THE UNIVERSITY OF MELBOURNE

Certified Adversarial Robustness via Randomized Alpha-Smoothing for Regression Models

NeurIPS 2024

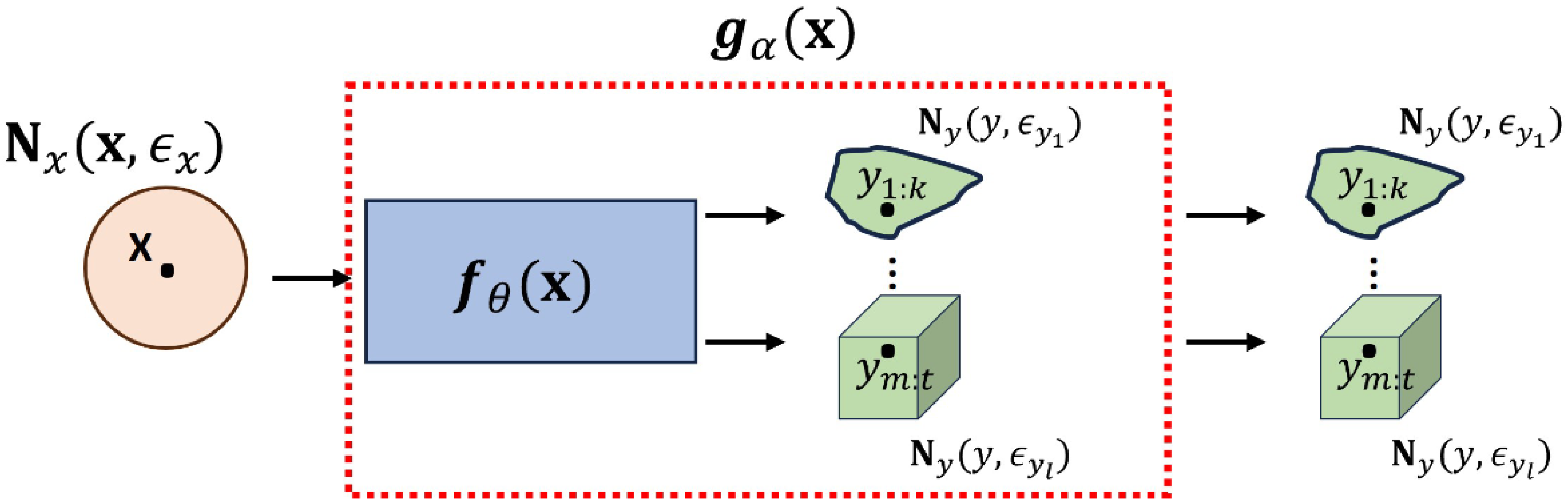
Aref Miri Rekavandi, Farhad Farokhi, Olga Ohrimenko, Benjamin Rubinstein

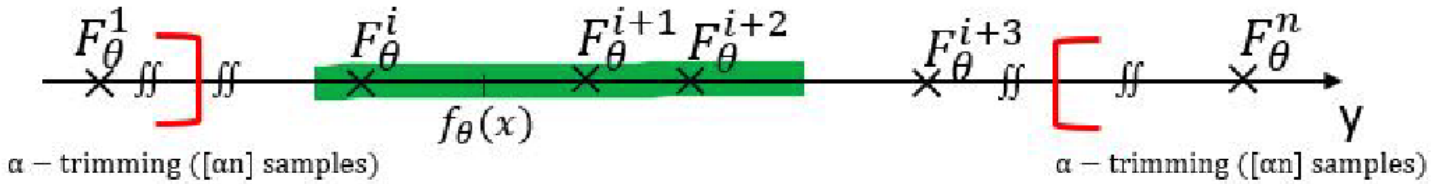
University of Melbourne, Australia

Motivation

- Randomized Smoothing (RS) has shown promising results in the certification of predictions in Large-scale classification models.
- Only bounded output regression models have been certified so far using RS with restrictive constraints.
- For the first time, a universal certification approach is developed for a broad class of regression models (bounded/unbounded outputs) and the certification is valid for both small and large sample regimes.

