A Additional experiments

Figure 5 and 6 plot the same phenomena as Figure 2 and 3. The only difference is that they strictly follows the setting described in Section 2 where the training data and test data has been normalized to $\|\mathrm{x}\| = 1$, and there's no bias term when initializing the neural network linear layer to consist of our problem setup. For each experimental setting (i.e, for each value of d, C_0, and m), we report results averaged over 5 independent random runs.

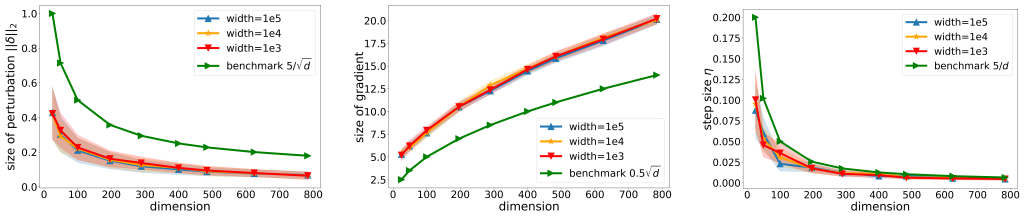

Figure 5: Normalize data $\|\mathrm{x}\| = 1$. Smallest size of perturbation to switch the prediction (**left**), norm of the gradient after training the neural networks (**middle**), smallest step size (**right**), as a function of input dimension d for fix $C_0 = 10$ with different width m.

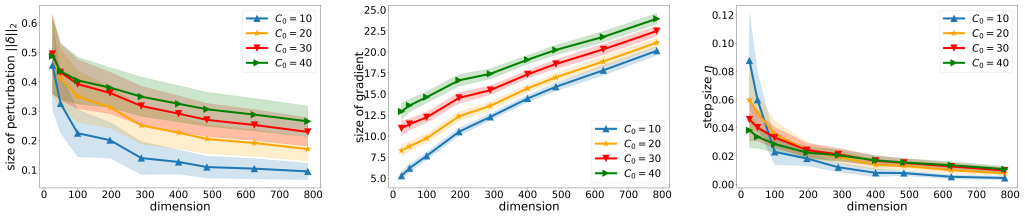

Figure 6: Normalize data $\|\mathrm{x}\| = 1$. Smallest size of perturbation to switch the prediction (**left**), norm of the gradient after training the neural networks (**middle**), smallest step size (**right**), as a function of input dimension d for fix $m = 10^5$ with different C_0.

Figure 7 is the histogram of smallest size of perturbation to switch the prediction, the norm of the gradient after training the neural networks, and the step size when $d = 784, m = 10^5, C_0 = 10$. The histogram exhibits a Gaussian distribution, and the step size to flip the prediction is small. (η concentrates around 0.045.)

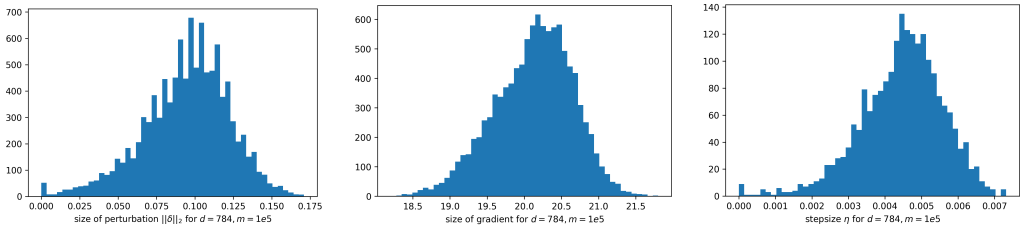

Figure 7: Histogram of smallest size of perturbation to switch the prediction, the norm of the gradient after training the neural networks, and the step size when $d = 784, m = 10^5, C_0 = 10$.

In Algorithm 1, we describe our projected adversarial training algorithm in details. For generating adversarial example with budget R at each round, we choose learning rate $\alpha = 2.5 \times R/100$ with $T_2 = 100$. This is a common choice, which is first introduced in [Madry et al., 2018]. For updating the weight matrix, we choose learning rate $\beta = 0.01$. We stop when the robust training accuracy is not increasing.

Algorithm 1 Projected Adversarial Training

Training samples $(\mathrm{X}, \mathrm{y}) = \{\mathrm{x}_i, y_i\}_{i=1}^n$. Initialize $\mathrm{w}_{s,0} \sim N(0, \mathrm{I}_d), a_s \sim \mathrm{unif}\{\pm 1\}, \forall s \in [m]$ with fixed a. Epochs T_1, T_2. Learning rate α, β, perturbation budget per sample R. Logistic loss ℓ. Batch size bs. $\mathcal{P}_\Delta$ is the projection operator onto the set Δ.
for $t = 1, \ldots, T_1$ **do**
  for $k = 1, \ldots, T_2$ **do**
    $\tilde{\mathrm{X}} = \mathrm{X} + \alpha \sum_{i=1}^n \nabla_{\mathrm{x}_i} \ell(y_i f(\mathrm{x}_i; \mathrm{a}, \mathrm{W}_{t-1}))$
    $\tilde{\mathrm{X}} = \mathcal{P}_{\mathcal{B}_{2,\infty}(\mathrm{X}, R)}(\tilde{\mathrm{X}})$ {generate adversarial examples}
  end for
  for $\lfloor \frac{n}{bs} \rfloor$ rounds **do**
    Sample a mini-batch of size bs from $(\tilde{\mathrm{X}}, \mathrm{y})$ as $(\tilde{\mathrm{x}}_{i_j}, y_{i_j})_{j=1}^{bs}$.
    $\mathrm{W}_t = \mathrm{W}_{t-1} - \beta \nabla_{\mathrm{W}_{t-1}} \sum_{j=1}^{bs} \ell(y_{i_j} f(\tilde{\mathrm{x}}_{i_j}; \mathrm{a}, \mathrm{W}_{t-1}))$
  end for
  $\mathrm{W}_t = \mathcal{P}_{\mathcal{B}_{2,\infty}(\mathrm{W}_0, \frac{C_0}{\sqrt{m}})}(\mathrm{W}_t)$ {Project the weight matrix to satisfy lazy regime.}
end for
return: W_{T_1}

B Proof of theorems

Proof of Theorem 3.1. From Lemma B.5, B.7, B.9, we have that

$$\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\| \geq C_1' \sqrt{d},$$
$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \leq C_2',$$
$$\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x} + \delta; \mathrm{a}, \mathrm{W})\| \leq C_3' \sqrt{d} \text{ for } \|\delta\| \leq o(\frac{1}{\sqrt{d}})$$

