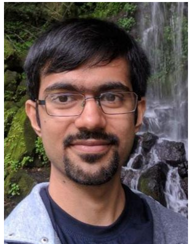
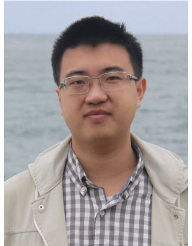
Can non-Lipschitz networks be robust?

November, 2024

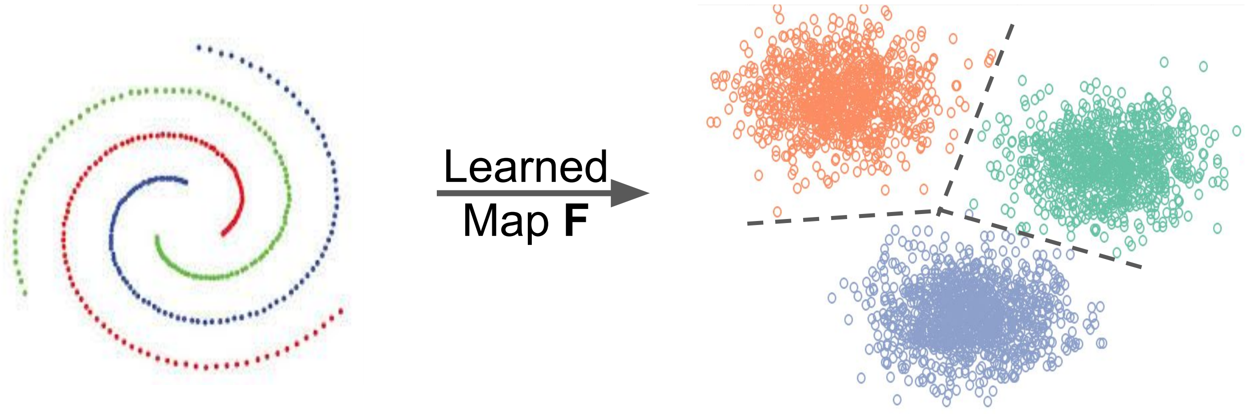
[1]Carnegie Mellon University [2]Toyota Technological Institute at Chicago
[3]University of Waterloo

Nina Balcan[1], Avrim Blum[2], **Dravy Sharma[2]**, Hongyang Zhang[3]

Motivation

Often classification involves embedding inputs $\mathcal{X}$ into feature space $\mathcal{F}$, where simple classifiers (e.g. linear boundaries) work.



