A Preliminaries

In this section, we recall some standard concepts and results in statistical learning theory.

Definition A.1 (growth function)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X} \subset \mathbb{R}^d$ to $\{-1, +1\}$. For any integer $m \geq 0$, we define the growth function of $\mathcal{F}$ to be*

$$\Pi_{\mathcal{F}}(m) = \max_{x_i \in \mathcal{X}, 1 \leq i \leq m} \left|\{(f(x_1), f(x_2), \cdots, f(x_m)) : f \in \mathcal{F}\}\right|.$$

In particular, if $\left|\{(f(x_1), f(x_2), \cdots, f(x_m)) : f \in \mathcal{F}\}\right| = 2^m$, then $(x_1, x_2, \cdots, x_m)$ is said to be shattered by $\mathcal{F}$.

Definition A.2 (Vapnik-Chervonenkis dimension)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X} \subset \mathbb{R}^d$ to $\{-1, +1\}$. The VC-dimension of $\mathcal{F}$, denoted by VC-dim$(\mathcal{F})$, is defined as the largest integer $m \geq 0$ such that $\Pi_{\mathcal{F}}(m) = 2^m$. For real-value function class $\mathcal{H}$, we define VC-dim$(\mathcal{H})$:= VC-dim$(\text{sgn}(\mathcal{H}))$.*

The following result gives a nearly-tight upper bound on the VC-dimension of neural networks.

Lemma A.3. *(Bartlett et al., 2019, Theorem 6) Consider a ReLU network with L layers and W total parameters. Let F be the set of (real-valued) functions computed by this network. Then we have VC-dim$(F) = O(W \log(WL))$.*

The growth function is connected to the VC-dimension via the following lemma; see e.g. (Anthony et al., 1999, Theorem 7.6).

Lemma A.4. *Suppose that VC-dim$(\mathcal{F}) = k$, then $\Pi_m(\mathcal{F}) \leq \sum_{i=0}^{k} \binom{m}{i}$. In particular, we have $\Pi_m(\mathcal{F}) \leq (em/k)^k$ for all $m > k + 1$.*

Lemma A.5. *(Mohri et al., 2018, Corollary 3.4) Let H be a family of functions taking values in $\{-1, +1\}$ with VC-dimension k. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over $m-$samples training dataset S i.i.d. drawn from the data distribution D, the following holds for all $h \in H$:*

$$\mathcal{L}_D(h) \leq \mathcal{L}_S(h) + \sqrt{\frac{2k \log \frac{em}{k}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

where $\mathcal{L}_D(h)$ and $\mathcal{L}_S(h)$ denote the standard test error and training error, respectively.

For deriving upper and lower bounds in the context of ℓ_2-robustness, we also need to introduce the following concepts.

Definition A.6 (ϵ-covering)**.** *Given a set $\Theta \subset \mathbb{R}^d$, we say that $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \subset \Theta$ is a δ-covering of Θ if $\Theta \subset \cup_{i=1}^{n} \mathcal{B}_2(\boldsymbol{x}_i, \delta)$. The covering number $\mathcal{C}(\Theta, \delta)$ is defined as the minimal size of a δ-covering set of Θ.*

The following proposition is straightforward from the definition.

Proposition A.7. *Let $\Theta \subset \mathbb{R}^d$ has volume (i.e. Lebesgue measure) V, then*

$$\mathcal{C}(\Theta, \delta) \geq v_d \cdot \delta^{-d} V,$$

where v_d is the volume of a d-dimensional unit ball.

Definition A.8 (ϵ-packing)**.** *Given a set $\Theta \subset \mathbb{R}^d$, we say that $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \subset \Theta$ is a δ-packing of Θ if $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geq \delta, \forall i \neq j$. The packing number $\mathcal{P}(\Theta, \delta)$ is defined as the maximal size of a δ-packing set of Θ.*

The relationship between the covering and packing number is given by the following result. For completeness, we also provide a simple proof.

Proposition A.9. *For any $\delta \geq 0$, we have $\mathcal{P}(\Theta, \delta) \geq \mathcal{C}(\Theta, \delta)$.*

Proof. Consider a maximal packing $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$. Pick any $\boldsymbol{x} \in \Theta$, then there must exists some $\boldsymbol{x}_i \in X$ such that $\|\boldsymbol{x} - \boldsymbol{x}_i\|_2 \leq \delta$; otherwise, $X \cup \{\boldsymbol{x}\}$ is a larger packing set, which contradicts the definition of X.

Hence it must holds that $\Theta \subset \cup_{i=1}^{n} \mathcal{B}_2(\boldsymbol{x}_i, \delta)$ i.e. X is a δ-covering of Θ. The conclusion follows. □

B Proofs for Section 2

To prove Theorem 2.2, we first recall some well-known results of neural networks for approximating simple functions.

Lemma B.1. *Let $\varepsilon > 0$, $0 < a < b$ and $B \geq 1$ be given.*

(1). (Yarotsky, 2017, Proposition 3) There exists a function $\widetilde{\times} : [0, B]^2 \to [0, B^2]$ computed by a ReLU network with $\mathcal{O}\left(\log^2\left(\varepsilon^{-1}B\right)\right)$ parameters such that

$$\sup_{x,y\in[0,B]} \left|\widetilde{\times}(x, y) - xy\right| \leq \varepsilon,$$

and $\widetilde{\times}(x, y) = 0$ if $xy = 0$.

(2). (Telgarsky, 2017, Lemma 3.5) There exists a function $R : [a, b] \to \mathbb{R}_+$ computed by a ReLU network with $\mathcal{O}\left(\log^4\left(a^{-1}b\right)\log^3(\varepsilon^{-1}b)\right)$ parameters such that $\sup_{[a,b]} \left|R(x) - \frac{1}{x}\right| \leq \varepsilon$.

The following lemma establishes uniform approximation of polynomials and is a slight generalization of (Telgarsky, 2017, Lemma 3.4).

Lemma B.2. *Let $\varepsilon \in (0, 1)$. Suppose that $P(x) = \sum_{k=1}^{s} \alpha_k \prod_{i=1}^{r_k} (x_{k,i} - a_{k,i})$ is a polynomial with $\max_k r_k = r$ and $\alpha_k, a_{k,i} \in [0, 1], \forall 1 \leq k \leq s, 1 \leq i \leq r_k$, and $P(x) \in [-1, +1]$ for $\forall x \in [0, 1]^d$. Then there exists a function $N(x)$ computed by a ReLU network with $\mathcal{O}\left(sr\log\left(\varepsilon^{-1}sr\right)\right)$ parameters such that $\sup_{[0,1]^d} |P(x) - N(x)| \leq \varepsilon$.*

Proof. It suffices to show that each monomial $P_k(x) = \prod_{i=1}^{r_k} (x_{k,i} - a_{k,i})$ can be ε-approximated using $\mathcal{O}\left(r\log\left(\varepsilon^{-1}r\right)\right)$ parameters. Firstly, we need at most $r_k \leq r$ parameters to obtain $x_{k,i} - a_{k,i}, 1 \leq i \leq r_k$ from a linear transformation. We can then apply Lemma B.1 to perform successive multiplication. Note that we still have $|x_{k,i} - a_{k,i}| \leq 1$, which can be used to control the cumulative error of $\widetilde{\times}$. □

We are now ready to prove Theorem 2.2. For convenience, we restate this theorem below.

