Adversarial Robustness of Sparse Local Lipschitz Predictors

Ramchandran Muthukumar

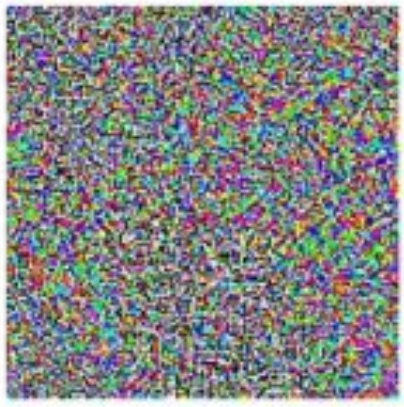
January 17, 2023

Adversarial Robustness

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Two central questions of Robustness

- *Certified Robustness* : What is the minimal size of an adversarial perturbation for a predictor h at input $\mathbf{x}$.

- *Robust Generalization* : When will a predictor h learnt on a training data S_T generalize to corrupted unseen data ?

Our Contribution

- Sensitivity of functions under structural invariance.
- Understanding robust properties of neural networks.

Preliminary Notation

- Input space : $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 \leq 1\}$
- Output space : $\mathcal{Y} := \{1, \ldots, C\}$.
- Perturbation Space : $\mathcal{B}_\nu := \{\boldsymbol{\delta} \in \mathbb{R}^d, \|\boldsymbol{\delta}\|_2 \leq \nu\}$
- Data Distribution : $\mathcal{D}_\mathcal{Z} := \mathcal{D}_\mathcal{X} \times \mathcal{D}_\mathcal{Y}$ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- Training sample (i.i.d): $\mathtt{S}_T := \{\mathbf{z}_i\}_{i=1}^m = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$
- Hypothesis class : $\mathcal{H} : \mathcal{X} \to \mathbb{R}^C$ with embedded norm $\|\cdot\|_\mathcal{H}$.

Notation - Representation-Linear Hypothesis

- We only consider *representation-linear hypothesis classes*.

$$\mathcal{H} := \{h_{\mathbf{A},\mathbf{W}}(\mathbf{x}) := \mathbf{A}\Phi_{\mathbf{W}}(\mathbf{x}),\ \forall\ (\mathbf{A},\mathbf{W}) \in \mathcal{A} \times \mathcal{W}\}\,.$$

  Here, $\Phi_{\mathbf{W}}$ is a representation map and $\mathbf{A}$ is a classification weight.

Notation - Representation-Linear Hypothesis

- We only consider *representation-linear hypothesis classes*.

$$\mathcal{H} := \{h_{\mathbf{A},\mathbf{W}}(\mathbf{x}) := \mathbf{A}\Phi_{\mathbf{W}}(\mathbf{x}),\ \forall\ (\mathbf{A},\mathbf{W}) \in \mathcal{A} \times \mathcal{W}\}\,.$$

  Here, $\Phi_{\mathbf{W}}$ is a representation map and $\mathbf{A}$ is a classification weight.

- Example : A feedforward neural networks with K hidden layers has the representation map $\Phi^{[K]}$,

$$\Phi^{[K]}(\mathbf{x}) := \sigma\left(\mathbf{W}^{K}\sigma\left(\mathbf{W}^{K-1}\cdots\sigma\left(\mathbf{W}^{1}\mathbf{x}+\mathbf{b}^{1}\right)\cdots+\mathbf{b}^{K-1}\right)+\mathbf{b}^{K}\right).$$

Sensitivity

- **Global Lipschitzness** : A constant L_{inp}, for all $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$ and $h \in \mathcal{H}$, we have that

$$\|h(\tilde{\mathbf{x}}) - h(\mathbf{x})\|_2 \leq \mathsf{L}_{\text{inp}} \|\tilde{\mathbf{x}} - \mathbf{x}\|_2$$

- **Local Lipschitzness** : A radius function r_{inp} and a Lipschitz scale function l_{inp} such that,

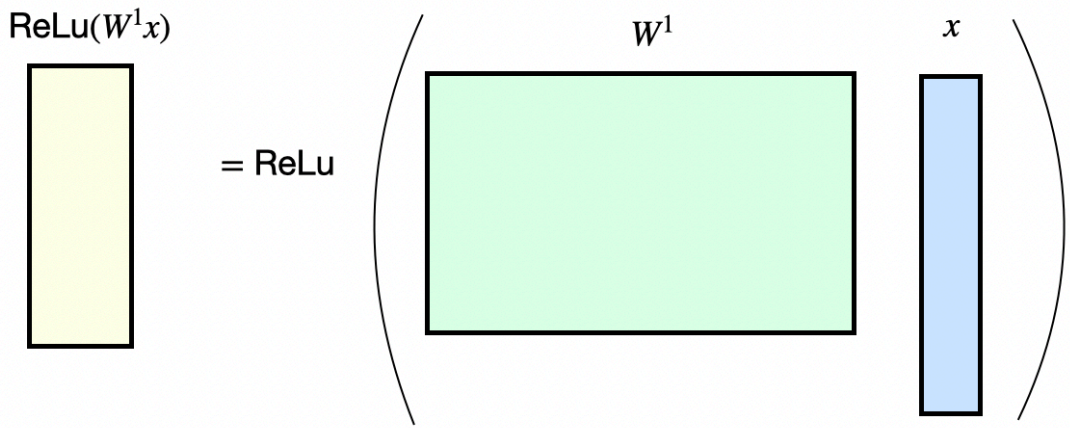
$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq r_{\text{inp}}(\mathbf{x}) \implies \|h(\tilde{\mathbf{x}}) - h(\mathbf{x})\|_2 \leq l_{\text{inp}}(\mathbf{x}) \|\tilde{\mathbf{x}} - \mathbf{x}\|_2 .$$

- If there is a structural property at a predictor output $h(\mathbf{x})$, within what radius can we gaurantee that $h(\tilde{\mathbf{x}})$ retains the property

- A structural property for neural networks - activation states of neurons in each layer.

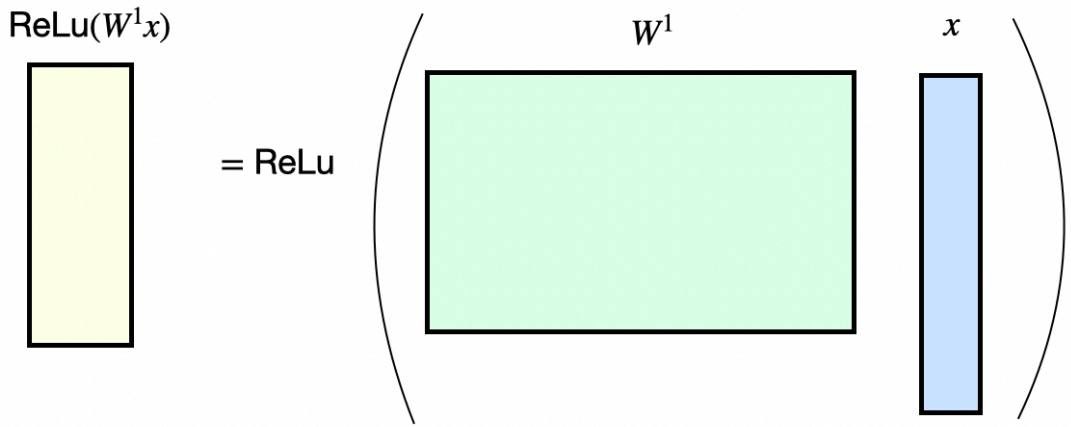
Motivation - Feedforward layers

For feedforward networks, each layer is a feed-forward map $\Phi^{(k)}(\mathbf{t}) := \sigma(\mathbf{W}^k\mathbf{t})$.



Motivation - Feedforward layers

For feedforward networks, each layer is a feed-forward map $\Phi^{(k)}(\mathbf{t}) := \sigma(\mathbf{W}^k\mathbf{t})$.



ReLu induces an **activation pattern** in the output of each layer $\Phi^{(k)}(\mathbf{t})$. We denote by $\mathcal{J}^k(\mathbf{t})$ and $\mathcal{I}^{(k)}$ the true support and co-support of the layer output.

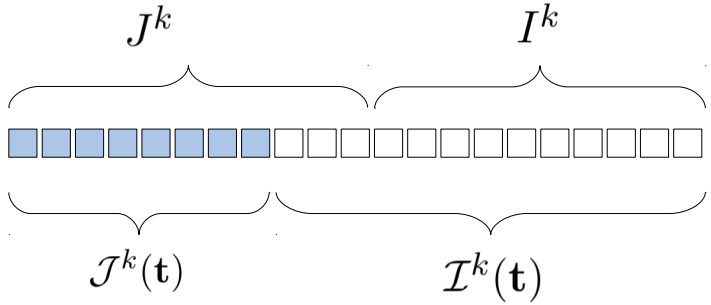


Figure: Illustration of the sets $\mathcal{J}^k(\mathbf{t})$, $\mathcal{I}^k(\mathbf{t})$, as well as I^k and J^k, for a given intermediate input $\sigma(\mathbf{W}^k\mathbf{t} + \mathbf{b}^k)$. Colored squares represent non-zero elements, ordered here without loss of generality.

