Supplemental: Superpolynomial Lower Bounds for Learning One-Layer Neural Networks using Gradient Descent

Surbhi Goel [1] Aravind Gollakota [1] Zhihan Jin [2] Sushrut Karmalkar [1] Adam Klivans [1]

A. Bounding the function norms under the Gaussian

Our goal in this section will be to give lower bounds on the norms of the functions in $\mathcal{C}_{\mathrm{orth}}(n,k)$, which is a technical requirement for our results to hold (see Lemma 4.4 and Corollary 4.6). Note that when learning with respect to L_2 error, such a lower bound is necessary if we wish to state SQ lower bounds, since if the target had small norm, say $\|f\|_D \le \epsilon$, then the zero function trivially achieves L_2 error ϵ.

All inner products and norms in this section will be with respect to the standard Gaussian, $\mathcal{N}(0,I)$. Since we will fix S throughout, for our purposes the only relevant part of the input is x_S and so we drop the subscripts and let $g = g_S$, $f = f_S$ and $x = x_S$, so that g and f are functions of $x \in \mathbb{R}^k$. Our approach will be as follows. In order to prove a norm lower bound on f, we will prove an anticoncentration result for g. To this end we first calculate the second moment of g in terms of the Hermite coefficients of ϕ.

Lemma A.1. *Under the distribution $\mathcal{N}(0,I_n)$, let the Hermite representation of ϕ be $\phi(x) = \sum_{i=0}^{\infty} \widehat{\phi}_i \tilde{H}_i(x)$, where $\tilde{H}_i(x)$ is the i^{th} normalized probabilists' Hermite polynomial. Then*

$$\mathbb{E}\left[g(x)^2\right] = 4^k \sum_{i\ge 0} \frac{\widehat{\phi}_i^{\,2}}{k^i} \sum_{\substack{i_1+\cdots+i_k=i \\ i_1,\ldots,i_k \text{ are odd}}} \binom{i}{i_1,\ldots,i_k}.$$

Proof. We use $\mathbb{E}$ in this proof instead of $\mathbb{E}_{x\sim\mathcal{N}(0,I_n)}$ for simplicity. Then we have

$$\begin{aligned}
&\mathbb{E}\left[g(x)^2\right] \\
&= \mathbb{E}\left[\sum_{\alpha\in\{\pm1\}^k} \chi(\alpha)\phi\left(\frac{\alpha\cdot x_S}{\sqrt{k}}\right)\right]\left[\sum_{\beta\in\{\pm1\}^k} \chi(\beta)\phi\left(\frac{\beta\cdot x_S}{\sqrt{k}}\right)\right] \\
&= \sum_{\alpha,\beta\in\{\pm1\}^k} \prod_{l=1}^{k} \alpha_l\beta_l\, \mathbb{E}\left[\phi\left(\frac{\alpha\cdot x_S}{\sqrt{k}}\right)\phi\left(\frac{\beta\cdot x_S}{\sqrt{k}}\right)\right] \\
&= \sum_{\alpha,\beta\in\{\pm1\}^k} \prod_{l=1}^{k} \alpha_l\beta_l\, \mathbb{E}\left[\sum_{i,j\ge0} \widehat{\phi_i}\widehat{\phi_j}\tilde{H}_i\left(\frac{\alpha\cdot x_S}{\sqrt{k}}\right)\tilde{H}_j\left(\frac{\beta\cdot x_S}{\sqrt{k}}\right)\right] \\
&= \sum_{\alpha,\beta\in\{\pm1\}^k} \prod_{l=1}^{k} \alpha_l\beta_l \sum_{i,j\ge0} \widehat{\phi_i}\widehat{\phi_j}\, \mathbb{E}\left[\tilde{H}_i\left(\frac{\alpha\cdot x_S}{\sqrt{k}}\right)\tilde{H}_j\left(\frac{\beta\cdot x_S}{\sqrt{k}}\right)\right].
\end{aligned}$$

