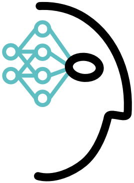
Can we certify adversarial robustness for classifiers learning high dimensional data?

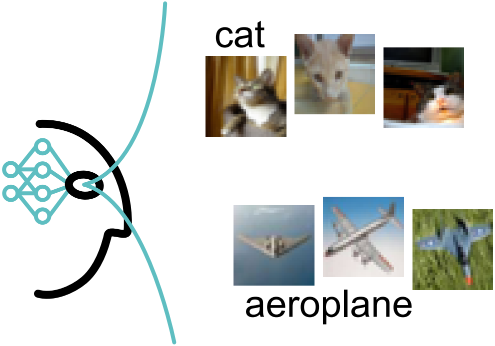
Oliver Sutton

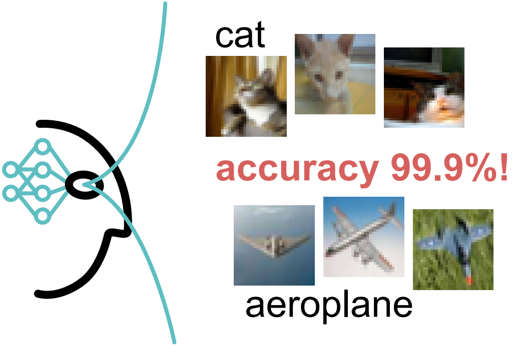
Department of Mathematics
King's College London

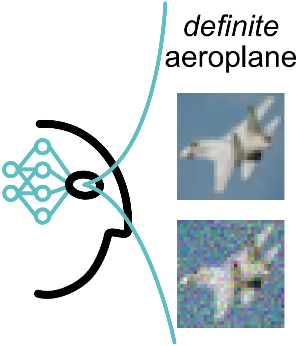
with **I. Tyukin** (King's College London),
Q. Zhou (King's College London),
A. Gorban (University of Leicester)
A. Bastounis (University of Leicester)
D. Higham (University of Edinburgh)

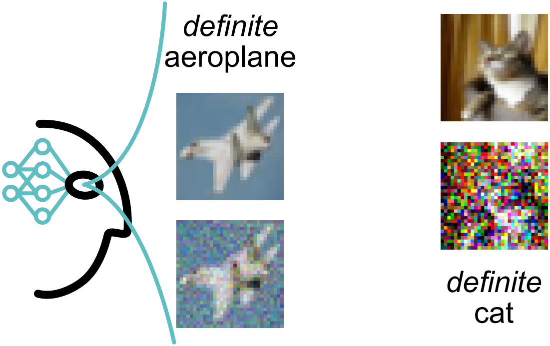
The Alan Turing Institute
King's College London
EPSRC
Engineering and Physical Sciences Research Council

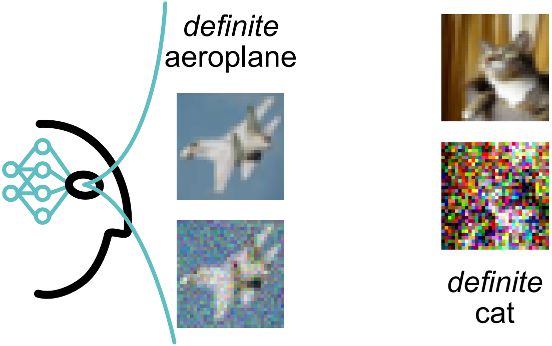



cat
aeroplane


cat
accuracy 99.9%!
aeroplane


definite
aeroplane

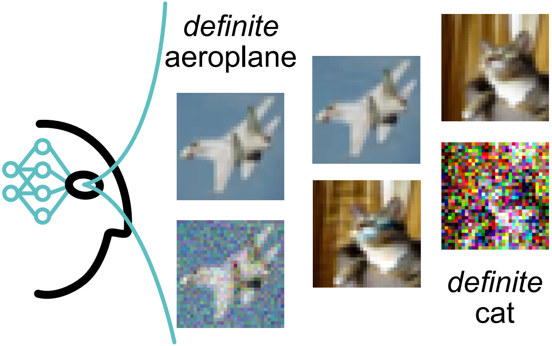

definite
aeroplane
definite
cat



Stable classifier

The classifier is robust to even very noisy inputs

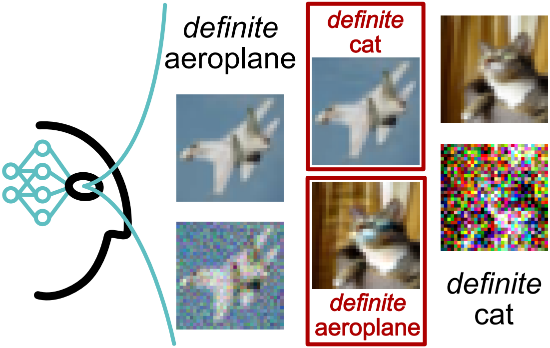
- 2000 random perturbations
  - Plane (with max pixel change 0.3): 4 (0.2%) caused misclassification
  - Cat (with max pixel change 1.6): 83 (4.15%) caused misclassification



Stable classifier

The classifier is robust to even very noisy inputs

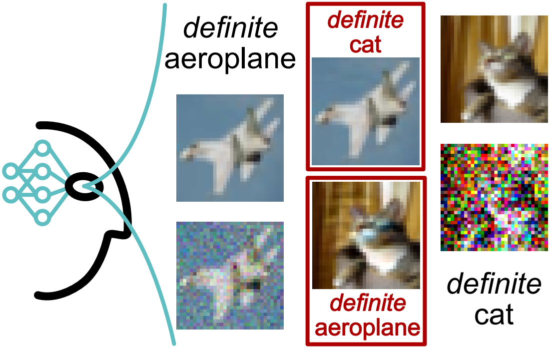
- 2000 random perturbations
  - Plane (with max pixel change 0.3): 4 (0.2%) caused misclassification
  - Cat (with max pixel change 1.6): 83 (4.15%) caused misclassification



Stable classifier

The classifier is robust to even very noisy inputs

- 2000 random perturbations
  - Plane (with max pixel change 0.3): 4 (0.2%) caused misclassification
  - Cat (with max pixel change 1.6): 83 (4.15%) caused misclassification

