Robust Nonparametric Regression under Poisoning Attack

Puning Zhao, Zhiguo Wan*

Zhejiang Lab
Hangzhou, Zhejiang, China
{pnzhao,wanzhiguo}@zhejianglab.com

Abstract

This paper studies robust nonparametric regression, in which an adversarial attacker can modify the values of up to q samples from a training dataset of size N. Our initial solution is an M-estimator based on Huber loss minimization. Compared with simple kernel regression, i.e. the Nadaraya-Watson estimator, this method can significantly weaken the impact of malicious samples on the regression performance. We provide the convergence rate as well as the corresponding minimax lower bound. The result shows that, with proper bandwidth selection, ℓ_∞ error is minimax optimal. The ℓ_2 error is optimal with relatively small q, but is suboptimal with larger q. The reason is that this estimator is vulnerable if there are many attacked samples concentrating in a small region. To address this issue, we propose a correction method by projecting the initial estimate to the space of Lipschitz functions. The final estimate is nearly minimax optimal for arbitrary q, up to a $\ln N$ factor.

Introduction

In the era of big data, it is common for some samples to be corrupted due to various reasons, such as transmission errors, system malfunctions, malicious attacks, etc. The values of these samples may be altered in any way, rendering many traditional machine learning techniques less effective. Consequently, evaluating the effects of these corrupted samples, and making corresponding robust strategies, have become critical tasks in the research community (Natarajan et al. 2013; Van Rooyen and Williamson 2017; Song et al. 2022).

Among all types of data contamination, adversarial attack is of particular interest in recent years (Biggio, Nelson, and Laskov 2012; Xiao et al. 2015; Jagielski et al. 2018; Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018; Mao et al. 2019), in which there exists a malicious adversary who aims at deteriorating our model performance. With this goal, the attacker alters the values of some samples using a carefully designed strategy. To cope with these attacks, robust statistics comes into being, which has been widely discussed in existing literatures (Huber 1981; Maronna et al. 2019). Several commonly used methods are trimmed mean, median-of-means and M-estimators. In recent years, many new methods are proposed for high dimensional problems with optimal statistical rates. These methods are summarized in (Steinhardt 2018; Diakonikolas and Kane 2019, 2023). For example, (Diakonikolas et al. 2016, 2017; Hopkins and Li 2018; Cheng et al. 2019) have solved some basic problems such as mean and covariance estimation. The idea of these research can then be used in machine learning problems with poisoning attack, which means that some training samples are modified by adversaries. (Bakshi and Prasad 2021; Diakonikolas, Kong, and Stewart 2019) designed some robust methods for linear regression. (Diakonikolas et al. 2019; Steinhardt, Koh, and Liang 2017) proposed a meta algorithm for robust learning with parametric models. There are also several other works that focus on general robust empirical risk minimization problems (Prasad et al. 2020; Jambulapati et al. 2021).

Despite these previous works toward robust learning problems, most of them focus on parametric models. However, for nonparametric methods such as kernel (Nadaraya 1964) and k nearest neighbor estimator, defense strategies against poisoning attack still need further exploration (Salibian-Barrera 2022). Actually, designing robust techniques is indeed more challenging for nonparametric methods than parametric one. For parametric models, the parameters are estimated using full dataset, while nonparametric methods have to rely on local training data around the query point. Even if the ratio of attacked samples among the whole dataset is small, the local anomaly ratio in the neighborhood of the query point can be large. As a result, the estimated function value at such query point can be totally wrong. Despite such difficulty, in many real scenarios, due to problem complexity or lack of prior knowledge, parametric models are not always available. Therefore, we hope to explore effective schemes to overcome the robustness issue of nonparametric regression.

In this paper, we provide a theoretical study about robust nonparametric regression problem under poisoning attack. In particular, we hope to investigate the theoretical limit of this problem, and design a method to achieve this limit. Towards this goal, we make the following contributions:

Firstly, we propose and analyze an estimator that minimizes a weighted Huber loss, which is quadratic with small input, and linear with large input. Such design achieves a

*Corresponding author

tradeoff between consistency and adversarial robustness. It was originally proposed in (Hall and Jones 1990), but to the best of our knowledge, it was not analyzed under adversarial setting. We show the convergence rate of both ℓ_2 and ℓ_∞ risk, under the assumption that the function to estimate is Lipschitz continuous, and the noise is sub-exponential. An interesting finding is that the maximum number of attacked samples (denoted as q) is not too large, then the convergence rate is not affected by adversarial samples, i.e. the influence of poisoning samples on the overall risk is only up to a constant factor.

Secondly, we provide an information theoretic minimax lower bound, which indicates the underlying limit one can achieve, with respect to q and N. The minimax lower bound without adversarial samples can be derived using standard information theoretic methods (Tsybakov 2009). Under adversarial attack, the estimation problem is harder, thus the lower bound in (Tsybakov 2009) may not be tight enough. We design some new techniques to derive a tighter one. The result shows that the initial estimator has optimal ℓ_∞ risk. With small q, the ℓ_2 risk is also minimax optimal. Nevertheless, for larger q, the ℓ_2 risk is not optimal, indicating that this estimator is still not perfect. We then provide an intuitive explanation of the suboptimality. Instead of attacking some randomly selected training samples, the best strategy for the attacker is to focus their attack within a small region. With this strategy, majority of training samples are altered here, resulting in wrong estimates. A simple remedy is to increase the kernel bandwidth to improve robustness, which can make ℓ_∞ risk optimal. However, this adjustment will introduce additional bias in other regions, thus the ℓ_2 risk is still suboptimal. The drawback of the initial estimator is that it does not make full use of the continuity of regression function, and thus unable to correct the estimation.

