Proceedings of Machine Learning Research vol 134:1–44, 2021 34th Annual Conference on Learning Theory

Robust learning under clean-label attack

Avrim Blum AVRIM@TTIC.EDU
Toyota Technological Institute at Chicago

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Jian Qian JIANQIAN@MIT.EDU
Massachusetts Institute of Technology

Han Shao HAN@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of robust learning under clean-label data-poisoning attacks, where the attacker injects (an arbitrary set of) *correctly-labeled* examples to the training set to fool the algorithm into making mistakes on *specific* test instances at test time. The learning goal is to minimize the attackable rate (the probability mass of attackable test instances), which is more difficult than optimal PAC learning. As we show, any robust algorithm with diminishing attackable rate can achieve the optimal dependence on ε in its PAC sample complexity, i.e., $O(1/\varepsilon)$. On the other hand, the attackable rate might be large even for some optimal PAC learners, e.g., SVM for linear classifiers. Furthermore, we show that the class of linear hypotheses is not robustly learnable when the data distribution has zero margin and is robustly learnable in the case of positive margin but requires sample complexity exponential in the dimension. For a general hypothesis class with bounded VC dimension, if the attacker is limited to add at most $t > 0$ poison examples, the optimal robust learning sample complexity grows almost linearly with t.

Keywords: adversarial machine learning, data poisoning, clean-label attack, PAC learning, sample complexity.

1. Introduction

Data poisoning is an attack on machine learning algorithms where the attacker adds examples to the training set with the goal of causing the algorithm to produce a classifier that makes specific mistakes the attacker wishes to induce at test time. In this paper, we focus on clean-label attacks in which an attacker, with knowledge of the training set S and the test instance x, injects a set of examples labeled by the target function into the training set with the intent of fooling the learner into misclassifying the test instance x. This type of attack is called a clean-label attack because the attacker can only add correctly-labeled examples to the training set, and it has been proposed and studied empirically by Shafahi et al. (2018).

In the realizable setting when the target function belongs to the hypothesis class $\mathcal{H}$, any empirical risk minimizer (ERM) will achieve error of $\widetilde{O}(\frac{\text{VCdim}(\mathcal{H})}{m})$ with training set size m. This means that an ERM learner will still have error rate at most $\widetilde{O}(\frac{\text{VCdim}(\mathcal{H})}{m})$ even in the presence of a clean-label attack; i.e., the attacker cannot significantly increase the overall *error rate*. However, an attacker could still cause the ERM learner to make *specific* mistakes that the attacker wishes. For example, consider an ERM learner for the hypothesis class of intervals over $[0, 1]$ that predicts

the positive interval of maximum length consistent with the training data, in the case that the target function labels all of $[0, 1]$ negative. Then any test instance not in the training set is attackable for this ERM learner by an adversary that adds enough poison examples so that the interval that the test instance is in becomes the largest interval in the training set. On the other hand, for any target interval, for the ERM learner that outputs the *smallest* consistent interval, the attackable test instances will only have probability mass $O(1/m)$ (see Example 1 for more details). Also, notice that for the hypothesis class of threshold functions over $[0, \infty)$, any ERM learner has a small portion of attackable test instances because the disagreement region of all consistent hypotheses is small and only test instances in the disagreement region are attackable.

From these examples, we can see that given an ERM learning algorithm $\mathcal{A}$ and a training set S, the probability mass of the attackable region (the set of attackable test instances) is at least as large as the error rate of the ERM learner and no greater than the disagreement region of all consistent hypotheses, and it depends on the specific algorithm $\mathcal{A}$. In this paper, we study the problem of whether we can obtain a small rate of attackable test instances in binary classification. In the process we find interesting connections to existing literature on the sample complexity of PAC learning, and complexity measures arising in that literature. Specifically, we study this problem in the realizable setting as it is unclear how to best define "clean-label" in the agnostic case.

Related work Clean-label data-poisoning attacks have been studied extensively in the literature (Shafahi et al., 2018; Suciu et al., 2018), and Shafahi et al. (2018) show that clean-label attacks can be very effective on neural nets empirically. For example, Shafahi et al. (2018) show that in natural image domains, given the knowledge of the training model and of the test point to be attacked, the attacker can cause the model retrained with an injection of clean-label poisoned data to misclassify the given test instance with high success rate. Moreover, the attacker is able to succeed even though the overall error rate of the trained classifier remains relatively unchanged.

Mahloujifar and Mahmoody (2017); Mahloujifar et al. (2018, 2019b) study a class of clean-label poisoning attacks called p-tampering attacks, where the attacker can substitute each training example with a correctly labeled poison example with independent probability p, and its variants. Mahloujifar and Mahmoody (2019); Mahloujifar et al. (2019a); Etesami et al. (2020) consider a more powerful adversary that can attack training examples of its choosing (rather than chosen at random) and show that the attacker can increase the probability of failing on a particular test instance from any non-negligible probability $\Omega(1/\operatorname{poly}(m))$ to ≈ 1 by replacing $\widetilde{O}(\sqrt{m})$ training examples with other correctly labeled examples. In contrast, in our setting the attacker cannot modify any of the existing training examples and can only add new ones. In addition, we mainly focus on attacks with an unlimited budget.

Data poisoning without requiring the poisoned data to be clean has been studied extensively (see Biggio et al. (2012); Barreno et al. (2006); Papernot et al. (2016); Steinhardt et al. (2017) for a non-exhaustive list). Robustness to data poisoning with a small portion of poison examples has been studied by Ma et al. (2019); Levine and Feizi (2020). The concurrent work of Gao et al. (2021) studies the instance-targeted poisoning risk (which is the probability mass of the attackable region in the classification task) by various attacker classes, which have a budget controlling the amount of training data points they can change. They mainly focus on the relationship between robust learnability and the budget.

There are other studied attacking methods, including perturbation over training examples (Koh and Liang, 2017), perturbation over test examples (Szegedy et al., 2013; Goodfellow et al., 2014;

Bubeck et al., 2019; Cullina et al., 2018; Montasser et al., 2019, 2020) and etc. Another different notion of robust learning is studied by Xu and Mannor (2012), where the data set is partitioned into several subsets and the goal is to ensure the losses of instances falling into the same subset are close. Another line of related work is covariate shift, where the training distribution is different from the test distribution (see Quionero-Candela et al. (2009) for an extensive study).

Notation For any vectors u, v, we let $\|u\|$ denote the ℓ_2 norm of u and $\theta(u, v)$ denote the angle of u and v. We denote by $e_i \in \mathbb{R}^n$ the one-hot vector with the i-th entry being one and others being zeros. We let $\mathcal{B}^n(c, r) = \{x | \|x - c\| \leq r\}$ denote the the ball with radius r centered at $c \in \mathbb{R}^n$ in the n-dimensional space and $\Gamma^n(c, r)$ denote the sphere of $\mathcal{B}^n(c, r)$. We omit the supscript n when it is clear from the context. For any $a, b \in \mathbb{R}$, denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. We use $\ln$ to represent natural logarithms and $\log$ to represent logarithms with base 2. Given a data set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ with size m, for any hypothesis h, we let $\mathrm{err}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[h(x_i) \neq y_i]$ denote the empirical error of h over S. For a data distribution $\mathcal{D}$, we let $\mathrm{err}_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}[h(x) \neq y]]$ denote the error of h. For any $A \subseteq \mathcal{X}$, we let $\mathcal{P}_{\mathcal{D}}(A) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in A)$ denote the probability mass of A. The subscript $\mathcal{D}$ is omitted when it is clear from the context. For any data set S, we let $S_{\mathcal{X}} = \{x | (x, y) \in S\}$ and for $(x, y) \sim \mathcal{D}$, we let $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution of x. For a finite set of hypotheses $\mathcal{H}$, we let $\mathrm{Major}(\mathcal{H})$ denote the majority vote of $\mathcal{H}$ and for simplicity denote $\mathrm{Major}(\mathcal{H}, x) = \mathrm{Major}(\mathcal{H})(x) \triangleq \mathbb{1}[\sum_{h\in\mathcal{H}} h(x) \geq \lceil |H|/2 \rceil]$.

2. Problem setup and summary of results

Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} = \{0, 1\}$ denote the label space. Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, we study the realizable case where there exists a deterministic target function $h^* \in \mathcal{H}$ such that the training set and the test set are realized by h^*. Let $D_{h^*} = \{(x, h^*(x)) | x \in \mathcal{X}\}$ denote the data space where every instance is labeled by h^*. A learning algorithm $\mathcal{A}$ is a map (possibly including randomization), from a labeled data set S (an unordered multiset) of any size, to a hypothesis h, and for simplicity we denote by $\mathcal{A}(S, x) = \mathcal{A}(S)(x)$ the prediction of $\mathcal{A}(S)$ at an instance x. An attacker Adv maps a target function h^*, a training data set S_{trn} and a specific test instance x to a data set $\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x)$ (a multiset) and injects $\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x)$ into the training set with the intent of making the learning algorithm misclassify x. We call Adv a *clean-label* attacker if $\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x)$ is consistent with h^*. Then for any deterministic algorithm $\mathcal{A}$, we say a point $x \in \mathcal{X}$ is attackable if there exists a clean-label attacker Adv such that

$$\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x) .$$

To be clear, we are defining $S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x)$ as an unordered multiset. Formally, we define clean-label attackable rate as follows.

Definition 1 (clean-label attackable rate) *For a target function h^*, a training data set S_{trn} and a (possibly randomized) algorithm $\mathcal{A}$, for any distribution $\mathcal{D}$ over D_{h^*}, the attackable rate by Adv for $(h^*, S_{\mathrm{trn}}, \mathcal{A})$ is defined as*

$$\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D},\mathcal{A}}[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x)]] .$$

The clean-label attackable rate is defined by the supremum over all clean-label attackers, i.e.,

$$\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \triangleq \sup_{\mathrm{Adv}} \mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) .$$

Then we define our learning problem as follows.

Definition 2 ((ε,δ)-robust learnability) *For any $\varepsilon,\delta \in (0,1)$, the sample complexity of (ε,δ)-robust learning of $\mathcal{H}$, denoted by $\mathcal{M}_{\mathrm{rbst}}(\varepsilon,\delta)$, is defined as the smallest $m \in \mathbb{N}$ for which there exists an algorithm $\mathcal{A}$ such that for every target function $h^* \in \mathcal{H}$ and data distribution over D_{h^*}, with probability at least $1-\delta$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$,*

$$\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \varepsilon\,.$$

If no such m exists, define $\mathcal{M}_{\mathrm{rbst}}(\varepsilon,\delta) = \infty$. We say that $\mathcal{H}$ is (ε,δ)-robust learnable if $\forall \varepsilon,\delta \in (0,1)$, $\mathcal{M}_{\mathrm{rbst}}(\varepsilon,\delta)$ is finite.

It is direct to see that the error of $\mathcal{A}(S_{\mathrm{trn}})$ is the attackable rate by attacker Adv_0 which injects an empty set to the training set, i.e., $\mathrm{Adv}_0(\cdot) = \emptyset$. Therefore, for any algorithm $\mathcal{A}$, we have

$$\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \geq \mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}_0) = \mathrm{err}_{\mathcal{D}}(\mathcal{A}(S_{\mathrm{trn}}))\,,$$

which indicates any hypothesis class that is not PAC learnable is not robust learnable. For any deterministic $\mathcal{A}$, let us define $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \triangleq \{x \in \mathcal{X} | \mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x)\}$ the attackable region by Adv. For any ERM learner and any clean-label attacker Adv, we have $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathrm{ERM}, \mathrm{Adv}) \subseteq \mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}})$, where $\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}}$ is the version space of S_{trn}, i.e., the set of all hypotheses in $\mathcal{H}$ that classify S_{trn} correctly and $\mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}}) = \{x | \exists h, h' \in \mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}}, h(x) \neq h'(x)\}$ is the disagreement region of the version space. Therefore, we have

$$\inf_{\mathcal{A}} \mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathcal{P}_{\mathcal{D}}(\mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}}))\,.$$

However, large $\mathcal{P}_{\mathcal{D}}(\mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}}))$ does not always result in large clean-label attackable rate. Below is an example showing the gap between them.

Example 1 (Interval over $[0,1]$) *The hypothesis class $\mathcal{H} = \{\mathbb{1}[(a,b)] : 0 \leq a \leq b \leq 1\} \cup \{\mathbb{1}[[a,b]] : 0 \leq a \leq b \leq 1\}$ contains all intervals on $[0,1]$. We consider the following two learners.*

- *$\mathcal{A}_1(S)$: return $\mathbb{1}[\emptyset]$ (the empty interval) if there are no positive examples in S and return the consistent positive closed interval with minimum length otherwise.*
- *$\mathcal{A}_2(S)$: return the consistent positive open interval with maximum length.*

Both are ERM learners for $\mathcal{H}$. For any $h^* \in \mathcal{H}$, let the data distribution $\mathcal{D}$ be a distribution on D_{h^*} and $S_{\mathrm{trn}} \sim \mathcal{D}^m$ for any $m > 0$, then $\mathcal{A}_1$'s attackable rate is $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}_1) = \mathrm{err}_{\mathcal{D}}(\mathcal{A}_1(S_{\mathrm{trn}})) = \widetilde{O}(1/m)$. However, consider algorithm $\mathcal{A}_2$ with $h^* = \mathbb{1}[\emptyset]$. For $S_{\mathrm{trn}} = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, w.l.o.g. assume that $x_1 \leq \ldots \leq x_m$ and let $x_0 = 0$, $x_{m+1} = 1$ for notation simplicity. Then for any $x \in (x_i, x_{i+1})$, the attacker can add enough poison data points to intervals $\{(x_j, x_{j+1})\}_{j \neq i}$ to make (x_i, x_{i+1}) be the interval with the maximum length. Therefore, so long as $\mathcal{D}$ has no point masses, $\mathcal{A}_2$'s attackable rate is $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}_2) = \mathcal{P}_{\mathcal{D}}(\mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}})) = 1$.

Main results We summarize the main contributions of this work.

- In Section 3, we present results on robust learnability under assumptions based on some known structural complexity measures, e.g., VC dimension $d = 1$, hollow star number $k_o = \infty$, etc. In addition, we show that all robust algorithms can achieve optimal dependence on ε in their PAC sample complexity.

- In Section 4, we show that the n-dimensional linear hypothesis class with $n \geq 2$ is not (ε, δ)-robust learnable. Then we study the linear problem in the case where the data distribution $\mathcal{D}$ has margin $\gamma > 0$. We propose one algorithm with sample complexity $O(n(2/\gamma)^n \log(2/\gamma))$ and show that the optimal sample complexity is $e^{\Omega(n)}$. We propose another algorithm in 2-dimensional space with sample complexity $O(\log(1/\gamma) \log\log(1/\gamma))$. We also show that even in the case where γ is large and the attacker is only allowed to inject one poison example into the training set, SVM requires at least $e^{\Omega(n)}$ samples to achieve low attackable rate.

- In Section 5, we show that for any hypothesis class $\mathcal{H}$ with VC dimension d, when the attacker is restricted to inject at most t poison examples, $\mathcal{H}$ is robust learnable with sample complexity $\widetilde{O}(\frac{dt}{\varepsilon})$. We also show that there exists a hypothesis class with VC dimension d such that any algorithm requires $\Omega(\frac{dt}{\varepsilon})$ samples to achieve ε attackable rate.

3. Connections to some known complexity measures and PAC learning

In this section, we analyze the robust learnability of hypothesis classes defined by a variety of known structural complexity measures. For some of these, we show they have the good property that there exists an algorithm such that adding clean-label points can only change the predictions on misclassified test instances and thus, the algorithm can achieve $\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathrm{err}(\mathcal{A}(S_{\mathrm{trn}}))$. For some other structure, we prove that there will be a large attackable rate for any consistent proper learner. We also show the connection to optimal PAC learning in Section 3.2.

3.1. Connections to some known complexity measures

Hypothesis classes with VC dimension $d = 1$ are (ε, δ)-robust learnable. First, w.l.o.g., assume that for every $x \neq x' \in \mathcal{X}$, there exists $h \in \mathcal{H}$ such that $h(x) \neq h(x')$ (otherwise, operate over the appropriate equivalence classes). Then we adopt the partial ordering $\leq_f^{\mathcal{H}}$ for any $f \in \mathcal{H}$ over $\mathcal{X}$ proposed by Ben-David (2015) defined as follows.

Definition 3 (partial ordering $\leq_f^{\mathcal{H}}$) *For any $f \in \mathcal{H}$,*

$$\leq_f^{\mathcal{H}} \triangleq \{(x, x') | \forall h \in \mathcal{H}, h(x') \neq f(x') \Rightarrow h(x) \neq f(x)\}.$$

By Lemma 5 of Ben-David (2015), $\leq_f^{\mathcal{H}}$ for $d = 1$ is a tree ordering. Due to this structural property of hypothesis classes with VC dimension $d = 1$, there is an algorithm originally proposed by Ben-David (2015) (Algorithm 4 in Appendix A.1) such that adding clean-label poison points can only narrow down the error region (the set of misclassified instances). Roughly, the algorithm finds a maximal (by $\leq_f^{\mathcal{H}}$) point x' in the data such that $h^*(x') \neq f(x')$, and outputs the classifier labeling all $x \leq_f^{\mathcal{H}} x'$ as $1 - f(x)$ and the rest as $f(x)$. We show that this algorithm can robustly learn $\mathcal{H}$ using m samples, where

$$m = \frac{2\ln(1/\delta)}{\varepsilon}.$$

The detailed algorithm and proof are given in Appendix A.1.

Intersection-closed hypothesis classes are (ε, δ)-robust learnable. A hypothesis class $\mathcal{H}$ is called intersection-closed if the collection of sets $\{\{x|h(x) = 1\}|h \in \mathcal{H}\}$ is closed under intersections, i.e., $\forall h, h' \in \mathcal{H}$, the classifier $x \mapsto \mathbb{1}[h(x) = h'(x) = 1]$ is also contained in $\mathcal{H}$. For intersection-closed hypothesis classes, there is a general learning rule, called the Closure algorithm (Helmbold et al., 1990; Auer and Ortner, 2007). For given data S, the algorithm outputs $\widehat{h} = \mathbb{1}[\{x|\forall h \in \mathrm{VS}_{\mathcal{H},S}, h(x) = 1\}]$. Since $\widehat{h}(x) = 1$ implies $h^*(x) = 1$, and since adding clean-label poison points will only increase the region being predicted as positive, we have $\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathrm{Closure}) = \mathrm{err}(\mathrm{Closure}(S_{\mathrm{trn}}))$. Then by Theorem 5 of Hanneke (2016a), for any intersection-closed hypothesis class $\mathcal{H}$ with VC dimension d, the Closure algorithm can robustly learn $\mathcal{H}$ using m samples, where

$$m = \frac{1}{\varepsilon}(21d + 16\ln(3/\delta)).$$

Unions of intervals are (ε, δ)-robust learnable. Let $\mathcal{H}_k = \cup_{k' \le k}\{\mathbb{1}[\cup_{i=1}^{k'}(a_i, b_i)]|0 \le a_i < b_i \le 1, \forall i \in [k']\}$ denote the union of at most k positive open intervals for any $k \ge 1$. This hypothesis class is a generalization of Example 1. There is a robust learning rule: output $\mathbb{1}[\emptyset]$ if there is no positive sample and otherwise, output the consistent union of minimum number of closed intervals, each of which has minimum length. More specifically, given input (poisoned) data $S = \{(x_1, y_1), \ldots, (x_{m'}, y_{m'})\}$ with $x_1 \le x_2 \le \ldots \le x_{m'}$ w.l.o.g., for notation simplicity, let $y_0 = y_{m'+1} = 0$. Then the algorithm $\mathcal{A}$ outputs $\widehat{h} = \mathbb{1}[X]$ where $X = \cup\{[x_i, x_j]|\forall i \le l \le j \in [m'], y_{i-1} = y_{j+1} = 0, y_i = y_l = y_j = 1\}$. The algorithm $\mathcal{A}$ can robustly learn union of intervals $\mathcal{H}_k$ using m samples, where

$$m = O\left(\frac{1}{\varepsilon}(k\log(1/\varepsilon) + \log(1/\delta))\right).$$

The detailed proof can be found in Appendix A.2.

Hypothesis classes with finite star number are (ε, δ)-robust learnable. The star number, proposed by Hanneke and Yang (2015), can measure the disagreement region of the version space.

Definition 4 (star number) *The star number $\mathfrak{s}$ is the largest integer s such that there exist distinct points $x_1, \ldots, x_s \in \mathcal{X}$ and classifiers $h_0, \ldots, h_s$ with the property that $\forall i \in [s]$, $\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \ldots, x_s\} = \{x_i\}$; if no such largest integer exists, define $\mathfrak{s} = \infty$.*

By Theorem 10 of Hanneke (2016a), for any $\mathcal{H}$ with star number $\mathfrak{s}$, with probability at least $1 - \delta$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$, $\mathcal{P}(\mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}})) \le \varepsilon$ where

$$m = \frac{1}{\varepsilon}(21\mathfrak{s} + 16\ln(3/\delta)).$$

As aforementioned, $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathrm{ERM}, \mathrm{Adv}) \subseteq \mathrm{DIS}(\mathrm{VS}_{\mathcal{H},S_{\mathrm{trn}}})$ for any clean-label attacker Adv and thus any ERM can robustly learn $\mathcal{H}$ using m samples.

Hypothesis classes with infinite hollow star number are not consistently properly (ε, δ)-robust learnable. The hollow star number, proposed by Bousquet et al. (2020), characterizes proper learnability. For any set $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$, $S^i = \{(x_1, y_1'), \ldots, (x_k, y_k')\}$ is said to be a neighbor of S if $y_i' \ne y_i$ and $y_j' = y_j$ for all $j \ne i$, for any $i \in [k]$.

Definition 5 (hollow star number) *The hollow star number k_o is the largest integer k such that there is a set $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ (called the hollow star set) which is not realizable by $\mathcal{H}$, however every set S' which is a neighbor of S is realizable by $\mathcal{H}$. If no such largest k exists, define $k_o = \infty$.*

For any hypothesis class $\mathcal{H}$ with hollow star number k_o, for any consistent proper learner $\mathcal{A}$, there exists a target function h^* and a data distribution $\mathcal{D}$ such that if $m \leq \lfloor (k_o - 1)/2 \rfloor$, then the expected attackable rate

$$\mathbb{E}_{S_{\text{trn}} \sim \mathcal{D}^m} \left[\text{atk}(h^*, S_{\text{trn}}, \mathcal{A}) \right] \geq 1/4 \,,$$

which implies $\mathbb{P}_{S_{\text{trn}}} \left(\text{atk}(h^*, S_{\text{trn}}, \mathcal{A}) > 1/8 \right) \geq 1/7$ by Markov's inequality. The construction of the target function, the data distribution and the attacker is as described below. Consider a hollow star set S as above, with size k. By definition, there exists a set of hypotheses $\{h_1, \ldots, h_k\} \subseteq \mathcal{H}$ such that each neighbor S^i is realized by h_i for any $i \in [k]$. Consider the target function being h_{i^*} where i^* is drawn uniformly at random from $[k]$ and the marginal data distribution is a uniform distribution over $\{x_i | i \in [k] \setminus \{i^*\}\}$. For any $\lfloor (k-1)/2 \rfloor$ i.i.d. samples from the data distribution, there are at least $k - \lfloor (k-1)/2 \rfloor$ instances in S not sampled. To attack an unseen instance x_i, the attacker adds all examples in S except x_i, x_{i^*}. Then any algorithm cannot tell whether h_{i^*} or h_i is the true target and any consistent proper learner will misclassify $\{x_i, x_{i^*}\}$ with probability $1/2$.

For hypothesis classes with $k_o = \infty$, there is a sequence of hollow star sets with increasing sizes $\{k_i\}_{i=1}^{\infty}$. Therefore, any hypothesis class with $k_o = \infty$ is not consistently properly robust learnable. The detailed proof is included in Appendix A.3.

3.2. All robust learners are optimal PAC learners

There is an interesting connection between algorithms robust to clean-label poisoning attacks and the classic literature on the sample complexity of PAC learning. Specifically, we can show that *any* learning algorithm that is robust to clean-label poisoning attacks necessarily obtains the optimal dependence on ε in its PAC sample complexity: that is, $O(1/\varepsilon)$. This is a very strong property, and not many such learning algorithms are known, as most learning algorithms have at least an extra $\log(1/\varepsilon)$ factor in their sample complexity (see e.g., Haussler, Littlestone, and Warmuth, 1994; Auer and Ortner, 2007; Hanneke, 2009, 2016b,a; Darnstädt, 2015; Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020). Thus, this property can be very informative regarding what types of learning algorithms one should consider when attempting to achieve robustness to clean-label poisoning attacks. This claim is formalized in the following result. Its proof is presented in Appendix A.4.

