Adversarial Robustness of Nonparametric Regression

Parsa Moradi[*]
University of Minnesota
moradi@umn.edu

Hanzaleh Akbarinodehi[*]
University of Minnesota
akbar066@umn.edu

Mohammad Ali Maddah-Ali
University of Minnesota
maddah@umn.edu

Abstract

In this paper, we investigate the adversarial robustness of nonparametric regression, a fundamental problem in machine learning, under the setting where an adversary can arbitrarily corrupt a subset of the input data. While the robustness of parametric regression has been extensively studied, its nonparametric counterpart remains largely unexplored. We characterize the adversarial robustness in nonparametric regression, assuming the regression function belongs to the second-order Sobolev space (i.e., it is square integrable up to its second derivative).

The contribution of this paper is two-fold: (i) we establish a minimax lower bound on the estimation error, revealing a fundamental limit that no estimator can overcome, and (ii) we show that, perhaps surprisingly, the classical smoothing spline estimator, when properly regularized, exhibits robustness against adversarial corruption. These results imply that if $o(n)$ out of n samples are corrupted, the estimation error of the smoothing spline vanishes as $n \to \infty$. On the other hand, when a constant fraction of the data is corrupted, no estimator can guarantee vanishing estimation error, implying the optimality of the smoothing spline in terms of maximum tolerable number of corrupted samples.

1 Introduction

In recent years, machine learning (ML) models have increasingly relied on data from diverse sources and are often deployed in distributed or decentralized computing environments [1–7]. These settings introduce new attack surfaces and create incentives for adversaries to corrupt data or disrupt learning algorithms [8–11]. This has motivated a growing body of research initiatives aimed at understanding and mitigating the impact of adversarial behavior [12–21].

One of the fundamental problems in ML is regression, which aims to estimate an unknown function f based on observed noisy data [22]. This task is generally categorized into two approaches: *parametric* regression, which assumes f is a parametric function with known structure, and *nonparametric* regression, which makes minimal assumptions on f, allowing it to belong to a wide class of functions such as Sobolev or Hölder spaces [23, 24]. Regression underpins many machine learning tasks, and understanding its robustness to adversarial corruption is a critical objective.

Adversarial robustness in parametric regression has been extensively studied [25–31]. Many approaches leverage tools from classical robust statistics [32, 33], adapting techniques like trimmed means, median-of-means, and M-estimators to modern high-dimensional settings [34–36]. These methods benefit from the structural constraints of parametric models, which narrow the hypothesis space and simplify the alleviation of adversarial attacks. In contrast, robustness in nonparametric regression is considerably more challenging due to the absence of such structure, which makes the models more vulnerable to adversarial attacks [37–40].

In this work, we address the problem of nonparametric regression under adversarial corruption. We consider a setting in which one observes n pairs $\{(x_i, \widetilde{y}_i)\}_{i=1}^n$, where the responses $\widetilde{y}_i$ may be

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

partially corrupted by an adversary. Specifically, the adversary arbitrarily choose $\widetilde{y}_i$, for all $i \in \mathcal{A}$, where $\mathcal{A}$ is an unknown subset of $\{1, \ldots, n\}$ with cardinality at most $q < n$. For each $i \notin \mathcal{A}$, the observed response is $\widetilde{y}_i = f(x_i) + \varepsilon_i$, where $f \colon \Omega \to \mathbb{R}$ is the unknown regression function with domain $\Omega \subset \mathbb{R}$, and $\{\varepsilon_i\}_{i \in [n] \setminus \mathcal{A}}$ are i.i.d. noise variables with zero mean and variance at most σ^2.

In this paper, we assume that regression function f belongs to the second-order Sobolev space, consisting of functions that are square-integrable up to the second derivative over Ω. The objective of non-parametric regression is to produce $\hat{f}$ as an estimation of f based on $\{(x_i, \widetilde{y}_i)\}_{i=1}^n$. To evaluate the performance of $\hat{f}$ in the presence of adversarial corruption, we use the following metrics [24]:

$$R_2(f, \hat{f}) := \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\mathcal{S}} \|f - \hat{f}\|_{L_2(\Omega)}^2 \right], \qquad R_\infty(f, \hat{f}) := \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\mathcal{S}} \|f - \hat{f}\|_{L_\infty(\Omega)}^2 \right],$$

where $\boldsymbol{\varepsilon} := [\epsilon_1, \ldots, \epsilon_n]$ and $\mathcal{S}$ denotes the adversarial strategy that can corrupt up to q samples. Our goal is to characterize $\inf_{\hat{f}} R_2(f, \hat{f})$ and $\inf_{\hat{f}} R_\infty(f, \hat{f})$ over all estimators, assuming f belongs to the second-order Sobolev space, under the setting where the adversary may corrupt up to q samples.

The contributions of this paper are two-fold:

- **A Computationally-Efficient Estimator (Theorem 1):** We prove that the classical smoothing spline estimator retains robustness against adversarial corruption. This estimator selects $\hat{f}$ from the second-order Sobolev space, by minimizing the empirical error $\frac{1}{n}\sum_{i=1}^n (g(x_i) - \widetilde{y}_i)^2$, regularized by $\lambda \int \hat{f}''(x)^2 \, dx$, where $\lambda > 0$ controls the level of smoothness [41]. Smoothing splines are computationally efficient, with $\mathcal{O}(n)$ complexity of fitting and evaluating, leveraging B-spline basis functions [42, 43], and have found wide applications in statistics and machine learning [44–47]. Note that classical nonparametric methods are not necessarily adversarially robust. For instance, the Nadaraya–Watson (NW) estimator [48] can be fragile even under a small number of adversarial corruptions [39]. It is therefore surprising that a computationally efficient nonparametric regression method, such as smoothing splines, also exhibits adversarial robustness.

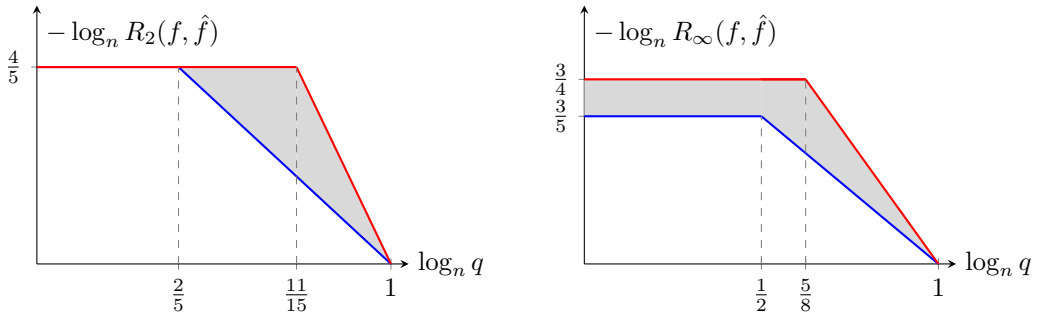
  While smoothing splines have been extensively studied in non-adversarial settings [41, 49–55], their robustness properties against adversarial corruption were not previously understood. In this paper, we show that if the adversary corrupts at most $q = o(n)$ samples and the regression function f belongs to a second-order Sobolev space, then the smoothing spline estimator achieves $R_2 \to 0$ and $R_\infty \to 0$ as $n \to \infty$. This result further provides an upper bound on $\inf_{\hat{f}} R_2(f, \hat{f})$ and $\inf_{\hat{f}} R_\infty(f, \hat{f})$ as functions of n and q.

- **Minimax Lower-Bound (Theorem 2):** We derive minimax lower bounds on $\inf_{\hat{f}} R_2(f, \hat{f})$ and $\inf_{\hat{f}} R_\infty(f, \hat{f})$, expressed as functions of n and q. These bounds characterize the fundamental limits of estimation accuracy: no estimator can achieve better rates over the second-order Sobolev space under adversarial corruption.

  A key implication of this result is that when $q = \Theta(n)$, no estimator can achieve vanishing error as $n \to \infty$. This highlights that smoothing splines are not only computationally efficient but also optimal in terms of the maximum number of tolerable adversarial corruptions (see Corollary 4).

To better understand the results of this paper, we examine their implications in the regime of large n. In this regime, our results can be concisely summarized in Figure 1. This figure illustrates the rate of convergence, defined as $-\log_n R_2(f, \hat{f})$ or respectively, $-\log_n R_\infty(f, \hat{f})$, as a function of q, or equivalently, $\frac{\log q}{\log n} = \log_n q$, as $n \to \infty$. The red curves represent the impossibility result, indicating that no estimator can achieve a convergence rate beyond this bound for all functions in second-order Sobolev space (Theorem 2). The blue curve shows the convergence rate of the smoothing spline estimator in the presence of adversarial samples, as established in Theorem 1.

It is worth noting that, as shown in Figure 1a, when $\log_n q < \frac{2}{5}$, the smoothing spline achieves the minimax-optimal convergence rate for metric R_2 . For $\frac{2}{5} \le \log_n(q) < 1$, in metric R_2 (Figure 1a), and for $0 \le \log_n(q) < 1$ in metric R_∞ (Figure 1b), while the smoothing splines offers vanishing estimation error for large n, the rate of convergence may not be optimum (there is a gap between the rate of convergence in smoothing splines and the minimax outer-bound). This theoretical results are supported with the simulation experiments results (see Section 4).



(a) Rate region for $R_2(f, \hat{f})$.

(b) Rate region for $R_\infty(f, \hat{f})$.

Figure 1: Rates of convergence for estimation error $R_2(f, \hat{f})$ and $R_\infty(f, \hat{f})$, as $n \to \infty$, and for any f belongs second-order Sobolev space (for non-asymptotic analysis, see Theorems 1 and 2). The blue curves represent the minimum rate achieved by the smoothing spline estimator. The red curves denote minimax outer bounds that are impossible to beat. Specifically, for $q = o(n)$, for the smoothing spline estimator, both R_2 and R_∞ converge to zero as $n \to \infty$. When $q = \Theta(n)$, we show that no estimator can achieve vanishing error, establishing a fundamental limit on robustness. This result highlights that smoothing splines are optimal in terms of the maximum tolerable number of adversarial corruptions (see Corollary 4).

This paper is organized as follows. Section 2 presents the problem formulation. Section 3 presents our main results by providing an upper and lower bounds under adversarial corruption. Section 4 provides simulation results, and Section 5 reviews related works.

Notation. Throughout the paper, we use $[n] := \{1, 2, \ldots, n\}$ and denote the cardinality of a set $\mathcal{A}$ by $|\mathcal{A}|$. Derivatives of scalar functions are written as f', f'', and, more generally, $f^{(k)}$ for the k-th derivative. The quantities $\|g\|_{L^2(\Omega)}$ and $\|g\|_{L^\infty(\Omega)}$ denote the L_2-norm and the supremum norm of a function $g(\cdot)$ over Ω. The space $\mathcal{W}^2(\Omega)$ refers to the second-order Sobolev space, consisting of square-integrable functions on Ω whose first and second derivatives are also square-integrable on Ω. We write $a \lesssim b$ to indicate that there exists a constant $C > 0$ such that $a \leq Cb$, and similarly $a \gtrsim b$ to mean $a \geq Cb$ for some constant $C > 0$.

2 Problem Formulation

Let $f : [a, b] \to \mathbb{R}$ be in $\mathcal{W}^2([a, b])$. The objective is to estimate f, from observations at fixed design points $x_i \in (a, b)$ for $i \in [n]$. Instead of observing a noisy version of responses (as in standard regression problem [22]), we are given (possibly) adversarially corrupted outputs $\{\widetilde{y}_i\}_{i=1}^n$, defined as:

$$\widetilde{y}_i = \begin{cases} f(x_i) + \varepsilon_i, & \text{if } i \notin \mathcal{A}, \\ *, & \text{if } i \in \mathcal{A}, \end{cases}$$

where $\{\varepsilon_i\}_{i \in [n] \setminus \mathcal{A}}$ are i.i.d. noise variables with zero mean and variance at most σ^2, and $\mathcal{A} \subseteq [n]$ is an unknown subset of indices corresponding to adversarially corrupted observations. Here, $*$ denotes an arbitrary value chosen strategically by the adversary to mislead the estimator. We assume $|\mathcal{A}| \leq q$, for some known $q \in \mathbb{N}$.