Since $x \sim \mathcal{N}(0,I_k)$, $\frac{\langle\alpha,x_S\rangle}{\sqrt{k}}$ and $\frac{\langle\beta,x_S\rangle}{\sqrt{k}}$ are both standard Gaussian and have correlation $\frac{\langle\alpha,\beta\rangle}{k}$, we then apply the following well-known property of the Hermite polynomials.

$$\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\left(\begin{smallmatrix}1&\rho\\ \rho&1\end{smallmatrix}\right)\right)} \tilde{H}_i(a)\tilde{H}_j(b) = \delta_{i,j}\rho^i,$$

where $\delta_{i,j}$ is the Dirac delta function.

$$\begin{aligned}
\mathbb{E}\left[g(x)^2\right] &= \sum_{\alpha,\beta\in\{\pm1\}^k} \prod_{l=1}^{k} \alpha_l\beta_l \sum_{i\ge0} \widehat{\phi}_i^{\,2}\left(\frac{\alpha\cdot\beta}{k}\right)^i \\
&= \sum_{w,\theta\in\{\pm1\}^k} \prod_{l=1}^{k} w_l \sum_{i\ge0} \widehat{\phi}_i^{\,2}\left(\frac{\sum_{l=1}^k w_l}{k}\right)^i \\
&= 2^k \sum_{w\in\{\pm1\}^k} \prod_{l=1}^{k} w_l \sum_{i\ge0} \widehat{\phi}_i^{\,2}\left(\frac{\sum_{l=1}^k w_l}{k}\right)^i,
\end{aligned}$$

where $w_i = \alpha_i\beta_i$ and $\theta_i = w_i\alpha_i$. Note that Assumption 3.3 implies that $\sum_{i=0}^{\infty}\widehat{\phi}_i^{\,2} < \infty$, the series above is absolute convergent. Then,

$$\begin{aligned}
&\mathbb{E}\left[g(x)^2\right] \\
&= 2^k \sum_{i\ge0} \widehat{\phi}_i^{\,2} \sum_{w\in\{\pm1\}^k} \prod_{l=1}^{k} w_l \left(\frac{\sum_{l=1}^k w_l}{k}\right)^i \\
&= 2^k \sum_{i\ge0} \frac{\widehat{\phi}_i^{\,2}}{k^i} \sum_{w\in\{\pm1\}^k} \prod_{l=1}^{k} w_l \sum_{i_1+\cdots+i_k=i} \prod_{l=1}^{k} w_l^{i_l}\binom{i}{i_1,\ldots,i_k}
\end{aligned}$$

[1] Department of Computer Science, University of Texas at Austin [2] Department of Computer Science, Shanghai Jiao Tong University. Correspondence to: Aravind Gollakota <aravindg@cs.utexas.edu>, Adam Klivans <klivans@cs.utexas.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

$$
\begin{aligned}
&= 2^k \sum_{i\geq 0} \frac{\widehat{\phi_i}^2}{k^i} \sum_{i_1+\cdots+i_k=i} \binom{i}{i_1,\ldots .i_k} \sum_{w\in\{\pm1\}^k} \prod_{l=1}^{k} w_l^{i_l+1} \\
&= 2^k \sum_{i\geq 0} \frac{\widehat{\phi_i}^2}{k^i} \sum_{i_1+\cdots+i_k=i} \binom{i}{i_1,\ldots,i_k} \prod_{l=1}^{k} \left[1^{i_l+1} + (-1)^{i_l+1}\right] \\
&= 4^k \sum_{i\geq 0} \frac{\widehat{\phi_i}^2}{k^i} \sum_{\substack{i_1+\cdots+i_k=i \\ i_1,\ldots,i_k \text{ are odd}}} \binom{i}{i_1,\ldots,i_k}
\end{aligned}
$$

since we consider all distinct monomials in $\left(\sum_{l=1}^{k} w_l\right)^i$. Note that $\sum_{\substack{i_1+\cdots+i_k=i \\ i_1,\ldots,i_k \text{ are odd}}} \binom{i}{i_1,\ldots,i_k}$ is always non-negative and is positive iff $i \geq k$ and $i \equiv k \pmod 2$. □