Finally, motivated by the issues of the initial method mentioned above, we propose a corrected estimator. If the attack focuses on a small region, then the initial estimate fails here. However, the estimate elsewhere is still reliable. With the assumption that the underlying function is continuous, the value at the severely corrupted region can be inferred using the surrounding values. With such intuition, we propose a nonlinear filtering method, which projects the estimated function to the space of Lipschitz functions with minimal ℓ_1 distance. The corrected estimate is then proved to be nearly minimax optimal up to only a $\ln N$ factor.

Preliminaries

In this section, we clarify notations and provide precise problem statements. Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathcal{X} \subset \mathbb{R}^d$ be N independently and identically distributed training samples, generated from a common probability density function (pdf) f. For each sample $\mathbf{X}_i$, we can receive a corresponding label Y_i:

$$Y_i = \begin{cases} \eta(\mathbf{X}_i) + W_i & \text{if } i \notin \mathcal{B} \\ \star & \text{otherwise,} \end{cases} \tag{1}$$

in which $\eta : \mathbb{R}^d \to \mathbb{R}$ is the unknown underlying function that we would like to estimate. W_i is the noise variable. For $i = 1, \ldots, N$, W_i are independent, with zero mean and finite variance. $\mathcal{B}$ is the set of indices of attacked samples. $\star$ means some value determined by the attacker. For each normal sample $\mathbf{X}_i$, the received label is $Y_i = \eta(\mathbf{X}_i) + W_i$. However, if a sample is attacked, then Y_i can be arbitrary value determined by the attacker. The attacker can manipulate up to q samples, thus $|\mathcal{B}| \leq q$.

Our goal is opposite to the attacker. We hope to find an estimate $\hat{\eta}$ that is as close to η as possible, while the attacker aims at reducing the estimation accuracy using a carefully designed attack strategy. We consider white-box setting here, in which the attacker has complete access to the ground truth η, $\mathbf{X}_i$ and W_i for all $i \in \{1, \ldots, N\}$, as well as our estimation algorithm. Under this setting, we hope to design a robust regression method that resists to any attack strategies.

The quality of estimation is evaluated using ℓ_2 and ℓ_∞ loss, which is defined as

$$R_2[\hat{\eta}] = \mathbb{E}\left[\sup_{\mathcal{A}}(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2\right], \tag{2}$$

$$R_\infty[\hat{\eta}] = \mathbb{E}\left[\sup_{\mathcal{A}}\sup_{\mathbf{x}}|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})|\right], \tag{3}$$

in which the expectation in (2) and (3) are taken over all training samples $(\mathbf{X}_i, Y_i), \ldots, (\mathbf{X}_N, Y_N)$. $\mathcal{A}$ denotes the attack strategy. The supremum over $\mathcal{A}$ is taken here because the adversary is assumed to be smart enough and the attack strategy is optimal. In (2), $\mathbf{X}$ denotes a random test sample that follows a distribution with pdf f. Our analysis can be easily generated to ℓ_p loss with arbitrary p.

Without any adversarial samples, η can be learned using kernel regression, also called the Nadaraya-Watson estimator (Nadaraya 1964; Watson 1964):

$$\hat{\eta}_{NW}(\mathbf{x}) = \frac{\sum_{i=1}^N K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) Y_i}{\sum_{i=1}^N K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}, \tag{4}$$

in which K is the Kernel function, h is the bandwidth that will decrease with the increase of sample size N. $\hat{\eta}_{NW}(\mathbf{x})$ can be viewed as a weighted average of the labels around $\mathbf{x}$. Without adversarial attack, such estimator converges to η (Devroye 1978). However, (4) fails even if a tiny fraction of samples are attacked. The attacked labels can just set to be sufficiently large. As a result, $\hat{\eta}_{NW}(\mathbf{x})$ could be far away from its truth.

The Initial Estimator

Now we build the estimator based on Huber loss minimization. Similar method was proposed in (Hall and Jones 1990). However, (Hall and Jones 1990) analyzed the case in which the distribution of label has heavy tails, instead of the case with corrupted samples. To the best of our knowledge, the performance under adversarial setting has not been analyzed. Now we use $\hat{\eta}_0$ to denote a slightly modified version of the estimator proposed in (Hall and Jones 1990):

$$\hat{\eta}_0(\mathbf{x}) = \arg\min_{|s| \leq M} \sum_{i=1}^N K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \phi(Y_i - s), \tag{5}$$

in which tie breaks arbitrarily if the minimum is not unique, and

$$\phi(u) = \begin{cases} u^2 & \text{if} \quad |u| \le T \\ 2T|u| - T^2 & \text{if} \quad |u| > T \end{cases} \tag{6}$$

is the Huber loss function.

Here we provide an intuitive understanding of this method. We hope the estimator to have two properties: robustness under attack, and consistency without attack. Robustness is guaranteed if we let ϕ be ℓ_1 loss, but the solution is the local median instead of mean. As long as the noise distribution is not symmetric, ℓ_1 minimizer is not consistent. On the contrary, letting ϕ be ℓ_2 loss just yields the kernel regression (4), which is not robust. Therefore, Huber cost (6) is designed to get a tradeoff between these two goals, which is quadratic with small input and linear with large input. The threshold parameter T can be set flexibly. Moreover, consider that there exists nonzero probability that $|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})|$ is arbitrarily large, we project the result into $[-M, M]$. [1]

There are several simple baselines for comparison. The first one is median-of-means (MoM) (Nemirovskij and Yudin 1983; Ben-Hamou and Guyader 2023), which divides samples into groups and calculates the median of the estimates in each group. MoM is inefficient because it fails even when there is only one attacked sample in each group. Another solution is trimmed mean (Bickel 1965; Welsh 1987; Dhar, Jha, and Rakshit 2022), which removes a fraction of samples with largest and smallest label values. The trim fraction parameter depends on the ratio of attacked samples. Unfortunately, such ratio is highly likely to be uneven over the support, while the trim fraction is set uniformly. This dilemma makes trimmed mean method not efficient. Robust regression with spline smoothing (Eubank 1999) is another alternative but is restricted to one dimensional problems. Finally, robust regression trees (Chaudhuri and Loh 2002) works practically but theoretical guarantee is not provided.

