Investigating the Robustness of Sequence Models in Deep Learning - Interim Report

Sanjif Shanmugavelu

Department of Physics, University of Warwick, CV4 7AL, Coventry, UK

1 Introduction

The accuracy of Machine Learning (ML) models have significantly improved in recent years with the increased volume of data & the development of powerful yet economic computational systems. These models can recognise complex patterns underlying a data set, make accurate predictions & generate further examples of existing data. Applications of ML models range a wide breadth of fields with examples including recommendation systems for video streaming services and road-ready self-driving cars.

While the accuracy of ML models have improved for a variety of applications, the reliability of these models have been put into question. A classifier that predicts the correct label of a datapoint with high confidence can be fooled into making a wrong classification with high confidence too [8, 7]. Corruptions of a data set imperceptible to the human eye have been shown to repeatedly fool state-of-the-art image classifiers [7]. In addition, [8] have shown the existence of universal adversarial perturbations which result in misclassification when applied to every datapoint. This significant shortcoming casts doubt on the practicality of applying machine learning models to safety-critical systems. The consequence of a false negative in a medical diagnostic tool or a false prediction in a self-driving system could possibly result in the loss of a life.

The robustness, or resilience of ML models to adversarial attack has been extensively studied in the context of image classification systems. This project aims to extend this work to sequence models trained on temporally variable or highly ordered data. Widely used deep learning sequence models include recurrent neural networks (RNNs), Long Short Term Memory networks (LSTMs) & transformers. Applications of these models include making predictions of trends in financial time series data and predicting exoplanets with light curve data from satellite missions.

The aims of this project include:

- Determining the robustness of sequence models in a deterministic and probabilistic setting, investigating upper and lower bounds for adversarial robustness in the de-

terministic case & identifying the probability of misclassification on the sphere of radius ρ for the probabilistic case.

- Characterising the geometric properties of the decision boundary & loss surface of robust sequence models. It has been shown that decision boundaries which are less curved result in more robust models for image classification systems [8]. We investigate if this holds for sequence models, providing theoretical justification.

- Investigate the effect of Jacobian & Hessian regularisation on the robustness of sequence models. If these regularisation techniques improve robustness, it would support the hypothesis that sequence models with less curved decision boundaries achieve greater robustness.

- Develop a framework based on the theory of condition numbers in numerical analysis and optimization to identify & characterise robustness.

2 Theory

2.1 Preliminaries

Let $f : \mathbb{R}^d \to \mathbb{R}^L$ be an arbitrary multi-class classifier. Given a datapoint $x \in \mathbb{R}^d$, the class which f predicts for x is given by $\hat{k}(x) = \arg\max_k f_k(x)$ where $f_k(x)$ is the k-th component of $f(x)$.

Definition 2.1.1 (Adversarial Perturbation) *An adversarial perturbation, $\Delta_{adv}(x; f)$ is defined by the following optimisation problem:*

$$\Delta_{adv}(x; f) = \arg\min_{\delta \in \mathbb{R}^d} \|\delta(x)\|_2 \ s.t \ \hat{k}(x + \delta) \neq \hat{k}(x)$$

In other words, $\Delta_{adv}(x; f)$ is the perturbation of minimal distance that results in misclassification. Given $\alpha \in \mathbb{R}$, if $\alpha \geqslant \|\Delta_{adv}(x; f)\|$ $\exists \delta \in \alpha\mathbb{S}$ where $\mathbb{S} = \left\{x \in \mathbb{R}^d \mid \|x\| = 1\right\}$ s.t $\hat{k}(x + \alpha\delta) \neq \hat{k}(x)$. In other words, there exists a perturbation vector of length α which causes misclassification. On the other hand, if $\alpha \leq \|\Delta_{adv}(x; f)\|, \forall r \in \alpha\mathbb{S}$ we have that $\hat{k}(x + \alpha\delta) = \hat{k}(x)$. All perturbations of magnitude less than α do not result in the classifier making a different prediction. In general the larger the value of $\Delta_{adv}(x; f)$, the more robust point $x \in \mathbb{R}^d$ is to adversarial perturbations.

