, u\rangle^2 \mathbb{1}\{E_u\}] &= 2\int_t^\infty s^2 d\Phi(s)\\ &= \Theta\big(t^2\Phi(t)\big) = O\Big(p\log\frac{1}{p}\Big).\end{aligned}$$

■

Lemma 30 (same as Lemma 15) *For all sufficiently negative b, and for any unit vector $u \perp v$, we have:*

$$u^\top M_1 u \leq \Phi(b).$$

Proof Let x_v, x_u be the component of x along v, u, respectively. Because $u \perp v$, by the property of isotropic Gaussian, we can integrate along the u-direction for each fixed $x_v = \langle x, v\rangle$:

$$\begin{aligned}u^\top M_1 u &= \mathbb{E}[\langle x, u\rangle^2 \mathbb{1}\{|y| \geq \tau, \langle x, v\rangle + b \geq 0\}]\\ &\leq \mathbb{E}[\langle x, u\rangle^2 \mathbb{1}\{\langle x, v\rangle + b \geq 0\}]\\ &= \int_{|b|}^\infty \int_{-\infty}^\infty x_u^2 \, d\Phi(x_u)\, d\Phi(x_v)\\ &= \Phi(b).\end{aligned}$$

■

The proof of the third lemma follows the intuitive description in Section 3.2: the adversary can only suppress a fraction of the ReLU below the threshold.

Lemma 31 (same as Lemma 16) *Suppose Assumption 6 holds. For all sufficiently negative b, we have:*

$$v^\top M_1 v = \Omega\big(b^2\Phi(b)\big).$$

Proof For convenience, we will often use random variable $z = \langle x, v\rangle$. Define event $A = \{z+b \geq \tau\}$, the outcomes on which the best-fit ReLU $\sigma(z+b)$ takes value at least τ. It now suffices to show the right hand side of the following expression is at least $\Omega\big(b^2\Phi(b)\big)$:

$$v^\top M_1 v = \mathbb{E}\big[\,\langle x, v\rangle^2\, \mathbb{1}\{|y| \geq \tau\}\,\big] \geq \mathbb{E}\big[z^2 \mathbb{1}\{y \geq \tau\}\,\mathbb{1}\{A\}\big].$$

Moreover, because $t \geq |b|$ on all of A, the proof is finished once we show that $\mathbb{P}[\{y \geq \tau\} \cap A] \geq 0.01\Phi(b)$.

Consider an adversary who wants to minimize the probability $\mathbb{P}[|y| \geq \tau \mid x \in A]$, under the constraint that $\mathbb{E}\big[\big(y - \sigma(z+b)\big)^2 \mathbb{1}\{A\}\big] \leq \mathrm{OPT}$. Note that the adversary's action can be recorded as function $p : [|b| + \tau, \infty) \in [0,1]$, where $p(z') = \mathbb{P}[y < \tau \mid z = z']$ is the probability that they suppress the y value below threshold, when $\langle x, v\rangle = z'$.

We first show that the optimal strategy for the adversary is to take $p(z) = \mathbb{1}\{z \leq t\}$, where $t \in [|b| + \tau, \infty)$ is the largest value for which the adversary does not exceed the budget:

$$t = \sup\left\{ t' \in [|b| + \tau, \infty) : \int_{|b|+\tau}^{|b|+\tau+t} (z + b - \tau)^2 \, d\Phi(z) \leq \mathrm{OPT} \right\},$$

which indicates the strategy to suppress exactly the *smallest ReLU values*. Consider any $p(z)$ the adversary picks, we WLOG suppose it exhausts all the budget:

$$\int_{|b|+\tau}^{\infty} (z + b - \tau)^2 p(z)\, d\Phi(z) = \mathrm{OPT} = \int_{|b|+\tau}^{t} (z + b - \tau)^2 \, d\Phi(z),$$

consequently,

$$\int_{|b|+\tau}^{t} (1 - p(z))(z + b - \tau)^2 p(z)\, d\Phi(z) = \int_{t}^{\infty} p(z)(z + b - \tau)^2 \, d\Phi(z), \tag{10}$$

Now, to show $\mathbb{1}\{z \leq t\}$ is optimal, it now suffices to show that it suppresses more contribution from the ReLU than this $p(z)$:

$$\int_{|b|+\tau}^{t} (1 - p(z)) z^2 \, d\Phi(z) \geq \int_{t}^{\infty} p(z) z^2 \, d\Phi(z). \tag{11}$$

But this is immediately true in light of Eq. (10): the value $\frac{z^2}{(z+b-\tau)^2}$ is always larger when $z \leq t$ than when $z \geq t$, hence the left hand side is indeed larger in Eq. (11). Intuitively, the adversary can gain more by paying less when the ReLU value is small.