Finally, we comment on the computation of the estimator (5). Note that ϕ is convex, therefore the minimization problem in (5) can be solved by gradient descent. The derivative of ϕ is

$$\phi'(u) = \begin{cases} 2u & \text{if} \quad |u| \le T \\ 2T & \text{if} \quad u > T \\ -2T & \text{if} \quad u < -T. \end{cases} \tag{7}$$

Based on (5) and (7), s can be updated using binary search. Denote ϵ as the required precision, then the number of iterations for binary search should be $O(\ln(M/\epsilon))$. Therefore, the computational complexity is higher than kernel regression up to a $\ln(M/\epsilon)$ factor.

Theoretical Analysis

This section proposes the theoretical analysis of the initial estimator (5) under adversarial setting. To begin with, we make some assumptions about the problem.

[1] Suppose that for some $\mathbf{x}$, there is only one training sample whose distance to $\mathbf{x}$ is smaller than h. This sample is controlled by adversary and the value is altered arbitrarily far away. This event can result in arbitrarily large estimation error, and happens with nonzero probability. Therefore, if we do not project the estimate output to $[-M, M]$, then both ℓ_2 and ℓ_∞ loss will be infinite.

Assumption 1. *(Problem Assumption) there exists a compact set $\mathcal{X}$ and several constants L, γ, f_m, f_M, D, α, σ, such that the pdf f is supported at $\mathcal{X}$, and*

(a) (Lipschitz continuity) For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $|\eta(\mathbf{x}_1) - \eta(\mathbf{x}_2)| \le L||\mathbf{x}_1 - \mathbf{x}_2||$;

(b) (Bounded f and η) For all $\mathbf{x} \in \mathcal{X}$, $f_m \le f(\mathbf{x}) \le f_M$ and $|\eta(\mathbf{x})| \le M$, in which M is the parameter used in (5)*;*

(c) (Corner shape restriction) For all $r < D$ and $\mathbf{x} \in \mathcal{X}$, $V(B(\mathbf{x}, r) \cap \mathcal{X}) \ge \alpha v_d r^d$, in which $B(\mathbf{x}, r)$ is the ball centering at $\mathbf{x}$ with radius r, v_d is the volume of d dimensional unit ball, which depends on the norm we use;

(d) (Sub-exponential noise) The noise W_i is subexponential with parameter σ,

$$\mathbb{E}[e^{\lambda W_i}] \le e^{\frac{1}{2}\sigma^2\lambda^2}, \forall|\lambda| \le \frac{1}{\sigma}, \tag{8}$$

for $i = 1, \ldots, N$.

In Assumption 1, (a) is a common assumption for smoothness. (b) is also commonly made and usually called "strong density assumption" in existing literatures on nonparametric statistics (Audibert and Tsybakov 2007; Döring, Györfi, and Walk 2017). This assumption requires the pdf to be both upper and lower bounded in its support. Although somewhat restrictive, this assumption facilitates theoretical analysis. Relaxing this assumption is possible but is not the focus of this paper. We refer to full paper (Zhao and Wan 2023) for some further analysis. (c) prevents the shape of the corner of the support from being too sharp. Without assumption (c), the samples around the corner may not be enough, and the attacker can just attack the samples at the corner of the support, which can result in large errors. (d) requires that the noise is sub-exponential. If the noise assumption is weaker, e.g. only requiring the bounded moments of W_i up to some order, then the noise can be disperse. In this case, it will be harder to distinguish adversarial samples from clean samples.

We then make some restrictions on the kernel function K.

Assumption 2. *(Kernel Assumption) the kernel need to satisfy:*

(a) $\int K(\mathbf{u})du = 1$;

(b) $K(\mathbf{u}) = 0, \forall||\mathbf{u}|| > 1$;

(c) $c_K \le K(\mathbf{u}) \le C_K$ for two constants c_K and C_K.

In Assumption 2, (a) is actually not necessary, since from (5), the estimated value will not change if the kernel function is multiplied by a constant factor. This assumption is only for convenience of proof. (b) and (c) require that the kernel need to be somewhat close to the uniform function in the unit ball. Intuitively, if the attacker wants to modify the estimate at some $\mathbf{x}$, the best way is to change the response of sample i with large $K((\mathbf{X}_i - \mathbf{x})/h)$, in order to make strong impact on $\hat{\eta}(\mathbf{x})$. To defend against such attack, the upper bound of K should not be too large. Besides, to ensure that clean samples dominate corrupted samples everywhere, the effect of each clean sample on the estimation should not be too small, thus K also need to be bounded from below in its support.

Furthermore, recall that (5) has three parameters, i.e. h, T and M. We assume that these three parameters satisfy the following conditions.

Assumption 3. *(Parameter Assumption) h, T, M need to satisfy*

(a)$h > \ln^2 N/N$;
(b)$T \geq 4Lh + 16\sigma \ln N$;
(c)$M > \sup_{\mathbf{x}\in\mathcal{X}} |\eta(\mathbf{x})|$.

In Assumption 3, (a) ensures that the number of samples whose distance to $\mathbf{x}$ less than h is not too small. It is necessary for consistency, but is not enough for a good tradeoff between bias, variance and robustness. The optimal dependence of h over N will be discussed later. (b) requires that $T \sim \ln N$. This rule is based on the sub-exponential noise condition in Assumption 1(d). If we use sub-Gaussian assumption instead, then it is enough for $T \sim \sqrt{\ln N}$. If the noise is further assumed to be bounded, then T can just be set to constant. On the contrary, if the noise has heavier tail, then T needs to grow with N faster. (b) is mainly designed for rigorous theoretical analysis. Practically, one may choose smaller T. (c) prevents the estimate from being truncated too much.