Theorem 1 *Fix any hypothesis class $\mathcal{H}$. Let $\mathcal{A}$ be a deterministic learning algorithm that always outputs a deterministic hypothesis. Suppose there exists a non-negative sequence $R(m) \to 0$ such that, $\forall m \in \mathbb{N}$, for every target function $h^* \in \mathcal{H}$ and every distribution $\mathcal{D}$ over D_{h^*}, for $S_{\text{trn}} \sim \mathcal{D}^m$, with probability at least $1/2$, $\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \mathcal{A}) \leq R(m)$. Then there exists an ($R$-dependent) finite constant c_R such that, for every $\delta \in (0,1)$, $m \in \mathbb{N}$, $h^* \in \mathcal{H}$, and every distribution $\mathcal{D}$ over D_{h^*}, for $S_{\text{trn}} \sim \mathcal{D}^m$, with probability at least $1 - \delta$, $\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \mathcal{A}) \leq \frac{c_R}{m} \log \frac{2}{\delta}$.*

An immediate implication of this result (together with Markov's inequality) is that any deterministic $\mathcal{A}$ outputting deterministic predictors, if $\mathbb{E}_{S_{\text{trn}} \sim \mathcal{D}^m}[\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \mathcal{A})] \leq R(m)/2 \to 0$ for all $h^* \in \mathcal{H}$ and $\mathcal{D}$ on D_{h^*}, then for any $h^* \in \mathcal{H}$, $\mathcal{D}$ on D_{h^*}, $\delta \in (0,1)$, $\text{err}_{\mathcal{D}}(\mathcal{A}(S_{\text{trn}})) \leq \frac{c_R}{m} \log \frac{2}{\delta}$ with probability at least $1 - \delta$. As mentioned, this is a strong requirement of the learning algorithm

$\mathcal{A}$; for instance, for many classes $\mathcal{H}$, many ERM learning rules would have an extra $\log(m)$ factor (Hanneke, 2016a). This also establishes a further connection to the hollow star number, which in some cases strengthens the result mentioned above (and detailed in Appendix A). Specifically, Bousquet, Hanneke, Moran, and Zhivotovskiy (2020) have shown that when $k_o = \infty$, for any fixed δ sufficiently small, any proper learning algorithm has, for some infinite sequence of m values, that $\exists h^* \in \mathcal{H}$ and $\mathcal{D}$ on D_{h^*} for which, with probability greater than δ, $\mathrm{err}_{\mathcal{D}}(\mathcal{A}(S_{\mathrm{trn}})) \geq \frac{c\log(m)}{m}$ for a numerical constant c. Together with Theorem 1, this implies that for such classes, any deterministic proper learning algorithm cannot have a sequence $R(m) \to 0$ as in the above theorem. Formally, using the fact that $\mathrm{atk}_{\mathcal{D}}(h^*, S, \mathcal{A})$ is non-increasing in S (see the proof of Theorem 1), we arrive at the following corollary, which removes the "consistency" requirement from the result for classes with $k_o = \infty$ stated above, but adds a requirement of being deterministic.

Corollary 1 *If $k_o = \infty$, then for any deterministic* proper *learning algorithm $\mathcal{A}$ that always outputs a deterministic hypothesis, there exists a constant $c > 0$ such that, for every $m \in \mathbb{N}$, $\exists h^* \in \mathcal{H}$ and distribution $\mathcal{D}$ on D_{h^*} such that $\mathbb{E}_{S_{\mathrm{trn}} \sim \mathcal{D}^m}[\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A})] > c$.*

4. Linear hypothesis class

In this section, we first show that n-dimensional linear classifiers $\mathcal{H} = \{\mathbb{1}[\langle w, x\rangle + b \geq 0] | (w, b) \in \mathbb{R}^{n+1}\}$ with $n \geq 2$ are not robust learnable. Then we study a restrictive case where the support of the data distribution has a positive margin to the boundary. We introduce two robust learners and prove a robust learning sample complexity lower bound. In addition, we also show the vulnerability of SVM.

4.1. Linear hypothesis class is not robust learnable

In this section, we show that the class of linear hypotheses is not robust learnable.

Theorem 2 *For $n \geq 2$, the class of linear hypotheses is not robust learnable.*

Proof sketch We present the proof idea in the case of $n = 3$ here and for simplicity, we allow the decision boundary to be either positive or negative. The construction details of limiting the boundary to be positive and the construction for $n = 2$ are deferred to Appendix B.

Consider the case where $\mathcal{X} = \Gamma^3(\mathbf{0}, 1)$ is the sphere of the 3-dimensional unit ball centered at the origin and the target function is uniformly randomly chosen from all linear classifiers with the decision boundary at distance $1/2$ from the origin, and the boundary labeled different from $\mathbf{0}$, i.e., $h^* \sim \mathrm{Unif}(\mathcal{H}^*)$, where $\mathcal{H}^* = \{\mathbb{1}[\langle w, x\rangle - \frac{1}{2} \geq 0] | \, \|w\| = 1\} \cup \{1 - \mathbb{1}[\langle w, x\rangle - \frac{1}{2} \geq 0] | \, \|w\| = 1\}$. W.l.o.g., suppose $h^* = \mathbb{1}[\langle w^*, x\rangle - \frac{1}{2} \geq 0]$. The data distribution is the uniform distribution over the intersection of the decision boundary and the sphere, i.e., $\mathcal{D}_{\mathcal{X}} = \mathrm{Unif}(C_{w^*})$, where $C_{w^*} = \{x | \, \langle w^*, x\rangle - \frac{1}{2} = 0\} \cap \Gamma^3(\mathbf{0}, 1)$. Then all training data come from the circle C_{w^*} and are labeled positive.

Given training data $S_{\mathrm{trn}} \sim \mathcal{D}^m$ and a test point $x_0 \in C_{w^*}$ (not in $S_{\mathrm{trn},\mathcal{X}}$), the attacker constructs a fake circle $C_{w'}$ tangent to C_{w^*} at point x_0, i.e., $C_{w'} = \{x | \, \langle w', x\rangle - \frac{1}{2} = 0\} \cap \Gamma^3(\mathbf{0}, 1)$ where $w' = 2\langle x_0, w^*\rangle x_0 - w^*$. Then the attacker adds m i.i.d. samples from the uniform distribution over $C_{w'}$ and labels them negative. Any algorithm cannot tell which circle is the true circle and which one of $\{\mathbb{1}[\langle w^*, x\rangle - \frac{1}{2} \geq 0], 1 - \mathbb{1}[\langle w', x\rangle - \frac{1}{2} \geq 0]\}$ is the true target. Hence, any algorithm will misclassify x_0 with probability $1/2$.

Algorithm 1 Robust algorithm for 2-dimensional linear classifiers

input: data S
initialize $l \leftarrow 0$, $h \leftarrow 2\pi$ and $\beta \leftarrow \frac{l+h}{2}$
if $\exists b \in [-2,2]$ s.t. $(0,b)$ is consistent **then** output $(0,b)$
while $\nexists b \in [-2,2]$ s.t. (β,b) is consistent with S **do**
  if $\exists \beta \in (l, \frac{l+h}{2})$ s.t. $\exists b \in [-2,2]$, (β,b) is consistent with S **then** let $h \leftarrow \frac{l+h}{2}$, $\beta \leftarrow \frac{l+h}{2}$
  else let $l \leftarrow \frac{l+h}{2}$, $\beta \leftarrow \frac{l+h}{2}$
end while
return (β,b) with any consistent b

4.2. Linear hypothesis class is robust learnable under distribution with margin

In this section, we discuss linear classifiers in the case where the distribution has a positive margin. Specifically, considering the instance space $\mathcal{X} \subseteq \mathcal{B}^n(\mathbf{0},1)$, we limit the data distribution $\mathcal{D}$ to satisfy that $\forall (x,y) \in \mathrm{supp}(\mathcal{D}), (2y-1)(\langle w^*, x\rangle + b^*) \geq \gamma \left\|w^*\right\|/2$ for some margin $\gamma \in (0,2]$ and target function $h^*(x) = \mathbb{1}[\langle w^*, x\rangle + b^* \geq 0]$.

4.2.1. A LEARNER FOR ARBITRARY $n > 0$

The learner $\mathcal{A}$ fixes a $\gamma/2$-covering V of $\mathcal{X}$, i.e., $\forall x \in \mathcal{X}, \exists v \in V, x \in \mathcal{B}(v,\gamma/2)$, where $|V| \leq (2/\gamma)^n$. It is easy to check that such a V always exists. Then given input data S, the learner outputs a classifier: for $x \in \mathcal{B}(v,\gamma/2)$, if $\exists (x',y') \in S$ s.t. $x' \in \mathcal{B}(v,\gamma/2)$, predicting $h(x) = y'$; otherwise, predicting randomly. Note that Adv does not necessarily need to be restricted to such margin.

Theorem 3 *The algorithm can robustly learn linear classifiers with margin γ using m samples where*

$$m = \frac{(2/\gamma)^n}{\varepsilon}\left(n\ln\frac{2}{\gamma} + \ln\frac{1}{\delta}\right).$$

Proof First, for every $v \in V$, at least one of $y \in \{0,1\}$ has $\mathcal{D}(x \in \mathcal{B}(v,\gamma/2) : h^*(x) = y) = 0$. Then with probability at least $1 - |V|(1-\varepsilon/|V|)^m$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$, for every ball $\mathcal{B}(v,\gamma/2)$ with probability mass at least $\varepsilon/|V|$, there exists $(x',y') \in S_{\mathrm{trn}}$ such that $x' \in \mathcal{B}(v,\gamma/2)$. Let $m = |V|\ln(|V|/\delta)/\varepsilon$, we have with probability at least $1-\delta$, $\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \varepsilon$. ■

4.2.2. A LEARNER FOR $n = 2$

In the 2-dimensional case, the hypothesis class can be represented as $\mathcal{H} = \{h_{\beta,b} | \beta \in [0,2\pi), b \in [-2,2]\}$ where $h_{\beta,b} = \mathbb{1}[(\cos\beta, \sin\beta)\cdot x + b \geq 0]$. When there is no ambiguity, we use (β,b) to represent $h_{\beta,b}$. The target is $h^* = h_{\beta^*,b^*}$. Then we propose a robust algorithm based on binary-search for the target direction β^* as shown in Algorithm 1.

Theorem 4 *For any data distribution $\mathcal{D}$, let $f(\varepsilon'') = \max\{s \geq 0 | \mathcal{P}(\{x | (\cos\beta^*, \sin\beta^*)\cdot x + b^* \in [-s,0]\}) \leq \varepsilon'', \mathcal{P}(\{x | (\cos\beta^*, \sin\beta^*)\cdot x + b^* \in [0,s]\}) \leq \varepsilon''\}$ for $\varepsilon'' \in [0,1]$ denote the maximum distance between the boundary and two parallel lines (on positive side and negative side respectively) such that the probability between the boundary and either line is no greater than ε''. With probability at least $1-\delta$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$, Algorithm 1 achieves*

$$\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \log(\frac{32}{f(\varepsilon'') \wedge 2})2\varepsilon' + 2\varepsilon'',$$

for any $\varepsilon'' \in [0,1]$ *using* m *samples where*

$$m = \frac{24}{\varepsilon'} \log \frac{13}{\varepsilon'} + \frac{4}{\varepsilon'} \log \frac{2}{\delta}.$$

Proof sketch First, by uniform convergence bound in PAC learning (Blumer et al., 1989), when $m \geq \frac{24}{\varepsilon'} \log \frac{13}{\varepsilon'} + \frac{4}{\varepsilon'} \log \frac{2}{\delta}$, every linear classifier consistent with S_{trn} has error no greater than ε'. For any fixed β, the probability mass of union of error region of all (β, b) consistent with the training data is bounded by $2\varepsilon'$. Then given a target (β^*, b^*), the binary-search path of β is unique and adding clean-label poison examples will only change the depth of search. When $h - l < \arctan(f(\varepsilon'')/2)$, the attackable rate caused by deeper search is at most $2\varepsilon''$. Combining these results together proves the theorem. The formal proof of Theorem 4 is included in Appendix C.

Theorem 5 *For any* $\gamma \in (0, 2]$*, Algorithm 1 can* (ε, δ)*-robustly learn 2-dimensional linear classifiers with margin* γ *using* m *samples where*

$$m = \frac{48 \log(64/\gamma)}{\varepsilon} \log \frac{26 \log(64/\gamma)}{\varepsilon} + \frac{8 \log(64/\gamma)}{\varepsilon} \log \frac{2}{\delta}.$$

Theorem 5 is the immediate result of Theorem 4 as $f(0) = \gamma/2$.

4.2.3. SVM REQUIRES $e^{\Omega(n)}/\varepsilon$ SAMPLES AGAINST ONE-POINT ATTACKER

SVM is a well-known optimal PAC learner for linear hypothesis class (Bousquet et al., 2020). In this section, we show that even in the case where $\gamma \geq 1/8$ and the attacker is limited to add at most one poison point, SVM requires $e^{\Omega(n)}/\varepsilon$ samples to achieve ε attackable rate.

Theorem 6 *For* n*-dim linear hypothesis class, for any* $\varepsilon < 1/16$*, there exists a target* $h^* \in \mathcal{H}$ *and a distribution* $\mathcal{D}$ *over* D_{h^*} *with margin* $\gamma = 1/8$ *such that* $\mathbb{E}_{S_{\text{trn}} \sim \mathcal{D}^m} [\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \text{SVM})] > \varepsilon$ *when the sample size* $m < \frac{e^{n/128}}{768\varepsilon} \vee \frac{1}{8\varepsilon}$.

Proof sketch Consider the case where $\mathcal{X} = \{x \in \mathbb{R}^3 | \, \|x\| = 1, \langle x, e_1 \rangle \geq 0\} \cup \{-e_1\}$ is the union of a half sphere and a point $-e_1$. The target function is $h^* = \mathbb{1}[\langle w^*, x \rangle \geq -\gamma/2]$ with $w^* = e_1$ and margin $\gamma = 1/8$. Note that h^* labels all points on the half sphere positive and $-e_1$ negative. Then we define the data distribution $\mathcal{D}_{\mathcal{X}}$ by putting probability mass $1 - 8\varepsilon$ on $-e_1$ and putting probability mass 8ε uniformly on the half sphere.

Then we draw training set $S_{\text{trn}} \sim \mathcal{D}^m$ and a test point $(x_0, y_0) \sim \mathcal{D}$. Condition on that x_0 is on the half sphere, with high probability, $\langle x_0, w^* \rangle \leq 1/8$. Then we define two base vectors $v_1 = w^*$ and $v_2 = \frac{x_0 - \langle x_0, w_* \rangle w^*}{\|x_0 - \langle x_0, w_* \rangle w^*\|}$ in the 2-dimensional space defined by w^* and x_0. With high probability over the choice of S_{trn}, for all positive training examples x on the half sphere, we have $\langle x, v_1 \rangle \leq 1/8$ and $\langle x, v_2 \rangle \leq 1/8$. Then the attacker injects a poison point at $-\gamma v_1 + \sqrt{1 - \gamma^2} v_2$, which is closer to x_0 than all the positive samples in S_{trn}. Since the poison point is classified as negative by the target function, SVM will misclassify x_0 as negative. The detailed proof can be found in Appendix D.

4.2.4. LOWER BOUND

Here we show that robust learning of linear hypothesis class under distribution with margin $\gamma > 0$ requires sample complexity $e^{\Omega(n)}/\varepsilon$.

Theorem 7 *For n-dimensional linear hypothesis class with $n > 256$, for any $\varepsilon \leq 1/16$ and for any algorithm $\mathcal{A}$, there exists a target function $h^* \in \mathcal{H}$ and a distribution $\mathcal{D}$ over D_{h^*} with margin $\gamma = 1/8$ such that $\mathbb{E}_{S_{\mathrm{trn}} \sim \mathcal{D}^m}[\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A})] > \varepsilon$ when the sample size $m \leq \frac{e^{\frac{n-1}{128}}}{192\varepsilon}$. For convenience, here we relax the instance space by allowing $\mathcal{X} \subseteq B^n(\mathbf{0}, 9/8)$.*

The construction of the target function and the data distribution is similar to that in the proof of Theorem 6. To attack a test instance x_0, the attacker adds the reflection points of all training points through the hyperplane defined by x_0 and w^* such that any algorithm will misclassify x_0 with probability $1/2$. The detailed proof is included in Appendix E.

5. Results for finite-point attackers

In this section, instead of considering the case where the attacker can add a set of poison examples of arbitrary size, we study a restrictive case where the attacker is allowed to add at most t poison examples for some $t < \infty$, i.e., $|\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x_0)| \leq t$ for any $h^*, S_{\mathrm{trn}}, x_0$. Following Definition 1 and 2, we define t-point clean-label attackable rate and (t, ε, δ)-robust learnability as follows.

Definition 6 (t-point clean-label attackable rate) *For a target function h^*, a training data set S_{trn} and a (possibly randomized) algorithm $\mathcal{A}$, for any distribution $\mathcal{D}$ over D_{h^*}, the t-point clean-label attackable rate is*

$$\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\mathrm{trn}}, \mathcal{A}) \triangleq \sup_{\mathrm{Adv}} \mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \text{ s.t. } |\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x)| \leq t, \forall x \in \mathcal{X}\,.$$

Definition 7 ((t, ε, δ)-robust learnability) *A hypothesis class $\mathcal{H}$ is (t, ε, δ)-robust learnable if there exists a learning algorithm $\mathcal{A}$ such that $\forall \varepsilon, \delta \in (0, 1)$, $\exists m(t, \varepsilon, \delta) \in \mathbb{N}$ such that $\forall h^* \in \mathcal{H}, \forall \mathcal{D}$ over D_{h^*}, with probability at least $1 - \delta$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$,*

$$\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \varepsilon\,.$$

5.1. Algorithms robust to t-point attacker

Robustness to a small number of poison examples has been studied by Ma et al. (2019); Levine and Feizi (2020). Ma et al. (2019) show that differentially-private learners are naturally resistant to data poisoning when the attacker can only inject a small number of poison examples. Levine and Feizi (2020) propose an algorithm called Deep Partition Aggregation (DPA), which partitions the training set into multiple sets by a deterministic hash function, trains base classifiers over each partition and then returns the majority vote of base classifiers. They show that for any instance x, the prediction on x is unchanged if the number of votes of the output exceeds half of the number of the total votes by t. But the attackable rate of DPA is not guaranteed. Here we propose several algorithms similar to DPA but with guarantees on the attackable rate. In Algorithm 2, we provide a protocol converting any given ERM learner $\mathcal{L}$ to a learner with small t-point clean-label attackable rate.

Theorem 8 *For any hypothesis class $\mathcal{H}$ with VC dimension d with any proper ERM learner $\mathcal{L}$, Algorithm 2 can (t, ε, δ)-robustly learn $\mathcal{H}$ using m samples where*

$$m = O\left(\frac{dt}{\varepsilon}\log\frac{dt}{\varepsilon} + \frac{d}{\varepsilon}\log\frac{1}{\delta}\right)\,.$$

Algorithm 2 A robust protocol for t-point attacker

1: **input**: A proper ERM learner $\mathcal{L}$, data S
2: divide S into $10t+1$ blocks $\{S^{(1)}, S^{(2)}, \ldots, S^{(10t+1)}\}$ with size $\left\lfloor \frac{|S|}{10t+1} \right\rfloor$ randomly without replacement (throw away the remaining $|S| - (10t+1)\left\lfloor \frac{|S|}{10t+1} \right\rfloor$ points)
3: **return** $\mathrm{Major}(\mathcal{H}')$ where $\mathcal{H}' = \{\mathcal{L}(S^{(i)}) | i \in [10t+1]\}$

Algorithm 3 A proper robust learner for t-point attacker given projection number k_p

1: **input**: A proper ERM learner $\mathcal{L}$, data S
2: Divide the data S into $10k_pt+1$ sets $\{S^{(1)}, S^{(2)}, \ldots, S^{(10k_pt+1)}\}$ with size $\left\lfloor \frac{|S|}{10k_pt+1} \right\rfloor$ randomly without replacement (throw away the remaining $|S| - (10k_pt+1)\left\lfloor \frac{|S|}{10k_pt+1} \right\rfloor$ points)
3: **return** $\widehat{h} = \mathrm{Proj}_{\mathcal{H}}(\mathcal{H}')$, where $\mathcal{H}' = \{h_i = \mathcal{L}(S^{(i)}) | i \in [10k_pt+1]\}$

Proof sketch For every misclassified point $x_0 \in \mathcal{X}$, there are at least $5t+1$ classifiers among $\{\mathcal{L}(S^{(i)})\}_{i=1}^{10t+1}$ misclassifying x_0. Since there are at most t blocks containing poison data, there are at least $4t+1$ non-contaminated classifiers (output by blocks without poison data) misclassifying x_0. Then t-point clean-label attackable rate is bounded by bounding the error of one non-contaminated classifier. The detailed proof is provided in Appendix F.

As we can see, Algorithm 2 is improper even if $\mathcal{L}$ is proper. Inspired by the projection number and the projection operator defined by Bousquet et al. (2020), we propose a proper robust learner in Algorithm 3. First, let us introduce the definitions of the projection number and the projection operator as follows. For a finite (multiset) $\mathcal{H}' \subseteq \mathcal{H}$, for $l \geq 2$, define the set $\mathcal{X}_{\mathcal{H}',l} \subseteq \mathcal{X}$ of all the points x on which less than $\frac{1}{l}$-fraction of all classifiers in $\mathcal{H}'$ disagree with the majority. That is,

$$\mathcal{X}_{\mathcal{H}',l} = \left\{ x \in \mathcal{X} : \sum_{h \in \mathcal{H}'} \mathbb{1}[h(x) \neq \mathrm{Major}(\mathcal{H}', x)] < \frac{|\mathcal{H}'|}{l} \right\}.$$

Definition 8 (projection number and projection operator) *The projection number of $\mathcal{H}$, denoted by k_p, is the smallest integer $k \geq 2$ such that, for any finite multiset $\mathcal{H}' \subseteq \mathcal{H}$ there exists $h \in \mathcal{H}$ that agrees with $\mathrm{Major}(\mathcal{H}')$ on the entire set $\mathcal{X}_{\mathcal{H}',k}$. If no such integer k exists, define $k_p = \infty$. If $k_p < \infty$, the projection operator $\mathrm{Proj}_{\mathcal{H}} : \mathcal{H}' \mapsto \mathcal{H}$ is a deterministic map from $\mathcal{H}'$ to $\mathcal{H}$ such that $\mathrm{Proj}_{\mathcal{H}}(\mathcal{H}', x) = \mathrm{Major}(\mathcal{H}', x), \forall x \in \mathcal{X}_{\mathcal{H}',k_p}$.*

Theorem 9 *For any hypothesis class $\mathcal{H}$ with VC dimension d and projection number k_p, with any proper ERM learner $\mathcal{L}$, Algorithm 3 can (t, ε, δ)-robustly learn $\mathcal{H}$ using m samples where*

$$m = O\left(\frac{k_p^2 dt}{\varepsilon} \log \frac{k_p dt}{\varepsilon} + \frac{k_p d}{\varepsilon} \ln \frac{1}{\delta} \right).$$

The proof adopts the same idea as the proof of Theorem 8 and is included in Appendix G. For the hypothesis class with infinite projection number, we can obtain a proper learner in a similar way: randomly selecting $\lfloor \varepsilon |S|/3t \rfloor$ samples with replacement from input data set S and run ERM over the selected data. We show that this algorithm can (t, ε, δ)-robustly learn $\mathcal{H}$ using $O(\frac{dt}{\varepsilon^2} \log \frac{d}{\varepsilon} + \frac{d}{\varepsilon} \log \frac{1}{\delta})$ samples. The details of the algorithm and the analysis can be found in Appendix G.

5.2. Lower bound

Theorem 10 *For any $d \geq 1$ and $\varepsilon \leq \frac{3}{8}$, there exists a hypothesis class $\mathcal{H}$ with VC dimension $5d$ such that for any algorithm $\mathcal{A}$, there exists a target function $h^* \in \mathcal{H}$ and a data distribution $\mathcal{D}$ on D_{h^*}, such that $\mathbb{E}_{S_{\mathrm{trn}} \sim \mathcal{D}^m}\left[\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\mathrm{trn}}, \mathcal{A})\right] > \varepsilon$ when the sample size $m < \frac{3td}{64\varepsilon}$.*

Proof sketch Consider d disjoint spheres in $\mathbb{R}^3$ and the target function is chosen by randomly selecting a circle on each sphere. Then label each circle differently from the rest of the sphere the circle lies on. Specifically, we flip d independent fair coins, one for each circle to decide whether the circle is labeled positive or negative. The data distribution puts probability mass $\frac{t}{8m}$ uniformly on each circle and $1 - \frac{td}{8m}$ probability mass on an irrelevant point (not on any of the spheres). Then we can show that with constant probability, every unseen point on each circle can be attacked by an attacker similar to the one in the proof sketch of Theorem 2. The detailed proof is included in Appendix H.

Remark 1 *Actually, our algorithms above even work for t-point* unclean-*label attackers (where the poison data are not necessarily labeled by the target function) as well, which indicates that cleanness of poison examples does not make the problem fundamentally easier in the worst-case over classes of a given VC dimension, in the t-point attack case (although it can potentially make a difference for particular algorithms or particular classes $\mathcal{H}$).*

6. Discussion and future directions

In this paper, we show the impossibility of robust learning in the presence of clean-label attacks for some hypothesis classes with bounded VC dimension, e.g., the class of linear separators, and the robust learnability of some hypothesis classes characterized by known complexity measures, e.g., finite star number. There are several interesting open questions.