Now, to finish the proof, we have to show that the adversary, even under this optimal strategy, cannot substantially harm the contribution from ReLU. We claim that for large α and sufficiently negative b wee have $t \leq \frac{1}{|b|}$, otherwise the adversary would exceed its budget:

$$\begin{aligned}\int_{|b|+\tau}^{|b|+\tau+\frac{1}{|b|}} (z+b-\tau)^2 \, d\Phi(z) &= \Theta\left(e^{-1}\sum_{k=3}^{\infty}\frac{1}{k!}\right)\int_{|b|+\tau}^{\infty}(z+b-\tau)^2 \, d\Phi(z) \\ &= \Theta\left(e^{-1}\sum_{k=3}^{\infty}\frac{1}{k!}\right)\frac{\Phi(b-\tau)}{(b-\tau)^2} \\ &= \Theta\left(e^{-2}\sum_{k=3}^{\infty}\frac{1}{k!}\right)\frac{\Phi(b)}{b^2}.\end{aligned}$$

Here the first equality is justified in Appendix B. Note that $\left(e^{-2}\sum_{k=3}^{\infty}\frac{1}{k!}\right)$ is a constant independent of α, hence by taking α large and $\mathrm{OPT} = O\left(\frac{\Phi(b)}{\alpha b^2}\right)$, we must have $t \leq \frac{1}{|b|}$. The proof is finished by noting that $\Omega(1)$ fraction of the ReLU's contribution is kept when $t \leq \frac{1}{|b|}$:

$$\int_{|b|+\tau+t}^{\infty} z^2 \, d\Phi(z) \geq \int_{|b|+\frac{2}{|b|}}^{\infty} z^2 \, d\Phi(z) = \Omega\big(b^2\Phi(b)\big).$$

■

E.2. Proof of main theorem.

The main theorem now follows from some linear algebra manipulation and concentration inequality.

Proof [Proof of Theorem 13] It's clear now that M has greater magnitude in v than in any perpendicular direction $u \perp v$, for sufficiently large α and negative b. We now show that the finite-sample estimation $\hat{M}$ using $n = \mathrm{poly}(d, 1/\varepsilon, 1/\delta)$ samples is a good estimator of M, in the sense that its top eigenvector still has a dominating component in v.

For each sample i, the matrix $\mathbb{E}[x_i x_i^\top \mathbb{1}\{|y| \geq \tau\}]$ is a 1-sub-Gaussian matrix, and by an analog of Hoeffding's inequality for matrices (see Tropp (2012) for related results):

$$\mathbb{P}\big[\|\hat{M} - M\|_{op} > t\big] \leq 2d\exp\big(-\Omega(nt^2)\big).$$

Hence, by taking $n = O\big(\frac{\log(d/\delta)}{\Phi(b)^2}\big) = \mathrm{poly}\big(\log d, \log\frac{1}{\delta}, \frac{1}{\varepsilon}\big)$. we have $\|\hat{M} - M\|_{op} \leq \Phi(b)$ with probability at least $1-\delta$. Combining this with previous lemmas, it follows that with the same probability we have:

$$\|\hat{M}u\| = O\left(\frac{1}{b^2} + \frac{\log\alpha}{\alpha}\right)\|\hat{M}v\|,$$

for all unit vector $u \perp v$.

Let $w \in \mathbb{R}^d$ be the output of thresholded PCA. We can decompose it into $w = \langle w, v\rangle \cdot v + \langle w, u\rangle \cdot u$, where u is again a unit vector perpendicular to v. Then:

$$\begin{aligned} \|\hat{M}v\| \leq \|\hat{M}w\| &\leq \langle w, v\rangle \|\hat{M}v\| + \langle w, u\rangle \|\hat{M}u\| \\ &\leq \left(\langle w, v\rangle + \langle w, u\rangle \frac{\|\hat{M}u\|}{\|\hat{M}v\|} \right) \|\hat{M}v\| \end{aligned}$$

Therefore,

$$1 \leq \langle w, v\rangle + \langle w, u\rangle \frac{\|\hat{M}u\|}{\|\hat{M}v\|} \leq \langle w, v\rangle + \frac{\|\hat{M}u\|}{\|\hat{M}v\|},$$

which means $\langle w, v\rangle \geq 1 - \frac{\|\hat{M}u\|}{\|\hat{M}v\|}$.