We simplify the notation as $f(\mathrm{x}) = f(\mathrm{x}; \mathrm{a}, \mathrm{W})$. Define $\tilde{\eta} = \eta \|\nabla f(\mathrm{x})\|^2$. Without loss of generality, assume $f(\mathrm{x}) > 0$, let

$$\tilde{\eta} = -\frac{2f(\mathrm{x})}{1 - \sup_{\|\delta\| \leq \frac{\tilde{\eta}}{\|\nabla f(\mathrm{x})\|}} \frac{\|\nabla f(\mathrm{x}+\delta)\| - \|f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|}},$$

we have for the point $\mathrm{x} + \tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}$,

$$\begin{aligned}
& f(\mathrm{x} + \tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) \\
&= f(\mathrm{x}) + \int_0^1 f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2})' dt && \text{(Fundamental theorem of calculus)} \\
&= f(\mathrm{x}) + \int_0^1 \tilde{\eta} \frac{\nabla f(\mathrm{x})^\top}{\|\nabla f(\mathrm{x})\|^2} \nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) dt \\
&= f(\mathrm{x}) + \tilde{\eta} + \int_0^1 \tilde{\eta} \frac{\nabla f(\mathrm{x})^\top}{\|\nabla f(\mathrm{x})\|} \frac{(\nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) - \nabla f(\mathrm{x}))}{\|\nabla f(\mathrm{x})\|} dt \\
&\leq f(\mathrm{x}) + \tilde{\eta} + |\tilde{\eta}| \int_0^1 \frac{\|\nabla f(\mathrm{x} + t\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(\mathrm{x})\|^2}) - \nabla f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|} dt && \text{(Cauchy–Schwarz)} \\
&\leq f(\mathrm{x}) + \tilde{\eta} + |\tilde{\eta}| \sup_{\delta \in \mathbb{R}^d : \|\delta\| \leq \frac{\tilde{\eta}}{\|\nabla f(\mathrm{x})\|}} \frac{\|\nabla f(\mathrm{x} + \delta) - \nabla f(\mathrm{x})\|}{\|\nabla f(\mathrm{x})\|} \\
&= -f(\mathrm{x}) < 0
\end{aligned}$$

Note $|\tilde{\eta}| = \Theta(1)$ and thus $\eta \leq O(\frac{1}{d})$, $\|\eta \nabla f(\mathrm{x})\| = \|\tilde{\eta} \frac{\nabla f(\mathrm{x})}{\|\nabla f(x)\|^2}\| \leq O(\frac{1}{\sqrt{d}})$.

□

In order to proof the main theorem, we start by giving some standard theorems, which will be used in later proof.

Theorem B.1 (Berstein's inequality)**.** Let $z_1, \ldots, z_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that

$$\sum_{i=1}^{n} \mathbb{E}[z_i^2] \leq v \text{ and } \sum_{i=1}^{n} \mathbb{E}[|z_i|^q] \leq \frac{q!}{2} v c^{q-2} \text{ for all integers } q \geq 3.$$

If $S = \sum_{i=1}^{n}(z_i - \mathbb{E}z_i)$, then for all $t > 0$,

$$\mathrm{P}(S \geq \sqrt{2vt} + ct) \leq e^{-t}.$$

We will also use repeatly that

$$\mathbb{E}_{z\sim\mathcal{N}(0,1)}[|z|^q] \leq (q-1)!! \leq \frac{q!}{2}, \tag{1}$$

as well as the following concentration of χ^2 random variables: let $z_1, \ldots, z_m$ be i.i.d. standard Gaussians, then with probability at least $1-\gamma$, one has

$$\left| \sum_{s=1}^{m} z_s^2 - m \right| \leq 4\sqrt{m \log(2/\gamma)} \tag{2}$$

Theorem B.2. (Chernoff's inequality) If $z_1, z_2, \ldots, z_N$ are independent Bernoulli random variables with parameters μ_i. Let $S_N = \sum_i z_i$ and $p = \mathbb{E}S_N = \sum_i \mu_i$, then for $t > p$,

$$P(S_N \geq t) \leq \exp(-p)\left(\frac{ep}{t}\right)^t$$

Theorem B.3. (Theorem 2.26 in Wainwright [2019]) Let $\mathrm{z} = (z_1, z_2, \ldots, z_n)$ be a vector of i.i.d standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t. the Euclidean norm. Then we have

$$P(|f(\mathrm{z}) - \mathbb{E}f(\mathrm{z})| \geq t) \leq 2e^{-\frac{t^2}{2L^2}} \text{ for all } t \geq 0$$

In Lemma B.4, we let S_v denote the set of neurons that change sign between weights w_0 and w for the x, S_v' as the set of neurons that change sign between weights w_0 and w for the $\mathrm{x}+\delta$. We show that the size of S_v and S_v' is small as long as the weights stay close to initialization.

Lemma B.4. Define the following

$$S_v := \left\{ s \middle| \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathbb{1}[\langle \mathrm{w}_s, \mathrm{x}\rangle > 0] \neq \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle > 0] \right\}$$
$$S_v' := \left\{ s \middle| \exists \delta, \|\delta\| \leq R \leq 0.5, \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathbb{1}[\langle \mathrm{w}_s, \mathrm{x}+\delta\rangle > 0] \neq \mathbb{1}[\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle > 0] \right\}$$

Then with probability at least $1-\gamma$, the following hold:

$$|S_v| \leq \left| \left\{ s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V \right\} \right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V \right] \leq Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}$$
$$|S_v'| \leq \left| \left\{ s \middle| |\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \leq V \right\} \right| = \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \leq V \right] \leq \left(Vm + \sqrt{\frac{m\log(1/\gamma)}{2}} \right)(1+R)$$

Proof of Lemma B.4. Define $\mathrm{v}_s = \mathrm{w}_s - \mathrm{w}_{s,0}$. Note that $s \in S_v$ implies $\forall 1 \leq s \leq m$,

$$\left| \langle \mathrm{w}_{s,0}, \mathrm{x}\rangle \right| \leq \sup_{\|\mathrm{v}_s\| \leq V} \left| \langle \|\mathrm{w}_s - \mathrm{w}_{s,0}\|, \mathrm{x}\rangle \right| \leq \sup_{\|\mathrm{v}_s\| \leq V} \|\mathrm{v}_s\|_2 \|\mathrm{x}\|_2 \leq V$$