A.1. ReLU Activation

The goal of this section is to give a lower-bound of $\|f\|$ for $\phi = \mathrm{ReLU}$ under the standard Gaussian distribution $\mathcal{N}(0, I)$. To this end, we prove an anti-concentration for g. We first give a lower bound on $\|g\|$ based on the Hermite coefficients of ϕ. If g were bounded, this alone would imply anti-concentration as in Appendix A.2. But since it is not, we first introduce g^T, where all activations are truncated at some T. We pick T large enough that g and g^T behave almost identically over $\mathcal{N}(0, I)$. We then show a lower bound on $\left\|g^T\right\|$, translate that into an anticoncentration result for g^T, and finally into one for g.

Let $T > 0$ be some constant to be determined later. Let

$$
\mathrm{ReLU}^T(x) = \min(\mathrm{ReLU}(x), T)
$$

and

$$
g^T(x) = \sum_{w\in\{\pm1\}^k} \chi(w)\, \mathrm{ReLU}^T\left(\frac{x\cdot w}{\sqrt{k}}\right).
$$

The following lemma from (Goel et al., 2019) describes the Hermite coefficients of ReLU.

Lemma A.2.

$$
\mathrm{ReLU}(x) = \sum_{i=0}^{\infty} c_i \tilde{H}_i(x)
$$

where

$$
\begin{aligned}
&c_0 = \sqrt{\frac{1}{2\pi}}, \quad c_1 = \frac{1}{2}, \\
&c_{2i-1} = 0, \quad c_{2i} = \frac{H_{2i}(0) + 2iH_{2i-2}(0)}{\sqrt{2\pi(2i)!}} \quad \textit{for } i \geq 2.
\end{aligned}
$$

In particular, $c_{2i}^2 = \Theta(i^{-2.5})$.

We can now derive a lower bound on the norm of g.

Lemma A.3. *When k is even,*

$$
\|g\| = \Omega\left(\left(\frac{4}{e}\right)^{(\frac{1}{2}+o(1))k}\right).
$$

Proof. Due to Lemma A.1,

$$
\begin{aligned}
\mathbb{E}\left[g(x)^2\right] &= 4^k \sum_{i\geq 0} \frac{c_i^2}{k^i} \sum_{\substack{i_1+\cdots+i_k=i \\ i_1,\ldots,i_k, \text{ are odd}}} \binom{i}{i_1,\ldots,i_k} \\
&\geq \frac{4^k c_k^2}{k^k} \sum_{\substack{i_1+\cdots+i_k=k \\ i_1,\ldots,i_k, \text{ are odd}}} \binom{k}{i_1,\ldots,i_k} \\
&\geq \frac{4^k c_k^2 k!}{k^k}.
\end{aligned}
$$

The lemma then follows by the Stirling's approximation,

$$
n! \geq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.
$$

and the bound on the Hermite coefficients,

$$
c_k^2 = \Theta(k^{-2.5}).
$$

□

For the difference of $g(x)$ and $g^T(x)$, we have

Lemma A.4.

$$
\left\|g - g^T\right\| \leq 2^k\, e^{-\frac{T^2}{4}} \sqrt{T^2 + 1 - \frac{T}{\sqrt{2\pi}}}
$$

