# Defending Support Vector Machines against Poisoning Attacks: the Hardness and Algorithm

**Hu Ding, Fan Yang, Jiawei Huang**
School of Computer Science and Technology
University of Science and Technology of China
He Fei, China
huding@ustc.edu.cn, {yang208, hjw0330}@mail.ustc.edu.cn

## Abstract

Adversarial machine learning has attracted a great amount of attention in recent years. In a poisoning attack, the adversary can inject a small number of specially crafted samples into the training data which make the decision boundary severely deviate and cause unexpected misclassification. Due to the great importance and popular use of support vector machines (SVM), we consider defending SVM against poisoning attacks in this paper. We consider two common strategies for defending: designing robust SVM algorithms and data sanitization. Though several robust SVM algorithms have been proposed before, most of them either are in lack of adversarial-resilience or rely on strong assumptions about the data distribution or the attacker's behavior. Moreover, the research on the complexities is still quite limited nowadays. We are the first, to the best of our knowledge, to prove that even the simplest hard-margin one-class SVM with outliers problem is NP-complete, and has no fully PTAS unless P=NP. For the data sanitization defense, we link it to the intrinsic dimensionality of data; in particular, we provide a sampling theorem in doubling metrics for explaining the effectiveness of DBSCAN (as a density-based outlier removal method) for defending against poisoning attacks. In our empirical experiments, we compare several defenses including the DBSCAN and robust SVM methods, and investigate the influences from the intrinsic dimensionality and data density to their performances.

## 1 Introduction

In the past decades we have witnessed enormous progress in machine learning. One driving force behind this is the successful applications of machine learning technology to many different fields, such as data mining, networking, and bioinformatics. However, with its territory rapidly enlarging, machine learning has also imposed a great deal of challenges for researchers to meet its new demands. In particular, the field of *adversarial machine learning* concerning about the potential vulnerabilities of the algorithms has attracted a great amount of attention [3, 29, 8, 27]. As mentioned in the survey paper [8], the very first work of adversarial machine learning dates back to 2004, in which Dalvi *et al.* [19] formulated the adversarial classification problem as a game between the classifier and the adversary. In general, the adversarial attacks against machine learning can be categorized to **evasion attacks** and **poisoning attacks** [8]. An evasion attack happens at test time, where the adversary aims to evade the trained classifier by manipulating test examples. For example, Szegedy *et al.* [50] observed that small perturbation to a test image can arbitrarily change the neural network's prediction.

In this paper, we focus mainly on poisoning attacks that happen at training time. Usually, the adversary injects a small number of specially crafted samples into the training data which make the

Preprint. Under review.

decision boundary severely deviate and cause unexpected misclassification. In particular, because of the fact that open datasets are commonly used to train our machine learning algorithms now, it has been considered as a key security issue that seriously limits real-world applications [8]. For instance, even a small number of poisoning samples can significantly increase the test error of support vector machine (SVM) [7, 39, 55, 33]. Beyond linear classifiers, a number of works studied the poisoning attacks on other machine learning problems, such as clustering [5], PCA [45], and regression [30].

Though lots of works focused on constructing poisoning attacks, our ultimate goal is to design defenses. Poisoning attacks can be regarded as *outliers*, and this leads to two natural approaches for defenses: **(1) data sanitization defense**, *i.e.,* first perform outlier removal on the data and then run an existing machine learning algorithm on the cleaned data [17], or **(2) directly design a robust optimization algorithm that is resilient against outliers** [15, 30].

Steinhardt *et al.* [49] studied two basic methods of data sanitization defense, which remove the points outside a specified sphere or slab, for binary classification; they showed that high dimensionality gives attacker more room for constructing attacks to evade outlier removal. Laishram and Phoha [35] applied the seminal DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method [21] to remove outliers for SVM and showed that it can successfully identify most of the poisoning data. However, their method is heuristic and lack of theoretical analysis. Several other outlier removal or detection methods for fighting poisoning attacks also have been studied recently, such as [43, 42]. We would like to point out that outlier removal is a topic that in fact has been extensively studied in various fields before (we refer the reader to the surveys [1, 34, 10]).

The other defense strategy, *i.e.,* designing robust optimization algorithms, also has a long history in machine learning. A substantial part of robust optimization algorithms rely on the idea of regularization. For example, Xu *et al.* [57] studied the relation between robustness and regularization for SVM; several other robust SVM algorithms were proposed as well [51, 58, 41, 56, 31]. However, as discussed in [39, 30], these approaches are not quite ideal to defend against poisoning attacks since the outliers can be located arbitrarily in the feature space by the adversary. Another idea for achieving the robustness guarantee is to add strong assumptions about the data distribution or the attacker's behavior [23, 53], but they are usually not well satisfied in practice. An alternative approach is to explicitly remove outliers during optimization, such as the "trimmed" version for robust regression recently proposed in [30]; but this model often results in a challenging **combinatorial optimization problem**: if $z$ of the input $n$ data items are outliers ($z < n$), we have to consider an exponentially large number $\binom{n}{z}$ of different possible cases when optimizing the objective function. As an example, the clustering with outliers problems, like $k$-means and $k$-center with outliers, have at least quadratic time complexities in general [12, 14]; of course, if we aim to obtain only a local optimum, we can formulate the problem as a bilevel optimization that alternatively removes outliers and minimizes the objective function on the remaining data in each iteration, like the $k$-means$--$ algorithm [13] (this alternating minimization idea also works for other problems, such as regression with outliers [24]).

### 1.1 Our Contributions

Due to the great importance and popular use of SVM [11], we consider defending SVM against poisoning attacks in this paper. Our contributions are twofold.

First, we consider the robust optimization approach. To study its complexity, we only consider the hard-margin case (because the soft-margin case is more complicated and thus should have an even higher complexity). As mentioned above, we can formulate the SVM with outliers problem as a combinatorial optimization problem for achieving the **adversarial-resilience**: finding an optimal subset of $n - z$ items from the poisoned input data to achieve the largest separating margin induced by the SVM (we will provide the formal definition in Section 2). Though its local optimum can be obtained by using the alternating minimization approach, we are still interested in its global optimal solution. We are unaware of any strong **hardness-of-approximation result** for this problem. In Section 3, we try to bridge the gap in the current state of knowledge. We prove that even the simplest one-class SVM with outliers problem is NP-complete, and has no fully polynomial-time approximation scheme (PTAS) unless P=NP. That is, it is quite unlikely that one can achieve a (nearly) optimal solution in polynomial time.

