Defending SVMs Against Poisoning Attacks: The Hardness and DBSCAN Approach

Hu Ding[1] **Fan Yang**[1] **Jiawei Huang**[1]

[1]School of Computer Science and Technology, University of Science and Technology of China, He Fei, China

Abstract

Adversarial machine learning has attracted a great amount of attention in recent years. Due to the great importance of support vector machines (SVM) in machine learning, we consider defending SVM against poisoning attacks in this paper. We study two commonly used 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 hardness of designing a quality-guaranteed adversarially-resilient SVM algorithm is still quite limited. We are the first, to the best of our knowledge, to prove that even the simplest hard-margin one-class SVM with adversarial outliers problem is NP-complete, and has no fully PTAS unless P=NP. For data sanitization, we explain the effectiveness of DBSCAN (as a density-based outlier removal method) for defending against poisoning attacks. In particular, we link it to the intrinsic dimensionality by proving a sampling theorem in doubling metrics. In our empirical experiments, we systematically compare several defenses including the DBSCAN and robust SVM methods, and investigate the influences from the intrinsic dimensionality and poisoned fraction 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 technologies to many different fields, such as data mining, networking, and bioinformatics. However, with its territory rapidly enlarging, machine learning has also imposed a number of new challenges. In particular, *adversarial machine learning* which concerns about the potential vulnerabilities of the algorithms, has attracted a great amount of attention [Barreno et al., 2006, Huang et al., 2011, Biggio and Roli, 2018, Goodfellow et al., 2018]. As mentioned in the survey paper [Biggio and Roli, 2018], the very first work of adversarial machine learning dates back to 2004, in which Dalvi et al. [2004] 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** [Biggio and Roli, 2018]. 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. [2014] observed that small perturbation to a test image can arbitrarily change the neural network's prediction.

In this paper, we focus on poisoning attacks that happen at training time. Usually, the adversary injects a small number of specially crafted samples into the training data which can make the decision boundary severely deviate; in particular, because open datasets are commonly used to train our machine learning algorithms nowadays, poisoning attack has become a key security issue that seriously limits real-world applications [Biggio and Roli, 2018]. For instance, even a small number of poisoning samples can significantly increase the test error of support vector machine (SVM) [Biggio et al., 2012, Mei and Zhu, 2015, Xiao et al., 2012]. Beyond linear classifiers, a number of works studied the poisoning attacks for other machine learning problems, such as clustering [Biggio et al., 2014], PCA [Rubinstein et al., 2009], and regression [Jagielski et al., 2018].

Though lots of works focused on constructing poisoning attacks, our ultimate goal is to design defenses. Poisoning samples can be regarded as *outliers*, and this leads to two natural approaches to defend: **(1) data sanitization defense**, *i.e.,* first perform outlier removal and then run an existing machine learning algorithm on the cleaned data [Cretu et al., 2008], or **(2) directly design a robust optimization algo-**

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

rithm that is resilient against outliers [Christmann and Steinwart, 2004, Jagielski et al., 2018].

Steinhardt et al. [2017] 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 [2016] applied the seminal DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method [Ester et al., 1996] to remove outliers for SVM and showed that it can successfully identify most of the poisoning data. However, their DBSCAN approach is lacking of theoretical analysis. Several other outlier removal methods for fighting poisoning attacks have also been studied recently [Paudice et al., 2018b,a]. Also, it is worth noting that outlier removal actually is an independent topic that has been extensively studied in various fields before [Chandola et al., 2009].

The other defense strategy, designing robust optimization algorithms, also has a long history in the machine learning community. A substantial part of robust optimization algorithms rely on the idea of regularization. For example, Xu et al. [2009] studied the relation between robustness and regularization for SVM; Zhu et al. [2003] proposed the 1-norm SVM to enhance the robustness to noise; other robust SVM algorithms include [Tax and Duin, 1999, Xu et al., 2006, Natarajan et al., 2013, Ding and Xu, 2015, Xu et al., 2017, Kanamori et al., 2017]. However, as discussed in [Mei and Zhu, 2015, Jagielski et al., 2018], 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 [Feng et al., 2014, Weerasinghe et al., 2019], but these assumptions are usually not well satisfied in practice. An alternative approach is to explicitly remove outliers during optimization, such as the "trimmed" method for robust regression [Jagielski et al., 2018]; but this is often a challenging **combinatorial optimization problem**: if z of the input n data items are outliers ($z < n$), (at first glance) we have to consider an exponentially large number $\binom{n}{z}$ of different possible cases in the adversarial setting.

1.1 OUR CONTRIBUTIONS

Due to the great importance in machine learning [Chang and Lin, 2011], we focus on defending SVM against poisoning attacks in this paper. Our contributions are twofold.

