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:

\[
\sup_{f \in F} |G(f) - \ell(f)| \leq \max \{ \sup_{f \in F} \{G(f) - \ell(f)\}, \sup_{f' \in F} \{\ell(f') - G(f')\}\}.
\]

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

\[
\sup_{f \in F} |G(f) - \ell(f)| \leq \sup_{f \in F} \{G(f) - \ell(f)\} + \sup_{f' \in F} \{\ell(f') - G(f')\}.
\]

To establish Lemma 1, it suffices to prove:

\[
E\left[\sup_{f \in F} \{G(f) - \ell(f)\}\right] \leq 2LR_k(F) \quad \text{and} \quad E\left[\sup_{f' \in F} \{\ell(f') - G(f')\}\right] \leq 2LR_k(F)
\]

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 \(E[\sup_{f \in F} \{G(f) - \ell(f)\}] \leq 2LR_k(F)\) 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

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

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

\[
E\left[\sup_{f \in F} \{G(f) - \ell(f)\}\right] \leq E\left[\sup_{f \in F} \left\{\frac{1}{k} \sum_{j=1}^{k} h(-y'_j f(x'_j)) - h(-y''_j f(x''_j))\right\}\right]
\]

\[
= E\left[\sup_{f \in F} \left\{\frac{1}{k} \sum_{j=1}^{k} \varepsilon_j (h(-y'_j f(x'_j)) - h(-y''_j f(x''_j)))\right\}\right]
\]

\[
\leq E\left[\sup_{f \in F} \left\{\frac{1}{k} \sum_{j=1}^{k} \varepsilon_j h(-y'_j f(x'_j))\right\} + \sup_{f \in F} \frac{1}{k} \sum_{j=1}^{k} \varepsilon_j (-y''_j f(x''_j))\right]\]

\[
= 2E\left[\sup_{f \in F} \left\{\frac{1}{k} \sum_{j=1}^{k} \varepsilon_j h(-y'_j f(x'_j))\right\}\right].
\]

(13)
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 \( \varphi_j(x) := h(-y_j^T x) \), the \( L \)-Lipschitz continuity of \( \varphi_j \) implies that

\[
E \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^{k} \varepsilon_j \varphi_j(f(x'_j)) \right] = E \left[ \sup_{f, f' \in \mathcal{F}} \left\{ \frac{\varphi_1(f(x'_1)) - \varphi_1(f'(x'_1))}{2} + \sum_{j=2}^{k} \varepsilon_j \left( \varphi_j(f(x'_j)) + \varphi_j(f'(x'_j)) \right) \right\} \right]
\]

\[
\leq E \left[ \sup_{f, f' \in \mathcal{F}} \left\{ \frac{L |f(x'_1) - f'(x'_1)|}{2} + \sum_{j=2}^{k} \varepsilon_j \left( \varphi_j(f(x'_j)) + \varphi_j(f'(x'_j)) \right) \right\} \right]
\]

\[
= E \left[ \sup_{f, f' \in \mathcal{F}} \left\{ \frac{L f(x'_1) - L f'(x'_1)}{2} + \sum_{j=2}^{k} \varepsilon_j \left( \varphi_j(f(x'_j)) + \varphi_j(f'(x'_j)) \right) \right\} \right].
\]

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

\[
\text{RHS} \leq \frac{1}{2} E \left[ \sup_{f \in \mathcal{F}} \left\{ L f(x'_1) + \sum_{j=2}^{k} \varepsilon_j \varphi_j(f(x'_j)) \right\} \right] + E \left[ \sup_{f' \in \mathcal{F}} \left\{ - L f(x'_1) + \sum_{j=2}^{k} \varepsilon_j \varphi_j(f'(x'_j)) \right\} \right]
\]

\[
= E \left[ \sup_{f \in \mathcal{F}} \left\{ \varepsilon_1 L f(x'_1) + \sum_{j=2}^{k} \varepsilon_j \varphi_j(f(x'_j)) \right\} \right].
\]

By repeating this argument for \( j = 2, 3, \ldots, k \), we obtain

\[
E \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^{k} \varepsilon_j \varphi_j(f(x'_j)) \right] \leq LE \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^{k} \varepsilon_j f(x'_j) \right]. \tag{14}
\]

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 \( R_k(\mathcal{N}_1) \leq \sqrt{\frac{q}{n}} B \). 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