Motivation : Effect of ReLu

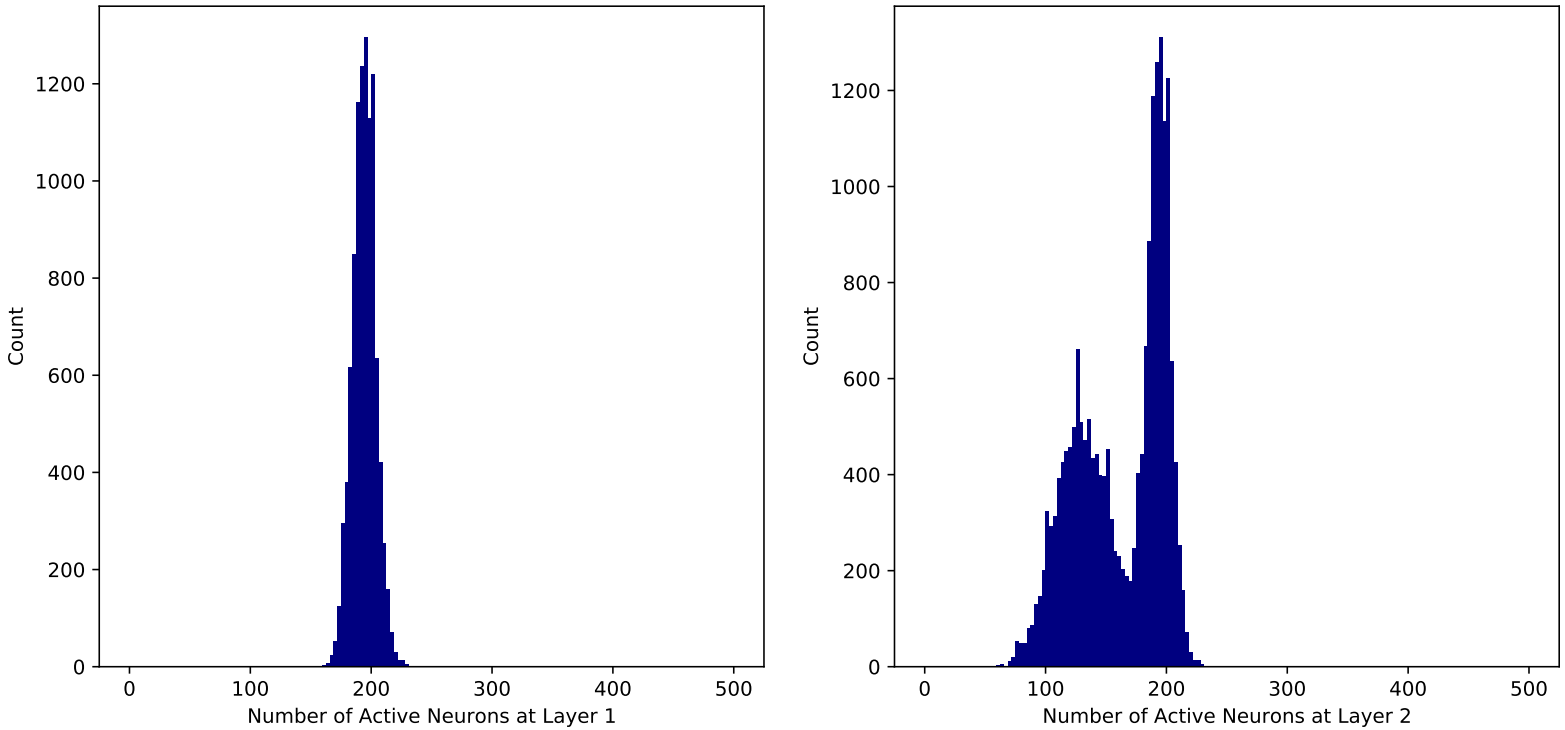
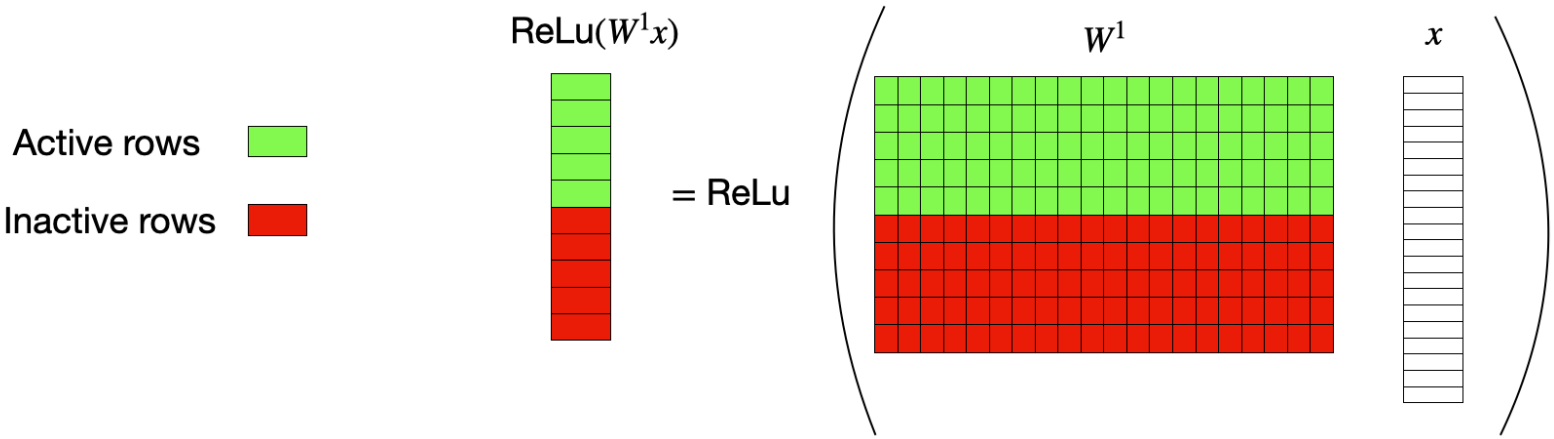


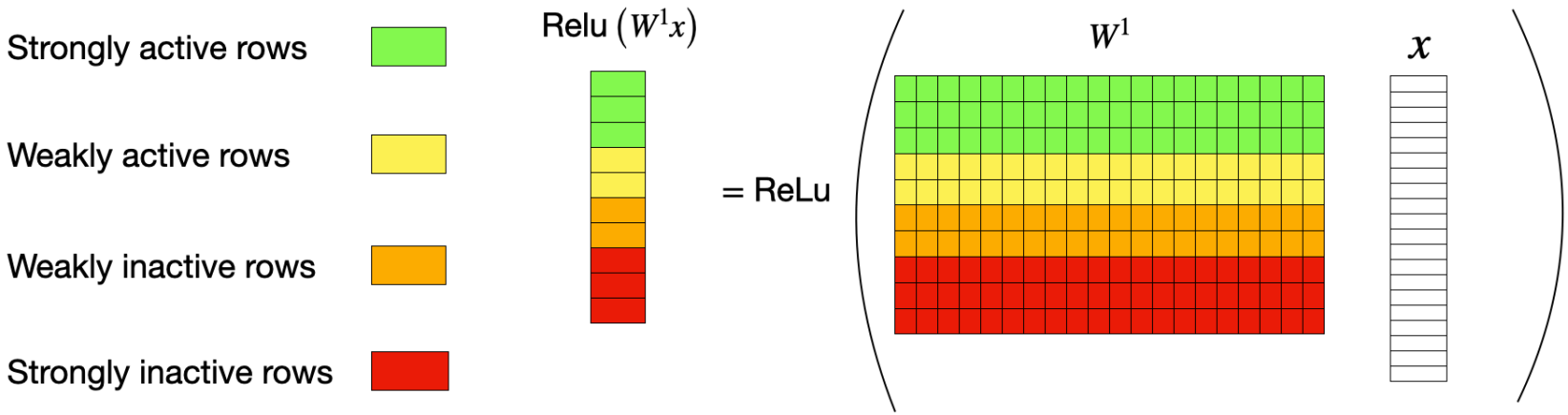
Figure: Distribution of neuron activity (size of $\mathcal{J}^k(\mathbf{t})$) in each layer k of a network trained on MNIST. At each layer only 40 percent of the neurons are activated.

Motivation - Effect of ReLu



Activation states are the result of interaction between rows of $\mathbf{W}^1$ and input $\mathbf{x}$.

Motivation - Effect of ReLu



For bounded perturbations, the strongly inactive rows remain inactive.

Sparse Local Lipschitz (SLL)

A representation map Φ is *SLL w.r.t inputs* if at **each** input $\mathbf{x} \in \mathcal{X}$ and sparsity level $s \in \mathfrak{S}$, there exists[1]

- A stable inactive index set $I(\mathbf{x}, s)$ of size s for the representation $\Phi(\mathbf{x})$
- A sparse local radius function $r_{\text{inp}} : \mathcal{X} \times \mathfrak{S} \to \mathbb{R}^{\geq 0}$
- A sparse local Lipschitz scale function $l_{\text{inp}} : \mathcal{X} \times \mathfrak{S} \to \mathbb{R}^{\geq 0}$

such that for any perturbation $\boldsymbol{\delta}$,

$$\|\boldsymbol{\delta}\|_2 \leq r_{\text{inp}}(\mathbf{x}, s) \implies \begin{cases} \|\Phi(\mathbf{x}+\boldsymbol{\delta}) - \Phi(\mathbf{x})\|_2 \leq l_{\text{inp}}(\mathbf{x}, s)\, \|\boldsymbol{\delta}\|_2 \\ I(\mathbf{x}, s) \text{ is inactive for } \Phi(\mathbf{x}+\boldsymbol{\delta}). \end{cases}$$

[1] Thus we necessarily only talk of $s \leq p - \|\Phi(\mathbf{x})\|_0$

Sparse Local Lipschitz (SLL)

A representation map Φ is *SLL w.r.t inputs* if at **each** input $\mathbf{x} \in \mathcal{X}$ and sparsity level $s \in \mathfrak{S}$, there exists[1]

- A stable inactive index set $I(\mathbf{x}, s)$ of size s for the representation $\Phi(\mathbf{x})$
- A sparse local radius function $r_{\text{inp}} : \mathcal{X} \times \mathfrak{S} \to \mathbb{R}^{\geq 0}$
- A sparse local Lipschitz scale function $l_{\text{inp}} : \mathcal{X} \times \mathfrak{S} \to \mathbb{R}^{\geq 0}$

such that for any perturbation $\boldsymbol{\delta}$,

$$\|\boldsymbol{\delta}\|_2 \leq r_{\text{inp}}(\mathbf{x}, s) \implies \begin{cases} \|\Phi(\mathbf{x}+\boldsymbol{\delta}) - \Phi(\mathbf{x})\|_2 \leq l_{\text{inp}}(\mathbf{x}, s)\, \|\boldsymbol{\delta}\|_2 \\ I(\mathbf{x}, s) \text{ is inactive for } \Phi(\mathbf{x}+\boldsymbol{\delta}). \end{cases}$$