(i). 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. Though its local optimum can be obtained by using various methods, such as the alternating minimization approach [Jagielski et al., 2018], it is often very challenging to achieve a quality guaranteed solution for such adversarial-resilience optimization problem. For instance, Simonov et al. [2019] showed that unless the Exponential Time Hypothesis (ETH) fails, it is impossible not only to solve the *PCA with outliers* problem exactly but even to approximate it within a constant factor. A similar hardness result was also proved for *linear regression with outliers* by Mount et al. [2014]. Some other hardness results for robust optimization problems were studied in [Bernholt, 2006]. But for SVM with outliers, we are unaware of any **hardness-of-approximation result** before. We try to bridge the gap in the current state of knowledge in Section 3. 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. So it is quite unlikely that one can achieve a (nearly) optimal solution in polynomial time.

(ii). Second, we investigate the DBSCAN based data sanitization defense 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 [Ester et al., 1996, Schubert et al., 2017]; 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 [Laishram and Phoha, 2016]. We assume the original input training data (before poisoning attack) is large and dense enough in the domain Ω; 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 sparser the outliers appear to be (and the more efficiently the DBSCAN performs).

Thus, to guarantee the effectiveness of the DBSCAN approach, a fundamental question in theory is what about **the lower bound of the data size** n (we can assume that the original input data is a set of *i.i.d.* samples drawn from the domain Ω). However, to achieve a favorable lower bound is a non-trivial task. The VC dimension [Li et al., 2001] 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 can be very large. Our idea is motivated by the recent observations on the link between the adversarial vulnerability and the intrinsic dimensionality [Khoury and Hadfield-Menell, 2019, Amsaleg et al., 2017, Ma et al., 2018]. We prove a lower bound of n that depends on the

intrinsic dimension of Ω and is independent of the feature space's dimensionality.

Our result strengthens the observation from Steinhardt et al. [2017] 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 on evasion attacks [Khoury and Hadfield-Menell, 2019, Amsaleg et al., 2017, Ma et al., 2018], our result links poisoning attacks to intrinsic dimensionality (independent of our work, Weerasinghe et al. [2021] recently also studied the relation between intrinsic dimension and poisoning attacks). In Section 5, we investigate several popular defending methods (including DBSCAN), where the intrinsic dimension of data demonstrates significant influence on their defending performances.

2 PRELIMINARIES

Given two point sets P^+ and P^- in $\mathbb{R}^d$, the problem of linear **support vector machine (SVM)** [Chang and Lin, 2011] is to find the maximum margin separating these two point sets (if they are separable). 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 [Cortes and Vapnik, 1995], ν-SVM [Scholkopf et al., 2000, Crisp and Burges, 1999], and Core-SVM [Tsang et al., 2005]. 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 $\mathcal{K}(p_1, p_2)$.

Poisoning attacks. An adversary can inject some bad points to the original data set $P^+ \cup P^-$. For instance, the adversary can take 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 $(0, 1)$). We can formulate the defense against poisoning attacks as the following combinatorial optimization problem. As mentioned in Section 1.1, it is sufficient to consider only the simpler hard-margin case for studying its hardness.

Definition 1 (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* (i.e., *the distance between the two parallel hyperplanes bounding the margin) separating P_1^+ and P_1^- is maximized.*

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

Remark 1. *The model proposed in Definition 1 follows the popular **data trimming** idea from robust statistics [Rousseeuw and Leroy, 1987]. As an example similar with Definition 1, Jagielski et al. [2018] proposed a data trimming based regression model to defend against poisoning attacks.*

We also need to clarify the intrinsic dimensionality for our following analysis. **Doubling dimension** is a measure of intrinsic dimensionality that has been widely adopted in the learning theory community [Bshouty et al., 2009]. Given a point p and $r \geq 0$, we use $\mathbb{B}(p, r)$ to indicate the ball of radius r around p in the space.

Definition 2 (Doubling Dimension). *The doubling dimension of a point set P from some metric space[1] is the smallest number ρ, such that for any $p \in P$ and $r \geq 0$, the set $P \cap \mathbb{B}(p, 2r)$ can always be covered by the union of at most 2^ρ balls with radius r in the space.*

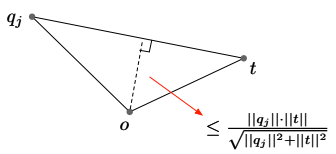
To understand doubling dimension, we consider the following simple case. If the points of P distribute uniformly in a d'-dimensional flat in $\mathbb{R}^d$, then it is easy to see that P has the doubling dimension $\rho = O(d')$, which is independent of the Euclidean dimension d (*e.g.,* d can be much higher than ρ). Intuitively, doubling dimension is used for describing the expansion rate of a given point set in the space. It is worth noting that the intrinsic dimensionality described in [Amsaleg et al., 2017, Ma et al., 2018] is quite similar to doubling dimension, which also measures expansion rate.

