

A. Appendix to Paper: Mathematical Theory of Adversarial Deep Learning

This appendix contains proofs for theorems in Section 3.

A.1. Proof of Theorem 3.4

We will show that $\text{RobM}(\mathcal{D}, \alpha)$ is computationally equivalent to the following NPC problem.

Definition A.1 (Reversible 6-SAT). Let φ be a Boolean formula and let $\bar{\varphi}$ denote the formula obtained from φ by negating each variable. The Boolean formula φ is called *reversible* if either both φ and $\bar{\varphi}$ are satisfiable or both are not satisfiable. The *reversible satisfiability problem* is to recognize the satisfiability of reversible formulae in conjunctive normal form (CNF). By the *reversible 6-SAT*, we mean the reversible satisfiability problem for CNF formulae with six variables per clause. In (Megiddo, 1988), it was shown that the reversible 6-SAT is NPC.

We restate Theorem 3.4 here for convenience.

Theorem A.2. $\text{RobM}(\mathcal{D}, \alpha)$ is NP-hard; that is, for $\alpha \in \mathbb{R}_+$ and a dataset $\mathcal{D} \subset \mathbb{R}^n \times \{-1, 1\}$, it is NP-hard to decide whether there exists a robust network in $\mathcal{H}_{n,2}$ for \mathcal{D} with budget α .

Proof. Let $\varphi(k, m) = \bigwedge_{i=1}^m \varphi_i(k, m)$ be a 6-SAT for k variables, where $\varphi_i(k, m) = \bigvee_{j=1}^6 \tilde{x}_{i,j}$ and $\tilde{x}_{i,j}$ is either x_s or $\neg x_s$ for $s \in [k]$ (refer to Definition A.1).

For $i \in [k]$, define $Q_i^\varphi \in \mathbb{R}^k$ as follows: $Q_i^\varphi[j] = 1$ if x_j occurs in $\varphi_i(k, m)$; $Q_i^\varphi[j] = -1$ if $\neg x_j$ occurs in $\varphi_i(k, m)$; $Q_i^\varphi[j] = 0$ otherwise. Then six entries of Q_i^φ are 1 or -1 and all other entries are zero. Also, let $\mathbf{1}_i \in \mathbb{R}^k$, whose i -th element is 1 and all other entries are 0.

We define a binary classification dataset $\mathcal{D}(\varphi) = \{(x_i, y_i)\}_{i=0}^{m+4k} \subset \mathbb{R}^k \times \{-1, 1\}$ as follows

- (1) $x_0 = 0, y_0 = -1$.
- (2) For $i \in [k]$, $x_i = k\mathbf{1}_i, y_i = -1$.
- (3) For $i \in \{k+1, k+2, \dots, 2k\}$, $x_i = -k\mathbf{1}_{i-k}, y_i = -1$.
- (4) For $i \in \{2k+1, 2k+2, \dots, 3k\}$, $x_i = 2.001k\mathbf{1}_{i-2k}, y_i = 1$.
- (5) For $i \in \{3k+1, 3k+2, \dots, 4k\}$, $x_i = -2.001k\mathbf{1}_{i-3k}, y_i = 1$.
- (6) For $i \in \{4k+1, 3k+2, \dots, 4k+m\}$, $x_i = k/4.1 \cdot Q_{i-4k}^\varphi, y_i = -1$.

The size of $\mathcal{D}(\varphi)$ is $O((m+k) \log k)$ and $\mathcal{D}(\varphi)$ has separation bound $k/4.1 > 1$, because $k \geq 6$ for 6-SAT problem.

We claim that $\text{RobM}(\mathcal{D}(\varphi), 0.5)$ has a solution \mathcal{F} if and only if the reversible 6-SAT $\varphi(k, m)$ has a solution $J = \{x_j = v_j\}_{j=1}^k$, and \mathcal{F} and J can be deduced from each other in polynomial time; that is, $\text{RobM}(\mathcal{D}(\varphi), 0.5)$ is computationally equivalent to $\varphi(k, m)$. Since reversible 6-SAT is NPC (Megiddo, 1988), by the claim, $\text{RobM}(\mathcal{D}(\varphi), 0.5)$ is NPC, which implies that $\text{RobM}(\mathcal{D}(\varphi), \alpha)$ is NP-hard. This proves the theorem.

Before proving the claim, we first introduce a notation. Let $J = \{x_j = v_j\}_{j=1}^k$ be a solution to the reversible 6-SAT problem φ and $\varphi_i(k, m) = \bigvee_{j=1}^6 \tilde{x}_{i,j}$ a clause of φ , where $v_i \in \{-1, 1\}$. Then denote $q(J, \varphi_i)$ to be the number of $\tilde{x}_{i,j}$ which has value 1 on the solution J . If $q(J, \varphi_i) = 0$, then φ_i is not true. If $q(J, \varphi_i) = 6$, then $\neg \varphi_i$ is not true. Since J is a solution to the reversible 6-SAT problem φ , we have $1 \leq q(J, \varphi_i) \leq 5$. It is easy to see that $q(J, \varphi_i) = |\{j \in [k] : Q_i^\varphi[j] = v_j\}|$.

The claim will be proved in two steps.

Step 1. We prove that if $\varphi(k, m)$ has a solution $J = \{x_j = v_j\}_{j=1}^k$, then $\text{RobM}(\mathcal{D}(\varphi), 0.5)$ has a solution \mathcal{F} , where $v_i \in \{-1, 1\}$. Let $U_1 = \frac{2}{3k}(v_1, v_2, \dots, v_k)$, $U_2 = -\frac{2}{3k}(v_1, v_2, \dots, v_k)$. Define $\mathcal{F} \in \mathbf{H}_{k,2}$ to be $\mathcal{F}(x) = \psi(\sigma(U_1 x - 1) + \sigma(U_2 x - 1))$, ψ means the Sgn. It is clear that \mathcal{F} can be obtained from J in $\text{Poly}(k)$. We will show that $\mathcal{F}(x)$ is a robust memorization of $\mathcal{D}(\varphi)$ with budget 0.5. The proof will be given in five steps: (c1) - (c5).

(c1) Since $\|U_1\|_1 = 2/3$, we have that $U_1 x - 1 > 1/3$ implies $U_1(x + \epsilon) - 1 > 0$ for any $\epsilon \in \mathbb{R}^k$ satisfying $\|\epsilon\|_\infty \leq 0.5$, and $U_1 x - 1 \leq -1/3$ implies $U_1(x + \epsilon) - 1 \leq 0$ for any $\|\epsilon\|_\infty \leq 0.5$. U_2 has similar properties.

(c2) Since $U_1 x_0 - 1 = -1 < -1/3$ and $U_2 x_0 - 1 = -1 < -1/3$, from (c1), for any $\|\epsilon\|_\infty \leq 0.5$, we have $\mathcal{F}(x_0 + \epsilon) =$

$\psi(\sigma(U_1(x_0 + \epsilon) - 1) + \sigma(U_2(x_0 + \epsilon) - 1)) = \psi(0 + 0) = 0 = y_0$. Thus \mathcal{F} is robust at x_0 with budget 0.5.

(c3) For $i \in [2k]$, we have that $U_1 x_i - 1 \leq -1 + |U_1||x_i| = -1 + k \cdot 2/(3k) = -1/3$ and $U_2 x_i - 1 \leq -1 + |U_2||x_i| = -1 + k \cdot 2/(3k) = -1/3$. By (c1), for any $\|\epsilon\|_\infty \leq 0.5$, we have

$$\mathcal{F}(x_i + \epsilon) = \psi(\sigma(U_1(x_i + \epsilon) - 1) + \sigma(U_2(x_i + \epsilon) - 1)) = \psi(0 + 0) = 0 = y_i.$$

Thus \mathcal{F} is robust at x_i with budget 0.5.

(c4) For $i \in \{2k + 1, 2k + 2, \dots, 4k\}$, since $U_1 = -U_2$, at least one of the following two inequalities $U_1 x_i - 1 = -1 + |U_1||x_i| = -1 + 2.001k \cdot 2/(3k) > 1/3$ and $U_2 x_i - 1 = -1 + |U_2||x_i| = -1 + 2.001k \cdot 2/(3k) > 1/3$ is true, say the first one is true. By (c1), for any $\|\epsilon\|_\infty \leq 0.5$, we have

$$\mathcal{F}(x_i + \epsilon) = \psi(\sigma(U_1(x_i + \epsilon) - 1) + \sigma(U_2(x_i + \epsilon) - 1)) \geq \psi(\sigma(U_1(x_i + \epsilon) - 1)) = 1 = y_i.$$

Thus \mathcal{F} is robust at x_i with budget 0.5.

