Investigating the Robustness of Sequence Models in Deep Learning

Sanjif Shanmugavelu, Department of Physics, University of Warwick, CV4 7AL, UK.

I) INTRODUCTION

Figure 1: What do you see? [1]

- Machine Learning (ML) classifiers repeatedly classify this as a 60-mph speed limit.
- *Adversarial perturbations* may fool classifiers into making false predictions.
- Easily fooled ML classifiers are **not** robust.
- The robustness of ML sequence models has **not** been extensively studied.

II) CONTRIBUTIONS

1. Formalise a framework for studying robustness for sequence models.
2. Prove classifiers with small decision boundary curvature are more robust.
3. Identify the probability of misclassification on the ball of radius ρ around a datapoint x.
4. Propose an efficient Hessian regulariser to improve robustness of sequence models.

III) PRELIMINARY THEORY

A multi-class classifier is a function $f : R^d \to R^L$.

The class which f predicts for $x \in R^d$ is $\hat{k}(x) = \mathrm{argmax}_k\, f_k(x)$.

A Recurrent Neural Network is the simplest sequence model.

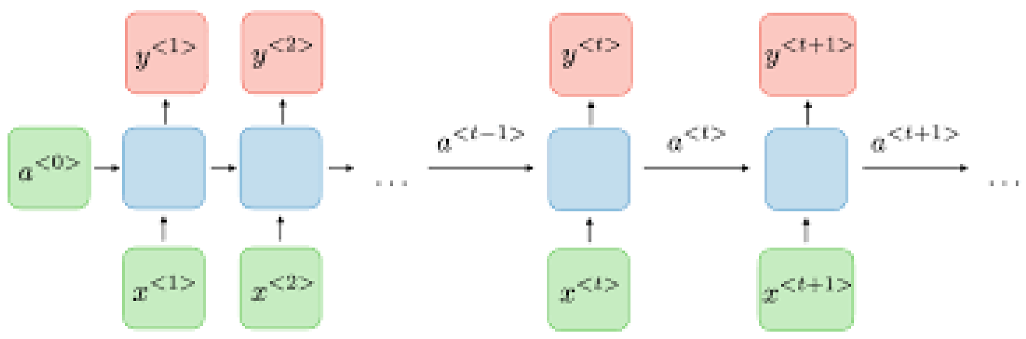

Figure 2 : RNN Architecure. [2]

The classifier f splits the input space into L regions.

The decision boundary, B is the set of points f is *equally likely* to classify into **two** distinct classes.

$$B = \left\{ x \in R^d \;\middle|\; F(x) = f_{\hat{k}(x)}(x) - \max_{k \neq \hat{k}(x)} f_k(x) = 0 \right\}$$

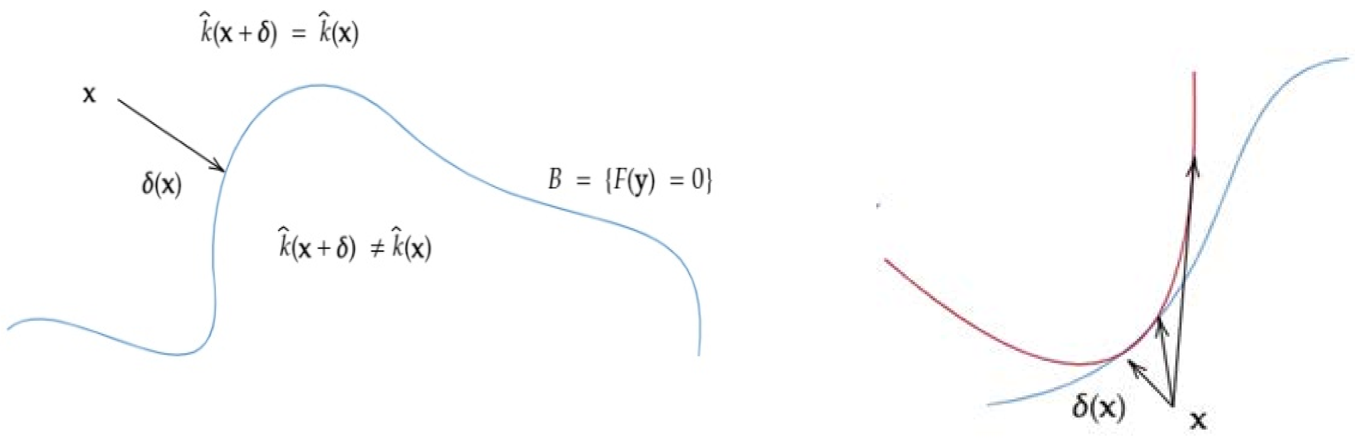

Figure 3: The decision boundary (blue) and second order approximation (red).