Theorem B.3. *Suppose that $\mathcal{D} \subset \mathcal{B}_p(0, 1)$ with $p \in \{2, +\infty\}$ consists of N data, and the two classes in $\mathcal{D}$ are 2ϵ-separated (cf. Definition 1.1), where $\epsilon \in \left(0, \frac{1}{2}\right)$ is a constant. Let robustness radius $\delta < \frac{1}{2}\epsilon$, then there exists a classifier f represented by a ReLU network with at most*

$$\mathcal{O}\left(Nd\log\left(\delta^{-1}d\right) + N \cdot \mathrm{polylog}(\delta^{-1}N)\right)$$

parameters, such that $\hat{\mathcal{L}}_{\mathcal{D}}^{p,\delta}(f) = 0$.

Proof. (1). The case $p = 2$. First, we choose $C, \varepsilon_1, \varepsilon_2 > 0$ and $m \in \mathbb{Z}_+$ that satisfy

$$C\left((\delta^2 + \varepsilon_1)^m + \varepsilon_2\right) \leq \frac{1}{4} < 4N \leq C\left((R^2 - \varepsilon_1)^m - \varepsilon_2\right). \tag{2}$$

These constants will be specified later. Since for $\forall \boldsymbol{x}_0 \in [0, 1]^d$, $\boldsymbol{x} \to \|\boldsymbol{x} - \boldsymbol{x}_0\|^2$ is a polynomial that consists of d monomials and with degree 2, satisfying the conditions in Lemma B.2, there exists a function ϕ_1 computed by a ReLU network with $\mathcal{O}\left(d\log\left(\varepsilon_1^{-1}d\right)\right)$ parameters such that $\sup_{\boldsymbol{x}\in[0,1]^d} \left|\phi_1(\boldsymbol{x}) - \|\boldsymbol{x} - \boldsymbol{x}_0\|^2\right| \leq \varepsilon_1$. We may further assume that $\phi\left([0, 1]^d\right) \subset [0, 1]$, or otherwise we can consider $\sigma\left(\phi_1(\boldsymbol{x})\right) - \sigma\left(\phi_1(\boldsymbol{x}) - 1\right)$ instead.

Applying Lemma B.2 again, we can see that the function $x \to x^m$ on $[0, 1]$ can be approximated with error ε_2 by a function ϕ_2 computed by a ReLU network with $\mathcal{O}\left(m\log\left(\varepsilon^{-1}m\right)\right)$ parameters. Now we can see that $1 + C \cdot \phi_2 \circ \phi_1$ is computable by a ReLU network and takes value in $\left[1, \frac{5}{4}\right]$ when $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}_0, \delta)$ and in $(4N + 1, C + 1)$ when $\boldsymbol{x} \notin \mathcal{B}(\boldsymbol{x}_0, R)$ (since $R \leq 1$).

The final step is to choose ϕ_3 computed by a ReLU network with $\mathcal{O}\left(\log^4 C\log^3(NC)\right)$ parameters such that it approximates $\frac{1}{x}$ on $[1, C + 1]$ with error $< \frac{1}{4N}$. Hence $\phi_3 \circ (1 + C \cdot \phi_2 \circ \phi_1)$ is larger

than $\frac{3}{4}$ inside $\mathcal{B}(\boldsymbol{x}_0, \delta)$ and smaller than $\frac{1}{2N}$ outside $\mathcal{B}(\boldsymbol{x}_0, R)$. This construction uses a total of $\mathcal{O}(W)$ parameters, where

$$W = d \log\left(\varepsilon_1^{-1} d\right) + m \log\left(\varepsilon_2^{-1} m\right) + \log^4 C \log^3(NC). \tag{3}$$

Finally, we choose

$$\varepsilon_1 = \frac{R\delta(R-\delta)}{R+\delta}, \quad m = \max\left\{1, \log \frac{32N\delta}{R}\right\}, \quad \varepsilon_2 = \frac{1}{33N}\left(\frac{R(R^2+\delta^2)}{R+\delta}\right)^m,$$

and $C = \frac{4N}{(R^2-\varepsilon_1)^m - \varepsilon_2} = \mathcal{O}\left(N\delta^{-2m}\right)$, which satisfies (2). Plugging all expressions into (3), we can see that

$$W = \mathcal{O}\left(d\left(\log d + \log \delta^{-1} + \log(R-\delta)^{-1}\right) + \log^7\left(\delta^{-1}N\right)\right).$$

We denote this construction by $\psi(\boldsymbol{x}; \boldsymbol{x}_0, \boldsymbol{\theta})$, where θ consists of all parameters. The arguments above show that there exists $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x_0})$ such that $\psi(\boldsymbol{x}; \boldsymbol{x_0}, \boldsymbol{\theta}) > \frac{3}{4}$ when $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}_0, \delta)$ and $\psi(\boldsymbol{x}; \boldsymbol{x}_0, \boldsymbol{\theta}) < \frac{1}{2N}$ when $\boldsymbol{x} \notin \mathcal{B}(\boldsymbol{x}_0, R)$. Consider the function $\Psi(\boldsymbol{x}; \boldsymbol{\theta}_{1:N}) = 4\sum_{i=1}^N \psi(\boldsymbol{x}; \boldsymbol{x}_i, \boldsymbol{\theta}_i) - \frac{5}{2}$. The total number of parameters in Ψ is $\widetilde{\mathcal{O}}(Nd)$. Moreover, if we choose $\boldsymbol{\theta}_i = \boldsymbol{\theta}(\boldsymbol{x}_i)$ when $y_i = 1$ and $\boldsymbol{\theta}_i = 0$ when $y_i = -1$, then Ψ satisfies the condition in Theorem B.3.

(2). The case $p = \infty$. To obtain the same result under the ℓ_∞ norm, it suffices to construct a neural network with size $\mathcal{O}(d)$ parameters to represent the function $x \to \|x - x_0\|_\infty$; the remaining steps are exactly the same with the ℓ_2 case.