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

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

\[
R_k(\mathcal{N}_1) = E \left[ \sup_{f_1, \ldots, f_d \in \mathcal{N}_{m-1}} \frac{1}{kd} \sum_{i=1}^{n} \sum_{j=1}^{d} w_j \sigma(f_j(x'_i)) \right] = E \left[ \sup_{f_1, \ldots, f_d \in \mathcal{N}_{m-1}} \frac{1}{k} \sum_{j=1}^{d} w_j \sum_{i=1}^{k} \varepsilon_i \sigma(f_j(x'_i)) \right]
\]

\[
\leq B E \left[ \sup_{f \in \mathcal{N}_{m-1}} \frac{1}{k} \sum_{i=1}^{k} \varepsilon_i \sigma(f(x'_i)) \right] = BR_k(\sigma \circ \mathcal{N}_{m-1}),
\]

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

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

which completes the proof.
As a consequence, we have
\[
\|\varphi(g) - \varphi(f^*)\|_2 \leq \|\varphi(g) - u\|_2 + \|\varphi(f^*) - u\|_2 \leq 2\|\varphi(f^*) - u\|_2.
\] (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 \(v\) 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 \(\sigma \circ N_{m-1}\) and every function \(f \in \sigma \circ N_{m-1}\) satisfies \(\|\varphi(f)\|_2 \leq \sqrt{k}\). By Lemma 4, there exist \(s\) functions in \(N_{m-1}\), say \(\tilde{f}_1, \ldots, \tilde{f}_s\), and a vector \(w \in \mathbb{R}^s\) satisfying \(\|w\|_1 \leq B\) such that
\[
\left\| \sum_{j=1}^{s} w_j \sigma(\varphi(\tilde{f}_j)) - \varphi(f^*) \right\|_2 \leq B \sqrt{\frac{k}{s}}.
\]

Let \(\varphi(\tilde{f}) := \sum_{j=1}^{s} w_j \sigma(\varphi(\tilde{f}_j))\). If we chose \(s = \lceil \frac{1}{\epsilon^2} \rceil\), then we have
\[
\|\varphi(\tilde{f}) - \varphi(f^*)\|_2 \leq \epsilon \sqrt{k}B. \quad (16)
\]

Recall that the function \(g\) satisfies \(g = \sum_{j=1}^{s} v_j \sigma \circ g_j\) for \(g_1, \ldots, g_s \in N_{m-1}\). Using the inductive hypothesis, we know that the following bound holds with probability at least \(p_m^{m-1}\):
\[
\|\sigma(\varphi(g_j)) - \sigma(\varphi(\tilde{f}_j))\|_2 \leq \|\varphi(g_j) - \varphi(\tilde{f}_j)\|_2 \leq (2m - 3)\epsilon \sqrt{k} B^{m-1}\text{ for any } j \in [s].
\]

As a consequence, we have
\[
\left\| \sum_{j=1}^{s} w_j \sigma(\varphi(g_j)) - \sum_{j=1}^{s} w_j \sigma(\varphi(\tilde{f}_j)) \right\|_2 \leq \sum_{j=1}^{s} |w_j| \cdot \|\sigma(\varphi(g_j)) - \sigma(\varphi(\tilde{f}_j))\|_2
\]
\[
\leq \|w\|_1 \cdot \max_{j \in [s]} \{\|\sigma(\varphi(g_j)) - \sigma(\varphi(\tilde{f}_j))\|_2\} \leq (2m - 3)\epsilon \sqrt{k} B^{m}. \quad (17)
\]

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
\[
\left\| \varphi(g) - \sum_{j=1}^{s} w_j \sigma(\varphi(g_j)) \right\|_2 \leq 2\|u - \sum_{j=1}^{s} w_j \sigma(\varphi(g_j))\|_2.
\]

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
\[
\left\| \varphi(g) - \sum_{j=1}^{s} w_j \sigma(\varphi(g_j)) \right\|_2 \leq \epsilon \sqrt{k}B. \quad (18)
\]
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
\[
\Pr \left[ \epsilon \cdot 2^{-m-1} \right] = \left( \frac{\epsilon}{4} \right) \left( \frac{2m-1}{s-1} + 1 \right) = \left( \frac{\epsilon}{4} \right)^{k(s^m-1)/(s-1)} = p_m,
\]
which completes the induction.

\[\textbf{B \ Proof of Theorem 2}\]