Definition 2.1.2 (Random Perturbation) *As stated in [3], given $\varepsilon \in [0, 1]$*

$$\Delta_{unif,\varepsilon}(x; f) = \max_{\delta \geq 0} \delta \ s.t \ \mathbb{P}_{\delta\mathbb{S}}\left(\arg\max_k f_k(x) = \hat{k}(x) \neq \hat{k}(x + \delta)\right) \leqslant \varepsilon$$

In other words, given a rate of misclassification $\epsilon \in [0, 1]$, $\Delta_{unif,\varepsilon}(x; f)$ denotes the maximal radius of the sphere centred at $x \in \mathbb{R}^d$ such that the probability of misclassification of datapoints on $x + \delta\mathbb{S}$ is less than or equal to ϵ. Note that $\mathbb{S}$ denotes the unit sphere centred at the origin in $\mathbb{R}^d$.

Definition 2.1.3 (Confidence of Classification) *The confidence of a classification $F(x)$ at $x \in \mathbb{R}^d$ is defined as:*

$$F(x) = \hat{k}(x) - \arg\max_{k \neq \hat{k}(x)} f_k(x)$$

$F(x)$ describes the difference between likelihood of classification of the most probable class and the next most probable class. Note that this second most probable class need not be unique. The larger the value of $F(x)$, the more confident we are in the prediction $\hat{k}(x) = \arg\max_k f_k(x)$ given by the classifier for point x.

Definition 2.1.4 (Decision Boundary) *The decision boundary of a classifer, B is defined as:*

$$B = \left\{ x \in \mathbb{R}^d \mid F(x) = \hat{k}(x) - \arg\max_{k \neq \hat{k}(x)} f_k(x) = 0 \right\}$$

The decision boundary is the set of all points the classifier f is equally likely to classify into (at least) 2 distinct classes. Note that this can be reformulated as:

$$B = \left\{ x \in \mathbb{R}^d \mid f_i(x) - f_j(x) = 0 \right\}$$

Definition 2.1.5 (Approximation of the decision boundary) *The linear, first order approximation of the decision boundary at $z = x + \delta(x)$ is given by the set:*

$$x + \{v : \delta(x)^T v = ||\delta(x)||_2^2\}$$

The quadratic, second order approximation of the decision boundary, is given by:

$$x + \{v : (v - \delta)^T (H_z)(v - \delta) + \alpha_x \delta^T (v - \delta) = 0\}$$

where $\alpha_x = \frac{||\nabla F(z)||}{||\delta(x)||}$ and H_z denotes the Hessian matrix at z.

Definition 2.1.6 (Decision Boundary with Confidence) *Given a confidence margin $t > 0$, the decision boundary with thresholding or confidence margin t is given by:*

$$B' = \left\{ x \in \mathbb{R}^d \; s.t \; \|f_i(x) - f_j(x)\| < t \right\} \; s.t \; t > 0$$

In other words $x \notin B'$ implies that x is assigned class $\hat{k}(x)$ by the classifier f with confidence $t > 0$. All points $x \in B'$ are not clssified by confidence at least t

2.2 Problem Setting

Let $f : \mathbb{R}^d \to \mathbb{R}^L$ be an arbitrary multi-class classifier. Given a datapoint $x \in \mathbb{R}^d$, the minimal displacement from x to the decision boundary $B = \left\{ x \in \mathbb{R}^d \mid f_i(x) - f_j(x) = 0 \right\}$ is given by $\Delta_{adv}(x; f) = \arg\min_{\delta \in \mathbb{R}^d} \|\delta(x)\|_2 \; s.t \; \hat{k}(x + \delta) \neq \hat{k}(x)$.