Let $x^{(i)}$ denote the i-th coordinate of $\boldsymbol{x}$, then $\|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty = \max_{1\le i\le d}\left|x^{(i)} - x_0^{(i)}\right|$. Since

$$|a| = \frac{1}{2}\left(\max\{a, 0\} + \max\{-a, 0\}\right),$$

we can see that $x^{(i)} \to \left|x^{(i)} - x_0^{(i)}\right|$ can be represented by a constant-size ReLU network. Moreover, the function $\max\{a, b\} = \frac{1}{2}\left(|a+b| + |a-b|\right)$, so that the function $(a_1, a_2, \cdots, a_d) \to \max_{1\le i\le d} a_d$ can be represented with $\mathcal{O}(d)$ parameters. To summarize, $\boldsymbol{x} \to \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty$ can be represented using a ReLU network of size $\mathcal{O}(d)$, as desired. □

In the following, we prove Theorem 2.3.

Theorem B.4 (Restatement of Theorem 2.3)**.** *Let $p \in \{2, +\infty\}$ and $\mathcal{F}_n$ be the set of functions represented by some ReLU network with at most n parameters. If for any 2ϵ-separated data set $\mathcal{D}$ under ℓ_p norm, there exists a classifier $f \in \mathcal{F}_n$ such that $\hat{\mathcal{L}}_{\mathcal{D}}^{p,\delta}(f) = 0$, then it must hold that $n = \Omega(\sqrt{Nd})$.*

Proof. It follows from the assumption that given any data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ which are pairwise 2ϵ-separated, there exists $f \in \mathcal{F}_n$ being able to achieve zero training error for any binary label. It directly follows from (Gao et al., 2019, Theorem 6.1) that

$$\text{VC-dim}(\mathcal{F}_n) = \Omega(Nd).$$

On the other hand, suppose that $L \neq n$ is the depth of the neural network, then we have

$$\text{VC-dim}(\mathcal{F}_n) = \mathcal{O}(nL\log(nL)) = \mathcal{O}(n^2).$$

As a result, it follows that $n = \tilde{\Omega}(\sqrt{Nd})$, as desired. □

C Proofs for Section 3

C.1 Proof of Theorem 3.3

The proof idea of Theorem 3.3 has two key steps. First, we construct a Lipschitz classifier f^* based on distance function between a point and a close set that can $\epsilon-$ robustly classify A, B. Then we regard f^* as the target function and use a ReLU network to approximate it to derive the $c\epsilon-$robust classifier. Before proving the theorem, we first introduce the two following useful conclusions, which also corresponding to the two steps of proof.

Proposition C.1. *For the separable $A, B \subset [0,1]^d$, we define $f^*(\boldsymbol{x}) := \frac{d_\infty(\boldsymbol{x},B)-d_\infty(\boldsymbol{x},A)}{d_\infty(\boldsymbol{x},A)+d_\infty(\boldsymbol{x},B)}$, which has the following properties:*

1. *$f^*(\boldsymbol{x})$ can classify A, B correctly i.e. $f^*(\boldsymbol{x}) = \begin{cases} 1, & x \in A \\ -1, & x \in B \end{cases}$.*
2. *$f^*(\boldsymbol{x})$ is a ϵ-robust classifier i.e. for any perturbed input x' that satisfies $||\boldsymbol{x}' - \boldsymbol{x}||_\infty \leq \epsilon$ can also be classified correctly.*
3. *$f^*(\boldsymbol{x})$ is $\frac{1}{\epsilon}$-Lipschitz w.r.t. ℓ_∞ norm.*

We can check these properties by the continuity and 1-Lipschitz property of distance function $d_\infty(p, S)$.

Lemma C.2. *For any $L-$lipschitz function f in $[0,1]^d$, there exists a function $\tilde{f}$ implemented by ReLU network with at most $c_1(c_2\epsilon/L)^{-d}(d^2+d\log d+d\log(1/\epsilon))$ parameters that satisfies $|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})| \leq \epsilon$ for any $\boldsymbol{x} \in [0,1]^d$, where c_1 and c_2 are constants.*

This lemma provides a useful approximation tool for us, which is an improved version of Theorem 1 in Yarotsky (2017). Compared with Theorem 1 in Yarotsky (2017), we use Lipschitz property of function instead of high-order differetiability and focus on not the bound order when ϵ goes to zero but also more accurate bound order depending on ϵ, L and d. By a refined analysis of total approximation error, we can derive this lemma.

Proof of Theorem 3.3. By Lemma C.2, we can approximate f^* in Lemma C.1 satisfying uniform error at most $1-c$ via a ReLU network f with at most $c_1(c_2(1-c)\epsilon)^{-d}(d^2+d\log d+d\log(1/(1-c)))$ parameters. Then, we prove the theorem by contradiction. Assume that there exists some perturbed input $\boldsymbol{x}'$ that is mis-classified and the original input x is in A. So we know $f(\boldsymbol{x}') < 0$ and $f^*(\boldsymbol{x}) < \epsilon'$. This impiles $d_\infty(\boldsymbol{x}', A) < d_\infty(\boldsymbol{x}', B) < \frac{1+\epsilon'}{1-\epsilon'} d_\infty(\boldsymbol{x}', A)$ Combined with $d_\infty(\boldsymbol{x}', A)+d_\infty(\boldsymbol{x}', B) \geq d_\infty(A, B) \geq 2\epsilon$, we have $d_\infty(\boldsymbol{x}', A) > (1-\epsilon')\epsilon = c\epsilon$, which is the contradiction. □

C.2 Proof of Theorem 3.4

The main idea of proof is to estimate the lower bound of the family's VC-dimension via the definition of $c\epsilon$-robust family.

Proof of Theorem 3.4. The key idea is to find some discrete points that can be shattered by the function family $\mathcal{F}_n$.

(1). The $p = \infty$ case. We use K to denote $\lfloor \frac{1}{2\epsilon} \rfloor + 1$, and we can divide $[0,1]^d$ into $(K-1)^d$ non-overlapping sub-cubes. Let S be the set of all the vertices of sub-cubes, which has K^d elements and can be represented by

$$S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{K^d}\} = \{(2\epsilon i_1, 2\epsilon i_2, \cdots, 2\epsilon i_d) | 0 \leq i_1, i_2, \cdots, i_d < K\}.$$

For any partition I, J of $[K^d]$ ($I \cap J = \Phi, I \cup J = [K^d]$), let $A = \{\boldsymbol{x}_i | i \in I\}$ and $B = \{\boldsymbol{x}_j | j \in J\}$ be the positive and negative classes. Then we have $d_\infty(A, B) \geq 2\epsilon$. By the definition of $c\epsilon$-robust classifier family, there exists a classifier $f \in F$ classify A, B correctly. Thus, the family F shatter the subset $S \subset [0,1]^d$. By using the conclusion of Lemma A.3, we have $K^d \leq \text{VC-dim}(F) = \mathcal{O}(WL\log(N)) = \mathcal{O}(W^2\log(W))$ where L is the depth of networks and W is the total number of parameters.