(c5) Let $i \in \{4k + 1, 4k + 2, \dots, 4k + m\}$. It is clear that $q(J, \varphi_{i-4k}) + q(J, \bar{\varphi}_{i-4k}) = 6$. Then

$$\begin{aligned} & U_1 Q_{i-4k}^\varphi \\ &= \sum_{j: x_j \in \varphi_{i-4k}} v_j Q_{i-4k}^\varphi[j] \\ &= \sum_{j: x_j \in \varphi_{i-4k}, \text{Sgn}(Q_{i-4k}^\varphi[j]) = \text{Sgn}(v_j)} \frac{2}{3k} - \\ & \quad \sum_{j: x_j \in \varphi_{i-4k}, \text{Sgn}(Q_{i-4k}^\varphi[j]) \neq \text{Sgn}(v_j)} \frac{2}{3k} \\ &= q(J, \varphi_{i-4k}) \frac{2}{3k} - q(J, \bar{\varphi}_{i-4k}) \frac{2}{3k} \\ &\in \{0, 2 \cdot (2/(3k)), 4 \cdot (2/(3k)), -2 \cdot (2/(3k)), -4 \cdot (2/(3k))\}, \end{aligned}$$

which means $|U_1 Q_{i-4k}^\varphi| \leq 8/(3k)$. Similarly, we also have $|U_2 Q_{i-4k}^\varphi| \leq 8/(3k)$. As a consequence, $U_1 x_i - 1 = -1 + U_1 Q_{i-4k}^\varphi \cdot k/4.1 \leq -1 + 8/(3k) \cdot k/4.1 \leq -1 + 2/3 = -1/3$, and similarly $U_2 x_i - 1 \leq -1/3$. By (c1), for any $\|\epsilon\|_\infty \leq 0.5$, we have that

$$\mathcal{F}(x_i + \epsilon) = \psi(\sigma(U_1(x_i + \epsilon) - 1) + \sigma(U_2(x_i + \epsilon) - 1)) = \psi(0 + 0) = y_i.$$

Thus \mathcal{F} is robust at x_i with budget 0.5.

From (c2) to (c5), \mathcal{F} is a robustness memorization of $\mathcal{D}(\varphi)$ with budget 0.5, and Step 1 is proved.

Step 2. We prove that if $\text{RobM}(\mathcal{D}(\varphi), 0.5)$ has a solution $\mathcal{F}(x) = \psi(\sigma(U_1 x + b_1) + \sigma(U_2 x + b_2)) \in \mathbf{H}_{k,2}$ which is a robust memorization of $\mathcal{D}(\varphi)$ with budget 0.5, then $\varphi(k, m)$ has a solution.

Without loss of generality, we can assume that $U_1 \neq 0$ and $U_2 \neq 0$. **We will show that $J = \{x_i = \text{Sgn}(U_1^{(i)})\}_{i=1}^k$ is the solution to the reversible 6-SAT problem $\varphi(k, m)$.** The proof is divided into six steps: (d1) - (d7).

(d1) We will show that it can be assumed $b_1 = b_2 = -1$. Since \mathcal{F} is robust at $x_0 = 0$ with budget 0.5 and $y_0 = 0$, we have that $0 = \mathcal{F}(x_0 + 0.1 \text{Sgn}(U_1)) = \mathcal{F}(0.1 \text{Sgn}(U_1)) \geq \psi(\sigma(0.1 U_1 \text{Sgn}(U_1) + b_1)) = \psi(\sigma(b_1 + 0.1 |U_1|_1))$, which implies $b_1 < 0$. Similarly, $b_2 < 0$. Then, we can assume $b_1 = b_2 = -1$, because

$$\begin{aligned} & \mathcal{F}(x) \\ &= \psi(\sigma(U_1 x + b_1) + \sigma(U_2 x + b_2)) \\ &= \psi(|b_1| \sigma(\frac{U_1}{|b_1|} x - 1) + |b_2| \sigma(\frac{U_2}{|b_2|} x - 1)) \\ &= \psi(\sigma(\frac{U_1}{|b_1|} x - 1) + \sigma(\frac{U_2}{|b_2|} x - 1)). \end{aligned}$$

(d2) We prove $\|U_1\|_1 \leq 2$ and $\|U_2\|_1 \leq 2$. Since \mathcal{F} is robust at $x_0 = 0$ with budget 0.5, we have that

$$\begin{aligned} 0 &= \mathcal{F}(0.5 \text{Sgn}(U_1)) \\ &= \psi(\sigma(0.5 U_1 \text{Sgn}(U_1) - 1) + \sigma(0.5 U_2 \text{Sgn}(U_1) - 1)) \\ &\geq \psi(\sigma(0.5 \|U_1\|_1 - 1)), \end{aligned}$$

which implies $\|U_1\|_1 \leq 2$. Similarly, we have $\|U_2\|_1 \leq 2$.

(d3) We prove that $U_1^{(j)}U_2^{(j)} < 0$ for all $j \in [k]$, where $U_i^{(j)}$ is the j -th component of U_i .

(d3.1) For $i \in \{2k+1, 2k+2, \dots, 4k\}$, since $\mathcal{F}(x_i) = \psi(\sigma(U_1x_i - 1) + \sigma(U_2x_i - 1)) = 1 > 0$, at least one of $U_1x_i - 1 > 0$ or $U_2x_i - 1 > 0$ is valid.

Assume $U_1^{(j)} \geq 0$ and $U_2^{(j)} \geq 0$ for some $j \in [k]$. Then we have $U_1x_{3k+j} - 1 = -1 - 2.001kU_1^{(j)} < 0$ and $U_2x_{3k+j} - 1 = -1 - 2.001kU_2^{(j)} < 0$, which contradict to (d3.1).

Assume $U_1^{(j)} \leq 0$ and $U_2^{(j)} \leq 0$ for some $j \in [k]$. Then we have $U_1x_{2k+j} - 1 = -1 + 2.001kU_1^{(j)} < 0$ and $U_2x_{2k+j} - 1 = -1 + 2.001kU_2^{(j)} < 0$, which contradicts to (d3.1). Then (d3) is proved.

(d4) For any $j \in [k]$, we have $k|U_1^{(j)}| \leq 1 - 0.5\|U_1\|_1$ and $k|U_2^{(j)}| \leq 1 - 0.5\|U_2\|_1$.

For $i \in [2k]$, because $\mathcal{F}(x_i) = \psi(\sigma(U_1(x_i + \epsilon) - 1) + \sigma(U_2(x_i + \epsilon) - 1)) = 0$ is stand for any $\|\epsilon\|_\infty \leq 0.5$, so we have $U_1(x_i + \epsilon) - 1 \leq 0$ and $U_2(x_i + \epsilon) - 1 \leq 0$; that is $U_1x_i - 1 \leq -0.5\|U_1\|_1$ and $U_2x_i - 1 \leq -0.5\|U_2\|_1$.

For any $j \in [k]$, we have $U_1x_j \leq 1 - 0.5\|U_1\|_1$ and $U_1x_{k+j} \leq 1 - 0.5\|U_1\|_1$, considering that the j -th component of x_j and x_{k+j} has different positive and negative, so one of them has the same positive and negative with the j -th component of $U_1^{(j)}$, so $\max\{U_1x_j, U_1x_{k+j}\} = k|U_1^{(j)}| \leq 1 - 0.5\|U_1\|_1$. Similar, we have $k|U_2^{(j)}| \leq 1 - 0.5\|U_2\|_1$.

(d5) For any $j \in [k]$, we have $2.001|kU_1^{(j)}| > 1 + 0.5\|U_1\|_1$ and $2.001|kU_2^{(j)}| > 1 + 0.5\|U_2\|_1$.

For $j \in [k]$, by (d3), we first assume that $U_1^{(j)} > 0$ and $U_2^{(j)} < 0$. We will prove the following conclusions about this j .

(d5.0) For $i \in \{2k+1, 2k+2, \dots, 4k\}$, since $\mathcal{F}(x_i) = \psi(\sigma(U_1(x_i + \epsilon) - 1) + \sigma(U_2(x_i + \epsilon) - 1)) = 1 > 0$ is valid for $\|\epsilon\|_\infty \leq 0.5$, at least one of $U_1(x_i + \epsilon) - 1 > 0$ or $U_2(x_i + \epsilon) - 1 > 0$ is valid.

(d5.1) $U_2^{(j)} < -\frac{1}{2.001k}$. By (d5.0), at least one of $U_1x_{3k+j} - 1 > 0$ and $U_2x_{3k+j} - 1 > 0$ is valid. Since $U_1x_{3k+j} - 1 = -1 - 2.001kU_1^{(j)} < -1 < 0$ (by $U_1^j > 0$), we have $U_2x_{3k+j} - 1 = -1 - 2.001kU_2^{(j)} > 0$; that is, $U_2^{(j)} < -\frac{1}{2.001k}$.

(d5.2) $U_2(x_{2k+j} + \epsilon) - 1 < 0$ for any $\|\epsilon\|_\infty \leq 0.5$. By (d2) and (d5.1), we have

$$\begin{aligned} U_2(x_{2k+j} + \epsilon) - 1 &\leq U_2x_{2k+j} + 0.5\|U_2\|_1 - 1 \\ &\stackrel{(d2)}{\leq} U_2x_{2k+j} + 0.5 \cdot 2 - 1 \\ &\stackrel{(d5.1)}{\leq} -\frac{1}{2.001k} \cdot 2.001k + 0.5 \cdot 2 - 1 \\ &< 0. \end{aligned}$$

(d5.3) $2.001kU_1^{(j)} > 1 + 0.5\|U_1\|_1$. By (d5.0), for any $\|\epsilon\|_\infty \leq 0.5$, at least one of $U_1(x_{2k+j} + \epsilon) - 1 > 0$ or $U_2(x_{2k+j} + \epsilon) - 1 > 0$ is valid. By (d5.2), we have $U_2(x_{2k+j} + \epsilon) - 1 < 0$ for any $\|\epsilon\|_\infty \leq 0.5$. So $U_1(x_{2k+j} + \epsilon) - 1 > 0$ is valid for any $\|\epsilon\|_\infty \leq 0.5$. Then we have $U_1x_{2k+j} - 1 > 0.5\|U_1\|_1$; that is $2.001kU_1^{(j)} > 1 + 0.5\|U_1\|_1$.