Given $x \in \mathbb{R}^d$ we aim to approximate the value of $\|\Delta_{adv}(x; f)\|$ with upper and lower bounds, U, L respectively s.t $L \leq \|\Delta_{adv}(x; f)\| \leq U$ For perturbations of radius,

$\rho < L \leq \|\Delta_{adv}(x; f)\|$ misclassification does not occur.

Perturbations of radius $\rho > U \geq \|\Delta_{adv}(x; f)\|$ however, do result in misclassification. We aim to describe the extent of misclassification as a function of this radius ρ. In other words, we aim to characterise the probability of misclassification of points on the ball of radius ρ centred at x. Note this probability is proportional to the fraction of misclassified points on the the sphere. This viewpoint is useful for experimental validation with Monte-Carlo simulations which will be described later.

We formalise this idea as follows. Given a fixed probability of misclassification, $\delta \in [0, 1]$, we seek a radius of perturbation $\rho \geq \|\Delta_{adv}(x; f)\|$ of minimal length such that $\mathbb{P}_{v \in \mathbb{S}}(\hat{k}(\boldsymbol{x} + \rho \boldsymbol{v}) \neq \hat{k}(\boldsymbol{x})) \leq \delta$. We aim to find the perturbation radius of minimal distance such that the probability of misclassification with this perturbation is δ.

2.3 Upper Bound of Average Adversarial Distance

The folllowing analysis is adapted from [3], applied to the multiclass case. Let μ denote the distribution of points in $\mathbb{R}^d$ that we wish to classify. Let $y(x) \in \{1, \ldots, \ell\}$ be the true classes of points $x \in \mathbb{R}^d$ and let μ_i denote the distribution of classes $i \in \{1, \ldots, \ell\}$ *in* $\mathbb{R}^d$. The risk of the classifier $R(f)$ is the probability of misclassification which is given by:

$$R(f) = \mathbb{P}_\mu \left(\arg\max_k f_k(x) \neq y(x) \right) = \sum_{i=1}^{\ell} p_i \mathbb{P}_{\mu_i} \left(\arg\max_k P_k(x) = i \right)$$

where $p_i = \mathbb{P}_\mu \left(\arg\max_k f_k(x) = i \right)$

Definition 2.3.1 (Assumption (A)) *There exist* $\tau > 0$ *and* $0 < \gamma \leqslant 1$ *s.t* $\forall x \in \mathbb{R}^d$.

$$dist\,(x, S_i) \leqslant \tau \max\{0, |f_i(x) - f_j(x)|\}^\gamma$$

$$\text{dist}\,(x, s_i) = \min y\,\{\|x - y\|_2 \mid y \in Si\}, S_i = \left\{ x \in \mathbb{R}^d \mid \arg\max_k f_k(x) = i \right\}$$

In the case where assumption (A) holds, we can bound the expectation of the adversarial robustness of all points in $\mathbb{R}^d$.

Lemma 2.3.1 *Let f be an arbitrary classifier that satisfies (A) with parameters (τ, γ). Let $F(x) = \hat{k}(x) - \arg\max_{k \neq \hat{k}(x)} f_k(x)$. Then,*

$$\rho_{adv}(f) = \mathbb{E}_\mu \left(\Delta_{\text{adv}}(x; f) \right) \leq \tau 2 L^{2(1-\gamma)} \left(R(f) \|f\|_\infty + \sum_{i,j=1}^{\ell} \mid p_i \mathbb{E}_{\mu_i}(F(x)) - p_j \mathbb{E}_{\mu_j}(F(x)) \mid \right)$$

The above result provides an upper bound on the average adversarial perturbation distance for an arbitrary classifier f. $\rho_{adv}(f)$ is the average length of minimal perturbations required to result in misclassification. This is a powerful general result. However, this bound is not of practical use since it is unlikely that we know an explicit form of the distribution of μ and μ_i for $i \in \{1, \ldots, \ell\}$. The above lemma assumes too much knowledge of the distribution of datapoints. Thus, We need to focus on methods which are more local.