(2). The $p = 2$ case. Similar to the case of $p = \infty$, we need to construct a set $S \subset \mathcal{B}_2(0,1)$ such that the ℓ_2-distance between any two points in S is as least 2ϵ.

Specifically, we choose S to be a 2ϵ-packing of $\mathcal{B}_2(0,1)$ with maximal size. Then we have that $|S| \geq \mathcal{P}(\mathcal{B}_2(0,1), 2\epsilon) \geq \mathcal{C}(\mathcal{B}_2(0,1), 2\epsilon) \geq (2\epsilon)^{-d}$, by Propositions A.7 and A.9. Similar to the $p = \infty$ case, robustness implies that S can be shattered by $\mathcal{F}_n$, so that $K^d = \mathcal{O}(W^2\log W)$ and the conclusion follows. □

D Proofs for Section 4

In this section, we present the proof of Theorem 4.3 and 4.4.

Theorem D.1 (Restatement of Theorem 4.3)**.** *Let $\epsilon \in (0,1)$ be a small constant, $p \in \{2,\infty\}$ and $\mathcal{F}_n$ be the set of functions represented by ReLU networks with at most n parameters. There exists a sequence $N_d = \Omega\left((2\epsilon)^{-\frac{d-1}{6}}\right), d \geq 1$ and a universal constant $C > 0$ such that the following holds: for any $c \in (0,1)$, there exists two linear separable sets $A, B \subset [0,1]^d$ that are 2ϵ-separated under ℓ_p-norm, such that for any μ_0-balanced distribution P on the supporting set $S = A \cup B$ and robust radius $c\epsilon$ we have*

$$\inf\left\{\mathcal{L}_P^{p,c\epsilon}(f) : f \in \mathcal{F}_{N_d}\right\} \geq C\mu_0.$$

Proof. (1). The $p = \infty$ case. Define

$$S_\phi = \left\{\left(\frac{i_1}{K}, \frac{i_2}{K}, \cdots, \frac{i_{d-1}}{K}, \frac{1}{2} + c\varepsilon \cdot \phi(i_1, i_2, \cdots, i_{d-1})\right) : 1 \leq i_1, i_2, \cdots, i_{d-1} \leq K\right\},$$

and

$$\widetilde{S} = \left\{\left(\frac{i_1}{K}, \frac{i_2}{K}, \cdots, \frac{i_{d-1}}{K}, \frac{1}{2}\right) : 1 \leq i_1, i_2, \cdots, i_{d-1} \leq K\right\},$$

where $K = \lfloor \frac{1}{2\varepsilon} \rfloor$, and $\phi : \{1, 2, \cdots, K\}^{d-1} \to \{-1, +1\}$ be an arbitrary mapping. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we use $x^{(i)}$ to denote its i-th component. Let $A_\phi = S_\phi \cap \left\{\boldsymbol{x} \in \mathbb{R}^d : x^{(d)} > \frac{1}{2}\right\}$, $B_\phi = S_\phi - A_\phi$ and μ be the uniform distribution on S. It's easy to see that A and B are linear separable by the hyperplane $x^{(d)} = \frac{1}{2}$. Moreover, we clearly have $d(A, B) \geq 2\varepsilon$. We will show that there exists some choice of ϕ such that robust classification of A_ϕ and B_ϕ with $(c\varepsilon, 1-\alpha)$-accuracy requires at least $\Omega\left(K^{(d-1)/6}\right)$ parameters.

Assume that for any choices of ϕ, the induced sets A_ϕ and B_ϕ can always be robustly classified with $(c\varepsilon, 1-\alpha)$-accuracy by a ReLU network with at most M parameters. Then, we can construct an *enveloping network* $F_{\boldsymbol{\theta}}$ with $M-1$ hidden layers, M neurons per layer and at most M^3 parameters such that any network with size $\leq M$ can be embedded into this envelope network. As a result, $F_{\boldsymbol{\theta}}$ is capable of $(c\varepsilon, 1-\alpha)$-robustly classify any sets A_ϕ, B_ϕ induced by arbitrary choices of ϕ. We use R_ϕ to denote the subset of $S_\phi = A_\phi \cup B_\phi$ satisfying $|R_\phi| = (1-\alpha)\,|S_\phi| = (1-\alpha)K^{d-1}$ such that R_ϕ can be $c\varepsilon$-robustly classified.

Consider the projection operator $\mathcal{P}$ onto the hyperplane $x^{(d)} = \frac{1}{2}$. For any set $C \in \mathbb{R}^d$, we use $\widetilde{C}$ to denote $\mathcal{P}(C)$. Then $c\varepsilon$-robustness implies that the labelled data set

$$R_\phi^+ = \left\{(\boldsymbol{x}, y) : \boldsymbol{x} \in \widetilde{R}_\phi, y = \phi\left(Kx^{(1)}, \cdots, Kx^{(d-1)}\right)\right\}$$

can be correctly classified by $F_{\boldsymbol{\theta}}$, with appropriate choices of parameters.

Let $V = \frac{1}{2}K^{d-1}$ and $\hat{\mathcal{R}}_\phi$ be the collection of all labelled V-subset (i.e. subset of size V) of R_ϕ^+. For each V-subset R of $\widetilde{S}$, we use $\mathcal{G}(R)$ to denote the set of all labelings of R, so that $|\mathcal{G}(R)| = 2^V$.

Note that for each labelled V-subset T, there exists at most $2^{K^{d-1}-V}$ different choices of ϕ such that $T \subset R_\phi^+$ (or, equivalently, $T \in \hat{\mathcal{R}}_\phi$): this is because the value of ϕ on data points in T has been specified by their labels, and there are two choices for each of the remaining $K^{d-1} - V$ points in $\{1, 2, \cdots, K\}^{d-1}$. As a result, we have

$$\left|\cup_\phi \hat{\mathcal{R}}_\phi\right| \geq 2^{-(K^{d-1}-V)} \sum_\phi \left|\hat{\mathcal{R}}_\phi\right| = 2^V \binom{(1-\alpha)K^{d-1}}{V}.$$