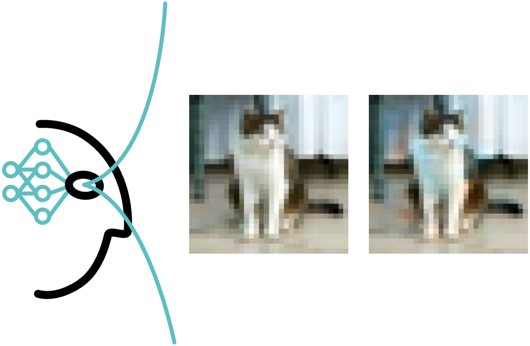


Stable classifier

The classifier is robust to even very noisy inputs

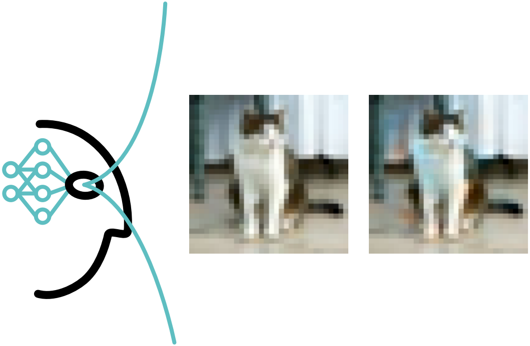
- 2000 random perturbations
  - Plane (with max pixel change 0.3): 4 (0.2%) caused misclassification
  - Cat (with max pixel change 1.6): 83 (4.15%) caused misclassification

...so what happened here?!?!



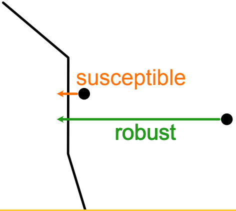
Adversarial attacks[1]

A *small* modification to an input which causes a classifier to confidently misclassify it



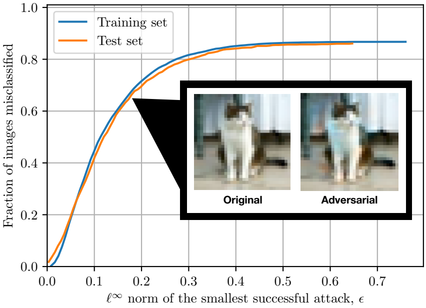
Adversarial attacks[1]

A *small* modification to an input which causes a classifier to confidently misclassify it

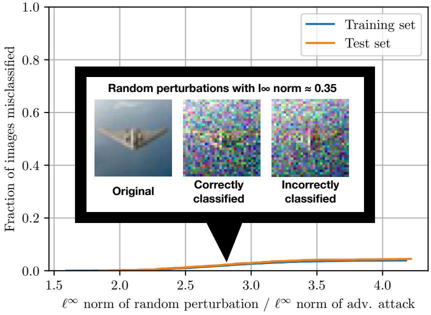


Aeorplane vs Cat see: [S. et al. (2023) arXiv:2309.03665]

Misclassified after adversarial attack



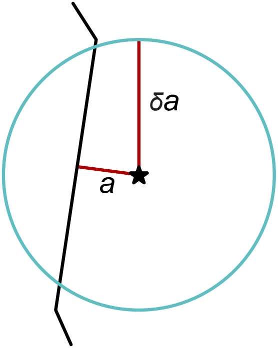
Misclassified after any of 2000 random noise samples

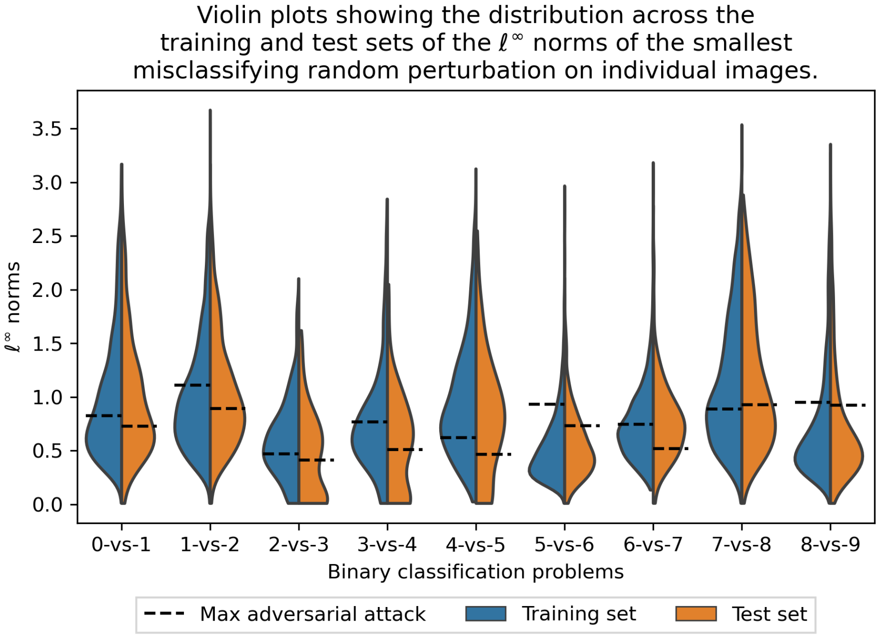


Other benchmarks see [S. et al. (2023) arXiv:2309.03665] for details

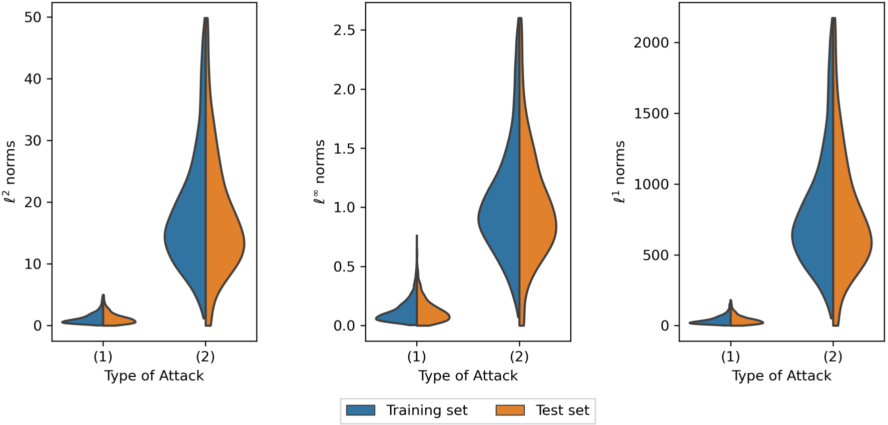
	CIFAR-10	Fashion MNIST	GTSRB
Accuracy	99.70, 95.80	99.51, 99.4	98.32, 98.51
Adversarial attack susceptibility	91.88, 89.96	53.58, 53.01	77.53, 77.00
Random attack susceptibility ($\delta = 2$)	0.02, 0.17	0.07, 0.09	0.36, 0.36
Random attack susceptibility ($\delta = 5$)	2.65, 2.57	10.71, 13.35	5.76, 5.1
Input dimension	$32 \times 32 \times 3$	$28 \times 28 \times 1$	$30 \times 30 \times 3$
Number of classes	10	10	6