The adversarial perturbation $\delta(x; f)$ of $x \in R^d$ is the **minimal** perturbation to B.

1. DECISION BOUNDARY CURVATURE AND ROBUSTNESS

Relax the decision boundary requirement and use the second order approximation to find a bound for perturbation distance:

Let $F(x) = t \geq 0$. Denote $J = \nabla F(x)$ and assume $\nu := \lambda_{max}(H) \geq 0$ where $Hu = \nu u$. Then,

$$\frac{|J|}{\nu}\left(\sqrt{1 + \frac{2\nu t}{|J|^2}} - 1\right) \leq |\delta| \leq \frac{|J^T u|}{\nu}\left(\sqrt{1 + \frac{2\nu t}{(J^T u)^2}} - 1\right)$$

Classifiers with small curvature ($|\boldsymbol{J}| and \lambda_{max}(\boldsymbol{H})$) have larger robustness $|\delta|$.

2. PROBABILISTIC BOUNDS ON ADVERSARIAL DISTANCE

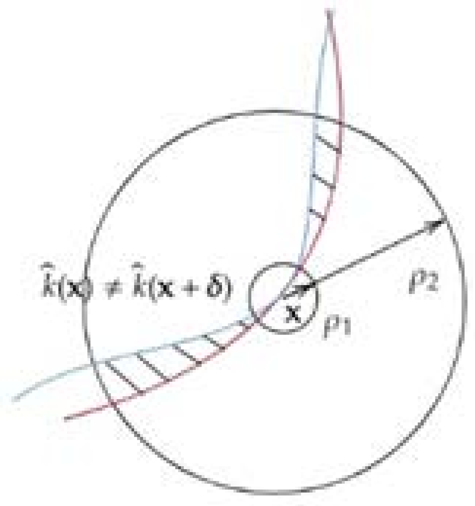

Figure 4: Second order approximation assumption.

Assume for small perturbations, a negative second order Taylor approximation of F implies misclassification.

Assume H is a Wigner matrix with independent eigenvalues. Then, we can find an explicit value for β such that given any curvature κ > 0,

$$P\_\{v \in R^d\} \left(\forall u \in R^2, u^T H^{span(\delta(x),v)} u \geq \kappa |u|_2^2\right) \geq 1 - \beta$$

We arrive at a probabilistic bound on misclassification as a function of perturbation radius

$$P\_\{v \sim \rho S\} \left(\hat{k}(x+v) \neq \hat{k}(x)\right) \leq 2\, exp\left(-\frac{d(\kappa\rho^2 - 1)^2}{2\rho^2(1-2\kappa)^2}\right) + \beta$$

For a given probability of misclassification, we can find an upper bound on perturbation radius ρ . Small misclassification probability requires small ρ.

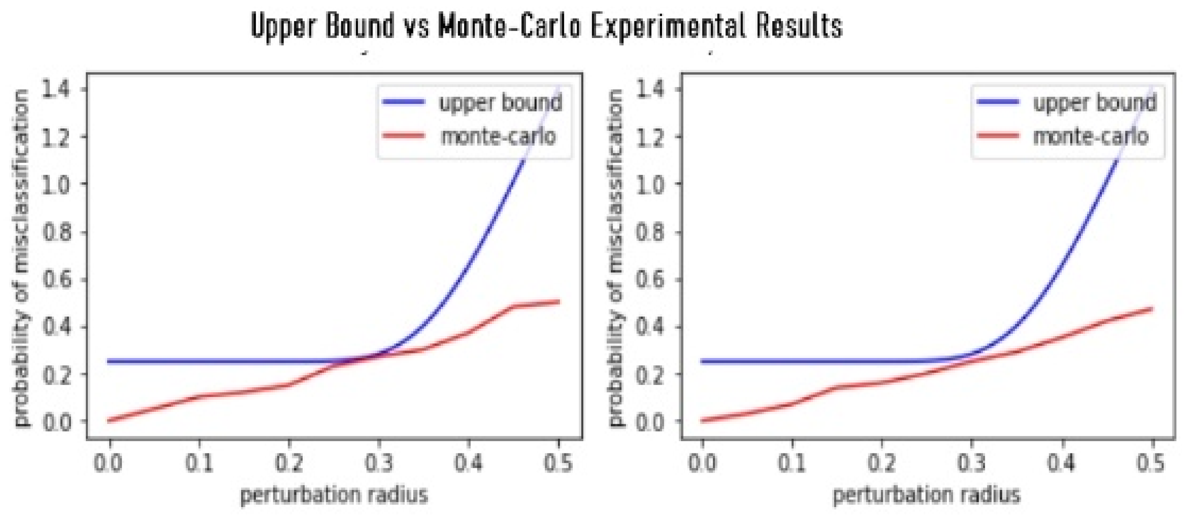

Figure 5: Experimental validation of probabilistic bound on a quadradic classifier.