Second, we investigate the DBSCAN based data sanitization defense proposed in [35] and explain its effectiveness in theory (Section 4). DBSCAN is one of the most popular density-based clustering methods and has been implemented for solving many real-world outlier removal problems [21, 47]; roughly speaking, the inliers are assumed to be located in some dense regions and the remaining points are recognized as the outliers. Actually, the intuition of using DBSCAN for data sanitization is straightforward. We assume the original input training data (before poisoning attack) is large and dense enough in the domain $\Omega$; thus the poisoning data should be the sparse outliers together with some small clusters located outside the dense regions, which can be identified by the DBSCAN. Obviously, if the attacker has a fixed budget $z$ (the number of poisoning points), the lager the data size $n$ is, the more efficient the DBSCAN performs (we can imagine the extreme case that $z/n$ is close to 1, where it is clearly inappropriate to use a density based clustering method to identify the outliers).

Thus a fundamental question in theory is what about **the lower bound of the data size** $n$ for guaranteeing the correctness of the DBSCAN (we can assume that the clean data is a set of *i.i.d.* samples drawn from the domain $\Omega$). However, to achieve a favorable lower bound is a non-trivial task. The VC dimension [36] of the range space induced by the Euclidean distance is high in a high-dimensional feature space, and thus the lower bound of the data size $n$ could be very large. Our idea is motivated by the recent observations on the link between the adversarial vulnerability of learning and the intrinsic dimensionality of the data [32, 2, 37]. We prove a lower bound of $n$ that depends on the intrinsic dimension of $\Omega$ and is independent of the feature space's dimensionality. Our result strengthens the observation from [49] who only considered the Euclidean space's dimensionality: more precisely, it is the "high intrinsic dimensionality" that gives attacker more room to evade outlier removal. In particular, different from the previous results [32, 2, 37] focusing on evasion attacks, our result is the first one linking poisoning attacks to intrinsic dimensionality, to the best of our knowledge.

## 2 Preliminaries

Given two point sets $P^+$ and $P^-$ in $\mathbb{R}^d$, the problem of linear **support vector machine (SVM)** [11] is to find the maximum margin (induced by two parallel hyperplanes) separating these two point sets (if they are separable). For convenience, we say that $P^+$ and $P^-$ are the sets of points labeled as "$+1$" and "$-1$", respectively. If $P^+$ (or $P^-$) is a single point, say the origin, the problem is called **one-class SVM**. The SVM can be formulated as a quadratic programming problem, and a number of efficient techniques have been developed in the past, such as the soft margin SVM [16], $\nu$-SVM [46, 18], and Core-SVM [52]. If $P^+$ and $P^-$ are not separable, we can apply the kernel method: each point $p \in P^+ \cup P^-$ is mapped to be $\phi(p)$ in a higher dimensional space; the inner product $\langle \phi(p_1), \phi(p_2) \rangle$ is defined by a kernel function $\mathcal{K}(p_1, p_2)$. Many existing SVM algorithms can be adapted to handle the non-separable case by using kernel functions.

**Poisoning attacks.** Usually, the adversary injects some bad points to the original data set $P^+ \cup P^-$. For instance, the adversary can drawn a sample $q$ from the domain of $P^+$, and flip its label to be "$-$"; therefore, this poisoning sample $q$ can be viewed as an outlier of $P^-$. Since poisoning attack is expensive, we often assume that the adversary can poison at most $z \in \mathbb{Z}^+$ points (or the poisoned fraction $\frac{z}{|P^+ \cup P^-|}$ is a fixed small number in $(0, 1)$). Overall, we can formulate the defense against poisoning attacks as the following combinatorial optimization problem. As mentioned in Section 1.1, we only consider the simpler hard-margin case for studying the complexity.

**Definition 1** (Support Vector Machine (SVM) with Outliers)**.** *Let $(P^+, P^-)$ be an instance of SVM in $\mathbb{R}^d$, and suppose $\left|P^+ \cup P^-\right| = n$. Given a positive integer $z < n$, the problem of SVM with outliers is to find two subsets $P_1^+ \subseteq P^+$ and $P_1^- \subseteq P^-$ with $\left|P_1^+ \cup P_1^-\right| = n - z$, such that the width of the margin separating $P_1^+$ and $P_1^-$ is maximized.*

*Suppose the optimal margin has the width $h_{opt}$. If we achieve a solution with margin width $h \geq (1-\epsilon)h_{opt}$ where $\epsilon$ is a small parameter in $(0, 1)$, we say that it is a $(1-\epsilon)$-approximation.*

**Remark 1.** *The model proposed in Definition 1 in fact follows the popular **data trimming** idea from robust statistics [44]. As an example, Jagielski* et al. *[30] proposed a robust regression model that is resilient against poisoning attacks based on data trimming.*

We also need to clarify the definition of intrinsic dimensionality for our following analysis. We consider the **doubling dimension** which is a measure of intrinsic dimensionality widely adopted in learning theory [9]. Given $p \in \mathbb{R}^d$ and $r \geq 0$, we use $\mathbb{B}(p, r) = \{q \in \mathbb{R}^d \mid ||q - p|| \leq r\}$ to indicate the ball of radius $r$ around $p$.

**Definition 2** (Doubling Dimension)**.** *The doubling dimension of a point set $P \subset \mathbb{R}^d$ is the smallest number $\rho$, such that for any $p \in P$ and $r \geq 0$, $P \cap \mathbb{B}(p, 2r)$ is always covered by the union of at most $2^\rho$ balls with radius $r$.*

The doubling dimension is often used for describing the expansion rates of point sets. Note that the intrinsic dimensionality described in [2, 37] is quite similar to the doubling dimension, which also measures the expansion rates of point sets.

## 3 The Hardness of SVM with Outliers

In this section, we prove that even the one-class SVM with outliers problem is NP-complete. Further, we show that there is no fully PTAS for the problem unless P=NP, that is, we cannot achieve a polynomial time $(1-\epsilon)$-approximation for any given $\epsilon \in (0, 1)$. Our idea is partly inspired by the result from Megiddo [38]. Given a set of points in $\mathbb{R}^d$, the "covering by two balls" problem is to determine that whether the point set can be covered by two unit balls. By the reduction from 3-SAT, Megiddo proved that the "covering by two balls" problem is NP-complete. In the proof of the following theorem, we modify Megiddo's construction of the reduction to adapt the one-class SVM with outliers problem.

