Adversarial Robustness of Nonparametric Regression

Parsa Moradi, Hanzaleh Akbarinodehi, Mohammad Ali Maddah-Ali

Motivation

- Robustness in **parametric regression** has been extensively studied in the literature.
- The adversarial robustness of **nonparametric regression** remains **largely unexplored**.
- We study adversarial robustness in nonparametric regression settings when the target regression function is in a **second-order Sobolev space**

Problem Setting

- Given dataset of $\{(x_i, \widetilde{y}_i)\}_{i=1}^n$ with $x_i, \tilde{y}_i \in \mathbb{R}$
- Fixed design: x_i are deterministic and fixed
- $f \in \mathcal{W}^2(\Omega)$ second-order Sobolev space over $\Omega \subset \mathbb{R}$
- Labels are adversarially corrupted

Square integrable over Ω up to second derivative

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

Set of adversarially corrupted sample indices

i.i.d noise with zero mean and variance at most σ^2

- $\mathcal{A}$ is unkown and $|\mathcal{A}| \leq q$
- Any nonparametric regression model produces $\hat{f}(\cdot)$ as an estimate of $f(\cdot)$

Metrics

- Measure estimation error over all possible adversarial strategies.

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

Adversary Strategy

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

Goal: Characterize $\inf_{\hat{f}} R_2(f, \hat{f})$ and $\inf_{\hat{f}} R_\infty(f, \hat{f})$

Main Result: Upper Bound
Smoothing Splines are Adversarially Robust

- Second-order smoothing spline estimator:

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

- Assumptions:
  1. Bounded function and adversarial corruption

$$|f(x)| \leq m_1 \qquad\qquad |\widetilde{y}_i| \leq m_2, \text{ for } i \in \mathcal{A}$$

  2. Emperical cumlative distribution F_n of design points uniformly converges to $F(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\}, \quad \sup_{x \in \Omega} |F_n(x) - F(x)| \to 0 \quad \text{as} \quad n \to \infty$$

Main Result: Upper Bound
Smoothing Splines are Adversarially Robust

- Second-order smoothing spline estimator:

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

Theorem (Upper Bound): Assume that $\lambda \to 0$ as $n \to \infty$ and $\lambda > n^{-2}$. Let $M = \max\{m_1, m_2\}$. Then, for sufficiently large n:

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

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

Main Result: Upper Bound
Smoothing Splines are Adversarially Robust

Theorem (Upper Bound): *Assume that $\lambda \to 0$ as $n \to \infty$ and $\lambda > n^{-2}$. Let $M = \max\{m_1, m_2\}$. Then, for sufficiently large n:*

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

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

- Scenario:

Optimal λ

$$q = \Theta(n^\beta) \begin{cases} R_2(f, \hat{f}^a_{\mathrm{SS}}) \leq \begin{cases} \mathcal{O}\left(n^{-4/5}\right) & \text{for } \beta \leq 0.4, \quad \lambda^* = \mathcal{O}(n^{-0.8}) \\ \mathcal{O}\left(n^{-4/3(1-\beta)}\right) & \text{for } \beta > 0.4, \quad \lambda^* = \mathcal{O}(n^{-4/3(1-\beta)}) \end{cases} \\ R_\infty(f, \hat{f}^a_{\mathrm{SS}}) \leq \begin{cases} \mathcal{O}\left(n^{-3/5}\right) & \text{for } \beta \leq 0.5, \quad \lambda^* = \mathcal{O}(n^{-0.8}) \\ \mathcal{O}\left(n^{-6/5(1-\beta)}\right) & \text{for } \beta > 0.5, \quad \lambda^* = \mathcal{O}(n^{-8/5(1-\beta)}) \end{cases} \end{cases}$$