The upper bound of ℓ_2 error is derived under these assumptions. Denote $a \lesssim b$ if $a \leq Cb$ for some constant C that depends only on $L, M, \gamma, f_m, f_M, D, \alpha, \sigma, c_K, C_K$.

Theorem 1. *Under Assumption 1, 2 and 3,*

$$\mathbb{E}\left[\sup_{\mathcal{A}} \left(\hat{\eta}_0(\mathbf{X}) - \eta(\mathbf{X})\right)^2\right] \lesssim \frac{T^2 q}{N} \min\left\{\frac{q}{Nh^d}, 1\right\} + h^2 + \frac{1}{Nh^d}. \quad (9)$$

The detailed proof of Theorem 1 is shown in the supplementary material. Here we provide an intuitive explanation of (9). The first term in (9) is caused by adversarial attack, while the remaining two terms are just the standard nonparametric regression error (Tsybakov 2009) for clean samples. Therefore we only discuss the first term here. The best strategy for the adversary is to concentrate its attack on a small region. Denote $B_h(\mathbf{x})$ as the ball centering at $\mathbf{x}$ with radius h, in which h is the bandwidth parameter in (5). Since the pdf f is both upper and lower bounded, the number of samples within $B_h(x)$ roughly scales as Nh^d. Now we discuss two cases. Firstly, if $q \lesssim Nh^d$, with q attacked samples around $\mathbf{x}$, the additional estimation error caused by these adversarial samples roughly scales as $Tq/(Nh^d)$. These attacked samples can affect $\hat{\eta}_0(\mathbf{x})$ for a region with radius roughly h, thus the overall additional ℓ_2 error is $(Tq/(Nh^d))^2 h^d = T^2q^2/(N^2h^d)$. Secondly, if $q \gtrsim Nh^d$, then the adversary can attack most of samples in a much broader region, whose volume scales as q/N. The additional estimation error in this region is proportional to T, thus the additional ℓ_2 error is T^2q/N. Combining these two cases yields (9), in which there is a phase transition between $q \lesssim Nh^d$ and $q \gtrsim Nh^d$.

Remark 1. *Theorem 1 is based on the assumption that the pdf f is bounded from below. For the case such that f has bounded support but can approach zero arbitrarily, we have provided an analysis in section 3 in the supplementary material. The result is that*

$$\mathbb{E}\left[\sup_{\mathcal{A}} \left(\hat{\eta}_0(\mathbf{X}) - \eta(\mathbf{X})\right)^2\right] \lesssim \frac{T^2 q}{N} + h^2 + \frac{1}{Nh^d}. \quad (10)$$

From (10)*, the bound is worse than the case with densities bounded from below if $q \lesssim Nh^d$. Intuitively, in this case, the best strategy for the adversary would be to attack samples in the region with low pdf values.*

The next theorem shows the bound of ℓ_∞ error:

Theorem 2. *Under Assumption 1, 2, 3, if $K(\mathbf{u})$ is monotonic decreasing with respect to $\|u\|$, then*

$$\mathbb{E}\left[\sup_{\mathcal{A}}\sup_{\mathbf{x}} |\hat{\eta}_0(\mathbf{x}) - \eta(\mathbf{x})|\right] \lesssim \frac{Tq}{Nh^d} + h + \frac{\ln N}{\sqrt{Nh^d}}. \quad (11)$$

The detailed proof is in section 4 in the supplementary material. Unlike ℓ_2 loss, the assumption that f is bounded from below can not be relaxed under ℓ_∞ loss. If f can approach zero, then the adversary can just attack the region with low density. As a result, we can only get a trivial bound $R_\infty \lesssim 1$.

We then show the minimax lower bound, which indicates the information theoretic limit of the adversarial nonparametric regression problem. In general, it is impossible to design an estimator with convergence rate faster than the following bound.

Theorem 3. *Let $\mathcal{F}$ be the collection of $f, \eta, \mathbb{P}_\mathbb{N}$ that satisfy Assumption 1, in which $\mathbb{P}_N$ is the distribution of the noise $W_1, \ldots, W_N$. Then*

$$\inf_{\hat{\eta}} \sup_{(f,\eta,\mathbb{P}_N)\in\mathcal{F}} \mathbb{E}\left[\sup_{\mathcal{A}} \left(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X})\right)^2\right] \gtrsim \left(\frac{q}{N}\right)^{\frac{d+2}{d+1}} + N^{-\frac{2}{d+2}}, \quad (12)$$

and

$$\inf_{\hat{\eta}} \sup_{(f,\eta,\mathbb{P}_N)\in\mathcal{F}} \mathbb{E}\left[\sup_{\mathcal{A}}\sup_{\mathbf{x}} |\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})|\right] \gtrsim \left(\frac{q}{N}\right)^{\frac{1}{d+1}} + N^{-\frac{1}{d+2}}. \quad (13)$$

In the right hand side of (12) and (13), $N^{-2/(d+2)}$ is the standard minimax lower bound for nonparametric estimation (Tsybakov 2009), which holds even if there are no adversarial samples. Therefore, we focus on the proof of the lower bound with the first term in the right hand side of (12). The basic idea is to construct two hypotheses on the regression function η. The total variation distance between these two hypotheses is not too large, thus the adversary can transform one of them to the other. As a result, after adversarial contamination with a carefully designed strategy, we are no longer able to distinguish between these hypotheses. The lower bounds in (12) and (13) can then be constructed accordingly.

