Supplementary Materials

A Technical lemmas for Theorem 1

In this appendix, we prove technical lemmas that appear in the proof of Theorem 1.

A.1 Proof of Lemma 1

The following inequality always holds:

$$\sup_{f\in\mathcal{F}} |G(f)-\ell(f)| \le \max\Big\{ \sup_{f\in\mathcal{F}}\{G(f)-\ell(f)\}, \sup_{f'\in\mathcal{F}}\{\ell(f')-G(f')\}\Big\}.$$

Since $\mathcal{F}$ contains the constant zero function, both $\sup_{f\in\mathcal{F}}\{G(f)-\ell(f)\}$ and $\sup_{f'\in\mathcal{F}}\{\ell(f')-G(f')\}$ are non-negative, which implies

$$\sup_{f\in\mathcal{F}} |G(f)-\ell(f)| \le \sup_{f\in\mathcal{F}}\{G(f)-\ell(f)\} + \sup_{f'\in\mathcal{F}}\{\ell(f')-G(f')\}.$$

To establish Lemma 1, it suffices to prove:

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\{G(f)-\ell(f)\}\Big] \le 2LR_k(\mathcal{F}) \quad \text{and} \quad \mathbb{E}\Big[\sup_{f'\in\mathcal{F}}\{\ell(f')-G(f')\}\Big] \le 2LR_k(\mathcal{F})$$

For the rest of the proof, we will establish the first upper bound. The second bound can be established through an identical series of steps.

The inequality $\mathbb{E}[\sup_{f\in\mathcal{F}}\{G(f)-\ell(f)\}] \le 2LR_k(\mathcal{F})$ follows as a consequence of classical symmetrization techniques [e.g. Bartlett and Mendelson, 2003] and the Talagrand-Ledoux concentration [e.g. Ledoux and Talagrand, 2013, Corollary 3.17]. However, so as to keep the paper self-contained, we provide a detailed proof here. By the definitions of $\ell(f)$ and $G(f)$, we have

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{G(f)-\ell(f)\Big\}\Big] = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\frac{1}{k}\sum_{j=1}^{k} h(-y'_j f(x'_j)) - \mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^{k} h(-y''_j f(x''_j))\Big]\Big\}\Big],$$

where (x''_j, y''_j) is an i.i.d. copy of (x'_j, y'_j). Applying Jensen's inequality yields

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{G(f)-\ell(f)\Big\}\Big] &\le \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\frac{1}{k}\sum_{j=1}^{k} h(-y'_j f(x'_j)) - h(-y''_j f(x''_j))\Big\}\Big] \\
&= \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\frac{1}{k}\sum_{j=1}^{k} \varepsilon_j (h(-y'_j f(x'_j)) - h(-y''_j f(x''_j)))\Big\}\Big] \\
&\le \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\frac{1}{k}\sum_{j=1}^{k} \varepsilon_j h(-y'_j f(x'_j)) + \sup_{f\in\mathcal{F}}\frac{1}{k}\sum_{j=1}^{k} \varepsilon_j h(-y''_j f(x''_j))\Big\}\Big] \\
&= 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\frac{1}{k}\sum_{j=1}^{k} \varepsilon_j h(-y'_j f(x'_j))\Big\}\Big]. \qquad (13)
\end{aligned}$$

We need to bound the right-hand side using the Rademacher complexity of the function class $\mathcal{F}$, and we use an argument following the lecture notes of Kakade and Tewari [2008]. Introducing the shorthand notation $\varphi_j(x) := h(-y_j'x)$, the L-Lipschitz continuity of φ_j implies that

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j=1}^k \varepsilon_j\varphi_j(f(x_j'))\Big] &= \mathbb{E}\Big[\sup_{f,f'\in\mathcal{F}}\Big\{\frac{\varphi_1(f(x_1'))-\varphi_1(f'(x_1'))}{2}+\sum_{j=2}^k\varepsilon_j\frac{\varphi_j(f(x_j'))+\varphi_j(f'(x_j'))}{2}\Big\}\Big]\\
&\le \mathbb{E}\Big[\sup_{f,f'\in\mathcal{F}}\Big\{\frac{L|f(x_1')-f'(x_1')|}{2}+\sum_{j=2}^k\varepsilon_j\frac{\varphi_j(f(x_j'))+\varphi_j(f'(x_j'))}{2}\Big\}\Big]\\
&= \mathbb{E}\Big[\sup_{f,f'\in\mathcal{F}}\Big\{\frac{Lf(x_1')-Lf'(x_1')}{2}+\sum_{j=2}^k\varepsilon_j\frac{\varphi_j(f(x_j'))+\varphi_j(f'(x_j'))}{2}\Big\}\Big].
\end{aligned}$$