Let $\Omega := [a, b]$ denote the domain of the design points, and $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i \in [n] \setminus \mathcal{A}}$ denote the noise vector. Following [24], we evaluate the performance of any estimator $\hat{f}$ using two metrics, $R_2(f, \hat{f})$ and $R_\infty(\hat{f})$, where

$$R_2(f, \hat{f}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\mathcal{S}} \left\| f - \hat{f} \right\|_{L_2(\Omega)}^2 \right], \tag{1}$$

$$R_\infty(f, \hat{f}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\mathcal{S}} \left\| f - \hat{f} \right\|_{L_\infty(\Omega)}^2 \right], \tag{2}$$

where $\mathcal{S}$ denotes the strategy, chosen by the adversary, in choosing the subset $\mathcal{A}$ and the value of $\tilde{y}_i$, for $i \in \mathcal{A}$, as long as $|\mathcal{A}| \leq q$. The supremum over $\mathcal{S}$ considers the worst-case adversarial attack, aiming to maximize estimation error for $\hat{f}$.

In this paper, the objective is to find $\hat{f}$, that minimizes $R_2(f, \hat{f})$ or $R_\infty(f, \hat{f})$, over all possible estimator functions $\hat{f}$, where f is an arbitrary function in $\mathcal{W}^2(\Omega)$.

3 Main Results

In this section, we present our main results on the adversarial robustness of nonparametric regression. Without loss of generality, we assume that $\Omega = [0, 1]$ [2]. Let $\{x_i\}_{i=1}^n \subset \Omega$ denote the set of design points, and $f \in \mathcal{W}^2(\Omega)$ be the regression function.

First, we evaluate the robustness of the classical cubic smoothing spline estimator under adversarial corruption. In Theorem 1, we show that this estimator, as a computationally efficient [42, 43] and widely popular estimator [44–47], exhibits robustness to adversarial corruption.

The cubic smoothing spline estimator is defined as the solution to the following optimization problem:

$$\hat{f}_{\mathrm{SS}}^a = \underset{g \in \mathcal{W}^2(\Omega)}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n (g(x_i) - \widetilde{y}_i)^2 + \lambda \int_\Omega \left(g''(x)\right)^2 dx \right\}. \tag{3}$$

Here, $\lambda > 0$ is a smoothing parameter that balances the fitness to the sample data, measured by $\frac{1}{n}\sum_{i=1}^n (g(x_i) - \widetilde{y}_i)^2$, and smoothness of the estimator, quantified by $\int_\Omega g''(x)^2 dx$.

We define the empirical distribution function F_n associated with the design points $\{x_i\}_{i=1}^n$ as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \leq x\}, \tag{4}$$

where $\mathbf{1}\{x_i \leq x\}$ is the indicator function. We assume that F_n converges uniformly to a continuously differentiable cumulative distribution function (CDF) F; that is,

$$\sup_{x \in \Omega} |F_n(x) - F(x)| \to 0 \quad \text{as} \quad n \to \infty.$$

This is a standard assumption in the related literature [24]. For the limiting CDF, i.e., $F(x)$, we assume that the density function $p(x) := F'(x)$ exists. In addition, similar to [39, 49], we assume that $p(x)$ is bounded away from zero, i.e., there exists a constant $p_{\min} > 0$ such that $\inf_{x \in \Omega} p(x) \geq p_{\min}$, and that $p(x)$ is three times continuously differentiable on Ω.

We assume that the function f is bounded; that is, for all $x \in \Omega$, $|f(x)| \leq m_1$ for some constant $m_1 \in \mathbb{R}$. Moreover, we assume that the adversary's corrupted values are also bounded, i.e., the adversary cannot inject arbitrarily large perturbations, satisfying $|\widetilde{y}_i| \leq m_2$ for $m_2 \in \mathbb{R}$ and $i \in \mathcal{A}$.

Finally, let $\Delta_{\max} := \sup_{x \in \Omega} \min_{i \in [n]} |x - x_i|$, $\Delta_{\min} := \min_{i \neq j} |x_i - x_j|$, denote the maximum gap from any point in Ω to the nearest design point, and the minimum separation between any two design points, respectively. Likewise to [50], we assume that their ratio is bounded by a constant, i.e., $\Delta_{\max}/\Delta_{\min} \leq k$, for some $k > 0$, ensuring that the design points are neither arbitrarily sparse nor overly clustered.

Theorem 1 (Upper Bound)**.** *Let $f \in \mathcal{W}^2(\Omega)$, and let $\hat{f}_{\mathrm{SS}}^a$ denote the smoothing spline estimator defined in* (3)*. Let $M = \max\{m_1, m_2\}$. Assume that $\lambda \to 0$ as $n \to \infty$ and $\lambda > n^{-2}$. Then, for sufficiently large n, we have*

$$R_2(f, \hat{f}_{\mathrm{SS}}^a) \lesssim \lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{1/4}} + \frac{q^2(M^2 + \sigma^2)}{n^2\lambda^{1/2}}, \tag{5}$$

and also

$$R_\infty(f, \hat{f}_{\mathrm{SS}}^a) \lesssim \lambda^{-1/4} \left(\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{1/4}} \right) + \frac{q^2(M^2 + \sigma^2)}{n^2\lambda^{1/2}}. \tag{6}$$

[2] Note that any function $f\colon [a, b] \to \mathbb{R}$ can be transformed with scaling and shifting into a function $\tilde{f}\colon [0, 1] \to \mathbb{R}$ without affecting its Sobolev regularity or the scaling of the associated metrics.

For the proof details, see Appendix A. Here, we present a proof sketch: We first decompose each metric into two components using the triangle inequality. Specifically, we have

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \leq 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] + 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right], \tag{7}$$

$$R_\infty(f, \hat{f}^a_{\mathrm{SS}}) \leq 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] + 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right], \tag{8}$$

where $\hat{f}_{\mathrm{SS}}$ denotes the smoothing spline estimator fitted on clean (uncorrupted) data. More precisely, we have

$$\hat{f}_{\mathrm{SS}} := \arg\min_{g \in \mathcal{W}^2(\Omega)} \left\{ \frac{1}{n}\sum_{i=1}^n \left(y_i - g(x_i)\right)^2 + \lambda \int_\Omega \left(g''(x)\right)^2\, dx \right\},$$

with $y_i = \tilde{y}_i$, for $i \in [n]\backslash\mathcal{A}$, and otherwise, for $i \in \mathcal{A}$, $y_i = f(x_i) + \epsilon_i$, for some i.i.d ϵ_i. In addition, we have $\hat{\boldsymbol{\epsilon}} = (\epsilon_i)_{i\in[n]}$.

The first term in each decompositions (7) and (8), quantifies the estimator's error in the absence of adversarial contamination, reflecting the classical estimation error. The second term, referred to as *adversarial deviation*, captures how much the adversarial estimator $\hat{f}^a_{\mathrm{SS}}$ deviates from its uncorrupted counterpart $\hat{f}_{\mathrm{SS}}$.

To bound the first term in decomposition (7), we rely on an established upper bounds for smoothing spline estimation [50, 51], which guarantee that, for sufficiently large n, we have

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f^{(j)} - \hat{f}^{(j)}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] \;\lesssim\; \lambda^{(2-j)/2} \int_\Omega \left(f''(x)\right)^2\, dx + \frac{\sigma^2}{n\lambda^{(2j+1)/4}}. \tag{9}$$

Applying (9) with $j = 0$ yields the desired bound for the first term in decomposition (7). For decomposition (8), the first term is bounded by combining (9) with $j = 0$ and $j = 1$, applying norm inequalities for Sobolev spaces [56], and using the Cauchy–Schwarz inequality, leading to

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] \;\lesssim\; \lambda^{-1/4}\left(\lambda \int_\Omega \left(f''(x)\right)^2\, dx + \frac{\sigma^2}{n\lambda^{1/4}}\right). \tag{10}$$

To bound the second term in decompositions (7) and (8), we leverage the fact that the smoothing spline estimator is a linear smoother [41]. Specifically, the solution to (3) can be expressed in a kernel form as

$$\hat{f}^a_{\mathrm{SS}}(x) = \frac{1}{n}\sum_{i=1}^n W_n(x, x_i)\, \widetilde{y}_i, \tag{11}$$

where $W_n(\cdot,\cdot)$ denotes the smoothing spline kernel (or weight function), which depends on the design points $\{x_i\}_{i\in[n]}$, sample size n, and the smoothing parameter λ. Using this representation, we have

$$\left|\hat{f}_{\mathrm{SS}}(x) - \hat{f}^a_{\mathrm{SS}}(x)\right| = \left|\frac{1}{n}\sum_{i=1}^n W_n(x, x_i)\,(y_i - \widetilde{y}_i)\right| \overset{(a)}{=} \left|\frac{1}{n}\sum_{i\in\mathcal{A}} W_n(x, x_i)\,(y_i - \widetilde{y}_i)\right|, \tag{12}$$

where (a) follows from the fact that $y_i = \tilde{y}_i$, for $i \in [n]\backslash\mathcal{A}$. Using the Hölder inequality, and taking expectation, we show that

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] \lesssim \frac{q^2(M^2+\sigma^2)}{n^2}\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}, x\in\Omega, j\in[n]} |W_n(x, x_j)|^2\right]. \tag{13}$$

Unfortunately, $W_n(\cdot,\cdot)$ does not admit an analytically tractable form [52, 53] for directly bounding its supremum in (13). However, a substantial body of research [52–55] has focused on approximating $W_n(\cdot,\cdot)$ with analytically tractable functions, known as *equivalent kernels*, denoted by $\widehat{W}_n(x, s)$. We leverage such approximations in our analysis to derive an upper bound for (13), leading to

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}^a_{\mathrm{SS}} - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] \;\lesssim\; \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}. \tag{14}$$

We also take similar steps to derive

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}}\left\|\hat{f}_{\mathrm{SS}}^{a}-\hat{f}_{\mathrm{SS}}\right\|_{L_\infty(\Omega)}^{2}\right] \lesssim \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}. \tag{15}$$

Combining (9), (10), (14), and (15) completes the proof of Theorem 1.

Corollary 1 (Convergence Rate of $R_2(f,\hat{f})$)**.** *Assume the conditions of Theorem 1 hold, and $q = \Theta(n^\beta)$ for some $\beta \in [0,1]$. Then, by choosing $\lambda = \mathcal{O}(n^{-4/5})$ for $\beta \le 0.4$ and $\lambda = \mathcal{O}(n^{-4/3(1-\beta)})$ for $\beta > 0.4$, we have*

$$\inf_{\hat{f}} R_2(f,\hat{f}) \le R_2(f,\hat{f}_{\mathrm{SS}}^{a}) \le \begin{cases} \mathcal{O}\left(n^{-4/5}\right) & \text{for } \beta \le 0.4, \\ \mathcal{O}\left(n^{-4/3(1-\beta)}\right) & \text{for } \beta > 0.4, \end{cases} \tag{16}$$

as depicted by the blue curve in Figure 1a.

Corollary 2 (Convergence Rate of $R_\infty(f,\hat{f})$)**.** *Under the same assumptions as in Corollary 1, by choosing $\lambda = \mathcal{O}(n^{-4/5})$ for $\beta \le 0.5$ and $\lambda = \mathcal{O}(n^{-8/5(1-\beta)})$ for $\beta > 0.5$, we have*

$$\inf_{\hat{f}} R_\infty(f,\hat{f}) \le R_\infty(f,\hat{f}_{\mathrm{SS}}^{a}) \le \begin{cases} \mathcal{O}\left(n^{-3/5}\right) & \text{for } \beta \le 0.5, \\ \mathcal{O}\left(n^{-6/5(1-\beta)}\right) & \text{for } \beta > 0.5, \end{cases} \tag{17}$$

as depicted by the blue curve in Figure 1b.