3 THE HARDNESS RESULT

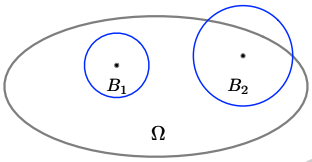
In this section, we prove that even the one-class SVM with outliers problem is NP-complete and has no fully PTAS unless P=NP (that is, we cannot achieve a polynomial time $(1 - \epsilon)$-approximation for any given $\epsilon \in (0, 1)$). Our idea is partly inspired by the result from Megiddo [1990]. 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.*

[1] The space can be a Euclidean space or an abstract metric space.



(a)



(b)

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

Let Γ 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_Γ 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 $q_j = (q_{j,1}, q_{j,2}, \cdots, q_{j,l+1})$ as follows. For $1 \leq i \leq l$,

$$q_{j,i} = \begin{cases} \alpha, & \text{if } u_i \text{ occurs in } E_j; \\ -\alpha, & \text{else if } \bar{u}_i \text{ occurs in } E_j; \\ 0, & \text{otherwise.} \end{cases}$$

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

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

The value of α will be determined later. Let Q denote the set $\{q_1, \cdots, q_m\}$. Now, we construct the instance $P_\Gamma = U \cup Q$ of one-class SVM with outliers, where the number of points $n = 2(l+1) + m$ and the number of outliers $z = l + 1$. Then we have the following lemma.

Lemma 1. *Let $\alpha > 1/2$. Γ has a satisfying assignment if and only if P_Γ has a solution with margin width $\frac{1}{\sqrt{l+1}}$.*

Proof. **First, we suppose there exists a satisfying assignment $\mathcal{A}(\Gamma)$ for Γ.** We define the set $S \subset P_\Gamma$ as follows. If u_i is true in $\mathcal{A}(\Gamma)$, we include e_i in S, else, we include $-e_i$ in S; we also include e_{l+1} in S. We claim that the set $S \cup Q$ yields a solution of the instance P_Γ 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 should be $\frac{1}{l+1}$, else, the i-th position of t should be $-\frac{1}{l+1}$; the $(l+1)$-th position of t is $\frac{1}{l+1}$. Obviously, $||t|| = \frac{1}{\sqrt{l+1}}$. Let $\mathcal{H}_t$ be the hyperplane that is orthogonal to the vector $t - o$ and passing through t. It is easy to know $\mathcal{H}_t$ separates S and o, and the margin width (*i.e.,* the distance between the origin and $\mathcal{H}_t$) is $||t|| = \frac{1}{\sqrt{l+1}}$. Furthermore, for any point $q_j \in Q$, since there exists at least one true variable in E_j, we have the inner product

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

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 $\frac{1}{\sqrt{l+1}}$.

Second, suppose the instance P_Γ 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 set of inliers so as to guarantee that there are $n - z$ inliers in total. We still use t to denote the mean point of S ($||t|| = \frac{1}{\sqrt{l+1}}$). Now, we design the assignment $\mathcal{A}(\Gamma)$ for Γ: if $e_i \in S$, we assign u_i to be true, else, we assign $\bar{u}_i$ to be true. We claim that Γ 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 the inner product

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

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 have the width at most

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

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

Overall, Γ has a satisfying assignment if and only if P_Γ has a solution with margin width $\frac{1}{\sqrt{l+1}}$. □

Now we are ready to prove the theorem.

Proof. (**of Theorem 1**) Since 3-SAT is NP-complete, Lemma 1 implies that the one-class SVM with outliers problem is NP-complete too; otherwise, we can determine that

whether a given instance Γ is satisfiable by computing the optimal solution of P_Γ. Moreover, the gap between $\frac{1}{\sqrt{l+1}}$ and $\frac{||q_j||\cdot||t||}{\sqrt{||q_j||^2+||t||^2}}$ (from the formula (4)) is

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

if we assume α 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 Γ is satisfiable if and only if any $(1-\epsilon)$-approximation of the instance P_Γ 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 Γ is satisfiable or not in polynomial time. In other words, we cannot even 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. 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 Ω who has the doubling dimension $\rho > 0$. For ease of presentation, we assume that Ω lies on a manifold $\mathcal{F}$ in the space. Let Δ denote the diameter of Ω, *i.e.,* $\Delta = \sup_{p_1,p_2\in\Omega}||p_1 - p_2||$. Also, we let f be the probability density function of the data distribution over Ω.