SLL $\implies$ local sensitivity to perturbation + invariance in representation sparsity pattern

[1] Thus we necessarily only talk of $s \leq p - \|\Phi(\mathbf{x})\|_0$

Feedforward Maps are SLL

Lemma

Any feedforward map, $\Phi(\mathbf{x}) := \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$ *is SLL w.r.t input.*

$$
\begin{aligned}
I(\mathbf{x}, s) &:= \underset{\substack{I \subseteq \mathcal{I}(\mathbf{x}), \\ |I| = s}}{\operatorname{argmax}} \ \min_{i \in I} \ \frac{|\mathbf{w}_i \mathbf{x} + \mathbf{b}_i|}{\|\mathbf{w}_i\|_2}, \\
r_{\mathrm{inp}}(\mathbf{x}, s) &:= \min_{i \in I} \ \frac{|\mathbf{w}_i \mathbf{x} + \mathbf{b}_i|}{\|\mathbf{w}_i\|_2}, \\
l_{\mathrm{inp}}(\mathbf{x}, s) &:= \|\mathbf{W}[J, :]\|_2 \, .
\end{aligned}
$$

$J = (I(\mathbf{x}, s))^c$ *is the complement index set.*

Note : The choice of index sets I (and hence the local Lipschitz scale) varies across inputs.

Sparse Local Radius at Layer k

For the feedforward map $\Phi^{(k)}$, the strongly inactive index set $I^k \subset \mathcal{I}^k(\mathbf{t})$ is uniquely identified at layer input $\mathbf{t}$ and sparsity level $\mathbf{s}^{(k)}$.

To compute I^k we sort the normalized pre-activation vector $\mathbf{q}^k := \left[\frac{\mathbf{w}_i^k \mathbf{t} + \mathbf{b}_i^k}{\|\mathbf{w}_i^k\|_2} \right]_{i=1}^{d^k}$.

Sparse Local Radius at Layer k

For the feedforward map $\Phi^{(k)}$, the strongly inactive index set $I^k \subset \mathcal{I}^k(\mathbf{t})$ is uniquely identified at layer input $\mathbf{t}$ and sparsity level $\mathbf{s}^{(k)}$.

To compute I^k we sort the normalized pre-activation vector $\mathbf{q}^k := \left[\frac{\mathbf{w}_i^k \mathbf{t} + \mathbf{b}_i^k}{\|\mathbf{w}_i^k\|_2}\right]_{i=1}^{d^k}$.

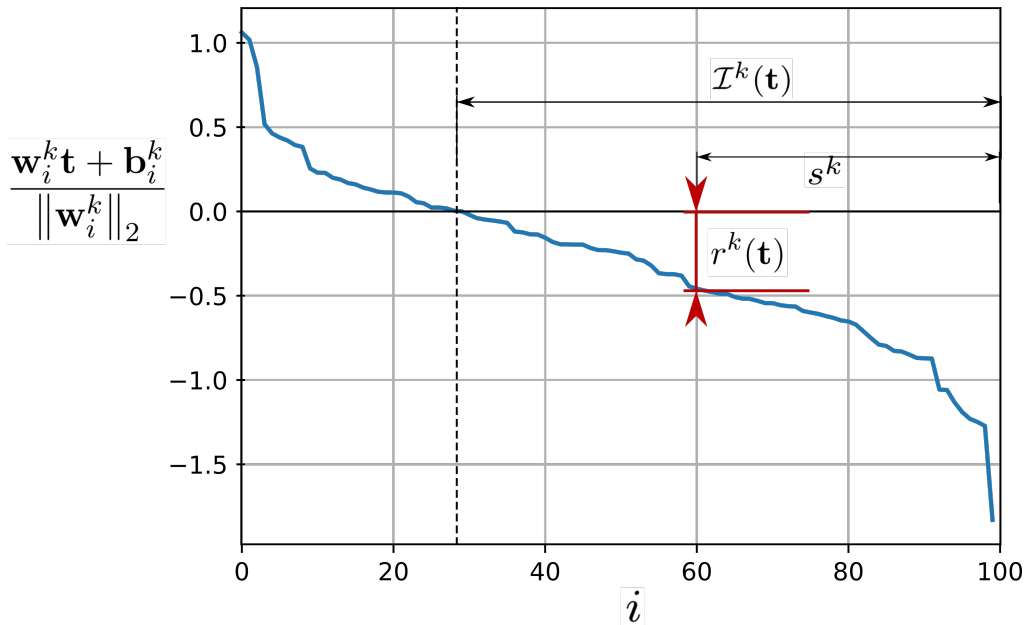
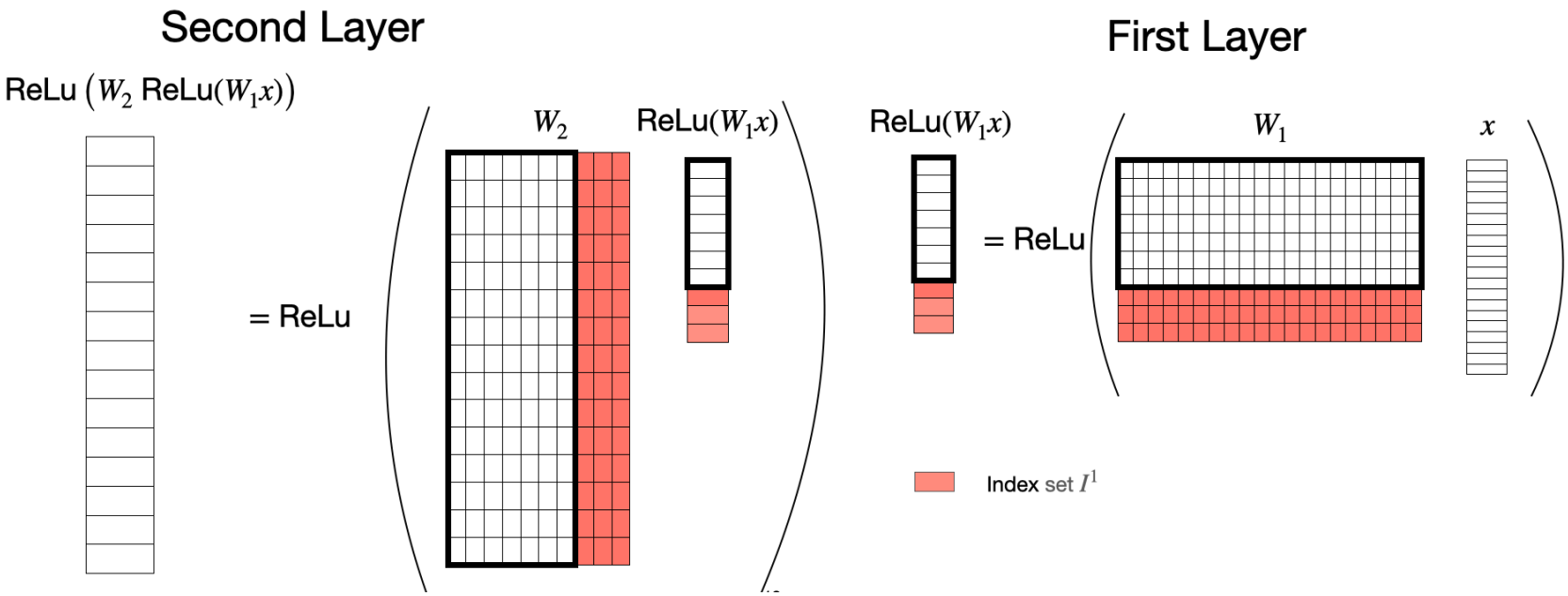


Figure: Illustration of the radius $r_{\text{inp}}^{(k)}(\mathbf{t}, \mathbf{s}^{(k)})$ for the intermediate feedforward representation $\Phi^{(k)}$, given the (sorted) values of the normalized pre-activations.

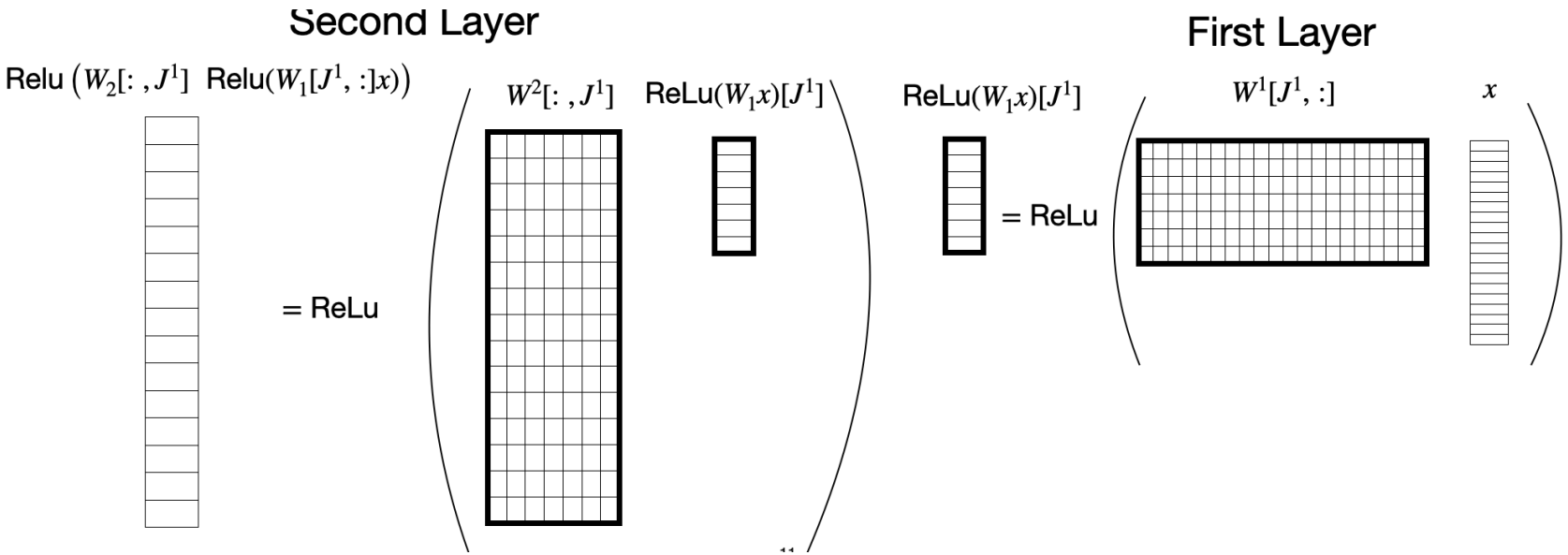
Impact of SLL analysis



Here the index set I^1 is the strongly inactive index set.
The stability of the set I^1 impacts the representation computed in subsequent layers.