**Theorem 1.** *The one-class SVM with outliers problem is NP-complete, and has no fully PTAS unless P=NP.*

*Proof.* Let $\Gamma$ be a 3-SAT instance with the literal set $\{u_1, \bar{u}_1, \cdots, u_l, \bar{u}_l\}$ and clause set $\{E_1, \cdots, E_m\}$. We construct the corresponding instance $P_\Gamma$ of one-class SVM with outliers. First, let $U = \{\pm e_i \mid i = 1, 2, \cdots, l+1\}$ be the $2(l+1)$ unit vectors of $\mathbb{R}^{l+1}$, where each $e_i$ has 1 in the $i$-th position and 0 in other positions. Also, for each clause $E_j$ with $1 \leq j \leq m$, we generate a point $q_j = (q_{j,1}, q_{j,2}, \cdots, q_{j,l+1})$: (1) if $u_i$ occurs in $E_j$, $q_{j,i} = \alpha$, (2) else if $\bar{u}_i$ occurs in $E_j$, $q_{j,i} = -\alpha$, (3) else, $q_{j,i} = 0$; in addition, $q_{j,l+1} = 3\alpha$. For example, if $E_j = u_{i_1} \vee \bar{u}_{i_2} \vee u_{i_3}$, the point

$$q_j = (0, \cdots, 0, \underbrace{\alpha}_{i_1}, 0, \cdots, 0, \underbrace{-\alpha}_{i_2}, 0, \cdots, 0, \underbrace{\alpha}_{i_3}, 0, \cdots, 0, 3\alpha). \quad (1)$$

We will determine the value of $\alpha$ below. Let $Q$ denote the set $\{q_1, \cdots, q_m\}$. Now, we construct the instance $P_\Gamma = U \cup Q$ with the number of points $n = 2(l+1) + m$ and the number of outliers $z = l + 1$. Below we prove that $\Gamma$ has a satisfying assignment if and only if $P_\Gamma$ has a solution with margin width $\frac{1}{\sqrt{l+1}}$.

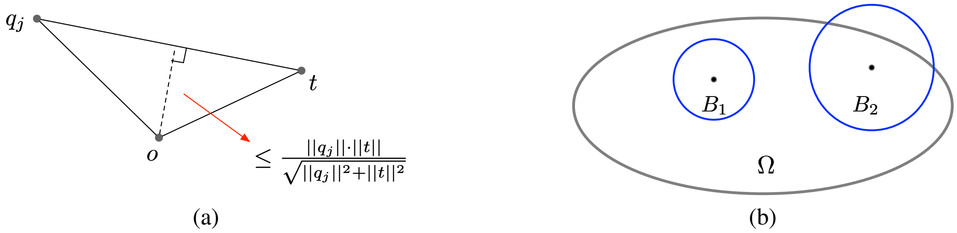
**Suppose there exists a satisfying assignment for** $\Gamma$**.** We define the set $S \subset P_\Gamma$ as follows. If $u_i$ is true in $\Gamma$, we include $e_i$ in $S$, else, we include $-e_i$ in $S$; we also include $e_{l+1}$ in $S$. Assume $\alpha > 1/2$. We claim that the set $S \cup Q$ is a solution of the instance $P_\Gamma$ with the margin width $\frac{1}{\sqrt{l+1}}$, that is, the size $|S \cup Q| = n - z$ and the margin separating the origin $o$ and $S \cup Q$ has width $\frac{1}{\sqrt{l+1}}$. It is easy to verify the size of $S \cup Q$. To compute the width, we consider the mean point of $S$ which is denoted as $t$. For each $1 \leq i \leq l$, if $u_i$ is true, the $i$-th position of $t$ is $\frac{1}{l+1}$, else, the $i$-th position of $t$ is $-\frac{1}{l+1}$; the $(l+1)$-th position of $t$ is $\frac{1}{l+1}$. Let $\mathcal{H}_t$ be the hyperplane that is orthogonal to the vector $t - o$ and passing through $t$. Obviously, $\mathcal{H}_t$ separates $S$ and $o$ with the margin width $||t|| = \frac{1}{\sqrt{l+1}}$. Furthermore, for any point $q_j \in Q$, since there exists at least one true variable in $E_j$, we have

$$\langle q_j, \frac{t}{||t||} \rangle \geq \frac{3\alpha}{\sqrt{l+1}} + \frac{\alpha}{\sqrt{l+1}} - \frac{2\alpha}{\sqrt{l+1}} = \frac{2\alpha}{\sqrt{l+1}} > \frac{1}{\sqrt{l+1}}, \quad (2)$$

where the last inequality comes from the fact $\alpha > 1/2$. Therefore, all the points from $Q$ lie on the same side of $\mathcal{H}_t$ as $S$, and then the set $S \cup Q$ can be separated from $o$ by a margin with width $\frac{1}{\sqrt{l+1}}$.

**Suppose the instance** $P_\Gamma$ **has a solution with margin width** $\frac{1}{\sqrt{l+1}}$**.** With a slight abuse of notations, we still use $S$ to denote the subset of $U$ that is included in the set of $n - z$ inliers. Since the number of outliers is $z = l + 1$, we know that for any pair $\pm e_i$, there exists exactly one point belonging to $S$; also, the whole set $Q$ should be included in the solution to keep that there are $n - z$ inliers in total. We still use $t$ to denote the mean point of $S$. Now, we have the assignment for $\Gamma$: if $e_i \in S$, we assign $u_i$ to be true, else, we assign $\bar{u}_i$ to be true. We claim that $\Gamma$ is satisfied by this assignment. For any clause $E_j$, if it is not satisfied, *i.e.,* all the three variables in $E_j$ are false, then we have

$$\langle q_j, \frac{t}{||t||} \rangle \leq \frac{3\alpha}{\sqrt{l+1}} - \frac{3\alpha}{\sqrt{l+1}} = 0. \quad (3)$$



(a) (b)

Figure 1: (a) An illustration for (4); (b) the ball $B_1$ is enclosed by $\Omega$ and the ball $B_2$ is not.

That means the angle $\angle q_j o t \geq \pi/2$. So any margin separating the origin $o$ and the set $S \cup Q$ should has the width at most

$$\frac{||q_j|| \cdot ||t||}{\sqrt{||q_j||^2 + ||t||^2}} < ||t|| = \frac{1}{\sqrt{l+1}}. \tag{4}$$

See Figure 1a. This is in contradiction to the assumption that $P_\Gamma$ has a solution with margin width $\frac{1}{\sqrt{l+1}}$.