Based on Corollaries 1 and 2, the thresholds at $\beta = 0.4$ for $R_2(f,\hat{f}_{\mathrm{SS}}^{a})$ and $\beta = 0.5$ for $R_\infty(f,\hat{f}_{\mathrm{SS}}^{a})$ indicate a phase transition: Below these points, the estimation error is dominated by noise, and adversarial corruption has no impact on the convergence rate. Beyond these thresholds, the adversary dictates the rate of convergence. In this scenario, we must choose a larger smoothing parameter λ to smooth out the adversarial contribution in the data points.

Furthermore, for all $\beta \in [0,1)$, the convergence rate of $R_\infty(f,\hat{f}_{\mathrm{SS}}^{a})$ is slower than that of $R_2(f,\hat{f}_{\mathrm{SS}}^{a})$, as established by Theorem 1. This reflects the greater sensitivity of R_∞ estimation error to adversarial attacks: while metric R_2 averages the estimation error across the entire domain, metric R_∞ is driven by the worst-case pointwise error, making it inherently more vulnerable to adversarial perturbations.

In the following theorem we provide a minimax lower bound for both metrics.

Theorem 2. *Let $P_{\boldsymbol{\epsilon}}$ denote the probability density function of the noise vector $\boldsymbol{\epsilon}$, with i.i.d zero-mean σ^2-variance entries, $f \in \mathcal{W}^2(\Omega)$ be the regression function. Then*

$$\inf_{\hat{f}}\sup_{f,\mathcal{S},P_{\boldsymbol{\epsilon}}} R_2(f,\hat{f}) \gtrsim (\frac{q}{n})^3 + \frac{1}{n^{4/5}}, \tag{18}$$

$$\inf_{\hat{f}}\sup_{f,\mathcal{S},P_{\boldsymbol{\epsilon}}} R_\infty(f,\hat{f}) \gtrsim (\frac{q}{n})^2 + (\frac{\log n}{n})^{3/4}. \tag{19}$$

The full proof of Theorem 2 is provided in Appendix B. Here, we present a proof sketch. To establish Theorem 2, we reduce the minimax risk in (18) and (19) to a hypothesis testing problem [57]. Specifically, we construct two functions f_1 and f_2 in $\mathcal{W}^2(\Omega)$ with L_2 and L_∞ distance, bounded away from zero (see Figure 2). However, given n samples from either function, an adversary can corrupt up to q of them, making it impossible for any estimator to reliably distinguish between f_1 and f_2. Consequently, no estimation approach can identify which function generated the data, and the average hypothesis testing error remains $1/2$. Applying [57, Proposition 5.1] yields the following lower bounds:

$$\inf_{\hat{f}}\sup_{f,\mathcal{S},P_{\boldsymbol{\epsilon}}} R_2(f,\hat{f}) \gtrsim \left(\frac{q}{n}\right)^3, \tag{20}$$

$$\inf_{\hat{f}}\sup_{f,\mathcal{S},P_{\boldsymbol{\epsilon}}} R_\infty(f,\hat{f}) \gtrsim \left(\frac{q}{n}\right)^2. \tag{21}$$

Furthermore, when $q = 0$, the adversarial model reduces to the classical non-adversarial setting, for which the minimax lower bounds have been established as $\inf_{\hat{f}}\sup_{f,P_{\boldsymbol{\epsilon}}} R_2(f,\hat{f}) \gtrsim n^{-4/5}$ and $\inf_{\hat{f}}\sup_{f,P_{\boldsymbol{\epsilon}}} R_\infty(f,\hat{f}) \gtrsim (\log n/n)^{3/4}$ [24]. Combining these with (20) and (21) completes the proof of Theorem 2.

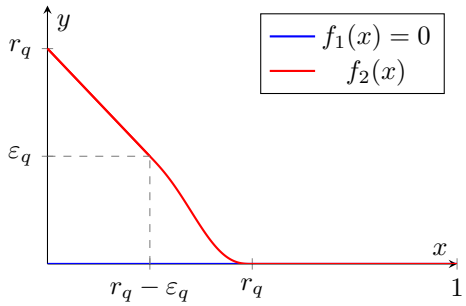


Figure 2: Construction of functions f_1 (blue) and f_2 (red) used in Theorem 2. Both functions belong to $\mathcal{W}^2([0,1])$, where $f_1(x) = 0$ for all x, and $f_2(x)$ differs from f_1 only on the interval $[0, r_q]$, with $r_q = q/n$. The function f_2 is linear on $[0, r_q - \varepsilon_q]$, where $\varepsilon_q = r_q^2$, and transitions smoothly to zero on $[r_q - \varepsilon_q, r_q]$ via a degree-5 polynomial, ensuring $f_2 \in \mathcal{W}^2([0,1])$. This construction induces a non-zero gap in both L_2 and L_∞ norms, while enabling the adversary to obscure the difference by corrupting only q samples, and making f_1, f_2 statistically indistinguishable. The details of this construction is provided in Appendix B.

Corollary 3. *Assuming $q = \Theta(n^\beta)$ for some $\beta \in [0,1]$, we conclude from Theorem 2 that*

$$\inf_{\hat{f}} \sup_{f,\mathcal{S},P_{\boldsymbol{\varepsilon}}} R_2(f,\hat{f})) \geq \begin{cases} \mathcal{O}\left(n^{-4/5}\right) & \text{for } \beta \leq \frac{11}{15}, \\ \mathcal{O}\left(n^{-3(1-\beta)}\right) & \text{for } \beta > \frac{11}{15}, \end{cases} \tag{22}$$

$$\inf_{\hat{f}} \sup_{f,\mathcal{S},P_{\boldsymbol{\varepsilon}}} R_\infty(f,\hat{f}) \geq \begin{cases} \tilde{\mathcal{O}}\left(n^{-3/4}\right) & \text{for } \beta \leq \frac{5}{8}, \\ \tilde{\mathcal{O}}\left(n^{-2(1-\beta)}\right) & \text{for } \beta > \frac{5}{8}, \end{cases} \tag{23}$$

as depicted by red curves in Figure 1.

Corollary 4 (On optimality of Smoothing Spline)**.** *Assume the conditions of Theorem 1 hold and that $q = o(n)$. Then, by selecting the smoothing parameter $\lambda = (\frac{q}{n})^{4/3}$ when $q \geq n^{0.4}$ and $\lambda = n^{-0.8}$ when $q < n^{0.4}$, $R_2(f, \hat{f}^a_{\mathrm{SS}})$ vanishes as $n \to \infty$. Similarly, for $R_\infty(f, \hat{f}^a_{\mathrm{SS}})$, setting $\lambda = (\frac{q}{n})^{6/5}$ when $q \geq n^{0.5}$ and $\lambda = n^{-0.8}$ when $q < n^{0.5}$ ensures that $R_\infty(f, \hat{f}^a_{\mathrm{SS}})$ also converges to zero. Consequently, as long as $q = o(n)$, then $R_2(f, \hat{f}^a_{\mathrm{SS}})$ and $R_\infty(f, \hat{f}^a_{\mathrm{SS}})$ go to zero, as $n \to \infty$. Conversely, if $q = \Theta(n)$, according to Corollary 3, there exists a function $f \in \mathcal{W}^2(\Omega)$ such that, for any estimator $\hat{f}$,* none *of $R_2(f,\hat{f})$ and $R_\infty(f,\hat{f})$ converges to zero as $n \to \infty$. This implies that the classical cubic smoothing spline estimator is optimal with respect to maximum tolerable number of adversarial corruptions.*

4 Experimental Results

In this section, we present numerical experiments to validate the theoretical results. All experiments are conducted on a single CPU-only machine. The smoothing spline estimator is implemented using the `SciPy` package [58]. We consider two regression functions: (i) $f(x) = x\sin(x)$ over the domain $\Omega = [-10, 10]$ with $M = 100$, and (ii) a three-layer MLP network with weights initialized in $[-1, 1]$ and $M = 500$. The noise vector $\boldsymbol{\varepsilon}$ is drawn independently from a Gaussian distribution with zero mean and variance $\sigma^2 = 1$.

To evaluate the adversarial robustness of the cubic smoothing spline estimator, we consider three distinct attack strategies:

- **Random Corruption Attack:** The adversary randomly selects q out of the n samples and replaces their response values with M.
- **Greedy Corruption Attack:** In this process, the attacker:
  1. Fits the baseline estimator on clean data.
  2. Computes $\ell_i = (\hat{f}(x_i) - y_i)^2$ for each sample.
  3. Identifies $i^\star = \arg\min_i \ell_i$.

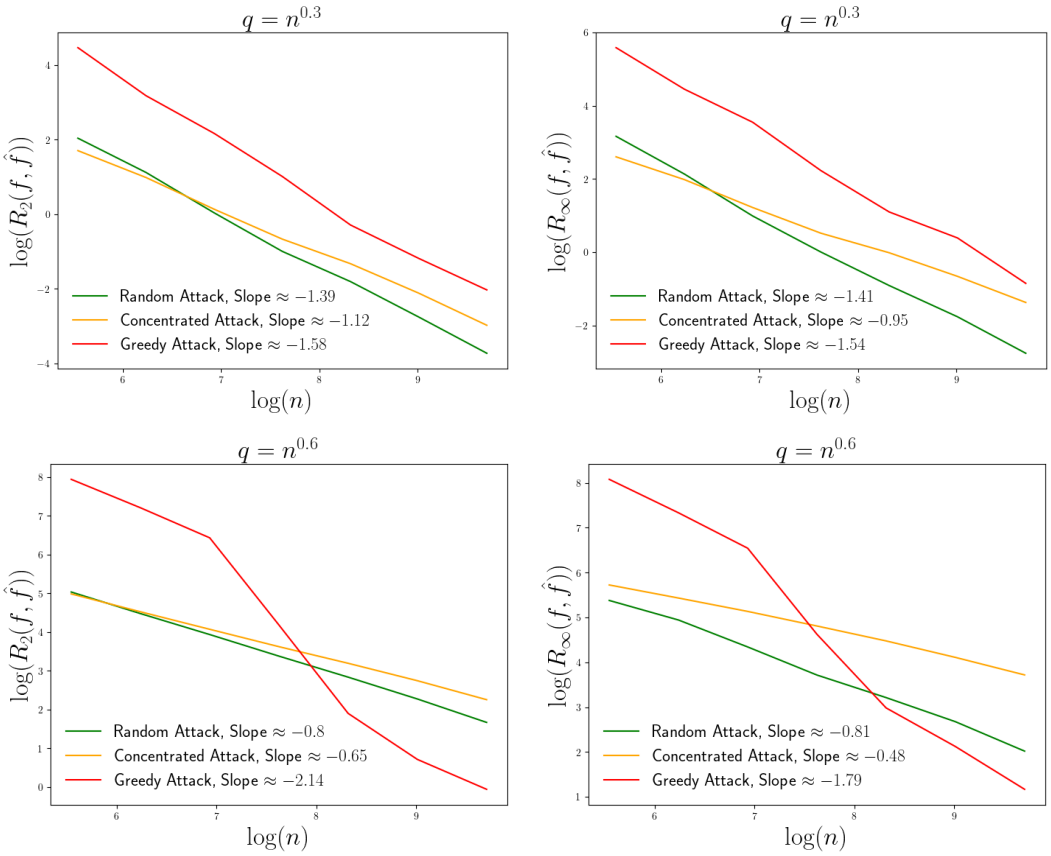


Figure 3: Log-log plots of error convergence rates for the cubic smoothing spline estimator $\hat{f} = \hat{f}^a_{\text{SS}}$ for $f(x) = x\sin(x)$ in the uniform design setting, where the input points converge to a uniform distribution. The top row shows $R_2(f, \hat{f})$ and $R_\infty(f, \hat{f})$ errors for $q = n^{0.3}$, along with the corresponding theoretical upper bounds of $\mathcal{O}(n^{-0.8})$ and $\mathcal{O}(n^{-0.6})$, respectively. The bottom row presents $R_2(f, \hat{f})$ and $R_\infty(f, \hat{f})$ errors for $q = n^{0.6}$, with theoretical upper bounds of $\mathcal{O}(n^{-0.53})$ and $\mathcal{O}(n^{-0.48})$, respectively.