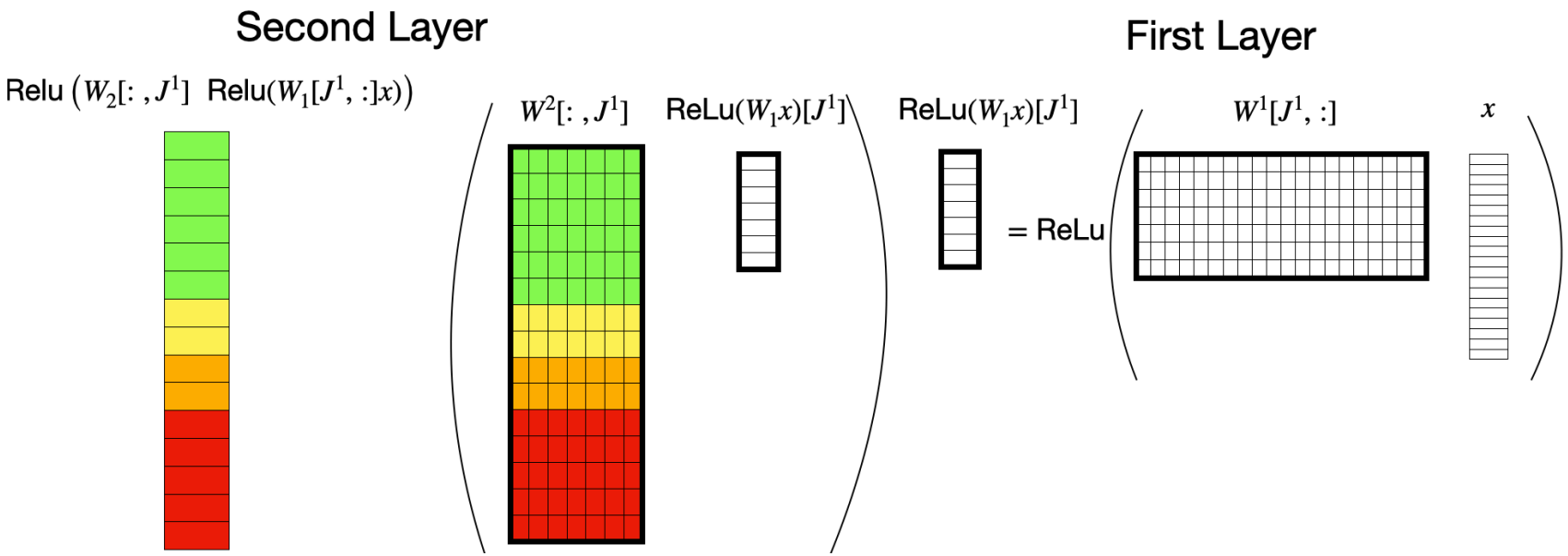
Motivation - Effect of ReLu

Let $J^1 = (I^1)^C$. For perturbations within the sparse local radius $r_{\text{inp}}(\mathbf{x}, s)$, the representation computed is equivalent to a reduced network without I^1 rows in $\mathbf{W}^1$ and I^1 columns in $\mathbf{W}^2$.



Hence the sensitivity in first layers propagates as $\left\|\mathbf{W}^1[J^1, :]\right\|_2)$

Motivation - Effect of ReLu



For the second layer there is sparsity pattern in the outputs as well as a sparsity pattern in the original layer input. We can propagate the same analysis.

Composition of SLL maps is SLL

Consider K intermediate layer representation maps $\Phi^{(k)}$ for $1 \leq k \leq K$, which are then composed to obtain $\Phi^{[K]}$,

$$\Phi^{[K]}(\mathbf{x}) := \Phi^{(K)} \circ \Phi^{(K-1)} \circ \cdots \circ \Phi^{(1)}(\mathbf{x}).$$

Lemma

Assume each $\Phi^{(k)}$ is SLL w.r.t. inputs with $r_{\text{inp}}^{(k)}$ and $l_{\text{inp}}^{(k)}$.

The composed maps (upto layer k) $\Phi^{[k]}$ are also SLL with radius $r_{\text{inp}}^{[k]}$ and Lipschitz scale $l_{\text{inp}}^{[k]}$ given by[2]

$$r_{\text{inp}}^{[k]}\left(\mathbf{x},\ \mathbf{s}^{[k]}\right) := \min_{1 \leq n \leq k} \frac{r_{\text{inp}}^{(n)}\left(\Phi^{[n-1]}(\mathbf{x}),\ \mathbf{s}^{(n)}\right)}{l_{\text{inp}}^{[n-1]}\left(\mathbf{x},\ \mathbf{s}^{[n-1]}\right)}$$

$$l_{\text{inp}}^{[k]}\left(\mathbf{x},\ \mathbf{s}^{[k]}\right) := \prod_{n=1}^{k} l_{\text{inp}}^{(n)}\left(\Phi^{[n-1]}(\mathbf{x}),\ \mathbf{s}^{(n)}\right).$$

For any perturbation $\boldsymbol{\delta}$ within $r_{\text{inp}}^{[k]}(\mathbf{x}, \mathbf{s}^{[k]})$, index sets $I^1, I^2, \ldots, I^k$ remain inactive.

---

[2] Here $(s^0, s^1, \ldots, s^K)$ are sparsity levels for each intermediate map, $\mathbf{s}^{(k)} := (s^{k-1}, s^k)$ is the layer-wise input-output sparsity levels and $\mathbf{s}^{[k]} := (s^0, s^k)$ is the cumulative input-output levels.

Reduced Dimensionality of SLL predictors

- The representation $\Phi^{[K]}$ computed by K feedforward layers is SLL with radius $r_{\text{inp}}^{[K]}$ and local Lipschitz scale $l_{\text{inp}}^{[K]}$.

Reduced Dimensionality of SLL predictors

- The representation $\Phi^{[K]}$ computed by K feedforward layers is SLL with radius $r_{\text{inp}}^{[K]}$ and local Lipschitz scale $l_{\text{inp}}^{[K]}$.

- Feedforward neural networks exhibit the reduced dimensionality. For for all perturbations $\tilde{\mathbf{x}}$ within the local radius,

$$
\begin{aligned}
h(\tilde{\mathbf{x}}) &= \mathbf{A}\sigma\left(\mathbf{W}^K\,\sigma\left(\mathbf{W}^{K-1}\cdots\sigma\left(\mathbf{W}^1\,\tilde{\mathbf{x}}+\mathbf{b}^1\right)\cdots+\mathbf{b}^{K-1}\right)+\mathbf{b}^K\right) \\
&= \mathbf{A}_{\text{red}}\sigma\left(\mathbf{W}_{\text{red}}^K\,\sigma\left(\mathbf{W}_{\text{red}}^{K-1}\cdots\sigma\left(\mathbf{W}_{\text{red}}^1\,\tilde{\mathbf{x}}+\mathbf{b}_{\text{red}}^1\right)\cdots+\mathbf{b}_{\text{red}}^{K-1}\right)+\mathbf{b}_{\text{red}}^K\right) \\
&=: h_{\text{red}}(\tilde{\mathbf{x}})
\end{aligned}
$$

where $\mathbf{W}_{\text{red}}^k := \mathbf{W}^k[J^k, J^{k-1}] \in \mathbb{R}^{(d^k-s^k)\times(d^{k-1}-s^{k-1})}$

Reduced Dimensionality of SLL predictors

- The representation $\Phi^{[K]}$ computed by K feedforward layers is SLL with radius $r_{\text{inp}}^{[K]}$ and local Lipschitz scale $l_{\text{inp}}^{[K]}$.

- Feedforward neural networks exhibit the reduced dimensionality. For for all perturbations $\tilde{\mathbf{x}}$ within the local radius,

$$
\begin{aligned}
h(\tilde{\mathbf{x}}) &= \mathbf{A}\sigma\left(\mathbf{W}^K\,\sigma\left(\mathbf{W}^{K-1}\cdots\sigma\left(\mathbf{W}^1\,\tilde{\mathbf{x}}+\mathbf{b}^1\right)\cdots+\mathbf{b}^{K-1}\right)+\mathbf{b}^K\right) \\
&= \mathbf{A}_{\text{red}}\sigma\left(\mathbf{W}_{\text{red}}^K\,\sigma\left(\mathbf{W}_{\text{red}}^{K-1}\cdots\sigma\left(\mathbf{W}_{\text{red}}^1\,\tilde{\mathbf{x}}+\mathbf{b}_{\text{red}}^1\right)\cdots+\mathbf{b}_{\text{red}}^{K-1}\right)+\mathbf{b}_{\text{red}}^K\right) \\
&=: h_{\text{red}}(\tilde{\mathbf{x}})
\end{aligned}
$$

  where $\mathbf{W}_{\text{red}}^k := \mathbf{W}^k[J^k, J^{k-1}] \in \mathbb{R}^{(d^k-s^k)\times(d^{k-1}-s^{k-1})}$

- A naive estimate of the global Lipschitz constant $\prod_{k=1}^{K+1}\left\|\mathbf{W}^k\right\|_2$.