2.4 Decision Boundary Curvature & Robustness

Suppose point $x \in \mathbb{R}^d$ *is predicted to belong to class* $\hat{k}(x) = \arg\max_k f_k(x)$ *with confidence margin* $F(x) = \hat{k}(x) - \arg\max_{k \neq \hat{k}(x)} f_k(x) = t \geq 0$. *Let* $\delta \in \mathbb{R}^d$ *be a perturbation vector such that*

$$|F(x+\delta) - F(x)| < t \tag{1}$$

In other words, points x *and* $x+\delta$ *are classified by* f *as being in the same class with similar confidence margin* t.

Extending the work of [11], assume further that F *is third-order differentiable with:*

$$F(x+\delta) = F(x) + J \cdot \delta + \frac{1}{2}\delta^T H \delta + \frac{1}{3!} T_{ijk}\delta^i\delta^j\delta^k \tag{2}$$

where J *is the Jacobian vector,* H *is the real symmetric Hessian matrix and* T *the Tressian, is the tensor of third order partial derivatives of* F *with respect to* x. *Note that we adopt Einstein summation notation when dealing with the Tressian.*

It is important to note that J, H *&* T *encode important geometric properties of the decision boundary. In the following we combine 1 & 2*

$$\max_{\delta \in \mathbb{R}^d} \left(\left| J \cdot \delta + \frac{1}{2}\delta^T H \delta + \frac{1}{3!} T_{ijk}\delta^i\delta^j\delta^k \right| \right) < t \tag{3}$$

Note that H *is diagonalisable. Let* $y = E^T\delta$ *where* E *is the matrix of eigenvectors. Then,*

$$\left| J \cdot \delta + \frac{1}{2}\delta^T H \delta + \frac{1}{3!} T_{ijk}\delta^i\delta^j\delta^k \right| \leq \sum_{i=1}^{n} |J_i| \cdot |\delta_i| + \frac{1}{2}\sum_{i=1}^{n} |\lambda_i| \cdot \left|y_i^2\right| + \frac{1}{6}\kappa|\delta|^3$$

where δ_i *and* y_i *are the components of* δ *and* y *respectively.* J_i *is the* i*-th entry of the Jacobian vector,* λ_i *is the* i*-th eigenvalue of Hessian* H *&* κ *is the standard condition number of* T. *The inequality above suugets the smaller the values of* J_i, λ_i *and* κ, *the larger the magnitude of the perturbation,* δ *can be without resulting in misclassification. In other words, a wider and flatter decision boundary with approximately constant curvature is more robust. The following Theorem extends the work of [9] to the third order setting.*

Theorem 2.4.1 (Bound on $\Delta_{adv}(x; f)$***)*** *Given confidence margin* $t > 0$, *let* x *be such that* $c := t - F(x) \geq 0$. *Assume that* $\nu := \lambda_{\max}(H) \geq 0$, *with corresponding eigenvector* $u \in \mathbb{R}^d$ *and* κ, *the condition number of* T *be greater than zero. Let* $\alpha = \Delta_{adv}(x; f)$. *If* $\Delta = -\frac{3}{2}\kappa\nu\|J\|c + \frac{1}{2}\nu^3 c + \frac{1}{4}\nu^2\|J\|^2 - \frac{2}{3}\kappa\|J\|^3 - \frac{9}{2}\kappa^2c^2$, *then*

$$\max\{-\frac{2}{\kappa}\left(\frac{\nu}{2} + C + \frac{\Delta_0}{C}\right), 0\} \leq \alpha \leq \left|\frac{2}{\kappa}\left(\frac{\nu}{2} + C' + \frac{\Delta_0'}{C'}\right)\right|$$

$$\Delta_0 = \frac{1}{2}(\frac{\nu}{2} - \kappa\|J\|), \quad \Delta_1 = \frac{1}{4}(\nu^3 - \frac{9}{4}\kappa\nu\|J\| - 3\kappa^2 c), \quad C = \sqrt[3]{\frac{\Delta_1 + \sqrt{\Delta_1^2 - 4\Delta_0^3}}{2}}$$

$$\Delta_0' = \frac{1}{2}(\frac{\nu}{2} - \kappa\|J \cdot u\|) \quad \Delta_1' == \frac{1}{4}(\nu^3 - \frac{9}{4}\kappa\nu\|J \cdot u\| - 3\kappa^2 c) \quad C' = \sqrt[3]{\frac{\Delta_1' + \sqrt{\Delta_1'^2 - 4\Delta_0'^3}}{2}}$$