On the other hand, the total number of V-subset of $\widetilde{S}$ is $\binom{K^{d-1}}{V}$, thus there must exists a V-subset $\mathcal{V}_0 \subset \widetilde{S}$, such that at least

$$\binom{K^{d-1}}{V}^{-1} \cdot 2^V \binom{(1-\alpha)K^{d-1}}{V} \geq \left(\frac{2\left((1-\alpha)k^{d-1} - V\right)}{K^{d-1} - V}\right)^V \tag{4}$$

different labelings of $\mathcal{V}_0$ are included in $\cup_\phi \hat{\mathcal{R}}_\phi$. Since $F_{\boldsymbol{\theta}}$ can correctly classify all elements (which are V-subsets) in $\cup_\phi \hat{\mathcal{R}}_\phi$, it can in particular classify the set $\mathcal{V}_0$ with at least $\left(\frac{2\left((1-\alpha)k^{d-1}-V\right)}{K^{d-1}-V}\right)^V$ different assignments of labels. Let d_{VC} be the VC-dimension of $F_{\boldsymbol{\theta}}$, then by Lemma A.4, either $d_{VC} \geq V = \frac{1}{2}K^{d-1}$, or

$$(2(1-2\alpha))^V \leq \left(\frac{2\left((1-\alpha)K^{d-1}-V\right)}{K^{d-1}-V}\right)^V \leq \Pi_{F_{\boldsymbol{\theta}}}(V) \leq \left(\frac{eV}{d_{VC}}\right)^{d_{VC}},$$

where Π is the growth function. The RHS is increasing in d_{VC} as long as $d_{VC} \leq V$. When $\alpha \leq \frac{1}{10}$, we have $2(1-2\alpha) > (10e)^{1/10}$, so that $d_{VC} \geq \frac{1}{10}V = \frac{1}{20}K^{d-1}$. Finally, since $F_{\boldsymbol{\theta}}$ has at most M^3 parameters, classical bounds on VC-dimension (Bartlett et al., 2019) imply that $M = \Omega\left(K^{(d-1)/6}\right)$, as desired.

(2). The $p = 2$ case. Let P be an 2ϵ-packing of the unit ball $\mathcal{B}_{d-1}$ in $\mathbb{R}^{d-1}$. Since the packing number $\mathcal{P}(\mathcal{B}_{d-1}, \|\cdot\|, 2\epsilon) \geq \mathcal{C}(\mathcal{B}_{d-1}, \|\cdot\|_2, 2\epsilon) \geq (2\epsilon)^{-(d-1)}$ by Propositions A.7 and A.9, where $\mathcal{C}(\Theta, \|\cdot\|, \epsilon)$ is the ϵ-covering number of a set Θ. For any $\lambda \in (0,1)$, we can consider the construction

$$S_\phi = \left\{\left(\boldsymbol{x}, \frac{1}{2} + \epsilon_0 \cdot \phi(\boldsymbol{x})\right) : \boldsymbol{x} \in P\right\},$$

where $\phi : P \to \{-1, +1\}$ is an arbitrary mapping. It's easy to see that all points in S_ϕ with first $d-1$ components satisfying $\|\boldsymbol{x}\|_2 \leq \sqrt{1-\varepsilon_0^2}$ are in the unit ball $\mathcal{B}_d$, so that by choosing ε_0 sufficiently small, we can guarantee that $|S_\phi \cap \mathcal{B}_d| \geq \frac{1}{2}(2\epsilon)^{-(d-1)}$. For convenience we just replace S_ϕ with $S_\phi \cap \mathcal{B}_d$ from now on.

Let $A_\phi = S_\phi \cap \left\{\boldsymbol{x} \in \mathbb{R}^d : x^{(d)} > \frac{1}{2}\right\}$, $B_\phi = S_\phi - A_\phi$. It's easy to see that for arbitrary ϕ, the construction is linear-separable and satisfies 2ϵ-separability. The remaining steps are just identical to the ℓ_∞ case. □

Theorem D.2 (Restatement of Theorem 4.4)**.** *For any two linear-separable $A, B \subset [0,1]^d$, a distribution P on the supporting set $S = A \cup B$, $\delta > 0$ and $\beta > 0$, let H be the family of $d-$dimensional hyperplane classifiers. Then, there exists a poly-time efficient algorithm $\mathcal{A} : 2^S \to H$, for $N = \Omega(d/\beta^2)$ training instances independently randomly sampled from P, with probability $1-\delta$ over samples, we can use the algorithm $\mathcal{A}$ to learn a classifier $\hat{f} \in F$ such that*

$$\mathcal{L}_P(\hat{f}) \leq \beta,$$

where $\mathcal{L}_P(f) := \mathbb{P}_{(x,y)\sim P}\{y \neq f(x)\}$ denotes the standard test error.

Proof. We i.i.d. sample N instances from the data distribution P, and use T to denote the training dataset. By Lemma A.5, with probabililty at least $1-\delta$, we have

$$\mathcal{L}_P(h) \leq \mathcal{L}_T(h) + \mathcal{O}\left(\sqrt{\frac{d}{N}}\right), \forall h \in H$$

By conclusions of Boser et al. (1992) and results of convex optimization, we have a poly-time algorithm $\mathcal{A} : 2^S \to H$ such that $\mathcal{L}_T(\mathcal{A}(T)) \leq \frac{\beta}{2}$, and we use $\hat{f}$ to denote $\mathcal{A}(T)$. Finally, when $N = \Omega(d/\beta^2)$ is sufficient large, we have $\mathcal{L}_P(\hat{f}) \leq \frac{\beta}{2} + \frac{\beta}{2} = \beta$. □

E Proofs for Section 5

E.1 Proof of Lemma 5.6

Theorem E.1 (Restatement of Lemma 5.6)**.** *Let $\mathcal{M} \subset [0,1]^d$ be a $k-$dimensional compact poly-partitionable Riemannian manifold with the condition number $\tau > 0$. For any small $\delta > 0$ and a $L-$lipschitz function $f : \mathcal{M} \to \mathbb{R}$, there exists a function $\hat{f}$ implemented by ReLU network with at most*

$$\tilde{\mathcal{O}}\left((\sqrt{d}L/\delta)^{-\tilde{k}}\right)$$

parameters, such that $|f - \hat{f}| < \delta$ *for any* $x \in \mathcal{M}$*, where* $\tilde{k}$ *is the same as Theorem 5.5.*

Proof. The full proof has six steps, and we finally construct a ReLU network as the following form

$$\hat{f} = \sum_{i=1}^{N} (\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i) \hat{\times} (\hat{I}_\theta \circ \hat{d}_i^2),$$

where all $\hat{f}_i, \phi_i, \hat{\rho}_i, T_i, \hat{I}_\Delta, \hat{d}_i^2$ and $\hat{\times}$ are implemented by sub ReLU networks, and these notation's detail will be introduced step by step.

As the above form shows, three sub-network groups will be combined by multiplication approximator $\hat{\times}$, where each group corresponds to a factor in the partition-of-unity-based decomposition of f (i.e. $f = \sum_{i=1}^{N} f \times \rho_i \times I\{x \in U_i\}$, and where $\{\rho_i\}_{i\in[N]}$ satisfying $\sum_{i\in[N]} \rho_i = 1$ is a partition of unity on an atlas $\{U_i\}_{i\in[N]}$).