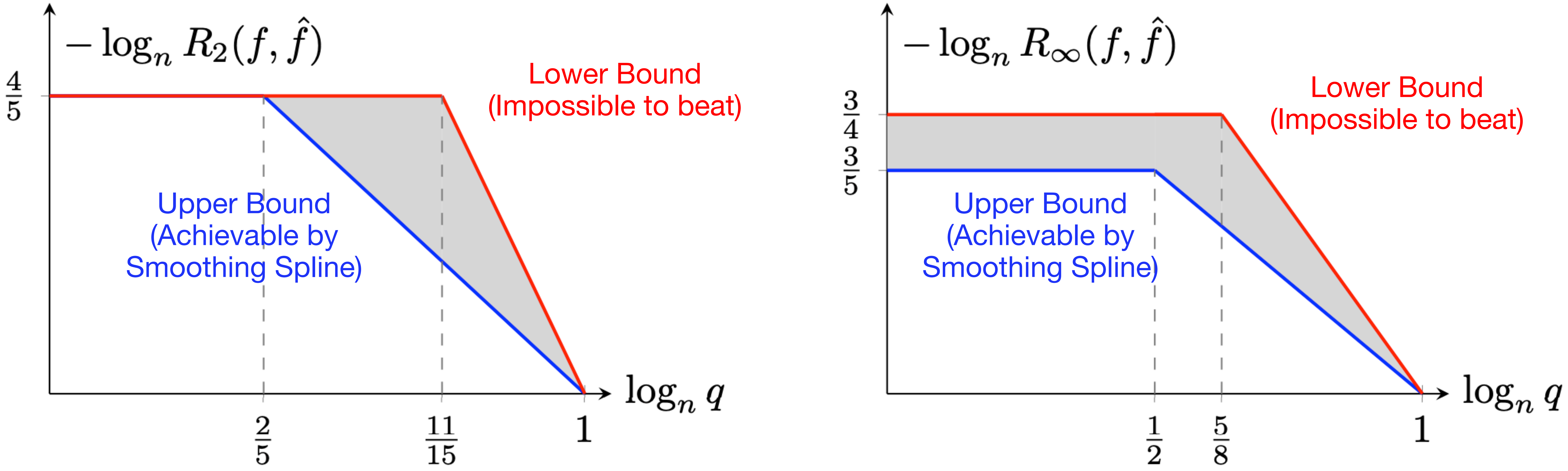
Main Result: Minimax Lower Bound

Theorem (Minimax Lower Bound): *Let $P_{\boldsymbol{\varepsilon}}$ denote the probability density function of the noise vector $\boldsymbol{\varepsilon}$, with i.i.d zero-mean and bounded variance. Then*

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

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

Convergence Rate Region



- **Optimality of convergence region:**
  1. Errors for smoothing spline converge to zero as long as $q = o(n)$
  2. When $q = \mu n$, no estimator can achieve vanishing error for any second-order Sobolev function
- **Optimality of convergence rate:** Smoothing Spline is **minimax-optimal** for R_2 when $\log_n(q) \leq 0.4$

Experiments

- Target functions: $x\sin(x)$ and 3-layer MLP.
- Adversarial attack strategies:

1. **Random Attack**: Randomly replaces the responses of q out of n samples with M.
2. **Greedy Attack**: Start from clean samples. Iteratively identifies the sample most aligned with the current estimator and change its label with M; repeats until q samples are corrupted.
3. **Concentrated Attack**: Corrupts q consecutive samples to M.