- CNNs trained for each binary classification problem in each benchmark
- Reporting are medians over all problems (train, test)
- **Adversarial attack susceptibility**: fraction of correctly classified images susceptible to an adversarial attack
- **Random attack susceptibility**: fraction of adversarially susceptible images which were misclassified after any of 2000 random perturbations sampled uniformly from a ball with radius δ times larger than the smallest adversarial attack found on that image




Violin plots showing the distribution across the training and test sets of the ℓ^∞ norms of the smallest misclassifying random perturbation on individual images.
ℓ^∞ norms
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0-vs-1
1-vs-2
2-vs-3
3-vs-4
4-vs-5
5-vs-6
6-vs-7
7-vs-8
8-vs-9
Binary classification problems
Max adversarial attack
Training set
Test set

Violin plots showing the distribution across the training and test sets of the norms of (1) smallest successful adversarial attacks, and (2) smallest misclassifying random perturbation on individual images.



Paradox of apparent stability

Seemingly stable classifier (probabilistic stability)

Even large random noise is unlikely to cause an input to be misclassified

Susceptible to adversarial attacks (deterministic instability)

A *small* modification can be made to most inputs which causes a classifier to confidently misclassify them

Paradox of apparent stability

Seemingly stable classifier (probabilistic stability)

Even large random noise is unlikely to cause an input to be misclassified

Susceptible to adversarial attacks (deterministic instability)

A *small* modification can be made to most inputs which causes a classifier to confidently misclassify them

Probabilistic stability does not prevent *deterministic instability*!

Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] Elan Rosenfeld [1] J. Zico Kolter [1 2]

(PMLR 2019)

Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] Elan Rosenfeld [1] J. Zico Kolter [1 2]

(PMLR 2019)

Certified Adversarial Robustness with Additive Noise

Bai Li
Department of Statistical Science
Duke University
bai.li@duke.edu

Changyou Chen
Department of CSE
University at Buffalo, SUNY
cchangyou@gmail.com

Wenlin Wang
Department of ECE
Duke University
wenlin.wang@duke.edu

Lawrence Carin
Department of ECE
Duke University
lcarin@duke.edu

(NeurIPS 2019)

Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] Elan Rosenfeld [1] J. Zico Kolter [1 2]

(PMLR 2019)

Certified Adversarial Robustness with Additive Noise

Bai Li
Department of Statistical Science
Duke University
bai.li@duke.edu

Changyou Chen
Department of CSE
University at Buffalo, SUNY
cchangyou@gmail.com

Wenlin Wang
Department of ECE
Duke University
wenlin.wang@duke.edu

Lawrence Carin
Department of ECE
Duke University
lcarin@duke.edu

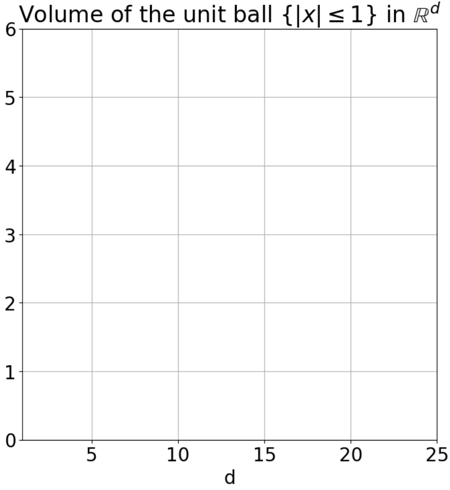
(NeurIPS 2019)

UniT: A Unified Look at Certified Robust Training against Text Adversarial Perturbation

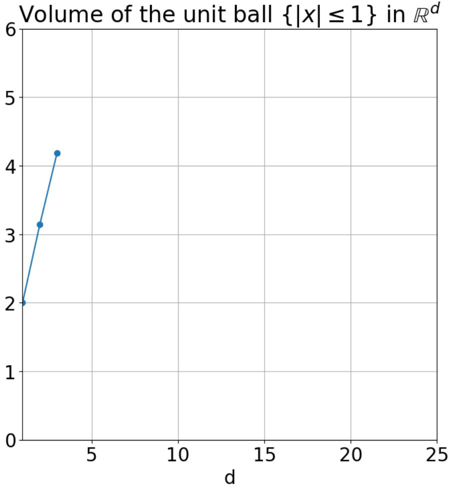
Muchao Ye[1] Ziyi Yin[1] Tianrong Zhang[1] Tianyu Du[2]
Jinghui Chen[1] Ting Wang[3] Fenglong Ma[1*]
[1]The Pennsylvania State University, [2]Zhejiang University, [3]Stony Brook University

(NeurIPS 2023)

