Appendix

A Proofs

We introduce some lemmata here, whose proofs can be found in the following sections.

Lemma 6 (Approximation error between sign and tanh)**.** *Under Assumption 1, w.p.* $1 - C_1 r\theta$

$$|\tilde{f}_{W^*,B^*}(\boldsymbol{x}) - f_{W^*,B^*}(\boldsymbol{x})| \le 8re^{-2\gamma\theta}$$

By taking $\theta = \delta/(C_1 r)$*, we have w.p.,* $1-\delta$

$$|\tilde{f}_{W^*,B^*}(\boldsymbol{x}) - f_{W^*,B^*}(\boldsymbol{x})| \le 8re^{-2\gamma\delta/(C_1 r)}$$

Lemma 7 (Lemma 2 in [12])**.** *If* $W \in \mathbb{R}^{r\times d}$ *is a random matrix, whose entries are sampled from* $\mathcal{N}(0,1)$ *i.i.d. and* $\|\boldsymbol{x}\| = \|\boldsymbol{x}'\| = 1$*, then w.p.* $1 - 2e^{-2\epsilon^2 r}$,

$$|\frac{1}{r}\rho_H(sign(W\boldsymbol{x}), sign(W\boldsymbol{x}')) - \rho(\boldsymbol{x},\boldsymbol{x}')| \le \epsilon \tag{18}$$

where $\rho(\cdot,\cdot)$ *is the Euclidean distance.*

Lemma 8 (Covering Spheres with Spheres. Corollary 1.2 in [5])**.** *For any* $0 < \phi \le \arccos(\frac{1}{\sqrt{d+1}})$*, a sphere* $\mathcal{S}^{d-1}$ *can be covered by*

$$\frac{C_2 d^{3/2}}{\sin^{d-1}\phi}\ln(d)$$

spherical balls of radius ϕ*, where* C_2 *is a global constant.*

A.1 Proof of Proposition 1

Proof. We use Lemma 7 to prove this lemma. There are mN data pairs $\{\boldsymbol{x}_i, \boldsymbol{c}_{y,p}\}$ for $i \in [N]$, $y \in \mathcal{Y}$ and $p \in [m_y]$. Then, w.p. $1 - 2mNe^{-2\epsilon^2 r}$, Eq. (18) holds for all the pairs. Set $2mNe^{-2\epsilon^2 r} \le \delta$. Then if $r \ge \frac{\log(2mN/\delta)}{2\epsilon^2}$, for all $i \in [N]$, $y \in \mathcal{Y}$ and $p \in [m_y]$, w.p. $1-\delta$,

$$|\frac{1}{r}\rho_H(\text{sign}(W\boldsymbol{x}_i), \text{sign}(W\boldsymbol{c}_{y,p})) - \rho(\boldsymbol{x}_i, \boldsymbol{c}_{y,p})| \le \epsilon. \tag{19}$$

Setting $\epsilon = \mu/4$ and applying the second assumption completes the proof. □

A.2 Proof of Theorem 1

First, by setting $\alpha = \nu = 32N^{-\frac{1}{16d^2}}$ and $\xi = O(N^{-\frac{1}{32d}})$ for large enough N such that $\xi \le 1/2$, Lemma 3 requires $r \ge \frac{Cd^{3/2}\log d}{2^d}N^{\frac{1}{16d}}$ and $\gamma \ge 16$. Setting $\delta = N^{-\frac{1}{32d}}$, Lemma 1 requires $\gamma \ge C_1(d)N^{\frac{3}{32d}}$ for some constant $C_1(d)$ depending on d. For Lemma 4, we set $t = N^{-\frac{1}{32d}}$. Finally, by setting $\xi = C_2(d)N^{-\frac{1}{32d}}$ for some constant $C_2(d)$ depending d and β, Eq. (13) in Lemma 2 and Eq. (15) in Lemma 3 will hold for $\epsilon = N^{-\frac{1}{32d}}$. By now we have shown that when N goes to ∞, the probabilities of Lemma 2 and Lemma 4 will go to 1 and the errors in the lemmata from Lemma 1 to Lemma 5 will go to zero. So we complete the proof.