(d5.4) $-2.001kU_2^{(j)} > 1 + 0.5\|U_2\|_1$. This can be proved similar to (d5.3).

If we replace $U_1^{(j)} > 0$ and $U_2^{(j)} < 0$ with $U_1^{(j)} < 0$ and $U_2^{(j)} > 0$, then we can obtain $-2.001kU_1^{(j)} > 1 + 0.5\|U_1\|_1$ and $2.001kU_2^{(j)} > 1 + 0.5\|U_2\|_1$. So, for any $j \in [k]$, we finally obtain $2.001|kU_1^{(j)}| > 1 + 0.5\|U_1\|_1$ and $2.001|kU_2^{(j)}| > 1 + 0.5\|U_2\|_1$.

(d6) For any $j \in [k]$, we have $\frac{2/3-0.001}{k} < |U_1^{(j)}| < \frac{2/3+0.001}{k}$ and $\frac{2/3-0.001}{k} < |U_2^{(j)}| < \frac{2/3+0.001}{k}$.

For any $j \in [k]$, by (d4) and (d5), we know that $\frac{1-0.5\|U_1\|_1}{k} \geq |U_1^{(j)}| > \frac{1+0.5\|U_1\|_1}{2.001k}$.

(d6.1) $|U_1^{(j)}| > \frac{2/3-0.001}{k}$. Let $e = \operatorname{argmin}_{j \in [k]} \{|U_1^{(j)}|\}$. Then by (d5), we have that $|U_1^{(e)}| > \frac{1+0.5\|U_1\|_1}{2.001k} \geq \frac{1+0.5k|U_1^{(e)}|}{2.001k}$; that is $|U_1^{(e)}| > \frac{2/3-0.001}{k}$. So for any $j \in [k]$, we have that $|U_1^{(j)}| \geq |U_1^{(e)}| > \frac{2/3-0.001}{k}$.

(d6.2) $|U_1^{(j)}| < \frac{2/3+0.001}{k}$. By (d6.1), we have $\|U_1\|_1 > k|U_1^{(e)}| > 2/3 - 0.001$. Then we have $|U_1^{(j)}| \leq \frac{1-0.5\|U_1\|_1}{k} <$

$$\frac{1-1/3+0.0005}{k} < \frac{2/3+0.001}{k}.$$

(d6.1) and (d6.2) prove the inequality for U_1 . The inequality for U_2 can be proved similarly.

(d7) We show that $\{x_i = \text{Sgn}(U_1^{(i)})\}_{i=1}^k$ is the solution of the reversible 6-SAT problem φ .

Assume that $J = \{x_i = \text{Sgn}(U_1^{(i)})\}_{i=1}^k$ is not the solution of φ . Then there exists a clause, say φ_1 , such that $\varphi_1 = 0$ at J , or $\bar{\varphi}_1 = 0$ at J ; that is, $q(J, \varphi_1) = 0$ or $q(J', \varphi_1) = 0$, where $J' = \{x_i = \text{Sgn}(U_2^{(i)})\}_{i=1}^k$. Then, we have

$$\begin{aligned} & \mathcal{F}(x_{4k+1} + 0.5\text{Sgn}(U_1)) \\ &= \psi(\sigma(U_1(x_{4k+1} + 0.5\text{Sgn}(U_1)) - 1) + \sigma(U_2(x_{4k+1} + 0.5\text{Sgn}(U_1)) - 1)) \\ &\geq \psi(\sigma(U_1(x_{4k+1} + 0.5\text{Sgn}(U_1)) - 1)) \\ &\stackrel{\text{(d6)}}{\geq} \psi(\sigma(U_1 x_{4k+1} + 0.5k \cdot \frac{2/3-0.001}{k} - 1)) \\ &\stackrel{\text{(d6)}}{\geq} \psi(\sigma(6 \cdot \frac{2/3-0.001}{k} \cdot \frac{k}{4.1} + 0.5k \cdot \frac{2/3-0.001}{k} - 1)) \\ &\geq \psi(\sigma(0.97 + 0.1 - 1)) \\ &= 1 \neq y_{4k+1}. \end{aligned}$$

Then the robustness budget of \mathcal{F} is smaller than 0.5 at x_{4k+1} , which contradicts to the fact that \mathcal{F} is a solution to $\text{RobM}(\mathcal{D}(\varphi), 0.5)$. Therefore, $J = \{x_i = \text{Sgn}(U_1^{(i)})\}_{i=1}^k$ is a solution to φ . This proves Step 2. \square

A.2. Proof of Theorem 3.6

We will show that $\text{RobM}(\mathcal{D}, \alpha)$ is computationally equivalent to the following NPC problem (Garey & Johnson, 1979).

Definition A.3 (Minimum vertex cover). A *vertex cover* in a graph G is a set of vertices of G such that each edge of G has at least one of its end point in this set. A *minimum vertex cover* in a graph G is a vertex cover that has the smallest number of vertices among all possible vertex covers of G .

We restate Theorem 3.6 for convenience.

Theorem A.4. *MinRob(\mathcal{D}, α) is NP-hard; that is, finding the smallest k such that there exists a robust network with budget α in $\mathcal{H}_{n,k}$ is NP-hard.*

Proof. Let $G = (V, E)$ be a graph with vertex set $V = \{v_1, \dots, v_p\}$ and edge set $E = \{e_1, \dots, e_q\}$. Denote $\mathbf{1}_{a,b}(m)$ to be the vector in \mathbb{R}^p , whose a -th and b -th entries are m , and all other entries are zero. Define a binary classification dataset $\mathcal{D}(G) = \{(x_i, y_i)\}_{i=0}^q \subset \mathbb{R}^p \times \{0, 1\}$ as follows:

- (1) $x_0 = 0$ and $y_0 = 0$;
- (2) For $i \in [q]$, if e_i has vertices v_a and v_b , then set $x_i = \mathbf{1}_{a,b}(1.001)$ and $y_i = 1$.

The separation bound of $\mathcal{D}(G)$ is 1.001.

We claim that the minimum vertex cover problem for G is computationally equivalent to problem $\text{MinRob}(\mathcal{D}(G), 0.5)$. Since the minimum vertex cover problem is NPC (Garey & Johnson, 1979), problem $\text{MinRob}(\mathcal{D}(G), 0.5)$ is NPC, which implies that $\text{MinRob}(\mathcal{D}, \alpha)$ is NP-hard. This proves the theorem.

We prove the claim in three steps.

Step 1. Let $V_1 \subseteq V$ be a minimum vertex cover of G and $k = |V_1|$. Then we can compute an $\mathcal{F} \in \mathbf{H}_{k,p}$ in time $\text{Poly}(p)$ such that \mathcal{F} is a robust memorization of $\mathcal{D}(G)$ with budget 0.5.

Let $\mathcal{F}(x) = \psi(\sum_{v_j \in V_1} \sigma(2\mathbf{1}_j x - 1)) \in \mathbf{H}_{k,p}$, where $\mathbf{1}_j \in \mathbb{R}^{1 \times p}$ is the vector whose j -th weight is 1 and others weights are 0. It is easy to see that \mathcal{F} can be computed in $\text{Poly}(p)$, since $k \leq p$. We will verify that this network is a robustness memorization of $\mathcal{D}(G)$ with budget 0.5, which comes from the facts (c1) and (c2) to be proved below.

(c1) For any $\chi \in \mathbb{R}^p$ satisfying $\|\chi\|_\infty \leq 0.5$, we have

$$\begin{aligned}
 & \mathcal{F}(x_0 + \chi) \\
 &= \psi(\sum_{v_j \in V_1} \sigma(2\mathbf{1}_j(x_0 + \chi) - 1)) \\
 &= \psi(\sum_{v_j \in V_1} \sigma(2\mathbf{1}_j\chi - 1)) \\
 &\stackrel{(s1)}{\leq} \psi(\sum_{v_j \in V_1} \sigma(2\|\mathbf{1}_j\|_1\|\chi\|_\infty - 1)) \\
 &\leq \psi(\sum_{v_j \in V_1} \sigma(1 - 1)) = 0.
 \end{aligned}$$

In step (c1), we use the inequality $ab \leq \|a\|_1\|b\|_\infty$ for $a \in \mathbb{R}^{1 \times p}$ and $b \in \mathbb{R}^{p \times 1}$. So, \mathcal{F} is robust at x_0 with budget 0.5.

