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.

As we saw in the demo, a small, deliberately chosen change to each input entry may flip a classifier's prediction. In this problem, we will compare deliberately chosen changes with random changes of the same size.

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

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

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

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

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. Suppose each input entry can change by at most ϵ, so $\|\boldsymbol{\delta}\|_\infty = \max_{1 \le j \le n} |\delta_j| \le \epsilon$. What is the largest possible absolute change $\left|s(\mathbf{X} + \boldsymbol{\delta}) - s(\mathbf{X})\right|$ in the logit?

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

3. Let's compare adversarial and random changes for a specific convolutional layer. For even n, a stride-two kernel $(1, 1)$ produces $h_i = X_{2i-1} + X_{2i}$ for $i = 1, \ldots, n/2$. If $s(\mathbf{X}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n/2} h_i = 1$ 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, both perturbations change every input entry by the same amount in absolute value. Their effects differ because adversarial changes reinforce one another, while random changes tend to cancel.*