A.3 Proof of Lemma 1

Proof.

$$\begin{aligned}&|\mathbb{E}[\mathbb{1}[yf_{W^*,B^*}(\boldsymbol{x})<0]] - \mathbb{E}[\mathbb{1}[y\tilde{f}_{W^*,B^*}(\boldsymbol{x})<0]]| \\ =&|\mathbb{E}[\mathbb{1}[yf_{W^*,B^*}(\boldsymbol{x})<0] - \mathbb{1}[y\tilde{f}_{W^*,B^*}(\boldsymbol{x})<0]]|\end{aligned} \tag{20}$$

Note that $f_{W^*,B^*}(\boldsymbol{x})$ can only take values in $\{\{-2r-\nu, -2r+1-\nu, \cdots, -1-\nu, -\nu, 1-\nu, \cdots, 2r-\nu\}\}$. So if we can show $|f_{W^*,B^*}(\boldsymbol{x}) - \tilde{f}_{W^*,B^*}(\boldsymbol{x})| \le \frac{\nu}{4}$, then $f_{W^*,B^*}(\boldsymbol{x})$ and $\tilde{f}_{W^*,B^*}(\boldsymbol{x})$ will have the same sign, and $\mathbb{1}[yf_{W^*,B^*}(\boldsymbol{x})<0] - \mathbb{1}[y\tilde{f}_{W^*,B^*}(\boldsymbol{x})<0] = 0$.

According to Lemma 6 with $\gamma \ge \frac{C_1 r}{2\delta}\log(\frac{32r}{\nu})$, we have $|f_{W^*,B^*}(\boldsymbol{x}) - \tilde{f}_{W^*,B^*}(\boldsymbol{x})| \le \frac{\nu}{4}$, w.p. at least $1-\delta$.

Therefore, we obtain

$$|\mathbb{E}[\mathbb{1}[yf_{W^*,B^*}(\boldsymbol{x})<0]] - \mathbb{E}[\mathbb{1}[y\tilde{f}_{W^*,B^*}(\boldsymbol{x})<0]]| \le \delta \tag{21}$$

□

A.4 Proof of Lemma 2

Proof. We use the Rademacher complexity to bound this quantity. First, let's apply Theorem 3.1 in [20], given $\epsilon > 0$,

$$\begin{aligned}&\mathbb{P}[\sup_{W,B}|\mathbb{E}[\Phi(y\tilde{f}_{W,B}(\boldsymbol{x}))] - \hat{\mathbb{E}}[\Phi(y\tilde{f}_{W,B}(\boldsymbol{x}))]| \\ &> \mathcal{R}_N(\Phi\circ\mathcal{F}_{W,B}) + \epsilon] \le e^{-N\epsilon^2/C_3},\end{aligned} \tag{22}$$

where $\mathcal{F}_{W,B}$ is the collection of functions formed by $\tilde{f}_{W,B}$ and $\mathcal{R}_N$ is the conditional Rademacher average. Since Φ is $\frac{1}{\alpha\xi}$-Lipschitz and $\tilde{f}_{W,B}$ is $2r\gamma$-Lipschitz, by Lemma 5.2, Lemma 5.4 in [9] ($\tilde{f}_{W,B}$ can be scaled such that the condition of Lemma 5.4 is satisfied) and the Talagrand's contraction lemma [17], we have

$$\begin{aligned}\mathcal{R}_N(\Phi\circ\mathcal{F}_{W,B}) &\le \frac{1}{\alpha\xi}\mathcal{R}_N(\mathcal{F}_{W,B}) \\ &\le \frac{1}{\alpha\xi}\inf_{\epsilon>0}\left(\epsilon + \sqrt{\frac{2(\frac{32r\gamma}{\epsilon})^{2d}\log(8/\epsilon)}{N}}\right) \\ &\le \frac{2^{(2d+3)/(2d+2)}(32r\gamma)^{2d/(2d+2)}}{\alpha\xi N^{1/2d+2}}\sqrt{\log(8/\kappa)},\end{aligned} \tag{23}$$

where $\kappa = \frac{2^{1/(2d+2)}(32r\gamma)^{2d/(2d+2)}}{\alpha\xi N^{1/(2d+2)}}$. As long as $\kappa < \frac{1}{4}$, we have $\sqrt{\log(8/\kappa)} \le 1/\sqrt{\kappa}$. Therefore, $\mathcal{R}_N(\Phi\circ\mathcal{F}_{W,B}) \le 2\sqrt{\kappa}$. We finish the proof by setting $2\sqrt{\kappa} \le \epsilon$. □