Step 1: Construct poly-partition of unity on $\mathcal{M}$

Consider a open cover $\{B_r(x)\}_{x\in\mathcal{M}}$ on $\mathcal{M}$, where we use $B_r(c)$ to denote the Euclidean neighborhood with center c and radius r. Due to the compactness of manifold $\mathcal{M}$, we know there exists a finite open cover $\{B_r(x_i)\}_{i\in I}$ indexed by a finite sub-index set I, which satisfies $\mathcal{M} \subset \bigcup_{i\in I} B_r(x_i)$.

Then, we estimate the cardinal number of index set. By the conclusions of Niyogi et al. (2008), when we select radius r satisfying $r < \tau/2$, we have the following lemma, which gives an lower bound of $k-$dimensional volume of the local neighborhood of $\mathcal{M}$.

Lemma E.2. *(Niyogi et al., 2008, Lemma 5.3) Let* $c \in \mathcal{M}$*. Now consider* $U = \mathcal{M} \cap B_r(c)$*. Then* $\mathrm{vol}(U) \geq (\cos(\theta))^k \mathrm{vol}\left(B_r^k(c)\right)$ *where* $B_r^k(c)$ *is the* k*-dimensional ball in* T_c *centered at* $c, \theta = \arcsin(r/2\tau)$*. All volumes are* k*-dimensional volumes where* k *is the dimension of* $\mathcal{M}$*.*

Recall the relation between the covering number $\mathcal{N}(\mathcal{M}, d_2, r)$ and the packing number $\mathcal{P}(\mathcal{M}, d_2, r)$, and then we have

$$\begin{aligned}
\mathcal{N}(\mathcal{M}, d_2, r) &\leq \mathcal{P}(\mathcal{M}, d_2, r/2) \\
&\leq \frac{\mathrm{vol}(\mathcal{M})}{(\cos(\theta))^k \mathrm{vol}\left(B_{r/2}^k(c)\right)} \\
&\leq c_N \frac{\mathrm{vol}(\mathcal{M})}{r^k},
\end{aligned}$$

where c_N is a constant that only exponentially depends on $k \log k$.

By the poly-partitionable and smoothness properties of Riemannian manifold $\mathcal{M}$, there exists a collection $\{U_i, T_i, \rho_i\}_{i\in\mathcal{N}(\mathcal{M},d_2,r)}$ such that $\{U_i, T_i\}$ compose a tangent-space-induced atlas and $\{\rho_i\}$ also compose a poly-partition of unity on $\mathcal{M}$. So we can decompose f as $f = \sum_{i=1}^{N} f\rho_i$, where we use notation N to denote $\mathcal{N}(\mathcal{M}, d_2, r)$ so as to simplify the written process.

Step 2: Local almost isotropic transformation via random projection

To achieve dimensional reduction of f in the local neighborhood U_i, we will use the following random projection technique proposed by Baraniuk and Wakin (2009).

Lemma E.3. *(Baraniuk and Wakin, 2009, Theorem 3.1)*

Let $\mathcal{M}$ *be a compact* k*-dimensional sub-manifold of* $\mathbb{R}^d$ *having condition number* $1/\tau$*. Fix* $0 < \delta < 1$ *and* $0 < \eta < 1$*. Let* A *be a random orthoprojector from* $\mathbb{R}^d$ *to* $\mathbb{R}^{\tilde{k}}$ *with*

$$\tilde{k} = O\left(\frac{k \log\left(d\,\mathrm{vol}(\mathcal{M})\tau^{-1}\delta^{-1}\right)\log(1/\eta)}{\delta^2}\right).$$

If $\tilde{k} \leq d$*, then with probability at least* $1 - \eta$ *the following statement holds: For every distinct pair of points* $x, y \in \mathcal{M}$*,*

$$(1-\delta)\sqrt{\frac{\tilde{k}}{d}} \leq \frac{\|Ax - Ay\|_2}{\|x - y\|_2} \leq (1+\delta)\sqrt{\frac{\tilde{k}}{d}}.$$

Since we can select η is very close to 1 in order to the probability $1-\eta > 0$, there exists a orthoprojector A_i for sub-manifold U_i by applying Lemma E.3. And we use V_r to denote the uniform upper bound of $\mathrm{vol}(U_i)$, which makes the uniform dimension $\tilde{k} = O(k \log d)$ for each $i \in [N]$. Let local almost isotropic transformation $\phi_i(x) = \frac{1}{2}A_i(x - c_i) + \frac{1}{2}\mathbb{1}$, where we use $\mathbb{1}$ to denote the vector $(1, 1, ...1) \in \mathbb{R}^{\tilde{k}}$, and then we know $\phi_i(U_i) \subset [0, 1]^{\tilde{k}}$.

Step 3: Approximate Lipschitz mapping $f \circ \phi_i^{-1}$ by $\hat{f}_i$

To approximate $f \circ \phi^{-1} : [0, 1]^{\tilde{k}} \to \mathbb{R}$ via ReLU networks, we first caculate the Lipschitzness of it. For any pair x, y of $\phi_i(U_i)$, we have

$$\begin{aligned} |f \circ \phi^{-1}(x) - f \circ \phi^{-1}(y)| &\leq L\|\phi^{-1}(x) - \phi^{-1}(y)\|_\infty \\ &\leq L\|\phi^{-1}(x) - \phi^{-1}(y)\|_2 \\ &\leq \frac{2L}{1-\delta}\sqrt{\frac{d}{\tilde{k}}}\|x - y\|_2 \\ &\leq \frac{2L\sqrt{d}}{1-\delta}\|x - y\|_\infty. \end{aligned}$$

The first inequality is due to the Lipschitzness of function f. The second and last equality is the equivalence between ℓ_2 norm and ℓ_∞ norm. The third inequality uses the isotropic property of the orthoprojector A_i. So $f \circ \phi_i^{-1}$ is a $\frac{2L\sqrt{d}}{1-\delta}$ Lipschitz mapping from $[0, 1]^{\tilde{k}}$ to $\mathbb{R}$.

By using Lemma C.2, there exists a ReLU network $\hat{f}_i$ with at most

$$c_1\left(\frac{c_2\epsilon_1(1-\delta)}{2L\sqrt{d}}\right)^{-\tilde{k}}(\tilde{k}^2 + \tilde{k}\log\tilde{k} + \tilde{k}\log\frac{1}{\epsilon_1})$$

parameters such that for any $x \in \phi_i(U_i)$, we have the uniform error ϵ_1 as

$$|f \circ \phi_i^{-1}(x) - \hat{f}_i(x)| \leq \epsilon_1.$$

Notice that ϕ_i is a linear mapping so that we can use a ReLU network with only one layer to represent it, which shows that we can approximate f efficiently in the local neighborhood U_i.