- The first question is what are necessary and sufficient conditions for (ε, δ)-robust learnability. Finite star number is a sufficient but not necessary condition. Here is an example where the instance space is $\mathcal{X} = \mathbb{N}$ and the hypothesis class is $\mathcal{H} = \{\mathbb{1}[i] | i \in \mathbb{N}\} \cup \{0\}$. The star number of $\mathcal{H}$ is $\mathfrak{s} = \infty$, but $\mathcal{H}$ is robust learnable since $\mathrm{VCdim}(\mathcal{H}) = 1$. One intriguing possible complexity measure is the largest number k such that there is a set of distinct points $S = \{x_1, \ldots, x_k\} \in \mathcal{X}^k$ and classifiers $\{h_0, \ldots, h_k\}$, where for any $i \in [k]$, there exists an involutory function $f_i : \mathcal{X} \mapsto \mathcal{X}$ (i.e., $f_i(f_i(x)) = x, \forall x \in \mathcal{X}$) such that $f_i(x_i) = x_i$ and $\mathrm{DIS}(\{h_0, h_i\}) \cap \{S \cup f_i(S)\} = \{x_i\}$.
- For proper robust learning, we prove that any hypothesis class with infinite hollow star number $k_o = \infty$ is neither consistently properly (ε, δ)-robust learnable nor deterministically properly (ε, δ)-robust learnable. Compared with the fact that $k_o = \infty$ only brings an extra $\Omega(\log(1/\varepsilon))$ in the optimal PAC sample complexity (Bousquet et al., 2020), we see that the hollow star number has a dramatically larger impact on proper robust learnability. On the other hand, finite hollow star number does not suffice for robust learnability (e.g., Bousquet et al., 2020, show linear classifiers on $\mathbb{R}^n$ have $k_o = n + 2$), and it is unclear what is the necessary and sufficient condition for proper robust learnability.
- For linear classifiers with margin $\gamma > 0$, the lower bound of the sample complexity presented in Section 4 ignores the dependence on γ. For the two learners introduced in Section 4, the one using the covering set has sample complexity of $O(n(2/\gamma)^n \log(1/\gamma))$ and the other one

designed for the 2-dimension has sample complexity of $O(\log(1/\gamma)\log\log(1/\gamma))$. There is a huge gap between the lower bound and the upper bound and thus far, the optimal dependence on γ remains unclear.

- For finite-point attacks, we construct a hypothesis class such that the t-point clean-label attackable rate is $\Omega(\frac{t}{m})$ in the proof of Theorem 10 and Algorithm 2 achieves $O(\frac{t\log(m)}{m})$ attackable rate. It is unclear to us for what kind of hypothesis class, there is an algorithm able to achieve $o(\frac{t}{m})$ attackable rate. At the same time, we are curious about its connection to (ε,δ)-robust learnablility. Notice that in all the proofs of the negative results in this paper, the attacker we construct never injects more than m poison examples. This triggers the following suspicion: are infinite-point attackers strictly more powerful than m-point attackers? Specifically, we have the following conjecture.

  Conjecture 1 (infinite to finite) *For any hypothesis class $\mathcal{H}$, for every target function $h^* \in \mathcal{H}$, data distribution $\mathcal{D}$ over D_{h^*}, there exist a pair of constants $c, c' > 0$ such that for any $m > 0$, any training data $S_{\mathrm{trn}} \in D_{h^*}^m$ and any algorithm $\mathcal{A}$, $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \geq c$ iff $\mathrm{atk}_{\mathcal{D}}(m, h^*, S_{\mathrm{trn}}, \mathcal{A}) \geq c'$.*

  Assuming that this conjecture holds, hence for any hypothesis class $\mathcal{H}$, if there exists an algorithm $\mathcal{A}$ able to (t,ε,δ)-robustly learn $\mathcal{H}$ with attackable rate $o(\frac{t}{m})$, then $\mathcal{H}$ is (m,ε,δ)-robust learnable and thus, (ε,δ)-robust learnable.
- Another open question is whether abstention helps. Considering the case where the algorithm is allowed to abstain on ε-fraction of inputs if the algorithm detects abnormality. That is to say, the algorithm outputs a selective classifier $(\widehat{h}, \mathrm{CR}(S_{\mathrm{trn}}))$, where the prediction hypothesis $\widehat{h}$ is a map from $\mathcal{X}$ to $\mathcal{Y}$ and $\mathrm{CR}(S_{\mathrm{trn}}) \subseteq \mathcal{X}$ is the confidence region of the prediction. The algorithm predicts $\mathcal{A}(S_{\mathrm{trn}}, x) = \widehat{h}(x)$ if $x \in \mathrm{CR}(S_{\mathrm{trn}})$ and $\mathcal{A}(S_{\mathrm{trn}}, x) = \perp$ if $x \notin \mathrm{CR}(S_{\mathrm{trn}})$, where $\perp$ means the algorithm predicts "I don't know". Then for any deterministic algorithm, we say a test instance $x \in \mathcal{X}$ is attackable if there is a clean-label attacker such that x is predicted incorrectly as well as x is in the confidence region, i.e.,

$$\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x) \qquad \& \qquad x \in \mathrm{CR}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x))\,.$$

  We define the event $\mathcal{E}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}, x, y) = \{\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x) \cap x \in \mathrm{CR}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x))\}$ and then define the selective attackable rate as

$$\sup_{\mathrm{Adv}} \mathbb{E}_{(x,y)\sim\mathcal{D},\mathcal{A}}\left[\mathbb{1}[\mathcal{E}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}, x, y)]\right].$$

  We are curious about the sample complexity required to achieve ε selective attackable rate while keeping the probability mass of the confidence region $\mathbb{P}_{(x,y)\sim\mathcal{D}}\left(x \in \mathrm{CR}(S)\right) \leq \varepsilon$ for any input $S \supseteq S_{\mathrm{trn}}$.

Acknowledgments

This work was supported in part by the National Science Foundation under grant CCF-1815011 and by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. Jian Qian acknowledges support of the ONR through grant # N00014-20-1-2336. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

References

Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007.

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31: 1–58, 1997.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25, 2006.

Shai Ben-David. 2 notes on classes with Vapnik-Chervonenkis dimension 1. *arXiv preprint arXiv:1507.05307*, 2015.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474, 2012.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the 33rd Annual Conference on Learning Theory*, 2020.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840. PMLR, 2019.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

Malte Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015.

Omid Etesami, Saeed Mahloujifar, and Mohammad Mahmoody. Computational concentration of measure: Optimal bounds, reductions, and more. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 345–363. SIAM, 2020.

Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Steve Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009.

Steve Hanneke. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016a.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016b.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

David Haussler, Nick Littlestone, and Manfred Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

D. Helmbold, R. Sloan, and M. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5(2):165–196, 1990.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*, 2020.

Yuzhe Ma, Xiaojin Zhu Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *International Joint Conference on Artificial Intelligence*, 2019.

Saeed Mahloujifar and Mohammad Mahmoody. Blockwise p-tampering attacks on cryptographic primitives, extractors, and learners. In *Theory of Cryptography Conference*, pages 245–279. Springer, 2017.

Saeed Mahloujifar and Mohammad Mahmoody. Can adversarially robust learning leverage computational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR, 2019.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. Learning under p-tampering attacks. In *Algorithmic Learning Theory*, pages 572–596. PMLR, 2018.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4536–4543, 2019a.

Saeed Mahloujifar, Mohammad Mahmoody, and Ameer Mohammed. Universal multi-party poisoning attacks. In *International Conference on Machine Learing (ICML)*, 2019b.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530, 2019.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. *arXiv preprint arXiv:2005.07652*, 2020.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, pages 3517–3529, 2017.

Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1299–1316, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Appendix A. Proof of results in Section 3

A.1. Hypothesis class with VC dimension $d = 1$

For any $f \in \mathcal{H}$, let $\max_{x\in S_0}^{\leq_f^{\mathcal{H}}} x$ denote the maximal element w.r.t. the partial ordering $\leq_f^{\mathcal{H}}$ in any non-empty ordered finite set S_0, i.e., $\forall x' \in S_0, x' \leq_f^{\mathcal{H}} \left(\max_{x\in S_0}^{\leq_f^{\mathcal{H}}} x\right)$. Then for any arbitratrily chosen but fixed $f \in \mathcal{H}$, the algorithm (originally proposed by Ben-David, 2015) is described as follows.

Algorithm 4 Robust algorithm for $\mathcal{H}$ with VC dimension $d = 1$

1: **input**: data S
2: If every $(x, y) \in S$ has $y = f(x)$, **return** $\widehat{h} = f$
3: Let $x_m = \max_{(x,y)\in S, y\neq f(x)}^{\leq_f^{\mathcal{H}}} x$
4: $\widehat{h}(x) = 1 - f(x)$ for $x \leq_f^{\mathcal{H}} x_m$ and $\widehat{h}(x) = f(x)$ otherwise
5: **return** $\widehat{h}$

By Lemma 5 of Ben-David (2015), $\leq_f^{\mathcal{H}}$ for $d = 1$ is a tree ordering. Thus, all points labeled differently by f and h^* should lie on one path, i.e., for every $x, x' \in \mathcal{X}$, if $h^*(x) \neq f(x)$ and $h^*(x') \neq f(x')$, then $x \leq_f^{\mathcal{H}} x'$ or $x' \leq_f^{\mathcal{H}} x$. Due to this structure property of hypothesis class with VC dimenison 1, adding clean-label attacking points can only narrow down the error region of Algorithm 4.

Theorem 11 *For any $\mathcal{H}$ with VC dimension $d = 1$, Algorithm 4 can (ε, δ)-robustly learn $\mathcal{H}$ using m samples, where*

$$m = \left\lceil \frac{2\ln(1/\delta)}{\varepsilon} \right\rceil .$$

Proof First, we prove that $X = \{x \in \mathcal{X} | h^*(x) \neq f(x)\}$ is totally ordered by $\leq_f^{\mathcal{H}}$. That is, for every $x, x' \in \mathcal{X}$, if $h^*(x) \neq f(x)$ and $h^*(x') \neq f(x')$, then $x \leq_f^{\mathcal{H}} x'$ or $x' \leq_f^{\mathcal{H}} x$. If it is not true, then there exists $h_1, h_2 \in \mathcal{H}$ such that $h_1(x) \neq f(x), h_1(x') = f(x')$ and $h_2(x') \neq f(x'), h_2(x) = f(x)$. Then, $\{f, h^*, h_1, h_2\}$ shatters $\{x, x'\}$, which contradicts that $d = 1$. Therefore the finite set $\{x|(x, y) \in S, h^*(x) \neq f(x)\} \subseteq \{x \in \mathcal{X}|h^*(x) \neq f(x)\}$ is also an ordered set. If the set is not empty, x_m in the algorithm is well-defined.

Now note that either $\widehat{h} = f$ or else x_m is defined and then every x with $\widehat{h}(x) \neq f(x)$ has $x \leq_f^{\mathcal{H}} x_m$, which implies $h^*(x) \neq f(x)$ as well (since $h^*(x_m) \neq f(x_m)$). In particular, if every $(x, y) \in S_{\text{trn}}$ has $y = f(x)$ then the attackable region $\text{ATK}(h^*, S_{\text{trn}}, \mathcal{A}, \text{Adv}) \subseteq X$. Otherwise let $x_{S_{\text{trn}}} = \max_{(x,y)\in S_{\text{trn}}, y\neq f(x)}^{\leq_f^{\mathcal{H}}} x$, the maximal element in $\{x|(x, y) \in S_{\text{trn}}, h^*(x) \neq f(x)\}$, we would have that $\text{ATK}(h^*, S_{\text{trn}}, \mathcal{A}, \text{Adv}) \subseteq \{x|x_{S_{\text{trn}}} \leq_f^{\mathcal{H}} x, h^*(x) \neq f(x)\}$.

In particular, if $\mathcal{P}_{\mathcal{D}}(X) \leq \varepsilon$ the above facts imply $\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \mathcal{A}) \leq \varepsilon$. Otherwise if $\mathcal{P}_{\mathcal{D}}(X) > \varepsilon$, then we let $X_\varepsilon \subseteq X$ be any minimal set such that $\mathcal{P}_{\mathcal{D}}(X_\varepsilon) \geq \frac{\varepsilon}{2}$ and for every $x' \in X \setminus X_\varepsilon$ and every $x \in X_\varepsilon$, $x' \leq_f^{\mathcal{H}} x$. If $\mathcal{P}_{\mathcal{D}}(X_\varepsilon) \geq \varepsilon$, there exists an element $x \in X_\varepsilon$ with probability mass at least $\frac{\varepsilon}{2}$. When $m \geq \frac{2\ln(1/\delta)}{\varepsilon}$, with probability at least $1 - \delta$, x is in S_{trn} and therefore $\text{ATK}(h^*, S_{\text{trn}}, \mathcal{A}, \text{Adv}) \subseteq X_\varepsilon \setminus \{x\}$, so that $\text{atk}_{\mathcal{D}}(h^*, S_{\text{trn}}, \mathcal{A}) \leq \frac{\varepsilon}{2}$. Otherwise, if $\frac{\varepsilon}{2} \leq \mathcal{P}_{\mathcal{D}}(X_\varepsilon) < \varepsilon$, then as long as S_{trn} contains at least one example from X_ε, then

$\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \subseteq X_\varepsilon$, so that $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathcal{P}_{\mathcal{D}}(X_\varepsilon) < \varepsilon$. Since S_{trn} contains an example from X_ε with probability at least $1 - (1 - \frac{\varepsilon}{2})^m$, when $m \geq \frac{2\ln(1/\delta)}{\varepsilon}$ we have that with probability at least $1 - \delta$, $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \varepsilon$. ■

A.2. Union of intervals

Theorem 12 *The algorithm described in Section 3 can (ε, δ)-robustly learn union of intervals $\mathcal{H}_k$ using m samples, where*

$$m = O\left(\frac{1}{\varepsilon}(k\log(1/\varepsilon) + \log(1/\delta))\right).$$

Proof We denote the target function by $h^* = \mathbb{1}[\cup_{i=1}^{k^*}(a_{2i-1}, a_{2i})]$ with $0 = a_0 \leq a_1 \leq \ldots \leq a_{2k^*+1} = 1$ and $a_{2i-1} \neq a_{2i}, \forall i \in [k^*]$ for some $0 \leq k^* \leq k$. In the following, we will construct two classifiers consistent with the training set and then prove that the attackable rate of our algorithm is upper bounded by the sum of the error rates of these two classifiers. For any $i \in [k^*]$, we define c_i^+ as the minimum consistent positive interval within (a_{2i-1}, a_{2i}), i.e.,

$$c_i^+ = \begin{cases} \left[\min_{x \in (a_{2i-1}, a_{2i}):(x,y) \in S_{\mathrm{trn}}} x, \max_{x \in (a_{2i-1}, a_{2i}):(x,y) \in S_{\mathrm{trn}}} x\right] & \text{if } S_{\mathrm{trn}} \cap (a_{2i-1}, a_{2i}) \times \mathcal{Y} \neq \emptyset\,, \\ \emptyset & \text{otherwise}\,. \end{cases}$$

Similarly, for $i = 0, \ldots, k^*$, we define c_i^- as the minimum consistent negative interval within $[a_{2i}, a_{2i+1}]$. Since $x = 0$ and $x = 1$ are labeled as 0 by every hypothesis in $\mathcal{H}_k$, we would like c_0^- to include $x = 0$ and $c_{k^*}-$ to include $x = 1$. Then we denote by $\overline{S_{\mathrm{trn}}} = S_{\mathrm{trn}} \cup \{(0,0), (1,0)\}$ and then define c_i^- as

$$c_i^- = \begin{cases} \left[\min_{x \in [a_{2i}, a_{2i+1}]:(x,y) \in \overline{S_{\mathrm{trn}}}} x, \max_{x \in [a_{2i}, a_{2i+1}]:(x,y) \in \overline{S_{\mathrm{trn}}}} x\right] & \text{if } (\overline{S_{\mathrm{trn}}}) \cap [a_{2i}, a_{2i+1}] \times \mathcal{Y} \neq \emptyset\,, \\ \emptyset & \text{otherwise}\,. \end{cases}$$

Let us define two classifiers: $h_c^+ = \mathbb{1}[\cup_{i=1}^{k^*} c_i^+]$ and $h_c^- = 1 - \mathbb{1}[\cup_{i=0}^{k^*} c_i^-]$. Then we extend $\mathcal{H}_k$ to $\overline{\mathcal{H}}_k = \cup_{k' \leq k}\{\mathbb{1}[\cup_{i=1}^{k'}[a_i, b_i]] | 0 \leq a_i < b_i \leq 1, \forall i \in [k']\} \cup \mathcal{H}_k$ by including union of closed intervals. Since both $h_c^+, h_c^- \in \overline{\mathcal{H}}_k$ are consistent with S_{trn} and the VC dimension of $\overline{\mathcal{H}}_k$ is $2k$, by classic uniform convergence results (Vapnik and Chervonenkis, 1974; Blumer et al., 1989), for any data distribution $\mathcal{D}$, with probability at least $1 - \delta$ over $S_{\mathrm{trn}} \sim \mathcal{D}^m$, $\mathrm{err}(h_c^+) \leq \frac{\varepsilon}{2}$ and $\mathrm{err}(h_c^-) \leq \frac{\varepsilon}{2}$ where $m = O(\frac{1}{\varepsilon}(2k\log(1/\varepsilon) + \log(1/\delta)))$.

It is easy to see that the algorithm (even under attack) will always predicts 1 over c_i^+ as the attacker cannot add negative instances into c_i^+. Then for any attacker Adv and any $i \in [k^*]$, for any $x \in \mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \cap (a_{2i-1}, a_{2i})$, we will have $x \notin c_i^+$, which is classified 0 by h_c^+. Therefore, $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \cap \{x | h^*(x) = 1\} \subseteq \{x | h_c^+(x) = 0, h^*(x) = 1\}$. We can prove a similar result for $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \cap \{x | h^*(x) = 0\}$. If the algorithm (under attack) predicts 1 on any $x \in [a_{2i}, a_{2i+1}]$ for $i = 1, \ldots, k^* - 1$, then $c_i^- = \emptyset$ and $h_c^-(x) = 1 \neq h^*(x)$. Note that the algorithm always correctly labels points in $[0, a_1]$ and $[a_{2k^*}, 1]$ as there are no positively-labeled points in these two intervals. Therefore, $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \cap \{x | h^*(x) = 0\} \subseteq \{x | h_c^-(x) = 1, h^*(x) = 0\}$. Then for any point in the attackable region, it is either in the error region of h_c^+ or in the error region of h_c^-. That is, $\mathrm{ATK}(h^*, S_{\mathrm{trn}}, \mathcal{A}, \mathrm{Adv}) \subseteq \{x | h_c^+(x) \neq h^*(x)\} \cup \{x | h_c^-(x) \neq h^*(x)\}$. Hence, the attackable rate $\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathrm{err}(h_c^+) + \mathrm{err}(h_c^-) \leq \varepsilon$. ■

A.3. Hypothesis class with infinite hollow star number

Theorem 13 *For any hypothesis class $\mathcal{H}$ with hollow star number $k_o \geq 3$, for any consistent proper learner $\mathcal{A}$, there exists a target function h^* and a data distribution $\mathcal{D}$ such that if $m \leq \lfloor (k_o-1)/2 \rfloor$, then $\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right] \geq 1/4$ and $\mathbb{P}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left(\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) > 1/8\right) \geq 1/7$.*

Proof For $\mathcal{H}$ with hollow star number k_o, there is an unrealizable set $S = \{(x_1, y_1), \ldots, (x_{k_o}, y_{k_o})\}$, such that for each $i \in \{1, \ldots, k_o\}$, there exists $h_i \in \mathcal{H}$ with $\{j \in [k_o] | h_i(x_j) \neq y_j\} = \{i\}$. The target function $h^* = h_{i^*}$. For notation simplicity, we let i represent point x_i and then $h(i) = h(x_i)$ for any classifier h and any $i \in [k_o]$. The marginal data distribution $\mathcal{D}_{\mathcal{X}}$ is a uniform distribution over $[k_o] \setminus \{i^*\}$. We define an attacker which injects all points except i, i^* to attack point i if it is unseen. Specifically, let $S_{i^*,i} = \{(j, h_{i^*}(j)) | j \in [k_o] \setminus \{i^*, i\}\}$ denotes the set of all points excluding i, i^* and we define the attacker as

$$\mathrm{Adv}(h_{i^*}, S_{\mathrm{trn}}, i) = \begin{cases} S_{i^*,i} & \text{if } (i, y_i) \notin S_{\mathrm{trn}}\,, \\ \emptyset & \text{else}\,. \end{cases}$$

Given the input data set $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{i^*}, S_{\mathrm{trn}}, i)$, if (i, y_i) is not sampled in the training set, then any algorithm cannot tell whether the true target function is h_i or it is h_{i^*}. Since $\{(1, y_1), \ldots, (k_o, y_{k_o})\}$ is unrealizable, any proper classifier consistent with S_{i,i^*} cannot predict i as $h_{i^*}(i) = y_i$ and i^* as $h_i(i^*) = y_{i^*}$ at the same time. Let S_{trn} be $m \leq \lfloor k_o/2 \rfloor$ i.i.d. samples from $\mathcal{D}$ and then we have

$$\begin{aligned}
&\sup_{i^*\in[k_o]} \mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right] \\
\geq&\mathbb{E}_{i^*\sim\mathrm{Unif}([k_o]), S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right] \\
\geq&\mathbb{E}_{i^*\sim\mathrm{Unif}([k_o]), S_{\mathrm{trn}}\sim\mathcal{D}^m, (i,y_i)\sim\mathcal{D}, \mathcal{A}}\left[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i) \neq h_{i^*}(i) \cap i \notin S_{\mathrm{trn},\mathcal{X}}]\right] \\
\geq&\mathbb{E}_{i^*,i\sim\mathrm{Unif}([k_o]\setminus\{i^*\})}\left[\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m, \mathcal{A}}\left[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i) \neq h_{i^*}(i) | i \notin S_{\mathrm{trn},\mathcal{X}}]\right] \cdot \mathbb{P}\left(i \notin S_{\mathrm{trn},\mathcal{X}}\right)\right] \\
\geq&\frac{1}{(k_o-1)k_o} \sum_{i^*=1}^{k_o} \sum_{i\neq i^*} \mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m, \mathcal{A}}\left[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i) \neq h_{i^*}(i) | i \notin S_{\mathrm{trn},\mathcal{X}}]\right] \cdot \frac{1}{2} \\
=&\frac{1}{2(k_o-1)k_o} \sum_{i^*=1}^{k_o} \sum_{i\neq i^*} \mathbb{E}_{S_{\mathrm{trn}}\sim\mathrm{Unif}^m(S_{i^*,i}), \mathcal{A}}\left[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i) \neq h_{i^*}(i)]\right] \\
=&\frac{\sum_{i^*<i} \mathbb{E}_{S_{\mathrm{trn}}\sim\mathrm{Unif}^m(S_{i^*,i}), \mathcal{A}}\left[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i) \neq h_{i^*}(i)] + \mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup S_{i^*,i}, i^*) \neq h_i(i^*)]\right]}{2(k_o-1)k_o} \\
\geq&\frac{(k_o-1)k_o}{4(k_o-1)k_o} = \frac{1}{4}\,.
\end{aligned}$$

For the second part, by Markov's inequality, we have

$$\begin{aligned}
&\mathbb{P}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left(\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) > 1/8\right) = 1 - \mathbb{P}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left(\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq 1/8\right) \\
\geq&1 - \frac{1 - \mathbb{E}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right]}{7/8} = \frac{1}{7}\,,
\end{aligned}$$

which completes the proof. ■

Theorem 14 *If $k_o = \infty$, then for any consistent* proper *learning algorithm $\mathcal{A}$, for every $m \in \mathbb{N}$, $\exists h^* \in \mathcal{H}$ and distribution $\mathcal{D}$ on D_{h^*} such that $\mathbb{E}_{S_{\mathrm{trn}} \sim \mathcal{D}^m}[\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A})] \geq 1/4$.*

Proof For any hypothesis class with $k_o = \infty$, there exists a sequence of hollow star set $\{S_i\}_{i=1}^{\infty}$ with increasing size $\{k_i\}_{i=1}^{\infty}$ with $k_1 \geq 3$. Then following the proof of Theorem 13, for any $m \leq \lfloor (k_i - 1)/2 \rfloor$ for some i, there exists a target function and a data distribution such that the expected attackable rate is at least $1/4$ with sample size m. Since $k_i \to \infty$ as $i \to \infty$, this theorem is proved. ■

A.4. Proof of Theorem 1

Here we present the proof of Theorem 1 establishing that any deterministic robust learner necessarily obtains a sample complexity with $O(1/\varepsilon)$ dependence on ε.

Proof of Theorem 1 Without loss of generality, we suppose $R(m) \leq 1$ and $R(m)$ is nonincreasing, since we can always replace it with $\sup_{m' \geq m} \min\{R(m'), 1\}$, which is monotone and inherits the other assumed properties of R. For convenience, let us also extend the function $R(m)$ to non-integer values of m by defining $R(\alpha) = R(\lfloor \alpha \rfloor)$, and defining $R(0) = 1$. Also define $\mathrm{Log}(x) = \lceil \log_2(x) \rceil$ for any $x \geq 1$.