Applying Jensen's inequality implies that the right-hand side is bounded by

$$\begin{aligned}
\text{RHS} &\le \frac{1}{2}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{Lf(x_1')+\sum_{j=2}^k\varepsilon_j\varphi_j(f(x_j'))\Big\}+\sup_{f'\in\mathcal{F}}\Big\{-Lf(x_1')+\sum_{j=2}^k\varepsilon_j\varphi_j(f'(x_j'))\Big\}\Big]\\
&= \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\varepsilon_1 Lf(x_1')+\sum_{j=2}^k\varepsilon_j\varphi_j(f(x_j'))\Big\}\Big].
\end{aligned}$$

By repeating this argument for $j = 2, 3, \dots, k$, we obtain

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j=1}^k\varepsilon_j\varphi_j(f(x_j'))\Big] \le L\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j=1}^k\varepsilon_j f(x_j')\Big]. \tag{14}$$

Combining inequalities (13) and (14), we have the desired bound.

A.2 Proof of Lemma 2

We prove the claim by induction on the number of layers m. It is known [Kakade et al., 2009] that $R_k(\mathcal{N}_1) \le \sqrt{\frac{q}{k}}\,B$. Thus, the claim holds for the base case $m = 1$. Now consider some $m > 1$, and assume that the claim holds for $m-1$. We then have

$$R_k(\mathcal{N}_1) = \mathbb{E}\left[\sup_{f\in\mathcal{N}_m}\frac{1}{k}\sum_{i=1}^k\varepsilon_i f(x_i')\right],$$

where $\varepsilon_1, \dots, \varepsilon_n$ are Rademacher variables. By the definition of $\mathcal{N}_m$, we may write the expression as

$$\begin{aligned}
R_k(\mathcal{N}_1) = \mathbb{E}\left[\sup_{f_1,\dots,f_d\in\mathcal{N}_{m-1}}\frac{1}{k}\sum_{i=1}^n\varepsilon_i\sum_{j=1}^d w_j\sigma(f_j(x_i'))\right] &= \mathbb{E}\left[\sup_{f_1,\dots,f_d\in\mathcal{N}_{m-1}}\frac{1}{k}\sum_{j=1}^d w_j\sum_{i=1}^k\varepsilon_i\sigma(f_j(x_i'))\right]\\
&\le B\mathbb{E}\left[\sup_{f\in\mathcal{N}_{m-1}}\frac{1}{k}\sum_{i=1}^k\varepsilon_i\sigma(f(x_i'))\right]\\
&= BR_k(\sigma\circ\mathcal{N}_{m-1}),
\end{aligned}$$

where the inequality follows since $\|w\|_1 \le B$. Since the function σ is 1-Lipschitz continuous, following the proof of inequality (14), we have

$$R_k(\sigma\circ\mathcal{N}_{m-1}) \le R_k(\mathcal{N}_{m-1}) \le \sqrt{\frac{q}{n}}\,B^m,$$

which completes the proof.

A.3 Proof of Lemma 3

We prove the claim by induction on the number of layers m. If $m = 1$, then f^* is a linear function and $\varphi(f^*) \in [-B_1, B_1]^n$. Since $\varphi(g)$ minimizes the ℓ_2-distance to vector u, we have

$$\|\varphi(g) - \varphi(f^*)\|_2 \leq \|\varphi(g) - u\|_2 + \|\varphi(f^*) - u\|_2 \leq 2\|\varphi(f^*) - u\|_2. \tag{15}$$

Since u is drawn uniformly from $[-B, B]^k$, with probability at least $(\frac{\epsilon}{4})^k$ we have $\|\varphi(f^*) - u\|_\infty \leq \frac{\epsilon B}{2}$, and consequently

$$\|\varphi(g) - \varphi(f^*)\|_2 \leq \sqrt{k}\|\varphi(g) - \varphi(f^*)\|_\infty \leq \epsilon\sqrt{k}B,$$

which establishes the claim.