Characterising high dimensional spaces



Characterising high dimensional spaces

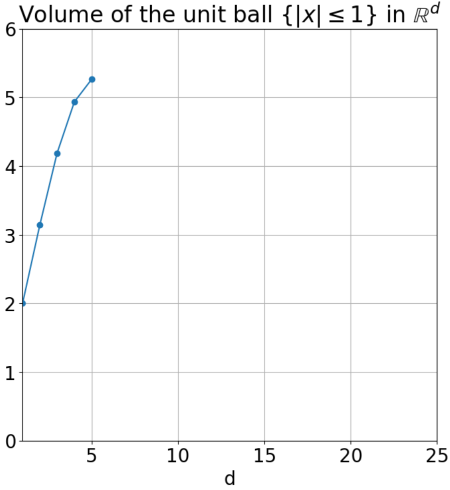


$$V_1 = 2$$

$$V_2 = \pi$$

$$V_3 = \frac{4}{3}\pi$$

Characterising high dimensional spaces



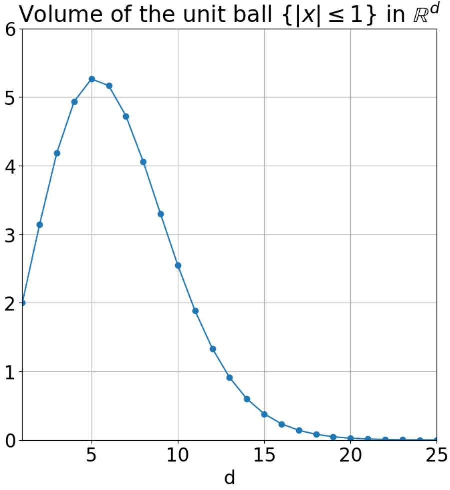
$$V_1 = 2$$

$$V_2 = \pi$$

$$V_3 = \frac{4}{3}\pi$$

$$\vdots$$

Characterising high dimensional spaces



$$V_1 = 2$$

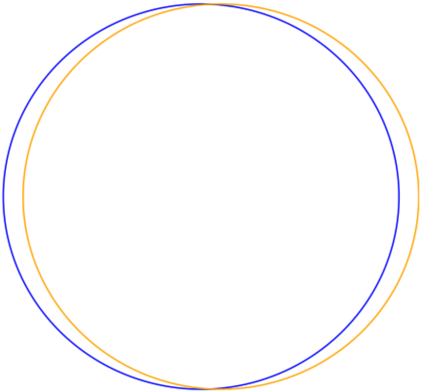
$$V_2 = \pi$$

$$V_3 = \frac{4}{3}\pi$$

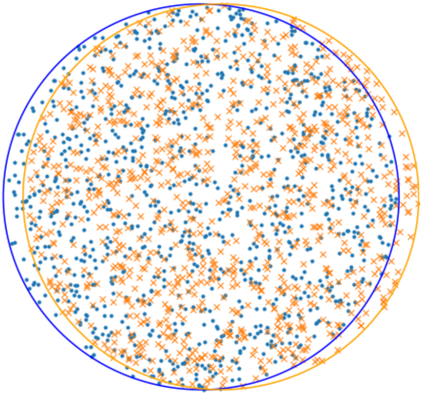
$$\vdots$$

$$V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

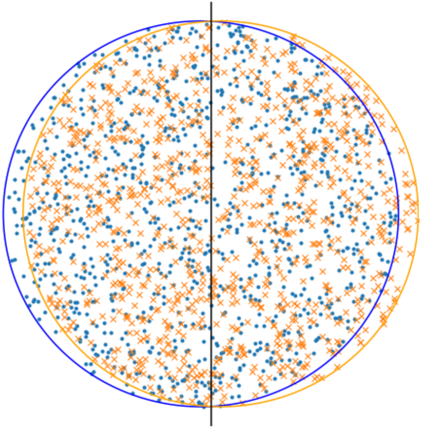
Characterising high dimensional spaces



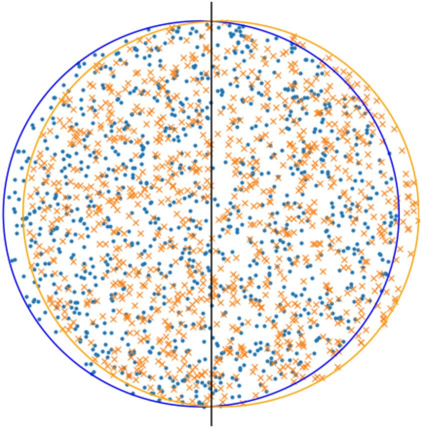
Characterising high dimensional spaces

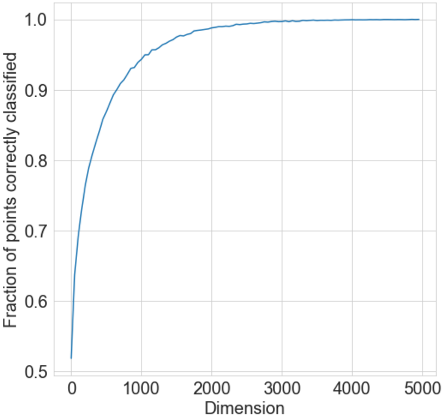


Characterising high dimensional spaces

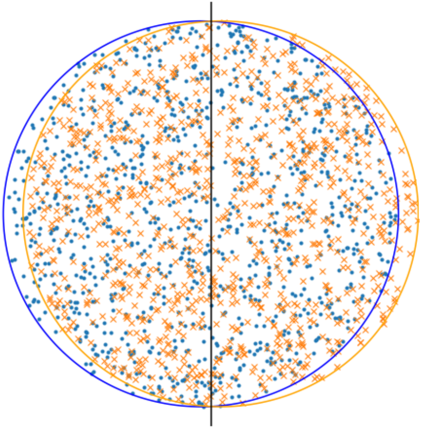


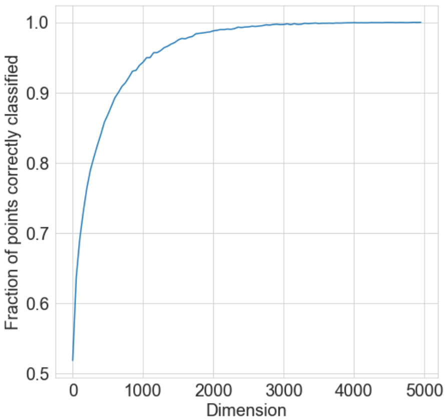
Characterising high dimensional spaces





Characterising high dimensional spaces





MNIST: 784 dimensions

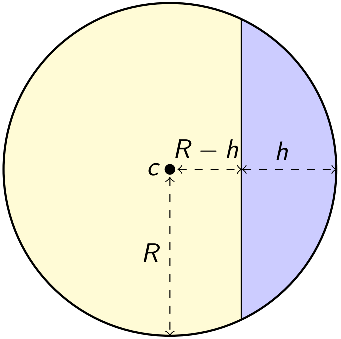
CIFAR-10: 3,072 dimensions

Llama-2: 4,096 dimensions

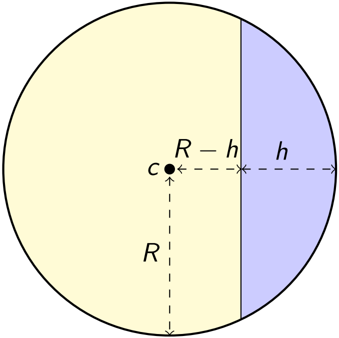
GPT3-Davinci: 12,288 dimensions

ImageNet: 196,608 dimensions

Characterising high dimensional spaces



Characterising high dimensional spaces



$$\frac{V_d^{\text{cap}}(R, h)}{V_d^{\text{ball}}(R)} \leq \frac{1}{2}\Big(\underbrace{1 - \Big(1 - \frac{h}{R}\Big)^2}_{<1}\Big)^{\frac{d}{2}}$$

$$\approx \exp\Big(- f\Big(\frac{h}{R}\Big) d\Big)$$

Characterising high dimensional spaces

Concentration of measure[2]

- Let x be sampled uniformly from the unit ball in $\mathbb{R}^d$. Then, for $0 \leq r \leq 1$

$$P(\|x\| > r) \geq 1 - r^d$$

[2] Ledoux (2001), Ball (1997), ...

[3] Kainen and Kůrková (1993), Gorban, Tyukin, Prokhorov, Sofeikov (2016), ...