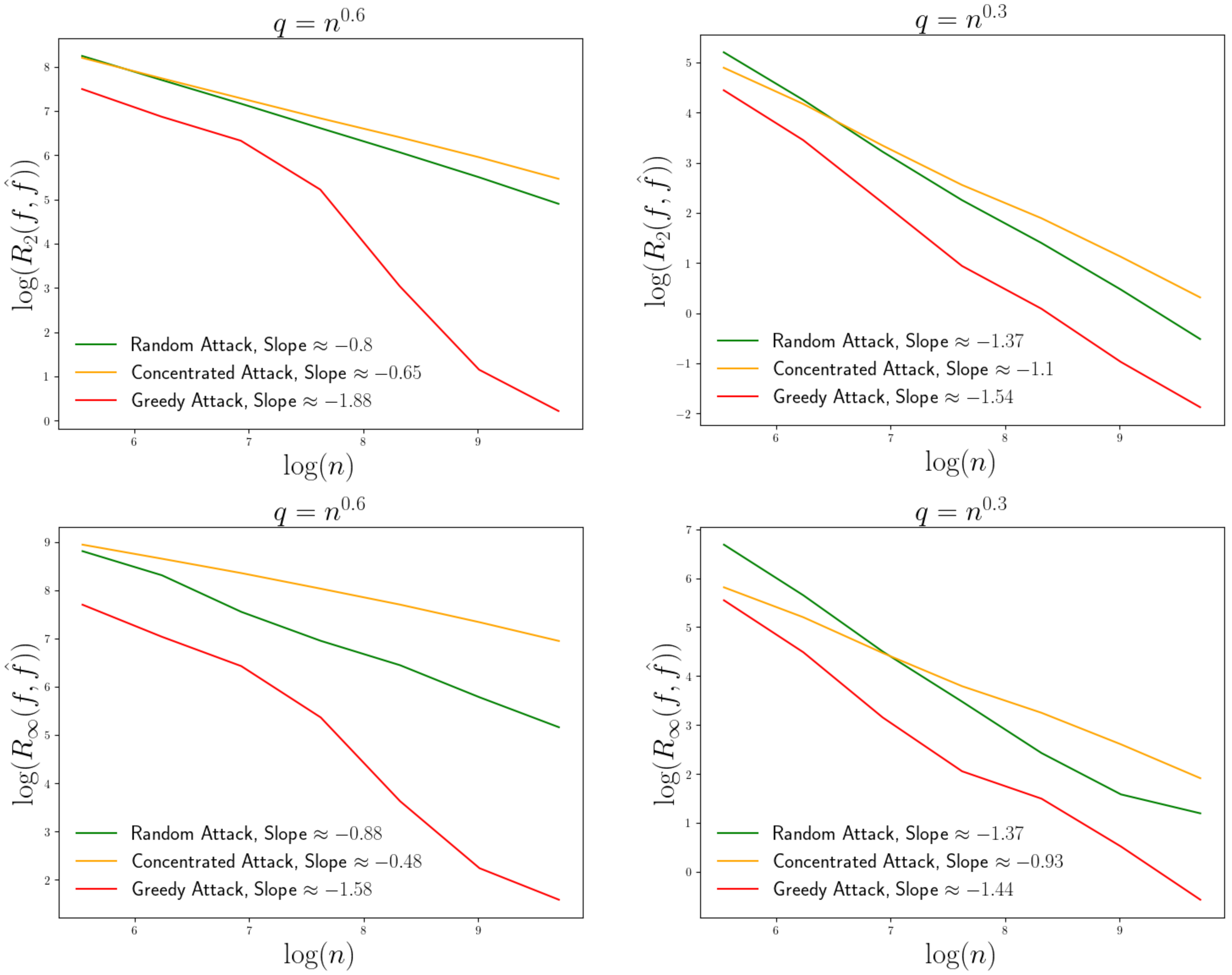
Experiments

Target functions: 3-layer MLP.

Theoritical upper bound rates:

$$R_\infty(f, \hat{f}^a_{\text{SS}}) \leq \begin{cases} \mathcal{O}\left(n^{-0.6}\right) & \text{for } q = n^{0.3}, \\ \mathcal{O}\left(n^{-0.48}\right) & \text{for } q = n^{0.6}, \end{cases}$$

$$R_2(f, \hat{f}^a_{\text{SS}}) \leq \begin{cases} \mathcal{O}\left(n^{-0.8}\right) & \text{for } q = n^{0.3}, \\ \mathcal{O}\left(n^{-0.53}\right) & \text{for } q = n^{0.6}, \end{cases}$$



Questions or Comments: Email moradi@umn.edu

Paper

Thank you 🙏