2.5 Jacobian Regularization, Curvature and Robustness

Theorem 2.5.1 *[4] Let $f : \mathbb{R}^d \to \mathbb{R}^L$ be a (twice) continuously differentiable multi-class classifier. Given $x \in \mathbb{R}^d$, suppose f is locally Lipschitz in $B_p(x, R) = \{y \in \mathbb{R}^d \mid \|x - y\|_p \leq R\}$. Let $p, q \in \mathbb{R}$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\forall \delta \in \mathbb{R}^d$ with*

$$\|\delta\|_p \leq \max_{R>0} \min \left\{ \min_{j \neq c} \frac{f_i(x) - f_j(x)}{\max_{y \in B_p(x,R)} \|\nabla f_i(y) - \nabla f_j(y)\|_q}, R \right\} := \alpha$$

it holds that $\hat{k}(x) = \arg\max_j f_j(x+\delta)$. The classifier decision does not change on $B_p(x, \alpha)$.

Provided f is Lipschitz continuous in a neighbourhood of x, $B_p(x, R)$, the above theorem gives an upper bound on $\Delta_{adv}(x; f)$ as a function of the Cross-Lipshitz terms, $||\nabla f_i(x) - \nabla f_j(x)||_q$. The following 2 lemmas allow us to make a connection between local Lipschitz continuity and bounded local decision boundary curvature. Note that we tract curvature as a bound on the eigenvalues (in absolute value) of the Hessian.

Lemma 2.5.1 *If the eigenvalues (in absolute value) of the hessian of a twice continuously differentiable function f is bounded by L, the following holds:*

$$-LI \leq \nabla^2 f(x) \leq LI \Rightarrow \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Lemma 2.5.2 *If a function f is twice continuously differentiable and locally Lipschitz continuous at x, then the eigenvalues of the Hessian (in absolute value) are bounded.*

In particular, bounded local curvature implies local Lipschitz continuity. This together with the statement of Lemma 2.5.1 implies we can tract a lower bound for $\Delta_{adv}(x; f)$ in this setting. $\forall \delta \in \mathbb{R}^d$ s.t $\|\delta\|_\mu \leqslant \alpha, \hat{k}(x) = \hat{k}(x + \delta)$. This supports our argument that is related to bounded curvature.

Note that that the smaller the Cross-Lipschitz terms $||\nabla f_i(x) - \nabla f_j(x)||_q$, the larger the quantity $||\delta||_p$. In other words, the classifier is more robust at x. The following lemma shows an important relationship between the Cross-Lipshitz terms and the Jacobian.

Lemma 2.5.3 *Suppose $\exists M > 0$ such that $||J(\boldsymbol{x}^\alpha)||_F^2 \leq M$, then $\exists C > 0$ s.t $||\nabla f_i(x) - \nabla f_j(x)||_q \leq C$.*

In other words, if the Frobenius norm of the Jacobian is bounded, the Cross-Lipschitz terms are bounded. This suggests incorporating a Jacobian regularisation term to our objective function during training will improve robustness.