Now the proof is finished by plugging in $b \leq -\sqrt{\alpha / \log \alpha}$, in which case we have:

$$\langle w, v\rangle \geq 1 - \frac{\|\hat{M}u\|}{\|\hat{M}v\|} = 1 - O\left(\frac{\log \alpha}{\alpha}\right),$$

as desired. ■

Appendix F. Putting things together

Now we state and prove the main algorithmic result of this paper:

Theorem 32 *There exists a constant α, such that for all $W > 0$ the following holds. Let $\mathcal{D}$ be a the joint distribution of $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, where the x-marginal is $\mathcal{N}(0, I_d)$. With population expectations replaced by finite-sample estimates, Algorithm 1 will run in $\mathrm{poly}(d, \frac{1}{\varepsilon}, \frac{1}{\delta}, W)$ time and samples, and with probability at least $1 - \delta$, outputs parameters $\hat{w} \in \mathbb{R}^n, \hat{b} \in \mathbb{R}$, such that:*

$$L(\hat{w}, \hat{b}) \leq \alpha \cdot \inf_{\substack{w \in \mathbb{R}^d, b \in \mathbb{R}: \\ \|w\|_2 \leq W}} L(w, b) + \varepsilon.$$

Proof The prior works that we use (Awasthi et al., 2023; Diakonikolas et al., 2022c) both guarantee *some* constant approximation factor. We can simply take the maximum between their factors and our algorithm's approximation factor to be the final approximation factor.

Regarding the special case of positive bias: the algorithm in Diakonikolas et al. (2022c) works for a class of unbounded activations, which includes positively biased ReLUs. Specifically, for any $b > 0$, the new activation function $\tilde{\sigma}(\langle x, w\rangle) = \sigma(\langle x, w\rangle + b)$ is a valid activation function under their framework. We note that this requires knowledge about b, but we can again do a grid search over the "guesses" of b in polynomial time, up to $b \leq \mathrm{poly}(W, \log 1/\varepsilon)$.

To deal with very positive b, it suffices to use the parameters w, b from a certain linear regression variant. In particular, we run linear regression with bias, and we limit bias to be $b \geq \Theta\big(\sqrt{\log(1/\varepsilon)}\big)$.

Let w, b be the output of that linear regression problem. Suppose the optimal ReLU is $\sigma(\langle v, x\rangle + b^*)$ with $b^* \geq \Theta\big(\sqrt{\log(1/\varepsilon)}\big)$. Then:

$$\begin{aligned}\mathbb{E}\big[\big(y - \sigma(\langle w, x\rangle + b)\big)^2\big] &\leq 2\mathbb{E}\big[\big(y - (\langle w, x\rangle + b)\big)^2\big] + 2\mathbb{E}\big[\big(\langle w, x\rangle + b - \sigma(\langle w, x\rangle + b)\big)^2\big] \\ &= 2\mathbb{E}\big[\big(y - (\langle w, x\rangle + b)\big)^2\big] + \Theta(W^2/b)\Phi(-b/W) \\ &\leq 2\mathbb{E}\big[\big(y - (\langle v, x\rangle + b^*)\big)^2\big] + \Theta(W^2/b)\Phi(-b/W) \\ &\leq 2\mathbb{E}\big[\big(y - \sigma(\langle v, x\rangle + b^*)\big)^2\big] + \Theta(W^2/b)\Phi(-b/W) + \Theta(W^2/b^*)\Phi(-b^*/W),\end{aligned}$$

it follows that the loss of this ReLU candidate is $\leq O(\mathrm{OPT}) + \varepsilon$.

Moreover, the algorithm (Awasthi et al., 2023) for moderately-biased ReLU can solve all constant-bounded bias, for any constant of our choice. The trade-off is that the approximation factor is larger as we allow for larger constants. This guarantee is good enough for our purpose, as our algorithm has provable guarantee when $b \leq b_\alpha < 0$, where b_α is constant.