- Reduced local dimensionality $\implies$ it is inefficient to directly compute the Lipschitz constant of the full original network. The local sensitivity scales with depth as $\prod_{k=1}^{K+1}\left\|\mathbf{W}_{\text{red}}^k\right\|_2$ i.e. $\prod_{k=1}^{K+1}\left\|\mathbf{W}^k[J^k, J^{k-1}]\right\|_2$.

Certified Robustness for SLL predictors

Theorem

Let $h(\mathbf{x}) := \mathbf{A}\Phi(\mathbf{x})$ *be a predictor such that the representation map* Φ *is SLL with radius function* r_{inp} *and Lipschitz scale function* l_{inp}*.*
The predicted label $\hat{y}(\mathbf{x})$ *at input* $\mathbf{x}$ *remains unchanged if an adversarial corruption is within the certified radius* $r_{\text{cert}}(\mathbf{x}, s)$*,*

$$r_{SLL}(\mathbf{x}, s) := \min \left\{ r_{\text{inp}}(\mathbf{x}, s), \ \frac{\rho(\mathbf{x})}{2 \left\| \mathbf{A} \right\|_2 \ l_{\text{inp}}(\mathbf{x}, s)} \right\}.$$

Here, $\rho(\mathbf{x})$ *is the classification margin.*

Certified Robustness for SLL predictors

Theorem

Let $h(\mathbf{x}) := \mathbf{A}\Phi(\mathbf{x})$ be a predictor such that the representation map Φ is SLL with radius function r_{inp} and Lipschitz scale function l_{inp}.
The predicted label $\hat{y}(\mathbf{x})$ at input $\mathbf{x}$ remains unchanged if an adversarial corruption is within the certified radius $r_{\text{cert}}(\mathbf{x}, s)$,

$$r_{SLL}(\mathbf{x}, s) := \min \left\{ r_{\text{inp}}(\mathbf{x}, s), \ \frac{\rho(\mathbf{x})}{2 \left\|\mathbf{A}\right\|_2 \ l_{\text{inp}}(\mathbf{x}, s)} \right\}.$$

Here, $\rho(\mathbf{x})$ is the classification margin.

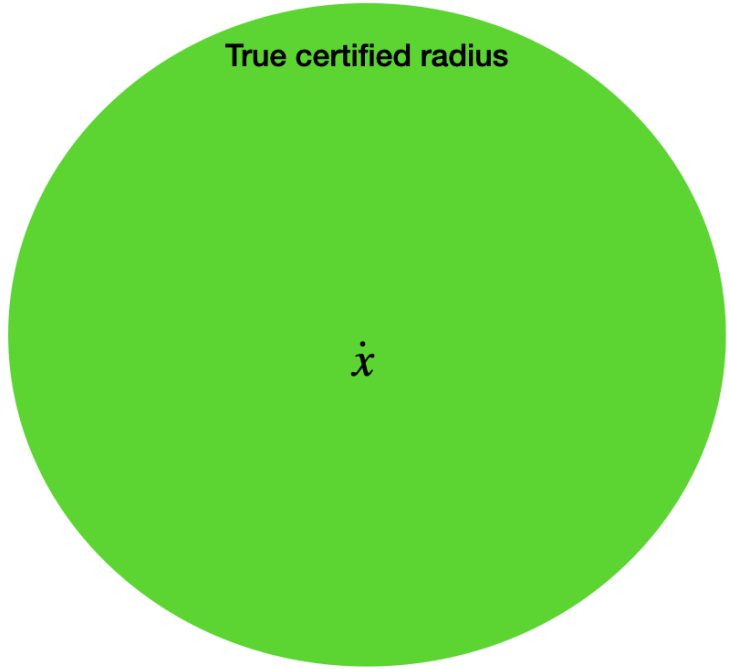
- For depth-$(K+1)$ feedforward networks, the local radius function is $r_{\text{inp}}^{[K]}$ and the local Lipschitz scale function is $l_{\text{inp}}^{[K]}$.
- Local /Global Lipschitz analysis correspond to $s = 0$.
- We can optimize over sparsity levels to get the best certified radius.

True certified radius

$\hat{y}(x) :=$ Label predicted by h on input $\mathbf{x}$.

$$r_{\text{cert}}(\mathbf{x}) := \min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_2$$
$$s.t. \hat{y}(\mathbf{x} + \boldsymbol{\delta}) \neq \hat{y}(\mathbf{x})$$

For all perturbations within $r_{\text{cert}}(\mathbf{x})$, the label remains unchanged.

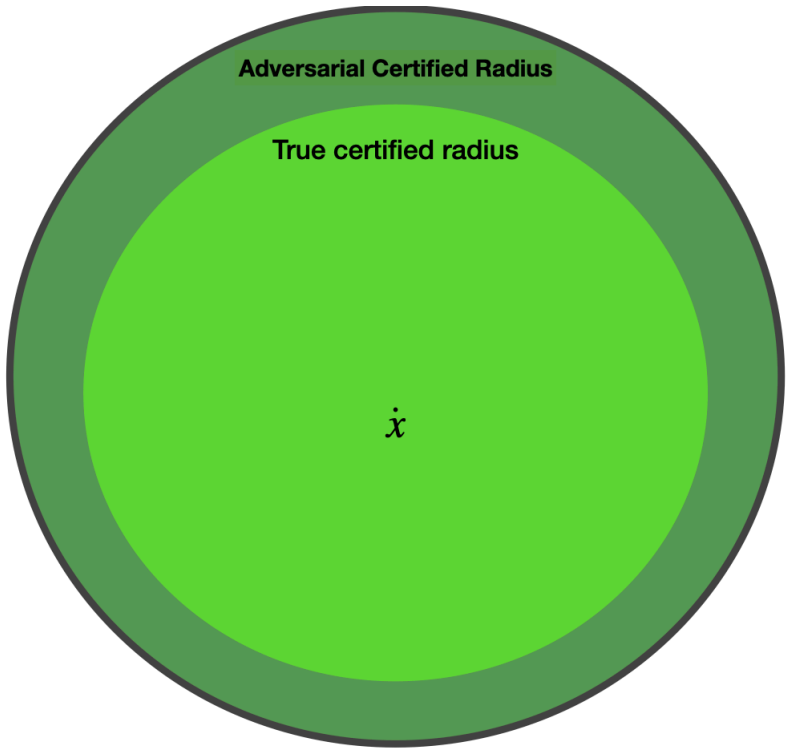


Adversarial upper bound

For any adversarial attack, pick the example with least energy.

$$r_{\text{adv}}(\mathbf{x}) := \min_{\text{adv attacks}} \|\boldsymbol{\delta}\|_2$$
$$s.t. \hat{y}(\mathbf{x} + \boldsymbol{\delta}) \neq \hat{y}(\mathbf{x})$$

Upper bound since PGD doesn't provably converge to optimal perturbation.



Global Lipschitz certificate

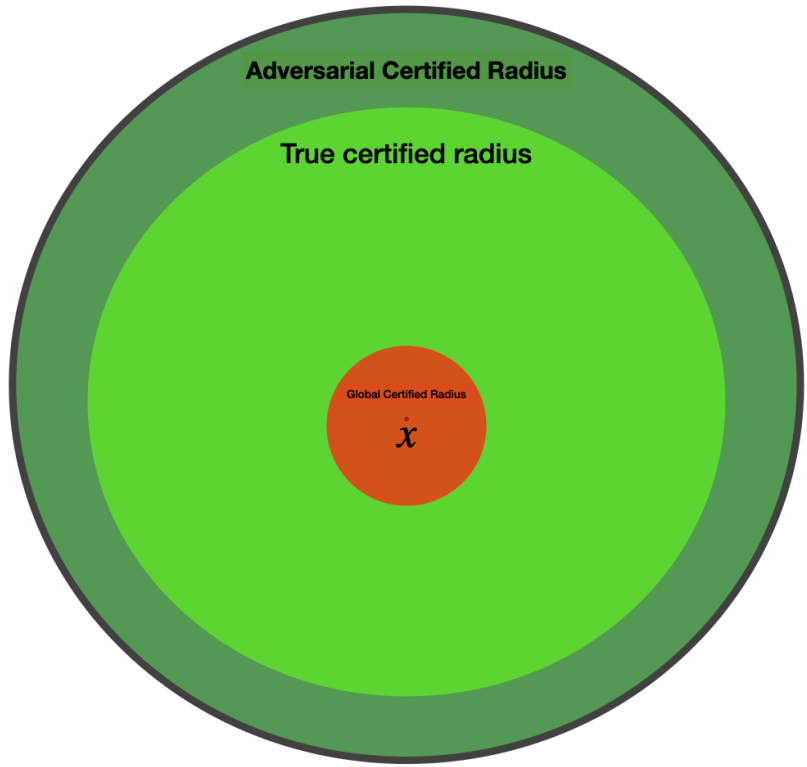
Let $\mathrm{L}_{\mathrm{inp}}$ be the global Lipschitz constant. For any perturbation,

$$\|h(\mathbf{x}+\boldsymbol{\delta}) - h(\mathbf{x})\|_2 \leq \mathrm{L}_{\mathrm{inp}} \|\boldsymbol{\delta}\|_2 .$$

The global certified radius

$$r_{\text{global}}(\mathbf{x}) := \frac{\rho(\mathbf{x})}{\mathrm{L}_{\mathrm{inp}}},$$

ensures perturbations don't cross decision boundaries.