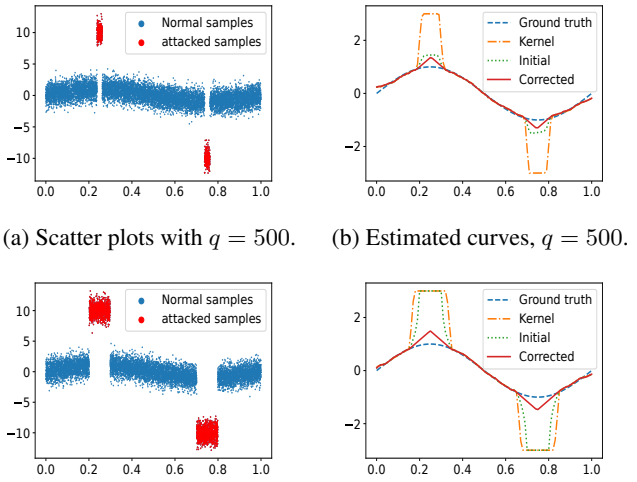
Compare Theorem 1, 2 and Theorem 3, we have the following findings. We claim that the upper and lower bound nearly match, if these two bounds match up to a polynomial of $\ln N$:

- The ℓ_∞ error is rate optimal. From (11) and (13), with $h \sim \max\{(q/N)^{1/(d+1)}, N^{-1/(d+2)}\}$ and $T \sim \ln N$, the upper and minimax lower bound of ℓ_∞ error nearly match.
- The ℓ_2 error is rate optimal if $q \lesssim \max\left\{\sqrt{N/\ln^2 N}, N^{d/(d+2)}/\ln^2 N\right\}$. From (9) and (12), let $h \sim N^{-\frac{1}{d+2}}$, the upper and minimax lower bound of ℓ_2 error match. In fact, in this case, the convergence rate of (5) is the same as ordinary kernel regression without adversarial samples, i.e. $h^2 + 1/(Nh^d)$. With optimal selection of h, ℓ_2 error scales as $N^{-2/(d+2)}$, which is just the standard rate for nonparametric statistics (Krzyzak 1986; Tsybakov 2009).
- The ℓ_2 error is not rate optimal if $q \gtrsim \max\left\{\sqrt{N/\ln^2 N}, N^{d/(d+2)}/\ln^2 N\right\}$. In this case, if $d \le 2$, the optimal h in (9) is $h \sim (q \ln N/N)^{2/(d+2)}$, and resulting ℓ_2 error is $R_2 \lesssim (q \ln N/N)^{4/(d+2)}$. If $d > 2$, then optimal h is $h \sim N^{-1/(d+2)}$, and the ℓ_2 error is $q \ln^2 N/N + N^{-2/(d+2)}$. For either $d \le 2$ or $d > 2$, these two bounds are worse than the lower bound in (12).

This result indicates that the initial estimator (5) is optimal under ℓ_∞, or under ℓ_2 with small q. However, under large number of adversarial samples, the ℓ_2 error becomes suboptimal.

Now we provide an intuitive understanding of the suboptimality of ℓ_2 risk with large q using a simple one dimensional example shown in Figure 1, in which $N = 10000$, $h = 0.05$, $M = 3$, $f(x) = 1$ for $x \in (0,1)$, $\eta(x) = \sin(2\pi x)$, and the noise follows standard normal distribution $\mathcal{N}(0,1)$. For each x, denote $q_h(x)$, $n_h(x)$ as the number of attacked samples and total samples within $(x-h, x+h)$, respectively. For robust mean estimation problems, the breakdown point is $1/2$ (Andrews and Hampel 2015), which also holds locally for nonparametric regression problem. Hence, if $q_h(x)/n_h(x) > 1/2$, the estimator will collapse and return erroneous values even if we use Huber cost. In Fig 1(a), $q = 500$, among which 250 attacked samples are around $x = 0.25$, while others are around $x = 0.75$. In this case, $q_h(x)/n_h(x) < 1/2$ over the whole support. The curve of estimated function is shown in Fig 1(b). The estimate with (5) is significantly better than kernel regression. Then we increase q to 2000. In this case, $q_h(x)/n_h(x) > 1/2$ around 0.25 and 0.75 (Fig 1(c)), thus the estimate fails. The estimated function curve shows an undesirable spike (Fig 1(d)).

The above example shows that the initial estimator (5) fails if the adversary focus its attack at a small region. In this case, the local ratio of attacked samples surpasses the breakdown point, resulting in spikes here. With such strategy and sufficiently large q, the initial estimator (5) fails to be optimal. Actually, getting a robust estimate of $\eta(\mathbf{x})$ using local training samples around $\mathbf{x}$ only is not enough. To improve the estimator, we exploit the continuity property of η (Assumption 1(a)), and use the estimate in neighboring regions to correct $\hat{\eta}_0(\mathbf{x})$. Based on such intuition, we propose a projection technique, which will close the gap between the upper and minimax lower bound. The details are shown in the next section.



(a) Scatter plots with $q = 500$. (b) Estimated curves, $q = 500$.

(c) Scatter plots with $q = 2000$. (d) Estimated curves, $q = 2000$.

Figure 1: A simple example with $q = 500$ and $q = 2000$. In (a) and (c), red dots are attacked samples, while blue dots are normal samples. In (b) and (d), four curves correspond to ground truth η, the result of kernel regression, initial estimate and corrected estimate, respectively. With $q = 500$, the initial estimate (5) works well. However, with $q = 2000$, the initial estimate fails, while the corrected regression works well.

Corrected Regression

As has been discussed in the previous section, while the initial estimator is already efficient in its own right with small q, it does not tolerate larger q. In particular, concentrated attack will generate undesirable spikes in $\hat{\eta}_0$. We hope to remove these spikes without introducing two much additional estimation error. Linear filters[2] do not work here since the profile of the regression estimate will be blurred. Therefore, we propose a nonlinear filter as following. It conducts minimum correction (in ℓ_1 sense) to the initial result $\hat{\eta}_0$, while ensuring that the corrected estimate is Lipschitz. Formally, given the initial estimate $\hat{\eta}_0(\mathbf{x})$, our method solves the following optimization problem

$$\hat{\eta}_c = \mathop{\arg\min}_{\|\nabla g\|_\infty \le L} \|\hat{\eta}_0 - g\|_1, \tag{14}$$

in which

$$\|\nabla g\|_\infty = \max\left\{\left|\frac{\partial g}{\partial x_1}\right|, \ldots, \left|\frac{\partial g}{\partial x_d}\right|\right\}. \tag{15}$$

[2]Here linear filter means that the output is linear in the input, i.e. an operator F is linear if for any function f_1, f_2 and any scalars λ_1 and λ_2, $F[\lambda_1 f_1 + \lambda_2 f_2] = \lambda_1 F[f_1] + \lambda_2 F[f_2]$. Alternatively, $F[f]$ is a convolution of f with another function K_F. Such convolution can blur the regression estimate.