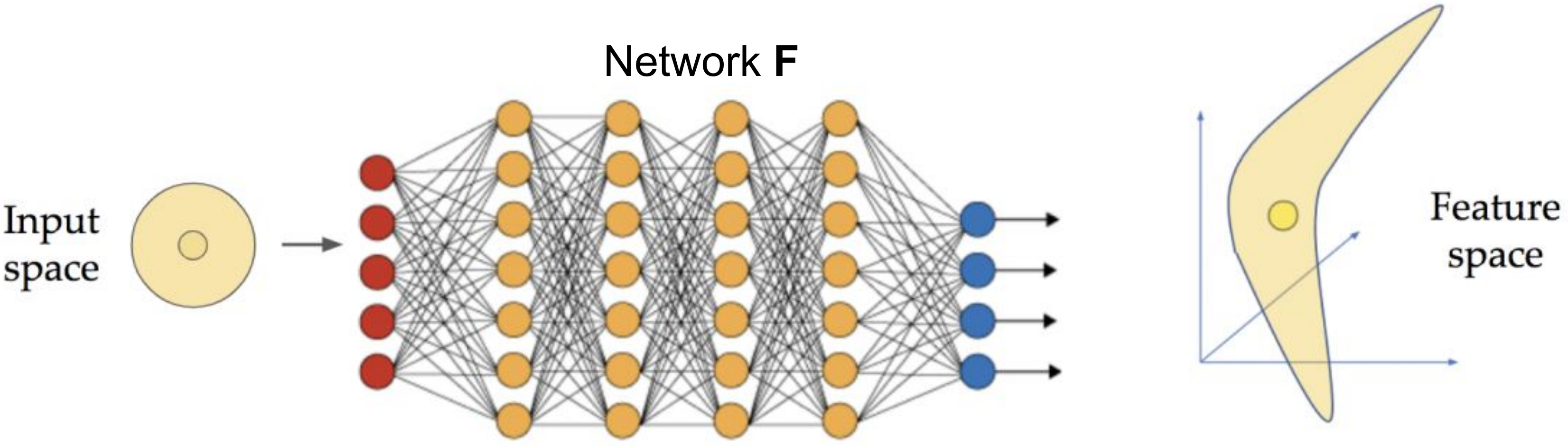
Input space, $\mathcal{X}$

Feature space, $\mathcal{F}$

Motivation

These embeddings are often non-Lipschitz, i.e. small movements in the input space can cause large movements in feature/output space.

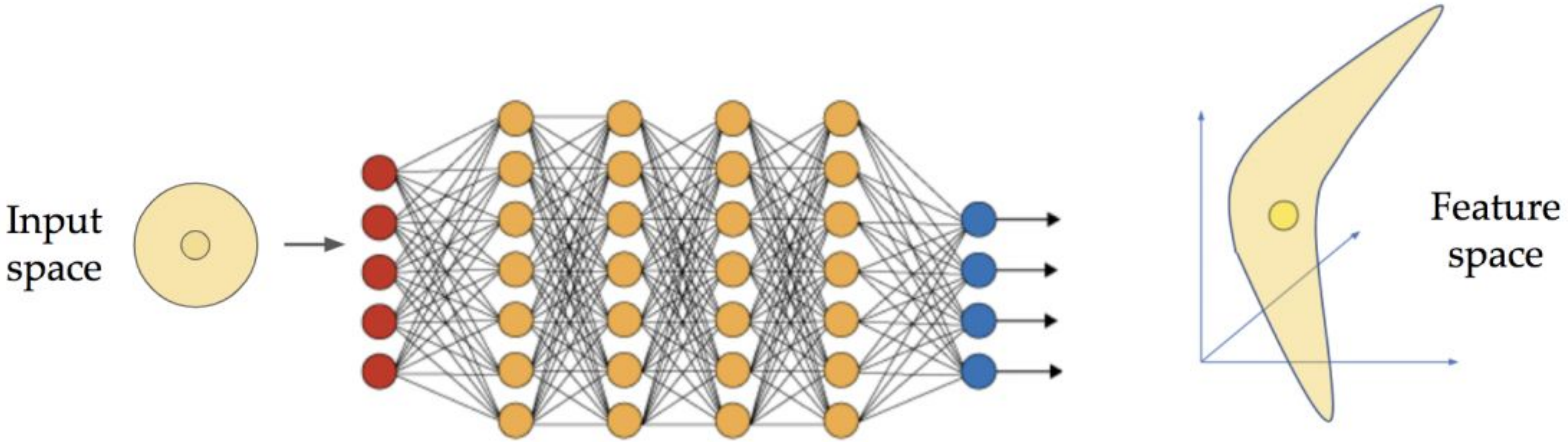
Exceptions: Hein et al. (NeurIPS 2017), Tsuzuku et al. (NeurIPS 2018)



Motivation

How to model non-Lipschitzness of network **F**?

- Intuitively, the adversary can make "large" movements in "some" directions **in the *feature space***.

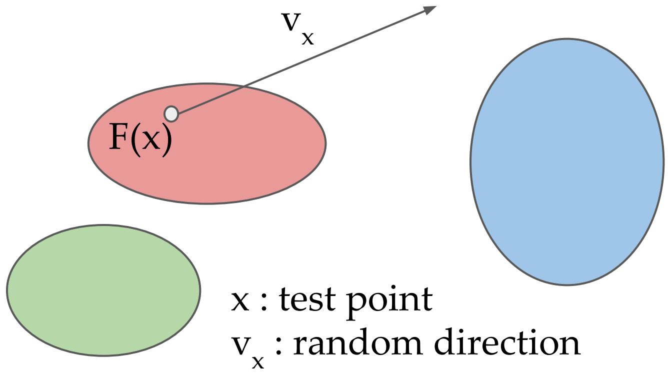


Motivation

How to model non-Lipschitzness of network **F**?