\textbf{Proof of Part (a)}

We first prove \( \hat{f} \in \mathcal{N}_m \). Indeed, the definition of \( b_T \) implies
\[
\sum_{t=1}^{T} \frac{B}{2b_T} \left| \log \left( \frac{1 - \mu_t}{1 + \mu_t} \right) \right| \leq B, \tag{19}
\]
Notice that \( \hat{f} = \sum_{t=1}^{T} \frac{B}{2b_T} \log \left( \frac{1 - \mu_t}{1 + \mu_t} \right) \hat{g}_t \), where \( \hat{g}_t \in \mathcal{N}_{m-1} \). Thus, combining inequality (19) with the definition of \( \mathcal{N}_m \) implies \( \hat{f} \in \mathcal{N}_m \). The time complexity bound is obtained by plugging in the bound from Theorem 1.

It remains to establish the correctness of \( \hat{f} \). 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 \( \sigma \) 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 \( \gamma \). 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
\[
G_t(f) = \sum_{i=1}^{n} \alpha_i \sigma(-y_i f(x_i)) \leq -\frac{\gamma}{B}.
\]
Hence, with probability at least \( 1 - \delta \), the sequence \( \mu_1, \ldots, \mu_T \) satisfies the relation
\[
\mu_t = G_t(\hat{g}_t) \leq -\frac{\gamma}{2B} \quad \text{for every } t \in [T]. \tag{20}
\]
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
\[
\frac{1}{n} \sum_{i=1}^{n} e^{-\beta \|[-y_i f_T(x_i)] \geq -\beta\|} \leq \frac{1}{n} \sum_{i=1}^{n} e^{-y_i f_T(x_i)} \leq \exp \left( -\frac{\sum_{t=1}^{T} \mu_t^2}{2} \right).
\]
Thus, the fraction of data that cannot be separated by \( f_T \) with margin \( \beta \) is bounded by \( \exp(\beta - \frac{\sum_{t=1}^{T} \mu_t^2}{2B}) \). 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 $\beta$. Recall that $\hat{f}$ is a scaled version of $f_T$. As a consequence, all points are separated by $\hat{f}$ with margin

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

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

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

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 $\ell(\hat{f}) = 0$ with probability at least $1 - \delta$. Note that $h$ is $(16/\gamma)$-Lipschitz continuous; the Rademacher complexity of $N_m$ with respect to $n$ i.i.d. samples is bounded by $\sqrt{q/nB^m}$ (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 $P$, then with probability at least $1 - \delta$ we have

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

Thus, in order to bound the generalization loss by $\epsilon$ with probability $1 - 2\delta$, it suffices to choose $n = \text{poly}(1/\epsilon, \log(1/\delta))$. Since $h(t)$ is an upper bound on the zero-one loss $\|t \geq 0\|$, 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 $\hat{g}$ that minimizes the following risk function:

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

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

$$
\ell(\hat{g}) \leq \ell(f^*) + \tau \quad \text{with probability at least } 1 - \delta/3.
$$
By the definition of $\tilde{y}$ and $\phi$, it is straightforward to verify that

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

(21)

where

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

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

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

Given the classifier $\hat{h}$, 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 $\{(x_i, \hat{h}(x_i))\}_{i=1}^n$ with probability at least $1 - 2\delta/3$. Let the BoostNet algorithm take $\{(x_i, \hat{h}(x_i))\}_{i=1}^n$ as its input. With sample size $n = \text{poly}(1/\epsilon, \log(1/\delta))$, Theorem 2 implies that the algorithm learns a neural network $\hat{f}$ such that $\mathbb{P}(\text{sign}(\hat{f}(x)) \neq y) \leq \epsilon$ with probability at least $1 - \delta$. Plugging in the assignments of $n$ and $\tau$, the overall sample complexity is $\text{poly}(1/\epsilon, 1/\delta)$ and the overall computation complexity is $\text{poly}(d, 1/\epsilon, 1/\delta)$.

### 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 $\neg$. The goal is to determine the maximum number of clauses that can be simultaneously satisfied by an assignment.

We consider the loss function:

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

(22)

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

$$\ell(\hat{w}) \leq \ell(w^*) + \frac{1}{(2n + 2)d},$$

(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 $\ell$ so that if any algorithm computes a vector $\hat{w}$ satisfying inequality (23), then the vector $\hat{w}$ solves MAX-2-SAT.