Fix any $h^* \in \mathcal{H}$. Since $\mathcal{A}$ is deterministic, note that for any finite multiset $S \subseteq D_{h^*}$ there is a set $\mathrm{ATK}_S \subseteq \mathcal{X}$ corresponding to the points that would be attackable for $\mathcal{A}$ if $S_{\mathrm{trn}} = S$. Moreover, we may note that the set ATK_S is *non-increasing* in S (subject to $S \subseteq D_{h^*}$), since adding any $(x, y) \in D_{h^*}$ to S is equivalent to constraining the adversary to include these points in its attack set.

Now we argue that $R\Big(\frac{m}{\mathrm{Log}(1/\delta)}\Big)$ is a $1 - \delta$ confidence bound on $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A})$. For any distribution $\mathcal{D}$ on D_{h^*}, and any $\delta \in (0, 1)$, if $m < \mathrm{Log}(1/\delta)$ then we trivially have $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq R\Big(\frac{m}{\mathrm{Log}(1/\delta)}\Big)$. Otherwise, if $m \geq \mathrm{Log}(1/\delta)$, then letting $S_{\mathrm{trn}} \sim \mathcal{D}^m$, letting S_1 be the first $\Big\lfloor \frac{m}{\mathrm{Log}(1/\delta)} \Big\rfloor$ elements of S_{trn}, S_2 the next $\Big\lfloor \frac{m}{\mathrm{Log}(1/\delta)} \Big\rfloor$ elements of S_{trn}, and so on up to $S_{\mathrm{Log}(1/\delta)}$, each $i \leq \mathrm{Log}(1/\delta)$ has, independently, probability at least $\frac{1}{2}$ of $\mathrm{atk}_{\mathcal{D}}(h^*, S_i, \mathcal{A}) \leq R\Big(\frac{m}{\mathrm{Log}(1/\delta)}\Big)$. In particular, this implies that, with probability at least $1 - (1/2)^{\mathrm{Log}(1/\delta)} \geq 1 - \delta$, at least one $i \leq \mathrm{Log}(1/\delta)$ will satisfy this inequality. Moreover, by the monotonicity property of ATK, we know that $\mathrm{ATK}_{S_{\mathrm{trn}}} \subseteq \bigcap_{i \leq \mathrm{Log}(1/\delta)} \mathrm{ATK}_{S_i}$. Thus, with probability at least $1 - \delta$,

$$\begin{aligned}\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) &= \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\mathrm{trn}}}) \leq \min_{i \leq \mathrm{Log}(1/\delta)} \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_i}) \\ &= \min_{i \leq \mathrm{Log}(1/\delta)} \mathrm{atk}_{\mathcal{D}}(h^*, S_i, \mathcal{A}) \leq R\left(\frac{m}{\mathrm{Log}(1/\delta)}\right).\end{aligned}$$

The remainder of the proof follows a familiar "conditioning" argument from the literature on log factors in the sample complexity of PAC learning (e.g., Hanneke, 2009, 2016a). Fix any distribution $\mathcal{D}$ over D_{h^*}. We proceed by induction on m, establishing for each m that $\forall \delta \in (0, 1)$, for $S_{\mathrm{trn}} \sim \mathcal{D}^m$, with probability at least $1 - \delta$, $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \frac{c}{m} \mathrm{Log}\big(\frac{1}{\delta}\big)$, where c is a finite R-dependent constant. Note that this suffices to establish the theorem by taking $c_R = c$ (assuming base 2 in the log). The claim is trivially satisfied for $m < 3\,\mathrm{Log}\big(\frac{1}{\delta}\big)$, as the claimed bound is vacuous (taking any $c \geq 3$). Now as an inductive hypothesis suppose $m \geq 3\,\mathrm{Log}\big(\frac{1}{\delta}\big)$ is such that, for every $m' < m$, for $S_{\mathrm{trn}} \sim \mathcal{D}^{m'}$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \frac{c}{m'} \mathrm{Log}\big(\frac{1}{\delta}\big)$.

Fix any $\delta \in (0,1)$ and let $S_{\mathrm{trn}} \sim \mathcal{D}^m$. Note that $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\mathrm{trn}}})$. Let $S_{\lfloor m/2 \rfloor}$ be the first $\lfloor m/2 \rfloor$ of the data points in S_{trn}, and let $T = (S \backslash S_{\lfloor m/2 \rfloor}) \cap (\mathrm{ATK}_{S_{\lfloor m/2 \rfloor}} \times \mathcal{Y})$: that is, T are the samples in the last $\lceil m/2 \rceil$ points in S_{trn} that are in the attackable region when $\mathcal{A}$ has training set $S_{\lfloor m/2 \rfloor}$.

Since, conditioned on $S_{\lfloor m/2 \rfloor}$ and $|T|$, the examples in T are conditionally i.i.d. with each sample having distribution $\mathcal{D}(\cdot | \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}} \times \mathcal{Y})$ on D_{h^*}, the property of $R(\cdot)$ established above implies that with conditional (given $S_{\lfloor m/2 \rfloor}$ and $|T|$) probability at least $1 - \frac{\delta}{3}$, we have $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_T | x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \leq R\Big(\frac{|T|}{\mathrm{Log}(3/\delta)}\Big)$. By the law of total probability, this inequality holds with (unconditional) probability at least $1 - \frac{\delta}{3}$.

Furthermore, a Chernoff bound (applied under the conditional distribution given $S_{\lfloor m/2 \rfloor}$) and the law of total probability imply that, with probability at least $1 - \frac{\delta}{3}$, if $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \geq \frac{16}{m} \ln \frac{3}{\delta}$, then $|T| \geq \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \frac{m}{4}$. Combining these two events with monotonicity of R, by the union bound, with probability at least $1 - \frac{2}{3}\delta$, either $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) < \frac{16}{m} \ln \frac{3}{\delta}$ or $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_T | x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \leq R\Big(\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \frac{m}{4\,\mathrm{Log}(3/\delta)}\Big)$.

Next, by monotonicity of ATK_S, we have $\mathrm{ATK}_{S_{\mathrm{trn}}} \subseteq \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}} \cap \mathrm{ATK}_T$. Therefore, $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_T | x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}})$. Thus, on the above event of probability at least $1 - \frac{2}{3}\delta$, either $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) < \frac{16}{m} \ln \frac{3}{\delta}$ or

$$\begin{aligned}
\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) &\leq \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) R\left(\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \mathrm{ATK}_{S_{\lfloor m/2 \rfloor}}) \frac{m}{4\,\mathrm{Log}(3/\delta)}\right) \\
&= \mathrm{atk}_{\mathcal{D}}(h^*, S_{\lfloor m/2 \rfloor}, \mathcal{A}) R\left(\mathrm{atk}_{\mathcal{D}}(h^*, S_{\lfloor m/2 \rfloor}, \mathcal{A}) \frac{m}{4\,\mathrm{Log}(3/\delta)}\right).
\end{aligned}$$

By the inductive hypothesis, with probability at least $1 - \frac{\delta}{3}$, we have that $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\lfloor m/2 \rfloor}, \mathcal{A}) \leq \frac{c}{\lfloor m/2 \rfloor} \mathrm{Log}\big(\frac{3}{\delta}\big) \leq \frac{3c}{m} \mathrm{Log}\big(\frac{3}{\delta}\big)$. For any $\alpha \geq 1$, define $R'(\alpha) = \frac{1}{\alpha} \sup_{1 \leq \alpha' \leq \alpha} \alpha' R(\alpha')$, and note that $R(\alpha) \leq R'(\alpha)$ for all $\alpha \geq 1$, and $\alpha R'(\alpha)$ is nondecreasing in $\alpha \geq 1$. Therefore, on the above event,

$$\begin{aligned}
&\mathrm{atk}_{\mathcal{D}}(h^*, S_{\lfloor m/2 \rfloor}, \mathcal{A}) R\left(\mathrm{atk}_{\mathcal{D}}(h^*, S_{\lfloor m/2 \rfloor}, \mathcal{A}) \frac{m}{4\,\mathrm{Log}(3/\delta)}\right) \\
&\leq \frac{3c}{m} \mathrm{Log}\left(\frac{3}{\delta}\right) R'\left(\frac{3c}{4}\right) \leq \frac{9c}{m} \mathrm{Log}\left(\frac{1}{\delta}\right) R'\left(\frac{3c}{4}\right).
\end{aligned}$$

Now note that $\lim_{\alpha\to\infty} R'(\alpha) = 0$. To see this, for the sake of contradiction, suppose $\exists \varepsilon > 0$ and a strictly increasing sequence $\alpha_t \geq 1$ with $\alpha_t \to \infty$ such that $R'(\alpha_t) \geq \varepsilon$, and let α'_t be any sequence with $1 \leq \alpha'_t \leq \alpha_t$ and $\frac{1}{\alpha_t} \alpha'_t R(\alpha'_t) \geq R'(\alpha_t)/2 \geq \varepsilon/2$. If there exists an infinite subsequence t_i with α'_{t_i} bounded above by some finite $\overline{\alpha}$, then $\lim_{i\to\infty} \frac{1}{\alpha_{t_i}} \alpha'_{t_i} R(\alpha'_{t_i}) \leq \lim_{i\to\infty} \frac{\overline{\alpha}}{\alpha_{t_i}} = 0$: a contradiction. Otherwise, we have $\alpha'_t \to \infty$, so that $\lim_{t\to\infty} \frac{1}{\alpha_t} \alpha'_t R(\alpha'_t) \leq \lim_{t\to\infty} R(\alpha'_t) = 0$: again, a contradiction. Thus, since we have just established that $\lim_{\alpha\to\infty} R'(\alpha) = 0$, there exists a sufficiently large choice of c for which $R'\big(\frac{3c}{4}\big) \leq \frac{1}{9}$, so that $\frac{9c}{m} \mathrm{Log}\big(\frac{1}{\delta}\big) R'\big(\frac{3c}{4}\big) \leq \frac{c}{m} \mathrm{Log}\big(\frac{1}{\delta}\big)$.

Altogether, by the union bound, we have established that with probability at least $1 - \delta$, either $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) < \frac{16}{m} \ln \frac{3}{\delta}$ or $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \frac{c}{m} \mathrm{Log}\big(\frac{1}{\delta}\big)$. Taking c sufficiently large so that $c \geq 16 \ln(3e)$, both cases imply that $\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \frac{c}{m} \mathrm{Log}\big(\frac{1}{\delta}\big)$. The theorem now follows by

the principle of induction. ■

Appendix B. Proof of Theorem 2

In this section, we first formally prove the statement in the case of $n = 3$, for which we already provided a proof sketch in Section 4. In the proof sketch, we relax the definition of linear hypothesis class by allowing the decision boundary to be either positive or negative. Here, we adopt the convention that the boundary is only allowed to be positive. Then we prove the statement in the case of $n = 2$, which requires a more delicate construction. Before proving Theorem 2, we first introduce a lemma.

Lemma 1 *For any hypothesis class $\mathcal{H}$, any algorithm $\mathcal{A}$ and any $m > 0$, if there exists a universal constant $c > 0$, a distribution μ over $\mathcal{H}$, and a set of distributions $\mathcal{D}(h)$ over D_h for every $h \in \mathrm{supp}(\mu)$, such that $\mathbb{E}_{h\sim\mu, S_{\mathrm{trn}}\sim\mathcal{D}(h)^m}\left[\mathrm{atk}_{\mathcal{D}(h)}(h, S_{\mathrm{trn}}, \mathcal{A})\right] \geq 2c$, then $\mathcal{H}$ is not (ε, δ)-robust learnable.*

Proof First, take $\varepsilon = c$, we have by the definition of sup, there exists an $h^* \in \mathcal{H}$ such that,

$$\begin{aligned}
&\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}(h^*)^m}\left[\mathrm{atk}_{\mathcal{D}(h^*)}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right] \\
\geq &\sup_{h\in\mathcal{H}} \mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}(h)^m}\left[\mathrm{atk}_{\mathcal{D}(h)}(h, S_{\mathrm{trn}}, \mathcal{A})\right] - \varepsilon \\
\geq &\mathbb{E}_{h\sim\mu, S_{\mathrm{trn}}\sim\mathcal{D}(h)^m}\left[\mathrm{atk}_{\mathcal{D}(h)}(h, S_{\mathrm{trn}}, \mathcal{A})\right] - \varepsilon \\
\geq &c\,.
\end{aligned}$$

Then by Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{atk}_{\mathcal{D}(h^*)}(h^*, S_{\mathrm{trn}}, \mathcal{A}) > c/2\right) = 1 - \mathbb{P}\left(\mathrm{atk}_{\mathcal{D}(h^*)}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq c/2\right) \\
\geq &1 - \frac{1 - \mathbb{E}\left[\mathrm{atk}_{\mathcal{D}(h^*)}(h^*, S_{\mathrm{trn}}, \mathcal{A})\right]}{1 - c/2} \geq \frac{c}{2 - c}\,.
\end{aligned}$$

Hence, $\mathcal{H}$ is not (ε, δ)-robust learnable. ■

Proof of Theorem 2 in $n = 3$ We divide the proof into three parts: a) the construction of the target function and the data distribution, b) the construction of the attacker and c) the analysis of the attackable rate.

The target function and the data distribution. We denote by $\Gamma = \Gamma^3(\mathbf{0}, 1)$ the sphere of the 3-dimensional unit ball centered at the origin. For some small $0 < \eta < 1/6$, let $\mathcal{H}_\eta = \{h(x) = \mathbb{1}[\langle w, x\rangle - \frac{1}{2} \geq 0] \,|\, \|w\| = 1\} \cup \{h(x) = \mathbb{1}[\langle w, x\rangle - \frac{1-\eta}{2} \leq 0] \,|\, \|w\| = 1\}$ denote a set of linear classifiers with boundary $1/2$ or $\frac{1-\eta}{2}$ away from the origin. Let $K_w = \{x \,|\, \langle w, x\rangle - \frac{1}{2} = 0\}$ denote the hyperplane of the boundary of $h = \mathbb{1}[\langle w, x\rangle - \frac{1}{2} \geq 0]$ and $C_w = K_w \cap \Gamma$ denote the intersection of K_w and Γ, which is a circle with radius $\sqrt{3}/2$ centered at $w/2$.

We consider the target function h^* selected uniformly at random from $\mathcal{H}_\eta$, which is equivalent to: randomly picking $w \sim \mathrm{Unif}(\Gamma)$ and randomly picking $j \sim \mathrm{Ber}(1/2)$; if $j = 1$, letting $h^* = h^*_{w,j} = \mathbb{1}[\langle w, x\rangle - \frac{1}{2} \geq 0]$; otherwise letting $h^* = h^*_{w,j} = \mathbb{1}[\langle w, x\rangle - \frac{1-\eta}{2} \leq 0]$. If the target function $h^* = h^*_{w,j}$, the data distribution $\mathcal{D} = \mathcal{D}_{w,j}$ is the uniform distribution over $C_w \times \{j\}$.

Note that all instances on the circle C_w are labeled as j by $h^*_{w,j}$. We will show that the expected attackable rate $\mathbb{E}_{h^*\sim\mathrm{Unif}(\mathcal{H}_\eta),S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}_{\mathcal{D}}(h^*,S_{\mathrm{trn}},\mathcal{A})\right]\geq 1/2$. Combining with Lemma 1, we prove Theorem 2 in $n=3$.

The attacker. Then we define the attacker Adv in the following way. We first define a map $m_{x_0}:\Gamma\mapsto\Gamma$ for some $x_0\in\Gamma$ such that $m_{x_0}(x)=2\langle x_0,x\rangle x_0-x$. Here, $m_{x_0}(x)$ is the reflection of x through the line passing the origin and x_0. Note that $m_{x_0}(m_{x_0}(x))=x$. This symmetric property will help to confuse algorithms such that no algorithm can distinguish the training data and the poisoning data. For $S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j}$, we define $m_{x_0}(S_{\mathrm{trn}})=\{(m_{x_0}(x),1-y)|(x,y)\in S_{\mathrm{trn}}\}$, and let

$$\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0)=\begin{cases}m_{x_0}(S_{\mathrm{trn}}) & \text{if } S_{\mathrm{trn},\mathcal{X}}\cap\mathcal{B}(x_0,\sqrt{3\eta/2})=\emptyset\,,\\ \emptyset & \text{else.}\end{cases}$$

Now we show that Adv is a clean-label attacker. In the second case of $\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0)=\emptyset$, it is clean-labeled trivially. In the first case of $S_{\mathrm{trn},\mathcal{X}}\cap\mathcal{B}(x_0,\sqrt{3\eta/2})=\emptyset$, we discuss two cases:

- The target function $h^*=h^*_{w,j}=\mathbb{1}[\langle w,x\rangle-\frac{1}{2}\geq 0]$ has its decision boundary $\frac{1}{2}$ away from the origin, i.e., $j=1$. Then every training instance is labeled by 1 and for any training instance x, $\langle w,m_{x_0}(x)\rangle-\frac{1}{2}=\langle x_0,x\rangle-1<0$. Hence $\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0)$ is clean-labeled.
- The target function $h^*=h^*_{w,j}=\mathbb{1}[\langle w,x\rangle-\frac{1-\eta}{2}\leq 0]$ has its decision boundary $\frac{1-\eta}{2}$ away from the origin, i.e., $j=0$. For each training instance x, since $x\notin\mathcal{B}(x_0,\sqrt{3\eta/2})$, we have $\|x-x_0\|_2^2\geq\frac{3\eta}{2}$ and thus, $\langle x,x_0\rangle\leq 1-\frac{3\eta}{4}$. Then $\langle w,m_{x_0}(x)\rangle-\frac{1-\eta}{2}=\langle x_0,x\rangle-(1-\frac{\eta}{2})\leq 1-\frac{3\eta}{4}-(1-\frac{\eta}{2})<0$. Hence $\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0)$ is clean-labeled.

Analysis. Let $\mathcal{E}_1(h^*_{w,j},S_{\mathrm{trn}},x_0)$ denote the event of $\{\mathcal{A}(S_{\mathrm{trn}}\cup\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0),x_0)\neq h^*_{w,j}(x_0)\}$ and $\mathcal{E}_2(S_{\mathrm{trn}},x_0)$ denote the event of $S_{\mathrm{trn},\mathcal{X}}\cap\mathcal{B}(x_0,\sqrt{3\eta/2})=\emptyset$. It is not hard to check that $\mathcal{E}_2(S_{\mathrm{trn}},x_0)=\mathcal{E}_2(m_{x_0}(S_{\mathrm{trn}}),x_0)$ due to the symmetrical property of the reflection. Besides, conditional on $\mathcal{E}_2(S_{\mathrm{trn}},x_0)$, the poisoned data set $S_{\mathrm{trn}}\cup\mathrm{Adv}(h^*_{w,j},S_{\mathrm{trn}},x_0)=m_{x_0}(S_{\mathrm{trn}})\cup\mathrm{Adv}(h^*_{m_{x_0}(w),1-j},m_{x_0}(S_{\mathrm{trn}}),x_0)$ and thus, any algorithm $\mathcal{A}$ will behave the same (under attack) at test instance x_0 given training set S_{trn} or $m_{x_0}(S_{\mathrm{trn}})$. Since $h^*_{w,j}(x_0)\neq h^*_{m_{x_0}(w),1-j}(x_0)$, we know that $\mathbb{1}[\mathcal{E}_1(h^*_{w,j},S_{\mathrm{trn}},x_0)]=\mathbb{1}[\neg\mathcal{E}_1(h^*_{m_{x_0}(w),1-j},m_{x_0}(S_{\mathrm{trn}}),x_0)]$ conditional on $\mathcal{E}_2(S_{\mathrm{trn}},x_0)$. Let $f_{w,j}(x)$ denote the probability density function of the marginal distribution of $\mathcal{D}_{w,j}$ (i.e., the uniform distribution over C_w) and then we have $f_{w,j}(x)=f_{m_{x_0}(w),1-j}(m_{x_0}(x))$. For any fixed x_0, the distributions of w and $m_{x_0}(w)$ and the distributions of j and $1-j$ are the same respectively. The training set S_{trn} are samples drawn from $\mathcal{D}_{w,j}$, and hence we can view $m_{x_0}(S_{\mathrm{trn}})$ as samples drawn from $\mathcal{D}_{m_{x_0}(w),1-j}$. Then for any algorithm $\mathcal{A}$, we have

$$\begin{aligned}
&\mathbb{E}_{h^*_{w,j}\sim\mathrm{Unif}(\mathcal{H}_\eta),S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j}}\left[\mathrm{atk}_{\mathcal{D}}(h^*_{w,j},S_{\mathrm{trn}},\mathcal{A})\right]\\
\geq&\mathbb{E}_{w\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(1/2),S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j},(x,y)\sim\mathcal{D}_{w,j},\mathcal{A}}\left[\mathbb{1}[\mathcal{E}_1(h^*_{w,j},S_{\mathrm{trn}},x)\cap\mathcal{E}_2(S_{\mathrm{trn}},x)]\right]\\
=&\int_{x\in\Gamma}\mathbb{E}_{w\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(1/2),S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j},\mathcal{A}}\left[f_{w,j}(x)\mathbb{1}[\mathcal{E}_1(h^*_{w,j},S_{\mathrm{trn}},x)\cap\mathcal{E}_2(S_{\mathrm{trn}},x)]\right]dx \qquad (1)\\
=&\int_{x\in\Gamma}\mathop{\mathbb{E}}_{w,j,S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j},\mathcal{A}}\left[f_{m_x(w),1-j}(x)\mathbb{1}[\neg\mathcal{E}_1(h^*_{m_x(w),1-j},m_x(S_{\mathrm{trn}}),x)\cap\mathcal{E}_2(m_x(S_{\mathrm{trn}}),x)]\right]dx \qquad (2)
\end{aligned}$$

$$= \int_{x\in\Gamma} \mathbb{E}_{w,j,S_{\mathrm{trn}}\sim\mathcal{D}^m_{w,j},\mathcal{A}} \left[f_{w,j}(x) \mathbb{1}[\neg\mathcal{E}_1(h^*_{w,j}, S_{\mathrm{trn}}, x) \cap \mathcal{E}_2(S_{\mathrm{trn}}, x)] \right] dx \tag{3}$$

$$= \frac{1}{2} \int_{x\in\Gamma} \mathbb{E}_{w\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(1/2),S_{\mathrm{trn}}\sim\mathcal{D}^m,\mathcal{A}} \left[f_{w,j}(x) \mathbb{1}[\mathcal{E}_2(S_{\mathrm{trn}}, x)] \right] dx \xrightarrow{\eta\to 0^+} \frac{1}{2}, \tag{4}$$

where Eq. (2) uses the fact $\mathcal{E}_2(S_{\mathrm{trn}}, x_0) = \mathcal{E}_2(m_{x_0}(S_{\mathrm{trn}}), x_0)$ and that conditional on $\mathcal{E}_2(S_{\mathrm{trn}}, x_0)$, $\mathbb{1}[\mathcal{E}_1(h^*_{w,j}, S_{\mathrm{trn}}, x_0)] = \mathbb{1}[\neg\mathcal{E}_1(h^*_{m_{x_0}(w),1-j}, m_{x_0}(S_{\mathrm{trn}}), x_0)]$; Eq. (3) uses the fact that for any fixed x_0, the distributions of w and $m_{x_0}(w)$ and the distributions of j and $1-j$ are the same respectively; and Eq. (4) is the average of Eq. (1) and Eq. (3). ■

Proof of Theorem 2 in $n = 2$ Again, we divide the proof into three parts.

The target function and the data distribution. In 2-dimensional space, we denote by $w = (\cos\theta, \sin\theta)$ and represent the target function $h^* = \mathbb{1}[\langle(\cos\theta^*, \sin\theta^*), x\rangle + b^* \geq 0]$ by (θ^*, b^*). Then the target function is selected in the following way: uniformly at random selecting a point o from a 2-dimensional ball centered at $\mathbf{0}$ with some large enough radius $r \geq 4$, i.e., $o \sim \mathrm{Unif}(\mathcal{B}^2(\mathbf{0}, r))$, then randomly selecting a direction $\theta^* \sim \mathrm{Unif}([0, 2\pi))$, and letting the target function be $h^* = \mathbb{1}[\langle(\cos\theta^*, \sin\theta^*), x - o\rangle \geq 0]$. Then for any $m \in \mathbb{N}$, we construct the data distribution over $2m$ discrete points, where all points are labeled the same and the distance between every two instances is independent of h^*. Specifically, the data distribution $\mathcal{D}$ is described as follows.

- We randomly draw $s \sim \mathrm{Ber}(1/2)$. We define two unit vectors $v_1 = (\sin\theta^*, -\cos\theta^*)$ and $v_2 = (2s-1)\cdot(\cos\theta^*, \sin\theta^*)$. Here v_1 is perpendicular to w^* and v_2 is in the same direction as w^* if $s = 1$ and in the opposite direction of w^* if $s = 0$.