Thus we have

$$\mathbb{E}\left[\frac{1}{m} \sum_{s=1}^{m} \mathbb{1}\left[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V \right] \right] = P\left(|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \leq V \right) \leq \frac{2V}{\|\mathrm{x}\|_2^2 \sqrt{2\pi}} \leq V,$$

where the expectation is with respect to the randomness in initialization, and the inequality holds since $\langle \mathrm{w}_{s,0}, \mathrm{x} \rangle$ is a Gaussian r.v. with variance 1. By Hoeffding inequality, with probability at least $1-\gamma$,

$$\begin{aligned}
&\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\big] \\
&\le \mathbb{E}\bigg[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\big]\bigg] + \sqrt{\frac{\log(1/\gamma)}{2m}} && \text{(Hoeffding inequality)} \\
&\le V + \sqrt{\frac{\log(1/\gamma)}{2m}}
\end{aligned}$$

As a result, with probability at least $1-\gamma$, we arrive at the following upperbound on the size of S_v:

$$|S_v| \le \Big|\big\{s \big| |\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\big\}\Big| = \sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}\rangle| \le V\big] \le Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}$$

Same way we can bound the size of S'_v. $s \in S'_v$ implies $\forall 1 \le s \le m$,

$$\big|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle\big| \le \sup_{\|\mathrm{v}_s\|\le V, \|\delta\|\le R} \big|\langle \mathrm{v}_s, \mathrm{x}+\delta\rangle\big| \le \sup_{\|\mathrm{v}_s\|\le V, \|\delta\|\le R} \|\mathrm{v}_s\|_2 \|\mathrm{x}+\delta\|_2 \le V(1+R)$$

Thus we have

$$\mathbb{E}\bigg[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\big]\bigg] = P\big(|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\big) \le V(1+R),$$

where the expectation is with respect to the randomness in initialization, and the inequality holds since $\langle \mathrm{w}_{s,0}, \mathrm{x} \rangle$ is a Gaussian r.v. with variance 1. By Hoeffding inequality, with probability at least $1-\gamma$,

$$\begin{aligned}
&\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\big] \\
&\le \mathbb{E}\bigg[\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V(1+R)\big]\bigg] + \sqrt{\frac{\log(1/\gamma)}{2m}}(1+R) && \text{(Hoeffding inequality)} \\
&\le V(1+R) + \sqrt{\frac{\log(1/\gamma)}{2m}}(1+R)
\end{aligned}$$

As a result, with probability at least $1-\gamma$, we arrive at the following upperbound on the size of S_v:

$$|S_v| \le \Big|\big\{s \big| |\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V\big\}\Big| = \sum_{s=1}^{m} \mathbb{1}\big[|\langle \mathrm{w}_{s,0}, \mathrm{x}+\delta\rangle| \le V\big] \le \bigg(Vm + \sqrt{\frac{m\log(1/\gamma)}{2}}\bigg)(1+R)$$

□

Now we begin our proof. Essentially we want to lower bound $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$, upper bound $|f(\mathrm{x}; \mathrm{a}, \mathrm{W})|$ and upper bound $\|\nabla f(\mathrm{x}; \mathrm{a}, \mathrm{W}) - \nabla f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W})\|$. Lemma B.5 gives the upper bound of $|f(\mathrm{x}; \mathrm{a}, \mathrm{W})|$ as $O(1)$.

Lemma B.5. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\Big(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\Big)$,

$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \le \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}} + C_0$$

Particularly, there exists $C > 0$ such that for $m \ge C\log(2/\gamma)$, we have

$$|f(\mathrm{x}; \mathrm{a}, \mathrm{W})| \le 2\sqrt{\log(2/\gamma)} + C_0$$

Proof of Lemma B.5. From Bubeck et al. [2021] we know that $|f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| \le \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}}$. Now we consider bound $|f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)|$,

$$\begin{aligned}
|f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| &\le |\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\sigma(\mathrm{w}_s^\top \mathrm{x}) - \sigma(\mathrm{w}_{s,0}^\top \mathrm{x}))| \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^{m} |\sigma(\mathrm{w}_s^\top \mathrm{x}) - \sigma(\mathrm{w}_{s,0}^\top \mathrm{x})| && \text{(Triangle Inequality)} \\
&\le \sqrt{m}|\mathrm{w}_s^\top \mathrm{x} - \mathrm{w}_{s,0}^\top \mathrm{x}| && (\sigma(\cdot) \text{ is 1-Lipschitz.}) \\
&\le \sqrt{m}\|\mathrm{w}_s - \mathrm{w}_{s,0}\|\|\mathrm{x}\| && (\|\mathrm{w}_s - \mathrm{w}_{s,0}\| \le \tfrac{C_0}{\sqrt{m}}) \\
&\le C_0
\end{aligned}$$

Thus,

$$|f(\mathrm{x};\mathrm{a},\mathrm{W})| = |f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| + |f(\mathrm{x};\mathrm{a},\mathrm{W}) - f(\mathrm{x};\mathrm{a},\mathrm{W}_0)| \le \sqrt{2\log(2/\gamma)} + \frac{2\log(2/\gamma)}{\sqrt{m}} + C_0$$

□

In Lemma B.6, we want to calculate the upper bound of the probability that the neuron flips sign due to (1) perturbation δ; (2) different perturbation δ, δ' on the data x at weight w_s.

Lemma B.6. For any δ such that $\|\delta\| \le R \le \frac{1}{2}$,

$$P\left(\mathrm{sign}(\mathrm{w}_{s,0}^\top \mathrm{x}) \ne \mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))\right) \le R\sqrt{2\log(d)} + \frac{1}{d} \tag{3}$$

$$\begin{aligned}
&P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) \ne \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\Big) \\
&\qquad \le 2\varepsilon\left(\sqrt{d} + 2\sqrt{d\log(2/\varepsilon)}\right) + (1+R+\varepsilon)V
\end{aligned} \tag{4}$$