First, we construct $n + 1$ vectors in $\mathbb{R}^d$. Define the vector $x_0 := \frac{1}{\sqrt{d}}1_d$, 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: $z_1 \wedge \neg z_2, \neg z_1 \wedge z_2, z_1 \wedge \neg z_2$. 

6
It is straightforward to verify that that $\|x_i\|_2 \leq 1$ for any $i \in \{0, 1, \ldots, n\}$. We consider the following minimization problem which is special case of the formulation (22):

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

The goal is to find a vector $w^* \in \mathbb{R}^d$ such that $\|w^*\|_2 \leq 1$ 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

$$\min_{\|w\|_2 \leq 1} (2n+2)\ell(w) = \min_{\|w\|_2 \leq 1} \sum_{i=0}^{n} \left( \min_{\alpha_i \in \{-1,1\}} \langle w, \alpha_i x_i \rangle \right) = \min_{\alpha \in \{-1,1\}^{n+1}} \min_{\|w\|_2 \leq 1} \sum_{i=0}^{n} \langle w, \alpha_i x_i \rangle$$

$$= \min_{\alpha \in \{-1,1\}^{n+1}} -\left\| \sum_{i=0}^{n} \alpha_i x_i \right\|_2$$

$$= - \left( \max_{\alpha \in \{-1,1\}^{n+1}} \sum_{j=1}^{d} \left( \sum_{i=0}^{n} \alpha_i x_{ij} \right)^2 \right)^{1/2}. \quad (24)$$

We claim that maximizing $\sum_{j=1}^{d} \left( \sum_{i=0}^{n} \alpha_i x_{ij} \right)^2$ with respect to $\alpha$ is equivalent to maximizing the number of satisfiable clauses. In order to prove this claim, we consider an arbitrary assignment to $\alpha$ to construct a solution to the MAX-2-SAT problem. For $i = 1, 2, \ldots, n$, let $z_i = \text{true}$ if $\alpha_i = \alpha_0$, and let $z_i = \text{false}$ 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

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

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

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

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 $(2n+2)\ell(w)$ 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 $\gamma$, and moreover that the margin is bounded as $\gamma = 1/\text{poly}(d)$.

To prove the claim, recall that $h^*(x) = 1$ if and only if $h_1(x) = \cdots = h_T(x) = 1$ for some $h_1, \ldots, h_T \in H$. 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 $w_t^T x - b_t - 1/2 \geq 1/2$, otherwise $w_t^T x - b_t - 1/2 \leq -1/2$. We consider the two possible choices of the activation function:

- **Piecewise linear function:** If $\sigma(x) := \min\{1, \max\{-1, x\}\}$, then let $g_t(x) := \sigma(c(w_t^T x - b_t - 1/2) + 1), \quad (23)$
for some quantity $c > 0$. The term inside the activation function can be written as $\langle \tilde{w}, x' \rangle$ where

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

Note that $\|x'\|_2 \leq 1$, and with a sufficiently small constant $c = 1/\text{poly}(d)$ we have $\|\tilde{w}\|_2 \leq 2$. Thus, $g_t(x)$ is the output of a one-layer neural network. If $h_t(x) = 1$, then $g_t(x) = 1$, otherwise $g_t(x) \leq 1 - c/2$. 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 $g_t(x) \leq 1 - c/2$ for at least one $t \in [T]$ which implies $f(x) \leq -c/4$. 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 $g_t(x) := \sigma(-c(w_t^T x - b_i - 1/2))$ for some quantity $c > 0$. We may write the term inside the activation function as $\langle \tilde{w}, x' \rangle$ where $\tilde{w} = (-c\sqrt{d + 1}w_t, c\sqrt{d + 1}(b_i + 1/2))$ and $x' = (x, 1)/\sqrt{d + 1}$. It is straightforward to verify that $\|x'\|_2 \leq 1$, and with a sufficiently small $c = 1/\text{poly}(d)$ we have $\|\tilde{w}\|_2 \leq 2$. Thus, $g_t(x)$ is the output of a one-layer neural network. If $h_t(x) = 1$, then $g_t(x) = 0$, otherwise $g_t(x) \geq c/2$. 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 $\tilde{f}$ in $\text{poly}(d, 1/\epsilon, 1/\delta)$ sample complexity and time complexity, and satisfies $\mathbb{P}(|\text{sign}(\tilde{f}(x)) - h^*(x)|) \leq \epsilon$ with probability $1 - \delta$. It contradicts with the assumption that the intersection of $T$ halfspaces is not efficiently learnable.

**References**