- Let $\mathcal{X}_m = \{x_1, \ldots, x_{2m}\}$ be a set of $2m$ points. For notation simplicity, we also define x_0 and x_{2m+1}. Let $x_0 = o$ and for all $i \in [2m+1]$, let $x_i = x_{i-1} + l\cos(\beta_{i-1})v_1 + l\sin(\beta_{i-1})v_2$, where $\beta_i = 7\beta_{i-1}$, $\beta_0 = 7^{-2m}\cdot\frac{\pi}{6}$ and $l = \frac{1}{2m}$.

- Let the marginal data distribution be a uniform distribution over $\mathcal{X}_m$. Note that if $s = 1$, all training points lie on the positive side of the decision boundary and are labeled by 1; if $s = 0$, all training points lie on the negative side and are labeled by 0.

Here (o, v_1, v_2) constructs a new coordinate system. For any $x \in \mathbb{R}^2$, we use $\widetilde{x} = ((x-o)^\top v_1, (x-o)^\top v_2)$ to represent x in this new coordinate system. Then the decision boundary of the target function is represented as $\langle\widetilde{x}, \widetilde{v_2}\rangle = 0$ and for any $x \in \mathcal{X}_m$ we have $\langle\widetilde{x}, \widetilde{v_2}\rangle > 0$. It is worth noting that for any $i \in [2m]$, if the positions of three points x_{i-1}, x_i, x_{i+1} are fixed, then o, s, v_1, v_2 are all fixed.

The attacker. For any $x_i \in \mathcal{X}_m$, we let $b_i = l\sum_{j=0}^{i}\cos\beta_j - \frac{l(\cos\beta_{i-1}+\cos\beta_i)\sum_{j=0}^{i}\sin\beta_j}{\sin\beta_{i-1}+\sin\beta_i}$ such that $z_i = b_i v_1 + o$, x_{i-1} and x_{i+1} are collinear. We denote by L_{x_i} the line passing z_i and x_i. Then for any $i \in [2m]$, let

$$\widetilde{\mathrm{Ref}_{L_{x_i}}}(x) = \frac{2\langle\widetilde{x} - \widetilde{z_i}, \widetilde{x_i} - \widetilde{z_i}\rangle}{\|\widetilde{x_i} - \widetilde{z_i}\|_2^2}(\widetilde{x_i} - \widetilde{z_i}) - \widetilde{x} + 2\widetilde{z_i},$$

be the reflection of x across L_{x_i} as illustrated in Fig. 1.

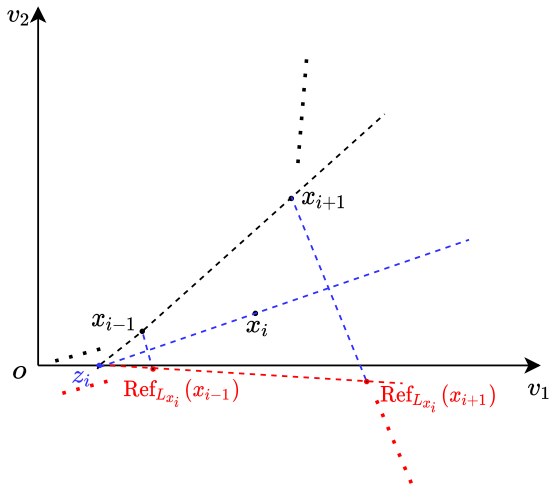

Figure 1: Illustration of $\mathrm{Ref}_{L_{x_i}}(\cdot)$.

For $S_{\mathrm{trn}} \sim \mathcal{D}^m$, we let $U = \mathcal{X}_m \setminus S_{\mathrm{trn},\mathcal{X}}$ denote the set of points not sampled in $\mathcal{X}_m$, where $|U| \geq m$. Then we define Adv as

$$\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x_i) = \begin{cases} \{(\mathrm{Ref}_{L_{x_i}}(x), 1-y) | (x,y) \in S_{\mathrm{trn}}\} & \text{if } x_i \in U\,, \\ \emptyset & \text{else.} \end{cases}$$

Now we need to show that Adv is a clean-label attacker. We will show that $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x_j)}, \widetilde{v_2} \right\rangle < 0$ for all $j \neq i \in [2m]$, which implies that the poison data is correctly labeled when $x_i \in U$. First, we claim that for any $j \neq i$, x_j lie in the polytope above the line passing x_{i-1}, x_{i+1}, the line passing x_{i-1}, x_i and the line passing x_i, x_{i+1}. Formally speaking, for any $j \neq i$, x_j satisfies

$$\begin{aligned} &\langle (-\sin\beta_{i-1} - \sin\beta_i, \cos\beta_{i-1} + \cos\beta_i), \widetilde{x_j} - \widetilde{x_{i+1}} \rangle \geq 0\,, \\ &\langle (-\sin\beta_{i-1}, \cos\beta_{i-1}), \widetilde{x_j} - \widetilde{x_i} \rangle \geq 0\,, \\ &\langle (-\sin\beta_i, \cos\beta_i), \widetilde{x_j} - \widetilde{x_i} \rangle \geq 0\,. \end{aligned}$$

This claim is not hard to prove. At a high level, we prove this claim by that $\{\beta_i\}_{i=0}^{2m}$ is monotonically increasing and that the polygon defined by connecting every pair of neighboring points in $\mathcal{X}_m \cup \{x_0, x_{2m+1}\}$ is convex. For the first constraint, it is satisfied trivially when $j = i-1, i+1$. If $j \geq i+2$, by direct calculation, we have

$$\begin{aligned} &\langle (-\sin\beta_{i-1} - \sin\beta_i, \cos\beta_{i-1} + \cos\beta_i), \widetilde{x_j} - \widetilde{x_{i+1}} \rangle \\ &= \left\langle (-\sin\beta_{i-1} - \sin\beta_i, \cos\beta_{i-1} + \cos\beta_i), \left(\sum_{k=i+1}^{j-1} \cos\beta_k, \sum_{k=i+1}^{j-1} \sin\beta_k \right) \right\rangle \\ &= \sum_{k=i+1}^{j-1} \cos\beta_k \left\langle (-\sin\beta_{i-1} - \sin\beta_i, \cos\beta_{i-1} + \cos\beta_i), \left(1, \frac{\sum_{k=i+1}^{j-1} \sin\beta_k}{\sum_{k=i+1}^{j-1} \cos\beta_k} \right) \right\rangle \\ &\geq \sum_{k=i+1}^{j-1} \cos\beta_k \left\langle (-\sin\beta_{i-1} - \sin\beta_i, \cos\beta_{i-1} + \cos\beta_i), (1, \tan\beta_{i+1}) \right\rangle \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\sum_{k=i+1}^{j-1} \cos\beta_k}{\cos\beta_{i+1}} (\sin(\beta_{i+1}-\beta_{i-1}) + \sin(\beta_{i+1}-\beta_i)) \\
&\geq 0\,.
\end{aligned}
$$

Similarly, if $j \leq i-2$, then

$$
\begin{aligned}
&\langle(-\sin\beta_{i-1}-\sin\beta_i, \cos\beta_{i-1}+\cos\beta_i), \widetilde{x_j} - \widetilde{x_{i+1}}\rangle \\
=&\langle(-\sin\beta_{i-1}-\sin\beta_i, \cos\beta_{i-1}+\cos\beta_i), \widetilde{x_j} - \widetilde{x_{i-1}}\rangle \\
=&-\left\langle(-\sin\beta_{i-1}-\sin\beta_i, \cos\beta_{i-1}+\cos\beta_i), \left(\sum_{k=j}^{i-2}\cos\beta_k, \sum_{k=j}^{i-2}\sin\beta_k\right)\right\rangle \\
=&\sum_{k=j}^{i-2}\cos\beta_k \left\langle(\sin\beta_{i-1}+\sin\beta_i, -\cos\beta_{i-1}-\cos\beta_i), \left(1, \frac{\sum_{k=j}^{i-2}\sin\beta_k}{\sum_{k=j}^{i-2}\cos\beta_k}\right)\right\rangle \\
\geq&\sum_{k=j}^{i-2}\cos\beta_k \langle(\sin\beta_{i-1}+\sin\beta_i, -\cos\beta_{i-1}-\cos\beta_i), (1, \tan\beta_{i-2})\rangle \\
\geq&\frac{\sum_{k=j}^{i-2}\cos\beta_k}{\cos\beta_{i-2}}(\sin(\beta_{i-1}-\beta_{i-2}) + \sin(\beta_i-\beta_{i-2})) \\
\geq&0\,.
\end{aligned}
$$

It is easy to check that x_j satisfies the second and the third constraints using the same way of computation, which is omitted here. Based on that x_j lie in the polytope for all $j \neq i$, then we only need to prove that $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x)}, \widetilde{v_2}\right\rangle < 0$ for the points lying on the faces of the polytope, which are $\widetilde{x} \in \{\widetilde{x_{i+1}} + \eta(\cos\beta_i, \sin\beta_i) | \eta \geq 0\}$, $\widetilde{x} \in \{\widetilde{x_{i-1}} - \eta(\cos\beta_{i-1}, \sin\beta_{i-1}) | \eta \geq 0\}$ and $\widetilde{x} \in \{\eta\widetilde{x_{i-1}} + (1-\eta)\widetilde{x_{i+1}} | \eta \in [0,1]\}$. Since $\widetilde{\mathrm{Ref}_{L_{x_i}}(\cdot)}$ is a linear transform, if we can show $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x_{i-1})}, \widetilde{v_2}\right\rangle < 0$ and $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x_{i+1})}, \widetilde{v_2}\right\rangle < 0$, then we have $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x)}, \widetilde{v_2}\right\rangle < 0$ for all points on the third face $\{\eta\widetilde{x_{i-1}} + (1-\eta)\widetilde{x_{i+1}} | \eta \in [0,1]\}$. Hence, we only need to prove the statement for points lying on the first two faces.

For any two vectors u, v, we denote by $\theta(u,v)$ the angle between u and v. Then let us denote by $\theta_1 = \theta(\widetilde{x_{i+1}} - \widetilde{z_i}, \widetilde{x_i} - \widetilde{z_i})$ the angle between $\widetilde{x_{i+1}} - \widetilde{z_i}$ and and $\widetilde{x_i} - \widetilde{z_i}$ and $\theta_2 = \theta(\widetilde{x_i} - \widetilde{z_i}, \widetilde{v_1})$ the angle between $\widetilde{x_i} - \widetilde{z_i}$ and $\widetilde{v_1}$. Then we have both $\theta_1 \leq \beta_i \leq \frac{\pi}{6}$ and $\theta_2 \leq \beta_i \leq \frac{\pi}{6}$. Then since $\|\widetilde{x_i} - \widetilde{x_{i-1}}\| = \|\widetilde{x_{i+1}} - \widetilde{x_i}\| = l$ and $\theta(\widetilde{x_{i+1}} - \widetilde{x_{i-1}}, \widetilde{x_i} - \widetilde{x_{i-1}}) = (\beta_i - \beta_{i-1})/2$ due to the construction, we have

$$
\sin\theta_1 = \frac{\|\widetilde{x_i} - (\widetilde{x_{i+1}} + \widetilde{x_{i-1}})/2\|}{\|\widetilde{x_i} - \widetilde{z_i}\|} = \frac{l\sin((\beta_i - \beta_{i-1})/2)}{\|\widetilde{x_i} - \widetilde{z_i}\|} = \frac{l\sin(3\beta_{i-1})}{\|\widetilde{x_i} - \widetilde{z_i}\|} \geq \frac{3\beta_{i-1}l}{2\|\widetilde{x_i} - \widetilde{z_i}\|}\,,
$$

where the last inequality is by $3\beta_{i-1} \leq \frac{\pi}{6}$. On the other hand, we have

$$
\sin\theta_2 = \frac{l\sum_{k=0}^{i-1}\sin\beta_k}{\|\widetilde{x_i} - \widetilde{z_i}\|} \leq \frac{l\sum_{k=0}^{i-1}\beta_k}{\|\widetilde{x_i} - \widetilde{z_i}\|} \leq \frac{7\beta_{i-1}l}{6\|\widetilde{x_i} - \widetilde{z_i}\|}\,.
$$

Combining these two equations, we have $\sin\theta_1 > \sin\theta_2$, which indicates that $\theta_1 > \theta_2$. Then since $\beta_i - \theta_2 = \theta(\widetilde{x_{i+1}} - \widetilde{x_i}, \widetilde{x_i} - \widetilde{z_i}) \geq \theta_1$, we have $\beta_i > 2\theta_2$. Let $\widetilde{w_i} = \frac{\widetilde{x_i} - \widetilde{z_i}}{\|\widetilde{x_i} - \widetilde{z_i}\|}$ denote the unit vector

in the direction of $\widetilde{x}_i - \widetilde{z}_i$. For $\widetilde{x} = \widetilde{x_{i+1}} + \eta(\cos\beta_i, \sin\beta_i)$ with $\eta \geq 0$,

$$\begin{aligned}
&\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x)}, \widetilde{v_2} \right\rangle \\
=&2\left\langle \widetilde{x} - \widetilde{z_i}, \widetilde{w_i} \right\rangle \left\langle \widetilde{w_i}, \widetilde{v_2} \right\rangle - \left\langle \widetilde{x} - \widetilde{z_i}, \widetilde{v_2} \right\rangle \\
=&2\left\langle \widetilde{x_{i+1}} - \widetilde{z_i} + \eta(\cos\beta_i, \sin\beta_i), \widetilde{w_i} \right\rangle \left\langle \widetilde{w_i}, \widetilde{v_2} \right\rangle - \left\langle \widetilde{x_{i+1}} - \widetilde{z_i} + \eta(\cos\beta_i, \sin\beta_i), \widetilde{v_2} \right\rangle \\
=&2\left\|\widetilde{x_{i+1}} - \widetilde{z_i}\right\| \cos\theta_1 \sin\theta_2 - \left\|\widetilde{x_{i+1}} - \widetilde{z_i}\right\| \sin(\theta_1 + \theta_2) + 2\eta\cos(\beta_i - \theta_2)\sin\theta_2 - \eta\sin\beta_i \\
=&\left\|\widetilde{x_{i+1}} - \widetilde{z_i}\right\| \sin(\theta_2 - \theta_1) + \eta\sin(2\theta_2 - \beta_i) \\
<&0\,.
\end{aligned}$$

It is easy to check that $\beta_{i-1} \leq \theta_2$ (let $\widetilde{p}$ denote the intersection of the line passing $\widetilde{x_{i-1}}$ and $\widetilde{x}_i$ and the line $\langle \widetilde{x}, \widetilde{v_2} \rangle = 0$, $\beta_{i-1} = \theta(\widetilde{x}_i - \widetilde{p}, \widetilde{v_1})$ and θ_2 is the external angle of triangle with vertices $\widetilde{p}, \widetilde{z}_i$ and $\widetilde{x}_i$). Then for $\widetilde{x} = \widetilde{x_{i-1}} - \eta(\cos\beta_{i-1}, \sin\beta_{i-1})$ with $\eta \geq 0$,

$$\begin{aligned}
&\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x)}, \widetilde{v_2} \right\rangle \\
=&2\left\langle \widetilde{x} - \widetilde{z_i}, \widetilde{w_i} \right\rangle \left\langle \widetilde{w_i}, \widetilde{v_2} \right\rangle - \left\langle \widetilde{x} - \widetilde{z_i}, \widetilde{v_2} \right\rangle \\
=&2\left\langle \widetilde{x_{i-1}} - \widetilde{z_i} - \eta(\cos\beta_{i-1}, \sin\beta_{i-1}), w_i \right\rangle \left\langle \widetilde{w_i}, \widetilde{v_2} \right\rangle - \left\langle \widetilde{x_{i-1}} - \widetilde{z_i} - \eta(\cos\beta_{i-1}, \sin\beta_{i-1}), \widetilde{v_2} \right\rangle \\
=&2\left\|\widetilde{x_{i-1}} - \widetilde{z_i}\right\| \cos\theta_1 \sin\theta_2 - \left\|\widetilde{x_{i-1}} - \widetilde{z_i}\right\| \sin(\theta_1 + \theta_2) - 2\eta\cos(\beta_{i-1} - \theta_2)\sin\theta_2 + \eta\sin\beta_{i-1} \\
=&\left\|\widetilde{x_{i-1}} - \widetilde{z_i}\right\| \sin(\theta_2 - \theta_1) + \eta\sin(\beta_{i-1} - 2\theta_2) \\
<&0\,.
\end{aligned}$$

Now we complete the proof of $\left\langle \widetilde{\mathrm{Ref}_{L_{x_i}}(x_j)}, \widetilde{v_2} \right\rangle < 0$ for all $j \neq i$ and that Adv is a clean-label attacker. It is worth noting that $L_{x_i'} = L_{x_i}$, where $L_{x_i'}$ is defined over $\{x_j' | j \in [2m]\}$ in the same way as L_{x_i} defined over $\{x_j | j \in [2m]\}$. This is because reflections of z_i and x_i over L_{x_i} are themselves. This symmetric property plays an important role in the analysis.

Analysis. Our probabilistic construction of the target function h^* and the data distribution $\mathcal{D}$ and the random sampling process of drawing m i.i.d. samples from $\mathcal{D}$ is equivalent to: sampling a multi-set of indexes $I_{\mathrm{trn}} \sim \mathrm{Unif}([2m])$ first; then selecting the target function and the data distribution to determine the positions of the m training points; mapping I_{trn} to S_{trn} by adding instance-label pair $(x_i, h^*(x_i))$ to S_{trn} for each i in I_{trn}. We let $I_u = [2m] \setminus I_{\mathrm{trn}}$ denote the indexes not sampled. As we know from the construction, for any $i \in [2m]$, once s and the positions of o and x_i is determined, the positions of other points in $\mathcal{X}_m$ and h^* are determined. Then we consider an equivalent way of determining the target function and the data distribution. That is, randomly selecting the position of x_i (dependent on the randomness of o, θ^*, s) and then considering the following two different processes of selecting s and o.

- Given a fixed x_i, randomly select $s \sim \mathcal{D}(s|x_i)$ and select $o \sim \mathcal{D}(o|s, x_i)$, where $\mathcal{D}(s|x_i)$ and $\mathcal{D}(o|s, x_i)$ denote the conditional distributions of s and o respectively. Note that when x_i satisfies $\|x_i\| \leq r - 2$, $\mathcal{D}(s|x_i) = \mathrm{Ber}(1/2)$ and $\mathcal{D}(o|s, x_i) = \mathrm{Unif}(\Gamma^2(x_i, r_i))$ is a uniform distribution over the circle with radius r_i centered at x_i, where r_i is the distance between o and x_i and is a constant according to the definition.

- Given (x_i, s, o) selected in the above process, if $\|x_i\| \leq r - 2$, we let $s' = 1 - s$ and $o' = \mathrm{Ref}_{L_{x_i}}(o)$ (where x_{i+1} and x_{i-1} is determined by (x_i, s, o)); otherwise we let $s' = s$

and $o' = o$. It is easy to check that the distribution of s' conditional on x_i is $\mathrm{Ber}(1/2)$ and the distribution of o' conditional on x_i is $\mathrm{Unif}(\Gamma^2(x_i, r_i))$ if $\|x_i\| \leq r-2$.

Therefore, the distributions of s and s' and the distributions of o and o' are the same given x_i respectively. Our following analysis depends on the event of $\|x_i\| \leq r-2$, the probability of which is $\mathbb{P}(\|x_i\| \leq r-2) \geq \mathbb{P}(\|o\| \leq r-3) = \frac{(r-3)^2}{r^2}$. We let $S_{\mathrm{trn}}(x_i, s, o)$ denote the training set by mapping I_{trn} to the positions determined by (x_i, s, o) and let $h^*(x_i, s, o)$ denote the target function determined by (x_i, s, o). Note that when $\|x_i\| \leq r-2$ and x_i is not in the training set, the poisoned data sets with the training sets generated in the above two different processes are the same, i.e., $S_{\mathrm{trn}}(x_i, s, o) \cup \mathrm{Adv}(h^*(x_i, s, o), S_{\mathrm{trn}}(x_i, s, o), x_i) = S_{\mathrm{trn}}(x_i, s', o') \cup \mathrm{Adv}(h^*(x_i, s', o'), S_{\mathrm{trn}}(x_i, s', o'), x_i)$. This is due to the symmetric property of the attacker. Hence, any algorithm will behave the same at point x_i no matter whether the training set is $S_{\mathrm{trn}}(x_i, s, o)$ or $S_{\mathrm{trn}}(x_i, s', o')$. In addition, the target functions produced in the two different processes classify x_i differently when $\|x_i\| \leq r-2$. Let $\mathcal{E}_2(x_i)$ denote the event of $\{\|x_i\| \leq r-2\}$ and $\mathcal{E}(h^*, \mathcal{A}, S_{\mathrm{trn}}, i)$ denote the event of $\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x_i), x_i) \neq h^*(x_i)$. Then for any $i \in I_u$, conditional on $\mathcal{E}_2(x_i)$, for any algorithm $\mathcal{A}$, we have

$$\mathbb{1}[\mathcal{E}(h^*(x_i, s, o), \mathcal{A}, S_{\mathrm{trn}}(x_i, s, o), i)] = \mathbb{1}[\neg\mathcal{E}(h^*(x_i, s', o'), \mathcal{A}, S_{\mathrm{trn}}(x_i, s', o'), i)]. \tag{5}$$

Similar to the proof in the case of $n = 3$, we have the expected attackable rate

$$\begin{aligned}
&\mathbb{E}_{h^*, s, S_{\mathrm{trn}} \sim \mathcal{D}^m}[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})] \\
\geq &\frac{1}{2m}\mathbb{E}_{o \sim \mathrm{Unif}(\{x:\|x\| \leq r\}), \theta^* \sim \mathrm{Unif}(2\pi), s \sim \mathrm{Ber}(1/2), S_{\mathrm{trn}} \sim \mathcal{D}^m, \mathcal{A}}\left[\sum_{x_i \in U} \mathbb{1}[\mathcal{E}(h^*, \mathcal{A}, S_{\mathrm{trn}}, i)]\right] \\
= &\frac{1}{2m}\mathbb{E}_{I_{\mathrm{trn}} \sim \mathrm{Unif}([2m])}\left[\sum_{i \in I_u} \mathbb{E}_{x_i, s, o, \mathcal{A}}[\mathbb{1}[\mathcal{E}(h^*(x_i, s, o), \mathcal{A}, S_{\mathrm{trn}}(x_i, s, o), i)]]\right] \\
\geq &\frac{1}{2m}\mathbb{E}_{I_{\mathrm{trn}}}\left[\sum_{i \in I_u} \mathbb{E}_{x_i}[\mathbb{E}_{s, o, \mathcal{A}}[\mathbb{1}[\mathcal{E}(h^*(x_i, s, o), \mathcal{A}, S_{\mathrm{trn}}(x_i, s, o), i)]|x_i]\,\mathbb{1}[\mathcal{E}_2(x_i)]]\right] \\
= &\frac{1}{4m}\left(\mathbb{E}_{I_{\mathrm{trn}}}\left[\sum_{i \in I_u} \mathbb{E}_{x_i}[\mathbb{E}_{s, o, \mathcal{A}}[\mathbb{1}[\mathcal{E}(h^*(x_i, s, o), \mathcal{A}, S_{\mathrm{trn}}(x_i, s, o), i)]|x_i]\,\mathbb{1}[\mathcal{E}_2(x_i)]]\right]\right. \\
&\left. + \mathbb{E}_{I_{\mathrm{trn}}}\left[\sum_{i \in I_u} \mathbb{E}_{x_i}\left[\mathbb{E}_{s', o', \mathcal{A}}\left[\mathbb{1}[\neg\mathcal{E}(h^*(x_i, s', o'), \mathcal{A}, S_{\mathrm{trn}}(x_i, s', o'), i)]|x_i\right]\,\mathbb{1}[\mathcal{E}_2(x_i)]\right]\right]\right) \qquad (6) \\
= &\frac{1}{4m}\mathbb{E}_{I_{\mathrm{trn}}}\left[\sum_{i \in I_u} \mathbb{E}_{x_i}[\mathbb{1}[\|x_i\| \leq r-2]]\right] \qquad (7) \\
\geq &\frac{m(r-3)^2}{4mr^2} \\
\geq &\frac{1}{64},
\end{aligned}$$

when $r \geq 4$. Here Eq. (6) holds due to Eq. (5) and Eq. (7) holds since the distributions of s and s' and the distributions of o and o' are the same given x_i respectively. Combining with Lemma 1, we

complete the proof. ■

Appendix C. Proof of Theorem 4

Proof The proof contains three steps. For notation simplicity, we sometimes use (β, b) to represent the linear classifier $h_{\beta,b}$. We say an angle β is consistent with a data set S if there exists $b \in [-2, 2]$ such that (β, b) is consistent with S. Also, for an fixed β, we say an offset b is consistent with S if (β, b) is consistent with S.