Proof of Lemma B.6. Equation (3) directly follows from Bubeck et al. [2021]. For equation (4), we have

$$\begin{aligned}
&P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) \ne \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\Big) \\
&\le P\Big(\exists \delta', \delta' \in \mathcal{B}_2(\delta,\varepsilon), \text{ and } \exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\delta'-\delta)| \ge t\Big) \\
&\quad + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\mathrm{x}+\delta)| \le t\Big) && \text{(Holds for any threshold } t.) \\
&\le P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \|\mathrm{w}_s\| \ge t/\varepsilon\Big) + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_s^\top(\mathrm{x}+\delta)| \le t\Big) \\
&\le P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), \|\mathrm{w}_{s,0}\| \ge t/\varepsilon - \|\mathrm{w}_s - \mathrm{w}_{s,0}\|\Big) \\
&\quad + P\Big(\exists \mathrm{W}, \mathrm{W} \in \mathcal{B}_{2,\infty}(\mathrm{W}_0, V), |\mathrm{w}_{s,0}^\top(x+\delta)| \le t + |(\mathrm{w}_s - \mathrm{w}_{s,0})^\top(\mathrm{x}+\delta)|\Big) \\
&\qquad\qquad \text{(Triangle Inequality, pick } t = \varepsilon\left(\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + V\right)) \\
&\le P\left(\|\mathrm{w}_{s,0}\| \ge \sqrt{d + 4\sqrt{d\log(2/\varepsilon)}}\right) + P\left(|\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)| \le \varepsilon\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + (1+R+\varepsilon)V\right) \\
&\qquad\qquad (\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta) \sim \mathcal{N}(0,\sigma^2) \text{ with } \sigma^2 \ge \tfrac{1}{2} \text{ since } \|\delta\| \le \tfrac{1}{2}) \\
&\le \varepsilon + \varepsilon\sqrt{d + 4\sqrt{d\log(2/\varepsilon)}} + (1+R+\varepsilon)V && \text{(Using equation (2).)} \\
&= 2\varepsilon(\sqrt{d} + 2\sqrt{d\log(2/\varepsilon)}) + (1+R+\varepsilon)V
\end{aligned}$$

□

Lemma B.7 gives lower bound on $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \Omega(\sqrt{d})$.

Lemma B.7. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \left(\frac{1}{2} - \left(\sqrt{\frac{2\log(4/\gamma)}{m}} + \frac{\log(4/\gamma)}{m}\right)\right)^{1/2}\left(d - 5\sqrt{d\log(8/\gamma)}\right)^{1/2} - C_0 - \sqrt{\frac{1}{\sqrt{m}}\left(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}\right)}\left(d + 4\sqrt{d\log(8m/\gamma)}\right)^{1/2}$$

Particularly, there exists $C > 0$ such that for $m \geq C\log(4/\gamma)$ and $d \geq C\frac{\log(8m/\gamma)}{m}$, we have

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \frac{1}{4}\sqrt{d}$$

Proof of Lemma B.7. Follow the process of Bubeck et al. [2021], let $\mathrm{P} = \mathrm{I}_d - \mathrm{x}\mathrm{x}^\top$ be the projection on the orthogonal complement of the span of x. We have $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\|$. Thus we have,

$$\begin{aligned}\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| &\geq \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \\ &= \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0) + \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \\ &\geq \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| - \|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|\end{aligned} \tag{5}$$

Since $a_s\mathrm{P}\mathrm{w}_{s,0}$ is distributed as $\mathcal{N}(0,\mathrm{I}_{d-1})$, denote $\mathrm{z} := a_s\mathrm{P}\mathrm{w}_{s,0}$, we have

$$\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\mathrm{P}\mathrm{w}_{s,0}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}) \overset{(d)}{=} \left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})^2}\right)\mathrm{z} \text{ where } \mathrm{z}\sim\mathcal{N}(0,\mathrm{I}_{d-1})$$

where $\overset{(d)}{=}$ means equal in distribution.

Using (2), we have with probability at least $1-\gamma$

$$\|\mathrm{z}\|^2 \geq d - 1 - 4\sqrt{d\log(2/\gamma)} \geq d - 5\sqrt{d\log(2/\gamma)} \qquad (d \geq 1, \gamma < 2/e)$$

Apply Bernstein's inequality with $v = m, c = 1$, we have with probability at least $1-\gamma$,

$$\begin{aligned}\frac{1}{m}\sum_{s=1}^{m}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})^2 &\geq \mathbb{E}_{X\sim\mathcal{N}(0,1)}[|\sigma'(X)|^2] - \left(\sqrt{\frac{2\log(1/\gamma)}{m}} + \frac{\log(1/\gamma)}{m}\right) \\ &= \frac{1}{2} - \left(\sqrt{\frac{2\log(1/\gamma)}{m}} + \frac{\log(1/\gamma)}{m}\right)\end{aligned}$$

Thus we can get

$$\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \geq \left(\frac{1}{2} - \left(\sqrt{\frac{2\log(2/\gamma)}{m}} + \frac{\log(2/\gamma)}{m}\right)\right)^{1/2}\left(d - 5\sqrt{d\log(4/\gamma)}\right)^{1/2}$$

Now we start upper bound $\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|$. Note that

$$
\begin{aligned}
&\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\| \\
&=\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_s^\top\mathrm{x}))\|+\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s(\mathrm{w}_s\sigma'(\mathrm{w}_s^\top\mathrm{x})-\mathrm{w}_{s,0}\sigma'(\mathrm{w}_s^\top\mathrm{x}))\| \\
&\qquad+\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le\sqrt{m}\|\mathrm{w}_s-\mathrm{w}_{s,0}\|+\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \\
&\le C_0+\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\| \qquad (6)
\end{aligned}
$$

Now we start bounding the second term of equation (6). We have that with probability at least $1-\gamma$,

$$
\begin{aligned}
&\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})) \\
&\overset{(d)}{=}\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))^2}\right)\mathrm{z}\quad(\mathrm{z}:=a_s\mathrm{P}\mathrm{w}_{s,0}\sim N(0,\mathrm{I}_{d-1})) \\
&=\left(\sqrt{\frac{1}{m}\sum_{s=1}^{m}\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))^2}\right)\mathrm{z} \\
&=\sqrt{\frac{1}{m}|S_v|}\cdot\mathrm{z}\qquad\text{(By the definition of } S_v \text{ in Lemma B.4 with } V=\tfrac{C_0}{\sqrt{m}}) \\
&\le\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(1/\gamma)}{2}})}\cdot\mathrm{z} \qquad (7)
\end{aligned}
$$

Using equation (2), we see that with probability at least $1-\gamma$, one has for all $1\le s\le m$,

$$\|\mathrm{z}\|=\|a_s\mathrm{P}\mathrm{w}_{s,0}\|\le\sqrt{d+4\sqrt{d\log(2m/\gamma)}}$$

Thus follow from equation (6), we have

$$\|\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\mathrm{P}\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}_0)\|\le C_0+\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(2/\gamma)}{2}})}\sqrt{d+4\sqrt{d\log(4m/\gamma)}}$$

And as a result, with probability at least $1-\gamma$

$$\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})\| \geq \left(\frac{1}{2}-\left(\sqrt{\frac{2\log(4/\gamma)}{m}}+\frac{\log(4/\gamma)}{m}\right)\right)^{1/2}\left(d-5\sqrt{d\log(8/\gamma)}\right)^{1/2}$$
$$-C_0-\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(4/\gamma)}{2}})\sqrt{d+4\sqrt{d\log(8m/\gamma)}}} \geq \Omega(\sqrt{d})$$

□

Now we start bounding $\|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W})-\nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\|$. In order to prove Lemma B.9, we will leverage the fact that for any $h\in\mathbb{R}^d$, we have $\|h\|=\sup_{r\in\mathbb{S}^{d-1}} r\cdot h$. We first start a lemma with fixed r,δ to prove Lemma B.8, then use the ε-net argument to prove Lemma B.9.