For $m > 1$, assume that the claim holds for $m - 1$. Our proof uses the following lemma:

Lemma 4 (Maurey-Barron-Jones lemma) *Consider any subset G of any Hilbert space H such that $\|g\|_H \leq b$ for all $g \in G$. Then for any point v is in the convex hull of G, there is a point v_s in the convex hull of s points of G such that $\|v - v_s\|_H^2 \leq b^2/s$.*

See the paper by Pisier [1980] for a proof.

Recall that f^*/B is in the convex hull of $\sigma \circ \mathcal{N}_{m-1}$ and every function $f \in \sigma \circ \mathcal{N}_{m-1}$ satisfies $\|\varphi(f)\|_2 \leq \sqrt{k}$. By Lemma 4, there exist s functions in $\mathcal{N}_{m-1}$, say $\widetilde{f}_1, \ldots, \widetilde{f}_s$, and a vector $w \in \mathbb{R}^s$ satisfying $\|w\|_1 \leq B$ such that

$$\Big\|\sum_{j=1}^s w_j\sigma(\varphi(\widetilde{f}_j)) - \varphi(f^*)\Big\|_2 \leq B\sqrt{\frac{k}{s}}.$$

Let $\varphi(\widetilde{f}) := \sum_{j=1}^s w_j\sigma(\varphi(\widetilde{f}_j))$. If we chose $s = \lceil \frac{1}{\epsilon^2} \rceil$, then we have

$$\|\varphi(\widetilde{f}) - \varphi(f^*)\|_2 \leq \epsilon\sqrt{k}B. \tag{16}$$

Recall that the function g satisfies $g = \sum_{j=1}^s v_j\sigma \circ g_j$ for $g_1, \ldots, g_s \in \mathcal{N}_{m-1}$. Using the inductive hypothesis, we know that the following bound holds with probability at least p_{m-1}^s:

$$\|\sigma(\varphi(g_j)) - \sigma(\varphi(\widetilde{f}_j))\|_2 \leq \|\varphi(g_j) - \varphi(\widetilde{f}_j)\|_2 \leq (2m-3)\epsilon\sqrt{k}B^{m-1} \quad \text{for any } j \in [s].$$

As a consequence, we have

$$\begin{aligned}
\Big\|\sum_{j=1}^s w_j\sigma(\varphi(g_j)) - \sum_{j=1}^s w_j\sigma(\varphi(\widetilde{f}_j))\Big\|_2 &\leq \sum_{j=1}^s |w_j| \cdot \|\sigma(\varphi(g_j)) - \sigma(\varphi(\widetilde{f}_j))\|_2 \\
&\leq \|w\|_1 \cdot \max_{j\in[s]}\{\|\sigma(\varphi(g_j)) - \sigma(\varphi(\widetilde{f}_j))\|_2\} \leq (2m-3)\sqrt{k}\epsilon B^m.
\end{aligned} \tag{17}$$

Finally, we bound the distance between $\sum_{j=1}^s w_j\sigma(\varphi(g_j))$ and $\varphi(g)$. Following the proof of inequality (15), we obtain

$$\Big\|\varphi(g) - \sum_{j=1}^s w_j\sigma(\varphi(g_j))\Big\|_2 \leq 2\Big\|u - \sum_{j=1}^s w_j\sigma(\varphi(g_j))\Big\|_2.$$

Note that $\sum_{j=1}^s w_j\sigma(\varphi(g_j)) \in [-B, B]^k$ and u is uniformly drawn from $[-B, B]^k$. Thus, with probability at least $(\frac{\epsilon}{4})^k$, we have

$$\Big\|\varphi(g) - \sum_{j=1}^s w_j\sigma(\varphi(g_j))\Big\|_2 \leq \epsilon\sqrt{k}B. \tag{18}$$

Combining inequalities (16), (17) and (18) and using the fact that $B \geq 1$, we have

$$\left\|\varphi(g) - \varphi(f^*)\right\|_\infty \leq (2m-1)\epsilon\sqrt{k}B^m,$$

with probability at least

$$p_{m-1}^s \cdot \left(\frac{\epsilon}{4}\right)^k = \left(\frac{\epsilon}{4}\right)^{k\left(\frac{s(s^{m-1}-1)}{s-1}+1\right)} = \left(\frac{\epsilon}{4}\right)^{k(s^m-1)/(s-1)} = p_m,$$

which completes the induction.