Proof. Let $\mathrm{ReLU}_w(x)$ be shorthand for $\mathrm{ReLU}(\frac{x\cdot w}{\sqrt{k}})$, and similarly ReLU_w^T. Observe that by the triangle inequality,

$$
\begin{aligned}
\left\|g - g^T\right\| &= \left\|\sum_{w\in\{\pm1\}^k} \chi(w)\left(\mathrm{ReLU}_w - \mathrm{ReLU}_w^T\right)\right\| \\
&\leq \sum_{w\in\{\pm1\}^k} \left\|\mathrm{ReLU}_w - \mathrm{ReLU}_w^T\right\| \\
&= 2^k \left\|\mathrm{ReLU} - \mathrm{ReLU}^T\right\|_{\mathcal{N}(0,1)},
\end{aligned}
$$

where the last equality holds because for any unit vector v and $x \sim \mathcal{N}(0, I)$, $x \cdot v$ has the distribution $\mathcal{N}(0, 1)$. Now,

$$
\left\|\mathrm{ReLU} - \mathrm{ReLU}^T\right\|_{\mathcal{N}(0,1)}^2 = \int_T^{\infty} (x - T)^2\, p(x)\, dx,
$$

where $p(x)$ is the probability density function of $\mathcal{N}(0, 1)$. Note that $p'(x) = -xp(x)$. We have

$$
\int_T^{\infty} x^2 p(x) dx = \int_T^{\infty} -x\, d(p(x))
$$

$$
\begin{aligned}
&= -x\,p(x)\Big|_T^\infty + \int_T^\infty p(x)dx \\
&\qquad\qquad\qquad\qquad \text{(integration by parts)} \\
&= T\,p(T) + \mathbb{P}_{x\sim\mathcal{N}(0,1)}(x > T),
\end{aligned}
$$

$$\int_T^\infty x\,p(x)dx = -p(x)\Big|_T^\infty = p(T),$$

$$\int_T^\infty p(x)dx = \mathbb{P}_{x\sim\mathcal{N}(0,1)}(x > T) \le e^{-\frac{T^2}{2}}.$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[g(x) - g^T(x)\right]^2 \\
&\le 4^k\,\left[(T^2+1)\mathbb{P}_{x\sim\mathcal{N}(0,1)}(x > T) - T\,p(T)\right] \\
&\le 4^k\,e^{-\frac{T^2}{2}}\left(T^2 + 1 - \frac{T}{\sqrt{2\pi}}\right).
\end{aligned}
$$

□

Lemma A.5.

$$\mathbb{P}[g(x) \neq g^T(x)] \le 2^k\,e^{-\frac{T^2}{2}}.$$

Proof. For any $w \in \{\pm 1\}^k$,

$$
\begin{aligned}
&\mathbb{P}_{x\sim\mathcal{N}(0,I)}\left[\mathrm{ReLU}(\frac{x\cdot w}{\sqrt{k}}) \neq \mathrm{ReLU}^T(\frac{x\cdot w}{\sqrt{k}})\right] \\
&= \mathbb{P}_{t\sim\mathcal{N}(0,1)}[t > T] \\
&\le e^{-\frac{T^2}{2}}.
\end{aligned}
$$

The lemma follows by a union bound. □

Lemma A.6.

$$\mathbb{P}\left[|g(x)| \ge 1\right] = \Omega(\exp(-\Theta(k))).$$

Proof. For large enough $T = \Omega(k)$, it holds from Lemmas A.3 and A.4 that

$$\|g^T\| = \Omega\left(\left(\frac{4}{e}\right)^{(\frac{1}{2}+o(1))k}\right).$$

Since $\left|g^T(x)\right| \le T\,2^k$,

$$\mathbb{E}[g^T(x)^2] \le 1\cdot\mathbb{P}[|g^T(x)| \le 1] + (T2^k)^2\cdot\mathbb{P}[|g^T(x)| \ge 1],$$

so that

$$
\begin{aligned}
\mathbb{P}\left[\left|g^T(x)\right| \ge 1\right] &= \frac{\Omega\Big(\left(\frac{4}{e}\right)^{(1+o(1))k}\Big) - 1}{(T\,2^k)^2} \\
&= \Omega(\exp(-\Theta(k)))
\end{aligned}
$$

Coupled with Lemma A.5, this means

$$\mathbb{P}\left[|g(x)| \ge 1\right] = \Omega(\exp(-\Theta(k)))$$

for large enough $T = \Omega(k)$. □

The lower bound on $\|f\|$ now follows easily.