(c2) For $i \in [q]$, $x_i = \mathbf{1}_{q,p}(1.001)$. Since V_1 is a vertex cover, we have $v_a \in V_1$ or $v_b \in V_1$, say $v_a \in V_1$. Then for any $\chi \in \mathbb{R}^p$ such that $\|\chi\|_\infty \leq 0.5$, we have

$$\begin{aligned}
 & \mathcal{F}(x_i + \chi) \\
 &= \psi(\sum_{v_j \in V_1} \sigma(2\mathbf{1}_j(x_i + \chi) - 1)) \\
 &\geq \psi(\sigma(2\mathbf{1}_a(x_i + \chi) - 1)) \\
 &\geq \psi(\sigma(2\mathbf{1}_a x_i - 1 - 2\|\mathbf{1}_a\|_1\|\chi\|_\infty)) \\
 &\geq \psi(\sigma(1.001 \cdot 2 - 1 - 1)) \\
 &> 0.
 \end{aligned}$$

So $\mathcal{F}(x_i + \chi) = 1 = y_i$ for any $i \in [q]$ and \mathcal{F} is robust at x_i with budget 0.5.

Step 2. Let $\mathcal{F} \in \mathbf{H}_{k,p}$ be a solution to problem $\text{MinRob}(\mathcal{D}(G), 0.5)$. Then we can compute a vertex cover $V_1 \subseteq V$ of G in time $\text{Poly}(p)$ such that $|V_1| \leq k$.

Let $\mathcal{F}(x) = \psi(\sum_{i=1}^k \sigma(U_i x + b_i))$ be such a network. Without loss of generality, we assume that $U_t \neq 0$ for any $t \in [k]$. Then we have three results (d1) to (d3).

(d1) For any $t \in [k]$, $b_t < 0$ and $\|U_t\|_1 \leq 2|b_t|$. This is because

$$\begin{aligned}
 0 &= y_0 \\
 &= \mathcal{F}(x_0 + 0.5\text{Sgn}(U_t)) \\
 &= \mathcal{F}(0.5\text{Sgn}(U_t)) \\
 &= \psi(\sum_{i=1}^k \sigma(0.5U_i(\text{Sgn}(U_t) + b_i)) \\
 &\geq \psi(\sigma(0.5U_t(\text{Sgn}(U_t) + b_t)) \\
 &= \psi(0.5\|U_t\|_1 + b_t)
 \end{aligned}$$

which implies $0.5\|U_t\|_1 + b_t \leq 0$ and we can deduce $b_t < 0$ and $\|U_t\|_1 \leq 2|b_t|$ from $0.5\|U_t\|_1 + b_t \leq 0$.

(d2) For any $t \in [q]$, let $x_t = \mathbf{1}_{a,b}(1.001)$. Then there exists at least one $s \in [k]$ such that $\frac{U_s^{(a)} + U_s^{(b)}}{|b_s|} > 2 - \frac{0.002}{0.501}$. We have $\|x_t - \mathbf{1}_{a,b}(0.501)\|_\infty = \|\mathbf{1}_{a,b}(1.001) - \mathbf{1}_{a,b}(0.501)\|_\infty = 0.5$. Since $\mathcal{F}(\hat{x}_t) = 1$ for any \hat{x}_t satisfying $\|\hat{x}_t - x_t\|_\infty \leq 0.5$, we have

$$\begin{aligned}
 1 &= \mathcal{F}(\mathbf{1}_{a,b}(0.501)) = \psi(\sum_{i=1}^k \sigma(U_i \mathbf{1}_{a,b}(0.501) + b_i)) \\
 &= \psi(\sum_{i=1}^k |b_i| \sigma(\frac{U_i \mathbf{1}_{a,b}(0.501)}{|b_i|} - 1))
 \end{aligned}$$

Thus at least one $s \in [k]$ satisfies $\frac{U_s \mathbf{1}_{a,b}(0.501)}{|b_s|} - 1 > 0$; that is, $\frac{U_s^{(a)} + U_s^{(b)}}{|b_s|} > 2 - \frac{0.002}{0.501}$.

(d3) Let $V_1 = \{v_j : \exists i \in [k], s.t. j = \widehat{U}_i\}$, where $j = \widehat{U}_i \in [p]$ means that $U_i^{(j)}$ is the largest component of U_i . It is clear that V_1 can be computed from \mathcal{F} in $\text{Poly}(p)$. **We will show that V_1 is a vertex cover of G .**

For any $m \in [q]$, let $e_m = (v_a, v_b)$. By (d2), there exists an $s \in [k]$ such that $\frac{U_s^{(a)} + U_s^{(b)}}{|b_s|} > 2 - \frac{0.002}{0.501}$. By (d1), we also have $\|U_s\|_1 \leq 2|b_s|$. So $\forall g \in [p]$ with $g \neq a, b$, we have

$$\begin{aligned} & |U_s^{(g)}| \\ & \leq \|U_s\|_1 - |U_s^{(a)}| - |U_s^{(b)}| \\ & \leq 2|b_s| - (U_s^{(a)} + U_s^{(b)}) \\ & \leq \frac{0.002|b_s|}{0.501} \\ & < (1 - \frac{0.001}{0.501})|b_s| \\ & < (U_s^{(a)} + U_s^{(b)})/2 \\ & \leq \max\{U_s^{(a)}, U_s^{(b)}\} \end{aligned}$$

which implies that $\hat{U}_s \in \{a, b\}$, so whether $v_a \in V_1$ or $v_b \in V_1$ and e_m is covered by V_1 . Thus, V_1 is a vertex cover of G .

Step 3. The two problems are computationally equivalence.

Let V_1 be a minimum vertex cover of the graph G . Then from Step 1, we can find a network \mathcal{F}_1 with width $|V_1|$, which is a robust memorization of $\mathcal{D}(G)$ with budget 0.5. On the other hand, by Step 2, if \mathcal{F}_1 is not a minimum-width robust memorization of $\mathcal{D}(G)$ with budget 0.5, then there exists a smaller vertex cover for G . So $|V_1|$ and \mathcal{F}_1 are a solution to $\text{MinRob}(\mathcal{D}, 0.5)$. Similarly, we can show that a solution of $\text{MinRob}(\mathcal{D}(G), 0.5)$ leads to a minimum vertex cover for G . \square

A.3. Proof of Proposition 3.7

We restate Proposition 3.7 for convenience.

Proposition A.5. *If $\mathbf{H} = \{\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}, \text{width}(\mathcal{F}) = w\}$ is an optimal robust memorization of any dataset $\mathcal{D} \in \mathcal{D}_{n,N,L}$ with $N > n$, then $\text{width}(\mathcal{F}) = w \geq n$.*

Proof. It suffices to show that there exists a dataset \mathcal{D} , if \mathcal{F} has width less than n and memorizes \mathcal{D} , then $\text{RA}_{\mathcal{D}}(\mathcal{F}, 0.4\lambda_{\mathcal{D}}) \leq 1 - \frac{1}{n+1}$; that is, \mathcal{F} is not a robust memorization of \mathcal{D} with budget $0.4\lambda_{\mathcal{D}}$.

Denote $\mathbf{1}$ to be the vector all of whose weights are 1 and $\mathbf{1}_k$ the vector whose k -th weight is 1 and all other weights are 0. Without loss of generality, let N satisfy $(n+1)|N$. We define a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ with separation bound 1 as follows:

(1) $x_1 = 0$ and $y_1 = 0$; $x_i = \mathbf{1}_{i-1}$ and $y_i = 1$ for $i = 2, 3, \dots, n+1$;

(2) for $i = k(n+1) + 1, \dots, k(n+1) + n + 1$ and $k = 1, \dots, \frac{N}{n+1} - 1$, $x_i = x_{\bar{i}} + \mathbf{1}$ and $y_i = y_{\bar{i}}$, where $\bar{i} = i \bmod (n+1)$ if $(n+1) \nmid i$ and $\bar{i} = n+1$ otherwise.

It is easy to see that $\lambda_{\mathcal{D}} = 1$.

Let $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ be a network which memorizes \mathcal{D} defined above. Let W_1 be the weight matrix of the first layer of \mathcal{F} . Then $W_1 \in \mathbb{R}^{K \times n}$. We will show that, there exists an s in $[n]$ such that

$$\exists \delta_1, \delta_s \in \mathbb{R}^n, \text{ satisfying } \|\delta_1\|_{\infty} < 0.4, \|\delta_s\|_{\infty} < 0.4, W_1(x_1 + \delta_1) = W_1(x_s + \delta_s).$$

Firstly, since $n > K$, $W_1 \in \mathbb{R}^{K \times n}$ is not of full row rank, and hence there exists a vector $v \in \mathbb{R}^n$ such that $W_1 v = 0$ and $\|v\|_{\infty} = 1$. For such a v , let $|v^{(s)}| = 1$ for some $s \in [n]$. We define $\delta_1, \delta_s \in \mathbb{R}^n$ as follows:

$$\delta_1^{(s)} = 1/3 \text{ and } \delta_1^{(k)} = -v^{(s)}v^{(k)}/3 \text{ for } k \neq s; \quad \delta_s^{(s)} = 0 \text{ and } \delta_s^{(k)} = v^{(s)}v^{(k)}/3 \text{ for } k \neq s.$$

It is clearly that $\|\delta_1\|_{\infty} = \frac{1}{3} < 0.4$ and $\|\delta_s\|_{\infty} = \frac{1}{3} < 0.4$. Also, $x_s + \delta_s - x_1 - \delta_1 = \frac{2}{3}v^{(s)}v$. Thus, $W_1(x_1 + \delta_1) - W_1(x_s + \delta_s) = W_1(x_1 + \delta_1 - x_s - \delta_s) = W_1(\frac{2}{3}v^{(s)}v) = 0$.