To measure the uniformity of f, we define a value λ as follows. For any $c \in \Omega$ and any $r > 0$, we say "the ball $\mathbb{B}(c, r)$ is enclosed by Ω" 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 Ω or the radius r is too large, the ball will not be enclosed by Ω. See Figure 1b for an illustration. We define $\lambda := \sup_{c,c',r}\frac{\int_{\mathbb{B}(c',r)}f(x)\,\mathrm{d}x}{\int_{\mathbb{B}(c,r)}f(x)\,\mathrm{d}x}$, where $\mathbb{B}(c, r)$ and $\mathbb{B}(c', r)$ are any two equal-sized balls, and $\mathbb{B}(c, r)$ is required to be enclosed by Ω. As a simple example, if Ω lies on a flat manifold and the data uniformly distribute over Ω, the value λ 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 λ can be high.

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

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

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

Remark 2. ***(i)*** *A highlight of Theorem 2 is that the lower bound of $|P|$ is independent of the dimensionality of the input space (which could be much higher than the intrinsic dimension). Moreover, our result holds for any metric space with bounded doubling dimension (not only for Euclidean space).*

(ii) *For the simplest case that Ω lies on a flat manifold and the data uniformly distribute over Ω, λ 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 ρ to the VC dimension $\mathtt{dim}$ of the range space consisting of all balls with different radii [Li et al., 2001]. Unfortunately, Huang et al. [2018] 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 "*ϵ-smoothed distance function*": $g(p, q) \in (1 \pm \epsilon)||p - q||$ for any two data points p and q, where $\epsilon \in (0, \frac{1}{8})$. Given a point p and $\delta > 0$, the ball defined by this distance function $g(\cdot,\cdot)$ is denoted by $\mathbb{B}_g(p, \delta) = \{q \in \text{the input space} \mid g(p, q) \le \delta\}$.

Theorem 3 (Huang et al. [2018])**.** *Suppose the point set P has the doubling dimension $\rho > 0$. There exists an ϵ-smoothed distance function "$g(\cdot,\cdot)$" such that the VC dimension[2] $\mathtt{dim}_\epsilon$ of the range space consisting of all balls with different radii is at most $\tilde{O}(\frac{\rho}{\epsilon^\rho})$, if replacing the distance by $g(\cdot,\cdot)$.*

Proof. **(of Theorem 2)** Let r be any positive number. First, since the doubling dimension of Ω is ρ, if recursively applying Definition 2 $\log\frac{\Delta}{r}$ times, we know that Ω can be

[2]Huang et al. [2018] 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 [Feldman and Langberg, 2011]. For example, if the shattering dimension is ρ_0, the VC dimension should be bounded by $O(\rho_0 \log\rho_0)$.

covered by at most $\Theta\Big(\big(\frac{\Delta}{r}\big)^{\rho}\Big)$ balls with radius r. Thus, if $\mathbb{B}(c,r)$ is enclosed by Ω, we have

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

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 ϵ-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 ϵ_0-sample of Ω; that is, for any point c 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[3] [Li et al., 2001]. Because $g(\cdot,\cdot)$ is an ϵ-smoothed distance function of the Euclidean distance, we have

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

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

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

The last inequality comes from (7) (since we assume the ball $\mathbb{B}\big(c,\delta\big)$ is enclosed by Ω, 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

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

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

[3]The exact probability comes from the success probability that P is an ϵ_0-sample of Ω. Let $\eta\in(0,1)$, and the size $|P|$ in (8) should be at least $\Theta\big(\frac{1}{\epsilon_0^2}(\mathtt{dim}_{\epsilon}\log\frac{\mathtt{dim}_{\epsilon}}{\epsilon_0}+\log\frac{1}{\eta})\big)$ to guarantee a success probability $1-\eta$. For convenience, we assume η is a fixed small constant and simply say "$1-\eta$" is a "constant probability".

4.2 THE DBSCAN APPROACH

For the sake of completeness, we briefly introduce the method of DBSCAN [Ester et al., 1996]. 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 core or border 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 DBSCAN is to identify these clusters and the outliers. Several efficient implementations for DBSCAN can be found in [Gan and Tao, 2015, Schubert et al., 2017].

Following Section 4.1, we assume that P is a set of *i.i.d.* samples drawn from the connected and compact domain Ω 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 Ω. 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 effectiveness of the DBSCAN with respect to the poisoned dataset $P\cup Q$. We assume the poisoned fraction $\frac{|Q|}{|P|}=\frac{z}{|P|}<1$.