In Appendix F in the full paper (Zhao and Wan 2023), we prove that the solution to the optimization problem (14) is unique.

From (14), $\hat{\eta}_c$ can be viewed as the projection of the output of initial estimator $\hat{\eta}_0$ into the space of Lipschitz functions. Here we would like to explain intuitively why we use ℓ_1 distance instead of other metrics in (14). Using the example in Fig.1(d) again, it can be observed that at the position of such spikes, $|\eta(\mathbf{x}) - g(\mathbf{x})|$ can be large. In order to ensure successful removal of spikes, we hope that the derivative of such cost should not be too large, otherwise the corrected estimate will tend to be closer to the original one to minimize the cost, thus spikes may not be fully removed. Based on such intuition, ℓ_1 cost is preferred here, since it has bounded derivatives, while other costs such as ℓ_2 distance have growing derivatives.

The estimation error of the corrected regression can be bounded by the following theorem.

Theorem 4. *(1) Under the same conditions as Theorem 1,*

$$\mathbb{E}\left[\sup_{\mathcal{A}}\left(\hat{\eta}_c(\mathbf{X}) - \eta(\mathbf{X})\right)^2\right] \lesssim \left(\frac{q \ln N}{N}\right)^{\frac{d+2}{d+1}} + h^2 + \frac{\ln N}{Nh^d}. \tag{16}$$

(2) Under the same conditions as Theorem 2,

$$\mathbb{E}\left[\sup_{\mathcal{A}}\sup_{\mathbf{x}}|\hat{\eta}_c(\mathbf{x}) - \eta(\mathbf{x})|\right] \lesssim \frac{Tq}{Nh^d} + h + \frac{\ln N}{\sqrt{Nh^d}}. \tag{17}$$

The proof is shown in Appendix G in the full paper (Zhao and Wan 2023). Here we provide a brief idea of the proof. For the error of $\hat{\eta}_0$ in (9), the first term is caused by adversarial samples, while the second and third term are just usual regression error. The latter one nearly remains the same after filtering, while the impact of the former error is significantly reduced. In particular, the ℓ_1 additional estimation error can be bounded first. This bound can then be used to infer ℓ_2 and ℓ_∞ error caused by adversarial samples, using the property that $\hat{\eta}_c$ is Lipschitz. From (16), compared with Theorem 3, with $T \sim \ln N$ and a proper h, the upper and lower bound nearly match.

Now we discuss the practical implementation. (14) can not be calculated directly for a continuous function. Therefore, we find an approximate numerical solution instead. The detail of practical implementation is shown in Appendix A in the full paper (Zhao and Wan 2023).

Despite the optimal sample complexity, the computation of the corrected estimator is expensive for high dimensional distributions. It would be an interesting future direction to improve the computational complexity on dimensionality. Currently, our method is designed mainly for low dimensional problems.

Numerical Examples

In this section we show some numerical experiments. In particular, we show the curve of the growth of mean square error over the attacked sample size q. More numerical results are shown in the full paper (Zhao and Wan 2023).

For each case, we generate $N = 10000$ training samples, with each sample follows uniform distribution in $[0, 1]^d$. The kernel function is

$$K(u) = 2 - |u|, \forall |u| \le 1. \tag{18}$$

We compare the performance of kernel regression, the median-of-means method, trimmed mean, initial estimate, and the corrected estimation under multiple attack strategies. For kernel regression, the output is $\max(\min(\hat{\eta}_{NW}, M), -M)$, in which $\hat{\eta}_{NW}$ is the simple kernel regression defined in (4). We truncate the result into $[-M, M]$ for a fair comparison with robust estimators. For the median-of-means method, we divide the training samples into $b = 20$ groups randomly, and then conduct kernel regression for each group and then find the median, i.e.

$$\hat{\eta}_{MoM} = \mathrm{Clip}(\mathrm{med}(\{\hat{\eta}_{NW}^{(1)}, \ldots, \hat{\eta}_{NW}^{(b)}\}), [-M, M]), \tag{19}$$

in which $\mathrm{Clip}(u, [-M, M]) = \max(\min(x, M), -M)$ projects the value onto $[-M, M]$, and med denotes the median. For trimmed mean regression, the trim fraction is 0.2.

For the initial estimator (5), the parameters are $T = 1$ and $M = 3$. The corrected estimator is described in the full paper (Zhao and Wan 2023). For $d = 1$, the grid count is $m = 50$. For $d = 2$, $m_1 = m_2 = 20$. Consider that the optimal bandwidth (h in (5)) need to increase with the dimension, in (4), the bandwidths of all these four methods are set to be $h = 0.03$ for one dimensional distribution, and $h = 0.1$ for two dimensional case. Here M and h satisfy Assumption 3(a) and (c), while T is smaller than the requirement in Assumption 3(b). As was already discussed earlier, the parameter selection rule in Assumption 3 is designed mainly for theoretical analysis, and does not need to be exactly satisfied in practice.

The attack strategies are designed as following. Let $q = 500k$ for $k = 0, 1, \ldots, 10$.