B Proof of Theorem 2

Proof of Part (a)

We first prove $\widehat{f} \in \mathcal{N}_m$. Indeed, the definition of b_T implies

$$\sum_{t=1}^{T} \frac{B}{2b_T} \left|\log(\frac{1-\mu_t}{1+\mu_t})\right| \leq B, \tag{19}$$

Notice that $\widehat{f} = \sum_{t=1}^{T} \frac{B}{2b_T} \log(\frac{1-\mu_t}{1+\mu_t})\widehat{g}_t$, where $\widehat{g}_t \in \mathcal{N}_{m-1}$. Thus, combining inequality (19) with the definition of $\mathcal{N}_m$ implies $\widehat{f} \in \mathcal{N}_m$. The time complexity bound is obtained by plugging in the bound from Theorem 1.

It remains to establish the correctness of $\widehat{f}$. We may write any function $f \in \mathcal{N}_m$ as

$$f(x) = \sum_{j=1}^{d} w_j \sigma(f_j(x)) \quad \text{where } w_j \geq 0 \text{ for all } j \in [d].$$

The constraints $w_j \geq 0$ are always satisfiable, otherwise since σ is an odd function we may write $w_j\sigma(f_j(x))$ as $(-w_j)\sigma(-f_j(x))$ so that it satisfies the constraint. The function f_j or $-f_j$ belongs to the class $\mathcal{N}_{m-1}$. We use the following result by Shalev-Shwartz and Singer [2010]: Assume that there exists $f^* \in \mathcal{N}_m$ which separate the data with margin γ. Then for any set of non-negative importance weights $\{\alpha_i\}_{i=1}^n$, there is a function $f \in \mathcal{N}_{m-1}$ such that $\sum_{i=1}^n \alpha_i \sigma(-y_i f(x_i)) \leq -\frac{\gamma}{B}$. This implies that, for every $t \in [T]$, there is $f \in \mathcal{N}_{m-1}$ such that

$$G_t(f) = \sum_{i=1}^{n} \alpha_{t,i} \sigma(-y_i f(x_i)) \leq -\frac{\gamma}{B}.$$

Hence, with probability at least $1-\delta$, the sequence $\mu_1, \ldots, \mu_T$ satisfies the relation

$$\mu_t = G_t(\widehat{g}_t) \leq -\frac{\gamma}{2B} \quad \text{for every } t \in [T]. \tag{20}$$

Algorithm 2 is based on running AdaBoost for T iterations. The analysis of AdaBoost Schapire and Singer [1999] guarantees that for any $\beta > 0$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} e^{-\beta}\mathbb{I}[-y_i f_T(x_i) \geq -\beta] &\leq \frac{1}{n}\sum_{i=1}^{n} e^{-y_i f_T(x_i)} \\ &\leq \exp\Big(-\frac{\sum_{t=1}^T \mu_t^2}{2}\Big).\end{aligned}$$

Thus, the fraction of data that cannot be separated by f_T with margin β is bounded by $\exp(\beta - \frac{\sum_{t=1}^T \mu_t^2}{8B^2})$. If we choose

$$\beta := \frac{\sum_{t=1}^T \mu_t^2}{2} - \log(n+1),$$

then this fraction is bounded by $\frac{1}{n+1}$, meaning that all points are separated with margin β. Recall that $\widehat{f}$ is a scaled version of f_T. As a consequence, all points are separated by $\widehat{f}$ with margin

$$\frac{B\beta}{b_T} = \frac{\sum_{t=1}^{T} \mu_t^2 - 2\log(n+1)}{\frac{1}{B}\sum_{t=1}^{T} \log(\frac{1-\mu_t}{1+\mu_t})}.$$