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. If we set $r=\delta$ and $\mathtt{MinPts}=m$, and run DBSCAN on the poisoned dataset $P\cup Q$, then the obtained largest cluster is exactly the set P. In other word, the set Q consists of 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 Ω, or p is covered by some ball $\mathbb{B}(q,\delta)$ enclosed by Ω. 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 $p\in P$ should not belong to the same cluster of the DBSCAN. Also, because the domain Ω is connected and compact, the set P must 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 ρ, the lower the poisoned fraction $\frac{z}{|P|}$ (and the easier corrupting the DBSCAN defense). In addition, when δ is large,* i.e., *each poisoning point has large perturbation distance and $\partial\Omega$ is sufficiently smooth, it will be relatively easy for 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 defending performance 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 cannot obtain the exact values of δ and m. We follow the strategy that was commonly used in the DBSCAN implementations [Gan and Tao, 2015, Schubert et al., 2017]; we set* `MinPts` *to be a small constant and tune the value of r until the largest cluster has $|P \cup Q| - z$ points.*

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 Ω^+ and Ω^- with doubling dimension ρ. 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 EMPIRICAL EXPERIMENTS

All the experiments were repeated 20 times on a Windows 10 workstation equipped with an Intel core $i5$-8400 processor and 8GB RAM. To generate the poisoning attacks, we use the **MIN-MAX** attack from [Koh et al., 2018] and the adversarial label-flipping attack **ALFA** from ALFASVMLib [Xiao et al., 2015]. We evaluate the defending performances of the basic SVM algorithms and several different defenses by using their publicly available implementations.

1. We consider both the cases that not using and using kernel. For SVM without kernel, we directly use **LINEAR SVM** as the basic SVM algorithm; for SVM with kernel, we consider RBF kernel (**RBF SVM**). Both the implementations are from [Chang and Lin, 2011].
2. The recently proposed robust SVM algorithm **RSVM-S** based on the rescaled hinge loss function [Xu et al., 2017]. The parameter "S" indicates the iteration number of the half-quadratic optimization (*e.g.,* we set $S = 3$ and 10 following their paper's setting). The algorithm also works fine when using a kernel.
3. The **DBSCAN** method [Schubert et al., 2017] implemented as Remark 3 (ii). We set `MinPts` $= 5$ (our empirical study finds that the difference is minor within the range $[3, 10]$).
4. The data sanitization defenses from [Koh et al., 2018] based on the spatial distribution of input data, which include **SLAB**, **L2**, **LOSS**, and **K-NN**.

For the data sanitization defenses, we run them on the poisoned data in the original input space; then, apply the basic SVM algorithm, LINEAR SVM or RBF SVM (if using RBF kernel), on the cleaned data to compute their final solutions.

Table 1: Datasets

Dataset	Size	Dimension
SYNTHETIC	10000	50-200
LETTER	1520	16
MUSHROOMS	8124	112
SATIMAGE	2236	36

Datasets. We consider both the synthetic and real-world datasets in our experiments. For each synthetic dataset, we generate two manifolds in $\mathbb{R}^d$, and each manifold is represented by a random polynomial function with degree d' (the values of d and d' will be varied in the experiments). Note that it is challenging to achieve the exact doubling dimensions of the datasets, and thus 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 randomly sample 5000 points; the data is randomly partitioned into 30% and 70% respectively for training and testing, and we report the classification accuracy on the test data. We also consider three real-world datasets from [Chang and Lin, 2011]. The details are shown in Table 1.

Results. First, we study the influence from the intrinsic dimensionality. We set the Euclidean dimensionality d to be 100 and vary the polynomial function's degree d' from 25 to 65 in Figure 2a and 2d. Then, we fix the degree d' to be 40 and vary the Euclidean dimensionality d in Figure 2b and 2e. We can observe that the accuracies of most methods dramatically decrease when the degree d' (intrinsic dimension) increases, and the influence from the intrinsic dimension is more significant than that from the Euclidean dimension.

We also study their classification performances under different poisoned fraction in Figure 2c and 2f. We can see that all the defenses yield lower accuracies when the poisoned fraction increases, while the performance of DBSCAN keeps much more stable compared with other defenses. Moreover, we calculate the widely used F_1 scores from the sanitization