Overall Structure





Main Results

Theorem 3. *(Certification of $\boldsymbol{g}_\alpha(\boldsymbol{x})$ against ℓ_p Attack). Let $\boldsymbol{f}_\theta(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^t$ be a deterministic or random base regressor and let $n \geq 1$, $0 \leq \alpha < 1/2$, $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and suppose the α-trimming function $\boldsymbol{g}_\alpha(\boldsymbol{x})$ defined in (8) is used for smoothing. Then given*

$$\mathbb{P}\{diss_y(\boldsymbol{f}_\theta(\boldsymbol{x}+\boldsymbol{e})_i, \boldsymbol{y}_i) \leq \epsilon_{y_i}\} \geq \underline{p_{A}}_i, \forall i \in [\![t]\!] \tag{13}$$

where $\underline{p_{A}}_i$ is the lower bound on the probability of accepting prediction in the i^{th} output variable, then $\boldsymbol{g}_\alpha(\boldsymbol{x}+\boldsymbol{\delta})$, $\forall \|\boldsymbol{\delta}\|_p \leq \epsilon_x$ $(p \geq 2)$ is within accepted region, i.e., $\mathbf{N}_{\boldsymbol{y}}(\boldsymbol{y}, \epsilon_y) = \prod_{i=1}^t \mathbf{N}_y(\boldsymbol{y}_i, \epsilon_{y_i})$, with the user-defined probability P, s.t. $I_{n,\alpha}^{-1}(P) \leq \underline{p_{A}}_i, \forall i \in [\![t]\!]$, where

$$\epsilon_x = \min_{i \in [\![t]\!]} \frac{\sigma}{d^{\frac{1}{2}-\frac{1}{p}}} \left(\Phi^{-1}(\underline{p_{A}}_i) - \Phi^{-1}(I_{n,\alpha}^{-1}(P))\right), \tag{14}$$

and where $I_{n,\alpha}^{-1}(x)$ is the inverse of the regularized beta function w.r.t Bernoulli success rate parameter.

Experimental Results

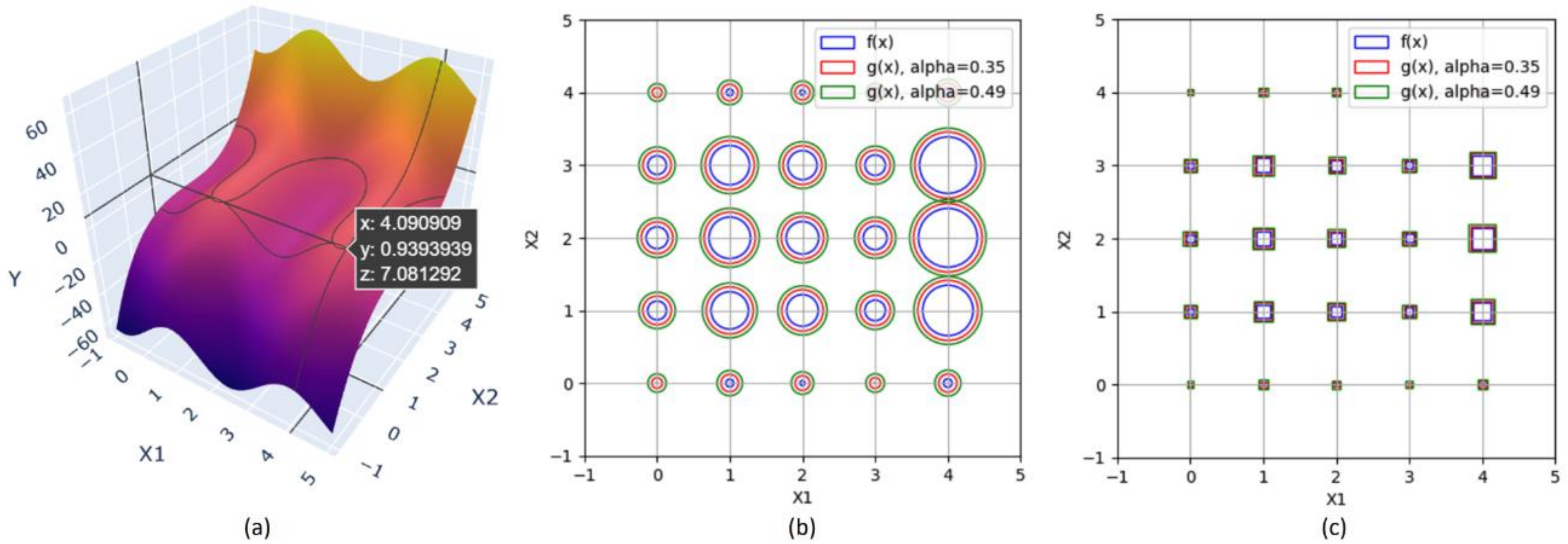


Figure 3: Adopted regression function (a) with the estimated certified radii (against ℓ_2 and ℓ_∞ attacks) for evaluated points in the center for both base and smoothed outputs (b & c).

Thank You!