Since $\mu_t \geq -1/2$, it is easy to verify that $\log(\frac{1-\mu_t}{1+\mu_t}) \leq 4|\mu_t|$. Using this fact and Jensen's inequality, we have

$$\frac{B\beta}{b_T} \geq \frac{(\sum_{t=1}^{T} |\mu_t|)^2/T - 2\log(n+1)}{\frac{4}{B}\sum_{t=1}^{T} |\mu_t|}.$$

The right-hand side is a monotonically increasing function of $\sum_{t=1}^{T} |\mu_t|$. Plugging in the bound in (20), we find that

$$\frac{B\beta}{b_T} \geq \frac{\gamma^2 T/(4B^2) - 2\log(n+1)}{2\gamma T/B^2}.$$

Plugging in $T = \frac{16B^2\log(n+1)}{\gamma^2}$, some algebra shows that the right-hand side is equal to $\gamma/16$ which completes the proof.

Proof of Part (b)

Consider the empirical loss function

$$\ell(f) := \frac{1}{n}\sum_{i=1}^{n} h(-y_i f(x_i)),$$

where $h(t) := \max\{0, 1 + 16t/\gamma\}$. Part (a) implies that $\ell(\widehat{f}) = 0$ with probability at least $1-\delta$. Note that h is $(16/\gamma)$-Lipschitz continuous; the Rademacher complexity of $\mathcal{N}_m$ with respect to n i.i.d. samples is bounded by $\sqrt{q/n}B^m$ (see Lemma 2). By the classical Rademacher generalization bound [Bartlett and Mendelson, 2003, Theorem 8 and Theorem 12], if (x, y) is randomly sampled form $\mathbb{P}$, then with probability at least $1-\delta$ we have

$$\mathbb{E}[h(-y\widehat{f}(x))] \leq \ell(\widehat{f}) + \frac{32B^m}{\gamma}\cdot\sqrt{\frac{q}{n}} + \sqrt{\frac{8\log(2/\delta)}{n}}.$$

Thus, in order to bound the generalization loss by ϵ with probability $1-2\delta$, it suffices to choose $n = \mathrm{poly}(1/\epsilon, \log(1/\delta))$. Since $h(t)$ is an upper bound on the zero-one loss $\mathbb{I}[t \geq 0]$, we obtain the claimed bound. ■

C Proof of Corollary 1

The first step is to use the improper learning algorithm [Zhang et al., 2015, Algorithm 1] to learn a predictor $\widehat{g}$ that minimizes the following risk function:

$$\ell(g) := \mathbb{E}[\phi(-\widetilde{y}g(x))] \quad \text{where} \quad \phi(t) := \begin{cases} -\frac{2\eta}{1-2\eta} + \frac{\eta(t+\gamma)}{(1-\eta)(1-2\eta)\gamma} & \text{if } t \leq -\gamma, \\ -\frac{2\eta}{1-2\eta} + \frac{t+\gamma}{(1-2\eta)\gamma} & \text{if } t > -\gamma. \end{cases}$$

Since $\eta < 1/2$, the function ϕ is convex and Lipschitz continuous. The activation function $\mathrm{erf}(x)$ satisfies the condition of [Zhang et al., 2015, Theorem 1]. Thus, with sample complexity $\mathrm{poly}(1/\tau, \log(1/\delta))$ and time complexity $\mathrm{poly}(d, 1/\tau, \log(1/\delta))$, the resulting predictor $\widehat{g}$ satisfies

$$\ell(\widehat{g}) \leq \ell(f^*) + \tau \quad \text{with probability at least } 1 - \delta/3.$$

By the definition of $\widetilde{y}$ and ϕ, it is straightforward to verify that

$$\ell(g) = \mathbb{E}[(1-\eta)\phi(-yg(x)) + \eta\phi(yg(x))] = \mathbb{E}[\psi(-yg(x))] \tag{21}$$

where

$$\psi(t) := \begin{cases} 0 & \text{if } t < -\gamma, \\ 1 + t/\gamma & \text{if } -\gamma \le t \le \gamma, \\ 2 + \frac{2\eta^2 - 2\eta + 1}{(1-\eta)(1-2\eta)\gamma}(t-\gamma) & \text{if } t > \gamma. \end{cases}$$