Step 1: For any fixed β, the probability mass of union of error region is bounded. For any β, if there exists any $b \in [-2, 2]$ such that (β, b) is consistent with S_{trn}, from this set of consistent b's, we denote by $b_{\sup}(\beta)$ the superior value of this set and $b_{\inf}(\beta)$ the inferior value of this set. By uniform convergence bound in PAC learning (Blumer et al., 1989), when $m \geq \frac{4}{\varepsilon'} \log \frac{2}{\delta} + \frac{24}{\varepsilon'} \log \frac{13}{\varepsilon'}$, we have that with probability at least $1 - \delta$, every linear classifier (β, b) consistent with S_{trn} has $\text{err}(h_{\beta,b}) \leq \varepsilon'$. Then the probability mass of the union of error region of all (β, b) consistent with S_{trn} for a fixed β is

$$
\begin{aligned}
&\mathcal{P}(\{x|\exists b \in [-2,2], h_{\beta,b}(x) \neq h^*(x), \forall (x', y') \in S_{\text{trn}}, y' = h_{\beta,b}(x')\}) \\
=&\mathcal{P}\Big(\bigcup_{b_{\inf}(\beta) \leq b \leq b_{\sup}(\beta)} \{x | b \text{ is consistent } \& \ h_{\beta,b}(x) \neq h^*(x)\}\Big) \\
=&\lim_{\delta \to 0^+} \mathcal{P}(\{x|h_{\beta, b_{\inf}(\beta)+\delta}(x) \neq h^*(x)\} \cup \{x|b_{\inf}(\beta) \text{ is consistent } \& \ h_{\beta,b_{\inf}(\beta)}(x) \neq h^*(x)\} \\
&\quad \cup \{x|h_{\beta, b_{\sup}(\beta)-\delta}(x) \neq h^*(x)\} \cup \{x|b_{\sup}(\beta) \text{ is consistent } \& \ h_{\beta,b_{\sup}(\beta)}(x) \neq h^*(x)\}) \\
\leq& 2\varepsilon' .
\end{aligned}
$$

If there does not exist any consistent $b \in [-2, 2]$ for β, then $\mathcal{P}(\{x|\exists b \in [-2, 2], h_{\beta,b}(x) \neq h^*(x), \forall (x', y') \in S_{\text{trn}}, y' = h_{\beta,b}(x')\}) = 0$.

Step 2: The binary-search path of β is unique and adding clean-label points can only change the depth of the search. That is, for any fixed target function (β^*, b^*), for $h - l = 2\pi, \pi, \frac{\pi}{2}, \ldots$, if $\frac{l+h}{2}$ is not consistent with the input (poisoned or not) data set S, then there cannot exist β consistent with the input data set S in both two intervals $(l, \frac{l+h}{2})$ and $(\frac{l+h}{2}, h)$. Since β^* is always consistent with S, only the interval containing β^* will contain β consistent with S. To prove this statement, assume that any $\beta \in \{l, h, \frac{l+h}{2}\}$ is not consistent with S and there exists (β_1, b_1) with $\beta_1 \in (l, \frac{l+h}{2})$ and (β_2, b_2) with $\beta_2 \in (\frac{l+h}{2}, h)$ consistent with S. If $\beta_2 - \beta_1 \leq \pi$, let $\beta_3 = \frac{l+h}{2}$; otherwise, let $\beta_3 = l$. Since (β_1, b_1) and (β_2, b_2) are consistent classifiers, for any $\alpha_1, \alpha_2 \geq 0$, $\mathbb{1}[(\alpha_1(\cos\beta_1, \sin\beta_1) + \alpha_2(\cos\beta_2, \sin\beta_2)) \cdot x + \alpha_1 b_1 + \alpha_2 b_2 \geq 0]$ is also a consistent classifier. By setting $\alpha_1 = \frac{\sin(\beta_2 - \beta_3)}{\sin(\beta_2 - \beta_1)}$ and $\alpha_2 = \frac{\sin(\beta_3 - \beta_1)}{\sin(\beta_2 - \beta_1)}$, we have $(\beta_3, \frac{\sin(\beta_2-\beta_3)b_1 + \sin(\beta_3-\beta_1)b_2}{\sin(\beta_2-\beta_1)})$ is consistent. If $\frac{\sin(\beta_2-\beta_3)b_1 + \sin(\beta_3-\beta_1)b_2}{\sin(\beta_2-\beta_1)} \in [-2, 2]$, this contradicts that any β_3 is not consistent with S; else, since $\mathcal{X} \subseteq \mathcal{B}^n(\mathbf{0}, 1)$, there must exist $b \in [-2, 2]$ such that (β_3, b) is consistent, which is a contradiction.

Step 3: When $h - l < \arctan(f(\varepsilon'')/2)$, the attackable rate caused by deeper search is at most $2\varepsilon''$. We consider two cases: $|b^*| > 1$ and $|b^*| \leq 1$. In the case of $|b^*| > 1$, the target function classifies $\mathcal{X}$ all positive or all negative and thus, there always exists a consistent b for $\beta = 0$. The

binary-search for β will not search in depth and output $(0, b)$ for some consistent b. Therefore, $\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A}) \leq \mathcal{P}(\{x|\exists b \in [-2, 2], h_{0,b}(x) \neq h^*(x), \forall (x', y') \in S_{\mathrm{trn}}, y' = h_{0,b}(x')\}) \leq 2\varepsilon'$.

In the case of $|b^*| \leq 1$, let $A_z \triangleq \{x|(\cos\beta^*, \sin\beta^*) \cdot x + b^* \in [-z, z]\}$ for $z \geq 0$. We will show that when $h - l \leq \arctan(z/2)$ for $z \leq 2$, the classifier boundary $K_{\beta,b}$ must lie in A_z, i.e., $K_{\beta,b} \triangleq \{x|(\cos\beta, \sin\beta) \cdot x + b = 0, \|x\| \leq 1\} \subseteq A_z$. This indicates that when $h - l \leq \arctan(z/2)$, the algorithm can only make mistakes at points inside A_z. If it is false, let $\{q_1, q_2\} = \{x|(\cos\beta^*, \sin\beta^*) \cdot x + b^* = 0\} \cap \Gamma(\mathbf{0}, 1)$ with $\beta^* - \theta((q_2 - q_1), e_1) = \frac{\pi}{2} \pmod{2\pi}$ if the intersection of the target function boundary and the unit circle has two different points; or let $q_1 = q_2 = q$ where $\{q\} = \{x|(\cos\beta^*, \sin\beta^*) \cdot x + b^* = 0\} \cap \Gamma(\mathbf{0}, 1)$ if intersection of the target function boundary and the unit circle has only one point. Similarly, we denote by $\{p_1, p_2\}$ the intersection of of $K_{\beta,b}$ and the unit circle. Since $h - l \leq \arctan(1) = \frac{\pi}{4}$, the input data set must contain both positive points and negative points and hence, $K_{\beta,b} \cap \Gamma(\mathbf{0}, 1)$ is not empty. We let $\{p_1, p_2\} = \{x|(\cos\beta, \sin\beta) \cdot x + b = 0\} \cap \Gamma(\mathbf{0}, 1)$ with $\beta - \theta((p_2 - p_1), e_1) = \frac{\pi}{2} \pmod{2\pi}$ if the intersection of the boundary of classifier (β, b) and the unit circle has two different points; or let $p_1 = p_2 = p$ where $\{p\} = \{x|(\cos\beta, \sin\beta) \cdot x + b = 0\} \cap \Gamma(\mathbf{0}, 1)$ if intersection of the boundary of classifier (β, b) and the unit circle has only one point. The definitions of q_1, q_2, p_1, p_2 are illustrated in Fig. 2. Then at least one of $\{p_1, p_2\}$ is not in A_z and w.l.o.g., assume that $p_2 \notin A_z$. It is easy to check that p_1, p_2 must lie on the same side of the boundary of the target function, otherwise $|\beta - \beta^*| = \theta((p_2 - p_1), (q_2 - q_1)) > \arctan(z/2)$, which contradicts $h - l \leq \arctan(z/2)$. Then there exists a consistent classifier $h' = \mathbb{1}[\langle w, x - p_2\rangle \geq 0]$ with $w = (q_{12} - p_{22}, p_{21} - q_{11})$ (whose boundary is the line passing q_1 and p_2). Then let β' denote the direction of h', i.e., $\cos\beta' = \frac{q_{12} - p_{22}}{\|w\|}$ and $\sin\beta' = \frac{p_{21} - q_{11}}{\|w\|}$, and we have $|\beta' - \beta^*| = \theta(p_2 - q_1, q_2 - q_1) > \arctan(z/2)$, which contradicts that $h - l \leq \arctan(z/2)$.

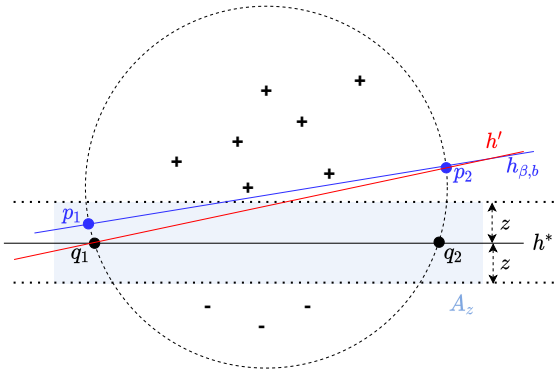

Figure 2: Illustration of p_1, p_2, q_1, q_2, h'.

We let $z = f(\varepsilon'')$. When $h - l \leq \arctan(f(\varepsilon'')/2)$, the classifier can only make mistakes inside $A_{f(\varepsilon'')}$ and thus the attackable rate is upper bounded by $2\varepsilon''$. In addition, when $h - l > \arctan(f(\varepsilon'')/2)$, the binary-search has searched to at most the $\left\lfloor \log_2(\frac{2\pi}{\arctan(f(\varepsilon'')/2)}) \right\rfloor + 1$-th depth, which leads to at most $2\varepsilon'(\left\lfloor \log_2(\frac{2\pi}{\arctan(f(\varepsilon'')/2)}) \right\rfloor + 1)$ attackable rate. Combining these results

together, when $m \geq \frac{4}{\varepsilon'} \log \frac{2}{\delta} + \frac{24}{\varepsilon'} \log \frac{13}{\varepsilon'}$, with probability at least $1-\delta$ over $S_{\text{trn}} \sim \mathcal{D}^m$, we have

$$\begin{aligned}
\text{atk}(h^*, S_{\text{trn}}, \mathcal{A}) \leq & 2\varepsilon' \left(\left\lfloor \log_2 \left(\frac{2\pi}{\arctan(f(\varepsilon'')/2)} \right) \right\rfloor + 1 \right) + 2\varepsilon'' \\
\leq & 2\varepsilon' \log_2 \left(\frac{4\pi}{\arctan(f(\varepsilon'')/2)} \right) + 2\varepsilon'' \\
\leq & \log_2 \left(\frac{4\pi}{\pi/8(f(\varepsilon'') \wedge 2)} \right) 2\varepsilon' + 2\varepsilon'' \qquad (8) \\
= & \log_2 \left(\frac{32}{f(\varepsilon'') \wedge 2} \right) 2\varepsilon' + 2\varepsilon'' ,
\end{aligned}$$

where Eq. (8) holds because $\arctan(x/2) \geq \pi/4$ when $x \geq 2$ and $\arctan(x/2) \geq \pi x/8$ when $x \in [0, 2]$. ■

Appendix D. Proof of Theorem 6

To prove Theorem 6, we first introduce a lemma on the behavior of uniform distribution on a unit sphere.

Lemma 2 (Lemma 2.2 by Ball et al. (1997)) *For any $a \in [0, 1)$, for $x \sim \text{Unif}(\Gamma^n(\mathbf{0}, 1))$, with probability at least $1 - e^{-na^2/2}$, we have $\langle x, e_1 \rangle \leq a$.*

Proof of Theorem 6 One essential hardness for robust learning of linear classifiers in high dimension is that for a fixed test instance x_0 on the sphere of a unit ball, with high probability over the selection of a set of training points from the uniform distribution over the sphere, every training instance has small component in the direction of x_0 as shown by Lemma 2. Taking advantage of this, the attacker can add a point (labeled differently from x_0) which is considerably closer to x_0 than all of the training instances, thus altering the behavior of SVM at x_0 as he wishes. In the following, we prove the theorem based on this idea. We divide the proof into three parts: a) the construction of the target function and the data distribution, b) the construction of the attacker and c) the analysis of the attackable rate.

The target function and the data distribution. The target function is $h^* = \mathbb{1}[\langle w^*, x \rangle \geq -\gamma/2]$ with $w^* = e_1$ and margin $\gamma = 1/8$. We define the marginal data distribution $\mathcal{D}_{\mathcal{X}}$ by putting probability mass $1 - 8\varepsilon$ on $-e_1$ and putting probability mass 8ε uniformly on the half sphere of a unit ball $\{x \,|\, \|x\| = 1, \langle x, e_1 \rangle \geq 0\}$. We let $\mathcal{D}^+$ denote the uniform distribution over this positive half sphere. We draw m i.i.d. training samples S_{trn} from $\mathcal{D}$ and then let S^+_{trn} denote the positive training samples. Let $m^+ = \left| S^+_{\text{trn}} \right|$ denote the number of positive training samples.

The attacker. For a given test instance $x_0 \in \{x \,|\, \|x\| = 1, \langle x, e_1 \rangle \geq 0\}$, we define two base vectors $v_1 = e_1$ and $v_2 = \frac{x_0 - \langle x_0, e_1 \rangle e_1}{\|x_0 - \langle x_0, e_1 \rangle e_1\|_2}$. Note that v_2 is well-defined almost surely. Then we define an attacker Adv as

$$\text{Adv}(h^*, S_{\text{trn}}, x_0) = \begin{cases} \{(-v_2, 1)\} & \text{if } m^+ = 0, \\ \{(-\gamma v_1 + \sqrt{1-\gamma^2} v_2, 0)\} & \text{else}. \end{cases}$$

Since $\langle w^*, v_2 \rangle = 0 > -\frac{\gamma}{2}$ and $\left\langle w^*, -\gamma v_1 + \sqrt{1-\gamma^2} v_2 \right\rangle = -\gamma < -\frac{\gamma}{2}$, Adv is a clean-label attacker.

Analysis. For $n \le 128$, if $m < \frac{1}{8\varepsilon} \vee \frac{e^{n/128}}{768\varepsilon} = \frac{1}{8\varepsilon}$ and $\varepsilon < 1/16$, then $\mathbb{P}\left(m^+ = 0\right) = (1-8\varepsilon)^m \ge \frac{1}{4}$. Furthermore, if $m^+ = 0$, SVM can only observe instance-label pairs of $(-v_1, 0)$ and $(-v_2, 1)$ and then output $\widehat{h}(x) = \mathbb{1}[\langle v_1 - v_2, x\rangle \ge 0]$. Therefore, if $\langle e_1, x_0\rangle < \frac{1}{\sqrt{2}}$, then x_0 is attackable. Therefore, for $m < \frac{1}{8\varepsilon}$, we have

$$\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathrm{SVM})\right] \ge 8\varepsilon \mathbb{P}_{x\sim\mathcal{D}^+}\left(\langle x, e_1\rangle < \frac{1}{\sqrt{2}}\right)\mathbb{P}\left(m^+ = 0\right) \ge \varepsilon\,.$$

For $n > 128$, if $m < \frac{1}{8\varepsilon}$, the analysis above works as well. Else suppose $\frac{1}{8\varepsilon} \le m \le \frac{e^{n/128}}{768\varepsilon}$, we know that $\mathbb{E}\left[m^+\right] = 8m\varepsilon$, thus by Chernoff bounds, we have $\mathbb{P}\left(m^+ > 32m\varepsilon\right) \le e^{-24m\varepsilon} \le e^{-3}$. Furthermore, by Lemma 2 and the union bound, drawing m_0 i.i.d. samples $S_0 \sim (\mathcal{D}^+)^{m_0}$, with probability at least $1 - 3m_0 e^{-n/128}$, every instance $x \in S_0$ satisfies $\langle x, v_1\rangle \le \frac{1}{8}$ and $\langle x, v_2\rangle \le \frac{1}{8}$. Let $\mathcal{E}$ denote the event of $\{\forall (x,y) \in S^+_{\mathrm{trn}}, \langle x, v_1\rangle \le \frac{1}{8}, \langle x, v_2\rangle \le \frac{1}{8}, 1 \le m^+ \le 32m\varepsilon\}$. If $\mathcal{E}$ holds, then there is a linear separator

$$\left(\sqrt{\frac{1+\gamma}{2}}v_1 - \sqrt{\frac{1-\gamma}{2}}v_2\right)^\top x + \frac{1}{2}\sqrt{\frac{1+\gamma}{2}} + \frac{1}{16}\sqrt{\frac{1-\gamma}{2}} \ge 0\,,$$

such that the distance between any point in $S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x_0)$ and the linear separator is no smaller than $\frac{3}{8} - \frac{\sqrt{7}}{64} \ge \frac{1}{4}$. Hence the distance between the points and the seperator output by SVM is also no smaller than $\frac{1}{4}$. When the test instance x_0 satisfies $\langle x_0, v_1\rangle \le \frac{1}{8}$, we have $\left\|x_0 - (-\gamma v_1 + \sqrt{1-\gamma^2}v_2)\right\| \le \frac{1}{4}$ and then x_0 is misclassified as negative by SVM. Hence, we have

$$\begin{aligned}
&\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathrm{SVM})\right] \\
\ge& 8\varepsilon \mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathbb{P}_{x\sim\mathcal{D}^+}\left(\mathrm{SVM}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \ne h^*(x)\right)\right] \\
\ge& 8\varepsilon \mathbb{E}_{x\sim\mathcal{D}^+, S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathbb{1}[\forall (x', y') \in S^+_{\mathrm{trn}}, \langle x', v_1\rangle \le \frac{1}{8}, \langle x', v_2\rangle \le \frac{1}{8}, \langle x, v_1\rangle \le \frac{1}{8}]\right] \\
\ge& 8\varepsilon \mathbb{E}_{x\sim\mathcal{D}^+}\left[\mathbb{E}_{S_{\mathrm{trn}}}\left[\mathbb{1}[\mathcal{E}]|x\right]\mathbb{1}[\langle x, v_1\rangle \le \frac{1}{8}]\right] \\
\ge& 8\varepsilon(1 - 2e^{-n/128})(1 - e^{-3} - e^{-8m\varepsilon})(1 - 96m\varepsilon e^{-n/128}) \\
\ge& \varepsilon\,,
\end{aligned}$$

when $\frac{1}{8\varepsilon} \le m \le \frac{e^{n/128}}{768\varepsilon}$ and $n > 128$. Thus in all we have shown that if $m < \frac{1}{8\varepsilon} \vee \frac{e^{n/128}}{768\varepsilon}$ then $\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}\left[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathrm{SVM})\right] > \varepsilon$. ■

Appendix E. Proof of Theorem 7

Proof of Theorem 7 The proof combines the idea of constructing a set of symmetrical poisoning instances in the proof of Theorem 2 and the idea that the training instances are far away from the test point in the proof of Theorem 6. Again, we divide the proof into three parts as we did in the previous proofs.

The target function and the data distribution. We denote every point in $\mathbb{R}^n$ by (x, z) for $x \in \mathbb{R}^{n-1}$ and $z \in \mathbb{R}$. The target function h^* is selected uniformly at random from $\mathcal{H}^*$, where $\mathcal{H}^* = \{\mathbb{1}[\langle (jw^*, 1), (x, z)\rangle \geq j\gamma/2] | j \in \{\pm 1\}, w^* \in \Gamma^{n-1}(\mathbf{0}, 1)\}$. Let $\gamma = \frac{1}{8}$. For target function $h^* = h_{w^*,j} = \mathbb{1}[\langle (jw^*, 1), (x, z)\rangle \geq j\gamma/2]$, the marginal data distribution $\mathcal{D}_{w^*,j,\mathcal{X}}$ puts probability mass $1 - 8\varepsilon$ on the point e_n, then put the remaining 8ε probability uniformly over the half sphere of a $(n-1)$-dimensional unit ball $\Gamma_{w^*,\gamma} \times \{0\}$, where $\Gamma_{w^*,\gamma} = \Gamma^{n-1}(\gamma w^*, 1) \cap \{x | \langle w^*, x\rangle \geq \gamma\}$. Then since every hypothesis in $\mathcal{H}^*$ predicts e_n positively, we only need to focus on the half sphere in the lower dimension. Note that the label of every point on the half sphere is determined by j. We sample $S_{\mathrm{trn}} \sim \mathcal{D}^m_{w^*,j}$ and $S_{\mathrm{trn},w^*,j}$ denote the samples on $\Gamma_{w^*,\gamma} \times \{0\} \times \mathcal{Y}$. Let $m_{w^*,j} = |S_{\mathrm{trn},w^*,j}|$.

The attacker. For any $u_1, u_2 \in \mathbb{R}^{n-1} \setminus \{\mathbf{0}\}$, let $K_{u_1}(u_2) = \{x | \|u_1\|^2 \langle u_2, x\rangle - \langle u_1, u_2\rangle \langle u_1, x\rangle = 0\}$ denote the homogeneous (passing through the origin) hyperplane perpendicular to the vector $u_2 - \langle u_1, u_2\rangle \frac{u_1}{\|u_1\|^2}$. For any given test instance $(x_0, 0)$ with $x_0 \in \Gamma_{w^*,\gamma}$, we define two base vectors $v_1 = w^*$ and $v_2 = \frac{x_0 - \langle x_0, w^*\rangle w^*}{\|x_0 - \langle x_0, w^*\rangle w^*\|}$. Note that v_2 is well-defined almost surely. Denote $K_{x_0} = K_{x_0}(v_1)$. Let $x_\parallel = \langle x, v_1\rangle v_1 + \langle x, v_2\rangle v_2$ denote x's component on the hyperplane defined by v_1, v_2 and $x_\perp = x - x_\parallel$ denote the component perpendicular to v_1, v_2 and then we define $\mathrm{Ref}_{K_{x_0}}(x) \triangleq x_\perp + \frac{2\langle x_\parallel, x_0\rangle}{\|x_0\|^2} x_0 - x_\parallel$ as the reflection of x through K_{x_0}. Then we define an attacker Adv as

$$\begin{aligned}
&\mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0)) \\
=&\begin{cases} \{((\mathrm{Ref}_{K_{x_0}}(x), 0), 1-y) | ((x, 0), y) \in S_{\mathrm{trn}}\} & \text{if } \mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j})\,, \\ \emptyset & \text{else}\,, \end{cases}
\end{aligned}$$

where

$$\mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j}) = \left\{\forall (x, 0) \in S_{\mathrm{trn},\mathcal{X}}, \langle x, w^*\rangle \leq \frac{1}{8} + \gamma, \left\langle x, \frac{x_0 - \langle x_0, w^*\rangle w^*}{\|x_0 - \langle x_0, w^*\rangle w^*\|}\right\rangle \leq \frac{1}{8}\right\}$$
$$\cap \left\{\langle x_0, w^*\rangle \leq \frac{1}{8} + \gamma\right\} \cap \{m_{w^*,j} \leq 32m\varepsilon\}\,.$$

Here $\mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j})$ is thought as a condition to attack x_0. Then we show that Adv is a clean-label attacker. If $\mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j})$ holds, we have

$$\begin{aligned}
&\left\langle \mathrm{Ref}_{K_{x_0}}(x), w^*\right\rangle \\
=&\left\langle x_\perp + \frac{2\left\langle x_\parallel, x_0\right\rangle}{\|x_0\|^2} x_0 - x_\parallel, v_1\right\rangle \\
=&\left\langle x - 2\left\langle x, v_1\right\rangle v_1 - 2\left\langle x, v_2\right\rangle v_2 + 2(\left\langle x, v_1\right\rangle \left\langle x_0, v_1\right\rangle + \left\langle x, v_2\right\rangle \left\langle x_0, v_2\right\rangle)\frac{x_0}{\|x_0\|^2}, v_1\right\rangle \\
=&-\left\langle x, v_1\right\rangle + 2\left\langle x, v_1\right\rangle \left\langle x_0, v_1\right\rangle^2 \frac{1}{\|x_0\|^2} + 2\left\langle x, v_2\right\rangle \left\langle x_0, v_2\right\rangle \left\langle x_0, v_1\right\rangle \frac{1}{\|x_0\|^2} \\
\leq&-\left\langle x, v_1\right\rangle + 2(\frac{1}{8} + \gamma)^2 \left\langle x, v_1\right\rangle + 2 \cdot \frac{1}{8}\left\langle x_0, v_1\right\rangle \\
\leq&\left(2(\frac{1}{8} + \gamma)^2 - 1\right)\gamma + \frac{1}{4}(\frac{1}{8} + \gamma) \\
<&0\,,
\end{aligned}$$

where the last inequality holds since $\gamma = \frac{1}{8}$. Therefore, $(\mathrm{Ref}_{K_{x_0}}(x), 0)$ is labeled different from $(x, 0)$ and Adv is a clean-label attacker.