Overall, $\Gamma$ has a satisfying assignment if and only if $P_\Gamma$ has a solution with margin width $\frac{1}{\sqrt{l+1}}$. Thus, the one-class SVM with outliers problem is NP-complete. Moreover, the gap between $\frac{1}{\sqrt{l+1}}$ and $\frac{||q_j||\cdot||t||}{\sqrt{||q_j||^2+||t||^2}}$ is

$$\begin{aligned}\frac{1}{\sqrt{l+1}} - \sqrt{\frac{12\alpha^2\frac{1}{l+1}}{12\alpha^2 + \frac{1}{l+1}}} &= (\frac{1}{l+1})^{3/2}\frac{1}{\sqrt{12\alpha^2 + \frac{1}{l+1}}(\sqrt{12\alpha^2 + \frac{1}{l+1}} + 2\sqrt{3}\alpha)} \\ &= \Theta\big((\frac{1}{l+1})^{3/2}\big) \end{aligned} \tag{5}$$

if we assume $\alpha$ is a fixed constant. Therefore, if we set $\epsilon = O\big(\frac{(\frac{1}{l+1})^{3/2}}{(\frac{1}{l+1})^{1/2}}\big) = O(\frac{1}{l+1})$, then $\Gamma$ is satisfiable if and only if any $(1-\epsilon)$-approximation of the instance $P_\Gamma$ has width $> \sqrt{\frac{12\alpha^2\frac{1}{l+1}}{12\alpha^2+\frac{1}{l+1}}}$. That means if we have a fully PTAS for the one-class SVM with outliers problem, we can determine that whether $\Gamma$ is satisfiable or not in polynomial time. It implies that we cannot achieve a fully PTAS for one-class SVM with outliers, unless P=NP. □

## 4 The Data Sanitization Defense

From Theorem 1, we know that it is extremely challenging to achieve the optimal solution even for one-class SVM with outliers. Therefore, we turn to consider the other approach, data sanitization defense, under some reasonable assumption in practice. First, we prove a general sampling theorem in Section 4.1 which can help us to analyze density-based clustering methods on data with low doubling dimensions. Then, we apply this theorem to explain the effectiveness of DBSCAN for defending against poisoning attacks in Section 4.2.

### 4.1 A Sampling Theorem

Let $P$ be a set of *i.i.d.* samples drawn from a connected and compact domain $\Omega$ who has the doubling dimension $\rho > 0$. For ease of presentation in our following analysis, we assume that $\Omega$ lies on a manifold $\mathcal{F}$ in the $\mathbb{R}^d$ space. Let $\Delta$ denote the diameter of $\Omega$, *i.e.,* $\Delta = \max_{p_1,p_2\in\Omega}||p_1 - p_2||$. Also, we let $f$ be the probability density function of the data distribution over $\Omega$.

To measure the uniformity of $f$, we define a value $\lambda$ as follows. For any $c \in \Omega$ and any $r > 0$, we say "the ball $\mathbb{B}(c, r)$ is enclosed by $\Omega$" if $\partial\mathbb{B}(c, r) \cap \mathcal{F} \subset \Omega$; intuitively, if the ball center $c$ is close to the boundary $\partial\Omega$ of $\Omega$ or the radius $r$ is too large, the ball will not be enclosed by $\Omega$. See Figure 1b for an illustration. We let $\lambda = \max \frac{\int_{\mathbb{B}(c',r)} f(x)\,\mathrm{d}x}{\int_{\mathbb{B}(c,r)} f(x)\,\mathrm{d}x}$, where $\mathbb{B}(c, r)$ and $\mathbb{B}(c', r)$ are any two equal-size balls, and $\mathbb{B}(c, r)$ is required to be enclosed by $\Omega$. As an example, if the data is uniformly distributed over $\Omega$ who lies on a flat manifold, the value $\lambda$ will be equal to 1. On the other hand, if the distribution is very imbalanced or the manifold $\mathcal{F}$ is very rugged, the value $\lambda$ will be high.

**Theorem 2.** *Let $m \in \mathbb{Z}^+$, $\epsilon \in (0, \frac{1}{8})$, and $\delta \in (0, \Delta)$. If the sample size*

$$|P| > \max\Big\{\Theta\big(\frac{m}{1-\epsilon} \cdot \lambda \cdot (\frac{1+\epsilon}{1-\epsilon}\frac{\Delta}{\delta})^{\rho}\big), \tilde{\Theta}\big(\rho \cdot \lambda^2 \cdot (\frac{1+\epsilon}{1-\epsilon}\frac{\Delta}{\delta})^{2\rho}(\frac{1}{\epsilon})^{\rho+2}\big)\Big\}, \tag{6}$$

*then with constant probability, for any ball $\mathbb{B}\big(c, \delta\big)$ enclosed by $\Omega$, the size $|\mathbb{B}\big(c, \delta\big) \cap P| > m$. The asymptotic notation $\tilde{\Theta}(f) = \Theta\big(f \cdot polylog(\frac{L\Delta}{\delta\epsilon})\big)$.*

**Remark 2.** *(i) A highlight of Theorem 2 is that the lower bound of $|P|$ is independent of the Euclidean dimensionality; so if the doubling dimension $\rho$ is a fixed number, the required sample size for $P$ is relatively low.*

*(ii) For the simplest case that the data is uniformly distributed over $\Omega$ who lies on a flat manifold, $\lambda$ will be equal to 1 and thus the lower bound of $|P|$ in Theorem 2 becomes $\max\Big\{\Theta\big(\frac{m}{1-\epsilon}(\frac{1+\epsilon}{1-\epsilon}\frac{\Delta}{\delta})^{\rho}\big), \tilde{\Theta}\big(\rho(\frac{1+\epsilon}{1-\epsilon}\frac{\Delta}{\delta})^{2\rho}(\frac{1}{\epsilon})^{\rho+2}\big)\Big\}$.*

Before proving Theorem 2, we need to relate the doubling dimension $\rho$ to the VC dimension $\mathtt{dim}$ of the range space consisting of all balls with different radii [36]. Unfortunately, Huang *et al.* [28] recently showed that "*although both dimensions are subjects of extensive research, to the best of our knowledge, there is no nontrivial relation known between the two*". For instance, they constructed a doubling metric having unbounded VC dimension, and the other direction cannot be bounded neither. However, if allowing a small distortion to the distance, we can achieve an upper bound on the VC dimension for a given metric space with bounded doubling dimension. For stating the result, they defined a distance function called "*$\epsilon$-smoothed distance function*": $g(p, q) \in (1 \pm \epsilon)||p - q||$ for any $p, q \in \mathbb{R}^d$, where $\epsilon \in (0, \frac{1}{8})$. Given $p \in \mathbb{R}^d$ and $\delta > 0$, the ball defined by this distance function $g(\cdot, \cdot)$ will be $\mathbb{B}_g(p, \delta) = \{q \in \mathbb{R}^d \mid g(p, q) \leq \delta\}$.