4. Updates $y_{i^\star} \leftarrow y_{i^\star} + M \cdot \text{sign}(\hat{f}(x_{i^\star}) - y_{i^\star})$.
5. Repeats the process until q points are corrupted.

- **Concentrated Corruption Attack:** The adversary targets q consecutive samples centered around the median of the design points and modifies their corresponding labels to M.

For each attack strategy, we evaluate both $R_2(f, \hat{f}^a_{\text{SS}})$ and $R_\infty(f, \hat{f}^a_{\text{SS}})$ across a range of sample sizes n, and examine how these metrics scale with n under varying levels of adversarial corruption. Additionally, for each experiment, we consider two settings for the design points: uniform and Gaussian. In the uniform and Gaussian settings, the design points $\{x_i\}_{i=1}^n$ converge to a uniform and a truncated Gaussian distribution over Ω, respectively, as $n \to \infty$.

For the function $f(x) = x\sin(x)$, Figures 3 and 5 illustrate the behavior of $R_2(f, \hat{f}^a_{\text{SS}})$ and $R_\infty(f, \hat{f}^a_{\text{SS}})$ under uniform and Gaussian designs, respectively, for two corruption levels, $q = n^{0.3}$ and $q = n^{0.6}$. Similarly, for the MLP network, Figures 4 and 6 present the corresponding results.

As shown in these figures, the empirical convergence rates align well with the theoretical upper bounds established in Theorem 1. Specifically, Theorem 1 establishes that $R_2(f, \hat{f}^a_{\text{SS}}) \leq \mathcal{O}(n^{-0.8})$ and $R_\infty(f, \hat{f}^a_{\text{SS}}) \leq \mathcal{O}(n^{-0.6})$ for $q = n^{0.3}$, and $R_2(f, \hat{f}^a_{\text{SS}}) \leq \mathcal{O}(n^{-0.53})$ and $R_\infty(f, \hat{f}^a_{\text{SS}}) \leq \mathcal{O}(n^{-0.48})$ for $q = n^{0.6}$. These theoretical predictions align with the empirical convergence trends observed in

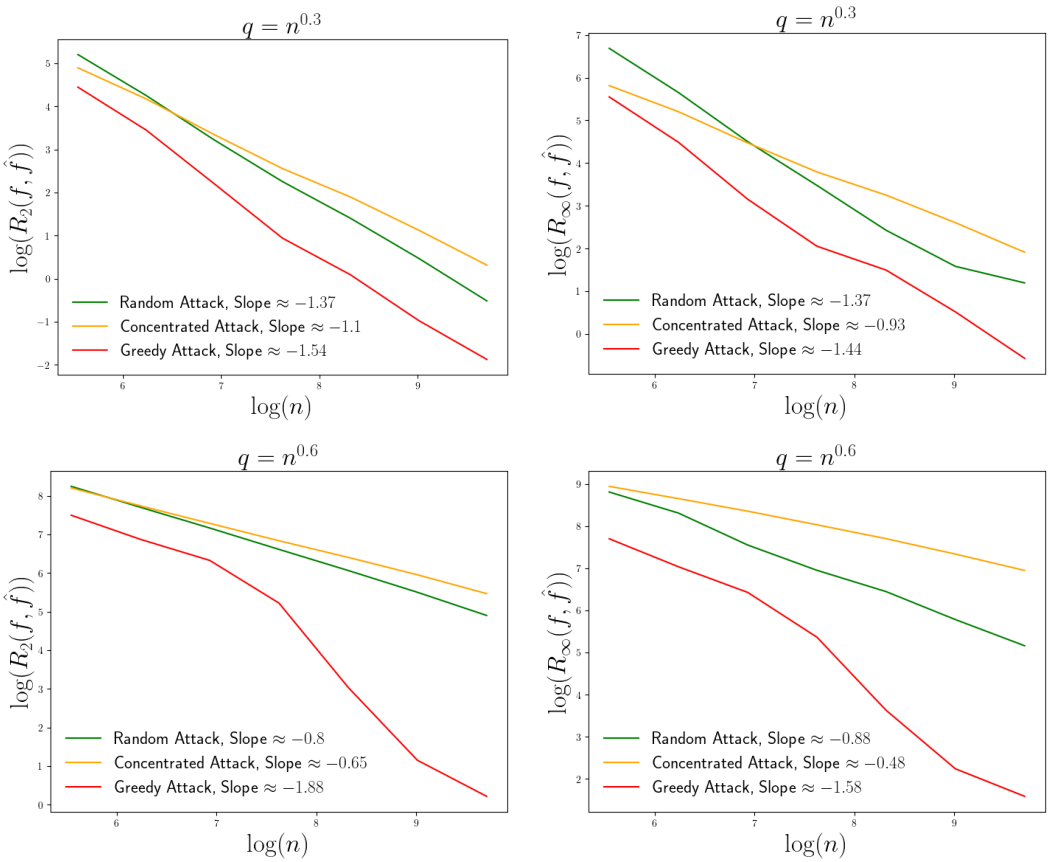


Figure 4: Log–log plots showing the convergence behavior of the cubic smoothing spline estimator $\hat{f} = \hat{f}^a_{\mathrm{SS}}$ when the ground-truth function is the MLP network, under the uniform design setting. The top row corresponds to the case $q = n^{0.3}$, with theoretical convergence rates of $\mathcal{O}(n^{-0.8})$ for $R_2(f, \hat{f})$ and $\mathcal{O}(n^{-0.6})$ for $R_\infty(f, \hat{f})$. The bottom row shows results for a higher corruption level, $q = n^{0.6}$, with respective theoretical upper bounds of $\mathcal{O}(n^{-0.53})$ and $\mathcal{O}(n^{-0.48})$.

the figures. Moreover, these figures show that the concentrated attack results in noticeably higher estimation error compared to the other two attack strategies.

It is important to note that these empirical rates are not expected to match the lower bounds from Theorem 2, since those bounds are minimax in nature. That is, they guarantee the existence of a worst-case function $f^\star \in \mathcal{W}^2(\Omega)$ for which no estimator can achieve faster convergence. Therefore, the lower bounds apply to such worst-case functions and not necessarily to all functions in $\mathcal{W}^2(\Omega)$, including the two functions in our experiments.

5 Related Work

Unlike parametric regression, which benefits from structural assumptions on the model class, nonparametric regression imposes minimal assumptions on the underlying function. This flexibility, makes it substantially more challenging to evaluate and guarantee adversarial robustness. Consequently, the literature on adversarial robustness in nonparametric settings remains relatively sparse. Nonetheless, several notable efforts have begun to address this gap.

As discussed earlier, the Nadaraya–Watson (NW) estimator is not robust to adversarial corruption and can fail even in the presence of a single corrupted sample [38, 39]. Classical robust estimation techniques, such as the Median-of-Means (MoM) estimator [59] and trimmed means [60], have been extended to nonparametric settings [37, 38, 61] to improve the robustness of the NW estimator. In the MoM approach, the data are partitioned into several groups, an NW estimator is fitted to each

group, and the median of the resulting estimates is taken. While this method enhances robustness to outliers, its performance degrades sharply when even a single corrupted sample appears in each group. Trimmed-mean methods, on the other hand, discard a fixed fraction of samples with extreme response values and fit the NW estimator on the remaining data. However, their effectiveness is limited when adversarial corruption is not uniformly distributed across the input space.

Zhao et al. [39] study adversarial robustness in kernel-based nonparametric regression by analyzing an M-estimator variant of the Nadaraya–Watson (NW) estimator [48, 62], deriving upper and minimax lower bounds for metrics based on L_2-norm and L_∞-norm. In comparison to our setting, which assumes the regression function lies in a second-order Sobolev space, their work considers a first-order Hölder class. Their proposed estimator requires gradient descent with $\mathcal{O}(n \log(1/\epsilon))$ complexity to produce an estimation with precision ϵ on new data point. In contrast, the cubic smoothing spline accurately evaluates new data point in $\mathcal{O}(n)$ time, offering greater computational efficiency.

Several works also have studied the robustness of nonparametric classification. In [40], the authors analyze the robustness of nonparametric linear classifiers under arbitrary norms and mild regularity assumptions. The robustness of nearest neighbor classifiers against adversarial perturbations has been studied in [63] and a general attack framework applicable to a wide class of nonparametric classifiers is introduced in [64] and a data-pruning defense strategy to mitigate such attacks is proposed.

6 Conclusion

In this paper, we study the adversarial robustness of nonparametric regression when the underlying regression function belongs to $\mathcal{W}^2(\Omega)$. We prove that the cubic smoothing spline achieves vanishing R_2 and R_∞ errors as long as the number of corrupted samples satisfies $q = o(n)$. We also establish lower bounds using a minimax argument. Notably, we show that cubic smoothing splines are optimal with respect to the maximum number of tolerable adversarial corruptions.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant CIF-2348638.

References

[1] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[3] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. Imagenet training in minutes. In *Proceedings of the 47th international conference on parallel processing*, pages 1–10, 2018.

[4] Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

[5] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[6] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[7] Hanzaleh Akbari Nodehi, Viveck R Cadambe, and Mohammad Ali Maddah-Ali. Game of coding: Sybil resistant decentralized machine learning with minimal trust assumption. *arXiv preprint arXiv:2410.05540*, 2024.

[8] Peiyu Xiong, Michael Tegegn, Jaskeerat Singh Sarin, Shubhraneel Pal, and Julia Rubin. It is all about data: A survey on the effects of data on adversarial robustness. *ACM Computing Surveys*, 56(7):1–41, 2024.

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

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

[11] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

[12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[13] Lianghe Shi and Weiwei Liu. Adversarial self-training improves robustness and generalization for gradual domain adaptation. *Advances in Neural Information Processing Systems*, 36: 37321–37333, 2023.

[14] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

[15] Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

[16] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[18] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

[19] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

[20] Parsa Moradi, Hanzaleh Akbarinodehi, and Mohammad Ali Maddah-Ali. General coded computing: Adversarial settings. *arXiv preprint arXiv:2502.08058*, 2025.

[21] Hanzaleh Akbari Nodehi, Viveck R Cadambe, and Mohammad Ali Maddah-Ali. Game of coding: Beyond trusted majorities. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2850–2855. IEEE, 2024.

[22] Wolfgang Härdle and Enno Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, pages 1926–1947, 1993.

[23] Wolfgang Härdle. *Applied nonparametric regression*. Cambridge university press, 1990.

[24] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

[25] Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

[26] Yue Xing, Ruizhi Zhang, and Guang Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.

[27] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[28] Aref Rekavandi, Farhad Farokhi, Olga Ohrimenko, and Benjamin Rubinstein. Certified adversarial robustness via randomized α-smoothing for regression models. *Advances in Neural Information Processing Systems*, 37:134127–134150, 2024.

[29] Yiling Xie and Xiaoming Huo. High-dimensional (group) adversarial training in linear regression. *arXiv preprint arXiv:2405.13940*, 2024.

[30] Antonio Ribeiro, Dave Zachariah, Francis Bach, and Thomas Schön. Regularization properties of adversarially-trained linear regression. *Advances in Neural Information Processing Systems*, 36:23658–23670, 2023.

[31] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *IEEE Transactions on Information Theory*, 69(12):7793–7822, 2023.

[32] Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

[33] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[34] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[35] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

[36] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pages 727–757. PMLR, 2019.

[37] Anna Ben-Hamou and Arnaud Guyader. Robust non-parametric regression via median-of-means. `https://hal.science/hal-03957385v1/document`, 2023. HAL preprint, hal-03957385.