It is easy to see that, for any $x, z \in \mathbb{R}^n$, $W_1 x = W_1 z$ implies $\mathcal{F}(x) = \mathcal{F}(z)$. Since $W_1(x_1 + \delta_1) = W_1(x_s + \delta_s)$, we have $\mathcal{F}(x_1 + \delta_1) = \mathcal{F}(x_s + \delta_s)$, and either $\mathcal{F}(x_1 + \delta_1) \neq 0$ or $\mathcal{F}(x_s + \delta_s) \neq 1$ must be valid. In other words, \mathcal{F} cannot be robust at x_1 or x_s for the robust budget 0.4. Similarly, \mathcal{F} cannot be robust for at least one point in $\{x_i\}_{i=k(n+1)+1}^{(k+1)(n+1)}$ for $k \in \{1, \dots, \frac{N}{n+1} - 1\}$. In summary, \mathcal{F} cannot be robust for at least $\frac{N}{n+1}$ points, so $\text{RA}_{\mathcal{D}}(\mathcal{F}, 0.4) \leq 1 - \frac{1}{n+1}$. \square

A.4. Proof of Theorem 3.9

We restate Theorem 3.9 for convenience.

Theorem A.6. *For any dataset $\mathcal{D} \in \mathcal{D}_{n,N,L}$, the hypothesis space $\mathbf{H}_{n,2N+1,3n+1,O(Nn)}$ is an optimal robust memorization for \mathcal{D} .*

Proof. It suffices to show that for any $\mu < 0.5\lambda_{\mathcal{D}}$, there exists a network with depth $2N + 1$, width $3n + 1$, and $O(Nn)$ nonzero parameters, which can robustly memorize \mathcal{D} with robust budget μ .

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^n \times [L]$. Let $C \in \mathbb{R}_+$ satisfy $C > |x_i^{(j)}| + \mu > 0$ for all $i \in [N]$ and $j \in [n]$.

\mathcal{F} will be defined in three steps for an input x .

Step 1. The first layer is used to check whether $x \in \mathbb{B}(x_1, \mu)$. The second layer is used to compute $E_1(x)$ in Property 2 given below. The two layers are given below.

$$(1-1.1) \mathcal{F}_1^0(x) = 0;$$

$$(1-1.2) \mathcal{F}_1^j(x) = \sigma(x_1^{(j)} - x^{(j)} - \mu), \mathcal{F}_1^{n+j}(x) = \sigma(x^{(j)} - x_1^{(j)} - \mu), \text{ where } j \in [n];$$

$$(1-1.3) \mathcal{F}_1^{2n+j}(x) = \sigma(x^{(j)} + C), \text{ where } j \in [n];$$

$$(1-2.1) \mathcal{F}_2^0(x) = 0;$$

$$(1-2.2) \mathcal{F}_2^1(x) = \sigma(y_1 - \frac{y_1}{\lambda_{\mathcal{D}} - 2\mu} \sum_{k=1}^{2n} \mathcal{F}_1^k(x));$$

$$(1-2.3) \mathcal{F}_2^{j+1}(x) = \sigma(\mathcal{F}_1^{2n+j}(x)), \text{ where } j \in [n].$$

Step 2. For $i = 2, 3, \dots, N$, the $(2i - 1)$ -th layer has width $3n + 1$ and is used to check whether $x \in \mathbb{B}(x_i, \mu)$. The $2i$ layer has width $n + 2$ and is used to compute $E_i(x)$ in Property 2 given below. These layers are given below.

$$(i-1.1) \mathcal{F}_{2i-1}^0(x) = \sigma(\mathcal{F}_{2i-2}^0(x) + \mathcal{F}_{2i-2}^1(x));$$

$$(i-1.2) \mathcal{F}_{2i-1}^j(x) = \sigma((x_i^{(j)} + C) - \mathcal{F}_{2i-2}^{j+1}(x) - \mu) \text{ and } \mathcal{F}_{2i-1}^{n+j}(x) = \sigma(\mathcal{F}_{2i-2}^{j+1}(x) - (x_i^{(j)} + C) - \mu), \text{ where } j \in [n];$$

$$(i-1.3) \mathcal{F}_{2i-1}^{2n+j}(x) = \sigma(\mathcal{F}_{2i-2}^{j+1}(x)), \text{ where } j \in [n];$$

$$(i-2.1) \mathcal{F}_{2i}^0(x) = \sigma(\mathcal{F}_{2i-1}^0(x));$$

$$(i-2.2) \mathcal{F}_{2i}^1(x) = \sigma(y_i - \frac{y_i}{\lambda_{\mathcal{D}} - 2\mu} \sum_{k=1}^{2n} \mathcal{F}_{2i-1}^k(x) - \mathcal{F}_{2i-1}^0(x));$$

$$(i-2.3) \mathcal{F}_{2i}^{j+1}(x) = \sigma(\mathcal{F}_{2i-1}^{2n+j}(x)), \text{ where } j \in [n].$$

Step 3. The output layer of \mathcal{F} is $\mathcal{F}(x) = \mathcal{F}_{2N}^0(x) + \mathcal{F}_{2N}^1(x)$.

Next, we will show that \mathcal{F} has the following properties.

Property 1. $\mathcal{F}_{2i}^{j+1}(x) = x^{(j)} + C$ for $i \in [N]$, $j \in [n]$, and $x \in \mathbb{R}^n$.

From (1-1.3) and (1-2.3), since $C + x_i^{(j)} > \mu > 0$ for all $i \in [N]$ and $j \in [n]$, we have that $\mathcal{F}_{2i}^{j+1}(x) = \mathcal{F}_1^{2n+j}(x) = \sigma(x^j + C) = x^j + C$.

From (i-2.3) and (i-1.3), we have that $\mathcal{F}_{2i}^{j+1}(x) = \sigma(\mathcal{F}_{2i-1}^{2n+j}(x)) = \sigma(\mathcal{F}_{2i-2}^{j+1}(x)) = \dots = \sigma(\mathcal{F}_2^{j+1}(x)) = x^{(j)} + C$, for all $i \in [N]$ and $j \in [n]$. Property 1 is proved.

Property 2. Let $E_i(x) = y_i - \frac{y_i}{\lambda_{\mathcal{D}} - 2\mu} \sum_{j=1}^{2n} \mathcal{F}_{2i-1}^j(x)$ for $i \in [N]$. Then $E_i(x) = y_i$ for $x \in \mathbb{B}_{\infty}(x_i, \mu)$, and $E_i(x) < y_i$ for $x \notin \mathbb{B}_{\infty}(x_i, \mu)$.

Due to Property 1, for $j \in [n]$, step (i-1.2) becomes

$$\begin{aligned} \mathcal{F}_{2i-1}^j(x) &= \sigma((x_i^{(j)} + C) - \mathcal{F}_{2i-2}^{j+1}(x) - \mu) \\ &= \sigma(x_i^{(j)} - x^{(j)} - \mu) \\ \mathcal{F}_{2i-1}^{n+j}(x) &= \sigma(\mathcal{F}_{2i-2}^{j+1}(x) - (x_i^{(j)} + C) - \mu) \\ &= \sigma(x^{(j)} - x_i^{(j)} - \mu). \end{aligned}$$

If $x \in \mathbb{B}_\infty(x_i, \mu)$, then $\sigma(x_i - x - \mu) = \sigma(x - x_i - \mu) = 0$, which means $\mathcal{F}_{2i-1}^j(x) = 0$ for $j \in [2n]$. Thus $E_i(x) = y_i$. If $x \notin \mathbb{B}_\infty(x_i, \mu)$, then $\|x_i - x - \mu\|_\infty > 0$ or $\|x - x_i - \mu\|_\infty > 0$ which means that $\mathcal{F}_{2i-1}^j(x) > 0$ for at least one $j \in [2n]$. Since $\mathcal{F}_i^j(x) \geq 0$ for all i and j , we have that $E_i(x) < y_i$.

Property 3. If $x \in \mathbb{B}_\infty(x_k, \mu)$ for $y_k \neq y_i$, then $E_i(x) \leq 0$.

Since $x \in \mathbb{B}_\infty(x_k, \mu)$ and $y_k \neq y_i$, we have that $\|x_i - x - \mu\|_\infty \geq \lambda_{\mathcal{D}} - 2\mu > 0$ or $\|x - x_i - \mu\|_\infty \geq \lambda_{\mathcal{D}} - 2\mu > 0$, since the separation bound is λ . Then $\mathcal{F}_{2i-1}^j(x) \geq \lambda_{\mathcal{D}} - 2\mu$ for at least one $j \in [2n]$ and thus $E_i(x) \leq y_i - \frac{y_i}{\lambda_{\mathcal{D}} - 2\mu}(\lambda_{\mathcal{D}} - 2\mu) = 0$.