**Theorem 3** ([28])**.** *Suppose the point set $P \subset \mathbb{R}^d$ has the doubling dimension $\rho > 0$. There exists an $\epsilon$-smoothed distance function "$g(\cdot, \cdot)$" such that the VC dimension[1] $\mathtt{dim}_\epsilon$ of the range space consisting of all balls with different radii is at most $\tilde{O}(\frac{\rho}{\epsilon^\rho})$, if replacing the Euclidean distance by $g(\cdot, \cdot)$.*

*Proof.* **(of Theorem 2)** Let $r$ be any positive number. First, since the doubling dimension of $\Omega$ is $\rho$, if recursively applying Definition 2 $\log\frac{\Delta}{r}$ times, we know that $\Omega$ can be covered by at most $\Theta\Big((\frac{\Delta}{r})^\rho\Big)$ balls with radius $r$. Thus, if $\mathbb{B}(c, r)$ is enclosed by $\Omega$, we have

$$\frac{\int_{\mathbb{B}(c,r)} f(x)\,\mathrm{d}x}{\int_{\Omega} f(x)\,\mathrm{d}x} \geq \Theta\Big(\frac{1}{\lambda} \cdot (\frac{r}{\Delta})^\rho\Big). \tag{7}$$

Now we consider the size $|\mathbb{B}(c, \delta) \cap P|$. From Theorem 3, we know that the VC dimension $\mathtt{dim}_\epsilon$ with respect to the $\epsilon$-smoothed distance is $\tilde{O}(\frac{\rho}{\epsilon^\rho})$. Thus, for any $\epsilon_0 \in (0, 1)$, if

$$|P| \geq \Theta\big(\frac{1}{\epsilon_0^2}\mathtt{dim}_\epsilon \log\frac{\mathtt{dim}_\epsilon}{\epsilon_0}\big), \tag{8}$$

the set $P$ will be an $\epsilon_0$-sample of $\Omega$; that is, for any $c \in \mathbb{R}^d$ and $\delta' \geq 0$,

$$\frac{|\mathbb{B}_g(c, \delta') \cap P|}{|P|} \in \frac{\int_{\mathbb{B}_g(c,\delta')} f(x)\,\mathrm{d}x}{\int_{\Omega} f(x)\,\mathrm{d}x} \pm \epsilon_0 \tag{9}$$

with constant probability [36]. Note the exact probability comes from the success probability that $P$ is an $\epsilon_0$-sample of $\Omega$; for convenience, we simply say it is a constant probability. Because $g(\cdot, \cdot)$ is an $\epsilon$-smoothed distance function of the Euclidean distance, we have

$$\mathbb{B}(c, \frac{\delta'}{1+\epsilon}) \subseteq \mathbb{B}_g(c, \delta') \subseteq \mathbb{B}(c, \frac{\delta'}{1-\epsilon}). \tag{10}$$

[1] In [28], the authors used "shattering dimension" to state their result. Actually, the shattering dimension is another measure for the complexity of range space, which is tightly related to the VC dimension [22]. For example, if the shattering dimension is $t$, the VC dimension is bounded by $O(t \log t)$.

So if we set $\epsilon_0 = \epsilon \cdot \Theta\Big(\frac{1}{\lambda} \cdot (\frac{1-\epsilon}{1+\epsilon}\frac{\delta}{\Delta})^\rho\Big)$ and $\delta' = (1-\epsilon)\delta$, (7), (9), and (10) jointly imply $\frac{|\mathbb{B}(c,\delta)\cap P|}{|P|} =$

$$\begin{aligned}
\frac{|\mathbb{B}(c, \frac{\delta'}{1-\epsilon}) \cap P|}{|P|} &\geq \frac{|\mathbb{B}_g(c, \delta') \cap P|}{|P|} \geq \frac{\int_{\mathbb{B}_g(c,\delta')} f(x)\,\mathrm{d}x}{\int_\Omega f(x)\,\mathrm{d}x} - \epsilon_0 \\
&\geq \frac{\int_{\mathbb{B}(c, \frac{\delta'}{1+\epsilon})} f(x)\,\mathrm{d}x}{\int_\Omega f(x)\,\mathrm{d}x} - \epsilon_0 \\
&\geq (1-\epsilon) \cdot \Theta\Big(\frac{1}{\lambda} \cdot (\frac{1-\epsilon}{1+\epsilon}\frac{\delta}{\Delta})^\rho\Big). \qquad (11)
\end{aligned}$$

The last inequality comes from (7); since we assume the ball $\mathbb{B}\big(c, \delta\big)$ is enclosed by $\Omega$, the shrunk ball $\mathbb{B}(c, \frac{\delta'}{1+\epsilon}) = \mathbb{B}(c, \frac{1-\epsilon}{1+\epsilon}\delta)$ should be enclosed as well. Moreover, if

$$|P| \geq \Theta\big(\frac{m}{1-\epsilon} \cdot \lambda \cdot (\frac{1+\epsilon}{1-\epsilon}\frac{\Delta}{\delta})^\rho\big), \qquad (12)$$

we have $|\mathbb{B}\big(c, \delta\big) \cap P| > m$ from (11). Combining (8) and (12), we obtain the lower bound of $|P|$. □

### 4.2 The DBSCAN Approach

For the sake of completeness, we briefly introduce the method of DBSCAN [21] below. Given two parameters $r > 0$ and $\mathtt{MinPts} \in \mathbb{Z}^+$, the DBSCAN divides the set $P$ into three classes: (1) $p$ is a **core point**, if $|\mathbb{B}(p, r) \cap P| > \mathtt{MinPts}$; (2) $p$ is a **border point**, if $p$ is not a core point but $p \in \mathbb{B}(q, r)$ of some core point $q$; (3) all the other points are **outliers**. Actually, we can imagine that the set $P$ forms a graph where any pair of points are connected if their pairwise distance is no larger than $r$; then the set of core points and border points form several clusters where each cluster is a connected component (a border point may belong to multiple clusters, but we can arbitrarily assign it to only one cluster). The goal of the DBSCAN is to identify these clusters and the remaining outliers.

Following Section 4.1, we assume that $P$ is a set of *i.i.d.* samples drawn from the connected and compact domain $\Omega$ who has the doubling dimension $\rho > 0$. We let $Q$ be the set of $z$ poisoning data items injected by the attacker to $P$, and suppose each $q \in Q$ has distance larger than $\delta_1 > 0$ to $\Omega$. In an evasion attack, we often use the adversarial perturbation distance to evaluate the attacker's capability; but in a poisoning attack the attacker can easily achieve a large perturbation distance (*e.g.*, in the SVM problem, if the attacker flips the label of some point $p$, it will become an outlier having the perturbation distance larger than $h_{opt}$ to its ground truth domain, where $h_{opt}$ is the optimal margin width). Also, we assume the boundary $\partial\Omega$ is smooth and has curvature radius at least $\delta_2 > 0$ everywhere. For simplicity, let $\delta = \min\{\delta_1, \delta_2\}$. The following theorem states the relation between the DBSCAN and the poisoned dataset $P \cup Q$. We assume the poisoned fraction $\frac{|Q|}{|P|} < 1$.