[38] Subhra Dhar, Prashant Jha, and Prabrisha Rakshit. The trimmed mean in non-parametric regression function estimation. *Theory of Probability and Mathematical Statistics*, 107:133–158, 2022.

[39] Puning Zhao and Zhiguo Wan. Robust nonparametric regression under poisoning attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17007–17015, 2024.

[40] Elvis Dohmatob. Consistent adversarially robust linear classification: non-parametric setting. In *Forty-first International Conference on Machine Learning*, 2024.

[41] Grace Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 24(5): 383–393, 1975.

[42] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[43] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121, 1996.

[44] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[45] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.

[46] Chuanbo Hua, Federico Berto, Michael Poli, Stefano Massaroli, and Jinkyoo Park. Learning efficient surrogate dynamic models with graph spline networks. *Advances in Neural Information Processing Systems*, 36:52523–52547, 2023.

[47] Junhui He, Ying Yang, and Jian Kang. Adaptive bayesian multivariate spline knot inference with prior specifications on model complexity. *arXiv preprint arXiv:2405.13353*, 2024.

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

[49] Felix Abramovich and Vadim Grinshtein. Derivation of equivalent kernel for general spline smoothing: a systematic approach. *Bernoulli*, 5(1):109–123, 1999.

[50] Florencio I Utreras. Convergence rates for multivariate smoothing spline functions. *Journal of approximation theory*, 52(1):1–27, 1988.

[51] David L Ragozin. Error bounds for derivative estimates based on spline smoothing of exact or noisy data. *Journal of approximation theory*, 37(4):335–355, 1983.

[52] Bernard W Silverman. Spline smoothing: the equivalent variable kernel method. *The annals of Statistics*, pages 898–916, 1984.

[53] K Messer. A comparison of a spline estimate to its equivalent kernel estimate. *The Annals of Statistics*, pages 817–829, 1991.

[54] Karen Messer and Larry Goldstein. A new class of kernels for nonparametric curve estimation. *The Annals of Statistics*, pages 179–195, 1993.

[55] Douglas Nychka. Splines as local smoothers. *The Annals of Statistics*, pages 1175–1197, 1995.

[56] Giovanni Leoni. *A first course in Sobolev spaces*, volume 181. American Mathematical Society, 2024.

[57] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[58] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[59] Arkadij Semenovich Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

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

[61] Pierre Humbert, Batiste Le Bars, and Ludovic Minvielle. Robust kernel density estimation with median-of-means principle. In *International Conference on Machine Learning*, pages 9444–9465. PMLR, 2022.

[62] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[63] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5133–5142. PMLR, 2018.

[64] Yao-Yuan Yang, Cyrus Rashtchian, Yizhen Wang, and Kamalika Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pages 941–951. PMLR, 2020.

[65] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[66] Parsa Moradi, Behrooz Tahmasebi, and Mohammad Maddah-Ali. Coded computing for resilient distributed computing: A learning-theoretic framework. *Advances in Neural Information Processing Systems*, 37:111923–111964, 2024.

[67] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[68] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

A Proof of Theorem 1

In this section, we prove Theorem 1. Without loss of generality, we assume that $\Omega = [0,1]$ [3]. Recall that the solution of (3) is unique and the explicit formula for $\hat{f}^a_{\mathrm{SS}}$ is given by

$$\hat{f}^a_{\mathrm{SS}}(x) = \frac{1}{n}\sum_{i=1}^{n} W_n(x, x_i)\, \widetilde{y}_i, \tag{24}$$

where $W_n(x, x_i)$ denotes the smoothing spline weight function depending on $\{x_i\}_{i=1}^n$, the sample size n, and the smoothing parameter λ.

To facilitate the analysis, we define a second scenario in which the adversarial strategy is to be *honest*, that is, for any $i \in \mathcal{A}$, the adversary does not deviate from the clean data generation process and behaves as if it were non-adversarial. This allows us to construct a one-to-one correspondence between the realizations of the adversarial and honest scenarios such that for each $i \notin \mathcal{A}$, the observed responses y_i are identical across both settings, while for $i \in \mathcal{A}$, the responses may differ: in Scenario 1 (adversarial), the adversary may introduce arbitrary deviations, whereas in Scenario 2 (honest), the responses follow the true underlying model.

In this second setting, we apply the same smoothing spline estimator to the uncorrupted data. The resulting estimator, which we denote by $\hat{f}_{\mathrm{SS}}$, is given by

$$\hat{f}_{\mathrm{SS}}(x) = \frac{1}{n}\sum_{i=1}^{n} W_n(x, x_i)\, y_i, \tag{25}$$

where y_i denotes the uncorrupted response corresponding to input x_i, i.e., $y_i = \tilde{y}_i$, for $i \in [n]\backslash\mathcal{A}$, and otherwise, for $i \in \mathcal{A}$, $y_i = f(x_i) + \epsilon_i$, for some i.i.d ϵ_i. We define $\hat{\boldsymbol{\epsilon}} := (\epsilon_i)_{i\in[n]}$.

We now proceed to prove the bounds stated in Theorem 1. We first establish the upper bound for $R_2(f, \hat{f}^a_{\mathrm{SS}})$ in (5), and subsequently turn to the bound for $R_\infty(f, \hat{f}^a_{\mathrm{SS}})$ in (6).

By the definition of $R_2(f, \hat{f}^a_{\mathrm{SS}})$, we have

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right], \tag{26}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i\in[n]\backslash\mathcal{A}}$. First, observe that

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] = \mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right],$$

which follows from the fact that $\left\| f - \hat{f}^a_{\mathrm{SS}} \right\|$ is independent of the noise terms $(\varepsilon_i)_{i\in\mathcal{A}}$. To proceed, we add and subtract $\hat{f}_{\mathrm{SS}}(x)$ inside the squared term:

$$\left(f(x) - \hat{f}^a_{\mathrm{SS}}(x)\right)^2 = \left(f(x) - \hat{f}_{\mathrm{SS}}(x) + \hat{f}_{\mathrm{SS}}(x) - \hat{f}^a_{\mathrm{SS}}(x)\right)^2. \tag{27}$$

Using AM-GM inequality, we obtain

$$\left(f(x) - \hat{f}^a_{\mathrm{SS}}(x)\right)^2 \le 2\left(f(x) - \hat{f}_{\mathrm{SS}}(x)\right)^2 + 2\left(\hat{f}_{\mathrm{SS}}(x) - \hat{f}^a_{\mathrm{SS}}(x)\right)^2. \tag{28}$$

Substituting this bound into the definition of $R_2(f, \hat{f}^a_{\mathrm{SS}})$, we get

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \le 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] + 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right]. \tag{29}$$

To prove the upper bound in (5), it suffices to find appropriate bounds for the two terms appearing in (29). We begin by analyzing the first term involving the honest estimator $\hat{f}_{\mathrm{SS}}$:

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right].$$

[3] Note that any function $f\colon [a,b] \to \mathbb{R}$ can be transformed with scaling and shifting into a function $\tilde{f}\colon [0,1] \to \mathbb{R}$ without affecting its Sobolev regularity or the scaling of the associated metrics.

To do so, we use the following theorem, which is a direct consequence of [50, Therorem 1.1], specialized to the second-order Sobolev space setting:

Lemma 1. *Let $I = [a, b] \subset \mathbb{R}$ be a bounded interval, and let the design points $\{x_i\}_{i=1}^n \subset I$ satisfy the quasi-uniformity condition*

$$\frac{\Delta_{\max}}{\Delta_{\min}} \leq k, \tag{30}$$

for some constant $k > 0$, where

$$\Delta_{\max} := \sup_{x \in I} \min_{i=1,\ldots,n} |x - x_i|, \quad \Delta_{\min} := \min_{i \neq j} |x_i - x_j|. \tag{31}$$

Then, for any $j = 0, 1, 2$, there exist constants $\lambda_0 > 0$, $P_0 > 0$, and $Q_0 > 0$, such that for all $n^{-4} \leq \lambda \leq \lambda_0$, we have

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}} \left[\left\| f^{(j)} - \hat{f}_{\mathrm{SS}}^{(j)} \right\|_{L_2(I)}^2 \right] \leq P_0 \, \lambda^{\frac{2-j}{2}} \int_I \left(f''(x) \right)^2 dx + \frac{Q_0 \, \sigma^2}{n \, \lambda^{\frac{2j+1}{4}}} \tag{32}$$

Here, $\hat{f}_{\mathrm{SS}}$ is the smoothing spline estimator applied to uncorrupted data, and $f^{(j)}$ denotes the j-th derivative of f. To bound the first term in (29), we invoke Lemma 1 with $j = 0$, corresponding to the $L_2(\Omega)$ error between the regression function f and the honest smoothing spline estimator $\hat{f}_{\mathrm{SS}}$. This yields:

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}} \left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|_{L_2(\Omega)}^2 \right] \leq P_0 \, \lambda \int_\Omega \left(f''(x) \right)^2 \, dx + \frac{Q_0 \, \sigma^2}{n \, \lambda^{1/4}}, \tag{33}$$

where λ is the regularization parameter, and $P_0, Q_0 > 0$ are constants from Lemma 1.

To complete the proof of (5), we now seek to find an upper bound for the second term in (29), which captures the deviation between the adversarial and honest estimators:

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}} \left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}_{\mathrm{SS}}^a \right\|_{L_2(\Omega)}^2 \right].$$

Note that from the kernel representations (24) and (25), we have

$$\hat{f}_{\mathrm{SS}}(x) - \hat{f}_{\mathrm{SS}}^a(x) = \frac{1}{n} \sum_{i=1}^n W_n(x, x_i) \, (y_i - \widetilde{y}_i). \tag{34}$$

Thus, for each $x \in \Omega$,

$$\left| \hat{f}_{\mathrm{SS}}(x) - \hat{f}_{\mathrm{SS}}^a(x) \right| = \left| \frac{1}{n} \sum_{i=1}^n W_n(x, x_i) \, (y_i - \widetilde{y}_i) \right|. \tag{35}$$

Note that for each i:

- If $i \notin \mathcal{A}$, there is no corruption, and $y_i = \widetilde{y}_i$.
- If $i \in \mathcal{A}$, the adversary may modify y_i, and since $f(x_i), \widetilde{y}_i \in [-M, M]$, we have

$$|y_i - \widetilde{y}_i| = |f(x_i) - \widetilde{y}_i + \epsilon_i| \leq |f(x_i) - \widetilde{y}_i| + |\epsilon_i| \leq 2M + |\epsilon_i| \, .$$

Thus, the sum above reduces to

$$\frac{1}{n} \sum_{i \in \mathcal{A}} W_n(x, x_i) \, (y_i - \widetilde{y}_i),$$

and we can bound

$$\left| \hat{f}_{\mathrm{SS}}(x) - \hat{f}_{\mathrm{SS}}^a(x) \right| \leq \frac{1}{n} \sum_{j \in \mathcal{A}} |W_n(x, x_j)| \, (2M + |\epsilon_i|) \leq \frac{1}{n} \sup_{x \in \Omega, j \in [n]} |W_n(x, x_j)| \cdot \sum_{j \in \mathcal{A}} (2M + \epsilon_i) \, . \tag{36}$$

This implies that

$$\begin{aligned}\left\|\hat{f}_{\mathrm{SS}}-\hat{f}_{\mathrm{SS}}^{a}\right\|_{L_2(\Omega)}^2 &\leq \left(\frac{\sup_{x\in\Omega,j\in[n]}|W_n(x,x_j)|}{n}\right)^2\cdot\left(\sum_{i\in\mathcal{A}}(2M+|\epsilon_i|)\right)^2\\ &\overset{\text{(a)}}{\leq}\left(\frac{\sup_{x\in\Omega,j\in[n]}|W_n(x,x_j)|}{n}\right)^2\cdot\left(\sum_{i\in\mathcal{A}}1^2\right)\cdot\left(\sum_{i\in\mathcal{A}}(2M+|\epsilon_i|)^2\right)\\ &\overset{\text{(b)}}{\leq}\left(\frac{\sup_{x\in\Omega,j\in[n]}|W_n(x,x_j)|}{n}\right)^2\cdot q\cdot\sum_{i\in\mathcal{A}}\left(8M^2+2|\epsilon_i|^2\right)\\ &=\left(\frac{\sup_{x\in\Omega,j\in[n]}|W_n(x,x_j)|}{n}\right)^2\cdot q\cdot\left(8M^2q+2\sum_{i\in\mathcal{A}}\epsilon_i^2\right),\end{aligned}\tag{37}$$

where (a) and (b) follow from the Cauchy–Schwarz and AM–GM inequalities, respectively.