Analysis. Observe that the probabilistic construction of the target function and the data distribution along with the random sampling of a test instance and the training set can be viewed in an equivalent way: first drawing the number of training samples on the half sphere m' from a binomial distribution $\mathrm{Bin}(m, 8\varepsilon)$; then on a fixed known half sphere, drawing a test instance and the training set with m' samples on the half sphere and $m - m'$ samples on e_n; and finally randomly selecting a coordinate system to decide the position of the true sphere and selecting a j to decide the labels of the training samples. Formally, let us fix a half sphere $\Gamma_+^{n-1} = \{x \in \Gamma^{n-1}(\mathbf{0}, 1) | \langle x, e_1 \rangle \geq 0\}$ and then sample $m' \sim \mathrm{Bin}(m, 8\varepsilon)$, $t_0 \sim \mathrm{Unif}(\Gamma_+^{n-1})$ and $Q_{\mathrm{trn}} \sim \mathrm{Unif}(\Gamma_+^{n-1})^{m'}$. We denote by $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$ the event of $\left\{\forall q \in Q_{\mathrm{trn}}, \langle q, e_1 \rangle \leq \frac{1}{8}, \left\langle q, \frac{t_0 - \langle t_0, e_1 \rangle e_1}{\|t_0 - \langle t_0, e_1 \rangle e_1\|} \right\rangle \leq \frac{1}{8}\right\} \cap \{\langle t_0, e_1 \rangle \leq \frac{1}{8}\} \cap \{m' = |Q_{\mathrm{trn}}| \leq 32m\varepsilon\}$. Then we sample $T_{n-1} \sim \mathrm{Unif}(O(n-1))$, where $O(n-1)$ is the orthogonal group. Finally we sample $j \sim \mathrm{Unif}(\{\pm 1\})$. We denote by R_{t_0} the linear isometry that reflects across the hyperplane $K_{\gamma e_1 + t_0}(e_1)$ in $\mathbb{R}^{n-1}$, i.e., $R_{t_0} u = \mathrm{Ref}_{K_{\gamma e_1 + t_0}(e_1)}(u)$, for any $u \in \mathbb{R}^n$.

Conditional on m', t_0 and Q_{trn} sampled in the above process, we consider two different coordinate systems and j's, which lead to two groups of random variables $(j, T_{n-1}, w^*, x_0, S_{\mathrm{trn},\mathcal{X}}, m_{w^*,j})$ and $(\widetilde{j}, \widetilde{T}_{n-1}, \widetilde{w}^*, \widetilde{x}_0, \widetilde{S}_{\mathrm{trn},\mathcal{X}}, \widetilde{m}_{\widetilde{w}^*,\widetilde{j}})$. Here for any random variable in the first group, we add a tilde to represent the corresponding random variable in the second group.

- In the first group, we have $j, T_{n-1}, w^* = T_{n-1}e_1, x_0 = T_{n-1}(\gamma e_1 + t_0), S_{\mathrm{trn},\mathcal{X}} = T_{n-1}(\gamma e_1 + Q_{\mathrm{trn}}) \times \{0\} \cup \{e_n\}^{m-m'}$ and $m_{w^*,j} = m'$.
- In the second group, we let $\widetilde{j} = -j, \widetilde{T}_{n-1} = T_{n-1}R_{t_0}, \widetilde{w}^* = \widetilde{T}_{n-1}e_1, \widetilde{x}_0 = \widetilde{T}_{n-1}(\gamma e_1 + t_0), \widetilde{S}_{\mathrm{trn},\mathcal{X}} = \widetilde{T}_{n-1}(\gamma e_1 + Q_{\mathrm{trn}}) \times \{0\} \cup \{e_n\}^{m-m'}$ and $\widetilde{m}_{\widetilde{w}^*,\widetilde{j}} = m'$.

The above two groups provide two ways of realizing the random process of selecting h^*, $(x_0, 0)$ and $S_{\mathrm{trn},\mathcal{X}}$: namely, $h^* = h_{w^*,j} = \mathbb{1}[\langle (jw^*, 1), (x, z) \rangle \geq j\gamma/2], (x_0, 0), S_{\mathrm{trn},\mathcal{X}}$ and $h^* = h_{\widetilde{w}^*,\widetilde{j}} = \mathbb{1}[\left\langle (\widetilde{j}\widetilde{w}^*, 1), (x, z) \right\rangle \geq \widetilde{j}\gamma/2], (\widetilde{x}_0, 0), \widetilde{S}_{\mathrm{trn},\mathcal{X}}$. Let $\widetilde{S}_{\mathrm{trn}} = \{(x, h_{\widetilde{w}^*,\widetilde{j}}(x)) | x \in \widetilde{S}_{\mathrm{trn},\mathcal{X}}\}$ denote the data set of instances in $\widetilde{S}_{\mathrm{trn},\mathcal{X}}$ labeled by $h_{\widetilde{w}^*,\widetilde{j}}$. Note that $(w^*, j, S_{\mathrm{trn}}, x_0)$ and $(\widetilde{w}^*, \widetilde{j}, \widetilde{S}_{\mathrm{trn}}, \widetilde{x}_0)$ are identical in distribution and that $x_0 = T_{n-1}(\gamma e_1 + t_0) = T_{n-1}R_{t_0}(\gamma e_1 + t_0) = \widetilde{x}_0$. We now argue that $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0))$ and $\widetilde{S}_{\mathrm{trn}} \cup \mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (\widetilde{x}_0, 0))$ are identical conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, \gamma e_1 + t_0)$. To prove this, we propose and prove the following three claims.

Claim I $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \Leftrightarrow \mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j}) \Leftrightarrow \mathcal{E}_1(\widetilde{w}^*, \widetilde{S}_{\mathrm{trn},\mathcal{X}}, \widetilde{x}_0, \widetilde{m}_{\widetilde{w}^*,\widetilde{j}})$.

Proof of Claim I This is true since $T_{n-1}, R_{t_0} \in O(n-1)$, thus they keep all the inner product properties. In particular, for any $(x, 0) \in S_{\mathrm{trn}}$, $x = T_{n-1}(\gamma e_1 + q)$ for some $q \in Q_{\mathrm{trn}}$ by definition of S_{trn}. Furthermore,

$$\langle x, w^* \rangle = \langle T_{n-1}(\gamma e_1 + q), T_{n-1}e_1 \rangle = \gamma + \langle q, e_1 \rangle \ .$$

Thus $\langle x, w^* \rangle \leq \frac{1}{8} + \gamma \Leftrightarrow \langle q, e_1 \rangle \leq \frac{1}{8}$. All the other equivalences can be derived similarly, thus omitted here. ■

Claim II For any homogeneous hyperplane $L_{u_1} \in \mathbb{R}^{n-1}$ with normal vector u_1, for any $u_2 \in \mathbb{R}^{n-1}$, we have $\mathrm{Ref}_{T_{n-1}L_{u_1}}(T_{n-1}u_2) = T_{n-1}\mathrm{Ref}_{L_{u_1}}(u_2)$.

Proof of Claim II We consider two cases. If $u_2 \in L_{u_1}$, then we have,

$$\mathrm{Ref}_{T_{n-1}L_{u_1}}(T_{n-1}u_2) = T_{n-1}u_2 = T_{n-1}\mathrm{Ref}_{L_{u_1}}(u_2).$$

Else if $u_2 \notin L_{u_1}$, we denote by $u_3 = \mathrm{Ref}_{T_{n-1}L_{u_1}}(T_{n-1}u_2)$. Thus u_3 is the only point such that $u_3 \neq T_{n-1}u_2$, $\langle u_3 - T_{n-1}u_2, T_{n-1}u_1\rangle = 0$ and $\|u_3\| = \|T_{n-1}u_2\|$. These immediately give us $T_{n-1}^\top u_3 \neq u_2$, $\left\langle T_{n-1}^\top u_3 - u_2, u_1\right\rangle = 0$ and $\left\|T_{n-1}^\top u_3\right\| = \|u_2\|$, which means $T_{n-1}^\top u_3 = \mathrm{Ref}_{L_{u_1}}(u_2)$. Thus

$$\mathrm{Ref}_{T_{n-1}L_{u_1}}(T_{n-1}u_2) = T_{n-1}T_{n-1}^\top u_3 = T_{n-1}\mathrm{Ref}_{L_{u_1}}(u_2)\,,$$

which completes the proof. ■

Claim III Conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$, the poisoned datasets $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0))$ and $\widetilde{S}_{\mathrm{trn}} \cup \mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (\widetilde{x}_0, 0))$ are identical.

Proof of Claim III Denote by $K = K_{\gamma e_1 + t_0}(e_1)$ the homogeneous hyperplane perpendicular to $e_1 - \langle \gamma e_1 + t_0, e_1\rangle \frac{\gamma e_1 + t_0}{\|\gamma e_1 + t_0\|^2}$. Thus we have,

$$\begin{aligned}
T_{n-1}K &= \left\{ T_{n-1}x \middle| \left\langle x, e_1 - \langle \gamma e_1 + t_0, e_1\rangle \frac{\gamma e_1 + t_0}{\|\gamma e_1 + t_0\|^2} \right\rangle = 0 \right\} \\
&= \left\{ x \middle| \left\langle x, T_{n-1}\left(e_1 - \langle \gamma e_1 + t_0, e_1\rangle \frac{\gamma e_1 + t_0}{\|\gamma e_1 + t_0\|^2} \right) \right\rangle = 0 \right\} \\
&= \left\{ x \middle| \left\langle x, w^* - \langle x_0, w^*\rangle \frac{x_0}{\|x_0\|^2} \right\rangle = 0 \right\} \\
&= K_{x_0}(w^*)\,.
\end{aligned}$$

By Claim I, we know that $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$ and $\mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j})$ are equivalent. Thus, conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$,

$$\begin{aligned}
\mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0)) &= \{((\mathrm{Ref}_{K_{x_0}(w^*)}(x), 0), 1-j) | ((x,0), j) \in S_{\mathrm{trn}}\} \\
&= \{((\mathrm{Ref}_{T_{n-1}K}(T_{n-1}(\gamma e_1 + q)), 0), 1-j) | q \in Q_{\mathrm{trn}}\} \\
&= \{((T_{n-1}\mathrm{Ref}_K(\gamma e_1 + q), 0), 1-j) | q \in Q_{\mathrm{trn}}\} \qquad (9) \\
&= \{((T_{n-1}R_{t_0}(\gamma e_1 + q), 0), 1-j) | q \in Q_{\mathrm{trn}}\} \\
&= \widetilde{S}_{\mathrm{trn}} \setminus \{(e_n, 1)\}^{m-m'}\,,
\end{aligned}$$

where Eq. (9) holds by applying Claim II. Similarly, for $\mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (x_0, 0))$, the plane of reflection is $K_{x_0}(\widetilde{w}^*) = K_{\widetilde{T}_{n-1}(\gamma e_1 + t_0)}(\widetilde{T}_{n-1}e_1) = \widetilde{T}_{n-1}K_{\gamma e_1 + t_0}(e_1) = \widetilde{T}_{n-1}K$ and

$$\begin{aligned}
\mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (x_0, 0)) &= \{((\mathrm{Ref}_{K_{x_0}(\widetilde{w}^*)}(x), 0), 1-j) | ((x,0), \widetilde{j}) \in \widetilde{S}_{\mathrm{trn}}\} \\
&= \{((\mathrm{Ref}_{\widetilde{T}_{n-1}K}(\widetilde{T}_{n-1}(\gamma e_1 + q)), 0), j) | q \in Q_{\mathrm{trn}}\} \\
&= \{((T_{n-1}(\gamma e_1 + q), 0), j) | q \in Q_{\mathrm{trn}}\} \qquad (10) \\
&= S_{\mathrm{trn}} \setminus \{(e_n, 1)\}^{m-m'}\,,
\end{aligned}$$

where Eq. (10) holds by applying Claim II and $\widetilde{T}_{n-1} = T_{n-1}R_{t_0}$. Thus,

$$\begin{aligned} S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0)) &= S_{\mathrm{trn}} \cup \widetilde{S}_{\mathrm{trn}} \setminus \{(e_n, 1)\}^{m-m'} \\ &= \widetilde{S}_{\mathrm{trn}} \cup \mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (\widetilde{x}_0, 0))\,. \end{aligned}$$

■

Now we have proved that $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0))$ and $\widetilde{S}_{\mathrm{trn}} \cup \mathrm{Adv}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (\widetilde{x}_0, 0))$ are identical conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, \gamma e_1 + t_0)$. Hence in this case any algorithm will behave the same given the input data being either S_{trn} or $\widetilde{S}_{\mathrm{trn}}$. Let $\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h^*, x_0)$ denote the event $\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}},\ (x_0, 0)), (x_0, 0)) \neq h^*((x_0, 0))$. Since $h_{w^*,j}((x_0, 0)) \neq h_{\widetilde{w}^*,\widetilde{j}}((x_0, 0))$, then conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$, for any algorithm $\mathcal{A}$, we have

$$\mathbb{1}[\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)] = \mathbb{1}[\neg\mathcal{E}_2(\mathcal{A}, \widetilde{S}_{\mathrm{trn}}, \mathrm{Adv}, h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{x}_0)]\,.$$

If $m < \frac{1}{8\varepsilon}$, then we have $\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]] \geq \mathbb{P}(m' = 0) = (1-8\varepsilon)^m > \frac{1}{4}$ when $\varepsilon \leq 1/16$. Else since $\mathbb{E}[m'] = 8m\varepsilon$, by Chernoff bounds, we have $\mathbb{P}(m' > 32m\varepsilon) \leq e^{-24m\varepsilon} \leq e^{-3}$. Furthermore, by Lemma 2 and the union bound, drawing m' i.i.d. samples $S_0 \sim \mathrm{Unif}(\Gamma_+^{n-1} \times \{0\})^{m'}$, with probability at least $1 - 3m'e^{-\frac{n-1}{128}}$, every $(x, 0) \in S_0$ satisfy $\langle x, e_1\rangle \leq \frac{1}{8}$ and $\left\langle x, \frac{t_0 - \langle t_0, e_1\rangle e_1}{\|t_0 - \langle t_0, e_1\rangle e_1\|}\right\rangle \leq \frac{1}{8}$. Thus in all, we have, for any algorithm $\mathcal{A}$,

$$\begin{aligned}
&\mathbb{E}_{w^*,j,S_{\mathrm{trn}}\sim\mathcal{D}^m_{w^*,j},(x,y)\sim\mathcal{D}_{w^*,j},\mathcal{A}}[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})] \\
=&\mathbb{E}_{w^*,j,S_{\mathrm{trn}}\sim\mathcal{D}^m_{w^*,j},(x,y)\sim\mathcal{D}_{w^*,j},\mathcal{A}}[\mathbb{1}[\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h^*, S_{\mathrm{trn}}, x), x) \neq h^*(x)]] \\
\geq&8\varepsilon\mathbb{E}_{w^*,j,S_{\mathrm{trn}},x_0\sim\mathrm{Unif}(\Gamma_{w^*,\gamma}),\mathcal{A}}[\mathbb{1}[\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)] \\
&\cdot \mathbb{1}[\mathcal{E}_1(w^*, S_{\mathrm{trn}}, x_0, m_{w^*,j}) \cup \{m_{w^*,j} = 0\}]] \\
=&8\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]\mathbb{E}_{T_{n-1},j,\mathcal{A}}[\mathbb{1}[\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)]]] \\
=&4\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]\mathbb{E}_{T_{n-1},j,\mathcal{A}}[\mathbb{1}[\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)]]] \\
&+ 4\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]\mathbb{E}_{T_{n-1},j,\mathcal{A}}[\mathbb{1}[\neg\mathcal{E}_2(\mathcal{A}, \widetilde{S}_{\mathrm{trn}}, \mathrm{Adv}, h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{x}_0)]]] \quad (11) \\
=&4\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]\mathbb{E}_{T_{n-1},j,\mathcal{A}}[\mathbb{1}[\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)]]] \\
&+ 4\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]\mathbb{E}_{T_{n-1},j,\mathcal{A}}[\mathbb{1}[\neg\mathcal{E}_2(\mathcal{A}, S_{\mathrm{trn}}, \mathrm{Adv}, h_{w^*,j}, x_0)]]] \quad (12) \\
=&4\varepsilon\mathbb{E}_{t_0,m',Q_{\mathrm{trn}}}[\mathbb{1}[\mathcal{E}_3(Q_{\mathrm{trn}}, t_0) \cup \{m' = 0\}]] \\
\geq&\begin{cases} 4\varepsilon(1 - 2e^{-\frac{n-1}{128}})(1 - e^{-3})(1 - 96m\varepsilon e^{-\frac{n-1}{128}}) & \text{when } m \geq \frac{1}{8\varepsilon} \\ 4\varepsilon(1-8\varepsilon)^m & \text{when } m < \frac{1}{8\varepsilon} \end{cases} \\
>&\varepsilon,
\end{aligned}$$

when $m \leq \frac{e^{\frac{n-1}{128}}}{192\varepsilon}$ and $n \geq 257$. Here Eq. (11) holds due to the fact that Adv will make $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{w^*,j}, S_{\mathrm{trn}}, (x_0, 0))$ and $\widetilde{S}_{\mathrm{trn}} \cup S_{\mathrm{Adv}}(h_{\widetilde{w}^*,\widetilde{j}}, \widetilde{S}_{\mathrm{trn}}, (\widetilde{x}_0, 0))$ identical conditional on $\mathcal{E}_3(Q_{\mathrm{trn}}, t_0)$ and Eq. (12) holds because $(w, j, S_{\mathrm{trn}}, x_0)$ is identical to $(\widetilde{w}^*, \widetilde{j}, \widetilde{S}_{\mathrm{trn}}, \widetilde{x}_0)$ in distribution. Thus in all, we have shown that for $n \geq 256$, if $m \leq \frac{e^{\frac{n-1}{128}}}{192\varepsilon}$ then for all algorithm $\mathcal{A}$, the expected attackable rate is $\mathbb{E}_{w,j,S_{\mathrm{trn}}\sim\mathcal{D}^m}[\mathrm{atk}(h^*, S_{\mathrm{trn}}, \mathcal{A})] > \varepsilon$. Thus there exists a target function $h^* \in \mathcal{H}$ and a distribution $\mathcal{D}$ over D_{h^*} with margin $\gamma = 1/8$ such that $\mathbb{E}_{S_{\mathrm{trn}}\sim\mathcal{D}^m}[\mathrm{atk}_{\mathcal{D}}(h^*, S_{\mathrm{trn}}, \mathcal{A})] > \varepsilon$. ■

Appendix F. Proof of Theorem 8

We first introduce two lemmas for the proof of the theorem.

Lemma 3 *For any hypothesis class $\mathcal{H}$ with finite VC dimensional d, any distribution $\mathcal{D}$, with probability at least $1-\delta$ over $S_{\text{trn}} \sim \mathcal{D}^m$, for all $h \in \mathcal{H}$,*

$$\text{err}(h) - \text{err}_{S_{\text{trn}}}(h) \leq \sqrt{\frac{18d(1-\text{err}_{S_{\text{trn}}}(h))\text{err}_{S_{\text{trn}}}(h)\ln(em/\delta)}{m-1}} + \frac{15d\ln(em/\delta)}{m-1}\,.$$

Proof This lemma is a direct result of empirical Bennett's inequality (Theorem 6 by Maurer and Pontil (2009)) and Sauer's lemma. Let $\Lambda(\cdot)$ denote the growth function of $\mathcal{H}$. Then by empirical Bennett's inequality (Theorem 6 by Maurer and Pontil (2009)), we have with probability at least $1-\delta$ over $S_{\text{trn}} \sim \mathcal{D}^m$,

$$\text{err}(h) - \text{err}_{S_{\text{trn}}}(h) \leq \sqrt{\frac{18(1-\text{err}_{S_{\text{trn}}}(h))\text{err}_{S_{\text{trn}}}(h)\ln(\Lambda(m)/\delta)}{m-1}} + \frac{15\ln(\Lambda(m)/\delta)}{m-1}, \forall h \in \mathcal{H}\,.$$

By Sauer's lemma, $\Lambda(m) \leq (\frac{em}{d})^d$, which completes the proof. ■

Lemma 4 *For any hypothesis class $\mathcal{H}$ with finite VC dimensional d, a fixed data set S with m elements, realizable by some $h^* \in \mathcal{H}$. Let S_0 be a set with size $m_0 < m$ drawn from S uniformly at random without replacement. Then with probability at least $1-\delta$, for all $h \in \mathcal{H}$ with $\text{err}_{S_0}(h) = 0$, we have*

$$\text{err}_S(h) \leq \frac{d\ln(em/d) + \ln(1/\delta)}{m_0}\,.$$

Proof For any $h \in \mathcal{H}$, we have

$$\mathbb{P}\left(\text{err}_S(h) > \varepsilon, \text{err}_{S_0}(h) = 0\right) \leq \frac{\binom{m-k}{m_0}}{\binom{m}{m_0}} \leq (1-k/m)^{m_0}\,,$$

where $k = \lceil \varepsilon \cdot m \rceil$. By Sauer's lemma, $\Lambda(m) \leq (\frac{em}{d})^d$. Taking the union bound completes the proof. ■

Proof of Theorem 8 Let $m = |S_{\text{trn}}|$ be the number of training samples. Let $h_i = \mathcal{L}(S^{(i)})$ denote the output hypothesis of block i. Let $N_c = \{i_1, \ldots, i_{n_c}\} \subseteq [10t+1]$ denote the set of index of non-contaminated blocks without poisoning points with $n_c = |N_c|$. Each block has $m_0 = \left\lfloor \frac{m}{10t+1} \right\rfloor$ or $m_0 = \left\lceil \frac{m}{10t+1} \right\rceil$ data points (dependent on the actual number of poison points injected by the attacker) and at least $9t+1$ blocks do not contain any poison points, i.e., $n_c \geq 9t+1$. If a point x is predicted incorrectly, then it is predicted incorrectly by more than $4t+1$ non-contaminated classifiers. Given training data S_{trn}, for any $x \in \mathcal{X}$, any m_0 and any t-point attacker Adv to make

each block has m_0 points, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(S_{\mathrm{trn}}, h^*, x), x) \neq h^*(x)\right) \\
=&\mathbb{P}_{\mathcal{A}}\left(\sum_{i=1}^{10t+1} \mathbb{1}[h_i(x) \neq h^*(x)] \geq 5t+1\right) \\
\leq&\mathbb{P}_{\mathcal{A}}\left(\sum_{i \in N_c} \mathbb{1}[h_i(x) \neq h^*(x)] \geq 4t+1\right) \\
\leq&\frac{1}{4t+1}\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{i \in N_c} \mathbb{1}[h_i(x) \neq h^*(x)] \middle| N_c\right]\right] \\
\leq&2.5\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[h_{i_1}(x) \neq h^*(x)] | N_c\right]\right] .
\end{aligned}$$

Notice here, if m_0 is fixed, the randomness of $\mathcal{A}$ can be regarded as selecting N_c first and drawing $n_c m_0$ samples uniformly at random from S_{trn} without replacement to construct $S^{(i_1)}, S^{(i_2)} \ldots, S^{(i_{n_c})}$. More specifically, conditioned on N_c, the randomness of $\mathcal{A}$ on h_{i_1} is only through drawing $S^{(i_1)}$, i.e., drawing m_0 samples without replacement from the clean training examples S_{trn}. The important thing is that if m_0 is fixed, this distribution does not depend on the attacker. Hence,

$$\begin{aligned}
&\mathrm{atk}(t, h^*, S_{\mathrm{trn}}, \mathcal{A}) \\
\leq&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\mathrm{Adv}} \mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(S_{\mathrm{trn}}, h^*, x), x) \neq h^*(x)\right)\right] \\
\leq&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\mathrm{Adv}: m_0 = \lfloor \frac{m}{10t+1} \rfloor} \mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(S_{\mathrm{trn}}, h^*, x), x) \neq h^*(x)\right)\right] \\
&+ \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\mathrm{Adv}: m_0 = \lceil \frac{m}{10t+1} \rceil} \mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(S_{\mathrm{trn}}, h^*, x), x) \neq h^*(x)\right)\right] \\
\leq&2.5\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\mathrm{Adv}: m_0 = \lfloor \frac{m}{10t+1} \rfloor} \mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[h_{i_1}(x) \neq h^*(x)] | N_c\right]\right]\right] \qquad (13) \\
&+ 2.5\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{\mathrm{Adv}: m_0 = \lceil \frac{m}{10t+1} \rceil} \mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[h_{i_1}(x) \neq h^*(x)] | N_c\right]\right]\right] \qquad (14) \\
\leq&2.5\mathbb{E}_{N_c}\left[\mathbb{E}_{S^{(i_1)}}\left[\mathrm{err}\left(h_{i_1}\right) \middle| N_c\right]; m_0 = \left\lfloor \frac{m}{10t+1} \right\rfloor\right] \\
&+ 2.5\mathbb{E}_{N_c}\left[\mathbb{E}_{S^{(i_1)}}\left[\mathrm{err}\left(h_{i_1}\right) \middle| N_c\right]; m_0 = \left\lceil \frac{m}{10t+1} \right\rceil\right] . \qquad (15)
\end{aligned}$$