Now we have reduced to the $b \to -\infty$ regime. Since we also consider the zero function $\mathbf{0}$ as a potential output in Algorithm 1, it also suffices to assume $\mathbf{0}$ incurs loss $> \alpha \cdot \mathrm{OPT} + \varepsilon$. This allows us to apply our main theorems Theorem 7 and Theorem 13, which take into account finite-sample estimations.

Finally, we bound the time complexity of our grid search approach. Suppose the optimal ReLU is some non-zero function $\sigma(\langle x, v\rangle + b)$. Then, from $\varepsilon = O\big(\Phi(b)/b^2\big)$ we know $b = O(\sqrt{\log 1/\varepsilon})$, and we also have $\|v\| \leq W$ by assumption. This means the grid search with accuracy $0.1\sqrt{\varepsilon}$ terminates in $\mathrm{poly}(W, 1/\varepsilon)$ rounds.

During the grid search, there must be some pair $(\beta_{\text{ind}}, \gamma_{\text{ind}})$ of parameters that correctly estimates the optimal $\|v\|, b$ each up to error at most $0.1\sqrt{\varepsilon}$. Since we apply the subroutines in a way that produces a smaller error margin than α and ε:

$$L_{\text{ind}} \leq 0.1\alpha \cdot \mathrm{OPT}_{\text{ind}} + 0.1\frac{\varepsilon}{\beta_{\text{ind}}^2},$$

this loss will indeed translate into $\alpha \cdot \mathrm{OPT} + \varepsilon$ via Theorem 3 and Theorem 4. ■

Appendix G. Omitted proofs from CSQ hardness (Section 4)

In this section we show the following theorem:

Theorem 33 (Same as Theorem 2) *There exists a function $F(\varepsilon)$ that goes to infinity as $\varepsilon \to 0$, such that for any $\varepsilon > 0$ and any constant $\alpha \geq 1$, there exists a family of instances with $\mathrm{OPT} \leq \varepsilon/\alpha$ such that any CSQ algorithm that can agnostically learn an arbitrary ReLU neuron with loss at most $\alpha \cdot \mathrm{OPT} + \varepsilon$ (as defined in Eq. (2)) must use either $2^{d^{\Omega(1)}}$ queries or queries of tolerance $d^{-F(\varepsilon)}$.*

In this section we will follow a standard procedure for showing CSQ hardness, which can be found in e.g. Diakonikolas et al. (2021b). We use many lemmas from previous works as black box.

G.1. Preliminaries on high-dimensional geometry and CSQ

Before proving the main theorem, we shall introduce some helpful lemmas from previous works. The first lemma quantifies the fact that, for large d, there are exponentially many near-perpendicular directions in $\mathbb{R}^d$.

Lemma 34 (Lemma 3.7 from Diakonikolas et al. (2017)) *For any constant $c \in (0, \frac{1}{2})$, there exists a set S of $2^{\Omega(d^c)}$ unit vectors in $\mathbb{R}^d$, such that for each distinct $u, v \in S$ we have:*

$$|\langle u, v\rangle| = O(d^{c-\frac{1}{2}}).$$

Next, we introduce some background in CSQ hardness. Most of these definitions and lemmas can be traced back to the seminal work Feldman et al. (2013).

Definition 35 (CSQ dimension) *Let $\mathcal{D}_x$ be a distribution over space $\mathcal{X}$, $\mathcal{G}$ a set of functions from $\mathcal{X}$ to $\mathbb{R}$, and β, γ two positive parameters. We define the correlational statistical query dimension of $\mathcal{G}$ w.r.t. $\mathcal{D}_x$ with pairwise correlation (γ, β), denoted $\mathrm{CSD}_{\mathcal{D}_x}(\mathcal{G}, \gamma, \beta)$, to be the largest integer D such that, there is a subset of D functions $\{f_1, \ldots, f_D\} \subseteq \mathcal{G}$, such that for all $i, j \in [D]$:*

$$\left|\mathbb{E}_{x\sim\mathcal{D}_x}[f_i(x)f_j(x)]\right| \le \begin{cases} \gamma & , \textit{if } i \neq j, \\ \beta & , \textit{if } i = j. \end{cases}$$