Characterising high dimensional spaces

Concentration of measure[2]

- Let x be sampled uniformly from the unit ball in $\mathbb{R}^d$. Then, for $0 \leq r \leq 1$

$$P(\|x\| > r) \geq 1 - r^d$$

Quasi-orthogonality[3]

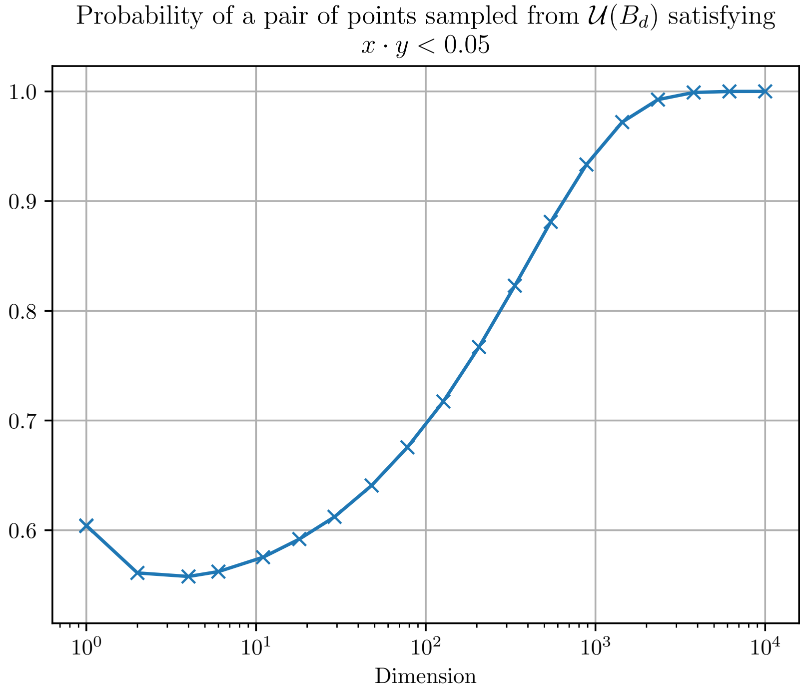
- In high dimensional spaces, randomly sampled points are typically nearly orthogonal
- For any $\epsilon > 0$ the number of points $x_i \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ such that $|(x_i, x_j)| \leq \epsilon$ for all $i \neq j$ grows exponentially with d
- For points x, y sampled independently and uniformly on the sphere in $\mathbb{R}^d$,

$$P(|(x, y)| < \epsilon) \geq 1 - \exp\left(-\frac{d\epsilon^2}{2}\right)$$

See notes: Ball (1997) 'An elementary introduction to modern convex geometry' for an introduction to high dimensional geometry

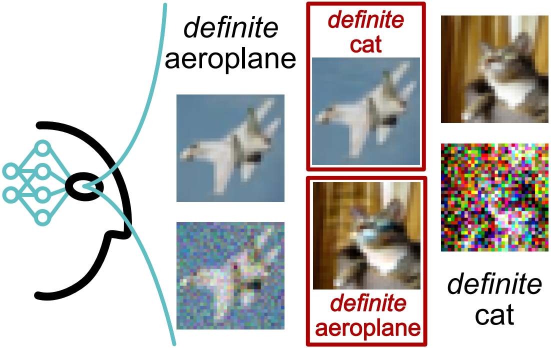
[2] Ledoux (2001), Ball (1997), ...

[3] Kainen and Kůrková (1993), Gorban, Tyukin, Prokhorov, Sofeikov (2016), ...



Probability of a pair of points sampled from $\mathcal{U}(B_d)$ satisfying $x \cdot y < 0.05$
1.0
0.9
0.8
0.7
0.6
10^0
10^1
10^2
10^3
10^4
Dimension

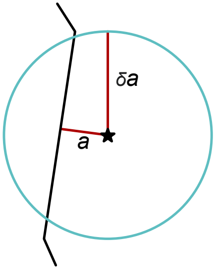
Explaining the paradox [S. et al. (2023) arXiv:2309.03665]



Paradox of apparent stability

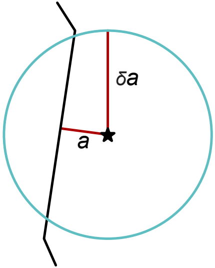
- Even *large random noise* is unlikely to cause an input to be misclassified
- Most inputs can be misclassified by adding a *small computed attack*

Explaining the paradox [S. et al. (2023) arXiv:2309.03665]



- adversarial attack walks straight to decision boundary
- adding random noise samples another point in this ball
- misclassified points are those from a spherical cap
- relative volume of a spherical cap is small in high dimensions

Explaining the paradox [S. et al. (2023) arXiv:2309.03665]