- Intuitively, the adversary can make “large” movements in “some” directions **in the *feature space***.

Abstraction: Model as ‘arbitrarily large’ movements (non-Lipschitz!) **in the *feature space***, but in ‘random’ directions.



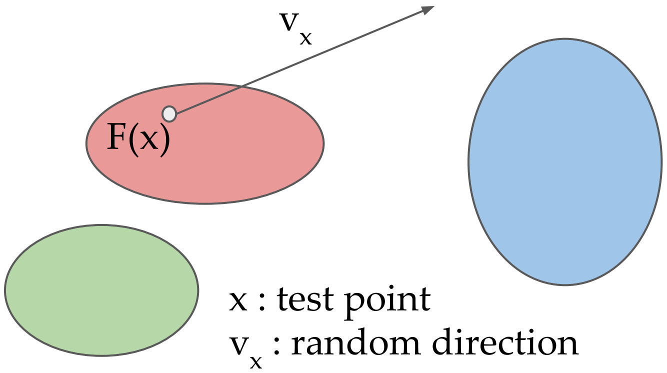
Motivation

How to model non-Lipschitzness of network **F**?

- Intuitively, the adversary can make “large” movements in “some” directions **in the *feature space***.

Abstraction: Model as ‘arbitrarily large’ movements (non-Lipschitz!) **in the *feature space***, but in ‘random’ directions.

‘large movements’ (non-Lipschitz F) in $\mathcal{F}$ +
all directions ⇒ adversary always wins!



Motivation

How to model non-Lipschitzness of network **F**?

- Intuitively, the adversary can make “large” movements in “some” directions in the *feature space*.

Abstraction: Model as ‘arbitrarily large’ movements (non-Lipschitz!) in the *feature space*, but in ‘random’ directions.

- Can be thought of as a ‘**smoothed analysis**’ [Spielman and Teng 2001].

Motivation

How to model non-Lipschitzness of network **F**?

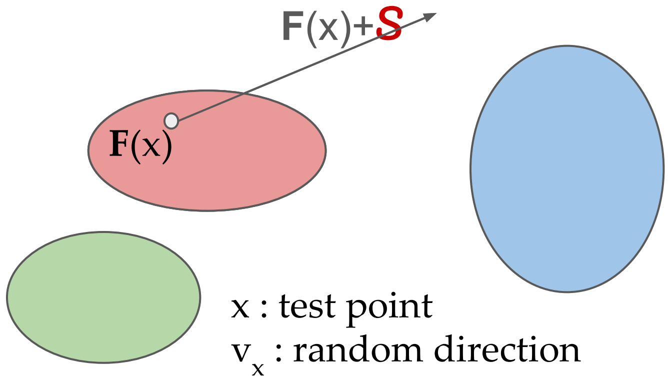
- Intuitively, the adversary can make “large” movements in “some” directions in the *feature space*.

Abstraction: Model as ‘arbitrarily large’ movements (non-Lipschitz!) in the *feature space*, but in ‘random’ directions.

- Can be thought of as a ‘**smoothed analysis**’ [Spielman and Teng 2001].
- We even allow the adversary to choose *any smooth distribution* over the directions.

Model

- Assume feature space $\mathcal{F}$ is n_1-dimensional.
- Sample a uniformly random n_2-dimensional affine subspace $\mathcal{S}$ of $\mathcal{F}$.
- Given test point x, the adversary can perturb **F**(x) to any point in **F**(x)+$\mathcal{S}$