Taking expectations and supremum over $\mathcal{S}$ yields

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}}\left\|\hat{f}_{\mathrm{SS}}-\hat{f}_{\mathrm{SS}}^{a}\right\|_{L_2(\Omega)}^2\right]\leq\frac{q^2(8M^2+2\sigma^2)}{n^2}\sup_{x\in\Omega,j\in[n]}|W_n(x,x_j)|^2.\tag{38}$$

Now, to complete the proof of (5), it remains to find an upper bound for the kernel supremum term

$$\sup_{x,j\in[n]}|W_n(x,x_j)|.$$

Unfortunately, $W_n(\cdot,\cdot)$ does not admit an analytically tractable form [52, 53] for directly bounding its supremum in (13). However, a substantial body of research [52–55] has focused on approximating $W_n(\cdot,\cdot)$ with analytically tractable functions, known as *equivalent kernels*, denoted by $\widehat{W}_n(x,s)$. We leverage such approximations in our analysis to derive an upper bound.

Recall that we define the empirical distribution function F_n as

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i\leq x\}.\tag{39}$$

We assume that the empirical distribution function F_n converges to a cumulative distribution function F, i.e., $\alpha(n):=\sup_{x\in\Omega}|F_n(x)-F(x)|$ satisfies $\alpha(n)\longrightarrow 0$ as $n\to\infty$. Moreover, we assume that $F(x)$ is differentiable on Ω with density $p(x)=F'(x)$, and that there exists a constant $p_{\min}>0$ such that

$$\inf_{x\in\Omega}p(x)\geq p_{\min}.\tag{40}$$

To proceed, according to [49], we define the equivalent kernel $\widehat{W}_n(x,s)$ as

$$\widehat{W}_n(x,s)=\frac{\lambda^{-1/4}}{2}\left(p(s)p(x)\right)^{-3/8}e^{-\lambda^{-1/4}\varphi_0(x,s)}\sin\left(\lambda^{-1/4}\varphi_0(x,s)+\frac{\pi}{4}\right),\tag{41}$$

where the phase function $\varphi_0(x,s)$ is given by

$$\varphi_0(x,s)=2^{-1/2}\int_{\min(x,s)}^{\max(x,s)}p(t)^{1/4}\,dt.\tag{42}$$

Based on [49, Theorem 1], for sufficiently large n, we have

$$\left|\widehat{W}_n(x,s)-W_n(x,s)\right|\leq C\left(\lambda^{-1/2}\alpha(n)+1\right),\tag{43}$$

where $C>0$ is a constant independent of n, and the bound holds uniformly over all $x\in[0,1]$ and $s\in[\tau_1,\tau_2]$, where $0<\tau_1<\tau_2<1$.

Now note that

$$\sup_{x\in\Omega, j\in[n]} |W_n(x,x_j)| = \sup_{x\in\Omega, j\in[n]} \left|\widehat{W}_n(x,x_j) + \left(W_n(x,x_j) - \widehat{W}_n(x,x_j)\right)\right| \tag{44}$$

$$\leq \sup_{x\in\Omega, j\in[n]} \left|\widehat{W}_n(x,x_j)\right| + \sup_{x\in\Omega, j\in[n]} \left|W_n(x,x_j) - \widehat{W}_n(x,x_j)\right|. \tag{45}$$

Using the uniform approximation property established in (43), we can bound the second term:

$$\sup_{x\in\Omega, j\in[n]} \left|W_n(x,x_j) - \widehat{W}_n(x,x_j)\right| \leq C\left(\lambda^{-1/2}\alpha(n) + 1\right). \tag{46}$$

Thus,

$$\sup_{x\in\Omega, j\in[n]} |W_n(x,x_j)| \leq \sup_{x\in\Omega, j\in[n]} \left|\widehat{W}_n(x,x_j)\right| + C\left(\lambda^{-1/2}\alpha(n) + 1\right)$$

$$\overset{(a)}{\leq} \frac{\lambda^{-1/4}}{2}\left(p_{\min}\right)^{-3/4} + C\left(\lambda^{-1/2}\alpha(n) + 1\right) \tag{47}$$

where (a) follows from the definition of $\widehat{W}_n(x,x_j)$ in (41), and the fact that $\inf_{x\in\Omega} p(x) \geq p_{\min}$. Combining the decomposition in (29), the bound on the honest estimator error from (33), and the adversarial deviation bounds from (38) and (47), we obtain the final upper bound for $R_2(f, \hat{f}^a_{\mathrm{SS}})$ stated in Theorem 1:

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \leq 2P_0\,\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{2Q_0\sigma^2}{n\lambda^{1/4}}$$

$$+ \frac{2q^2(8M^2+2\sigma^2)}{n^2}\left[\frac{\lambda^{-1/4}}{2}(p_{\min})^{-3/4} + C\left(\lambda^{-1/2}\alpha(n)+1\right)\right]^2. \tag{48}$$

Therefore, in the regime where $\lambda \to 0$ as $n \to \infty$ and $\lambda > n^{-2} > n^{-4}$, there exist constants E_1, E_2, E_3 such that for sufficiently large n,

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \leq E_1\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{E_2\sigma^2}{n\lambda^{1/4}} + \frac{E_3 q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}\left(1 + \lambda^{-1/4}\alpha(n) + \lambda^{1/4}\right)^2. \tag{49}$$

Since $\lambda^{1/4} \to 0$ as $n \to \infty$, the additive term $\lambda^{1/4}$ becomes negligible compared to 1 for sufficiently large n. Dropping this term and absorbing constants, we obtain

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \lesssim \lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{1/4}} + \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}\left(1 + \lambda^{-1/4}\alpha(n)\right)^2. \tag{50}$$

For a continuous cumulative distribution function F, Serfling [65] shows that $\alpha(n) = n^{-1/2}\log\log n$ almost surely. Since $\lambda > n^{-2}$, it follows that $\lambda^{-1/4}\alpha(n) \to 0$ as $n \to \infty$. Therefore, for sufficiently large n, we have $1 + \lambda^{-1/4}\alpha(n) < 2$. As a result, we obtain

$$R_2(f, \hat{f}^a_{\mathrm{SS}}) \lesssim \lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{1/4}} + \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}. \tag{51}$$

This concludes the proof of the upper bound on $R_2(f, \hat{f}^a_{\mathrm{SS}})$ in Theorem 1.

To complete the proof of Theorem 1, it remains to prove (6). To do so, we adopt a similar strategy as in the L_2 case, but adapted to the squared supremum norm. By the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\left\|f - \hat{f}^a_{\mathrm{SS}}\right\|^2_{L_\infty(\Omega)} \leq 2\left\|f - \hat{f}_{\mathrm{SS}}\right\|^2_{L_\infty(\Omega)} + 2\left\|\hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}}\right\|^2_{L_\infty(\Omega)}. \tag{52}$$

Taking expectation and supremum over $\mathcal{S}$, we substitute into the definition of R_∞ and obtain

$$R_\infty(f, \hat{f}^a_{\mathrm{SS}}) = \mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right]$$
$$\leq 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] + 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right]. \tag{53}$$

From the pointwise bound established in (36), we have

$$\left| \hat{f}_{\mathrm{SS}}(x) - \hat{f}^a_{\mathrm{SS}}(x) \right| \leq \frac{2q(M + \max_i |\epsilon_i|)}{n} \sup_{x\in\Omega, j\in[n]} |W_n(x, x_j)|\,. \tag{54}$$

Applying the kernel estimate from (47), we conclude that

$$\left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|_{L_\infty(\Omega)} \leq \frac{2q(M + \max_i |\epsilon_i|)}{n} \left(\frac{\lambda^{-1/4}}{2} (p_{\min})^{-3/4} + C\left(\lambda^{-1/2}\alpha(n) + 1\right)\right). \tag{55}$$

Squaring both sides and taking expectation and supremum over $\mathcal{S}$, we obtain

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| \hat{f}_{\mathrm{SS}} - \hat{f}^a_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] \lesssim \frac{q^2(M^2 + \sigma^2)}{n^2\lambda^{1/2}} \left(1 + \lambda^{-1/4}\alpha(n) + \lambda^{1/4}\right)^2. \tag{56}$$

To complete the proof of (6), it remains to find an upper bound for the first term in (53), namely

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right].$$

To do so, Since $f - \hat{f}_{SS} \in \mathcal{W}^2(\Omega)$, we can leverage Sobolev norms inequalities [56] and use the same arguments as in [66, Lemma 5] and obtain:

$$\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)} \leq 2 \left\| f - \hat{f}_{\mathrm{SS}} \right\|_{L_2(\Omega)} \cdot \left\| f' - \hat{f}'_{\mathrm{SS}} \right\|_{L_2(\Omega)} \tag{57}$$

Taking expectations on both sides of (57) and applying the Cauchy–Schwarz inequality, we obtain:

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] \leq 2\,\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|_{L_2(\Omega)} \cdot \left\| f' - \hat{f}'_{\mathrm{SS}} \right\|_{L_2(\Omega)}\right]$$
$$\leq 2\left(\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right]\right)^{1/2} \cdot \left(\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f' - \hat{f}'_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right]\right)^{1/2}. \tag{58}$$

Applying Lemma 1 with $j = 0$ and $j = 1$, we can bound the right-hand side using:

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] \leq P_0\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{Q_0\sigma^2}{n\lambda^{1/4}}, \tag{59}$$

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f' - \hat{f}'_{\mathrm{SS}} \right\|^2_{L_2(\Omega)}\right] \leq P_0\lambda^{1/2} \int_\Omega \left(f''(x)\right)^2 dx + \frac{Q_0\sigma^2}{n\lambda^{3/4}}. \tag{60}$$

Substituting the bounds from (59) and (60) into (57), we obtain

$$\mathbb{E}_{\hat{\boldsymbol{\epsilon}}}\left[\left\| f - \hat{f}_{\mathrm{SS}} \right\|^2_{L_\infty(\Omega)}\right] \leq 2\left(P_0\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{Q_0\sigma^2}{n\lambda^{1/4}}\right)^{1/2}$$
$$\times \left(P_0\lambda^{1/2} \int_\Omega \left(f''(x)\right)^2 dx + \frac{Q_0\sigma^2}{n\lambda^{3/4}}\right)^{1/2}. \tag{61}$$

Combining the decomposition in (53) with the bounds from (61) and (56), we obtain the following upper bound in the regime where $\lambda \to 0$ as $n \to \infty$ and $\lambda > n^{-2} \geq n^{-4}$:

$$R_\infty(f, \hat{f}^a_{\mathrm{SS}}) \lesssim \left(\lambda \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{1/4}}\right)^{1/2} \times \left(\lambda^{1/2} \int_\Omega \left(f''(x)\right)^2 dx + \frac{\sigma^2}{n\lambda^{3/4}}\right)^{1/2}$$
$$+ \frac{q^2(M^2 + \sigma^2)}{n^2\lambda^{1/2}} \left(1 + \lambda^{-1/4}\alpha(n) + \lambda^{1/4}\right)^2. \tag{62}$$

We now multiply and divide the first term by $\lambda^{1/4}$, yielding:

$$R_\infty(f, \hat{f}^a_{\mathrm{SS}}) \lesssim \lambda^{-1/4} \left(\lambda \int_\Omega \left(f''(x)\right)^2 \, dx + \frac{\sigma^2}{n\lambda^{1/4}} \right) + \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}} \left(1 + \lambda^{-1/4}\alpha(n) + \lambda^{1/4}\right)^2.$$

By arguments similar to those used in the bound for $R_2(f, \hat{f}^a_{\mathrm{SS}})$, we can neglect both $\lambda^{1/4}$ and $\lambda^{-1/4}\alpha(n)$ compared to 1 for sufficiently large n. Thus, we obtain

$$R_\infty(f, \hat{f}^a_{\mathrm{SS}}) \lesssim \lambda^{-1/4} \left(\lambda \int_\Omega \left(f''(x)\right)^2 \, dx + \frac{\sigma^2}{n\lambda^{1/4}} \right) + \frac{q^2(M^2+\sigma^2)}{n^2\lambda^{1/2}}. \tag{63}$$

This completes the proof of the upper bound on $R_\infty(f, \hat{f}^a_{\mathrm{SS}})$ in (6), and thereby concludes the proof of Theorem 1.