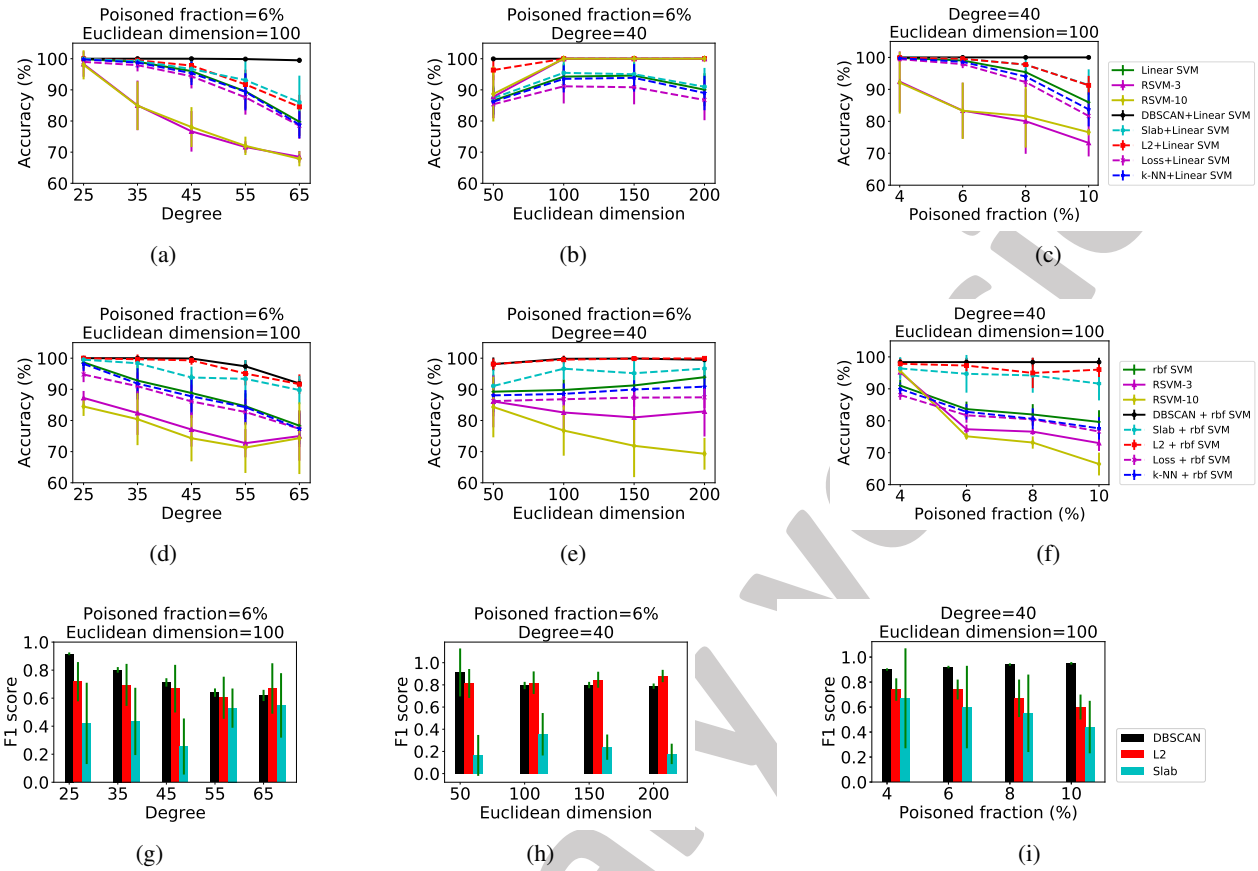


Figure 2: The classification accuracy on the SYNTHETIC datasets of Linear SVM (the first line) and SVM with RBF kernel (the second line) under MIN-MAX attack. The third line are the average F_1 scores.