- adversarial attack walks straight to decision boundary
- adding random noise samples another point in this ball
- misclassified points are those from a spherical cap
- relative volume of a spherical cap is small in high dimensions

Theorem (Random noise is a bad way to detect adversarial attacks)

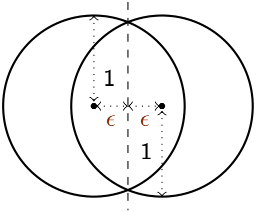
Let $f : \mathbb{R}^d \to \{0, 1\}$ be a linear classifier, let $x \in \mathbb{R}^d$ with $f(x) = 0$, and let

$$a = \inf_{v \in \mathbb{R}^d \text{ such that } f(x+v)=1} \|v\|.$$

Then, for any $\delta > 1$,

$$P(s \sim \mathcal{U}(\mathbb{B}^d_{\delta a}) \, : \, f(x+s) \neq f(x)) \leq \frac{1}{2}\left(1 - \frac{1}{\delta^2}\right)^{\frac{d}{2}}.$$

Building up the model [S. et al. (2023) arXiv:2309.03665]



decision surface of f

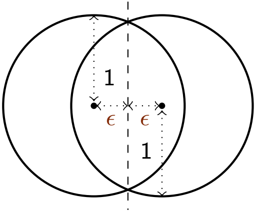
- Data sampled in dimension d
- Balls B_0 and B_1, unit radius, centres 2ϵ apart
- Points of class 0 sampled from distribution D_0 supported in B_0
- Points of class 1 sampled from distribution D_1 supported in B_1
- Distributions D_0 and D_1 don't have pathological accumulation points
  - they have densities p_0 and p_1 which are bounded[4]: there exists $A \geq 1$ such that

$$p_i(x) \leq \frac{A}{V^d(B_i)} \quad \text{for all } x \in B_i$$

- Combined distribution $\mathcal{D}_\epsilon$ samples labelled point $(x, \ell) \in \mathbb{R}^d \times \{0, 1\}$; each label has probability $\frac{1}{2}$
- f is the optimal (balanced) classifier for this

[4] A simplified version of the Smeard Absolute Continuity (SmAC) condition: [Gorban et al. (2018) Information Sciences]

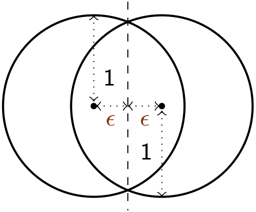
Building up the model [S. et al. (2023) arXiv:2309.03665]



decision surface of f

- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

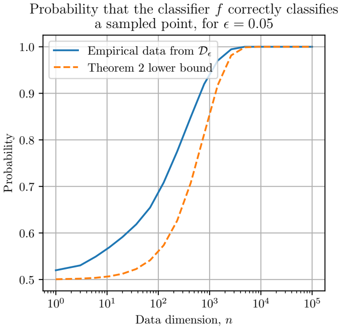
- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (The classifier is accurate)

For any $\epsilon > 0$, the probability that the classifier applies the correct label to a randomly sampled data point grows exponentially to 1 with dimension n, specifically

$$P((x, \ell) \sim \mathcal{D}_\epsilon : f(x) = \ell) \geq 1 - \frac{1}{2} A (1 - \epsilon^2)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]

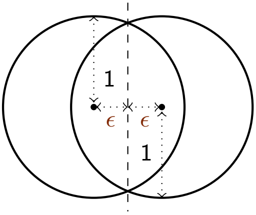


Theorem (The classifier is accurate)

For any $\epsilon > 0$, the probability that the classifier applies the correct label to a randomly sampled data point grows exponentially to 1 with dimension n, specifically

$$P((x, \ell) \sim \mathcal{D}_\epsilon : f(x) = \ell) \geq 1 - \frac{1}{2} A (1 - \epsilon^2)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

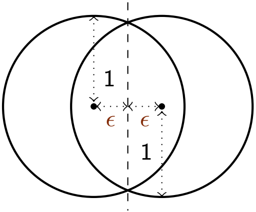
- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (The classifier is accurate)

For any $\epsilon > 0$, the probability that the classifier applies the correct label to a randomly sampled data point grows exponentially to 1 with dimension n, specifically

$$P((x, \ell) \sim \mathcal{D}_\epsilon : f(x) = \ell) \geq 1 - \frac{1}{2} A (1 - \epsilon^2)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

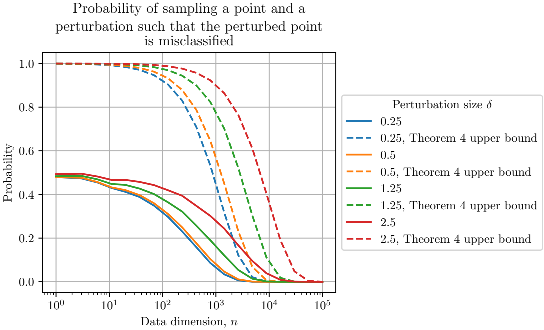
- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (Destabilising perturbations are rare)

For any fixed $\delta > \epsilon \geq 0$, the probability that a randomly selected perturbation with Euclidean norm δ causes a randomly sampled data point to be misclassified converges exponentially to 0 with the dimension d, specifically

$$P\Big((x,\ell) \sim \mathcal{D}_\epsilon, s \sim \mathcal{U}(\mathbb{B}^d{}_\delta) : f(x+s) \neq \ell\Big) \leq A\bigg(1 - \Big(\frac{\epsilon}{1+\delta}\Big)^2\bigg)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]

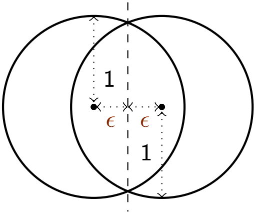


Theorem (Destabilising perturbations are rare)

For any fixed $\delta > \epsilon \geq 0$, the probability that a randomly selected perturbation with Euclidean norm δ causes a randomly sampled data point to be misclassified converges exponentially to 0 with the dimension d, specifically

$$P\Big((x, \ell) \sim \mathcal{D}_\epsilon, s \sim \mathcal{U}(\mathbb{B}^d{}_\delta) : f(x+s) \neq \ell\Big) \leq A\Big(1 - \Big(\frac{\epsilon}{1+\delta}\Big)^2\Big)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