Lemma B.8. Fix $r\in\mathbb{S}^{d-1}$, and $\delta\in\mathbb{R}^d$ such that $\|\delta\|\leq R\leq\frac{C_1}{\sqrt{d}}$, C_1 is a constant. For any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W}\in\mathcal{B}_{2,\infty}\left(\mathrm{W}_0,\frac{C_0}{\sqrt{m}}\right)$,

$$\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$$
$$\leq 2\sqrt{\log(4/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4}+\sqrt{\frac{\log(4/\gamma)}{m}}\right)+C_0$$
$$+3\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(4/\gamma)}{2}})\sqrt{d+4\sqrt{d\log(8m/\gamma)}}+5(C_0+\sqrt{\log(4/\gamma)})\cdot\sqrt{\log(4m/\gamma)}}$$

Proof of Lemma B.8. Let $X_{s,0}=a_s(\mathrm{w}_{s,0}\cdot r)(\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))$, $X_s=a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$. We have $\mathbb{E}[X_{s,0}]=0$, and equation (1) gives us $\mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^{2q}]\leq\frac{(2q)!}{2}$. Thus, we have

$$\begin{aligned}
\mathbb{E}[|X_{s,0}|^q] &\leq \mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^q|\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))|^q]\\
&\leq \sqrt{\mathbb{E}[|\mathrm{w}_{s,0}\cdot r|^{2q}]\cdot P(\mathrm{sign}(\mathrm{w}_{s,0}^\top\mathrm{x})\neq\mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))} \qquad \text{(Using equation (3))}\\
&\leq \sqrt{\frac{(2q)!}{2}}\cdot\sqrt{R\sqrt{2\log(d)}+\frac{1}{d}}\\
&\leq \frac{q!}{2}2^q\cdot\sqrt{2R\sqrt{2\log(d)}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[|X_{s,0}|^2] &\leq \sqrt{\mathbb{E}[(\mathrm{w}_{s,0}\cdot r)^4]\cdot P(\mathrm{sign}(\mathrm{w}_{s,0}^\top\mathrm{x})\neq\mathrm{sign}(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))}\\
&\leq 2\sqrt{R\sqrt{2\log(d)}} \qquad (\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|X|^4]\leq 3!!\leq 4)\\
&\leq 4\sqrt{R\sqrt{\log(d)}}
\end{aligned}$$

Using Theorem B.1 with $v=4m\sqrt{R\sqrt{\log(d)}}, c=2$, we have

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{s=1}^{m}X_{s,0} &\leq \sqrt{8\sqrt{R\sqrt{\log(d)}}\log(1/\gamma)}+\frac{2\log(1/\gamma)}{\sqrt{m}}\\
&= 2\sqrt{\log(1/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4}+\sqrt{\frac{\log(1/\gamma)}{m}}\right)
\end{aligned}$$

Now we start bounding $|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}(X_s-X_{s,0})|$. In fact, here we no longer fix r,δ and directly bound $|\sup_{r\in\mathbb{S}^{d-1},\delta\in\mathbb{R}^d,\|\delta\|\leq\frac{C_1}{\sqrt{d}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}(X_s-X_{s,0})|$.

$$\Big|\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}(X_s - X_{s,0})\Big|$$

$$=\Big|\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\Big(\mathrm{w}_s^\top r(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big)\Big|$$

$$\le\Big|\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\Big(\mathrm{w}_s^\top r(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))\Big)\Big|$$

$$+\Big|\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\Big(\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))-\mathrm{w}_{s,0}^\top r(\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big)\Big|$$

(Triangle Inequality)

$$\le\frac{1}{\sqrt{m}}\sum_{s=1}^{m}\|\mathrm{w}_s-\mathrm{w}_{s,0}\||\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))|+\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|$$

$$+\Big\|\sup_{\delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big\|$$

(Cauchy-Schwarz, triangle Inequality)

$$\le C_0+\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\| \tag{8}$$

$$+\Big\|\sup_{\delta\in\mathbb{R}^d,\|\delta\|\le\frac{C_1}{\sqrt{d}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))-\sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))\Big\| \tag{9}$$

Define $\mathrm{P}=\mathrm{I}_d-\mathrm{x}\mathrm{x}^\top$ as the projection on the orthogonal complement of the span of x. For the second term of equation (9), we have

$$\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|$$

$$\le\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|+\Big\|\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{x}\mathrm{x}^\top\mathrm{w}_{s,0}(\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|$$

$$\le\Big\|\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|$$

$$+\frac{1}{\sqrt{m}}\sum_{s=1}^{m}\|\mathrm{x}\mathrm{x}^\top\mathrm{w}_{s,0}\|\cdot|\sigma'(\mathrm{w}_s^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x})|$$

$$\le\Big\|\sup_{\substack{\forall s\in[m],\mathrm{v}_s\in\mathbb{R}^d,\\ \|\mathrm{v}_s\|\le\frac{C_0}{\sqrt{m}}}}\frac{1}{\sqrt{m}}\sum_{s=1}^{m}a_s\mathrm{P}\mathrm{w}_{s,0}(\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top\mathrm{x})-\sigma'(\mathrm{w}_{s,0}^\top\mathrm{x}))\Big\|+\frac{1}{\sqrt{m}}|S_v|\max_{1\le s\le m}\|\mathrm{x}\mathrm{x}^\top\mathrm{w}_{s,0}\|$$

(By the definition of S_v in Lemma B.4, Hölder's inequality)

$$\le\sqrt{\frac{1}{\sqrt{m}}(C_0+\sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d+4\sqrt{d\log(8m/\gamma)}}+(C_0+\sqrt{\log(4/\gamma)})\cdot\sqrt{\log(4m/\gamma)}$$

(Equation (7), $\mathrm{x}^\top\mathrm{w}_{s,0}\sim N(0,1)$)

where the last line holds because with probability at least $1-\gamma$, one has for all $1 \le s \le m$, $|\mathrm{x}^\top \mathrm{w}_{s,0}| \le \sqrt{\log(m/\gamma)}$.