**Theorem 4.** *We let $m$ be any absolute constant number larger than $1$, and assume that the size of $P$ satisfies the lower bound of Theorem 2 (with respect to $m$ and $\delta$). If we set $r = \delta$ and $\mathtt{MinPts} = m$, and run the DBSCAN on the poisoned dataset $P \cup Q$, the obtained largest cluster should be exactly $P$. In other word, the set $Q$ should be formed by the outliers and the clusters except the largest one from the DBSCAN.*

*Proof.* Since $\delta \leq \delta_2$, for any $p \in P$, either the ball $\mathbb{B}(p, \delta)$ is enclosed by $\Omega$, or $p$ is covered by some ball $\mathbb{B}(q, \delta)$ enclosed by $\Omega$. We set $r = \delta$ and $\mathtt{MinPts} = m$, and hence from Theorem 2 we know that all the points of $P$ will be core points or border points. Moreover, any point $q$ from $Q$ has distance larger than $r$ to the points of $P$, that is, any two points $q \in Q$ and $p \in P$ will not belong to the same cluster. Also, because we assume that the domain $\Omega$ is connected and compact, the set $P$ will form the largest cluster. □

**Remark 3.** *(i) We often adopt the poisoned fraction $\frac{z}{|P|}$ as the measure to indicate the attacker's capability. If we fix the value of $z$, the bound of $|P|$ from Theorem 2 reveals that the larger the doubling dimension $\rho$, the lower the poisoned fraction $\frac{z}{|P|}$ (and the easier corrupting the DBSCAN defense). In addition, when $\delta$ is large,* i.e., *each poisoning point has large perturbation distance and $\partial\Omega$ is sufficiently smooth, it will be relatively easy for the DBSCAN to defend.*

*But we should point out that this theoretical bound probably is overly conservative, since it requires a "perfect" sanitization result that removes all the poisoning samples (this is not always a necessary condition for achieving a good performance of the defending in practice). In our experiments, we show that the DBSCAN method can achieve promising performance, even when the poisoned fraction is higher than the threshold.*

*(ii) In practice, we usually cannot obtain the exact values of $\delta$ and $m$; instead, we may only estimate a reasonable lower bound $\hat{\delta}$ for $\delta$. Thus, we can set $r = \hat{\delta}$ and tune the value of* `MinPts` *until the largest cluster has $|P \cup Q| - z$ points.*

Directly solving such a high-dimensional DBSCAN instance is very expensive. A bottleneck of the original DBSCAN algorithm is that it needs to perform a range query for each data item, *i.e.,* computing the number of neighbors within the distance $r$, and the overall time complexity can be as large as $O(n^2 d)$ in the worst case, where $n$ is the number of data items. To speed up the step of range query, a natural idea is using some efficient index structures, such as $R^*$-tree [4], though the overall complexity in the worst case is still $O(n^2 d)$ (we refer the reader to the recent articles that systematically discussed this issue [25, 47]).

**Putting it all together.** Let $(P^+, P^-)$ be an instance of SVM with $z$ outliers, where $z$ is the number of poisoning points. We assume that the original input point sets $P^+$ and $P^-$ (before the poisoning attack) are *i.i.d.* samples drawn respectively from the connected and compact domains $\Omega^+$ and $\Omega^-$ with doubling dimension $\rho$. Then, we perform the DBSCAN procedure on $P^+$ and $P^-$ respectively (as Remark 3 (ii)). Suppose the obtained largest clusters are $\tilde{P}^+$ and $\tilde{P}^-$. Finally, we run an existing SVM algorithm on the cleaned instance $(\tilde{P}^+, \tilde{P}^-)$.

## 5 Discussion

In this paper, we study two different strategies for protecting SVM against poisoning attacks. To achieve the adversarial-resilience, the defense can be formulated as a combinatorial optimization problem called "SVM with outliers". We show for the first time that even the simplest hard-margin one-class SVM with outliers is NP-complete, and has no fully PTAS unless P=NP. We then focus on the data sanitization defense. Under the assumption that the original input data (before poisoning attack) are drawn from the domains with low doubling dimensions, we provide the lower bound of the data size to ensure that the DBSCAN can correctly identify the poisoning samples.

We leave the detailed experimental results on the synthetic and real datasets to our supplement. We compare several defenses including the DBSCAN and robust SVM methods, and study the trends of their classification accuracies with varying three values: the poisoned fraction, the intrinsic dimensionality, and the Euclidean dimensionality. All the experimental results were obtained by using publicly available implementations on a Windows 10 workstation equipped with an Intel core $i5$-$8265U$ processor and 8GB RAM.

In future, there are also several open questions deserving to study. To name a few:

(1) How about the effectiveness of the DBSCAN for protecting other machine learning problems? For the SVM, we can assume that each poisoning point has a perturbation distance at least $\delta_1$ (since any simple label flipping will result in a large distance); but for other problems, such as regression, we cannot simply follow the same assumption.

(2) How about the ensemble methods? For example, can we take the ensemble of different robust SVM methods or outlier removal methods to achieve a more convincing result? Cretu *et al.* [17] and Biggio *et al.* [6] proposed the "voting" and "bagging" ideas for fighting attacks, but our understanding on their effectiveness in theory is still far from being satisfactory.

(3) What about the complexities of other machine learning problems under the adversarially-resilient formulations as Definition 1. Mount *et al.* [40] proved that it is impossible to achieve even an approximate solution for the linear regression with outliers problem within polynomial time under the conjecture of *the hardness of affine degeneracy* [20], if the dimensionality $d$ is not fixed. Simonov *et al.* [48] showed that unless Exponential Time Hypothesis fails, it is impossible not only to solve the PCA with outliers problem exactly but even to approximate it within a constant factor. For a large number of other adversarial machine learning problems, however, the study of their complexities is still in its infancy.

## References

[1] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. *ACM Sigmod Record*, 30(2):37–46, 2001.

[2] L. Amsaleg, J. Bailey, D. Barbe, S. M. Erfani, M. E. Houle, V. Nguyen, and M. Radovanovic. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *2017 IEEE Workshop on Information Forensics and Security, WIFS 2017, Rennes, France, December 4-7, 2017*, pages 1–6. IEEE, 2017.

[3] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In F. Lin, D. Lee, B. P. Lin, S. Shieh, and S. Jajodia, editors, *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006, Taipei, Taiwan, March 21-24, 2006*, pages 16–25. ACM, 2006.

[4] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The r*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, USA, May 23-25, 1990*, pages 322–331, 1990.

[5] B. Biggio, S. R. Bulò, I. Pillai, M. Mura, E. Z. Mequanint, M. Pelillo, and F. Roli. Poisoning complete-linkage hierarchical clustering. In P. Fränti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, editors, *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, volume 8621 of *Lecture Notes in Computer Science*, pages 42–52. Springer, 2014.