Local Lipschitz certificate

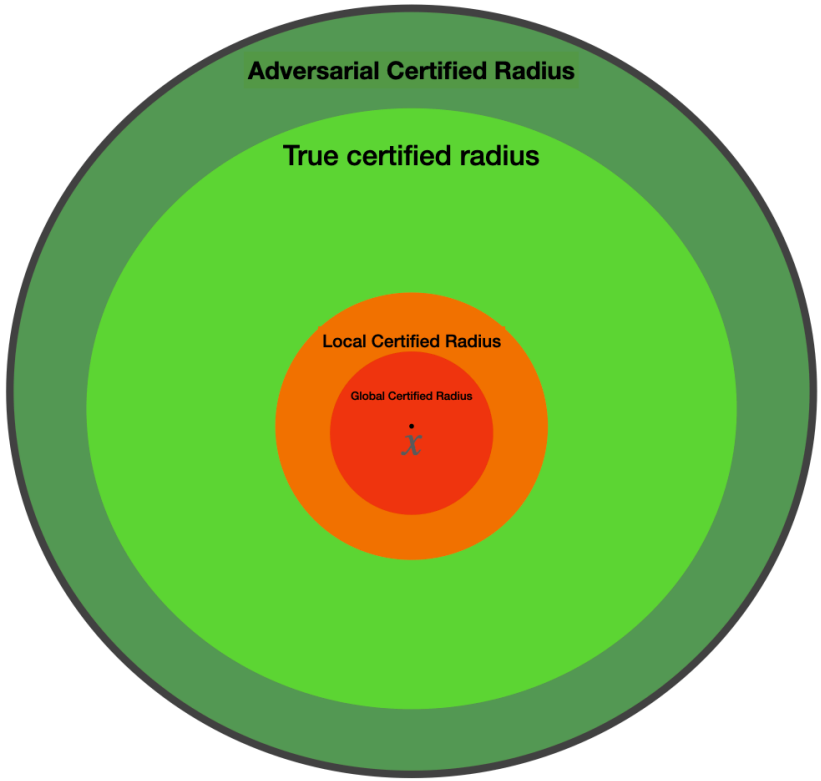
If Φ is local Lipschitz,

$$\|\boldsymbol{\delta}\|_2 \leq r_{\text{inp}}(\mathbf{x})$$
$$\implies \|h(\mathbf{x}+\boldsymbol{\delta}) - h(\mathbf{x})\|_2 \leq \|\mathbf{A}\|_2 \, l_{\text{inp}}(\mathbf{x}).$$

The local certified radius is

$$r_{\text{local}}(\mathbf{x}) := \min \left\{ r_{\text{inp}}(\mathbf{x}), \frac{\rho(\mathbf{x})}{2\,\|\mathbf{A}\|_2 \, l_{\text{inp}}(\mathbf{x})} \right\}$$

ensures perturbations don't exceed local Lipschitz radius or margin in output space.

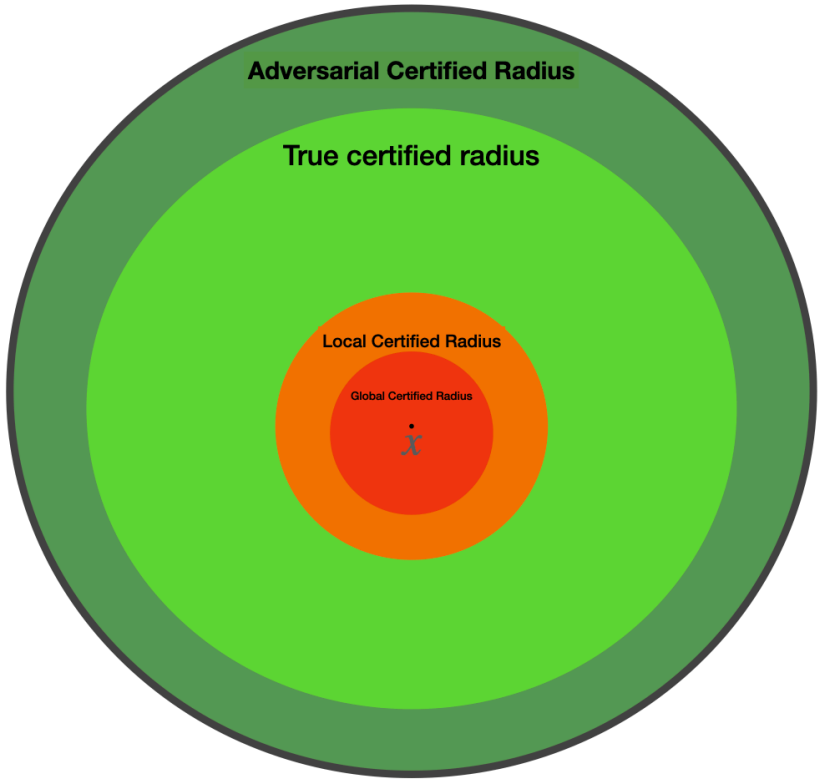


Sparse Local Lipschitz Certificate

The sparse certificate is,

$$r_{SLL}(\mathbf{x}, s) := \min\left\{ r_{\text{inp}}(\mathbf{x}, s),\ \frac{\rho(\mathbf{x})}{2\,\|\mathbf{A}\|_2\ l_{\text{inp}}(\mathbf{x}, s)} \right\}.$$

Equivalent to local Lipschitz analysis for $s = 0$.

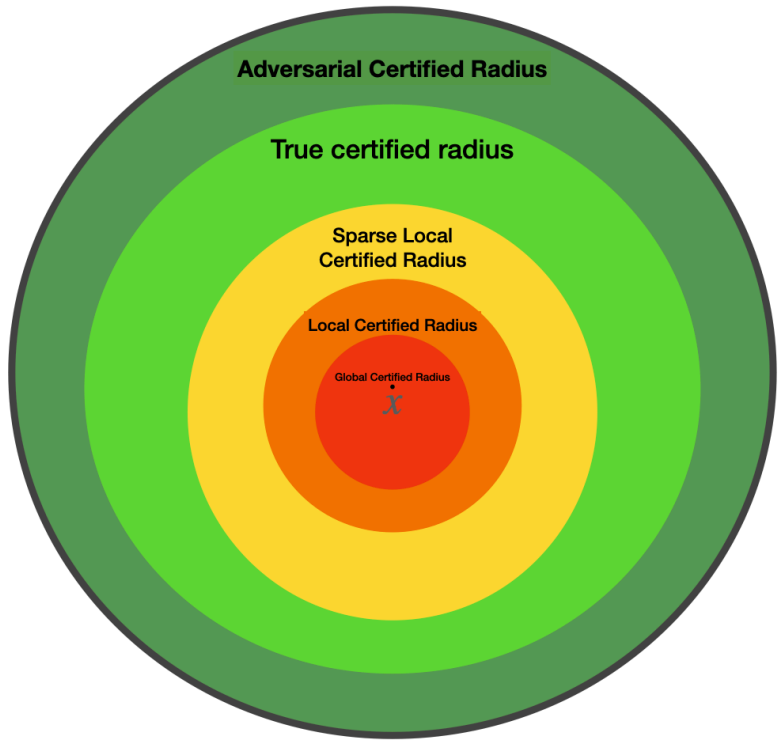


Sparse Local Lipschitz Certificate

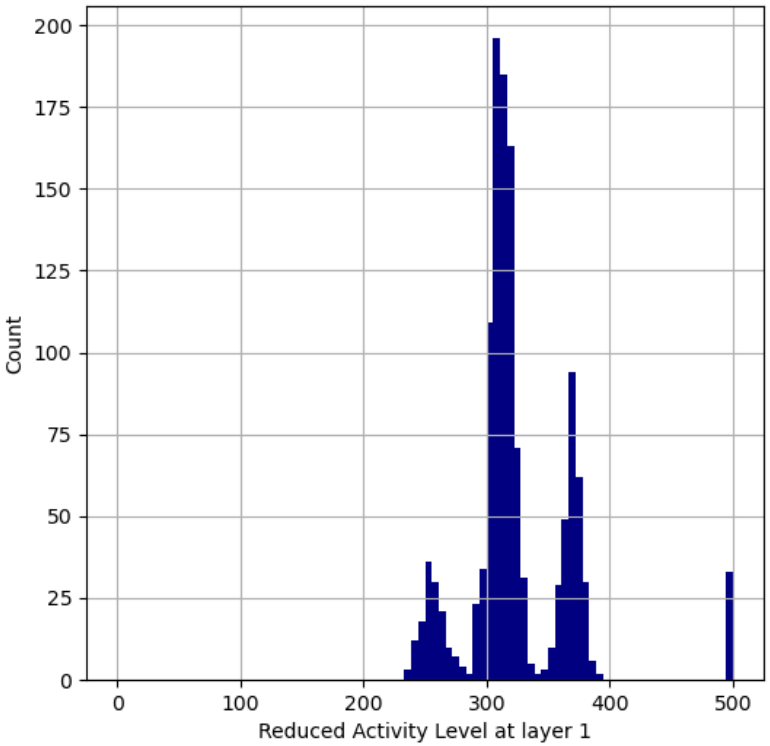
Optimize over sparsity levels for best certificate,

$$
\begin{aligned}
& r_{sparse}(\mathbf{x}) \\
& \quad := \max_{s} r_{SLL}(\mathbf{x}, s) \\
& \quad = \max_{s} \min \left\{ r_{\text{inp}}(\mathbf{x}, s), \ \frac{\rho(\mathbf{x})}{2 \left\| \mathbf{A} \right\|_2 \ l_{\text{inp}}(\mathbf{x}, s)} \right\}.
\end{aligned}
$$

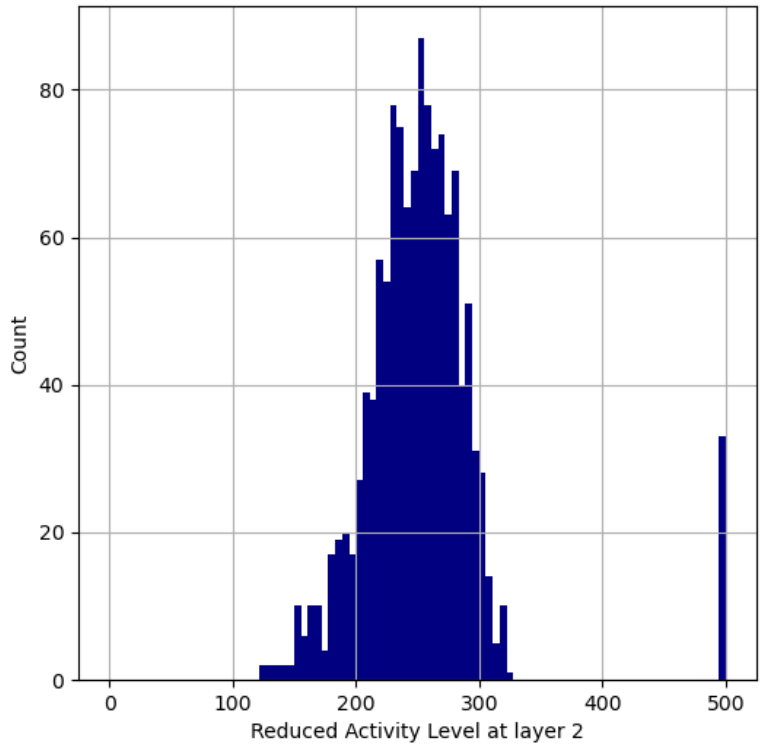
At each $\mathbf{x}$, the optimal sparsity level s^* gives a specific reduced network.