Recall that $yf^*(x) \ge \gamma$ almost surely. From the definition of ψ, we have $\ell(f^*) = 0$, so that $\ell(\widehat{g}) \le \ell(f^*) + \tau$ implies $\ell(\widehat{g}) \le \tau$. Also note that $\psi(t)$ upper bounds the indicator $\mathbb{I}[t \ge 0]$, so that the right-hand side of equation (21) provides an upper bound on the probability $\mathbb{P}(\mathrm{sign}(g(x)) \neq y)$. Consequently, defining the classifier $\widehat{h}(x) := \mathrm{sign}(g(x))$, then we have

$$\mathbb{P}(\widehat{h}(x) \neq y) \le \ell(\widehat{g}) \le \tau \quad \text{with probability at least } 1 - \delta/3.$$

Given the classifier $\widehat{h}$, we draw another random dataset of n points taking the form $\{(x_i, y_i)\}_{i=1}^n$. If $\tau = \frac{\delta}{3n}$, then this dataset is equal to $\{(x_i, \widehat{h}(x_i))\}_{i=1}^n$ with probability at least $1 - 2\delta/3$. Let the BoostNet algorithm take $\{(x_i, \widehat{h}(x_i))\}_{i=1}^n$ as its input. With sample size $n = \mathrm{poly}(1/\epsilon, \log(1/\delta))$, Theorem 2 implies that the algorithm learns a neural network $\widehat{f}$ such that $\mathbb{P}(\mathrm{sign}(\widehat{f}(x)) \neq y) \le \epsilon$ with probability at least $1-\delta$. Plugging in the assignments of n and τ, the overall sample complexity is $\mathrm{poly}(1/\epsilon, 1/\delta)$ and the overall computation complexity is $\mathrm{poly}(d, 1/\epsilon, 1/\delta)$.

D Proof of Proposition 1

The following MAX-2-SAT problem is known to be NP-hard [Papadimitriou and Yannakakis, 1991].

Definition 1 (MAX-2-SAT) *Given n literals $\{z_1, \ldots, z_n\}$ and d clauses $\{c_1, \ldots, c_d\}$. Each clause is the conjunction of two arguments that may either be a literal or the negation of a literal* *. *The goal is to determine the maximum number of clauses that can be simultaneously satisfied by an assignment.*

We consider the loss function:

$$\ell(w) := -\frac{1}{n}\sum_{i=1}^n \max\{0, \langle w, x_i\rangle\} = \frac{1}{n}\sum_{i=1}^n \min\{0, \langle w, -x_i\rangle\}. \tag{22}$$

It suffices to prove that: it is NP-hard to compute a vector $\widehat{w} \in \mathbb{R}^d$ such that $\|\widehat{w}\|_2 \le 1$ and

$$\ell(\widehat{w}) \le \ell(w^*) + \frac{1}{(2n+2)d}, \tag{23}$$

To prove this claim, we reduce MAX-2-SAT to the minimization problem. Given a MAX-2-SAT instance, we construct a loss function ℓ so that if any algorithm computes a vector $\widehat{w}$ satisfying inequality (23), then the vector $\widehat{w}$ solves MAX-2-SAT.

First, we construct $n+1$ vectors in $\mathbb{R}^d$. Define the vector $x_0 := \frac{1}{\sqrt{d}}\mathbf{1}_d$, and for $i = 1, \ldots, n$, the vectors $x_i := \frac{1}{\sqrt{d}}x'_i$, where $x'_i \in \mathbb{R}^d$ is given by

$$x'_{ij} = \begin{cases} 1 & \text{if } z_i \text{ appears in } c_j, \\ -1 & \text{if } \neg z_i \text{ appears in } c_j, \\ 0 & \text{otherwise.} \end{cases}$$

*In the standard MAX-2-SAT setup, each clause is the disjunction of two literals. However, any disjunction clause can be reduced to three conjunction clauses. In particular, a clause $z_1 \vee z_2$ is satisfied if and only if one of the following is satisfied: $z_1 \wedge z_2$, $\neg z_1 \wedge z_2$, $z_1 \wedge \neg z_2$.