I would like to thank my partner Peter Fazekas for formulating the following definition based on the main result of [5]:

Definition 2.5.1 *Let $J_{i,j}(x) = \frac{\partial f_i}{\partial x_j}(x)$ be the entries of the Jacobian matrix. [5] defines the Jacobian regularization by minimizing a joint loss function during training,*

$$\mathcal{L}^{\mathcal{B}}_{joint}(\theta) = \mathcal{L}_{bare}(\{x^\alpha, y^\alpha\}_{\alpha\in\mathcal{B}}; \theta) + \frac{\lambda_{JR}}{2}\left[\frac{1}{|\mathcal{B}|}\sum_{\alpha\in\mathcal{B}}||J(x^\alpha)||^2_F\right]$$

where $\mathcal{L}_{bare}$ defines the regular loss function during training. Minimising $\mathcal{L}_{joint}$, decreases the absolute values of the Jacobian matrix entries.

Following a similar setting as [8] Peter Fazekas & myself formulated the following theorem providing justification of the link between robustness and curvature of the decision boundary.

Theorem 2.5.2 *Fix a datapoint $x \in \mathbb{R}^d$ Let $\kappa > 0, \delta > 0, \tilde{\delta} > 0$, $v \in \mathbb{B}(x, \rho)$ and $m \in \mathbb{N}$. Approximate the decision boundary to second order by the set $z = x + r(x) = x + \{v : (v-r)_i (H_z)^i_j (v-r)^j + \alpha_x r_i (v-r)^i = 0\}$. Given $v \in \mathbb{B}(x, \rho)$,*

$$\mathbb{P}_{v\in\mathbb{S}}(\hat{k}(\boldsymbol{x} + \rho\boldsymbol{v}) \neq \hat{k}(\boldsymbol{x})) \leq \tilde{\delta} \tag{4}$$

holds true given $\mathbb{P}_{v\in\mathbb{S}}(h||\rho v - r||^2 + \alpha_x r^T(\rho v - r) \geq 0) \leq \delta$ where the constant h is given by the Frobenius norm $||H_z||_F$ evaluated at z and the radius ρ satisfies the quadratic equation:

$$h\rho^2 + ((\frac{f}{r} - h)\sqrt{\frac{16\ln(\frac{2}{\delta})}{m}}\rho + r(hr - 2f) \leq 0 \tag{5}$$

where $r = \|r(x)\|$ and $f = \alpha_x r = ||\nabla F(z)||_q r$.

Theorem 2.5.2 gives the probability of misclassification on the sphere of radius ρ centred at $x \in \mathbb{R}^d$. Given a fixed probability of misclassification, the larger the radius ρ the more robust the model.

Please note that due to the page restriction of this interim report, I have omitted all proofs in favour of describing the general framework of the project. All proofs will be present in the final report.

3 Work Done

The majority of Term 1 was spent developing a deterministic and probabilistic representation of robustness. Work done include:

- *Extending the upper bound for adversarial robustness for arbitrary binary classifiers described in [3] to the multi-class case, generalising assumptions to the multi-class setting in the process. Following a similar approach, I showed that cubic classifiers satisfy the assumptions given by [3] and determined an upper bound for adversarial robustness for classifiers of degree 3.*

- *A Monte Carlo simulation to test the above bounds. I approximated the probability of misclassification on the sphere of radius ρ, centred at $x \in \mathbb{R}^d$ as the ratio of misclassified points to the number of total points uniformly sampled from the sphere. I randomly sampled points $x \in \mathbb{R}^d$ and took the average of the probability of misclassification as an approximation to the average adversarial robustness. The results of this simulation agree with the theoretical bounds. However, the results indicate that the bounds are loose. Deriving bounds for higher degrees do not provide further insight.*

Next, I focus on the geometric properties of the decision boundary and it's effect on the adversarial robustness of an arbitrary classifier. Work done include:

- *Developing on a probabilistic setting for robustness. My partner & I modified the framework of [8], moving from the universal perturbation case to local adversarial perturbation setting. We developed a concentration inequality for the probability of misclassification on the sphere of radius ρ about $x \in \mathbb{R}^d$ as a function of perturbation radius ρ. The constants appearing in this inequality imply the robustness of a classifier improves as the curvature of the desision boundary decreases.*