Define $\mathrm{P}' = \mathrm{I}_d - \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2}$ as the projection on the orthogonal complement of the span of $\mathrm{x}+\delta$. We bound the third term of equation (9) the same way as follows.

$$
\begin{aligned}
&\left\| \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \mathrm{w}_{s,0} (\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))) \right\| \\
&\le \left\| \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \mathrm{P}' \mathrm{w}_{s,0} (\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))) \right\| \\
&\quad + \left\| \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0} (\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))) \right\| \\
&\le \left\| \sup_{\substack{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}} \\ \forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m a_s \mathrm{P}' \mathrm{w}_{s,0} (\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))) \right\| \\
&\quad + \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} \sum_{s=1}^m |\sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta))| \left\| \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0} \right\| \\
&\le \sqrt{\sup_{\substack{\forall s\in[m], \mathrm{v}_s\in\mathbb{R}^d, \\ \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}}} \frac{1}{m} \sum_{s=1}^m (\sigma'((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta)) - \sigma'(\mathrm{w}_{s,0}^\top(\mathrm{x}+\delta)))} \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \|\mathrm{z}(\delta)\| \\
&\qquad\qquad \text{(Define } \mathrm{z}(\delta) := a_s \mathrm{P}' \mathrm{w}_{s,0} \sim N(0, \mathrm{I}_{d-1})\text{)} \\
&\quad + \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \frac{1}{\sqrt{m}} |S_v'| \left\| \frac{(\mathrm{x}+\delta)(\mathrm{x}+\delta)^\top}{\|\mathrm{x}+\delta\|^2} \mathrm{w}_{s,0} \right\| \qquad \text{(By definition of } S_v' \text{ in Lemma B.4)} \\
&\le \sqrt{\frac{1}{m}|S_v'|} \cdot \sup_{\delta\in\mathbb{R}^d, \|\delta\| \le \frac{C_1}{\sqrt{d}}} \|\mathrm{z}(\delta)\| \\
&\quad + \frac{1}{\sqrt{m}} |S_v'| \cdot 2 \max_{1\le s\le m} \left| \mathrm{x}^\top \mathrm{w}_{s,0} + \delta^\top \mathrm{w}_{s,0} \right| \qquad (\|\mathrm{z}(\delta)\| \le \|\mathrm{w}_{s,0}\|) \\
&\le \sqrt{\frac{1+R}{\sqrt{m}} (C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}) \sqrt{d + 4\sqrt{d\log(8m/\gamma)}}} \\
&\quad + (C_0 + \sqrt{\log(4/\gamma)})(1+R) \cdot 2 \left(\sqrt{\log(4m/\gamma)} + \frac{\sqrt{d + 4\sqrt{d\log(8m/\gamma)}}}{\sqrt{d}} \right) \\
&\qquad\qquad (R \le \tfrac{C_1}{\sqrt{d}}, d > \log(m/\gamma)\log(1/\gamma)) \\
&\le \sqrt{\frac{2}{\sqrt{m}} (C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}) \sqrt{d + 4\sqrt{d\log(8m/\gamma)}}} + 4(C_0 + \sqrt{\log(4/\gamma)}) \left(\sqrt{\log(4m/\gamma)} + 3 \right)
\end{aligned}
$$

As a result, we get

$$
\begin{aligned}
\frac{1}{\sqrt{m}} \sum_{s=1}^m X_s &\le \frac{1}{\sqrt{m}} \sum_{s=1}^m X_{s,0} + |\frac{1}{\sqrt{m}} \sum_{s=1}^m (X_s - X_{s,0})| \\
&\le 2\sqrt{\log(4/\gamma)} \left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{\log(4/\gamma)}{m}} \right) + C_0 \\
&\quad + 3\sqrt{\frac{1}{\sqrt{m}} (C_0 + \sqrt{\frac{\log(4/\gamma)}{2}}) \sqrt{d + 4\sqrt{d\log(8m/\gamma)}}}
\end{aligned}
$$

$$+ 5(C_0 + \sqrt{\log(4/\gamma)})\left(\sqrt{\log(4m/\gamma)} + 3\right)$$

□

Lemma B.9. Let $\|\delta\| \le R \le \frac{C_1}{\sqrt{d}}, \bar{C} = \max\{1, C_0\}, m \ge d^{2.4}$. Then, for any x, with probability at least $1-\gamma$, the following holds for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right)$,

$$\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d:\|\delta\|\le R}} \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta))) \le 9\left(C_1 d^2\log^2(md)\sqrt{\frac{\log(d)}{d}}\right)^{1/4} + \frac{15d\log(md)}{\sqrt{m}} + 327\bar{C}C_0 d^{0.25}$$

provided that $d \ge \log(4/\gamma)^2$. Particularly, for any $c > 0$, there exists C, C' such that if $\log^2(md)\sqrt{\frac{\log(d)}{d}} \le C$, $\frac{d\log(md)}{\sqrt{m}} \le C'\sqrt{d}$, then we have,

$$\sup_{\substack{r\in\mathbb{S}^{d-1},\\ \delta\in\mathbb{R}^d,\|\delta\|\le R}} \|\nabla f(\mathrm{x};\mathrm{a},\mathrm{W}) - \nabla f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W})\| \le c\sqrt{d}$$