In the following, we will bound the error of h_{i_1} for each value of m_0. Let $\mathcal{E}$ denote the event of $\mathrm{err}_{S_{\mathrm{trn}}}(h_{i_1}) \leq \frac{(d+1)\ln(em/d)}{m_0}$ and by Lemma 4 we have $\mathbb{P}_{\mathcal{A}}\left(\neg\mathcal{E}|N_c\right) \leq \frac{d}{em}$. Then with probability

at least $1-\delta$ over the choice of S_{trn}, for each fixed value of m_0, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})|N_c\right] \\
=&\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})\mathbb{1}[\mathcal{E}]|N_c\right]+\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})\mathbb{1}[\neg\mathcal{E}]|N_c\right] \\
=&\mathbb{E}_{\mathcal{A}}\left[(\mathrm{err}(h_{i_1})-\mathrm{err}_{S_{\mathrm{trn}}}(h_{i_1})+\mathrm{err}_{S_{\mathrm{trn}}}(h_{i_1}))\mathbb{1}[\mathcal{E}]|N_c\right]+\mathbb{P}_{\mathcal{A}}\left(\neg\mathcal{E}|N_c\right) \\
\leq&\sqrt{\frac{18d\ln(em/\delta)(d+1)\ln(em/d)}{(m-1)m_0}}+\frac{15d\ln(em/\delta)}{m-1}+\frac{(d+1)\ln(em/d)}{m_0}+\frac{d}{em} \qquad (16)\\
\leq&\frac{1}{2}\left(\frac{6d\ln(em/\delta)}{m-1}+\frac{6d\ln(em/d)}{m_0}\right)+\frac{15d\ln(em/\delta)}{m-1}+\frac{(d+1)\ln(em/d)}{m_0}+\frac{d}{em} \\
\leq&\frac{24d\ln(em)}{m_0}+\frac{19d\ln(1/\delta)}{10tm_0}, \qquad (17)
\end{aligned}$$

where Eq. (16) applies Lemma 3. Then when $m_0 \geq \frac{960d}{\varepsilon}\ln\frac{2640etd}{\varepsilon}+\frac{19d\ln(1/\delta)}{\varepsilon t}$, we have that $\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})|N_c\right] \leq \frac{24d\ln(em)}{m_0}+\frac{19d\ln(1/\delta)}{10tm_0} \leq 0.2\varepsilon$. Combined with Eq. (13), we have that when $m \geq (10t+1)(\frac{960d}{\varepsilon}\ln\frac{2640etd}{\varepsilon}+\frac{19d\ln(1/\delta)}{\varepsilon t}+1)$, the t-point attackable rate is $\mathrm{atk}(t,h^*,S_{\mathrm{trn}},\mathcal{A}) \leq \varepsilon$. ■

Appendix G. Proofs and discussions for (t,ε,δ)-robust proper learners

G.1. Proof of Theorem 9

Proof Similar to the proof of Theorem 8, we let $m=|S_{\mathrm{trn}}|$ be the number of training samples, and let $N_c=\{i_1,\ldots,i_{n_c}\}\subseteq[10tk_p+1]$ denote the set of index of blocks without poisoning points. Each block has $m_0=\left\lfloor\frac{m}{10tk_p+1}\right\rfloor$ or $m_0=\left\lceil\frac{m}{10tk_p+1}\right\rceil$ data points and at least $t(10k_p-1)+1$ blocks do not contain any attacking points, i.e., $n_c \geq t(10k_p-1)+1$. Given any fixed $x\in\mathcal{X}$, if $\sum_{i=1}^{10tk_p+1}\mathbb{1}[h_i(x)\neq h^*(x)] \leq 10t < \frac{10tk_p+1}{k_p}$, then $x\in\mathcal{X}_{\mathcal{H}',k_p}$, thus $\widehat{h}(x)=\mathrm{Major}(\mathcal{H}',x)=h^*(x)$. Thus we have, given training data S_{trn}, for any $x\in\mathcal{X}$, any m_0 and any t-point attacker Adv to make each block has m_0 points, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}}\cup\mathrm{Adv}(S_{\mathrm{trn}},h^*,x),x)\neq h^*(x)\right) \\
\leq&\mathbb{P}_{\mathcal{A}}\left(\sum_{i=1}^{10tk_p+1}\mathbb{1}[h_i(x)\neq h^*(x)]\geq 10t+1\right) \\
\leq&\mathbb{P}_{\mathcal{A}}\left(\sum_{i\in N_c}\mathbb{1}[h_i(x)\neq h^*(x)]\geq 9t+1\right) \\
\leq&\frac{1}{9t+1}\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{i\in N_c}\mathbb{1}[h_i(x)\neq h^*(x)]\middle|N_c\right]\right] \\
\leq&\frac{10}{9}k_p\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[h_{i_1}(x)\neq h^*(x)]\middle|N_c\right]\right],
\end{aligned}$$

which indicates

$$\begin{aligned}\mathrm{atk}(t,h^*,S_{\mathrm{trn}},\mathcal{A}) \le \frac{10}{9}k_p\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})\,|N_c\right];m_0=\left\lfloor\frac{m}{10tk_p+1}\right\rfloor\right]\\ +\frac{10}{9}k_p\mathbb{E}_{N_c}\left[\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})\,|N_c\right];m_0=\left\lceil\frac{m}{10tk_p+1}\right\rceil\right].\end{aligned}$$

Then we bound $\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})|N_c\right]$ in the same way as the proof of Theorem 8. Following the same calculation process of Eq. (17), we have with probability at least $1-\delta$, $\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})|N_c\right] \le \frac{24d\ln(em)}{m_0}+\frac{19d\ln(1/\delta)}{10k_ptm_0}$ by using Lemma 4 and Lemma 3. Then when $m_0 \ge \frac{960dk_p}{\varepsilon}\ln\frac{2640etdk_p^2}{\varepsilon}+\frac{19d\ln(1/\delta)}{\varepsilon t}$, we have $\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_{i_1})|N_c\right] \le \frac{24d\ln(em)}{m_0}+\frac{19d\ln(1/\delta)}{10k_ptm_0} \le \frac{0.2}{k_p}\varepsilon$. Therefore, we have that when $m \ge (10k_pt+1)(\frac{960dk_p}{\varepsilon}\ln\frac{2640etdk_p^2}{\varepsilon}+\frac{19d\ln(1/\delta)}{\varepsilon t}+1)$, the t-point attackable rate is $\mathrm{atk}(t,h^*,S_{\mathrm{trn}},\mathcal{A}) \le \varepsilon$. ■

G.2. A proper learner for hypothesis class with no limitation over k_p

Algorithm 5 A robust proper learner for t-point attacker

1: **input**: a proper ERM learner $\mathcal{L}$, data S
2: uniformly at random pick $\left\lfloor\frac{|S|}{3t/\varepsilon}\right\rfloor$ points S_0 from S *with replacement*
3: **return** $\mathcal{L}(S_0)$

Theorem 15 *For any hypothesis class with VC dimension d, with any proper ERM learner $\mathcal{L}$, Algorithm 5 can (t,ε,δ)-robustly learn $\mathcal{H}$ using m samples where*

$$m = O\left(\frac{dt}{\varepsilon^2}\log\frac{d}{\varepsilon}+\frac{d}{\varepsilon}\log\frac{1}{\delta}\right).$$

Proof Let $\mathcal{E}$ denote the event that every point in S_0 is selected from the training data S_{trn}. Let $m=|S_{\mathrm{trn}}|$ and $h_0=\mathcal{L}(S_0)$. Let us denote the size of S_0 by $m_0=\left\lfloor\frac{|S|}{3t/\varepsilon}\right\rfloor$, which can be $\left\lfloor\frac{m}{3t/\varepsilon}\right\rfloor$ or $\left\lceil\frac{m}{3t/\varepsilon}\right\rceil$. Since $\mathbb{P}_{\mathcal{A}}(\mathcal{E}) \ge (1-\frac{t}{3tm_0/\varepsilon})^{m_0} \ge 1-\frac{\ln 4}{3}\varepsilon$, we have $\mathbb{P}_{\mathcal{A}}(\neg\mathcal{E}) \le \frac{\ln 4}{3}\varepsilon$. Then for any t-point attacker Adv, we have

$$\begin{aligned}&\mathbb{P}_{\mathcal{A}}\left(\mathcal{A}(S_{\mathrm{trn}}\cup\mathrm{Adv}(h^*,S_{\mathrm{trn}},x),x)\neq h^*(x)\right)\\ =&\mathbb{P}_{\mathcal{A}}\left(h_0(x)\neq h^*(x)\cap\mathcal{E}\right)+\mathbb{P}_{\mathcal{A}}\left(h_0(x)\neq h^*(x)\cap\neg\mathcal{E}\right)\\ \le&\mathbb{P}_{\mathcal{A}}\left(h_0(x)\neq h^*(x)|\mathcal{E}\right)+\mathbb{P}_{\mathcal{A}}\left(\neg\mathcal{E}\right)\\ \le&\mathbb{P}_{\mathcal{A}}\left(h_0(x)\neq h^*(x)|\mathcal{E}\right)+\frac{\ln 4}{3}\varepsilon,\end{aligned}$$

which indicates $\mathrm{atk}(t,h^*,S_{\mathrm{trn}},\mathcal{A}) \le \mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)\middle|\mathcal{E};m_0=\left\lfloor\frac{m}{3t/\varepsilon}\right\rfloor\right]+\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)\middle|\mathcal{E};m_0=\left\lceil\frac{m}{3t/\varepsilon}\right\rceil\right]$ $+\frac{\ln 4}{3}\varepsilon$. Conditioned on $\mathcal{E}$, S_0 is a set of i.i.d. samples uniformly drawn from S_{trn}. By classic uniform convergence bound, $\mathrm{err}_{S_{\mathrm{trn}}}(h_0) \le \frac{2}{m_0}(d\log(2em_0/d)+\log(2/\delta_0))$ with probability at

least $1-\delta_0$ over the choice of S_0 (for a fixed $S_{\rm trn}$). Let $\mathcal{E}_1$ denote the event of $\mathrm{err}_{S_{\rm trn}}(h_0) \le \frac{2}{m_0}(d+1)\log(2em_0/d)$ and it is easy to check that $\mathbb{P}_{\mathcal{A}}(\neg\mathcal{E}_1) \le \frac{d}{em_0}$. Similar to the proof of Theorem 8, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)|\mathcal{E}\right] \\
=&\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)\mathbb{1}[\mathcal{E}_1]|\mathcal{E}\right] + \mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)\mathbb{1}[\neg\mathcal{E}_1]|\mathcal{E}\right] \\
\le&\sqrt{\frac{36d\ln(em/\delta)(d+1)\log(2em_0/d)}{(m-1)m_0}} + \frac{15d\ln(em/\delta)}{m-1} + \frac{2}{m_0}(d+1)\log\frac{2em_0}{d} + \frac{d}{em_0} \qquad (18)\\
\le&\frac{1}{2}\left(\frac{6d\ln(em/\delta)}{m-1} + \frac{12d\ln(2em_0/d)}{m_0}\right) + \frac{15d\ln(em/\delta)}{m-1} + \frac{6d\ln(2em_0/d)}{m_0} + \frac{d}{em_0} \\
\le&\frac{13d\ln(2em_0/d)}{m_0} + \frac{18d\ln(em/\delta)}{m-1} \\
\le&\frac{31d\ln(2em_0)}{m_0} + \frac{18d\ln(1/\delta)}{(3t/\varepsilon-1)m_0},
\end{aligned}$$

where Eq. (18) adopts Lemma 3. When $m_0 \ge \frac{1120d}{\varepsilon}\ln\frac{560ed}{\varepsilon} + \frac{72d\ln(1/\delta)}{t}$, $\mathbb{E}_{\mathcal{A}}\left[\mathrm{err}(h_0)|\mathcal{E}\right] \le 0.25\varepsilon$. Hence, with probability at least $1-\delta$, the t-point attackable rate is $\mathrm{atk}(t, h^*, S_{\rm trn}, \mathcal{A}) \le \varepsilon$ by using m training samples where

$$m = \frac{3t}{\varepsilon}\left(\frac{1120d}{\varepsilon}\ln\frac{560ed}{\varepsilon} + \frac{72d\ln(1/\delta)}{t} + 1\right).$$

■

Appendix H. Proof of Theorem 10

Proof of Theorem 10 Now we show that for any sample size $m > 0$, there exists a hypothesis class $\mathcal{H}$ with VC dimension $5d$, a target function $h^* \in \mathcal{H}$ and a data distribution $\mathcal{D}$ on D_{h^*} such that $\mathbb{E}_{S_{\rm trn}\sim\mathcal{D}^m}\left[\mathrm{atk}(t, h^*, S_{\rm trn}, \mathcal{A})\right] \ge \min(\frac{3td}{64m}, \frac{3}{8})$. We start with proving this statement in the base case of $d = 1$ and then extend it to $d \ge 2$. We divide the proof into four parts: a) construction of the hypothesis class, the target function and the data distribution in $d = 1$, b) computation of the VC dimension of the hypothesis class, c) construction of the attacker, and d) generalization to $d \ge 1$.

The hypothesis class, the target function and the data distribution. We denote by $\Gamma = \Gamma^3(\mathbf{0}, 1)$ the sphere of the 3-d unit ball centered at the origin. First, consider a base case where the domain $\mathcal{X} = \Gamma \cup \mathbf{0}$, which is the union of the sphere of a unit ball centered at the origin and the origin. For any point $q \in \Gamma$, We let $C_q = \Gamma^3(q, 1) \cap \Gamma$ denote the circle of intersection of the sphere of two unit balls. Then we define $h_{q,1} = \mathbb{1}[C_q]$, which only classifies the circle C_q positive and $h_{q,0} = \mathbb{1}[\Gamma \setminus C_q]$ only classifies the circle and the origin negative. Our hypothesis class is $\mathcal{H} = \{h_{q,j} | q \in \Gamma, j \in \{0,1\}\}$. We draw our target h^* uniformly at random from $\mathcal{H}$, which is equivalent to drawing $q \sim \mathrm{Unif}(\Gamma)$ and $j \sim \mathrm{Ber}(1/2)$. The marginal data distribution $\mathcal{D}_{q,j,\mathcal{X}}$ puts probability mass $\zeta \in (0, \frac{t}{8m}]$ uniformly on the circle C_q and puts the remaining probability mass on $\mathbf{0}$, where the value of ζ is determined later. We draw $S_{\rm trn} \sim \mathcal{D}^m_{q,j}$.

The VC dimension of the hypothesis class. Then we show that the VC dimension of $\mathcal{H}$ is 5. Since all classifiers in $\mathcal{H}$ will classify $\mathbf{0}$ as negative, $\mathbf{0}$ cannot be shattered and thus, we only need to find shattered points on the sphere. Then we show that $\mathcal{H}$ can shatter 5 points. It is not hard to check that the following set of 5 points can be shattered: $\left\{(\frac{1}{2}, \frac{\sqrt{3}}{2}\cos(\frac{2k\pi}{5}), \frac{\sqrt{3}}{2}\sin(\frac{2k\pi}{5}))\right\}_{k=1}^{5}$.

Then we show that $\mathcal{H}$ cannot shatter 6 points. For any 6 points $P = \{p_1, \ldots, p_6\}$, if the 6 points can be shattered, then for any subset $P_1 \subseteq P$ with size 3, there exists a hypothesis classifying P_1 as 0s and $P \setminus P_1$ as 1s. That is, there exists a circle of radius $\frac{\sqrt{3}}{2}$ such that either only P_1 is on the circle or only $P \setminus P_1$ is on the circle. Then we claim that no 4 points can be on a circle of radius $\frac{\sqrt{3}}{2}$. If there are 4 points, w.l.o.g., $\{p_1, p_2, p_3, p_4\}$ on a circle of radius $\frac{\sqrt{3}}{2}$, then $\{p_i, p_5, p_6\}$ has to be on a circle C_i of radius $\frac{\sqrt{3}}{2}$, where $1 \le j \ne i \le 4$, $p_j \notin C_i$. But since the radius is fixed, there are only two different circles passing through $\{p_5, p_6\}$. Hence, there exists $1 \le i \ne j \le 4$ such that $C_i = C_j$, which contradicts that $p_j \notin C_i$.

Then w.l.o.g., if $\{p_1, p_2, p_3\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$. Consider $\{p_1, p_2, p_4\}$ and $\{p_3, p_5, p_6\}$, if $\{p_1, p_2, p_4\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$, then $\{p_3, p_4, p_5, p_6\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$ (to label $\{p_1, p_2\}$ different from $\{p_3, p_4, p_5, p_6\}$); if $\{p_3, p_5, p_6\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$, then there are three sub-cases: $\{p_1, p_3, p_5\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$, $\{p_2, p_3, p_5\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$ and both $\{p_2, p_4, p_6\}, \{p_1, p_4, p_6\}$ are on two circles of radius $\frac{\sqrt{3}}{2}$. For the first case, $\{p_1, p_2, p_4, p_6\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$ (to label $\{p_3, p_5\}$ different from $\{p_1, p_2, p_4, p_6\}$). For the second case, similarly $\{p_1, p_2, p_4, p_6\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$. For the third case, $\{p_1, p_2, p_3, p_5\}$ is on a circle of radius $\frac{\sqrt{3}}{2}$. Therefore, any 6 points cannot be shattered.

The attacker. We adopt the reflection function $m_{x_0}(\cdot)$ defined in the proof of Theorem 2 where $m_{x_0}(x) = 2\langle x_0, x\rangle x_0 - x$ for $x \in \Gamma$. For $S_{\mathrm{trn}} \sim \mathcal{D}_{q,j}^m$, we let $S_q = C_q \cap S_{\mathrm{trn},\mathcal{X}}$ denote the training instances in C_q (with replicants) and we further define $m_{x_0}(S_{\mathrm{trn}}) = \{(m_{x_0}(x), 1-y) | (x,y) \in S_q \times \mathcal{Y}\}$, and let

$$\mathrm{Adv}(h^*, S_{\mathrm{trn}}, x_0) = \begin{cases} m_{x_0}(S_{\mathrm{trn}}) & \text{if } x_0 \notin S_{\mathrm{trn},\mathcal{X}}, |S_q| \le t\,, \\ \emptyset & \text{else}\,. \end{cases}$$

If $x_0 \notin S_{\mathrm{trn},\mathcal{X}}$, then $h_{q,j}$ is consistent with $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{q,j}, S_{\mathrm{trn}}, x_0)$. That is, $\mathrm{Adv}(h_{q,j}, S_{\mathrm{trn}}, x_0)$ is clean-labeled.

Analysis. Due to the construction, we have

$$\mathbb{E}_{S_{\mathrm{trn}} \sim \mathcal{D}_{q,j}^m}[|S_q|] = m\zeta\,.$$

Then by Markov's inequality, we have

$$\mathbb{P}_{S_{\mathrm{trn}} \sim \mathcal{D}_{q,j}^m}(|S_q| \ge t) \le \frac{m\zeta}{t} < \frac{1}{4}\,.$$

Let $\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\mathrm{trn}}, x_0)$ denote the event of $\{\mathcal{A}(S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{q,j}, S_{\mathrm{trn}}, x_0), x_0) \ne h_{q,j}(x_0)\}$ and let $\mathcal{E}_2(S_{\mathrm{trn}}, x_0, q)$ denote the event of $\{|S_q| \le t \cap x_0 \notin S_{\mathrm{trn},\mathcal{X}}\}$. It is easy to check that $\mathcal{E}_2(S_{\mathrm{trn}}, x_0, q) = \mathcal{E}_2(m_{x_0}(S_{\mathrm{trn}}), x_0, m_{x_0}(q))$. Besides, conditional on $\mathcal{E}_2(S_{\mathrm{trn}}, x_0, q)$, we have the poisoned data set $S_{\mathrm{trn}} \cup \mathrm{Adv}(h_{q,j}, S_{\mathrm{trn}}, x_0) = m_{x_0}(S_{\mathrm{trn}}) \cup \mathrm{Adv}(h_{m_{x_0}(q),1-j}, m_{x_0}(S_{\mathrm{trn}}), x_0)$ and

thus, any algorithm $\mathcal{A}$ will behave the same at the test instance x_0 no matter whether the training set is $S_{\rm trn}$ or $m_{x_0}(S_{\rm trn})$. Since $h_{q,j}(x_0) \neq h_{m_{x_0}(q),1-j}(x_0)$, we have $\mathbb{1}[\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\rm trn}, x_0)] = \mathbb{1}[\neg\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{m_{x_0}(q),1-j}, m_{x_0}(S_{\rm trn}), x_0)]$ conditional on $\mathcal{E}_2(S_{\rm trn}, x_0, q)$. Let $f_q(x)$ denote the probability density function of $\mathrm{Unif}(C_q)$ and then we have $f_q(x) = f_{m_{x_0}(q)}(m_{x_0}(x))$. For any fixed x_0, the distributions of q and $m_{x_0}(q)$ and the distributions of j and $1-j$ are the same respectively. Since $S_{\rm trn}$ are samples drawn from $\mathcal{D}^m_{q,j}$, $m_{x_0}(S_{\rm trn})$ are actually samples drawn from $\mathcal{D}^m_{m_{x_0}(q),1-j}$. Then we have

$$
\begin{aligned}
&\mathbb{E}_{h^*\sim\mathrm{Unif}(\mathcal{H}),S_{\rm trn}\sim\mathcal{D}^m}\left[\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\rm trn}, \mathcal{A})\right] \\
=&\zeta\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j},x\sim\mathrm{Unif}(C_q),\mathcal{A}}\left[\mathbb{1}[\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\rm trn}, x)]\right] \\
\geq&\zeta\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j},x\sim\mathrm{Unif}(C_q),\mathcal{A}}\left[\mathbb{1}[\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\rm trn}, x)\cap\mathcal{E}_2(S_{\rm trn}, x, q)]\right] \\
=&\zeta\int_{x\in\Gamma}\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j},\mathcal{A}}\left[f_q(x)\mathbb{1}[\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\rm trn}, x)\cap\mathcal{E}_2(S_{\rm trn}, x, q)]\right]dx \\
=&\zeta\int_{x\in\Gamma}\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j},\mathcal{A}}[f_{m_x(q)}(x)\mathbb{1}[\neg\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{m_x(q),1-j}, m_x(S_{\rm trn}), x)] \\
&\cdot\mathbb{1}[\mathcal{E}_2(m_x(S_{\rm trn}), x, m_x(q))]]dx \\
=&\zeta\int_{x}\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j},\mathcal{A}}\left[f_q(x)\mathbb{1}[\neg\mathcal{E}_1(\mathcal{A}, \mathrm{Adv}, h_{q,j}, S_{\rm trn}, x)]\cdot\mathbb{1}[\mathcal{E}_2(S_{\rm trn}, x, q)]\right]dx \\
=&\frac{\zeta}{2}\int_{x}\mathbb{E}_{q\sim\mathrm{Unif}(\Gamma),j\sim\mathrm{Ber}(\frac{1}{2}),S_{\rm trn}\sim\mathcal{D}^m_{q,j}}\left[f_q(x)\mathbb{1}[\mathcal{E}_2(S_{\rm trn}, x, q)]\right]dx \\
>&\frac{3\zeta}{8},
\end{aligned}
$$

which completes the proof for $d = 1$ by setting $\zeta = \min(\frac{t}{8m}, 1)$.

Extension to general $d \geq 1$. To extend the base case to $d > 1$, we construct d separate balls and repeat the above construction on each ball individually. For $i \in [d]$, let $\Gamma_i = \Gamma^3(3ie_1, 1)$ denote the sphere of a ball with radius 1 centered at $3ie_1$. Consider the domain $\mathcal{X} = \cup_{i\in[d]}\Gamma_i\cup\{\mathbf{0}\}$ as the union of d non-overlapping unit balls and the origin. For $q_i \in \Gamma_i$, let $h^1_{q_i} = \mathbb{1}[\Gamma^3(q_i, 1)\cap\Gamma_i]$ denote the hypothesis classifying only points on the circle of $\Gamma^3(q_i, 1)\cap\Gamma_i$ positive and $h^0_{q_i} = \mathbb{1}[\Gamma_i\setminus\Gamma^3(q_i, 1)]$ denote the hypothesis classifying only points on Γ_i positive except the circle $\Gamma^3(q_i, 1)\cap\Gamma_i$. Let $h^s_{q_1,\ldots,q_d} = \sum_{i\in[d]} h^{s_i}_{q_i}$, where $s \in \{0,1\}^d$ denote the hypothesis combining all d balls and $\mathcal{H} = \{h^s_{q_1,\ldots,q_d} | q_i \in \Gamma_i, \forall i \in [d], s \in \{0,1\}^d\}$. We have the VC dimension of $\mathcal{H}$ is $5d$. Our target function is selected uniformly at random from $\mathcal{H}$ and similar to the case of $d = 1$, we assign probability $\zeta = \min(\frac{1}{d}, \frac{t}{8m})$ to each circle on the balls and the remaining probability mass on the origin. Since every ball is independent with other balls and thus, we have $\mathbb{E}_{h^*\sim\mathrm{Unif}(\mathcal{H}),S_{\rm trn}\sim\mathcal{D}^m}\left[\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\rm trn}, \mathcal{A})\right] > \frac{3d\zeta}{8} = \min(\frac{3td}{64m}, \frac{3}{8})$.

In all, there exists a target function $h^* \in \mathcal{H}$ and a data distribution $\mathcal{D}$ over D_{h^*} such that $\mathbb{E}_{S_{\rm trn}}\left[\mathrm{atk}_{\mathcal{D}}(t, h^*, S_{\rm trn}, \mathcal{A})\right] > \varepsilon$ when $m < \frac{3td}{64\varepsilon}$ for $\varepsilon \leq \frac{3}{8}$. ■