Corollary A.7.

$$\|f\| = \Omega(\exp(-\Theta(k))).$$

Proof. Since $f = \psi\circ g$, from Lemma A.6 and the fact that ψ is odd and increasing, we have that

$$
\begin{aligned}
\|f\| &\ge |\psi(1)|\,\mathbb{P}[g(x) \ge 1] + |\psi(-1)|\,\mathbb{P}[g(x) \ge 1] \\
&= \psi(1)\,\mathbb{P}[|g(x)| \ge 1] \\
&= \Omega(\exp(-\Theta(k))).
\end{aligned}
$$

□

A.2. Sigmoid Activation

Here we consider g and f with $\phi(x) = \sigma(x) = \frac{1}{1+e^{-x}}$. For the asymptotic bound of Hermite polynomial coefficients, we need the following theorem from (Boyd, 1984).

Theorem A.8. *For a function $f(z)$ whose convergence is limited by simple poles at the roots of $z^2 = -\gamma^2$ with residue R, the non-zero expansion coefficients $\{a_n\}$ of $f(z)$ as a series of normalized Hermite functions have magnitudes asymptotically given by*

$$|a_n| \sim 2^{\frac{5}{4}}\,\pi^{\frac{1}{2}}\,R\,n^{-\frac{1}{4}}\,e^{-\gamma(2n+1)^{\frac{1}{2}}},$$

Here the normalized Hermite function $\{\psi_n(x)\}_{n\in\mathbb{N}}$ is defined by

$$\psi_n(z) = e^{-\frac{z^2}{2}}\pi^{-\frac{1}{4}}\tilde{H}_n(\sqrt{2}z).$$

Applying this to $f(x) = e^{-\frac{x^2}{2}}\sigma(\sqrt{2}x)$ and translating the Hermite coefficients for the series in terms of Hermite functions to those in terms of Hermite polynomials, we have

Lemma A.9.

$$\sigma(x) = \sum_{i=0}^{\infty} c_i\tilde{H}_i(x),$$

where $c_0 = 0.5, c_{2i} = 0$ for $i \ge 1$ and all non-zero odd terms satisfies

$$c_{2i-1} = e^{-\Theta(\sqrt{i})}.$$

Corollary A.10. *There is an infinite increasing sequence $\{k_i\}_{i\in\mathbb{N}}$ such that k_i's are all odd and*

$$c_{k_i} = e^{-\Theta(\sqrt{k_i})}.$$

Proof. It follows simply from the fact that σ is not a polynomial and there should be infinitely many non-zero terms in $\{c_k\}_{k\in\mathbb{N}}$. □

Remark A.11. Experimental evidence strongly indicates that in fact all odd Hermite coefficients of sigmoid are nonzero and decay as above, but this is laborious to formally establish. So we state our norm lower bound only for $k \in \{k_i\}_{i \in \mathbb{N}}$ (and the associated $n \in \{2^{k_i}\}_{i \in \mathbb{N}}$, since we end up taking $k = \log n$). Since this is nevertheless an infinite sequence, it still establishes that no better asymptotic bound holds.

Similar to Lemma A.3, we can derive a lower bound of $\|g\|$ for some k's.

Lemma A.12. *For $k \in \{k_i\}_{i \in \mathbb{N}}$,*

$$\|g(x)\| = \Omega\left(\left(\frac{4}{e}\right)^{(\frac{1}{2}+o(1))k}\right).$$

Proof. Due to Lemma A.1,

$$\begin{aligned}\mathbb{E}\left[g(x)^2\right] &= 4^k \sum_{i \geq 0} \frac{c_i^2}{k^i} \sum_{\substack{i_1+\cdots+i_k=i \\ i_1,\cdots,i_k, \text{ are odd}}} \binom{i}{i_1 \cdots i_k} \\ &\geq \frac{4^k c_k^2}{k^k} \sum_{\substack{i_1+\cdots+i_k=k \\ i_1,\cdots,i_k, \text{ are odd}}} \binom{k}{i_1 \cdots i_k} \\ &\geq \frac{4^k c_k^2 k!}{k^k}.\end{aligned}$$