Property 4. $\mathcal{F}(x) = \max_{i \in [N]} \{E_i(x), 0\}$ for $x \in \mathbb{R}^n$.

Since $\max\{x, y\} = x + \sigma(y - x)$ for $x, y \in \mathbb{R}$ and $\mathcal{F}_i^j(x) \geq 0$ for all i and j , we have that

$$\begin{aligned} \sigma(\mathcal{F}_{2i}^0(x) + \mathcal{F}_{2i}^1(x)) &= \mathcal{F}_{2i}^0(x) + \mathcal{F}_{2i}^1(x) \\ &= \sigma(\mathcal{F}_{2i-1}^0(x)) + \sigma(E_{i-1}(x) - \mathcal{F}_{2i-1}^0(x)) \\ &= \max\{\mathcal{F}_{2i-1}^0(x), E_i(x)\} \\ &= \max\{\sigma(\mathcal{F}_{2i-2}^0(x) + \mathcal{F}_{2i-2}^1(x)), E_i(x)\}. \end{aligned}$$

Using the above equation repeatedly, we have that $\mathcal{F}(x) = \sigma(\mathcal{F}_{2N}^0(x) + \mathcal{F}_{2N}^1(x)) = \max_{i=1}^N \{E_i(x), \mathcal{F}_2^0(x)\} = \max_{i=1}^N \{E_i(x), 0\}$.

We now show that \mathcal{F} satisfies the conditions of the theorem. Let $x \in \mathbb{B}_\infty(x_s, \mu)$ for $s \in [N]$. By Property 2, $E_s(x) = y_s$; and if $i \neq s$ and $y_i = y_s$, then $E_i(x) < y_s$. By Property 3, if $y_i \neq y_s$, then $E_i(x) \leq 0$. By Property 4, $\mathcal{F}(x) = \max_{i \in [N]} \{E_i(x)\} = E_s(x) = y_s$; that is, \mathcal{F} is robust at x_s with budget μ .

We now estimate the number of nonzero parameters. For $i \in [N]$, constructions (i-1.1) and (i-2.1) need 3 parameters; (i-1.2) needs $8n$ parameters; (i-1.3) and (i-2.3) need $2n$ parameters; (i-2.2) need $2n + 2$ parameters. Totally, $(N - 1)(12n + 5) + 2$ parameters are needed. \square

A.5. Proofs for Theorem 3.11

We give a lemma below.

Lemma A.7. *There exists a network*

$\mathcal{F} \in \mathbf{H}_{n, O(\log n), O(n), O(n)}$ such that $\mathcal{F}(x) = \|x\|_\infty$; that is, there exists a network $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ with depth $O(\log n)$, width $O(n)$, and $O(n)$ nonzero parameters such that $\mathcal{F}(x) = \|x\|_\infty$.

Proof. Let $e = \lceil \log_2 n \rceil$. Without loss of generality, we assume that $n = 2^e$. Then \mathcal{F} has depth $2e$ and for $i \in [e + 1]$, the $(2i - 1)$ -th layer has width 2^{e-i+2} , and the $2i$ -th layer has width 2^{e-i+1} .

Denote W_i and b_i to be the weight matrix and the bias of the i -th layer of \mathcal{F} . The first and second layers will change x to $|x|$. The first layer has width 2^{e+1} and the second layer has width 2^e , which are defined below.

$$W_1^{2i,i} = 1 \text{ and } W_1^{2i+1,i} = -1; \text{ other entries of } W_1 \text{ are } 0. \quad b_1 = 0.$$

$$W_2^{i,2i} = 1 \text{ and } W_2^{i,2i+1} = 1; \text{ other entries of } W_2 \text{ are } 0. \quad b_2 = 0.$$

Since $\sigma(x) + \sigma(-x) = |x|$ for any $x \in \mathbb{R}$, it is easy to check that $\mathcal{F}_2(x) = \sigma(W_2 \sigma(W_1 x)) = |x|$.

For $i \in [e]$, the $(2i + 1)$ -th and the $(2i + 2)$ -th layers are defined below.

$$\mathcal{F}_{2i+1}^{2m}(x) = \sigma(\mathcal{F}_{2i}^{2m}(x)), \text{ where } m = 0, 1, \dots, 2^{e-i} - 1.$$

$$\mathcal{F}_{2i+1}^{2m+1}(x) = \sigma(\mathcal{F}_{2i}^{2m+1}(x) - \mathcal{F}_{2i}^{2m}(x)), \text{ where } m = 0, 1, \dots, 2^{e-i} - 1.$$

$$\mathcal{F}_{2i+2}^m(x) = \sigma(\mathcal{F}_{2i+1}^{2m}(x) + \mathcal{F}_{2i+1}^{2m+1}(x)), \text{ where } m = 0, 1, \dots, 2^{e-i} - 1.$$

For $i \in [e + 1]$, using $\sigma(x - y) + y = \max\{x, y\}$ for any $x, y \in \mathbb{R}$, we have that

$$\begin{aligned}
 & \mathcal{F}_{2i+2}^m(x) \\
 = & \sigma(\mathcal{F}_{2i+1}^{2m}(x) + \mathcal{F}_{2i+1}^{2m+1}(x)) \\
 = & \mathcal{F}_{2i+1}^{2m}(x) + \mathcal{F}_{2i+1}^{2m+1}(x) \\
 = & \sigma(\mathcal{F}_{2i}^{2m}(x)) + \sigma(\mathcal{F}_{2i}^{2m+1}(x) - \mathcal{F}_{2i}^{2m}(x)) \\
 = & \mathcal{F}_{2i}^{2m}(x) + \sigma(\mathcal{F}_{2i}^{2m+1}(x) - \mathcal{F}_{2i}^{2m}(x)) \\
 = & \max\{\mathcal{F}_{2i}^{2m}(x), \mathcal{F}_{2i}^{2m+1}(x)\}.
 \end{aligned}$$

The $(2e + 2)$ -th layer has width 1 and is the output

$$\begin{aligned}
 \mathcal{F}(x) &= \mathcal{F}_{2e+2}^1(x) \\
 &= \max\{\mathcal{F}_{2e}^2(x), \mathcal{F}_{2e}^1(x)\} \\
 &= \max\{\mathcal{F}_{2e-2}^4(x), \mathcal{F}_{2e-2}^3(x), \mathcal{F}_{2e-2}^2(x), \mathcal{F}_{2e-2}^1(x)\} \\
 &= \dots \\
 &= \max\{\mathcal{F}_2^{2^e}(x), \mathcal{F}_2^{2^e-1}(x), \dots, \mathcal{F}_2^2(x), \mathcal{F}_2^1(x)\} \\
 &= \|x\|_\infty.
 \end{aligned}$$

We now estimate the number of parameters. The first two layers need $4d$ nonzero parameters. For $i \in [e]$, the $(2i + 1)$ -th layer and $(2i + 2)$ -th layer need $5 \cdot 2^{e-i}$ parameters. So, we need $\sum_{i=1}^e 5 \cdot 2^{e-i} = O(2^e) = O(n)$ parameters. Then the lemma is proved. \square

We restate the theorem for convenience.

Theorem A.8. *For any dataset $\mathcal{D} \in \mathcal{B}_{n,N}$, the hypothesis space $\mathbf{H}_{n,O(N \log(n)),O(n),O(Nn \log(n))}$ contains a network \mathcal{F} which is an optimal robust memorization of \mathcal{D} via Lipschitz; that is, \mathcal{F} is a memorization of \mathcal{D} and $\text{Lip}_\infty(\mathcal{F}) = 2/\lambda_{\mathcal{D}}$.*

Proof. Let \mathcal{D} be defined in section 3.4 and $C \in \mathbb{R}_+$ satisfy $C + x_i^{(k)} - 0.5\lambda_{\mathcal{D}} > 0$ for all $i \in [N], k \in [n]$. The network has $N(2\lceil \log(n) \rceil + 5) + 1$ hidden layers which will be defined below.

Step 1. The first layer has width $n + 1$: $\mathcal{F}_1^0(x) = 2$ and $\mathcal{F}_1^j(x) = \sigma(x^{(j)} + C) = x^{(j)} + C$, where $j \in [n]$.

Step 2. Let $s_k = (2\lceil \log(n) \rceil + 5)(k - 1) + 2$. For $k \in [N]$, we will use the s_k -th layer to the $(s_k + 2\lceil \log(n) \rceil + 4)$ -th layer to check if $\|x - x_k\|_\infty < 0.5\lambda_{\mathcal{D}}$. Step 2 consists of three sub-steps.