Reduced Dimensionality



(a) Histogram of reduced widths at layer 1



(b) Histogram of reduced widths at layer 2

Figure: For an off-the-shelf trained network h, (a) and (b) represent the distribution of widths of the particular reduced network h_{red} at each input $\mathbf{x}$. The reduced widths at each layer correspond to the choice of **optimal sparsity level**.

Reduced Lipschitz constant

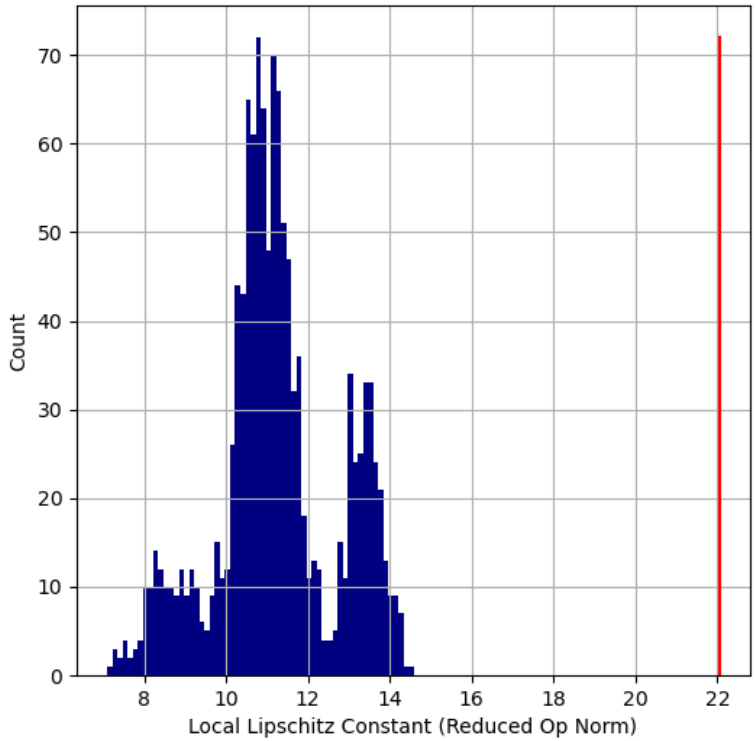
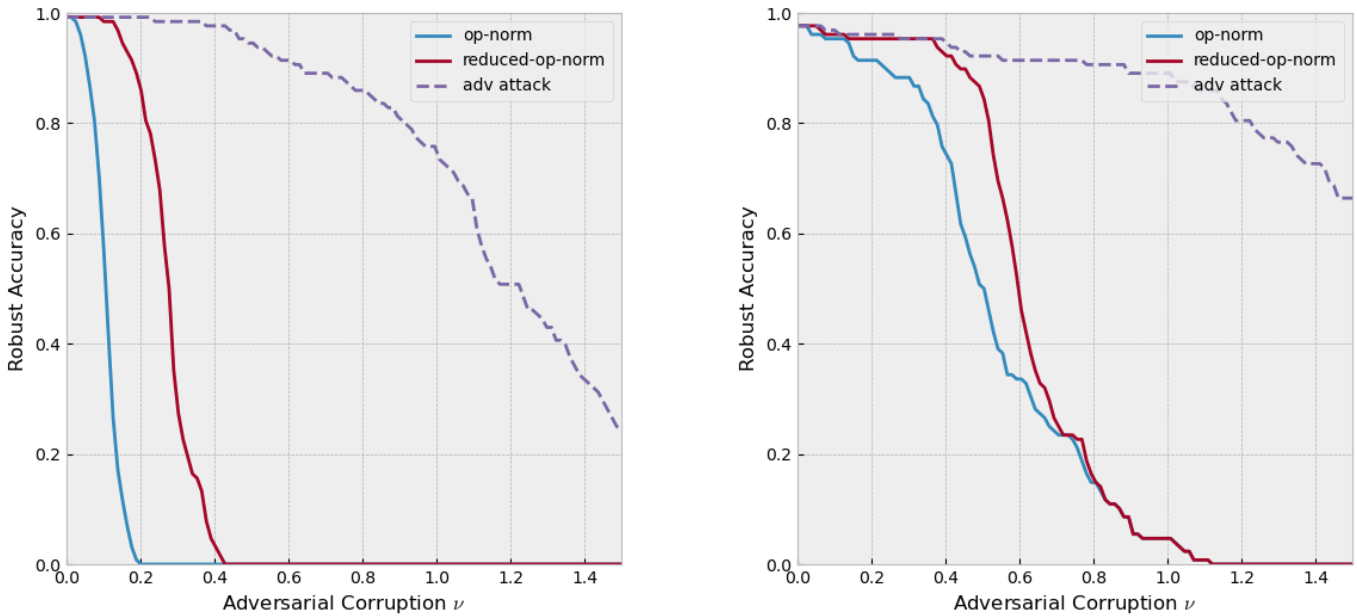


Figure: Histogram of optimal sparse local Lipschitz scale across inputs. At each input, the size of the reduced network corresponds to $\mathbf{s}^*(\mathbf{x})$. The red line marks the naive estimate of global Lipschitz constant.

Certified Robustness for Feed-forward Neural Networks

We plot the *certified accuracy* of a trained predictor using,

- Naive certificate with global Lipschitz constant $= \prod_{k=1}^{K+1} \left\| \mathbf{W}^k \right\|_2$.
- SLL certificate with local Lipschitz constant $= \prod_{k=1}^{K+1} \left\| \mathbf{W}^k [J^k, J^{k-1}] \right\|_2$.
- Heuristic upper bound from common adversarial attacks.



(a) Off-the-shelf

(b) Regularized

Figure: Security curves for feed-forward neural networks on MNIST.

Sparse Local Lipschitz w.r.t Parameters and Inputs

- Analysis can be extended to perturbations to both weights and inputs.
- Sparse local radius again quantifies stability of inactive index sets.
- Similar reduced dimensionality effect for a perturbed input $\tilde{\mathbf{x}}$ and perturbed weight $\hat{\mathbf{W}}$ within local radius.

Robust Generalization Bound for Feedforward Neural Networks

Theorem

With probability at least $(1-\alpha)$ *over the choice of i.i.d training sample* S_T *and unlabeled data* S_U*, for any multi-layered neural network predictor* $h \in \mathcal{H}^{K+1}$ *with parameters* $\{\mathbf{W}^k\}$ *the robust stochastic risk is bounded as,*

$$R_{\text{rob}}(h) - \hat{R}_{\text{rob}}(h) \leq$$

$$\tilde{\mathcal{O}}\left(b\sqrt{\frac{\ln\left(\mathcal{N}\left(\frac{1}{m(K+1)}, \mathcal{H}^{K+1}\right)\right) + \ln(\frac{2}{\alpha})}{2m}} + \frac{\mathsf{L}_{\text{loss}}(1+\nu)}{m} \prod_{k=1}^{K+1} \left\|\mathbf{W}^k\right\|_{2,\infty} \sqrt{1 + \mu_{s^k, s^{k-1}}(\mathbf{W}^k)} \right)$$

Here, $\mathbf{s} = (s^1, \ldots, s^K)$ *is an optimal sparsity level chosen based on* S_T *and* S_U*.* $\mu_{s^k, s^{k-1}}(\mathbf{W}^k)$ *is a reduced babel function and* $\|\mathbf{W}\|_{2,\infty}$ *is the maximal* ℓ_2 *norm of a row in* $\mathbf{W}$*.*

Thank you for attending my talk :)

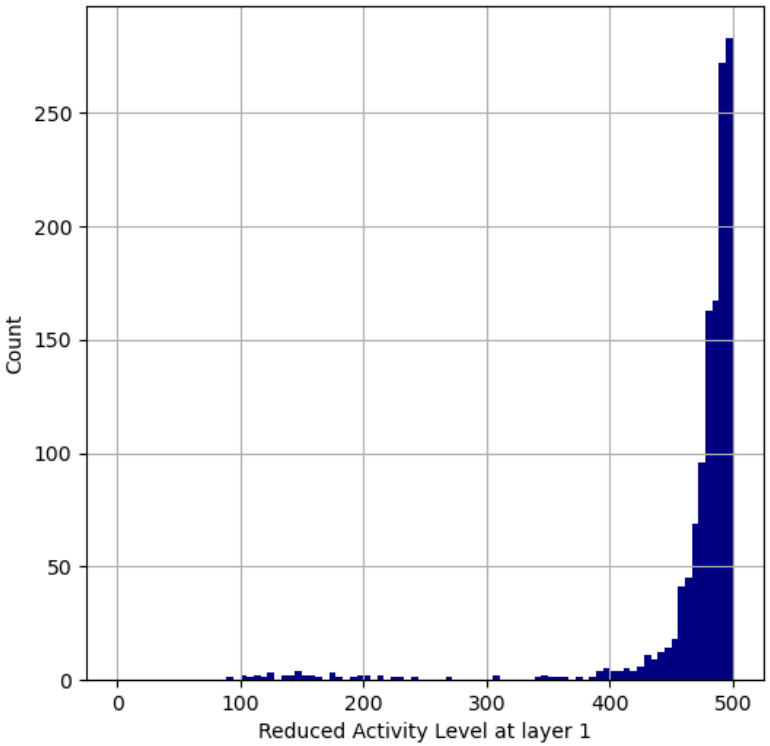
Corollary

Consider a trained depth-$K+1$ feed-forward neural network h.
Let $\mathbf{s} = (s^1, \ldots, s^K)$ be a choice of sparsity levels at each layer
Let $\mathbf{v}^{(k)} := (s^{k-1}, s^k)$ be the corresponding layer-wise input-output sparsity levels.
The predicted label remains unchanged, whenever $\|\boldsymbol{\delta}\|_2 \leq r_{\text{cert}}(\mathbf{x}, \mathbf{s})$, where