- *Proving upper and lower bounds for adversarial robustness in the case where the decision boundary is approximated to third order, extending the result of [9]. Note these bounds are derived from solving for the roots of a cubic and this result only holds true for the case of a positive determinant. In agreement with the previous theorem proved, the less curved the decision boundary, the more robust the classifier.*

The end of the term was spent doing further literature review. I explored the connection of robustness of sequence models to statistical mechanics & random matrix theory. The work done is outlined below:

- *Investigating the connection of robustness of sequence models to statistical mechanics. In particular, I developed a spin glass model framework to describe bidirectional RNNs. I believe bidirectional sequence models provide a more natural analogy to spin glass models compared to feed forward neural networks as discussed in [1, 2].*

- *I explored the connection of robustness to random matrix theory. I confirmed the eigenvalues of the hessian of a RNN trained on MINST data follow the Marchenko-Pastur distribution as predicted by [10].*

4 Term 2 plans

I aim to develop a more fundamental framework for robustness incorporating the theory of condition numbers. In addition, I aim to link the ideas developed in Term 1 to Weyl's tube formula [6]. Weeks 1 and 2 will be spent on literature review while weeks 3 to 7 will be spent on formalising the framework. I believe this time-scale is realistic as our supervisor Martin Lotz has a clear direction in mind.

I plan to spent the majority of term 2 experimentally verifying the theory we developed:

- *From weeks 1 to 3 I plan to compare the curvature of the decision boundary and adversarial robustness of sequence models after Hessian regularisation and adversarial training respectively. Note Hessian regularisation aims to decrease curvature while adversarial training aims to improve robustness. If Hessian regularisation improves robustness and adversarial training decreases the curvature of the decision boundary, this supports the idea that curvature and robustness are intricately linked. In particular, a decrease in the curvature of decision boundary results in an increase in robustness.*

- *From weeks 4 to 6 I plan to investigate if Hessian regularisation can be used as an alternative to adversarial training. [9] show that projected gradient descent along the eigenvectors of the Hessian converge to adversarial examples efficiently. I aim to investigate if this implies regions of the decision boundary that are more curved are the more susceptible to adversarial attacks. I aim to investigate this property in the context of sequence models. Note that I will use a sequence model trained on stock data as a case study. I have developed this model in Term 1 but some fine tuning is necessary.*

- *From weeks 7 to 8 I would like to identify the link between adversarial perturbations for image classifiers and sequence models. In particular, I aim to investigate whether universal perturbations to MNIST data which fool a CNN would fool an PixelRNN as well.*

- *Weeks 8 to 10 will be dedicated to writing up the final report and ironing out the details with my partner and supervisor. I plan to complete all the experiments by week 8 to set time for this. Note that the Final Report is due Thursday Week 10 of Term 2.*

Time permitting, I would like to explore the connection of robustness of sequence models to spin glass models in statistical physics. I hope to draw an analogy between adversarial perturbations and nucleation theory in statistical physics. If successful, I can translate well established results/predictions of statistical mechanics to the language of robustness and test these predictions experimentally.

References

[1] Elena Agliari, Adriano Barra, Andrea Galluzzi, Daniele Tantari, and Flavia Tavani. A walk in the statistical mechanical formulation of neural networks, 2014.

[2] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks, 2015.

[3] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations, 2016.

[4] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation, 2017.

[5] Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization, 2019.

[6] Martin Lotz. On the volume of tubular neighborhoods of real algebraic varieties, 2013.

[7] Yan Luo, Xavier Boix, Gemma Roig, Tomaso A. Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *CoRR*, abs/1511.06292, 2015.

[8] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: A geometric perspective. In *International Conference on Learning Representations*, 2018.

[9] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa, 2018.

[10] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 2637–2646. Curran Associates, Inc., 2017.

[11] Fuxun Yu, Chenchen Liu, Yanzhi Wang, Liang Zhao, and Xiang Chen. Interpreting adversarial robustness: A view from decision surface in input space, 2018.