Step 2a. We use the s_k -th layer and the $(s_k + 1)$ -th layer to calculate $|x - x_k|$. The s_k -th layer has width $3n + 1$ and is defined below.

$$\begin{aligned}
 \mathcal{F}_{s_k}^0(x) &= \sigma(\mathcal{F}_{s_k-1}^0(x)); \\
 \mathcal{F}_{s_k}^j(x) &= \sigma(\mathcal{F}_{s_k-1}^j(x) - x_k^{(j)} - C), \text{ where } j \in [n]; \\
 \mathcal{F}_{s_k}^{n+j}(x) &= \sigma(-\mathcal{F}_{s_k-1}^j(x) + x_k^{(j)} + C), \text{ where } j \in [n]; \\
 \mathcal{F}_{s_k}^{2n+j}(x) &= \sigma(\mathcal{F}_{s_k-1}^j(x)), \text{ where } j \in [n].
 \end{aligned}$$

The $(s_k + 1)$ -th layer has width $2n + 1$ and is defined below.

$$\begin{aligned}
 \mathcal{F}_{s_k+1}^0(x) &= \sigma(\mathcal{F}_{s_k}^0(x)); \\
 \mathcal{F}_{s_k+1}^j(x) &= \sigma(\mathcal{F}_{s_k}^j(x) + \mathcal{F}_{s_k}^{n+j}(x)), \text{ where } j \in [n]; \\
 \mathcal{F}_{s_k+1}^{n+j}(x) &= \sigma(\mathcal{F}_{s_k}^{2n+j}(x)), \text{ where } j \in [n].
 \end{aligned}$$

The s_k -th layer needs $5n + 1$ nonzeros parameters and $(s_k + 1)$ -th layer needs $3n + 1$ nonzeros parameters.

Step 2b. Lemma A.7 is used to calculate $\|x - x_k\|_\infty$. By Lemma A.7, there exists a network $\mathcal{H} : \mathbb{R}^n \rightarrow \mathbb{R}$ with $2\lceil \log(n) \rceil$ hidden layers, width $O(n)$, and $O(n)$ nonzero parameters to compute $\mathcal{H}(x) = \|x\|_\infty$ for $x \in \mathbb{R}^n$. Since \mathcal{H} has $2\lceil \log(n) \rceil$ hidden layers, we set the output of the $(s_k + 2\lceil \log(n) \rceil + 1)$ -th layer to be

$$\begin{aligned} \mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^0(x) &= \sigma(\mathcal{F}_{s_k+1}^0(x)); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^1(x) &= \mathcal{H}(\mathcal{F}_{s_k+1}^1(x), \dots, \mathcal{F}_{s_k+1}^n(x)) = \|\mathcal{F}_{s_k+1}(x)\|_\infty; \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^{j+1}(x) &= \sigma(\mathcal{F}_{s_k+1}^{n+j}(x)), \text{ where } j \in [n]. \end{aligned}$$

Step 2c. Use the $(s_k + 2\lceil \log(n) \rceil + 2)$ -th to the $(s_k + 2\lceil \log(n) \rceil + 4)$ -th layers to check if $\|x - x_k\|_\infty < 0.5\lambda_{\mathcal{D}}$. The $(s_k + 2\lceil \log(n) \rceil + 2)$ -th layer has width $n + 4$ and is defined below

$$\begin{aligned} \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^0(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^0(x)); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^1(x) &= \sigma(-2/\lambda_{\mathcal{D}}\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^1(x) + 1); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^2(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^0(x) - 2); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^3(x) &= \sigma(-\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^0(x) + 2); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^{j+3}(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^{j+1}(x)), \text{ where } j \in [n]. \end{aligned}$$

The $(s_k + 2\lceil \log(n) \rceil + 3)$ -th layer has width $n + 3$ and is defined below

$$\begin{aligned} \mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^0(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^0(x) + y_k\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^1(x)); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^1(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^1(x)); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^2(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^1(x) - (\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^2(x) + \mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^3(x))); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^{j+2}(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^{j+3}(x)), \text{ where } j \in [n]. \end{aligned}$$

The $(s_k + 2\lceil \log(n) \rceil + 4)$ -th layer has width $n + 1$ and is defined as

$$\begin{aligned} \mathcal{F}_{s_k+2\lceil \log(n) \rceil+4}^0(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^0(x) - y_k(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^1(x) - \mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^2(x))); \\ \mathcal{F}_{s_k+2\lceil \log(n) \rceil+4}^j(x) &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^{j+2}(x)), \text{ where } j \in [n]. \end{aligned}$$

It is easy to check that if $\mathcal{F}_{s_k+1}^j(x) = |x^{(j)} - x_k^{(j)}|$. Then

$$\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^1(x) = \sigma(-2/\lambda_{\mathcal{D}}\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^1(x) + 1) > 0$$

if and only if $\|x - x_k\|_\infty < 0.5\lambda_{\mathcal{D}}$. These three layers need $3n + 16$ nonzeros parameters.

Step 3. The output is $\mathcal{F}(x) = \mathcal{F}_{s_N+2\lceil \log(n) \rceil+4}^0(x) - 2$. The network \mathcal{F} has width $O(n)$, depth $O(N \log(n))$, and $O(Nn \log(n))$ nonzeros parameters.

We now show that \mathcal{F} satisfies the condition of the theorem; that is \mathcal{F} memorizes \mathcal{D} and satisfies $\text{Lip}_\infty(\mathcal{F}) = 2/\lambda_{\mathcal{D}}$.

Property 1. $\mathcal{F}_{s_k-1}^j(x) = x^{(j)} + C$ for $j \in [n]$ and $k \in [N]$. When $k = 1$, $s_k - 1 = 1$. By Step 1, we have that $\mathcal{F}_{s_1-1}^j(x) = \mathcal{F}_1^j(x) = x^{(j)} + C$. When $k > 1$, we have that

$$\begin{aligned} &\mathcal{F}_{s_k+1-1}^j(x) \\ &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+4}^{j+2}(x)) = \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+3}^{j+2}(x)) \\ &= \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+2}^{j+3}(x)) = \sigma(\mathcal{F}_{s_k+2\lceil \log(n) \rceil+1}^{j+1}(x)) \\ &= \sigma(\mathcal{F}_{s_k+1}^{n+j}(x)) = \sigma(\mathcal{F}_{s_k}^{2n+j}(x)) = \sigma(\mathcal{F}_{s_k-1}^j(x)) \\ &= \mathcal{F}_{s_k-1}^j(x). \end{aligned}$$

Then, $\mathcal{F}_{s_{k+1}-1}^j(x) = \mathcal{F}_{s_k-1}^j(x) = \dots = \mathcal{F}_{s_1-1}^j(x) = \mathcal{F}_1^j(x) = x^{(j)} + C$.

Property 2. $\mathcal{F}_{s_k+1}^j(x) = |x^{(j)} - x_k^{(j)}|$ and $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x) = \|x - x_k\|_\infty$ for $j \in [n]$.

Since $\sigma(x) + \sigma(-x) = |x|$ for any $x \in \mathbb{R}$, from Step 2a, $\mathcal{F}_{s_k+1}^j(x) = |\mathcal{F}_{s_k-1}^j(x) - x_k^{(j)} - C|$ for $j \in [n]$. By Property 1, $\mathcal{F}_{s_k-1}^j(x) = x^{(j)} + C$ for $j \in [n]$. Then, $\mathcal{F}_{s_k+1}^j(x) = |x^{(j)} - x_k^{(j)}|$ for $j \in [n]$. From Step 2b, we have that $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x) = \|x - x_k\|_\infty$ for $j \in [n]$.

Property 3. $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) = 2 + y_{w_k} \sigma(1 - 2/\lambda_{\mathcal{D}} \|x - x_{w_k}\|_\infty)$, where $w_k = \operatorname{argmin}_{i \in [k]} \|x - x_i\|_\infty$.

We prove the property by induction on k . We first show that the statement is valid for $k = 1$. We have that $w_k = 1$ and $\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^0(x) = \mathcal{F}_{s_1+2\lceil\log(n)\rceil+1}^0(x) = \mathcal{F}_{s_1+1}^0(x) = \mathcal{F}_{s_1}^0(x) = \mathcal{F}_{s_1-1}^0(x) = 2$. From Step 2c and Property 2,

$$\begin{aligned} & \mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^0(x) \\ &= \sigma(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^0(x) + y_0 \mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^1(x)) \\ &= \sigma(2 + y_0 \sigma(1 - 2/\lambda_{\mathcal{D}} \mathcal{F}_{s_1+2\lceil\log(n)\rceil+1}^1(x))) \\ &= 2 + y_0 \sigma(-2/\lambda_{\mathcal{D}} \mathcal{F}_{s_1+2\lceil\log(n)\rceil+1}^1(x) + 1) \\ &= 2 + y_0 \sigma(1 - 2/\lambda_{\mathcal{D}} \|x - x_0\|_\infty). \end{aligned}$$

Since $\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^2(x) = \sigma(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+1}^0(x) - 2) = \sigma(2 - 2) = 0$ and $\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^3(x) = \sigma(-\mathcal{F}_{s_1+2\lceil\log(n)\rceil+1}^0(x) + 2) = \sigma(2 - 2) = 0$, we have that $\mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^2(x) = \sigma(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^1(x) - (\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^2(x) + \mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^3(x))) = \sigma(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+2}^1(x)) = \mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^1$. Then

$$\begin{aligned} & \mathcal{F}_{s_1+2\lceil\log(n)\rceil+4}^0(x) \\ &= \sigma(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^0(x) - y_0(\mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^1(x) - \mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^2(x))) \\ &= \mathcal{F}_{s_1+2\lceil\log(n)\rceil+3}^0(x) \\ &= 2 + y_0 \sigma(1 - 2/\lambda_{\mathcal{D}} \|x - x_0\|_\infty). \end{aligned}$$

We have proved the statement for $k = 1$.