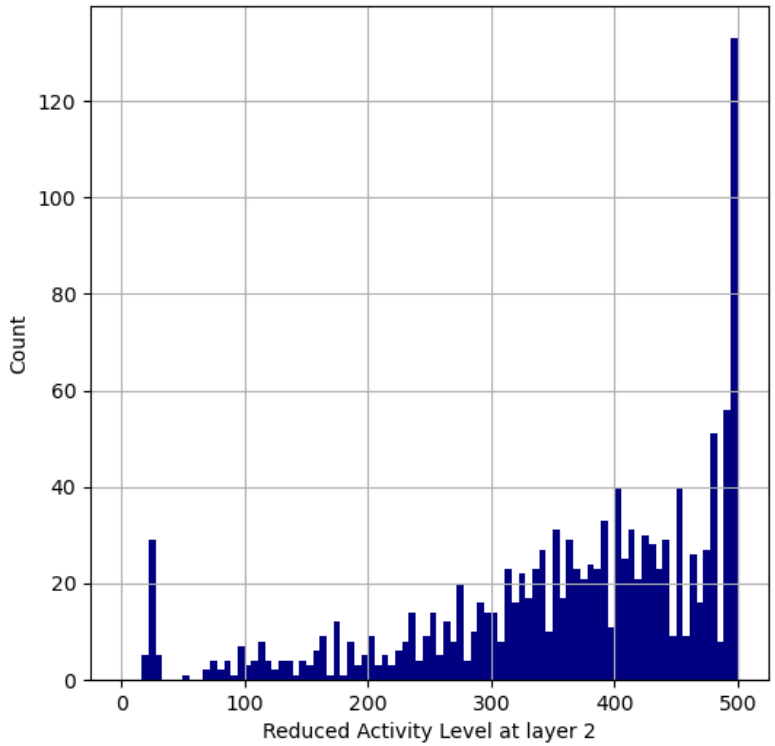
$$r_{\text{cert}}(\mathbf{x}, \mathbf{s}) := \min \left\{ \min_{1 \leq k \leq K} \frac{r_{\text{inp}}^{(k)}(\Phi^{[k-1]}(\mathbf{x}), \mathbf{v}^{(k)})}{\prod_{n=1}^{k} \left\| \mathcal{P}_{J^n, J^{n-1}}(\mathbf{W}^n) \right\|_2}, \frac{\rho(\mathbf{x})}{2 \|\mathbf{A}\|_2 \prod_{k=1}^{K} \left\| \mathcal{P}_{J^k, J^{k-1}}(\mathbf{W}^k) \right\|_2} \right\}$$

Here, $r_{\text{inp}}^{(k)}$ is the local radius for the feedforward map at layer k.
and $\mathcal{P}_{J^k, J^{k-1}}(\mathbf{W}^k)$ is the activated weight at layer k.

Reduced Widths for regularized networks



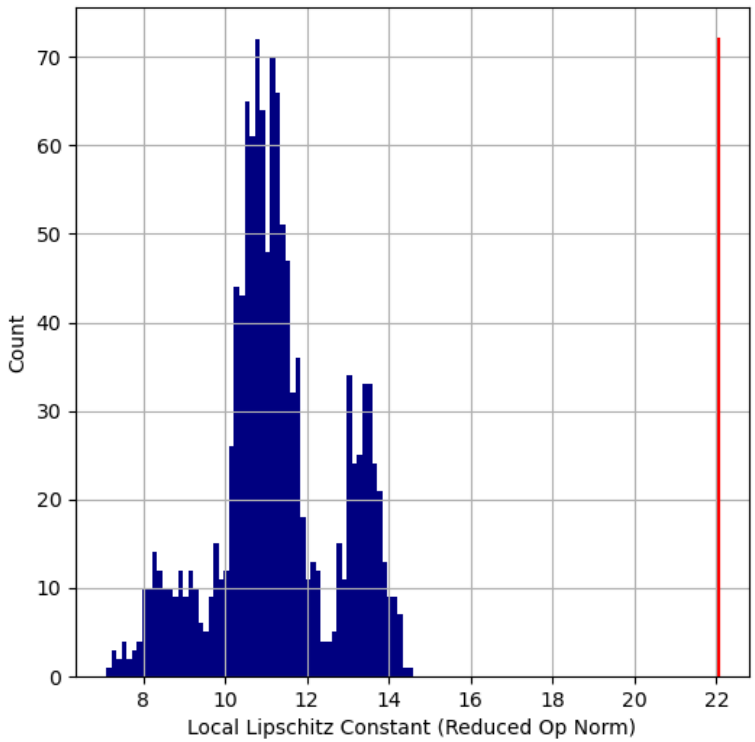
(a) Histogram of reduced widths at layer 1



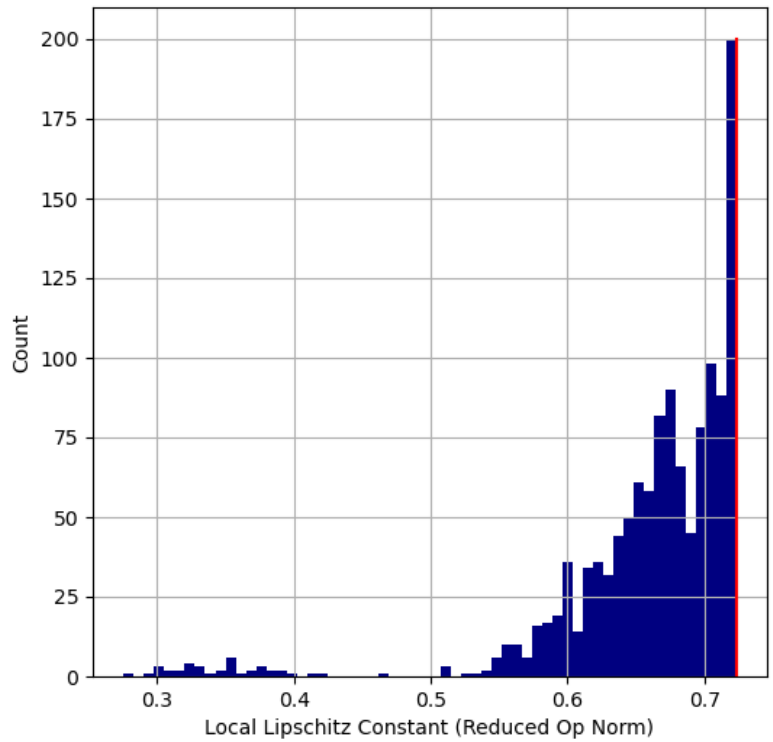
(b) Histogram of reduced widths at layer 2

Figure: For an original regularized trained network h, this plot is a histrogram of the size of a particular reduced network h_{red} at each input $\mathbf{x}$. The reduced widths at each layer correspond to the choice of optimal sparsity level.

Reduced Lipschitz constant for regularized networks



(a) Off-the-shelf



(b) Regularized

Figure: Histogram of optimal sparse local Lipschitz scale across inputs. At each input, the size of the reduced network corresponds to $\mathbf{s}^*(\mathbf{x})$.

Reduced Babel Function

Definition

For any matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$, we define the reduced babel function at row sparsity level $s_1 \in \{0, \ldots, d_1 - 1\}$ and column sparsity level $s_2 \in \{0, \ldots, d_2 - 1\}$ as,

$$\mu_{s_1,s_2}(\mathbf{W}) := \max_{\substack{J_1 \subset [d_1], \\ |J_1| = d_1 - s_1}} \max_{j \in J_1} \left[\sum_{\substack{i \in J_1, \\ i \neq j}} \max_{\substack{J_2 \subseteq [d_2] \\ |J_2| = d_2 - s_2}} \frac{|\mathcal{P}_{J_2}(\mathbf{w}_i)\mathcal{P}_{J_2}(\mathbf{w}_j)^T|}{\|\mathcal{P}_{J_2}(\mathbf{w}_i)\|_2 \|\mathcal{P}_{J_2}(\mathbf{w}_j)\|_2} \right],$$

the maximum cumulative mutual coherence between a reference row in J_1 of size $(d_1 - s_1)$ and any other row in J_1, each restricted to any subset of columns J_2 of size[3] $(d_2 - s_2)$.

Lemma

For any matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$, the operator norm of any non-trivial[4] sub-matrix indexed by sets $J_1 \subseteq [d_1]$ of size $(d_1 - s_1)$ and $J_2 \subseteq [d_2]$ of size $(d_2 - s_2)$ can be bounded as

$$\|\mathcal{P}_{J_1,J_2}(\mathbf{W})\|_2 \leq \sqrt{1 + \mu_{s_1,s_2}(\mathbf{W})} \cdot \|\mathbf{W}\|_{2,\infty}.$$

[3] When $s_1 = d_1 - 1, |J_1| = 1$, we simply define $\mu_{(s_1,s_2)}(\mathbf{W}) := 0$.

[4] That is $0 \leq s_1 \leq d_1 - 1$ and $0 \leq s_2 \leq d_2 - 1$.