Step 4: Approximate simple piecewise polynomial $\rho_i \circ T_i^{-1}$ by $\hat{\rho}_i$

According to the poly-partitionable property of manifold M and Lemma B.2, there exists a ReLU network $\hat{\rho}_i$ with at most $O(k \log(k/\epsilon_2))$ parameters such that for any $x \in T_i(U_i) \subset [0, 1]^k$, we have the uniform error ϵ_2 as

$$|\rho_i \circ T_i^{-1}(x) - \hat{\rho}_i(x)| \leq \epsilon_2,$$

where T_i is composed by the tangent vectors of c_i and is scaled and translated to ensure $T_i(U_i) \subset [0, 1]^k$.

Step 5: Determine the corresponding neighborhood for input

Notice that $\mathrm{supp}(\rho_i) \subset U_i$ but $\hat{\rho}_i$ may be non-zero for some point $[0, 1]^k / T_i(U_i)$, so we need to determine the corresponding chart for input $x \in \mathcal{M}$ by ReLU networks. Inspired by Chen et al. (2019), we construct indicate approximator $\hat{I}_\theta$ and ℓ_2 distance approximators $\{\hat{d}_i^2\}_{i\in[N]}$ based on quadratic approximator in Lemma B.1 to approximate the neighborhood's indicator $I\{x \in U_i\}$, which relies upon the following identical equations

$$I\{x \in U_i\} = I\{\|x - c_i\|_2^2 < r^2\} = I\{(\cdot) < r^2\} \circ d_i^2(x),$$

where $d_i^2(x)$ denotes the square of ℓ_2 distance between x and c_i. Then, if $\hat{I}_\theta \approx I\{(\cdot) < r^2\}$ and $\hat{d}_i^2 \approx d_i^2$, we have $\hat{I}_\theta \circ \hat{d}_i^2 \approx I\{(\cdot) < r^2\} \circ d_i^2 = I\{x \in U_i\}$, which determines the corresponding chart approximately.

Assume that the uniform error of square distance approximator is ϵ_q (i.e. $|d_i^2 - \hat{d}_i^2| \leq \epsilon_q$ for any $x \in [0, 1]^d$). In fact, functions computed by ReLU networks are piecewise linear but the indicator functions are not continuous, so we need to relax the indicator such that $\hat{I}_\theta(x) = 1$ for $x \leq r^2 + \epsilon_q - \theta$, $\hat{I}_\theta(x) = 0$ for $x \geq r^2 - \epsilon_q$ and $\hat{I}_\theta$ is linear in $(r^2 + \epsilon_q - \theta, r^2 - \epsilon_q)$.

To correct the difference between indicator and its approximator, we will bound the value of function f such that the magnitude of $f(x)$ is sufficient small when x is nearly on the boundary of U_i. Intuitively, for any $y \in \partial(U_i)$, we have

$$f\rho_i(y) = 0.$$

This is due to $\mathrm{supp}(\rho_i) \subset U_i$, which implies that we only need estimate the upper bound of $\|x-y\|_2$ for the Lipshcitzness of f and smoothness of ρ_i, where x is nearly on ∂U_i. Indeed, we can prove that for any $xU_i' :=\in U_i/B_{\sqrt{r^2-\theta}}(c_i)$, there exists $y \in \partial U_i$ such that $\|x-y\|_2 = O(\theta)$ (Chen et al., 2019, Lemma 3).

Step 6: Estimate the total error

We combine three sub-network groups as

$$\hat{f} = \sum_{i=1}^{N} (\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i) \hat{\times} (\hat{I}_\theta \circ \hat{d}_i^2).$$

Next, we estimate the total error between f and $\hat{f}$. For any $x \in M$, we use g_i to denote $(\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i)$, I_i to denote $I\{x \in U_i\}$ and $\hat{I}_i$ to denote $\hat{I}_\theta \circ \hat{d}_i^2$, then we have

$$\begin{aligned}
|f(x) - \hat{f}(x)| &= \left| \sum_{i=1}^{N} f\rho_i - \sum_{i=1}^{N} g_i \hat{\times} \hat{I}_i \right| \\
&\le \left| \sum_{i=1}^{N} f\rho_i - g_i I_i \right| + \left| \sum_{i=1}^{N} g_i \times I_i - g_i \hat{\times} \hat{I}_i \right| \\
&= \left| \sum_{i:x\in U_i} f\rho_i - (\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i) \right| + \left| \sum_{i=1}^{N} g_i \times I_i - g_i \hat{\times} \hat{I}_i \right| \\
&\le \left| \sum_{i:x\in U_i} ((f \circ \phi_i^{-1} - \hat{f}_i) \circ \phi_i) \rho_i \right| + \left| \sum_{i:x\in U_i} (\hat{f}_i \circ \phi_i) \times \rho_i - (\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i) \right| \\
&\quad + \left| \sum_{i=1}^{N} g_i \times I_i - g_i \hat{\times} \hat{I}_i \right|.
\end{aligned}$$

The second identical equation is due to $\mathrm{supp}(\rho_i) \subset U_i$. Notice that $\sum_{i\in[N]} \rho_i = 1$, then the first term satisfies that

$$\left| \sum_{i:x\in U_i} ((f \circ \phi_i^{-1} - \hat{f}_i) \circ \phi_i) \rho_i \right| \le \left(\sum_{i:x\in U_i} \rho_i \right) \max_{i:x\in U_i} \{|f \circ \phi_i^{-1} - \hat{f}_i|\} \le \epsilon_1.$$

By the approximation of $\hat{\times}$, the second term satisfies that

$$\left| \sum_{i:x\in U_i} (\hat{f}_i \circ \phi_i) \times \rho_i - (\hat{f}_i \circ \phi_i) \hat{\times} (\hat{\rho}_i \circ T_i) \right| \lesssim \left| \sum_{i:x\in U_i} (\hat{f}_i \circ \phi_i) \times ((\rho_i \circ T_i^{-1} - \hat{\rho}_i) \circ T_i) \right| \le c_f N \epsilon_2.$$

where c_f is the uniform upper bound the value of $\{\hat{f}_i\}_{i\in[N]}$. And the third term satisfies that

$$\left| \sum_{i=1}^{N} g_i \times I_i - g_i \hat{\times} \hat{I}_i \right| \lesssim \left| \sum_{i=1}^{N} g_i \times (I_i - \hat{I}_i) \right| \le \sum_{i=1}^{N} \max_{x\in U_i'} |g_i| \lesssim \sum_{i=1}^{N} \max_{x\in U_i'} |f\rho_i| = O(N\theta).$$

Finally, we choose $\epsilon_1 = O(\epsilon)$ and $\epsilon_2 = \theta = O(\epsilon/N)$ to control the total error boounded by ϵ and derive the upper bound for the size of network in Lemma E.1.