Using Stirling's approximation,

$$n! \geq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

and Corollary A.10,

$$c_k = e^{-\Theta(\sqrt{k})},$$

we obtain

$$\mathbb{E}\left[g(x)^2\right] = \Omega\left(\frac{4^k\sqrt{2\pi k}}{k^k}\left(\frac{k}{e}\right)^k e^{-\Theta(\sqrt{k})}\right)$$

and hence

$$\mathbb{E}\left[g(x)^2\right] = \Omega\left(\left(\frac{4}{e}\right)^{(1+o(1))k}\right).$$

□

Lemma A.13. *For $k \in \{k_i\}_{i \in \mathbb{N}}$,*

$$\mathbb{P}\left(|g(x)| \geq 1\right) = \Omega(\exp(-\Theta(k))).$$

Proof. Since $|g(x)| \leq 2^k$,

$$\|g\|^2 = \mathbb{E}[g(x)^2] \leq 1 \cdot \mathbb{P}[|g(x)| \leq 1] + (2^k)^2 \cdot \mathbb{P}[|g(x)| \geq 1],$$

and so

$$\mathbb{P}\left(|g(x)| \geq 1\right) = \frac{\Omega\left(\left(\frac{4}{e}\right)^{(1+o(1))k}\right) - 1}{(2^k)^2}.$$

The lemma then follows. □

Using the same argument as Corollary A.7, we have the following bound.

Corollary A.14.

$$\|f\| = \Omega(\exp(-\Theta(k))).$$

A.3. General activations

It is not hard to see that the norm analysis of ReLU and sigmoid extends to any activation function for which a suitable lower bound on the Hermite coefficients holds, and which is either bounded or grows at a polynomial rate, so that under the standard Gaussian it behaves essentially identically to its truncated form. In particular, a lower bound of α^{-j} for any constant $\alpha < 4/e$ on the j^{th} Hermite coefficient suffices to give $\|g\| \geq \exp(\Theta(k))$, by the same argument as in Lemma A.3 and Lemma A.12. This then suffices to give $\|f\| \geq \exp(-\Theta(k))$, as above.

In fact, even a very weak lower bound on $\|f\|$ yields *some* superpolynomial bound on learning. Suppose we only had $\|f\| \geq 1/\exp(\exp(\Theta(k)))$, for instance. Then we can take $k = \log\log n$ and have $\|f\| \geq 1/\operatorname{poly}(n)$ and still obtain a lower bound of $n^{\log\log n} = n^{\omega(1)}$ (see Theorem 3.9). Any lower bound on $\|f\|$ will be a function only of k, so a similar argument applies.

B. SQ lower bound for real-valued functions proof

We give a self-contained variant of the elegant proof of (Szörényi, 2009) for the reader's convenience. For simplicity, we include the 0 function in our class $\mathcal{C}$ — this can only negligibly change the SDA, and it makes the core argument cleaner.

Theorem B.1. *Let D be a distribution on X, and let $\mathcal{C}$ be a real-valued concept class over a domain X such that $0 \in \mathcal{C}$, and $\|c\|_D > \epsilon$ for all $c \in \mathcal{C}, c \neq 0$. Consider any SQ learner that is allowed to make only inner product queries to an SQ oracle for the labeled distribution D_c for some unknown $c \in \mathcal{C}$. Let $d = \mathrm{SDA}_D(\mathcal{C}, \gamma)$. Then any such SQ learner needs at least $d/2$ queries of tolerance $\sqrt{\gamma}$ to learn $\mathcal{C}$ up to L_2 error ϵ.*

Proof. Consider the adversarial strategy where we respond to every query $h : X \to \mathbb{R}$ ($\|h\|_D \leq 1$) with 0. This corresponds to the true expectation if the target were the 0 function. By the norm lower bound, outputting any other c would then mean L_2 error greater than ϵ. Thus we must rule out all other $c \in \mathcal{C}$.