Definition 36 (average CSQ dimension) *Let $\mathcal{D}_x$ be a distribution over space $\mathcal{X}$, $\mathcal{G}$ a finite set of functions from $\mathcal{X}$ to $\mathbb{R}$, and γ a positive parameter. We define the average pairwise correlation of functions in $\mathcal{G}$ to be:*

$$\rho(\mathcal{G}) = \frac{1}{|\mathcal{G}|^2} \sum_{f,g\in\mathcal{G}} \left|\mathbb{E}_{x\sim\mathcal{D}_x}[f(x)g(x)]\right|.$$

Then, the average correlational statistical query dimension of $\mathcal{G}$ w.r.t. $\mathcal{D}_x$ with parameter γ, denoted $\mathrm{CSDA}_{\mathcal{D}_x}(\mathcal{G}, \gamma)$, is the largest integer D such that every subset $\mathcal{G}' \subseteq \mathcal{G}$ of size at least $\frac{|\mathcal{G}|}{D}$ has average pairwise correlation $\rho(\mathcal{G}') \ge \gamma$.

The next lemma shows large CSQ dimension implies large average CSQ dimension:

Lemma 37 *Let $\mathcal{D}_x$ be a distribution over space $\mathcal{X}$, and let $\mathcal{G}$ a finite set of functions from $\mathcal{X}$ to $\mathbb{R}$ with $\mathrm{CSD}_{\mathcal{D}_x}(\mathcal{G}, \gamma, \beta) = D$ for some $\gamma, \beta > 0$. Then, for all $\gamma' > 0$, we have:*

$$\mathrm{CSD}_{\mathcal{D}_x}(\mathcal{G}, \gamma + \gamma') \ge \frac{D\gamma'}{\beta - \gamma}.$$

Finally, large *Average* CSQ dimension implies the following CSQ hardness result:

Lemma 38 (Theorem B.1 in Goel et al. (2020a)) *Suppose $x \sim \mathcal{D}_x$ and let $\mathcal{H}$ be a real-valued function class that includes the zero-function, with $\mathbb{E}_{x\sim\mathcal{D}_x}[f(x)^2] \ge \eta$ for all non-zero $f \in \mathcal{H}$. Let $D = \mathrm{CSDA}_{\mathcal{D}_x}(\mathcal{H}, \gamma)$ for some $\gamma > 0$, then any CSQ algorithm that realizably learns $\mathcal{H}$ up to L^2 error strictly smaller than η needs at least $\frac{D}{2}$ queries or queries of tolerance $\sqrt{\gamma}$.*

G.2. Hermite expansion of negatively biased ReLU

We first define some new notations. For any $k \in \mathbb{N}$, let $H_k(x)$ be the kth probabilist's Hermite polynomial:

$$H_k(x) = (-1)^k \exp\left(\frac{x^2}{2}\right) \cdot \frac{d^k}{dx^k} \exp\Big(-\frac{x^2}{2}\Big),$$

Note that $\int_{-\infty}^{\infty} H_k(s)^2 \, d\Phi(s) = k!$, so the set

$$\left\{\frac{H_k(x)}{\sqrt{k!}} : k \in \mathbb{N}\right\}$$

forms an orthonormal basis for Hilbert space $L^2_{\mathcal{N}}$, the space of all square-integrable functions under to the standard Gaussian measure. We use $\langle f, g\rangle_{\mathcal{N}}$ and $\|g\|_{\mathcal{N}}$ to denote the inner product and norm of this space, defined by:

$$\begin{cases} \langle f, g\rangle_{\mathcal{N}} & := \int_{-\infty}^{\infty} f(s)g(s)\, d\Phi(s), \\ \|f\|^2_{\mathcal{N}} & := \langle f, f\rangle_{\mathcal{N}}. \end{cases}$$

We are now equipped to analyze the Hermite expansion of negatively-biased ReLUs:

Lemma 39 (Lemma 3.5 of Awasthi et al. (2021)) *Let $f_b(s) = \sigma(s-b)$ for some $b < 0$. Then,*

$$\begin{cases} \langle f_b, H_0\rangle_{\mathcal{N}} & = \varphi(b) - |b|\Phi(b), \\ \langle f_b, H_1\rangle_{\mathcal{N}} & = \Phi(b), \end{cases}$$

and for all $k \geq 2$:

$$\begin{aligned} \langle f_b, H_k\rangle_{\mathcal{N}} &= (-1)^k \cdot H_{k-2}(b) \cdot \varphi(b) \\ &= H_{k-2}(|b|) \cdot \varphi(b) \end{aligned}$$