B Proof of Theorem 2

To prove Theorem 2, we first state and prove Lemma 2.

Lemma 2. *Let P_1 and P_2 denote two probability density functions of two distributions with common variance $\sigma^2 > 0$. Then, there exists $\alpha \in [0,1]$, and two probability density functions Q_1 and Q_2 such that*

$$(1-\alpha)P_1 + \alpha Q_1 = (1-\alpha)P_2 + \alpha Q_2, \tag{64}$$

where Q_1 and Q_2 are explicitly constructed from P_1 and P_2.

Proof. Define α as:

$$\alpha = \frac{\int_{\{u:P_2(u)\geq P_1(u)\}} \left(P_2(u) - P_1(u)\right) \, du}{1 + \int_{\{u:P_2(u)\geq P_1(u)\}} \left(P_2(u) - P_1(u)\right) \, du} \leq 1. \tag{65}$$

Next, define Q_1 and Q_2 as:

$$Q_1(u) = \frac{1-\alpha}{\alpha} \left(P_2(u) - P_1(u)\right) \mathbf{1}\{P_2(u) \geq P_1(u)\}, \tag{66}$$

$$Q_2(u) = \frac{1-\alpha}{\alpha} \left(P_1(u) - P_2(u)\right) \mathbf{1}\{P_1(u) > P_2(u)\}, \tag{67}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

By construction, both $Q_1(u)$ and $Q_2(u)$ are non-negative since the indicator functions restrict the support to regions where the corresponding differences are non-negative. We now show that Q_1 and Q_2 are valid probability density functions. Consider:

$$\begin{aligned} \int Q_1(u)\, du &= \frac{1-\alpha}{\alpha} \int \left(P_2(u) - P_1(u)\right) \mathbf{1}\{P_2(u) \geq P_1(u)\} \, du \\ &= \frac{1-\alpha}{\alpha} \int_{\{u:P_2(u)\geq P_1(u)\}} \left(P_2(u) - P_1(u)\right) \, du = 1. \end{aligned} \tag{68}$$

By symmetry, the same argument shows that $\int Q_2(u)\, du = 1$ as well.

Hence, both Q_1 and Q_2 are valid densities. With this choice of α, the following identity holds:

$$(1-\alpha)P_1 + \alpha Q_1 = (1-\alpha)P_2 + \alpha Q_2. \tag{69}$$

This completes the proof. □

We now prove Theorem 2, building on Lemma 2. We begin by establishing the lower bound for the metric R_2, as stated in (18); the proof for R_∞, given in (19), follows by a similar argument. To do so, we reduce the minimax risk in (18) and (19) to a hypothesis testing problem [57]. Specifically, we construct two functions f_1 and f_2 in $\mathcal{W}^2(\Omega)$ with L_2 and L_∞ distance, bounded away from zero (see Figure 2). However, given n samples from either function, an adversary can corrupt up to q of them, making it impossible for any estimator to reliably distinguish between f_1 and f_2. Consequently,

no estimation approach can identify which function generated the data, and the average hypothesis testing error remains $1/2$. Applying [57, Proposition 5.1] yields the lower bounds in Theorem 2. The details of the proof is as follows.

Throughout the proof, we assume a fixed design given by $x_i = i/n$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d noise samples drawn from a normal distribution with zero mean and variance σ^2, for $i \in [n]$.

Let $r_q = \frac{q}{n}$ and define $\varepsilon_q = r_q^2$. We construct two functions, f_1 and f_2, as follows. Set

$$f_1(x) = 0 \quad \text{for all } x \in [0, 1].$$

To define f_2, we construct a degree-5 polynomial $g(x)$ on the interval $[r_q - \varepsilon_q, r_q]$ that satisfies the following conditions:

$$g(r_q - \varepsilon_q) = \varepsilon_q, \tag{70}$$
$$g'(r_q - \varepsilon_q) = -1, \tag{71}$$
$$g''(r_q - \varepsilon_q) = 0, \tag{72}$$
$$g(r_q) = 0, \tag{73}$$
$$g'(r_q) = 0, \tag{74}$$
$$g''(r_q) = 0. \tag{75}$$

These six conditions uniquely determine a polynomial of degree 5, since there are six coefficients to solve for. Hence, such a polynomial g exists and can be explicitly constructed. Now, define f_2 on the interval $[0, 1]$ by

$$f_2(x) = \begin{cases} r_q - x, & \text{if } x \in [0, r_q - \varepsilon_q], \\ g(x), & \text{if } x \in [r_q - \varepsilon_q, r_q], \\ 0, & \text{if } x > r_q. \end{cases}$$

It is straightforward to verify that $f_2 \in \mathcal{W}^2([0, 1])$, since both f_2 and its first and second derivatives have bounded norms over Ω (See Figure 2).

Note that f_1 and f_2 are close but not identical; their differences are concentrated on the interval $[0, r_q]$, and will be used to construct the lower bound.

For each sample x_i, the adversary proceeds as follows:

- If $x_i \geq r_q$, then $f_1(x_i) = f_2(x_i)$, so no corruption is needed: both models produce identical distributions for $\widetilde{y}_i$.
- If $x_i < r_q$, then $f_1(x_i) \neq f_2(x_i)$, and the adversary applies Lemma 2 to the pair of normal distributions
$$P_1^{(i)} := \mathcal{N}(f_1(x_i), \sigma^2), \qquad P_2^{(i)} := \mathcal{N}(f_2(x_i), \sigma^2),$$
obtaining a scalar $\alpha_i \in [0, 1]$ and auxiliary distributions $Q_1^{(i)}$ and $Q_2^{(i)}$ such that
$$(1 - \alpha_i)P_1^{(i)} + \alpha_i Q_1^{(i)} = (1 - \alpha_i)P_2^{(i)} + \alpha_i Q_2^{(i)}.$$
For each such i, the adversary acts:
  - With probability $1 - \alpha_i$, leave y_i uncorrupted (i.e., drawn from $P_1^{(i)}$ if $f = f_1$, or from $P_2^{(i)}$ if $f = f_2$).
  - With probability α_i, the adversary replaces y_i by a draw from $Q_1^{(i)}$ if the true function is f_1, and from $Q_2^{(i)}$ if the true function is f_2.

For the above adversarial strategy, we have $|\mathcal{A}| \leq r_q n = q$. In addition, note that under model f_1, conditionally on x_i, the corrupted response $\widetilde{y}_i$ is distributed according to $(1 - \alpha_i)P_1^{(i)} + \alpha_i Q_1^{(i)}$, and under model f_2, it is distributed according to $(1 - \alpha_i)P_2^{(i)} + \alpha_i Q_2^{(i)}$. By construction of $Q_1^{(i)}$ and $Q_2^{(i)}$ in Lemma 2, these two mixtures are identical for each i.

Therefore, after adversarial corruption, the distribution of all observed data $\{\widetilde{y}_i\}_{i=1}^n$ is identical under f_1 and f_2. More precisely:

- For all i with $x_i > r_q$, we have $f_1(x_i) = f_2(x_i)$, and hence $P_1^{(i)} = P_2^{(i)}$; no corruption is needed, and the distribution of $\widetilde{y}_i$ is the same under both models.
- For all i with $x_i \le r_q$, the adversary modifies the responses exactly so that the overall conditional distribution of $\widetilde{y}_i$ is matched across the two models.

Note that the constructed functions f_1 and f_2 are not identical: by definition, their difference measured by the metrics introduced in (1) and (2) is nonzero. However, the adversarial corruption strategy described above renders the corrupted data distribution identical under both f_1 and f_2. Consequently, no estimator can achieve better performance than random guessing between the two hypotheses. As a result, the minimax error under adversarial corruption remains bounded away from zero, establishing a nontrivial lower bound.

To prove (18), by starting from the definition of $R_2(f, \hat{f})$, we have

$$R_2(f, \hat{f}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\mathcal{S}} \int_0^1 \left(f(x) - \hat{f}(x) \right)^2 dx \right], \tag{76}$$

where the expectation is over the noise $\boldsymbol{\varepsilon}$, and the supremum is taken over all admissible adversarial strategies $\mathcal{S}$. Since Theorem 2 considers the worst-case function f, we obtain

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega), \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_2(f, \hat{f}) \ge \inf_{\hat{f}} \sup_{f \in \{f_1, f_2\}} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\int_0^1 \left(f(x) - \hat{f}(x) \right)^2 dx \right]. \tag{77}$$

As established earlier, the adversary makes the corrupted data distribution identical under both f_1 and f_2. Formally, let $\mathbb{P}_{f_1}^{(\mathcal{A})}$ and $\mathbb{P}_{f_2}^{(\mathcal{A})}$ denote the distributions over the corrupted datasets when the ground truth is f_1 or f_2, respectively. Thus, we have:

$$\mathbb{P}_{f_1}^{(\mathcal{A})} = \mathbb{P}_{f_2}^{(\mathcal{A})}.$$

That is, the total variation distance satisfies:

$$\mathrm{TV}(\mathbb{P}_{f_1}^{(\mathcal{A})}, \mathbb{P}_{f_2}^{(\mathcal{A})}) = 0. \tag{78}$$

This guarantees that no estimator can distinguish between them better than random guessing. To formalize this, we use Le Cam's two-point method [67, 68] (the hypothesis testing between two points), which states that for any estimator $\hat{f}$ and any pair f_1, f_2,

$$\inf_{\hat{f}} \sup_{f \in \{f_1, f_2\}} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\|\hat{f} - f\|_{L^2(\Omega)}^2 \right] \ge \frac{\|f_1 - f_2\|_{L^2(\Omega)}^2}{4} \cdot \left(1 - \mathrm{TV}(\mathbb{P}_{f_1}^{(\mathcal{A})}, \mathbb{P}_{f_2}^{(\mathcal{A})}) \right).$$

Using (78), we obtain the following lower bound:

$$\inf_{\hat{f}} \sup_{f \in \{f_1, f_2\}} \mathbb{E} \left[\|\hat{f} - f\|_{L^2(\Omega)}^2 \right] \ge \frac{1}{4} \|f_1 - f_2\|_{L^2(\Omega)}^2.$$

Consequently, following (77) we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega), \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_2(f, \hat{f}) \ge \frac{1}{4} \int_0^1 \left(f_1(x) - f_2(x) \right)^2 dx. \tag{79}$$

Recall that $f_1(x) = 0$, and

$$f_2(x) = \begin{cases} r_q - x, & x \in [0, r_q - \varepsilon_q], \\ g(x), & x \in [r_q - \varepsilon_q, r_q], \\ 0, & x > r_q, \end{cases}$$

where $g(x)$ is a degree-5 polynomial satisfying the smoothness and boundary conditions described earlier. Therefore,

$$\begin{aligned} \int_0^1 \left(f_1(x) - f_2(x) \right)^2 dx = \int_0^{r_q} f_2(x)^2 dx &= \int_0^{r_q - \varepsilon_q} (r_q - x)^2 dx + \int_{r_q - \varepsilon_q}^{r_q} g(x)^2 dx \\ &\ge \int_0^{r_q - \varepsilon_q} (r_q - x)^2 dx. \end{aligned} \tag{80}$$

Note that since $\varepsilon_q = r_q^2$, we have

$$\int_0^{r_q-\varepsilon_q} (r_q - x)^2 dx = \int_{\varepsilon_q}^{r_q} u^2 du = \frac{r_q^3 - \varepsilon_q^3}{3} \gtrsim r_q^3 = \left(\frac{q}{n}\right)^3. \tag{81}$$

Therefore, we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega), \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_2(f, \hat{f}) \gtrsim r_q^3 = \left(\frac{q}{n}\right)^3. \tag{82}$$

Moreover, even in the absence of adversarial corruption (i.e., $q = 0$), it is well known from classical minimax theory in nonparametric regression [24] that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega)} \mathbb{E}\left[\left\|f - \hat{f}\right\|_{L_2(\Omega)}^2\right] \gtrsim n^{-4/5}. \tag{83}$$

Combining the two regimes, we obtain the following lower bound on the adversarial error:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega),\, \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_2(f, \hat{f}) \gtrsim \left(\frac{q}{n}\right)^3 + n^{-4/5}. \tag{84}$$

This completes the proof of (18). To complete the proof of Theorem 2, we now establish a lower bound for R_∞. Recall that

$$R_\infty(f, \hat{f}) = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathcal{S}} \left\|f - \hat{f}\right\|_{L_\infty(\Omega)}^2\right], \tag{85}$$

where the expectation is taken over the noise $\boldsymbol{\varepsilon}$, and the supremum is over all adversarial corruption strategies $\mathcal{S}$. The norm $\|\cdot\|_{L_\infty(\Omega)}$ denotes the supremum norm over the interval $[0, 1]$.