Definition 1. *There are three strategies, namely, random attack, one directional attack, and concentrated attack, which are defined as following:*

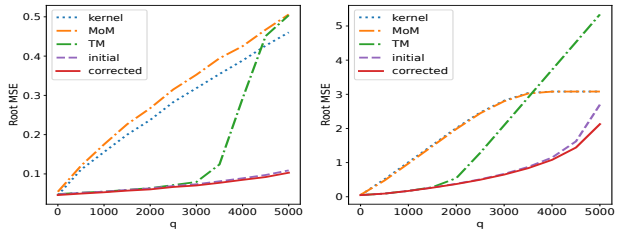
(1) Random Attack. The attacker randomly select q samples among the training data to attack. The value of each attacked sample is -10 or 10 with equal probability;

(2) One directional Attack. The attacker randomly select q samples among the training data to attack. The value of all attacked samples are 10;

(3) Concentrated Attack. The attacker pick two random locations $\mathbf{c}_1$, $\mathbf{c}_2$ that are uniformly distributed in $[0, 1]^d$. For $\lfloor q/2 \rfloor$ samples that are closest to $\mathbf{c}_1$, modify their values to 10. For $\lfloor q/2 \rfloor$ samples that are closest to $\mathbf{c}_2$, modify their values to -10.

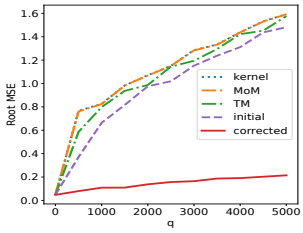
For one dimensional distribution, let the ground truth be $\eta_1(x) = \sin(2\pi x)$. For two dimensional distribution, let $\eta(\mathbf{x}) = \sin(2\pi x_1) + \cos(2\pi x_2)$.

The noise follows standard Gaussian distribution $\mathcal{N}(0, 1)$. The performances are evaluated using square root of ℓ_2 error, as well as ℓ_∞ error. The results are shown in Figure 2 and 3 for one and two dimensional distributions, respectively. In these figures, each point is the average over 1000 independent trials.

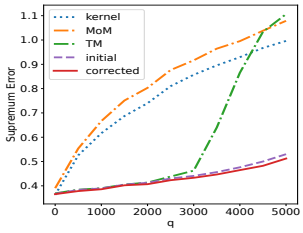


(a) Squared root of ℓ_2 error, under random attack.

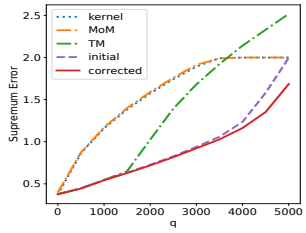
(b) Squared root of ℓ_2 error, under one directional attack.



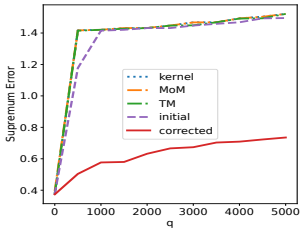
(c) Squared root of ℓ_2 error, under concentrated attack.



(d) ℓ_∞ error, under random attack.

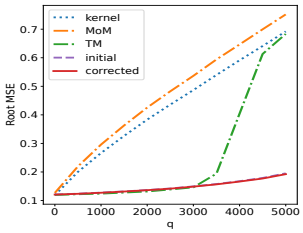


(e) ℓ_∞ error, under one directional attack.

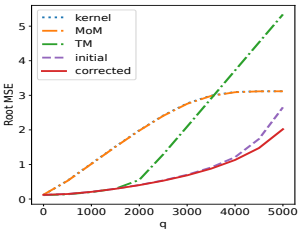


(f) ℓ_∞ error, under concentrated attack.

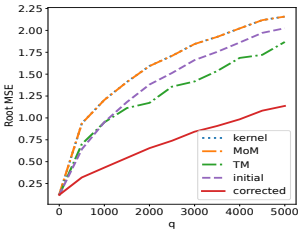
Figure 2: Comparison of ℓ_2 and ℓ_∞ error between various methods for $d = 1$.



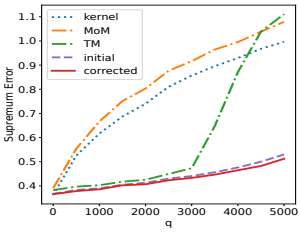
(a) Squared root of ℓ_2 error, under random attack.



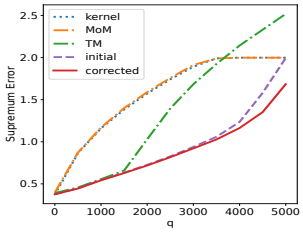
(b) Squared root of ℓ_2 error, under one directional attack.



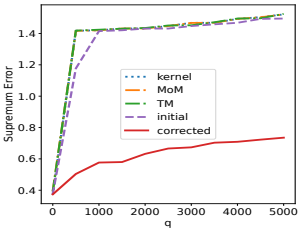
(c) Squared root of ℓ_2 error, under concentrated attack.



(d) ℓ_∞ error, under random attack.



(e) ℓ_∞ error, under one directional attack.



(f) ℓ_∞ error, under concentrated attack.

Figure 3: Comparison of ℓ_2 and ℓ_∞ error between various methods for $d = 2$.

Figure 2 and 3 show that the simple kernel regression (blue dotted line) fails under poisoning attack. The ℓ_2 and ℓ_∞ error grows fast with the increase of q. Besides, traditional median-of-means (orange dash-dot line) does not improve over kernel regression. Trimmed mean estimator works well under random or one directional attack with small q, but fails otherwise. Moreover, the initial estimator (5) (purple dashed line) shows significantly better performance than kernel estimator under random and one directional attack, as are shown in Fig.2 and 3, (a), (b), (d), (e). However, if the attacked samples concentrate around some centers, then the initial estimator fails. Compared with kernel regression, there is some but limited improvement for (5). Finally, the corrected estimator (red solid line) performs well under all attack strategies. Under random attack, the corrected estimator performs nearly the same as initial one. For one directional attack, the corrected estimator performs better than the initial one with large q. Under concentrated attack, the correction shows significant improvement. These results are consistent with our theoretical analysis.