If h_k is the k^{th} query, let $S_k = \{c \in \mathcal{C} \mid \langle c, h_k \rangle_D > \tau\}$ be the functions ruled out by our response of 0. Let $\Phi = \langle h_k, \sum_{c \in S_k} c \rangle_D$. Take $\tau = \sqrt{\gamma}$, and we claim that $|S_k| \geq$

$|\mathcal{C}|/d$. Suppose not, then $\rho_D(S_k) \leq \gamma$ by Definition 2.4.

$$\begin{aligned}\Phi &\leq \|h_k\|_D \left\| \sum_{c \in S_k} c \right\|_D \\ &\leq \sqrt{\sum_{c,c' \in S_k} \langle c, c' \rangle_D} \\ &= \sqrt{|S_k|^2 \rho_D(S_k)} \\ &\leq \sqrt{\gamma} |S_k|,\end{aligned}$$

contradicting the fact that $\Phi > |S_k|\tau$ by definition of S_k.

A similar argument holds for $S'_k = \{c \in \mathcal{C} \mid \langle c, h_k \rangle_D < -\tau\}$. Thus we rule out at most a $2/d$ fraction of functions with each query and hence need at least $d/2$ queries. □

C. Further experiments

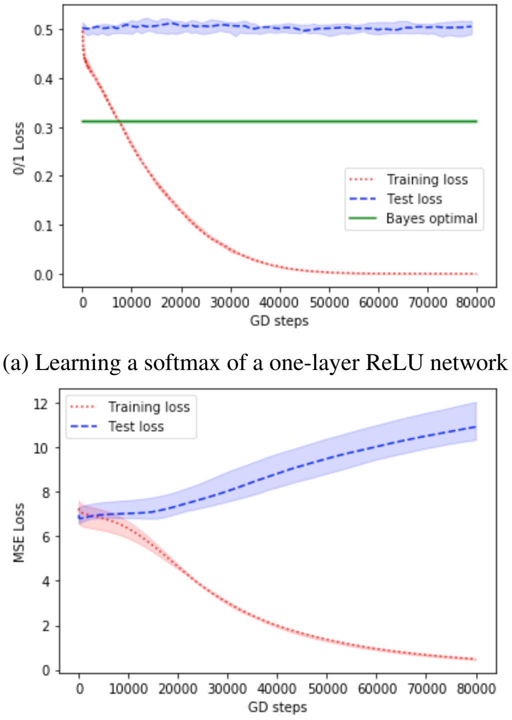
Fig. 1 shows further experiments conducted exactly as in the Experiments section of the main paper, but with the inner activation functions being ReLU instead of tanh.

References

Boyd, J. P. Asymptotic coefficients of hermite function series. *Journal of Computational Physics*, 54(3):382–410, 1984.

Goel, S., Karmalkar, S., and Klivans, A. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems*, pp. 8582–8591, 2019.

Szörényi, B. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pp. 186–200. Springer, 2009.



(a) Learning a softmax of a one-layer ReLU network

(b) Learning a linear combination of ReLUs

Figure 1. In (a) the target function is a softmax (±1 labels) of a sum of 2^8 ReLU activations with $n = 14$; in (b) the labels are obtained similarly but without the softmax. In both cases, we train a 1-layer neural network with $5 \cdot 2^8 = 1280$ ReLU units (hence 20481 parameters) using a training set of size 6000 and a test set of size 1000, withe the learning rate set to 0.005 for classification and 0.002 for regression. For (a) we take the sign of this trained network and measure its training and testing 0/1 loss; for (b) we measure the train and test square loss of the learned network directly. In (a) we also plot the test error of the bayes optimal network (sign of the target function).