[6] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In *Multiple Classifier Systems - 10th International Workshop, MCS 2011, Naples, Italy, June 15-17, 2011. Proceedings*, pages 350–359, 2011.

[7] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[8] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[9] N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *J. Comput. Syst. Sci.*, 75(6):323–335, 2009.

[10] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[11] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3), 2011.

[12] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 642–651. Society for Industrial and Applied Mathematics, 2001.

[13] S. Chawla and A. Gionis. k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 189–197. SIAM, 2013.

[14] K. Chen. A constant factor approximation algorithm for k-median clustering with outliers. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 826–835. Society for Industrial and Applied Mathematics, 2008.

[15] A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, 5:1007–1034, 2004.

[16] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273, 1995.

[17] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, and A. D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 81–95. IEEE Computer Society, 2008.

[18] D. J. Crisp and C. J. C. Burges. A geometric interpretation of v-SVM classifiers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *NIPS*, pages 244–250. The MIT Press, 1999.

[19] N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, and D. Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 99–108, 2004.

[20] J. Erickson and R. Seidel. Better lower bounds on detecting affine and spherical degeneracies. *Discrete & Computational Geometry*, 13:41–57, 1995.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

[22] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In L. Fortnow and S. P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578. ACM, 2011.

[23] J. Feng, H. Xu, S. Mannor, and S. Yan. Robust logistic regression and classification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 253–261, 2014.

[24] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[25] J. Gan and Y. Tao. DBSCAN revisited: mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 519–530. ACM, 2015.

[26] B. Gao and J. Wang. A fast and robust TSVM for pattern classification. *CoRR*, abs/1711.05406, 2017.

[27] I. J. Goodfellow, P. D. McDaniel, and N. Papernot. Making machine learning robust against adversarial inputs. *Commun. ACM*, 61(7):56–66, 2018.

[28] L. Huang, S. Jiang, J. Li, and X. Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 814–825, 2018.

[29] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[30] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 19–35, 2018.

[31] T. Kanamori, S. Fujiwara, and A. Takeda. Breakdown point of robust support vector machines. *Entropy*, 19(2):83, 2017.

[32] M. Khoury and D. Hadfield-Menell. Adversarial training with voronoi constraints. *CoRR*, abs/1905.01019, 2019.

[33] P. W. Koh, J. Steinhardt, and P. Liang. Stronger data poisoning attacks break data sanitization defenses. *CoRR*, abs/1811.00741, 2018.

[34] H.-P. Kriegel, P. Kröger, and A. Zimek. Outlier detection techniques. *Tutorial at PAKDD*, 2009.

[35] R. Laishram and V. V. Phoha. Curie: A method for protecting SVM classifier from poisoning attack. *CoRR*, abs/1606.01584, 2016.

[36] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.

[37] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[38] N. Megiddo. On the complexity of some geometric problems in unbounded dimension. *J. Symb. Comput.*, 10(3/4):327–334, 1990.

[39] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In B. Bonet and S. Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2871–2877. AAAI Press, 2015.

[40] D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. On the least trimmed squares estimator. *Algorithmica*, 69(1):148–183, 2014.

[41] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1196–1204, 2013.

[42] A. Paudice, L. Muñoz-González, A. György, and E. C. Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *CoRR*, abs/1802.03041, 2018.

[43] A. Paudice, L. Muñoz-González, and E. C. Lupu. Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops - Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings*, pages 5–15, 2018.

[44] P. J. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley, 1987.

[45] B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S. Lau, S. Rao, N. Taft, and J. D. Tygar. ANTIDOTE: understanding and defending against poisoning of anomaly detectors. In A. Feldmann and L. Mathy, editors, *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference, IMC 2009, Chicago, Illinois, USA, November 4-6, 2009*, pages 1–14. ACM, 2009.

[46] B. Scholkopf, A. J. Smola, K. R. Muller, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[47] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):19, 2017.

[48] K. Simonov, F. V. Fomin, P. A. Golovach, and F. Panolan. Refined complexity of PCA with outliers. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5818–5826, 2019.

[49] J. Steinhardt, P. W. Koh, and P. Liang. Certified defenses for data poisoning attacks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3517–3529, 2017.

[50] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[51] D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognit. Lett.*, 20(11-13):1191–1199, 1999.

[52] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[53] S. Weerasinghe, S. M. Erfani, T. Alpcan, and C. Leckie. Support vector machines resilient against training data integrity attacks. *Pattern Recognit.*, 96, 2019.

[54] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.

[55] H. Xiao, H. Xiao, and C. Eckert. Adversarial label flips attack on support vector machines. In L. D. Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. J. F. Lucas, editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier,*

*France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 870–875. IOS Press, 2012.

[56] G. Xu, Z. Cao, B. Hu, and J. C. Príncipe. Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognit.*, 63:139–148, 2017.

[57] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, 2009.

[58] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *AAAI*, pages 536–542. AAAI Press, 2006.

## 6 Empirical Experiments

We compare several defenses including the DBSCAN and robust SVM methods, and study the trends of their classification accuracies with varying three values: the poisoned fraction, the intrinsic dimensionality, and the Euclidean dimensionality. All the experimental results were obtained on a Windows 10 workstation equipped with an Intel core $i5$-$8265U$ processor and 8GB RAM.

Table 1: Datasets

| Dataset | Size | Dimension |
|---|---|---|
| SYNTHETIC | 10000 | $50 - 200$ |
| LETTER | 1520 | 16 |
| MUSHROOMS | 8124 | 112 |

**Datasets.** We consider both the synthetic and real datasets in our experiments. For each synthetic dataset, we generate two manifolds in $\mathbb{R}^d$, where $d$ is between 50 and 200, and each manifold is represented by a random polynomial function with degree ranging from 25 to 65. Note that it is challenging to achieve the exact doubling dimensions of the datasets, so we use the degree of the polynomial function as a “rough indicator” for the doubling dimension (the higher the degree, the larger the doubling dimension). In each of the manifolds, we sample 5000 points; specifically, the data is randomly partitioned into 30% and 70% respectively for training and testing. We also consider two real datasets: the LETTER and MUSHROOMS datasets from LibSVM [11]. The details of the datasets are shown in Table 1.