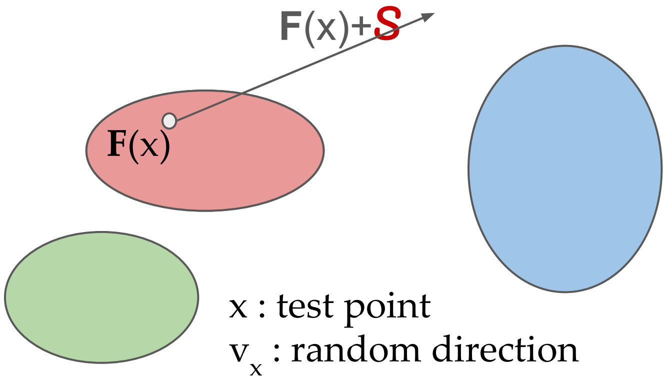


Model

- Assume feature space $\mathcal{F}$ is n_1-dimensional.
- Sample a uniformly random n_2-dimensional affine subspace $\boldsymbol{S}$ of $\mathcal{F}$.
- Given test point x, the adversary can perturb **F**(x) to any point in **F**(x)+$\boldsymbol{S}$

Usual/natural loss

$$\mathcal{L} = \mathbb{E}\left[\ell\,(y, h(x))\right];\ (x, y) \sim \mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}$$

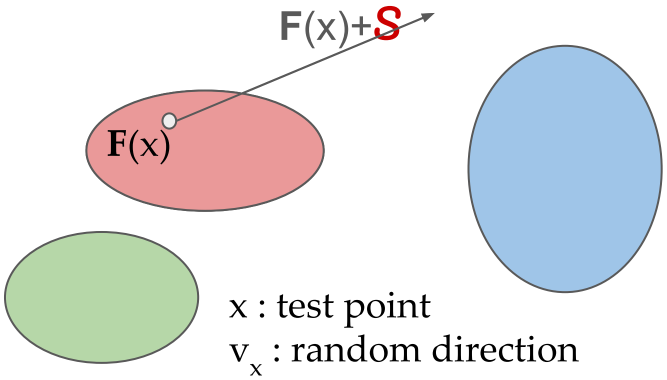


Model

- Assume feature space $\mathcal{F}$ is n_1-dimensional.
- Sample a uniformly random n_2-dimensional affine subspace $\mathcal{S}$ of $\mathcal{F}$.
- Given test point x, the adversary can perturb **F**(x) to any point in **F**(x)+$\mathcal{S}$

Adversarial loss

$$\mathcal{L}_{\mathbf{A}} = \mathbb{E}\left[\ell\,(y, h(\mathbf{A}(x)))\right];\ (x, y) \sim \mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}$$



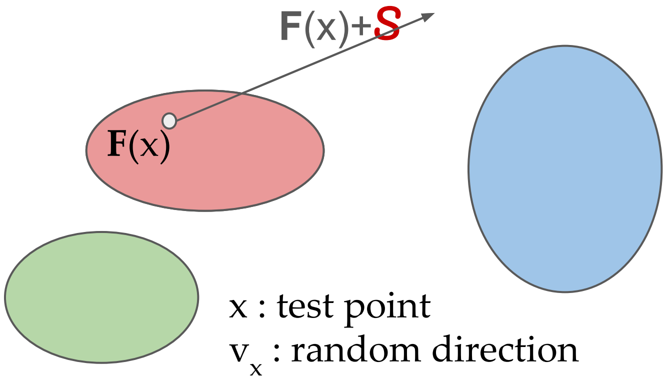
Model

- Assume feature space $\mathcal{F}$ is n_1-dimensional.
- Sample a uniformly random n_2-dimensional affine subspace $\mathcal{S}$ of $\mathcal{F}$.
- Given test point x, the adversary can perturb **F**(x) to any point in **F**(x)+$\mathcal{S}$

Adversarial loss with 'abstain' option

$\mathcal{L}_A = \mathbb{E}\left[\ell\,(y, h(\mathbf{A}(x) \neq \perp))\right]$; $(x, y) \sim \mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

Also want $\mathcal{P}_{\mathcal{D}}(h(x) = \perp)$ is small



Summary of results

Can non-Lipschitz networks be robust?

Worst-case adversary, or classifier without abstention ⇒ NO!

Smoothed adversary ⇒ Possible with abstention!

How?

Threshold-equipped nearest neighbor **in the feature space**

Threshold for abstention may be set using a **data-driven approach**

On real datasets…

Contrastive sampling: A recently popular technique, which minimizes the ratio of intra-class and inter-class distances in training

Consider 1D adversary