3. CURVATURE REGULARISATION

We use the Hessian Frobenius norm, $|\boldsymbol{H}(\boldsymbol{x})|_F^2$ as a regulariser to penalize large curvature.

$$|H(x)|_F^2 = \sum_{\{e\}} \left[\frac{\partial^2 (e \cdot f)}{\partial x^2}\right]^T \left[\frac{\partial^2 (e \cdot f)}{\partial x^2}\right]$$

Where $\{e\}$ is an orthonormal basis. We express the Frobenius norm as above to exploit automatic differentiation in ML libraries.

Consider a two-layer MNIST RNN model. The MNIST dataset consists of 60000 labelled handwritten digits ranging from 0 to 9. Each image is 28 x 28 pixels which we treat as a sequence of data.

Figure 6: MNIST [3]

Projected Gradient Descent (PGD) attack is performed for 30 epochs with perturbation size $\epsilon \in [1,30]$, step size $\alpha = 2$.

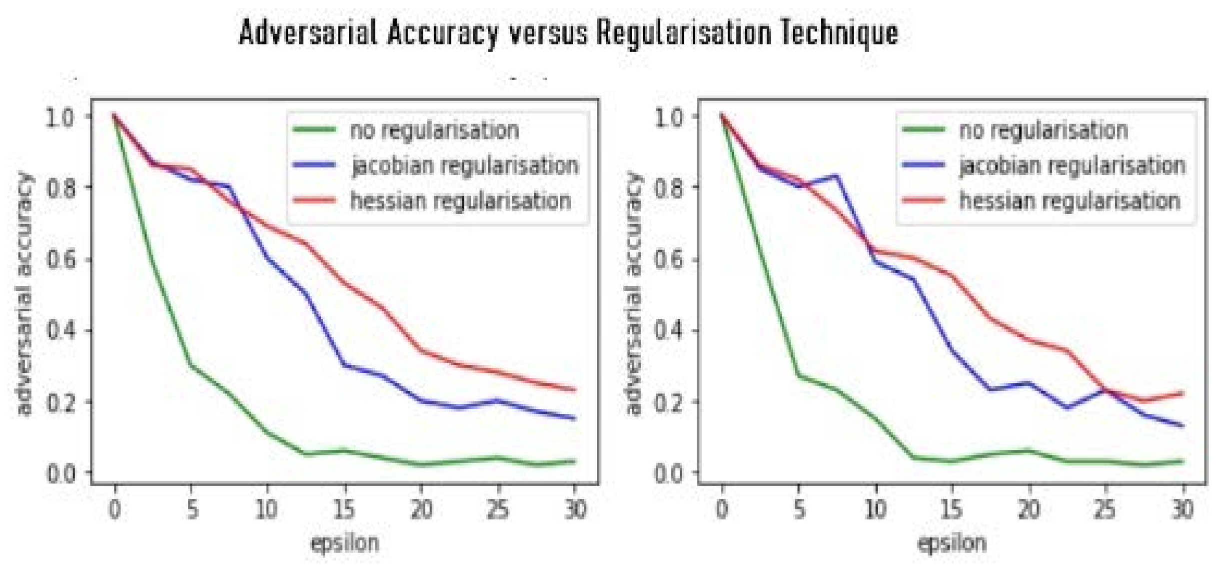

Figure 7: Hessian regularisation outperforms regular training and Jacobian regularisation

REFERENCES

[1] K. Eykholt, I. Evtimov, Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR (2018).

[2] A Amidi, S Amidi RNN Cheatsheet https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

[3] Bottou L. Bengio Y. LeCun, Y. and P. Haffner. The MNIST dataset of handwritten digits (images) (1998).

IV) CONCLUSION AND FUTURE WORK

1. Robustness of sequence models increases as decision boundary curvature decreases.
2. We propose an efficient Hessian regularisation algorithm. We hope to see this applied to sequential data (not just image MNIST).