How Far Can Small Input Changes Move a Logit?

Due: Wednesday, November 25 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

A small change to each input entry may flip a classifier's prediction. We will compare deliberately chosen changes with random changes of the same size.

A simplified classifier convolves its input and flattens the feature map:

$$\mathbf{h} = \mathbf{C}\mathbf{X}, \qquad \mathbf{X} \in \mathbb{R}^n, \quad \mathbf{C} \in \mathbb{R}^{m \times n}, \quad \mathbf{h} \in \mathbb{R}^m.$$

A final linear layer, with $\mathbf{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$, produces the logit

$$s(\mathbf{X}) = \mathbf{w}^\top \mathbf{C}\mathbf{X} + b.$$

There is no activation between these layers. The prediction is $+1$ when $s(\mathbf{X}) > 0$ and -1 otherwise. Fix an input with $s(\mathbf{X}) > 0$ and let $\epsilon > 0$. An adversary replaces $\mathbf{X}$ by $\mathbf{X} + \boldsymbol{\delta}$.

1. With a Euclidean budget $\|\boldsymbol{\delta}\|_2 \le \epsilon$, Cauchy–Schwarz gives a maximum logit decrease of $\epsilon\|\mathbf{C}^\top \mathbf{w}\|_2$. What is the largest decrease $s(\mathbf{X}) - s(\mathbf{X} + \boldsymbol{\delta})$ if each input entry can change by at most ϵ?

2. Now suppose the changes $\delta_1, \dots, \delta_n$ are independent, each equally likely to be $+\epsilon$ or $-\epsilon$. What is $\mathbb{E}[(s(\mathbf{X} + \boldsymbol{\delta}) - s(\mathbf{X}))^2]$?

3. For even n, a stride-two kernel $(1, 1)$ produces $h_i = X_{2i-1} + X_{2i}$ for $i = 1, \dots, n/2$. If $s(\mathbf{X}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n/2} h_i + b = 1$ and $\epsilon = 2/\sqrt{n}$, how large is the worst per-entry decrease relative to the root mean square random change?

Takeaway: *In the example, an optimal adversarial perturbation and every random perturbation have the same ℓ_2 and ℓ_∞ norms. Their effects differ because adversarial changes reinforce one another, while random changes tend to cancel.*