

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}_{SS}^a is given by

$$\hat{f}_{\text{SS}}^a(x) = \frac{1}{n} \sum_{i=1}^n W_n(x, x_i) \tilde{y}_i, \quad (24)$$

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}_{SS} , is given by

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

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

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

By the definition of $R_2(f, \hat{f}_{\text{SS}}^a)$, we have

$$R_2(f, \hat{f}_{\text{SS}}^a) = \mathbb{E}_\epsilon \left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\text{SS}}^a \right\|_{L_2(\Omega)}^2 \right], \quad (26)$$

where $\epsilon = (\epsilon_i)_{i \in [n] \setminus \mathcal{A}}$. First, observe that

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

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

$$\left(f(x) - \hat{f}_{\text{SS}}^a(x) \right)^2 = \left(f(x) - \hat{f}_{\text{SS}}(x) + \hat{f}_{\text{SS}}(x) - \hat{f}_{\text{SS}}^a(x) \right)^2. \quad (27)$$

Using AM-GM inequality, we obtain

$$\left(f(x) - \hat{f}_{\text{SS}}^a(x) \right)^2 \leq 2 \left(f(x) - \hat{f}_{\text{SS}}(x) \right)^2 + 2 \left(\hat{f}_{\text{SS}}(x) - \hat{f}_{\text{SS}}^a(x) \right)^2. \quad (28)$$

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

$$R_2(f, \hat{f}_{\text{SS}}^a) \leq 2 \mathbb{E}_{\hat{\epsilon}} \left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{\text{SS}} \right\|_{L_2(\Omega)}^2 \right] + 2 \mathbb{E}_{\hat{\epsilon}} \left[\sup_{\mathcal{S}} \left\| \hat{f}_{\text{SS}} - \hat{f}_{\text{SS}}^a \right\|_{L_2(\Omega)}^2 \right]. \quad (29)$$

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}_{SS} :

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

³Note that any function $f: [a, b] \rightarrow \mathbb{R}$ can be transformed with scaling and shifting into a function $\tilde{f}: [0, 1] \rightarrow \mathbb{R}$ 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, Theorem 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*

$$\frac{\Delta_{\max}}{\Delta_{\min}} \leq k, \quad (30)$$

for some constant $k > 0$, where

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

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

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

Here, \hat{f}_{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}_{SS} . This yields:

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

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:

$$\mathbb{E}_{\hat{\epsilon}} \left[\sup_S \left\| \hat{f}_{\text{SS}} - \hat{f}_{\text{SS}}^a \right\|_{L_2(\Omega)}^2 \right].$$

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

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

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

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

Note that for each i :

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

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

Thus, the sum above reduces to

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

and we can bound

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

This implies that

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

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

Taking expectations and supremum over \mathcal{S} yields

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

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 $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$. 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

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

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

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

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

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

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

Thus,

$$\begin{aligned} \sup_{x \in \Omega, j \in [n]} |W_n(x, x_j)| &\leq \sup_{x \in \Omega, j \in [n]} \left| \widehat{W}_n(x, x_j) \right| + C \left(\lambda^{-1/2} \alpha(n) + 1 \right) \\ &\stackrel{(a)}{\leq} \frac{\lambda^{-1/4}}{2} (p_{\min})^{-3/4} + C \left(\lambda^{-1/2} \alpha(n) + 1 \right) \end{aligned} \quad (47)$$

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 $R_2(f, \hat{f}_{\text{SS}}^a)$ stated in Theorem 1.

$$\begin{aligned} R_2(f, \hat{f}_{\text{SS}}^a) &\leq 2P_0 \lambda \int_{\Omega} (f''(x))^2 dx + \frac{2Q_0 \sigma^2}{n \lambda^{1/4}} \\ &\quad + \frac{2q^2(8M^2 + 2\sigma^2)}{n^2} \left[\frac{\lambda^{-1/4}}{2} (p_{\min})^{-3/4} + C \left(\lambda^{-1/2} \alpha(n) + 1 \right) \right]^2. \end{aligned} \quad (48)$$