It is straightforward to verify that that $\|x_i\|_2 \le 1$ for any $i \in \{0, 1, \dots, n\}$. We consider the following minimization problem which is special case of the formulation (22):

$$\ell(w) = \frac{1}{2n+2} \sum_{i=0}^{n} \Big(\min\{0, \langle w, x_i \rangle\} + \min\{0, \langle w, -x_i \rangle\} \Big).$$

The goal is to find a vector $w^* \in \mathbb{R}^d$ such that $\|w^*\|_2 \le 1$ and it minimizes the function $\ell(w)$.

Notice that for every index i, at most one of $\min\{0, \langle w, x_i \rangle\}$ and $\min\{0, \langle w, -x_i \rangle\}$ is non-zero. Thus, we may write the minimization problem as

$$\begin{aligned} \min_{\|w\|_2 \le 1} (2n+2)\ell(w) = \min_{\|w\|_2 \le 1} \sum_{i=0}^{n} \Big(\min_{\alpha_i \in \{-1,1\}} \langle w, \alpha_i x_i \rangle \Big) &= \min_{\alpha \in \{-1,1\}^{n+1}} \min_{\|w\|_2 \le 1} \sum_{i=0}^{n} \langle w, \alpha_i x_i \rangle \\ &= \min_{\alpha \in \{-1,1\}^{n+1}} - \Big\| \sum_{i=0}^{n} \alpha_i x_i \Big\|_2 \\ &= - \left(\max_{\alpha \in \{-1,1\}^{n+1}} \sum_{j=1}^{d} \Big(\sum_{i=0}^{n} \alpha_i x_{ij} \Big)^2 \right)^{1/2}. \end{aligned} \tag{24}$$

We claim that maximizing $\sum_{j=1}^{d} (\sum_{i=0}^{n} \alpha_i x_{ij})^2$ with respect to α is equivalent to maximizing the number of satisfiable clauses. In order to prove this claim, we consider an arbitrary assignment to α to construct a solution to the MAX-2-SAT problem. For $i = 1, 2, \dots, n$, let $z_i = \texttt{true}$ if $\alpha_i = \alpha_0$, and let $z_i = \texttt{false}$ if $\alpha_i = -\alpha_0$. With this assignment, it is straightforward to verify the following: if the clause c_j is satisfied, then the value of $\sum_{i=0}^{n} \alpha_i x_{ij}$ is either $3/\sqrt{d}$ or $-3/\sqrt{d}$. If the clause is not satisfied, then the value of the expression is either $1/\sqrt{d}$ or $-1/\sqrt{d}$. To summarize, we have

$$\sum_{j=1}^{d} \Big(\sum_{i=0}^{n} \alpha_i x_{ij} \Big)^2 = 1 + \frac{8 \times (\text{\# of satisfied clauses})}{d}. \tag{25}$$

Thus, solving problem (24) determines the maximum number of satisfiable clauses:

$$(\text{max \# of satisfied clauses}) = \frac{d}{8} \Big(\big(\min_{\|w\|_2 \le 1} (2n+2)\ell(w) \big)^2 - 1 \Big).$$

By examining equation (24) and (25), we find that the value of $(2n+2)\ell(w)$ ranges in $[-3, 0]$. Thus, the MAX-2-SAT number is exactly determined if $(2n+2)\ell(\widehat{w})$ is at most $1/d$ larger than the optimal value. This optimality gap is guaranteed by inequality (23), which completes the reduction.

E Proof of Proposition 2

We reduce the PAC learning of intersection of T halfspaces to the problem of learning a neural network. Assume that $T = \Theta(d^\rho)$ for some $\rho > 0$. We claim that for any number of pairs taking the form $(x, h^*(x))$, there is a neural network $f^* \in \mathcal{N}_2$ that separates all pairs with margin γ, and moreover that the margin is bounded as $\gamma = 1/\mathrm{poly}(d)$.