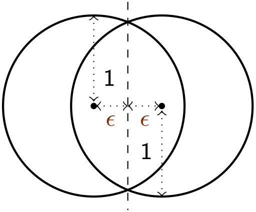
- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (Destabilising perturbations are rare)

For any fixed $\delta > \epsilon \geq 0$, the probability that a randomly selected perturbation with Euclidean norm δ causes a randomly sampled data point to be misclassified converges exponentially to 0 with the dimension d, specifically

$$P\big((x,\ell) \sim \mathcal{D}_\epsilon, s \sim \mathcal{U}(\mathbb{B}^d{}_\delta) : f(x+s) \neq \ell\big) \leq A\Big(1 - \Big(\frac{\epsilon}{1+\delta}\Big)^2\Big)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

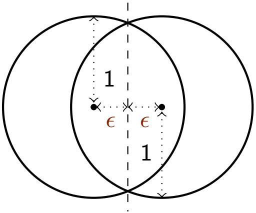
- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (Susceptible data points are typical)

For any $\epsilon \geq 0$ and $\delta \in [\epsilon, 1+\epsilon]$, the probability that a randomly sampled data point is susceptible to an adversarial attack with Euclidean norm δ grows exponentially to 1 with the dimension d, specifically

$$P\big((x, \ell) \sim \mathcal{D}_\epsilon : \text{there exists } s \in \mathbb{B}^d{}_\delta \text{ such that } f(x+s) \neq \ell\big)$$
$$\geq 1 - \frac{1}{2} A (1 - (\delta - \epsilon)^2)^{\frac{d}{2}}.$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



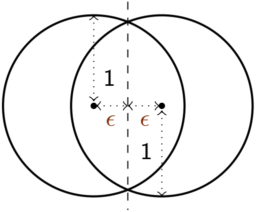
decision surface of f

- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Theorem (Gradient-based methods find the optimal adversarial attack)

Let $L : \mathbb{R}_{>0} \to \mathbb{R}$ denote any differentiable, monotonically increasing loss function, and let $(x, \ell) \sim \mathcal{D}_\epsilon$. Then, with probability 1 with respect to the sample (x, ℓ), the gradient of the loss $L(|\tilde{f}(x) - \ell|)$ with respect to the components of x corresponds to a positive multiple of the optimal attack direction.

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]



decision surface of f

- Class i sampled from distribution D_i in $B_i \subset \mathbb{R}^d$
- Centre distance: 2ϵ
- Bounded densities: $A \in \mathbb{R}$ s.t. $p_i(x) \leq \frac{A}{V^d(B_i)}$
- Combined distribution $\mathcal{D}_\epsilon$ samples (x, ℓ); each label has probability $\frac{1}{2}$

Let $d_\ell(x)$ measure how far x is on the wrong side of the decision boundary for class ℓ.

Theorem (Adversarial attacks are universal)

Let $\epsilon \geq 0$ and suppose that $x, z \sim D_\ell$ are independently sampled with the same label ℓ. For any $\gamma \in (0, 1]$, the probability that x is destabilised by all perturbations $s \in \mathbb{R}^d$ which destabilise z with margin $d_\ell(z + s) > \gamma$ converges exponentially to 1 with d. Specifically, let $S_z = \{s \in \mathbb{R}^d : d_\ell(z + s) > \gamma\}$. Then,

$$P(x, z \sim \mathcal{D}_\ell : f(x + s) \neq \ell \text{ for all } s \in S_z) \geq \left(1 - A\left(1 - \frac{\gamma^2}{4}\right)^{\frac{d}{2}}\right)^2$$

The model predicts the observations [S. et al. (2023) arXiv:2309.03665]

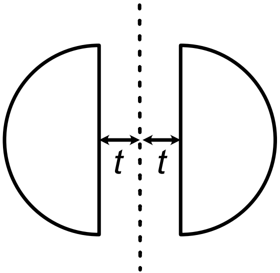
In summary, in high dimensions:

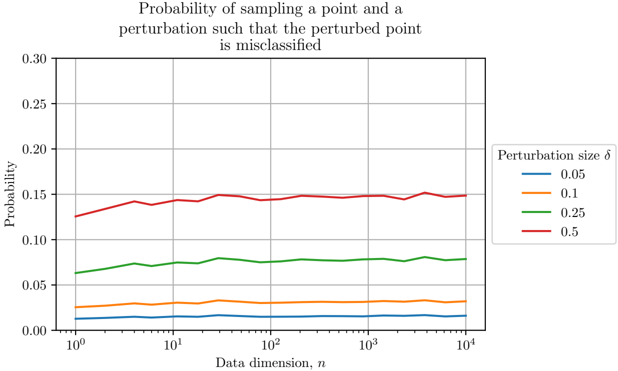
- The classifier is **accurate**
- Destabilising random perturbations are **rare**
- Typical data points sampled from either class are **susceptible** to small adversarial attacks which can be **easily constructed** and which **universally** affect most points from the same class

Extensions discussed in the paper:

- General data distributions
- Non-flat decision surfaces
- Multi-class setting

Coda: Classification with no margin





Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.
$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.
$\mathcal{L}$: loss function $\mathcal{T}$: training data $\mathcal{V}$: test data $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.

$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.

$\mathcal{L}$: loss function $\qquad$ $\mathcal{T}$: training data $\qquad$ $\mathcal{V}$: test data $\qquad$ $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

1. *There exists a network which correctly classifies the training data $\mathcal{T}$ and the test data $\mathcal{V}$, satisfying*

$$f \in \arg \min_{\varphi \in \mathcal{NN}} \sum_{(x,\ell) \in \mathcal{T} \cup \mathcal{V}} \mathcal{L}(x, \ell; f)$$

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.

$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.