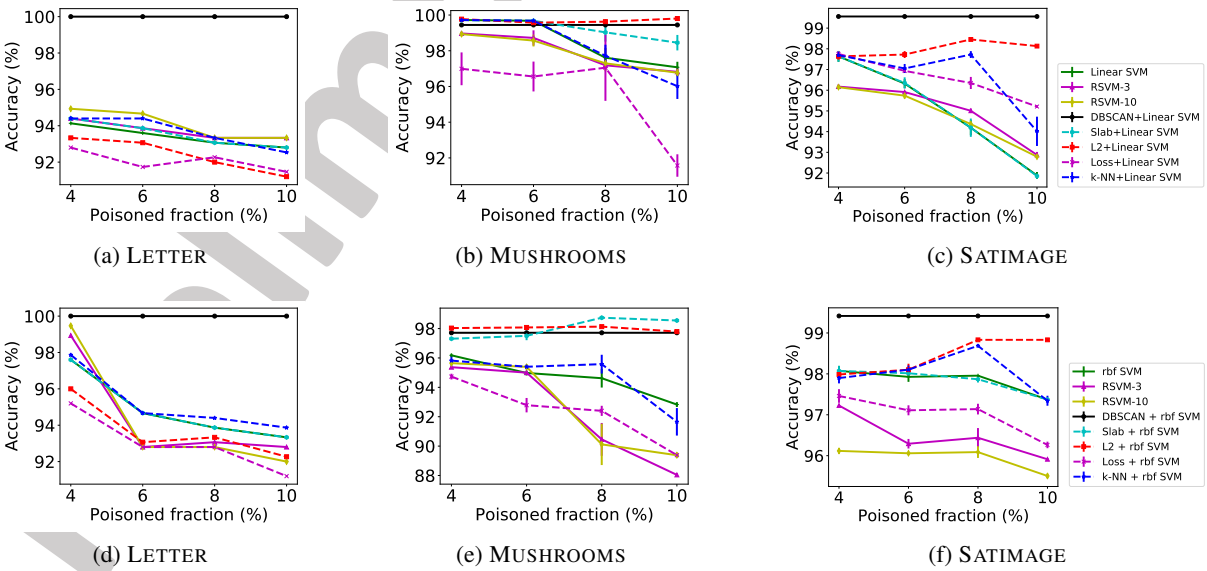
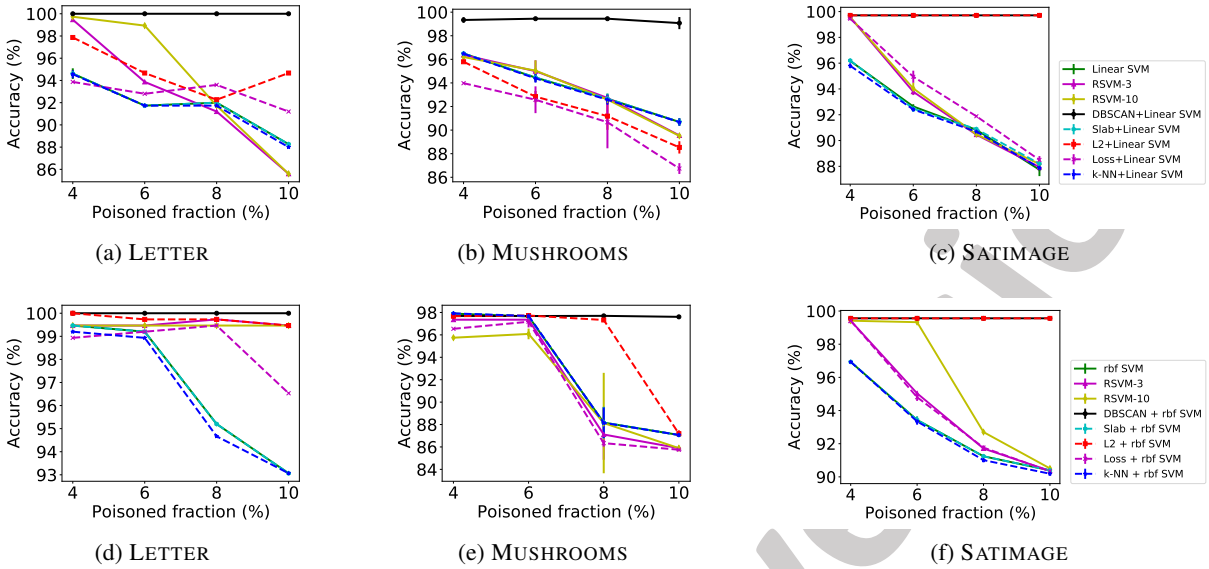


Figure 3: The classification accuracy on the real datasets of linear SVM (the first line) and SVM with RBF kernel (the second line) under MIN-MAX attack.



(a) LETTER (b) MUSHROOMS (c) SATIMAGE

(d) LETTER (e) MUSHROOMS (f) SATIMAGE

Figure 4: The classification accuracy of linear SVM (the first line) and SVM with RBF kernel (the second line) under ALFA attack.

Table 2: F_1 scores on MUSHROOM dataset.

	ALFA				MIN-MAX			
	4%	6%	8%	10%	4%	6%	8%	10%
DBSCAN	**0.72**	**0.79**	**0.84**	**0.86**	**0.72**	**0.79**	**0.84**	**0.87**
SLAB	< 0.1	< 0.1	< 0.1	< 0.1	0.17	0.22	0.27	0.30
L2	0.34	0.37	0.40	0.40	0.67	0.66	0.65	0.69
LOSS	0.11	0.60	0.37	< 0.1	0.14	0.28	0.37	< 0.1
KNN	< 0.1	< 0.1	< 0.1	< 0.1	0.17	0.11	< 0.1	< 0.1

defenses for identifying the outliers. LOSS and K-NN both yield very low F_1 scores (< 0.1); that means they are not quite capable to identify the real poisoning data items. The F_1 scores yielded by DBSCAN, L2 and SLAB are shown in Figure 2g-2i, where DBSCAN in general outperforms the other two sanitization defenses for most cases.

We also perform the experiments on the real datasets under MIN-MAX attack and ALFA attack with the poisoned fraction ranging from 4% to 10%. The experimental results (Figure 3 and 4) reveal the similar trends as the results for the synthetic datasets, and DBSCAN keeps considerably better performance compared with other defenses. The F_1 scores on MUSHROOM dataset are shown in Table 2 (due to the space limit, the F_1 scores on the other two real datasets are placed in our full paper).