The above can realize when $m \le O(\exp(d^{0.24}))$

Proof. Define $\Phi(r,\delta) = \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s(\mathrm{w}_s\cdot r)(\sigma'(\mathrm{w}_s^\top \mathrm{x}) - \sigma'(\mathrm{w}_s^\top(\mathrm{x}+\delta)))$, and N_ε an ε-net for $\Omega = \{(r,\delta), \|r\| = 1, \|\delta\| \le R\}$. In Lemma B.8 we bound $\Phi(r,\delta)$ for a fixed r and δ, here we bounded it uniformly over Ω. We know that $|N_\varepsilon| \le (10/\varepsilon)^{2d}$. Using Lemma B.8, we obtain with probability at least $1-\gamma$,

$$\begin{aligned}
\sup_{(r,\delta)\in\Omega}\Phi(r,\delta) &\le \sup_{(r,\delta)\in N_\varepsilon}\Phi(r,\delta) + \sup_{(r,\delta),(r',\delta')\in\Omega:\|r-r'\|+\|\delta-\delta'\|\le\varepsilon}|\Phi(r,\delta)-\Phi(r',\delta')| \\
&\le 2\sqrt{2d\log(10/\varepsilon)+\log(4/\gamma)}\left(\left(4R\sqrt{\log(d)}\right)^{1/4} + \sqrt{\frac{2d\log(10/\varepsilon)+\log(4/\gamma)}{m}}\right) \\
&\quad + C_0 + 3\sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})\sqrt{d+4\sqrt{d\log(8m/\gamma)}}} \\
&\quad + 5(C_0+\sqrt{\log(4/\gamma)})\cdot\left(\sqrt{\log(4m/\gamma)}+3\right) + \sup_{\substack{(r,\delta),(r',\delta')\in\Omega:\\ \|r-r'\|+\|\delta-\delta'\|\le\varepsilon}}|\Phi(r,\delta)-\Phi(r',\delta')|
\end{aligned} \tag{10}$$

For r, r', one has

$$|\Phi(r,\delta)-\Phi(r',\delta)| \le \frac{\|r-r'\|}{\sqrt{m}}\sum_{s=1}^m\|\mathrm{w}_s\| \le \frac{\|r-r'\|}{\sqrt{m}}\sum_{s=1}^m(\|\mathrm{w}_{s,0}\| + \frac{C_0}{\sqrt{m}})$$

Using (2), we know with probability at least $1-\gamma$, one has for all $s \in [m]$,

$$\|\mathrm{w}_{s,0}\|^2 \le d + 4\sqrt{d\log(m/\gamma)},$$

so that in this event we have,

$$|\Phi(r,\delta)-\Phi(r',\delta)| \le \|r-r'\|\left(\sqrt{md+4m\sqrt{d\log(m/\gamma)}} + C_0\right)$$

On the other hand, for δ, δ', we write

$$\begin{aligned}
&|\Phi(r,\delta)-\Phi(r,\delta')| \\
&\le \frac{1}{\sqrt{m}}\left|\sum_{s=1}^m \mathbb{1}\{\mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta)) > \mathrm{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\}a_s\mathrm{w}_s\cdot r\right|
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{m}}\left|\sum_{s=1}^{m}\mathbb{1}\{\operatorname{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta))<\operatorname{sign}(\mathrm{w}_s^\top(\mathrm{x}+\delta'))\}a_s\mathrm{w}_s\cdot r\right| \\
&\le \frac{1}{\sqrt{m}}\left\|\sup_{\substack{\forall s\in[m],\\ \mathrm{v}_s\in\mathbb{R}^d,\|\mathrm{v}_s\|\le V}}\sum_{s=1}^{m}\mathbb{1}\{\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta))>\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\}a_s\mathrm{w}_s\right\| \\
&\quad+\frac{1}{\sqrt{m}}\left\|\sup_{\substack{\forall s\in[m],\\ \mathrm{v}_s\in\mathbb{R}^d,\|\mathrm{v}_s\|\le V}}\sum_{s=1}^{m}\mathbb{1}\{\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta))<\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\}a_s\mathrm{w}_s\right\|
\end{aligned}
$$

Letting $X_s(\delta)=\mathbb{1}\{\exists\delta':\|\delta-\delta'\|\le\varepsilon \text{ and } \exists\mathrm{v}_s\in\mathbb{R}^d,\|\mathrm{v}_s\|\le V,\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta))\ne\operatorname{sign}((\mathrm{w}_{s,0}+\mathrm{v}_s)^\top(\mathrm{x}+\delta'))\}$, we now control with exponentially high probability $\sum_{s=1}^m X_s(\delta)$. By (4) in Lemma B.6, we know that $X_s(\delta)$ is a Bernoulli of parameter at most $p=2\varepsilon(\sqrt{d}+2\sqrt{d\log(2/\varepsilon)})+(1+R+\varepsilon)V$. Using Theorem B.2 and apply a union bound, we have

$$
P\left(\exists(r,\delta)\in N_\varepsilon:\sum_{s=1}^{m}X_s(\delta)\ge k\right)\le\left(\frac{10}{\varepsilon}\right)^{2d}\exp(-pm)\left(\frac{epm}{k}\right)^k:=g(p)
$$

We would like $g(p)\le\gamma$. Since $\sqrt{\varepsilon}(1+2\sqrt{\log(2/\varepsilon)})\le 4\varepsilon^{3/8}$ holds for $\varepsilon\in(0,1)$, we keep upper bound $p\le 8\sqrt{d}\varepsilon^{7/8}+3V$. Note that $p\le\frac{k}{m}$ to make sure $g(p)$ increases as p increases. Choose $\varepsilon=m^{-4/7}d^{-4/7}$, $V=\frac{C_0}{\sqrt{m}}$, $\bar{C}=\max\{C_0,1\}$, $k=44\bar{C}\sqrt{m}$, with $m\ge 58, m\ge d^{2.4}$, we have

$$
\begin{aligned}
\log g(8\sqrt{d}\varepsilon^{7/8}+\frac{3C_0}{\sqrt{m}}) &= -pm+2d\log\left(\frac{10}{\varepsilon}\right)+44\bar{C}\sqrt{m}\log\left(\frac{epm}{44\bar{C}\sqrt{m}}\right) \\
&\qquad\qquad\text{(Here } p=8\sqrt{d}\varepsilon^{7/8}+\tfrac{3C_0}{\sqrt{m}}\le\tfrac{11\bar{C}}{\sqrt{m}}\text{)} \\
&= -(8\sqrt{d}\varepsilon^{7/8}+\frac{3C_0}{\sqrt{m}})m+2d\log(10m^{4/7}d^{4/7})+44\bar{C}\sqrt{m}\log\left(\frac{11e\bar{C}\sqrt{m}}{44\bar{C}\sqrt{m}}\right) \\
&\le -8\sqrt{m}+2.3d\log(md) && (m\ge 58) \\
&\le -8\sqrt{m}+7\sqrt{m} && (m\ge d^{2.4}\implies\sqrt{m}\ge 0.33d\log(md)) \\
&\le -\sqrt{m} \\
&\le \log(\gamma)
\end{aligned}
$$

The last line holds for $m\ge d^{2.4}$ and $d\ge\log(4m/\gamma)^2\ge\log(4/\gamma)^2$.