Consequently, as $b \to -\infty$, the b-biased ReLU will correlate less with low-degree Hermite polynomials, in the following sense:

Lemma 40 *For any fixed $t \in \mathbb{N}$, Let $f_b(s) = \sigma(s-b)$, then as $b \to -\infty$, we have:*

$$\sum_{k=0}^{t} \left\langle f_b, \frac{H_k}{\sqrt{k!}} \right\rangle^2 = o(1) \cdot \|f\|^2_{\mathcal{N}}.$$

Proof By Parseval's identity:

$$\sum_{k=0}^{\infty} \left\langle f_b, \frac{H_k}{\sqrt{k!}} \right\rangle = \|f_b\|^2_{\mathcal{N}} = \big(2 + o(1)\big)\frac{\varphi(b)}{|b|^3}.$$

Applying Theorem 39, we have:

$$\begin{aligned} \frac{\sum_{k=0}^{t} \left\langle f_b, \frac{H_k}{\sqrt{k!}} \right\rangle^2}{\|f_b\|^2_{\mathcal{N}}} &= \frac{o(\varphi(b)^2) + \sum_{k=2}^{t} H_{k-2}(|b|)^2 \cdot \varphi(b)^2/k!}{(2+o(1))\varphi(b)/|b|^3} \\ &= O\big(p(|b|)\big) \cdot \varphi(b), \end{aligned}$$

for some polynomial p of bounded degree. Since $\varphi(b)$ decreases exponentially with $|b|$, this value vanishes as $b \to -\infty$. ■

Finally, we prove the following lemma that is tailor-made for our problem.

Lemma 41 (same as Lemma 17) *The following holds for all sufficiently small ε. Let $g_\varepsilon(s) = \sigma(s - b_\varepsilon)$, where b_ε is chosen so that $\|g_\varepsilon\|_{\mathcal{N}}^2 = 3\varepsilon$. Let integer $t_\varepsilon \in \mathbb{N}$ be:*

$$t_\varepsilon := \max\left\{ t \in \mathbb{N} : \sum_{k=0}^{t} \left\langle g_\varepsilon, \frac{H_k}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2 \le \frac{\varepsilon}{\alpha} \right\},$$

then, we have $t_\varepsilon \to \infty$ as $\varepsilon \to 0$.

Proof By the previous lemma, for any fixed t we have

$$\sum_{k=0}^{t} \left\langle g_\varepsilon, \frac{H_k}{\sqrt{n!}} \right\rangle_{\mathcal{N}}^2 = o(1) \cdot \|g_\varepsilon\|_{\mathcal{N}}^2 = o(\varepsilon).$$

Hence t_ε cannot be bounded as $\varepsilon \to 0$. ■

G.3. Proof of main theorem

To bound the pairwise correlation of two functions defined on almost-perpendicular directions, we introduce one final lemma which we use as a black box:

Lemma 42 (Lemma 2.3 from Diakonikolas et al. (2021b)) *For function $g : \mathbb{R} \to \mathbb{R}$ and unit vectors $u, v \in \mathbb{R}^d$, we have:*

$$\mathop{\mathbb{E}}_{x \in \mathcal{N}(0, I_d)} [g(\langle x, u \rangle) g(\langle x, v \rangle)] \le \sum_{k=0}^{\infty} |\langle u, v \rangle|^k \cdot \left\langle g, \frac{H_k(x)}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2.$$

We are now ready to prove the main hardness result. Suppose some CSQ algorithm $\mathcal{A}$ can agnostically learn ReLUs with unbounded bias up to error $\alpha \cdot \mathrm{OPT} + \varepsilon$. We first fix some sufficiently small $\varepsilon > 0$, and consider the following set:

$$\mathcal{G} = \{G_{v,\varepsilon} = \tilde{g}_\varepsilon(\langle x, v \rangle) : v \in S\},$$

where S is the set of $2^{d^{\Omega(1)}}$ almost-perpendicular unit vectors in Theorem 34, and $\tilde{g}_\varepsilon$ is a modification of g_ε in Theorem 17, by removing the *first t_ε Hermite components* from the latter:

$$\tilde{g}_\varepsilon(s) = \sum_{k=t_\varepsilon+1}^{\infty} \left\langle g_\varepsilon, \frac{H_k}{\sqrt{k!}} \right\rangle_{\mathcal{N}} \frac{H_k(s)}{\sqrt{k!}}.$$