A.5 Proof of Lemma 3

Proof. We decompose

$$\hat{\mathbb{E}}[\Phi(y\tilde{f}_{W^*,B^*}(\boldsymbol{x}))] - \hat{\mathbb{E}}_\beta[\Phi(yf^*_{2\alpha}(\boldsymbol{x}))] \tag{24a}$$

$$=\hat{\mathbb{E}}[\Phi(y\tilde{f}_{W^*,B^*}(\boldsymbol{x}))] - \hat{\mathbb{E}}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}))] \tag{24b}$$

$$+\hat{\mathbb{E}}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \hat{\mathbb{E}}_\beta[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] \tag{24c}$$

$$+\hat{\mathbb{E}}_\beta[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \hat{\mathbb{E}}_\beta[\Phi(yf^*_{2\alpha}(\boldsymbol{x}))] \tag{24d}$$

where $\tilde{W}, \tilde{B}$ will be defined later.

Eq. (24b) is less than zero because of the definition of W^*, B^*.

Eq. (24c) can be further decomposed into

$$\begin{aligned} &\hat{\mathbb{E}}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \hat{\mathbb{E}}_\beta[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] && (25a)\\ =&\hat{\mathbb{E}}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \mathbb{E}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] && (25b)\\ &+ \mathbb{E}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \hat{\mathbb{E}}_\beta[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] && (25c) \end{aligned}$$

Since Lemma 2 holds for any W, B, if Eq. (13) holds, w.p. $1 - e^{-N\epsilon^2/C_3}$,

$$|\hat{\mathbb{E}}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})] - \mathbb{E}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}})]| \le 2\epsilon,$$

For the second term, we need to slightly modify this bound as we have βN data points rather than N. It can be presented as, if ϵ satisfies

$$\frac{2^{1+1/(4d+4)}(32r\gamma)^{d/(2d+2)}}{\sqrt{\alpha\xi}(\beta N)^{1/(4d+4)}} < \epsilon < 1, \tag{26}$$

we have w.p. $1 - e^{-\beta N\epsilon^2/C_3}$

$$|\mathbb{E}[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}))] - \hat{\mathbb{E}}_\beta[\Phi(y\hat{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}))]| \le 2\epsilon, \tag{27}$$

where C_3 is a constant. So now we can bound Eq. (24c) by 4ϵ w.p. $1 - 2e^{-\beta N\epsilon^2/C_3}$ given that Eq. (15) holds for ϵ.

Next we show that given the conditions in the lemma, Eq. (24d) will be less than zero. Define $\tilde{S} \subset \Omega_\beta$,

$$\tilde{S} := \{\boldsymbol{x}_i \in \Omega_\beta | y_i f^*_{2\alpha}(\boldsymbol{x}_i) \ge \alpha(1-\xi)\}.$$

Then

$$\begin{aligned} \hat{\mathbb{E}}_\beta[\Phi(yf^*_{2\alpha}(\boldsymbol{x}))] =& \frac{1}{|\Omega_\beta|}\sum_{i\in\Omega_\beta}\Phi(y_i f^*_{2\alpha}(\boldsymbol{x}_i))\\ \ge& \frac{1}{|\Omega_\beta|}\sum_{i\in\Omega_\beta}\mathbb{1}[y_i f^*_{2\alpha}(\boldsymbol{x}_i) < \alpha(1-\xi)]\\ =& \frac{1}{|\Omega_\beta|}\sum_{i\in\Omega_\beta}(1 - \mathbb{1}[y_i f^*_{2\alpha}(\boldsymbol{x}_i) \ge \alpha(1-\xi)])\\ \ge& 1 - \frac{|\tilde{S}|}{\beta N} \end{aligned} \tag{28}$$

By the definition of Φ, we also have

$$\hat{\mathbb{E}}_\beta[\Phi(y\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}))] \le \hat{\mathbb{E}}_\beta[\mathbb{1}[y\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}) < \alpha]] \tag{29}$$

So in the following we will show that under the condition given in the lemma, there exists a pair of $\tilde{W}$ and $\tilde{B}$ such that

$$\hat{\mathbb{E}}_\beta[\mathbb{1}[y\hat{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}) < \alpha]] \le 1 - \frac{|\tilde{S}|}{\beta N} \tag{30}$$