We have also conducted numerical experiments using real data. In particular, we obtain and compare the root MSE score of the median-of-means, trimmed mean, our initial estimator and the corrected estimator under all three types of attacks. All experiments show that our methods have desirable performance. The initial estimator significantly improves over median-of-means and trimmed mean estimator. The performance is further improved using our correction technique. The detailed implementation and results are shown in Appendix I in the full paper (Zhao and Wan 2023).

Conclusion

In this paper, we have provided a theoretical analysis of robust nonparametric regression problem under adversarial attack. In particular, we have derived the convergence rate of an M-estimator based on Huber loss minimization. We have also derived the information theoretic minimax lower bound, which is the underlying limit of robust nonparametric regression. The result shows that the initial estimator has minimax optimal ℓ_∞ risk. With small q, which is the number of adversarial samples, ℓ_2 risk is also optimal. However, for large q, the initial estimator becomes suboptimal. Finally, we have proposed a correction technique, which is a nonlinear filter that projects the estimated function into the space of Lipschitz functions. Our theoretical analysis shows that the corrected estimator is minimax optimal even for large q. Numerical experiments on both synthesized and real data validate our theoretical analysis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62132008, 62272425 and 61972229), the Key Research Project of Zhejiang Lab (No.2022PD0AC02, 2022PD0AC01 and K2022PD1BB01), the Natural Science Foundation of Jiangsu Province (BK20220075) and the Fok Ying-Tong Education Foundation for Young Teachers in the Higher Education Institutions of China (No.20193218210004).

References

Andrews, D. F.; and Hampel, F. R. 2015. *Robust estimates of location: Survey and advances*, volume 1280. Princeton University Press.

Audibert, J.-Y.; and Tsybakov, A. B. 2007. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35(2): 608–633.

Bakshi, A.; and Prasad, A. 2021. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 102–115.

Ben-Hamou, A.; and Guyader, A. 2023. Robust nonparametric regression via median-of-means. *arXiv preprint arXiv:2301.10498*.

Bickel, P. J. 1965. On some robust estimates of location. *The Annals of Mathematical Statistics*, 847–858.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*.

Chaudhuri, P.; and Loh, W.-Y. 2002. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 561–576.

Cheng, Y.; Diakonikolas, I.; Ge, R.; and Woodruff, D. P. 2019. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, 727–757. PMLR.

Devroye, L. P. 1978. The uniform convergence of the nadaraya-watson regression function estimate. *Canadian Journal of Statistics*, 6(2): 179–191.

Dhar, S.; Jha, P.; and Rakshit, P. 2022. The trimmed mean in non-parametric regression function estimation. *Theory of Probability and Mathematical Statistics*, 107: 133–158.

Diakonikolas, I.; Kamath, G.; Kane, D.; Li, J.; Steinhardt, J.; and Stewart, A. 2019. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, 1596–1606. PMLR.

Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; and Stewart, A. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *57th Annual Symposium on Foundations of Computer Science*, 655–664.

Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; and Stewart, A. 2017. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, 999–1008. PMLR.

Diakonikolas, I.; and Kane, D. M. 2019. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Diakonikolas, I.; and Kane, D. M. 2023. *Algorithmic high-dimensional robust statistics*. Cambridge University Press.

Diakonikolas, I.; Kong, W.; and Stewart, A. 2019. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2745–2754. SIAM.

Döring, M.; Györfi, L.; and Walk, H. 2017. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(1): 8485–8500.

Eubank, R. L. 1999. *Nonparametric regression and spline smoothing*. CRC press.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Hall, P.; and Jones, M. 1990. Adaptive M-estimation in nonparametric regression. *Annals of Statistics*, 1712–1728.

Hopkins, S. B.; and Li, J. 2018. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 1021–1034.

Huber, P. J. 1981. *Robust Statistics*. John Wiley & Sons.

Jagielski, M.; Oprea, A.; Biggio, B.; Liu, C.; Nita-Rotaru, C.; and Li, B. 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy (SP)*, 19–35. IEEE.

Jambulapati, A.; Li, J.; Schramm, T.; and Tian, K. 2021. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34: 4475–4488.

Krzyzak, A. 1986. The rates of convergence of kernel regression estimates and classification rules. *IEEE Transactions on Information Theory*, 32(5): 668–679.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mao, C.; Zhong, Z.; Yang, J.; Vondrick, C.; and Ray, B. 2019. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32.

Maronna, R. A.; Martin, R. D.; Yohai, V. J.; and Salibián-Barrera, M. 2019. *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26.

Nemirovskij, A. S.; and Yudin, D. B. 1983. Problem complexity and method efficiency in optimization. *Wiley-Interscience Series in Discrete Mathematics*.

Prasad, A.; Suggala, A. S.; Balakrishnan, S.; and Ravikumar, P. 2020. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3): 601–627.

Salibian-Barrera, M. 2022. Robust nonparametric regression: review and practical considerations. *arXiv preprint arXiv:2211.08376*.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Steinhardt, J. 2018. *Robust learning: Information theory and algorithms*. Stanford University.

Steinhardt, J.; Koh, P. W. W.; and Liang, P. S. 2017. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tsybakov, A. B. 2009. *Introduction to Nonparametric Estimation*. Springer Series in Statistics.

Van Rooyen, B.; and Williamson, R. C. 2017. A Theory of Learning with Corrupted Labels. *J. Mach. Learn. Res.*, 18(1): 8501–8550.

Watson, G. S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Welsh, A. 1987. The trimmed mean in the linear model. *The Annals of Statistics*, 15(1): 20–36.

Xiao, H.; Biggio, B.; Brown, G.; Fumera, G.; Eckert, C.; and Roli, F. 2015. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, 1689–1698. PMLR.

Zhao, P.; and Wan, Z. 2023. Robust Nonparametric Regression under Poisoning Attack. *arXiv preprint arXiv:2305.16771*.