To prove the claim, recall that $h^*(x) = 1$ if and only if $h_1(x) = \cdots = h_T(x) = 1$ for some $h_1, \dots, h_T \in H$. For any h_t, the definition of H implies that there is a (w_t, b_t) pair such that if $h_t(x) = 1$ then $w_t^T x - b_t - 1/2 \ge 1/2$, otherwise $w_t^T x - b_t - 1/2 \le -1/2$. We consider the two possible choices of the activation function:

- **Piecewise linear function:** If $\sigma(x) := \min\{1, \max\{-1, x\}\}$, then let

$$g_t(x) := \sigma(c(w_t^T x - b_t - 1/2) + 1),$$

for some quantity $c > 0$. The term inside the activation function can be written as $\langle \widetilde{w}, x' \rangle$ where

$$\widetilde{w} = (c\sqrt{2d+2}w_t, -c\sqrt{2d+2}(b_t + 1/2), \sqrt{2}) \quad \text{and} \quad x' = (\frac{x}{\sqrt{2d+2}}, \frac{1}{\sqrt{2d+2}}, \frac{1}{\sqrt{2}}).$$

Note that $\|x'\|_2 \le 1$, and with a sufficiently small constant $c = 1/\text{poly}(d)$ we have $\|\widetilde{w}\|_2 \le 2$. Thus, $g_t(x)$ is the output of a one-layer neural network. If $h_t(x) = 1$, then $g_t(x) = 1$, otherwise $g_t(x) \le 1 - c/2$. Now consider the two-layer neural network $f(x) := c/4 - T + \sum_{t=1}^T g_t(x)$. If $h^*(x) = 1$, then we have $g_t(x) = 1$ for every $t \in [T]$ which implies $f(x) = c/4$. If $h^*(x) = -1$, then we have $g_t(x) \le 1 - c/2$ for at least one $t \in [T]$ which implies $f(x) \le -c/4$. Thus, the neural network f separates the data with margin $c/4$. We normalize the edge weights on the second layer to make f belong to $\mathcal{N}_2$. After normalization, the network still has margin $1/\text{poly}(d)$.

- **ReLU function:** if $\sigma(x) := \max\{0, x\}$, then let $g_t(x) := \sigma(-c(w_t^T x - b_t - 1/2))$ for some quantity $c > 0$. We may write the term inside the activation function as $\langle \widetilde{w}, x' \rangle$ where $\widetilde{w} = (-c\sqrt{d+1}w_t, c\sqrt{d+1}(b_t + 1/2))$ and $x' = (x, 1)/\sqrt{d+1}$. It is straightforward to verify that $\|x'\|_2 \le 1$, and with a sufficiently small $c = 1/\text{poly}(d)$ we have $\|\widetilde{w}\|_2 \le 2$. Thus, $g_t(x)$ is the output of a one-layer neural network. If $h_t(x) = 1$, then $g_t(x) = 0$, otherwise $g_t(x) \ge c/2$. Let $f(x) := c/4 - \sum_{t=1}^T g_t(x)$, then this two-layer neural network separates the data with margin $c/4$. After normalization the network belongs to $\mathcal{N}_2$ and it still separates the data with margin $1/\text{poly}(d)$.

To learn the intersection of T halfspaces, we learn a neural network based on n i.i.d. points taking the form $(x, h^*(x))$. Assume that the neural network is efficiently learnable. Since there exists $f^* \in \mathcal{N}_m$ which separates the data with margin $\gamma = 1/\text{poly}(d)$, we can learn a network $\widehat{f}$ in $\text{poly}(d, 1/\epsilon, 1/\delta)$ sample complexity and time complexity, and satisfies $\mathbb{P}(\text{sign}(\widehat{f}(x)) \neq h^*(x)) \le \epsilon$ with probability $1 - \delta$. It contradicts with the assumption that the intersection of T halfspaces is not efficiently learnable.

References

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

S. Kakade and A. Tewari. Lecture note: Rademacher composition and linear prediction. 2008.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, volume 21, pages 793–800, 2009.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 2013.

C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of computer and system sciences*, 43(3):425–440, 1991.

G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse Fonctionnelle*, pages 1–12, 1980.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine Learning*, 80(2-3):141–163, 2010.

Y. Zhang, J. D. Lee, and M. I. Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. *arXiv:1510.03528*, 2015.