Define

$$\tilde{S}^+ := \{\boldsymbol{x}_i \in \tilde{S} | y_i = 1, f_{2\alpha}(\boldsymbol{x}_i) \ge \alpha(1-\xi)\}$$

and

$$\tilde{S}^- := \{\boldsymbol{x}_i \in \tilde{S} | y_i = -1, f_{2\alpha}(\boldsymbol{x}_i) \le -\alpha(1-\xi)\}.$$

Therefore, $\tilde{S} = \tilde{S}^+ \cup \tilde{S}^-$. Now given any $\boldsymbol{x}_i \in \tilde{S}^+$ and $\boldsymbol{x}_j \in \tilde{S}^-$, $f^*_{2\alpha} \in \mathcal{F}_2$ implies

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \ge |f_{2\alpha}(\boldsymbol{x}_i) - f_{2\alpha}(\boldsymbol{x}_j)|/2 \ge \alpha(1-\xi).$$

For some small $\tau > 0$, set $r = C_2 d^{3/2}\log(d)/\tau^{d-1}$. According to Lemma 8, the sphere $\mathcal{S}^{d-1}$ can be covered by r spherical balls with radius $\arcsin\tau$. Let $\{\boldsymbol{w}_k\}_{k\in[r]}$ be the centers of these spherical balls. Then for any $\boldsymbol{x}_i \in \mathcal{S}^{d-1}$, there exists a $\boldsymbol{w}_k$, such that $\|\boldsymbol{w}_k - \boldsymbol{x}_i\| \le 2\tau$. Set $\tilde{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_K]^T$

Let $\tilde{B} = \text{sign}\left(\tilde{W}\tilde{S}\right)$, i.e., $\tilde{B} = \{\text{sign}\left(\tilde{W}\boldsymbol{x}\right) | \boldsymbol{x} \in \tilde{S}\}$ and the labels of $\tilde{B}$ follows the corresponding $\boldsymbol{x}$. Note that the size of $\tilde{B}$ is less than βN, but is in order of $O(\beta N)$, so for simplicity, we set $m = \beta N$. Let $\tilde{B}^+ = \text{sign}\left(\tilde{W}\tilde{S}^+\right)$ and $\tilde{B}^- = \text{sign}\left(\tilde{W}\tilde{S}^-\right)$.

$$\begin{aligned} &\hat{\mathbb{E}}_\beta[\mathbb{1}[y\hat{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}) < \alpha]]\\ =& \frac{1}{|\Omega_\beta|}\sum_{i\in\Omega_\beta}\mathbb{1}[y_i\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}_i) < \alpha]\\ \le& 1 - \frac{|\tilde{S}|}{\beta N} + \frac{1}{\beta N}\sum_{\boldsymbol{x}_i\in\tilde{S}}\mathbb{1}[y_i\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}_i) < \alpha] \end{aligned} \tag{31}$$

We are now going to show $y_i\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}_i) \ge \alpha$ holds for all $\boldsymbol{x}_i \in \tilde{S}$. We now just consider the case when $y_i = 1$ and the case for $y_i = -1$ is similar. For $\boldsymbol{x}_i \in \tilde{S}^+$.

$$\begin{aligned} &\tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}_i)\\ =& \max_{j\in\tilde{B}^+}\left(\tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\boldsymbol{b}_j\right) - \max_{j\in\tilde{B}^-}\left(\tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\boldsymbol{b}_j\right) - \nu\\ \ge& \tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\text{sign}\left(\tilde{W}\boldsymbol{x}_i\right) - \tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\text{sign}\left(\tilde{W}\boldsymbol{x}_{j^*_-}\right) - \nu\\ \ge& \tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\left(\text{sign}\left(\tilde{W}\boldsymbol{x}_i\right) - \text{sign}\left(\tilde{W}\boldsymbol{x}_{j^*_-}\right)\right) - \nu \end{aligned} \tag{32}$$

where $j^*_- = \arg\max_{j\in\tilde{B}^-}\left(\tanh(\gamma\tilde{W}\boldsymbol{x}_i)^T\boldsymbol{b}_j\right)$. For any $k \in [r]$, we have

$$\tanh(\gamma\boldsymbol{w}_k^T\boldsymbol{x}_i)\left(\text{sign}\left(\boldsymbol{w}_k^T\boldsymbol{x}_i\right) - \text{sign}\left(\boldsymbol{w}_k^T\boldsymbol{x}_{j^*_-}\right)\right) \ge 0$$