Assume that the statement is valid for $k - 1$; that is, $\mathcal{F}_{s_{k-1}+2\lceil\log(n)\rceil+4}^0(x) = 2 + y_{w_{k-1}} \sigma(1 - 2/\lambda_{\mathcal{D}} \|x - x_{w_{k-1}}\|_\infty)$. We have that $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^0(x) = \mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^0(x) = \mathcal{F}_{s_k+1}^0(x) = \mathcal{F}_{s_k}^0(x) = \mathcal{F}_{s_k-1}^0(x) = 2 + y_{w_{k-1}} \sigma(1 - 2/\lambda_{\mathcal{D}} \|x - x_{w_{k-1}}\|_\infty) \geq 1$, and we also have that $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) = \sigma(-2/\lambda_{\mathcal{D}} \mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x) + 1) \leq 1$. Then

$$\begin{aligned} & \mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^0(x) \\ &= \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^0(x) + y_k \mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x)) \\ &= \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^0(x) + y_k \sigma(1 - 2/\lambda_{\mathcal{D}} \mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x))) \\ &= \mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^0(x) + y_k \sigma(1 - 2/\lambda_{\mathcal{D}} \mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x)) \\ &= \mathcal{F}_{s_k-1}^0(x) + y_k \sigma(1 - 2/\lambda_{\mathcal{D}} \mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x)) \\ &= \mathcal{F}_{s_k-1}^0(x) + y_k \mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) \end{aligned} \tag{23}$$

Since $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^2(x) = \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^0(x) - 2)$ and $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^3(x) = \sigma(-\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^0(x) + 2)$, we have that $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^2(x) = \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - (\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^2(x) + \mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^3(x))) = \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x)) = \mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^1$.

935 $|\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^0(x) - 2|$). Then

$$\begin{aligned}
 &936 \\
 &937 \\
 &938 \\
 &939 \\
 &940 \\
 &941 \\
 &942 \\
 &943 \\
 &944 \\
 &945 \\
 &946 \\
 &947 \\
 &948 \\
 &949
 \end{aligned}$$

$$\begin{aligned}
 &= \mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) \\
 &= \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^0(x) - \\
 &\quad y_k(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^1(x) - \mathcal{F}_{s_k+2\lceil\log(n)\rceil+3}^2(x))) \\
 &= \sigma(\mathcal{F}_{s_k-1}^0(x) + y_k\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) \\
 &\quad - y_k(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - \\
 &\quad |\mathcal{F}_{s_k-1}^0(x) - 2|))) .
 \end{aligned}$$

We divide the proof into two cases.

950 **Case 1.** If $x \notin \mathbb{B}_\infty(x_k, 0.5\lambda_{\mathcal{D}})$, then $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) = \sigma(-2/\lambda_{\mathcal{D}}\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x) + 1) = \sigma(1 - 2/\lambda_{\mathcal{D}}\|x -$
 951 $x_k\|_\infty) = 0$ and

$$\begin{aligned}
 &952 \\
 &953 \\
 &954 \\
 &955 \\
 &956 \\
 &957 \\
 &958 \\
 &959 \\
 &960 \\
 &961 \\
 &962 \\
 &963 \\
 &964 \\
 &965 \\
 &966
 \end{aligned}$$

$$\begin{aligned}
 &\mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) \\
 &= \sigma(\mathcal{F}_{s_k-1}^0(x) + y_k\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) \\
 &\quad - y_k(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - \\
 &\quad \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - |\mathcal{F}_{s_k-1}^0(x) - 2|))) \\
 &= \mathcal{F}_{s_k-1}^0(x) \\
 &= \mathcal{F}_{s_k-1+2\lceil\log(n)\rceil+4}^0(x) \\
 &= 2 + y_{w_{k-1}}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty) \\
 &= 2 + y_{w_k}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_k}\|_\infty).
 \end{aligned}$$

967 **Case 2.** If $x \in \mathbb{B}_\infty(x_k, 0.5\lambda_{\mathcal{D}})$, then $\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) = \sigma(-2/\lambda_{\mathcal{D}}\mathcal{F}_{s_k+2\lceil\log(n)\rceil+1}^1(x) + 1) = \sigma(1 - 2/\lambda_{\mathcal{D}}\|x -$
 968 $x_k\|_\infty) \geq 0$ and using equation 23:

$$\begin{aligned}
 &969 \\
 &970 \\
 &971 \\
 &972 \\
 &973 \\
 &974 \\
 &975 \\
 &976 \\
 &977 \\
 &978 \\
 &979 \\
 &980 \\
 &981 \\
 &982 \\
 &983 \\
 &984 \\
 &985 \\
 &986 \\
 &987 \\
 &988 \\
 &989
 \end{aligned}$$

$$\begin{aligned}
 &\mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) \\
 &= \sigma(\mathcal{F}_{s_k-1}^0(x) + y_k\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) \\
 &\quad - y_k(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - \sigma(\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) - \\
 &\quad - |\mathcal{F}_{s_k-1}^0(x) - 2|))) \\
 &= \sigma(\mathcal{F}_{s_k-1}^0(x) + y_k\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x) \\
 &\quad - y_k(\min\{\mathcal{F}_{s_k+2\lceil\log(n)\rceil+2}^1(x), |2 - \mathcal{F}_{s_k-1}^0(x)|\})) \\
 &= \sigma(2 + y_{w_{k-1}}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty) + \\
 &\quad y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty) \\
 &\quad - y_k(\min\{1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty, \\
 &\quad \sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty)\})).
 \end{aligned}$$

Consider two sub-cases:

990 **Case 2.1.** If $\|x - x_{w_{k-1}}\|_\infty > 0.5\lambda_{\mathcal{D}}$, then $w_k = k$ and hence

$$\begin{aligned}
 & \mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) \\
 &= \sigma(2 + y_{w_{k-1}}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty) + \\
 & \quad y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty) \\
 & \quad - y_k(\min\{1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty, \\
 & \quad \quad \sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty)\})) \\
 &= \sigma(2 + y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty)) \\
 &= 2 + y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty) \\
 &= 2 + y_{w_k}(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_k}\|_\infty).
 \end{aligned}$$

1004 **Case 2.2.** If $\|x - x_{w_{k-1}}\|_\infty \leq 0.5\lambda_{\mathcal{D}}$, then $y_{w_{k-1}} = y_k$ and hence

$$\begin{aligned}
 & \mathcal{F}_{s_k+2\lceil\log(n)\rceil+4}^0(x) \\
 &= \sigma(2 + y_{w_{k-1}}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty) + \\
 & \quad y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty) \\
 & \quad - y_k(\min\{1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty, \\
 & \quad \quad \sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty)\})) \\
 &= \sigma(2 + y_{w_{k-1}}(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty) \\
 & \quad + y_k(1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty) \\
 & \quad - y_k(\min\{1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty, \\
 & \quad \quad 1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty\})) \\
 &= 2 + y_k \max\{1 - 2/\lambda_{\mathcal{D}}\|x - x_k\|_\infty, \\
 & \quad 1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_{k-1}}\|_\infty\} \\
 &= 2 + y_{w_k}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_k}\|_\infty).
 \end{aligned}$$

1024 The property is proved.

1025 **Property 4.** \mathcal{F} is a memorization \mathcal{D} and has $\text{Lip}_\infty(\mathcal{F}) = 2/\lambda_{\mathcal{D}}$.

1026 By Property 3, the output is

$$1027 \mathcal{F}(x) = \mathcal{F}_{s_N+2\lceil\log(n)\rceil+4}^1(x) - 2 = y_{w_N}\sigma(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_N}\|_\infty)$$

1030 where $w_N = \operatorname{argmin}_{i \in [N]} \|x - x_i\|_\infty$.

1032 If $x = x_s$, then $w_N = s$ and $\mathcal{F}(x) = y_s$; that is, \mathcal{F} memorizes \mathcal{D} . If $x \in \mathbb{B}(x_s, 0.5\lambda_{\mathcal{D}})$ for some $s \in [N]$, then $w_N \in [N]$
 1033 and $\mathcal{F}(x) = y_{w_N}(1 - 2/\lambda_{\mathcal{D}}\|x - x_{w_N}\|_\infty)$ such that the local $\text{Lip}_\infty(\mathcal{F}) = 2/\lambda_{\mathcal{D}}$ over $\mathbb{B}(x_{w_N}, 0.5\lambda_{\mathcal{D}})$. If x is not in
 1034 $\cup_{i=1}^N \mathbb{B}(x_s, 0.5\lambda_{\mathcal{D}})$, then $\|x - x_{w_N}\|_\infty > 0.5\lambda_{\mathcal{D}}$. Hence $\mathcal{F}(x) = 0$ and the local $\text{Lip}_\infty(\mathcal{F}) = 0$. The theorem is proved. \square