**Attack and Defenses.** We generate the adversarial label-flipping attacks through the free software ALFASVMLib [54]. We evaluate the performances of 9 different defenses using their publicly available implementations:

- The basic SVM classification algorithm **C-SVC** [11];
- The robust SVM algorithm **RSVM-**$S$ based on the rescaled hinge loss function [56], where the parameter $S$ indicates the iteration number of the half-quadratic optimization (*e.g.,* we set $S = 3$ and 10 following the paper [56]);
- The fast and robust twin support vector machine **FRTSVM** [26];
- The **L2** defense [33], which removes points that are far from their class centroids in $L_2$ distance;
- The **SLAB** defense [49], which first projects points onto the line between the class centroids, and then removes points that are too far from the class centroids;
- The **LOSS** defense [33], which discards points that are not well fit by a model trained (without any data sanitization) on the full dataset;
- The **K-NN** defense [33], which removes points that are far from their k nearest neighbors (we set $k = 5$ as [33]);
- The **SVD** defense [33], which assumes that the clean data lies in some low-rank subspace, and that poisoned data therefore will have a large component out of this subspace [45];
- The **DBSCAN** method [47] that is implemented as Remark 3 (ii).

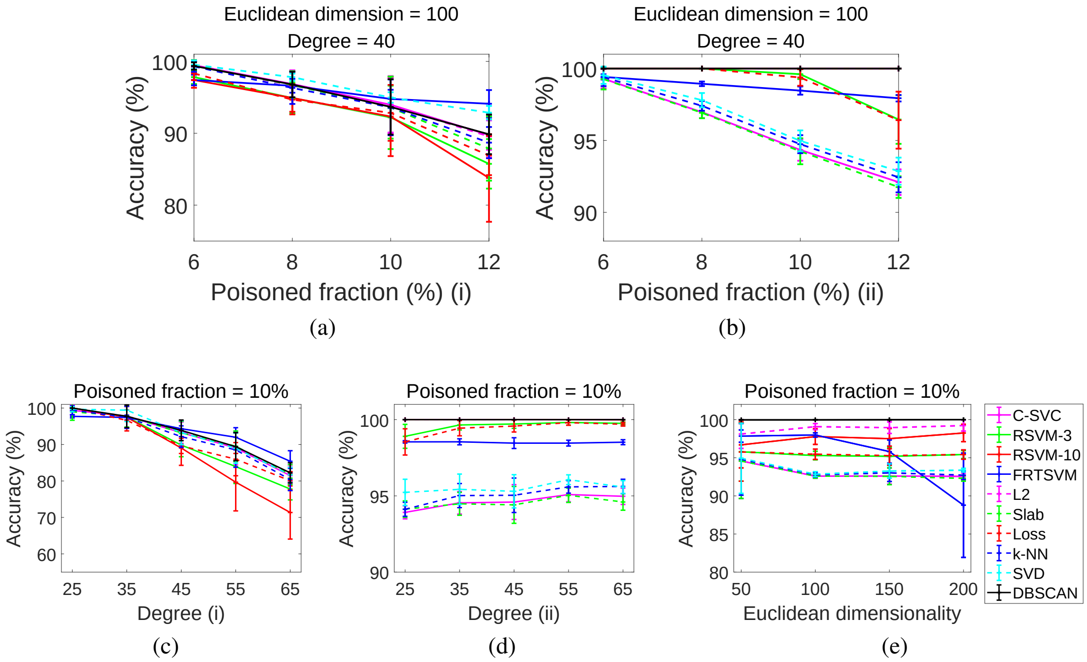
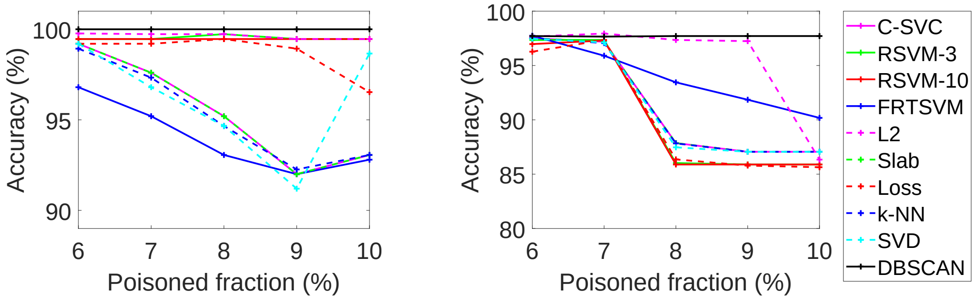


Figure 2: The performances on the synthetic datasets.

For the 6 data sanitization defenses, we run the SVM algorithm **C-SVC** on the cleaned data to compute their final solutions.

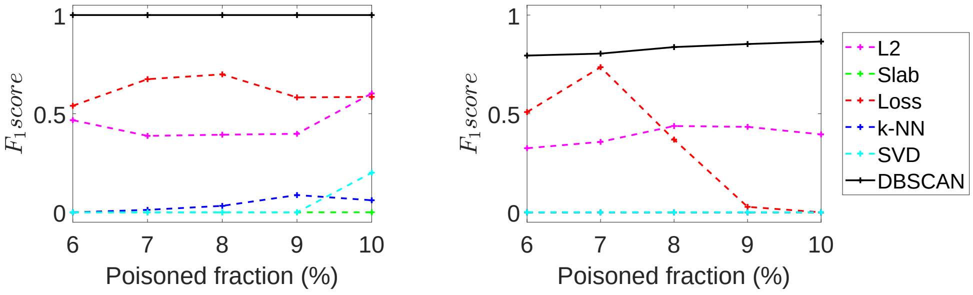
**The results.** In Figure 2, we illustrate the results on the synthetic datasets. We consider two different cases: (i) the two manifolds are overlapped with each other and (ii) the two manifolds are separated. For case (i), all the defenses obtain lower accuracies when the poisoned fraction increases as seen in Figure 2a; for case (ii), the performance of DBSCAN keeps much more stable comparing with other defenses when varying the poisoned fraction in Figure 2b. We also study the influence from the intrinsic dimensionality in these two cases. We set the Euclidean dimensionality to be $100$ and vary the polynomial function's degree from $25$ to $65$. In case (i), from Figure 2c we can observe that the accuracy of each defense dramatically decreases when the degree increases, which is in agreement with our theoretical analysis. However, for case (ii), from Figure 2d we can see that the accuracies are not substantially affected by the intrinsic dimensionality; we believe that it is due to the fact that case (ii) is relatively easier for classification, as long as the two classes are well separated. Finally, we fix the degree to be $20$ and vary the Euclidean dimensionality in Figure 2e; we can see that the influence from the Euclidean dimensionality is small (except for FRTSVM).

For the real datasets, we set the poisoned fraction to be $6\%$-$10\%$ in Figure 3; the experimental results reveal the similar trends with the synthetic datasets. To further investigate the performances of these data sanitization defenses, we plot their $F_1$ score curves in Figure 4 (the score is the harmonic mean of precision and recall for the outlier removal). We can see that DBSCAN always outperforms other data sanitization defenses.



(a) LETTER (b) MUSHROOMS

Figure 3: The performances on the real datasets.



(a) LETTER (b) MUSHROOMS

Figure 4: $F_1$ scores.