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

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

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

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

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 $\lambda^{-1/4} \alpha(n) \rightarrow 0$ as $n \rightarrow \infty$. Therefore, for sufficiently large n , we have $1 + \lambda^{-1/4} \alpha(n) < 2$. As a result, we obtain

$$R_2(f, \hat{f}_{\text{SS}}^a) \lesssim \lambda \int_{\Omega} (f''(x))^2 dx + \frac{\sigma^2}{n \lambda^{1/4}} + \frac{q^2 (M^2 + \sigma^2)}{n^2 \lambda^{1/2}}. \quad (51)$$

This concludes the proof of the upper bound on $R_2(f, \hat{f}_{\text{SS}}^a)$ 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

$$\left\| f - \hat{f}_{\text{SS}}^a \right\|_{L_{\infty}(\Omega)}^2 \leq 2 \left\| f - \hat{f}_{\text{SS}} \right\|_{L_{\infty}(\Omega)}^2 + 2 \left\| \hat{f}_{\text{SS}} - \hat{f}_{\text{SS}}^a \right\|_{L_{\infty}(\Omega)}^2. \quad (52)$$

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

$$\begin{aligned} R_\infty(f, \hat{f}_{SS}^a) &= \mathbb{E}_{\hat{\epsilon}} \left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{SS}^a \right\|_{L_\infty(\Omega)}^2 \right] \\ &\leq 2 \mathbb{E}_{\hat{\epsilon}} \left[\sup_{\mathcal{S}} \left\| f - \hat{f}_{SS} \right\|_{L_\infty(\Omega)}^2 \right] + 2 \mathbb{E}_{\hat{\epsilon}} \left[\sup_{\mathcal{S}} \left\| \hat{f}_{SS} - \hat{f}_{SS}^a \right\|_{L_\infty(\Omega)}^2 \right]. \end{aligned} \quad (53)$$

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

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

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

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

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

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

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

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

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:

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

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

$$\begin{aligned} \mathbb{E}_{\hat{\epsilon}} \left[\left\| f - \hat{f}_{SS} \right\|_{L_\infty(\Omega)}^2 \right] &\leq 2 \mathbb{E}_{\hat{\epsilon}} \left[\left\| f - \hat{f}_{SS} \right\|_{L_2(\Omega)} \cdot \left\| f' - \hat{f}'_{SS} \right\|_{L_2(\Omega)} \right] \\ &\leq 2 \left(\mathbb{E}_{\hat{\epsilon}} \left[\left\| f - \hat{f}_{SS} \right\|_{L_2(\Omega)}^2 \right] \right)^{1/2} \cdot \left(\mathbb{E}_{\hat{\epsilon}} \left[\left\| f' - \hat{f}'_{SS} \right\|_{L_2(\Omega)}^2 \right] \right)^{1/2}. \end{aligned} \quad (58)$$

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

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

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

Substituting the bounds from (59) and (60) into (58), we obtain

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

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

$$\begin{aligned} R_\infty(f, \hat{f}_{SS}^a) &\lesssim \left(\lambda \int_{\Omega} (f''(x))^2 dx + \frac{\sigma^2}{n \lambda^{1/4}} \right)^{1/2} \times \left(\lambda^{1/2} \int_{\Omega} (f''(x))^2 dx + \frac{\sigma^2}{n \lambda^{3/4}} \right)^{1/2} \\ &\quad + \frac{q^2(M^2 + \sigma^2)}{n^2 \lambda^{1/2}} \left(1 + \lambda^{-1/4} \alpha(n) + \lambda^{1/4} \right)^2. \end{aligned} \quad (62)$$

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

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

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

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

This completes the proof of the upper bound on $R_\infty(f, \hat{f}_{\text{SS}}^a)$ 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*

$$(1 - \alpha)P_1 + \alpha Q_1 = (1 - \alpha)P_2 + \alpha Q_2, \quad (64)$$

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

Proof. Define α as:

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

Next, define Q_1 and Q_2 as:

$$Q_1(u) = \frac{1 - \alpha}{\alpha} (P_2(u) - P_1(u)) \mathbf{1}_{\{P_2(u) \geq P_1(u)\}}, \quad (66)$$

$$Q_2(u) = \frac{1 - \alpha}{\alpha} (P_1(u) - P_2(u)) \mathbf{1}_{\{P_1(u) > P_2(u)\}}, \quad (67)$$

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:

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

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:

$$(1 - \alpha)P_1 + \alpha Q_1 = (1 - \alpha)P_2 + \alpha Q_2. \quad (69)$$

This completes the proof. \square

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 \tilde{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

$$P_1^{(i)} := \mathcal{N}(f_1(x_i), \sigma^2), \quad P_2^{(i)} := \mathcal{N}(f_2(x_i), \sigma^2),$$

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 \tilde{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 $\{\tilde{y}_i\}_{i=1}^n$ 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 \tilde{y}_i is the same under both models.
- For all i with $x_i \leq r_q$, the adversary modifies the responses exactly so that the overall conditional distribution of \tilde{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

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

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

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

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:

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

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 ,

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

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

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

Consequently, following (77) we have

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

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,

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

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

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

Therefore, we have

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

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

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

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

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

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

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

where the expectation is taken over the noise ε , 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:

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

Therefore, we have:

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

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,

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

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])

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

Combining both contributions, we conclude that

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

This completes the proof of [19], and thereby the proof of Theorem 2.