6 DISCUSSION

In this paper, we study two different strategies for protecting SVM against poisoning attacks. We also have several open questions to study in future. For example, what about the complexities of other machine learning problems under the adversarially-resilient formulations as Definition 1? For many other adversarial machine learning problems, the study on their complexities is still in its infancy.

Acknowledgements

The authors would like to thank Ruomin Huang and the anonymous reviewers for their helpful discussions and suggestions on improving this paper.

References

Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah M. Erfani, Michael E. Houle, Vinh Nguyen, and Milos Radovanovic. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *2017 IEEE Workshop on Information Forensics and Security, WIFS*, pages 1–6. IEEE, 2017.

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

Thorsten Bernholt. Robust estimators are hard to compute. Technical report, Technical report, 2006.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

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

Battista Biggio, Samuel Rota Bulò, Ignazio Pillai, Michele Mura, Eyasu Zemene Mequanint, Marcello Pelillo, and Fabio Roli. Poisoning complete-linkage hierarchical clustering. In *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR*, 2014.

Nader H. Bshouty, Yi Li, and Philip M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *J. Comput. Syst. Sci.*, 75(6):323–335, 2009.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3), 2011.

Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, 5:1007–1034, 2004.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273, 1995.

Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy*, pages 81–95. IEEE Computer Society, 2008.

David J. Crisp and Christopher J. C. Burges. A geometric interpretation of v-SVM classifiers. In *NIPS*, pages 244–250. The MIT Press, 1999.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108, 2004.

Hu Ding and Jinhui Xu. Random gradient descent tree: A combinatorial approach for svm with outliers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2561–2567, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei 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*, pages 226–231, 1996.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC*, pages 569–578. ACM, 2011.

Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. In *Advances in Neural Information Processing Systems*, pages 253–261, 2014.

Junhao Gan and Yufei 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, 2015.

Ian J. Goodfellow, Patrick D. McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Commun. ACM*, 61(7):56–66, 2018.

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

Lingxiao Huang, Shaofeng Jiang, Jian Li, and Xuan Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 814–825, 2018.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *Symposium on Security and Privacy, SP*, pages 19–35, 2018.

Takafumi Kanamori, Shuhei Fujiwara, and Akiko Takeda. Breakdown point of robust support vector machines. *Entropy*, 19(2):83, 2017.

Marc Khoury and Dylan Hadfield-Menell. Adversarial training with voronoi constraints. *CoRR*, abs/1905.01019, 2019.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *CoRR*, abs/1811.00741, 2018.

Ricky Laishram and Vir Virander Phoha. Curie: A method for protecting SVM classifier from poisoning attack. *CoRR*, abs/1606.01584, 2016.

Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *6th International Conference on Learning Representations, ICLR*. OpenReview.net, 2018.

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

Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, pages 2871–2877. AAAI Press, 2015.

David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. On the least trimmed squares estimator. *Algorithmica*, 69(1):148–183, 2014.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

Andrea Paudice, Luis Muñoz-González, András György, and Emil C. Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *CoRR*, abs/1802.03041, 2018a.

Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops*, pages 5–15, 2018b.

Peter J. Rousseeuw and Annick Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

Benjamin I. P. Rubinstein, Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J. D. Tygar. ANTIDOTE: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference, IMC*, pages 1–14. ACM, 2009.

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

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):19, 2017.

Kirill Simonov, Fedor V. Fomin, Petr A. Golovach, and Fahad Panolan. Refined complexity of PCA with outliers. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pages 5818–5826, 2019.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Neural Information Processing System*, pages 3517–3529, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

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

Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

Sandamal Weerasinghe, Sarah M. Erfani, Tansu Alpcan, and Christopher Leckie. Support vector machines resilient against training data integrity attacks. *Pattern Recognit.*, 96, 2019.

Sandamal Weerasinghe, Tansu Alpcan, Sarah M. Erfani, and Christopher Leckie. Defending support vector machines against data poisoning attacks. *IEEE Trans. Inf. Forensics Secur.*, 16:2566–2578, 2021.

Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial label flips attack on support vector machines. In *20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence*, volume 242, pages 870–875. IOS Press, 2012.

Huang Xiao, Battista Biggio, Blaine Nelson, Han Xiao, Claudia Eckert, and Fabio Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.

Guibiao Xu, Zheng Cao, Bao-Gang Hu, and José C. Príncipe. Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognit.*, 63:139–148, 2017.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, 2009.

Linli Xu, Koby Crammer, and Dale Schuurmans. Robust support vector machine training via convex outlier ablation. In *AAAI*, pages 536–542. AAAI Press, 2006.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Robert Tibshirani. 1-norm support vector machines. In *Advances in Neural Information Processing Systems 16 [NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 49–56. MIT Press, 2003.