$\mathcal{L}$: loss function $\quad \mathcal{T}$: training data $\quad \mathcal{V}$: test data $\quad M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

1. *There exists a network which correctly classifies the training data $\mathcal{T}$ and the test data $\mathcal{V}$, satisfying*

$$f \in \arg\min_{\varphi \in \mathcal{NN}} \sum_{(x,\ell) \in \mathcal{T} \cup \mathcal{V}} \mathcal{L}(x, \ell; f)$$

2. *Yet, for any $q \in (0, 1/2)$, with probability greater than or equal to $1 - \exp(-2q^2 M)$ there exists a multi-set $\mathcal{U} \subset \mathcal{T} \cup \mathcal{V}$ of cardinality at least $\lfloor (1/2 - q)M \rfloor$ on which f is unstable in the sense that for any $(x, \ell) \in \mathcal{U}$ and any $\alpha \in (0, \varepsilon/2)$, there exists a perturbation $\zeta \in \mathbb{R}^n$ with $\|\zeta\| \leq \alpha/\sqrt{n}$ and*

$$|f(x) - f(x + \zeta)| = 1.$$

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.

$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.

$\mathcal{L}$: loss function $\mathcal{T}$: training data $\mathcal{V}$: test data $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

3. *Moreover, such destabilising perturbations are typical in the sense that if vectors ζ are sampled from the equidistribution in $\mathbb{B}_n(\alpha/\sqrt{n}, 0)$, then for $(x, \ell) \in \mathcal{U}$*

$$|f(x) - f(x+\zeta)| = 1 \quad \textit{with probability at least} \quad 1 - \frac{1}{2^n}.$$

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.

$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.

$\mathcal{L}$: loss function $\quad$ $\mathcal{T}$: training data $\quad$ $\mathcal{V}$: test data $\quad$ $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

3. *Moreover, such destabilising perturbations are typical in the sense that if vectors ζ are sampled from the equidistribution in $\mathbb{B}_n(\alpha/\sqrt{n}, 0)$, then for $(x, \ell) \in \mathcal{U}$*

$$|f(x) - f(x + \zeta)| = 1 \quad \textit{with probability at least} \quad 1 - \frac{1}{2^n}.$$

4. *Furthermore, there exist universal destabilising perturbations, in the sense that a single perturbation ζ drawn from the equidistribution in $\mathbb{B}_n(\alpha/\sqrt{n}, 0)$ destabilises $m \leq |\mathcal{U}|$ points from the set $\mathcal{U}$ with probability at least*

$$1 - \frac{m}{2^n}.$$

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.
$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.
$\mathcal{L}$: loss function $\mathcal{T}$: training data $\mathcal{V}$: test data $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

5. *For the* **same distribution** $\mathcal{D}_\delta$ *there is a* **robust network with the same architecture** *as f, satisfying*

$$\tilde{f} \in \arg \min_{\varphi \in \mathcal{NN}_{\mathbf{N},L}} \mathcal{L}(\mathcal{T} \cup \mathcal{V}, \varphi)$$

with $\mathcal{L}(\mathcal{T} \cup \mathcal{V}, \tilde{f}) = 0$, which is robust in the sense that for all $(x, \ell) \in \mathcal{T} \cup \mathcal{V}$

$$\tilde{f}(x) = \tilde{f}(x + \zeta)$$

for any $\zeta \in \mathbb{R}^n$ with $\|\zeta\| \leq \alpha/\sqrt{n}$, even when $|\mathcal{T} \cup \mathcal{V}| = \infty$.

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.
$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.
$\mathcal{L}$: loss function $\quad$ $\mathcal{T}$: training data $\quad$ $\mathcal{V}$: test data $\quad$ $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

6 *Moreover, there exist pairs of unstable and robust networks, $f_\lambda, \tilde{f}_\lambda$ and $f_\Lambda, \tilde{f}_\Lambda$, satisfying the statements above such that the maximum absolute difference between their weights and biases is either arbitrarily small or arbitrarily large. That is, for any $\lambda > 0, \Lambda > 0$:*

$$\|\Theta(f_\lambda) - \Theta(\tilde{f}_\lambda)\|_\infty < \lambda, \ \|\Theta(f_\Lambda) - \Theta(\tilde{f}_\Lambda)\|_\infty > \Lambda.$$

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.
$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.
$\mathcal{L}$: loss function $\mathcal{T}$: training data $\mathcal{V}$: test data $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

7 *However, for the above robust solution $\tilde{f}$,*

a) *there exists an uncountably large family of distributions $\tilde{D}_\delta \in \mathcal{F}$ on which $\tilde{f}$ correctly classifies both the training and test data, yet fails in the same way*

b) *there exists an uncountably large family of distributions $\hat{D}_\delta \in \mathcal{F}$ such that the map $\tilde{f}$ is robust on $\mathcal{T} \cup \mathcal{V}$ (with respect to perturbations ζ with $\|\zeta\| \leq \alpha/\sqrt{n}$, $\alpha \in (0, \varepsilon/2)$) with probability*

$$\left(1 - \frac{1}{2^{n+1}}\right)^{Mk}$$

but is unstable to arbitrarily small perturbations on future samples with probability $k/2^{n+1}$.

Can we certify models are free from adversarial attacks?

$\mathcal{NN}$: networks with input dimension d, first hidden layer has $2d$ neurons, $L \geq 2$ layers, ReLU activations inside, step function for binary classification on output.
$\mathcal{F}$: family of 2-class data distributions, margin at least $\delta > 0$ between opposite classes.
$\mathcal{L}$: loss function $\quad$ $\mathcal{T}$: training data $\quad$ $\mathcal{V}$: test data $\quad$ $M = |\mathcal{T} \cup \mathcal{V}|$

Theorem (Inevitability, typicality and undetectability of instability)

Let $\varepsilon \in (0, \sqrt{d} - 1)$ and fix $0 < \delta \leq \varepsilon/\sqrt{d}$. Then, there is an uncountably large family of distributions $\mathcal{D}_\delta \in \mathcal{F}$ such that for any $\mathcal{D}_\delta \in \mathcal{F}$, any training and validation data $\mathcal{T}$, $\mathcal{V}$ drawn independently from $\mathcal{D}_\delta$:

1. *A network perfectly classifies the data, and minimises the loss*
2. *The training/test points are susceptible to small adversarial attacks*
3. *Nearly half the training/test points are susceptible to small adversarial attacks*

Conclusions

- Stability to *random* perturbations is not the same as stability to *adversarial* perturbations!
- In high dimensions, the two are very different

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As the
unprocessed data that should have been added to the final page this e
has been added to receive it.
If you rerun the document (without altering it) this surplus page will g
because LaTeX now knows how many pages to expect for this documen