By the concentration of Lipschitz functions of Gaussians (Theorem B.3) and a union bound, we have

$$
P\left(\exists S\in[m],|S|\le 44\bar{C}\sqrt{m},\left\|\frac{1}{\sqrt{m}}\sum_{s\in S}a_s\mathrm{w}_{s,0}\right\|\ge\sqrt{\frac{|S|}{m}}(\sqrt{d}+t)\right)\le m^{44\bar{C}\sqrt{m}}e^{-\frac{t^2}{2}}
$$

By setting $t=\sqrt{88\bar{C}\sqrt{m}\log(m)+2\log(\frac{8}{\gamma})}$, we get that with probability at least $1-\gamma/8$,

$$
\forall S\subset[m],|S|\le 44\bar{C}\sqrt{m}:\left\|\frac{1}{\sqrt{m}}\sum_{s\in S}a_s\mathrm{w}_{s,0}\right\|\le\sqrt{\frac{44\bar{C}}{\sqrt{m}}}\sqrt{88\bar{C}\sqrt{m}\log(m)+2\log(\frac{8}{\gamma})}+\sqrt{\frac{44\bar{C}d}{\sqrt{m}}}
$$

$$\forall S \subset [m], |S| \le 44\bar{C}\sqrt{m} : \left\| \sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{w}_s \right\|$$

$$\le \left\| \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{w}_{s,0} \right\| + \left\| \sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \frac{1}{\sqrt{m}} \sum_{i \in S} a_s \mathrm{v}_s \right\|$$

$$\le \sqrt{\frac{44\bar{C}}{\sqrt{m}}} \sqrt{88\bar{C}\sqrt{m}\log(m) + 2\log(\frac{8}{\gamma})} + \sqrt{\frac{44\bar{C}d}{\sqrt{m}}} + \frac{44\bar{C}C_0}{\sqrt{m}}$$

$$\le 113\bar{C}C_0 \sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}$$

Note that for all $(r, \delta) \in N$, $\|\delta - \delta'\| \le \varepsilon$,

$$\sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \mathbb{1}\{\mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top(\mathrm{x} + \delta)) < \mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top(\mathrm{x} + \delta'))\} \le X_s(\delta)$$

$$\sup_{\mathrm{v}_s \in \mathbb{R}^d, \|\mathrm{v}_s\| \le \frac{C_0}{\sqrt{m}}} \mathbb{1}\{\mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top(\mathrm{x} + \delta)) > \mathrm{sign}((\mathrm{w}_{s,0} + \mathrm{v}_s)^\top(\mathrm{x} + \delta'))\} \le X_s(\delta)$$

With probability at least $1-\gamma$, we have for all δ, δ', r, r' with $\|\delta - \delta'\| \le m^{-4/7}d^{-4/7}$ and $\|r - r'\| \le m^{-4/7}d^{-4/7}$,

$$|\Phi(r, \delta) - \Phi(r', \delta)| \le m^{-4/7}d^{-4/7}\left(\sqrt{md + 4m\sqrt{d\log(m/\gamma)}} + C_0\right)$$

$$|\Phi(r, \delta) - \Phi(r, \delta')| \le 226\bar{C}C_0\sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}$$

Combining this with (10) we obtain with probability at least $1 - \gamma$,

$$\sup_{(r,\delta)\in\Omega} \Phi(r, \delta)$$

$$\le 2\sqrt{2d\log(10m^{4/7}d^{4/7}) + \log(4/\gamma)}\left((4R\sqrt{\log(d)})^{1/4} + \sqrt{\frac{2d\log(10m^{4/7}d^{4/7}) + \log(4/\gamma)}{m}}\right)$$

$$+ C_0 + 3\sqrt{\frac{1}{\sqrt{m}}(C_0 + \sqrt{\frac{\log(4/\gamma)}{2}})}\sqrt{d + 4\sqrt{d\log(8m/\gamma)}}$$

$$+ 5(C_0 + \sqrt{\log(4/\gamma)}) \cdot \left(\sqrt{\log(4m/\gamma)} + 3\right) + m^{-4/7}d^{-4/7}\left(\sqrt{md + 4m\sqrt{d\log(m/\gamma)}} + C_0\right)$$

$$+ 226\bar{C}C_0\sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}$$

$$\le 6\sqrt{d\log(md)}\left((4R\sqrt{\log(d)})^{1/4} + \sqrt{\frac{6d\log(md)}{m}}\right) + 50d^{0.25} + 227\bar{C}C_0\sqrt{\log(m) + \frac{2}{\sqrt{m}}\log(\frac{8}{\gamma})}$$

$$\le 9\left(C_1 d^2 \log^2(md)\sqrt{\frac{\log(d)}{d}}\right)^{1/4} + \frac{15d\log(md)}{\sqrt{m}} + 327\bar{C}C_0 d^{0.25} \qquad (R \le \tfrac{C_1}{\sqrt{d}})$$

$$= o(\sqrt{d}) \quad \text{(holds when } \log^2(md)\sqrt{\tfrac{\log(d)}{d}} = o(1), \tfrac{d\log(md)}{\sqrt{m}} = o(\sqrt{d}) \implies m \le O(\exp(d^{0.24}))\text{)}$$

□

Proof of Corollary 3.2. From Theorem 3.1, we have that for any point $\mathrm{x} \in \mathbb{S}^{d-1}$, with probability at least $1-\gamma$, there exists an adversarial example for the neural network, with parameters (W, a), at x for perturbation size R. Thus, with probability at least $1-\gamma$, the robust error of W at x with perturbation R is one: $\ell_R(\mathrm{W}; \mathrm{x}, y) = 1$. This gives us that $\mathbb{E}_{\mathrm{W}}[\inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)] \geq 1-\gamma$. From Markov's inequality,

$$
\begin{aligned}
\mathbb{P}\left[\inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W}) < 0.9\right] &= \mathbb{P}\left[1 - \inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W}) > 0.1\right] \\
&\leq 10\mathbb{E}_{\mathrm{W}}\left[1 - \inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} L_R(\mathrm{W})\right] \\
&= 10\mathbb{E}_{\mathrm{x}}\mathbb{E}_{\mathrm{W}}\left[1 - \inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)\right] \\
&\leq 10 \sup_{\mathrm{x}} \mathbb{E}_{\mathrm{W}}\left[1 - \inf_{\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{d}}\right)} \ell_R(\mathrm{W}; \mathrm{x}, y)\right] \\
&\leq 10\gamma
\end{aligned}
$$

where in the second equality we swap the expectations using Fubini's theorem. □