Let

$$k^* = \arg\min_{k\in[r]}\left\{\|\boldsymbol{w}_k - \frac{\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}}{\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\|}\|\right\}.$$

Then

$$\begin{aligned} \boldsymbol{w}_{k^*}^T\boldsymbol{x}_i =& \boldsymbol{x}_i^T(\boldsymbol{w}_{k^*} - \frac{\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}}{\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\|}) + \boldsymbol{x}_i^T\frac{\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}}{\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\|}\\ \ge& \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\| - 2\tau \ge \frac{\alpha(1-\xi)}{2} - 2\tau. \end{aligned}$$

And

$$\begin{aligned} \boldsymbol{w}_{k^*}^T\boldsymbol{x}_{j^*_-} =& \boldsymbol{x}_{j^*_-}^T(\boldsymbol{w}_{k^*} - \frac{\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}}{\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\|}) + \boldsymbol{x}_{j^*_-}^T\frac{\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}}{\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\|}\\ \le& -\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_{j^*_-}\| + 2\tau \le -\frac{\alpha(1-\xi)}{2} + 2\tau. \end{aligned}$$

By setting $\tau = \frac{\alpha(1-\xi)}{8}$, we see that

$$\tanh(\gamma \boldsymbol{w}_{k^*}^T \boldsymbol{x}_i) \left(\text{sign}\left(\boldsymbol{w}_{k^*}^T \boldsymbol{x}_i\right) - \text{sign}\left(\boldsymbol{w}_{k^*}^T \boldsymbol{x}_{j_-^*}\right)\right) \geq \frac{\gamma\alpha(1-\xi)}{4}$$

Therefore, as long as $\gamma \geq \frac{4(\nu+\alpha)}{\alpha(1-\xi)}$, we have $\mathbb{1}[y_i \tilde{f}_{\tilde{W},\tilde{B}}(\boldsymbol{x}_i) < \alpha] = 0$ for all $\boldsymbol{x}_i \in \tilde{S}$, and Eq. (30) holds.

Finally by combining Eq. (28), Eq. (29) and Eq. (30), we have Eq. (24d) ≤ 0. This completes the proof. □

A.6 Proof of Lemma 4

Proof. Since $f_{2\alpha}^*$ is independent of $\boldsymbol{x}_i$ and $0 \leq \Phi \leq 1$, by Hoeffding bound, w.p. $1 - 2e^{-2\beta N t^2}$

$$|\hat{\mathbb{E}}[\Phi(y f_{2\alpha}^*(\boldsymbol{x}))] - \mathbb{E}[\Phi(y f_{2\alpha}^*(\boldsymbol{x}))]| \leq t \tag{33}$$

□

A.7 Proof of Lemma 6

Proof.

$$\begin{aligned}
&|\tilde{f}_{W^*,B^*}(\boldsymbol{x}) - f_{W^*,B^*}(\boldsymbol{x})| \\
\leq &\max_{j \in B^-} \left(|\tanh(\gamma W^* \boldsymbol{x})^T \boldsymbol{b}_j^* - \text{sign}\left(W^* \boldsymbol{x}\right)^T \boldsymbol{b}_j^*|\right) \\
&+ \max_{j \in B^+} \left(|\tanh(\gamma W^* \boldsymbol{x})^T \boldsymbol{b}_j^* - \text{sign}\left(W^* \boldsymbol{x}\right)^T \boldsymbol{b}_j^*|\right) \\
\leq &2 \max_{j \in B} \left(|\tanh(\gamma W^* \boldsymbol{x})^T \boldsymbol{b}_j^* - \text{sign}\left(W^* \boldsymbol{x}\right)^T \boldsymbol{b}_j^*|\right) \\
\leq &4r \max_{k \in [r]} |\tanh(\gamma \boldsymbol{w}_k^{*T} \boldsymbol{x}) - \text{sign}\left(\boldsymbol{w}_k^{*T} \boldsymbol{x}\right)|
\end{aligned}$$

Given Assumption 1, we have w.p. at least $1 - c_1 r\theta$, $|\boldsymbol{w}_k^{*T} \boldsymbol{x}| \geq \theta$ for all $k \in [r]$ and

$$|\tanh(\gamma \boldsymbol{w}_k^{*T} \boldsymbol{x}) - \text{sign}\left(\boldsymbol{w}_k^{*T} \boldsymbol{x}\right)| \leq 2e^{-2\gamma\theta}$$

□