As in the case of R_2, the adversary can construct corrupted data distributions under f_1 and f_2 that are indistinguishable. Consequently, no estimator can distinguish between the two hypotheses better than random guessing. Applying Le Cam's two-point method [67, 68] to the L_∞ loss, we obtain:

$$\inf_{\hat{f}} \sup_{f \in \{f_1, f_2\}} \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\|\hat{f} - f\|_{L^\infty(\Omega)}^2\right] \geq \frac{\|f_1 - f_2\|_{L^\infty(\Omega)}^2}{4} \cdot \left(1 - \mathrm{TV}(\mathbb{P}_{f_1}^{(\mathcal{A})}, \mathbb{P}_{f_2}^{(\mathcal{A})})\right).$$

Therefore, we have:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega), \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_\infty(f, \hat{f}) \geq \frac{\|f_1 - f_2\|_{L_\infty(\Omega)}^2}{4}. \tag{86}$$

Since $f_1(x) = 0$, we have $\|f_1 - f_2\|_{L_\infty(\Omega)} = \|f_2\|_{L_\infty(\Omega)} \geq f_2(0)$. From the definition of f_2, we have $f_2(0) = r_q$. Therefore,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega),\, \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_\infty(f, \hat{f}) \gtrsim r_q^2 = \left(\frac{q}{n}\right)^2. \tag{87}$$

Moreover, in the absence of adversarial corruption (i.e., $q = 0$), the standard minimax rate for estimation under the supremum norm is known to satisfy (see [24])

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega)} \mathbb{E}\left[\left\|f - \hat{f}\right\|_{L_\infty(\Omega)}^2\right] \gtrsim \left(\frac{\log n}{n}\right)^{3/4}. \tag{88}$$

Combining both contributions, we conclude that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}^2(\Omega),\, \mathcal{S}, P_{\boldsymbol{\varepsilon}}} R_\infty(f, \hat{f}) \gtrsim \left(\frac{q}{n}\right)^2 + \left(\frac{\log n}{n}\right)^{3/4}. \tag{89}$$

This completes the proof of (19), and thereby the proof of Theorem 2.

C Gaussian Setting Experiments

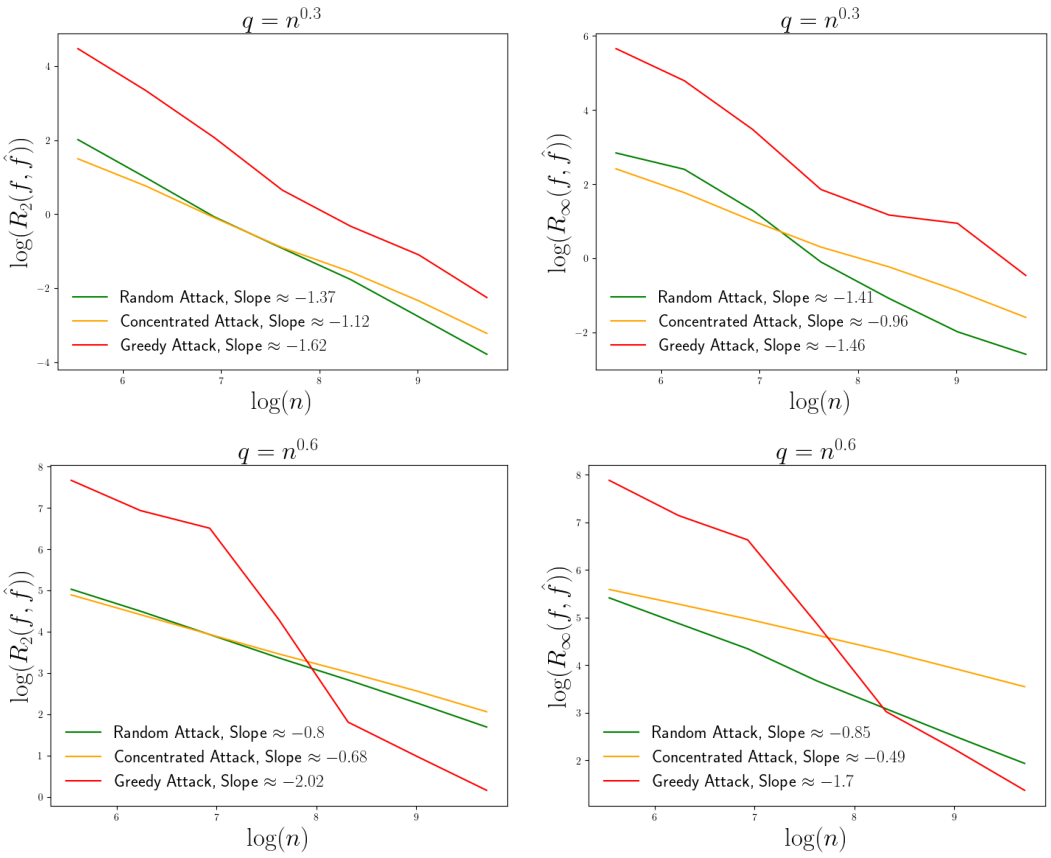


Figure 5: Log-log plots showing the convergence rate of the cubic smoothing spline estimator $\hat{f} = \hat{f}^a_{\text{SS}}$ for $f(x) = x\sin(x)$ under a Gaussian design. The top row plots are results for $q = n^{0.3}$, with theoretical rates of $\mathcal{O}(n^{-0.8})$ for $R_2(f, \hat{f})$ and $\mathcal{O}(n^{-0.6})$ for $R_\infty(f, \hat{f})$. The bottom row corresponds to a higher corruption level, $q = n^{0.6}$, with respective theoretical upper bounds of $\mathcal{O}(n^{-0.53})$ and $\mathcal{O}(n^{-0.48})$.

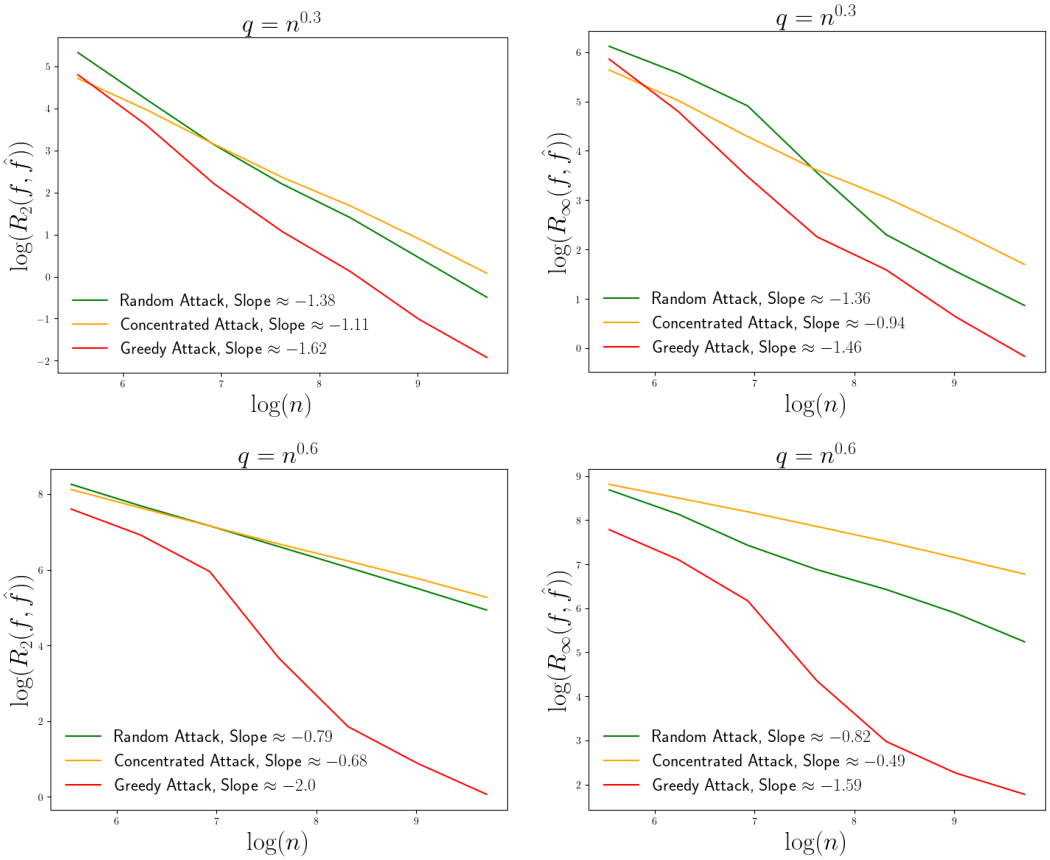


Figure 6: Log–log plots showing the convergence behavior of the cubic smoothing spline estimator $\hat{f} = \hat{f}^a_{\mathrm{SS}}$ when the ground-truth function is an MLP, under the Gaussian design. The top row corresponds to the case $q = n^{0.3}$, with theoretical convergence rates of $\mathcal{O}(n^{-0.8})$ for $R_2(f, \hat{f})$ and $\mathcal{O}(n^{-0.6})$ for $R_\infty(f, \hat{f})$. The bottom row shows results for a higher corruption level, $q = n^{0.6}$, with respective theoretical upper bounds of $\mathcal{O}(n^{-0.53})$ and $\mathcal{O}(n^{-0.48})$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We detailed our contributions clearly in the abstract and the introduction sections of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provided all the details regarding the assumptions, conditions, and limitations of the main problem (Section 2), theorems (Section 3), the experiments (Section 4), as well as the conclusion (Section 6) in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our paper includes theoretical results in Section 3. In our theorems, we clearly mentioned all the required assumptions, and a complete (and correct) proof of them is available in appendices. Please see Section 2 for a full definition of the problem and the notations used in the paper are described in introduction 1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided details regarding our empirical evaluations in Section 4 in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We provided references to all packages that we used in the paper. Regarding the code, we are happy to share it later if required.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided full experimental details in the paper (see Section 4).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The details are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided all the details regarding our experiments in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS code of ethics in our paper.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper is focused on developing a theoretical foundation for adversarial robustness of non-parametric regression. We believe this work has no direct societal impact that should be explained in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is not applicable to our work and this paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve these.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve these.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We just use LLM for writing and editing purposes.

Guidelines: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.