□

E.2 Proof of Theorem 5.8

Theorem E.4 (Restatement of Theorem 5.8)**.** *Let $\epsilon \in (0,1)$ be a small constant. There exists a sequence $\{N_k\}_{k\geq 1}$ that satisfies $N_k = \Omega\left((2\epsilon\sqrt{d/k})^{-\frac{k}{2}}\right)$. and a universal constant $C_1 > 0$ such that the following holds: let $\mathcal{M} \subset [0,1]^d$ be a complete and compact $k-$dimensional Riemannian manifold with non-negative Ricci curvature , then there exists two 2ϵ-separated sets $A, B \subset \mathcal{M}$ under ℓ_∞ norm, such that for any μ_0-balanced distribution P on the supporting set $S = A \cup B$ and robust radius $c \in (0,1)$, we have*

$$\inf\{\mathcal{L}_P^{\infty, c\epsilon}(f) : f \in F_{N_k}\} \geq C_1\mu_0.$$

Proof. Our proof relies on the following propositions.

Lemma E.5. *(Niyogi et al., 2008, Proposition 6.3)*

Let $\mathcal{M}$ be a sub-manifold of $\mathbb{R}^d$ with condition number $1/\tau$. Let p and q be two points in $\mathcal{M}$ such that $\|x-y\|_2 = r$. Then for all $r \leq \tau/2$, the geodesic distance $d_{\mathcal{M}}(p,q)$ is bounded by

$$d_{\mathcal{M}}(x,y) \leq \tau - \tau\sqrt{1-2r/\tau}.$$

By Lemma E.5, we know that $d_{\mathcal{M}}(x,y) \leq \tau - \tau\sqrt{1-2r/\tau} \leq 2r$ when $r \leq \tau/2$.

Lemma E.6. *(Bishop, 1964, Bishop-Gromov Volume Comparison Theorem) Let $\mathcal{M}$ is a complete Riemannian manifold with Ricci curvature $Ric \geq (k-1)l$, and $p \in \mathcal{M}$ is an arbitrary point. Then the function*

$$r \mapsto \frac{\mathrm{vol}\left(B_{\mathcal{M},r}(p)\right)}{\mathrm{vol}\left(B_r^l\right)}$$

is a non-increasing function which tends to 1 as r goes to 0 , where $B_{\mathcal{M},r}(p)$ is the $\mathcal{M}$'s geodesic ball of radius r and center p, and B_r^l is a geodesic ball of radius r in the space form $\mathcal{M}_l^k$. In particular, $\mathrm{vol}\left(B_{\mathcal{M},r}(p)\right) \leq \mathrm{vol}\left(B_r^l\right)$.

By Lemma E.6 and the non-negativeness of $\mathcal{M}$'s Ricci curvature, we know $\mathrm{vol}(B_{\mathcal{M},r}(c)) \leq \mathrm{vol}(B_r^0) = r^k V_k$, where V_k denotes the volume of the unit ball in $\mathbb{R}^k$. Recall the relation between the covering number $\mathcal{N}_{\mathcal{M}}(r)$ and the packing number $\mathcal{P}_{\mathcal{M}}(r)$ on the manifold $\mathcal{M}$, then we have

$$\mathcal{P}_{\mathcal{M}}(r) \geq \mathcal{N}_{\mathcal{M}}(2r) \geq \frac{\mathrm{vol}(\mathcal{M})}{(2r)^k V_k} = \Omega\left(\frac{\mathrm{vol}(\mathcal{M})k^{\frac{k}{2}}}{r^k}\right).$$

By choosing $r = 2\epsilon\sqrt{d}$, we know that there are at least $\Omega\left((2\epsilon\sqrt{d/k})^{-k}\right)$ points on $\mathcal{M}$ such that the ℓ_∞ distance between each pair points of these is more than 2ϵ, where we use $\mathcal{Q}$ to denote the set of these selected points. The remain of proof is similar to the latter half of proof for Theorem D.1.

Let $S = \mathcal{Q}$ be the supporting set. Assume that for any partition A, B of S such that $A \cup B = S$ and $A \cap B = \emptyset$, there exists a classifier $f \in F_{N_k}$ that robustly classifies A and B with at least $1-\alpha$ accuracy. Next, we estimate the lower and upper bounds for the cardinal number of the vector set

$$R := \{(f(x))_{x\in\mathcal{Q}} | f \in F_{N_k}\}.$$

Let n denote $|\mathcal{Q}|$, then we have

$$R = \{(f(x_1), f(x_2), ...f(x_n)) | f \in F_{N_k}\},$$

where $\mathcal{Q} = \{x_1, x_2, ..., x_n\}$.

On one hand, we know that for any $u \in \{-1,1\}^n$, there exists a $v \in R$ such that $d_H(u,v) \leq \alpha n$, where $d_H(\cdot,\cdot)$ denotes the Hamming distance, then we have

$$|R| \geq \mathcal{N}(\{-1,1\}^n, d_H, \alpha n) \geq \frac{2^n}{\sum_{i=0}^{\alpha n}\binom{n}{i}}.$$

On the other hand, by applying Lemma A.4, we have

$$\frac{2^n}{\sum_{i=1}^{\alpha n}\binom{n}{i}} \leq |R| \leq \Pi_{F_{N_k}}(n) \leq \sum_{j=0}^{l}\binom{n}{j}.$$

where l is the VC-dimension of F_{N_k}. In fact, we can derive $l = \Omega(n)$ when α is a small constant. Assume that $l < n-1$, then we have $\sum_{j=0}^{l}\binom{n}{j} \leq (en/l)^l$ and $\sum_{i=1}^{\alpha n}\binom{n}{i} \leq (e/\alpha)^{\alpha n}$, so

$$\frac{2^n}{(e/\alpha)^{\alpha n}} \leq |R| \leq (en/l)^l.$$

We define a function $h(x)$ as $h(x) = (e/x)^x$, then we derive

$$2 \leq \left(\frac{e}{\alpha}\right)^{\alpha}\left(\frac{e}{l/n}\right)^{l/n} = h(\alpha)h(l/n).$$

When α is sufficient small, $l/n \geq C(\alpha)$ that is a constant only depending on α, which implies $l = \Omega(n)$. Finally, by using Lemma A.3 and $n = |\mathcal{Q}| = \Omega\left((2\epsilon\sqrt{d/k})^{-k}\right)$, we know $N_k = \Omega\left((2\epsilon\sqrt{d/k})^{-\frac{k}{2}}\right)$. Combined with the definition of balanced distribution, we conclude the proof of Theorem E.4. □