We now show that the functions $\{G_{v,\varepsilon} : v \in S\}$ have small pairwise correlation. For each distinct $u, v \in S$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \in \mathcal{N}(0,I_d)}[G_{u,\varepsilon}(x)G_{v,\varepsilon}(x)] &\leq \sum_{k=0}^{\infty} |\langle u, v\rangle|^k \cdot \left\langle \tilde{g}_\varepsilon, \frac{H_k(x)}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2 \\
&= \sum_{k=t_\varepsilon+1}^{\infty} |\langle u, v\rangle|^k \cdot \left\langle \tilde{g}_\varepsilon, \frac{H_k(x)}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2 \\
&\leq |\langle u, v\rangle|^{t_\varepsilon+1} \sum_{k=t_\varepsilon+1}^{\infty} \left\langle \tilde{g}_\varepsilon, \frac{H_k(x)}{\sqrt{k!}} \right\rangle_{\mathcal{N}}^2 \\
&\leq d^{-\Omega(t_\varepsilon)} \cdot \|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2.
\end{aligned}$$

Here, the first step is by Theorem 42, and the last step is by property of S in Theorem 34. Since we also have $\mathop{\mathbb{E}}_{x \in \mathcal{N}(0,I_d)}[G_{u,\varepsilon}(x)^2] = \|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2$ by definition, this means:

$$\mathrm{CSD}_{\mathcal{N}(0,I_d)}\big(\mathcal{G},\ d^{-\Omega(t_\varepsilon)}\|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2,\ \|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2\big) \geq 2^{d^{\Omega(1)}}.$$

Applying Theorem 37 with $\gamma = \gamma' = d^{-\Omega(t_\varepsilon)}\|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2$, this translates into:

$$\begin{aligned}
\mathrm{CSDA}_{\mathcal{N}(0,I_d)}\big(\mathcal{G}, d^{-\Omega(t_\varepsilon)}\|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2\big) &\geq \frac{2^{d^{\Omega(1)}} \cdot d^{-\Omega(t_\varepsilon)} \cdot \|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2}{(1 - d^{-\Omega(t_\varepsilon)})\|\tilde{g}_\varepsilon\|_{\mathcal{N}}^2} \\
&= 2^{d^{\Omega(1)}} \text{ for all sufficiently large } d.
\end{aligned}$$

Now we show that $\mathcal{A}$ can learn this low-correlation class $\mathcal{G}$ up to nontrivial accuracy. Given CSQ-oracle access to $x \sim \mathcal{N}(0, I_d)$ and y labeled by some $G_{v,\varepsilon}(x)$, $\mathcal{A}$ outputs a function h with error at most:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,I_d)}\left[\big(h(x) - G_{v,\varepsilon}(x)\big)^2\right] &\leq \alpha \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,I_d)}\left[\big(\sigma(\langle x, v\rangle + b_\varepsilon) - G_{v,\varepsilon}(x)\big)^2\right] + \varepsilon \\
&= \alpha \cdot \|g_\varepsilon - \tilde{g}_\varepsilon\|_{\mathcal{N}}^2 + \varepsilon \\
&\leq \alpha \cdot \frac{\varepsilon}{\alpha} + \varepsilon \\
&< \left(3 - \frac{1}{\alpha}\right)\varepsilon \\
&\leq \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,I_d)}[G_{v,\varepsilon}(x)^2].
\end{aligned}$$

We can thus apply Theorem 38 with $\eta = \|\tilde{g}_\varepsilon\|_{\mathcal{N}} = \mathbb{E}_{x \sim \mathcal{N}(0,I_d)}[G_{v,\varepsilon}(x)^2]$, and conclude that $\mathcal{A}$ must use $2^{d^{\Omega(1)}}$ queries or queries of tolerance $d^{-\Omega(t_\varepsilon)}$. Now the proof is finished since $t_\varepsilon \to \infty$ as $\varepsilon \to 0$.