C Gaussian Setting Experiments

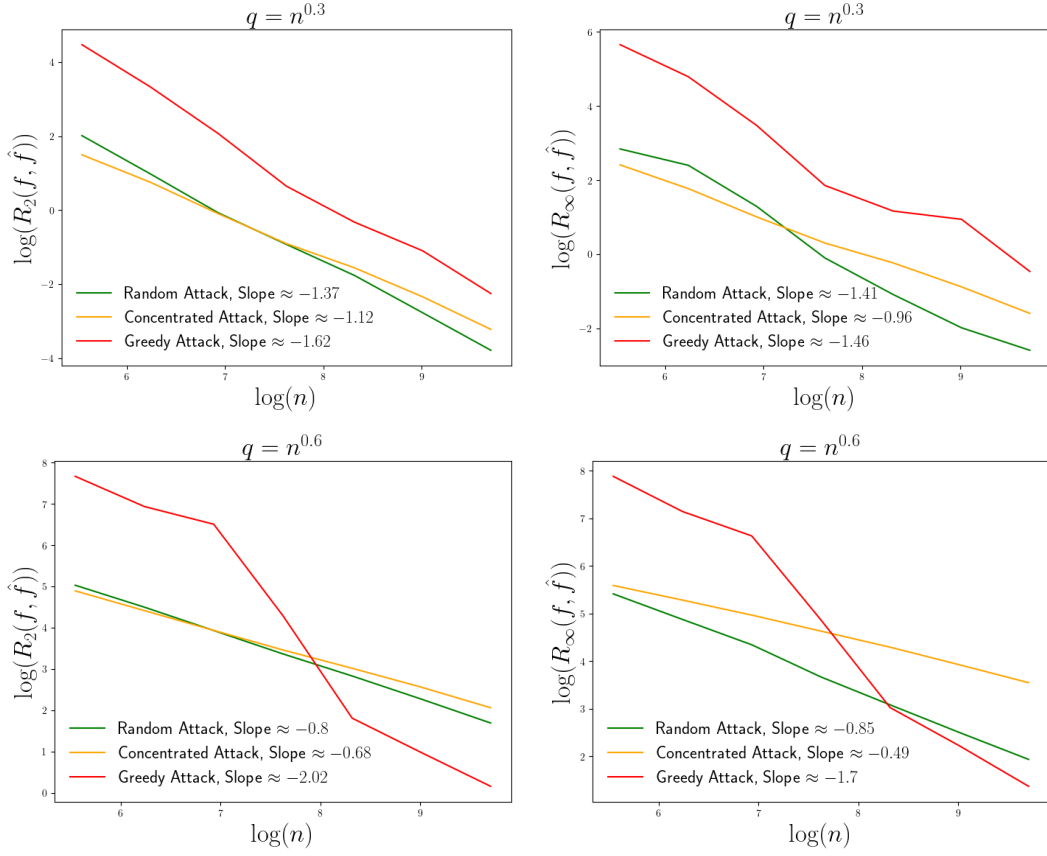


Figure 5: Log-log plots showing the convergence rate of the cubic smoothing spline estimator $\hat{f} = \hat{f}_{\text{SS}}^a$ for $f(x) = x \sin(x)$ under a Gaussian design. The top row plots are results for $q = n^{0.3}$, with theoretical rates of $\mathcal{O}(n^{-0.8})$ for $R_2(f, \hat{f})$ and $\mathcal{O}(n^{-0.6})$ 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 $\mathcal{O}(n^{-0.48})$.

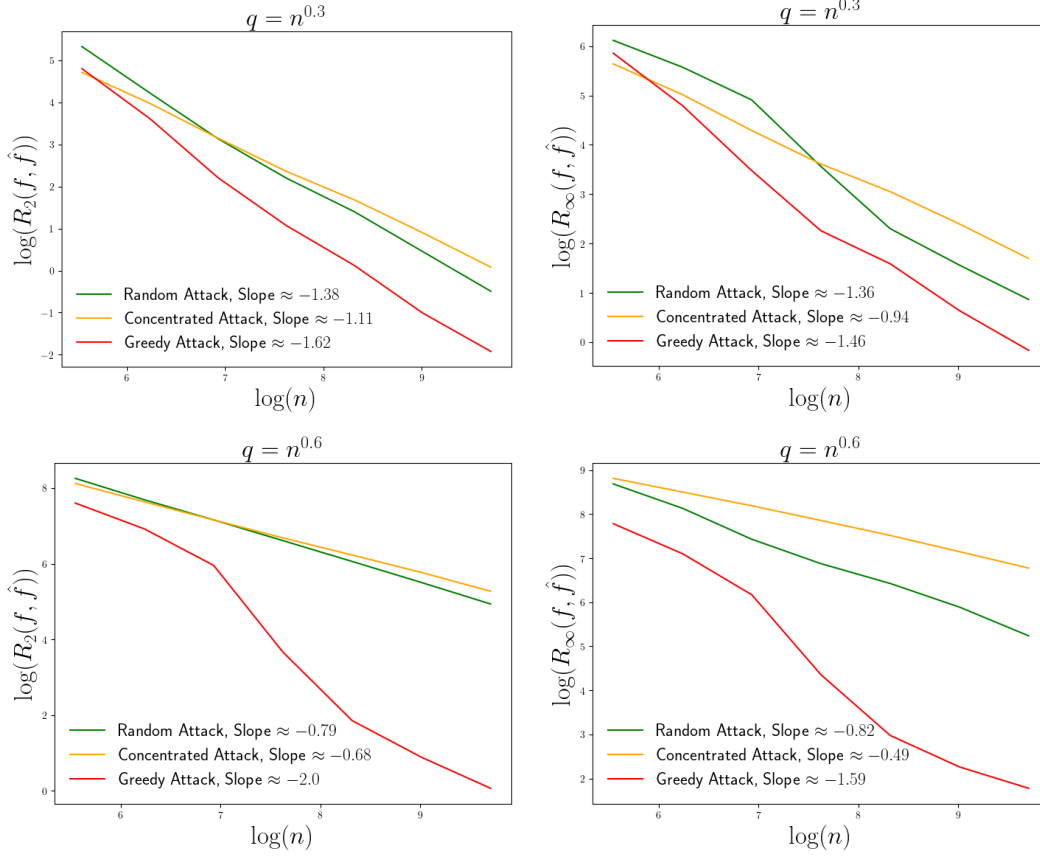


Figure 6: Log-log plots showing the convergence behavior of the cubic smoothing spline estimator $\hat{f} = \hat{f}_{SS}^a$ when the ground-truth function is an MLP, under the Gaussian design. The top row corresponds to the case $q = n^{0.3}$, with theoretical convergence rates of $\mathcal{O}(n^{-0.8})$ 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 $\mathcal{O}(n^{-0.48})$.