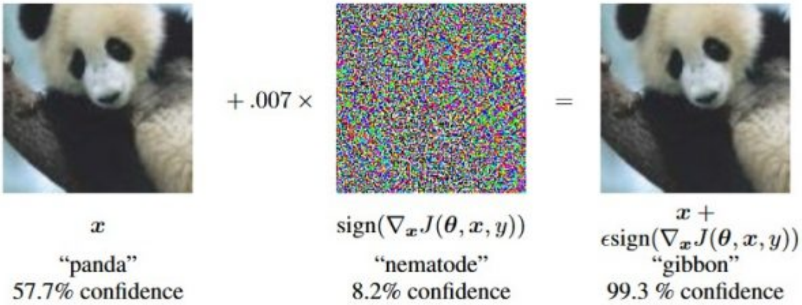
Adversarial Robustness of Sparse Local Lipschitz Predictors

Ramchandran Muthukumar

January 17, 2023

Adversarial Robustness



Two central questions of Robustness

- *G0}§_»0L |n@« §j0* : What is the minimal size of an adversarial perturbation for a predictor \ at input ².
- *|n@« § W0j0}9℮_¶9§_nj* : When will a predictor \ learnt on a training data ~¦ generalize to corrupted unseen data ?

Our Contribution

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

Preliminary Notation

- Input space : $X := f{}^2 \ 2 \ \mathrm{R}^L; k{}^2 k_2 \quad 1g$
- Output space : $Y := f1; : : : ; Gg$.
- Perturbation Space : $B \ := f \quad 2 \ \mathrm{R}^L; k \ \ k_2 \qquad g$
- Data Distribution : $D_Z := D_X \quad D_Y$ on $Z := X \quad Y$.
- Training sample (i.i.d): $\sim_{\text{¦}} := f\P_{\_} g^{h}_{\_=1} = f({}^2_{\_}; {}´_{\_}) g^{h}_{\_=1}$
- Hypothesis class : $H : X \ \ ! \ \ \mathrm{R}^G$ with embedded norm $k \ k_H$.

Notation - Representation-Linear Hypothesis

- We only consider }0q}0 0j§9§_nj]e_j09} \´qn§\0 _ He9 0 .

$$H := f \backslash _{8\,;} - (^2) := 8 \quad - (^2);\ 8\ (8\,;\ ^{-}\ )\ 2\ A \quad W g :$$

Here, – is a representation map and 8 is a classification weight.

Notation - Representation-Linear Hypothesis

- We only consider }0q}0 0j§9§_nj]e_j09} \´qn§\0 _ He9 0 .

$$H := f \backslash_{8;} - (^2) := 8 \quad - (^2);\ 8\,(8;\ ^{-}\ )\ 2\ A \quad W g:$$

  Here, $-$ is a representation map and 8 is a classification weight.

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

$$^{[b]}(^2) := \quad - ^{b} \quad - ^{b}\ ^{1} \quad - ^{1}\ _{2} + @^{1} \quad + @^{b}\ ^{1} \quad + @^{b} \quad :$$

Sensitivity

- **Wen@9e d_q H_§¶j0** : A constant L_jq, for all ²;² 2 X and \ 2 H, we have that

$$k\backslash(^2) \quad \backslash(^2)k_2 \quad \mathsf{L}\_{jq}\, k^2 \quad {}^2k_2$$

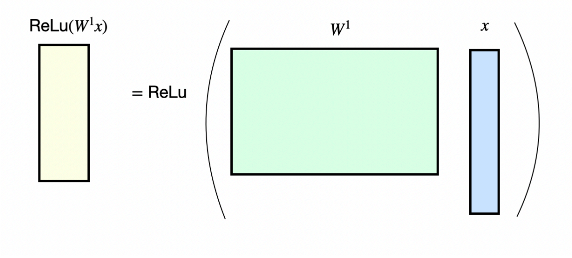
- **dnH9e d_q H_§¶j0** : A radius function }_jq and a Lipschitz scale function ℓ_jq such that,

$$k^2 \quad {}^2k_2 \quad \}\_{jq}(^2) \;\; Q) \quad k\backslash(^2) \quad \backslash(^2)k_2 \quad \ell\_{jq}(^2)\,k^2 \quad {}^2k_2 :$$

- If there is a structural property at a predictor output \(²), within what radius can we gaurantee that \(²) retains the property
- A structural property for neural networks - activation states of neurons in each layer.

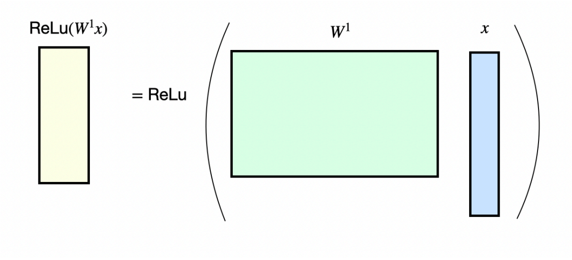
Motivation - Feedforward layers

For feedforward networks, each layer is a feed-forward map $^{(c)}(\S) := \; (\,^{-}\; {}^{c}\S)$.



Motivation - Feedforward layers

For feedforward networks, each layer is a feed-forward map $^{(c)}(\S) :=\ (\,^{-}\,^{c}\S)$.



ReLu induces an **9H§_®9§_nj q9§§0}j** in the output of each layer $^{(c)}(\S)$. We denote by $J^{\,c}(\S)$ and $I^{\,(c)}$ the true support and co-support of the layer output.

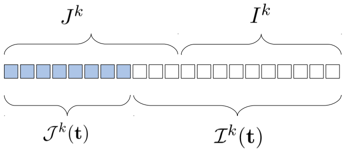


Figure: Illustration of the sets $J^{\,c}(\S)$, $I^{\,c}(\S)$, as well as $^{\wedge c}$ and $`^{c}$, for a given intermediate input $(\,^{-}\,^{c}\S + @^{c})$. Colored squares represent non-zero elements, ordered here without loss of generality.

Motivation : Effect of ReLu

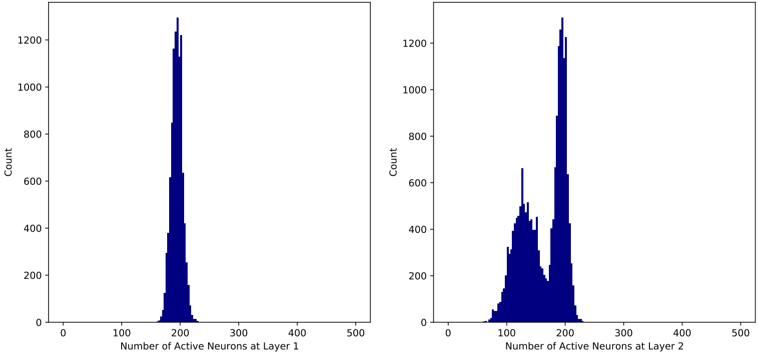
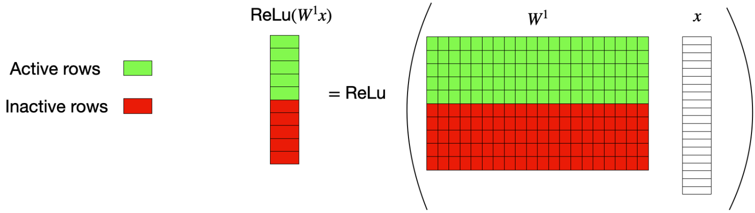


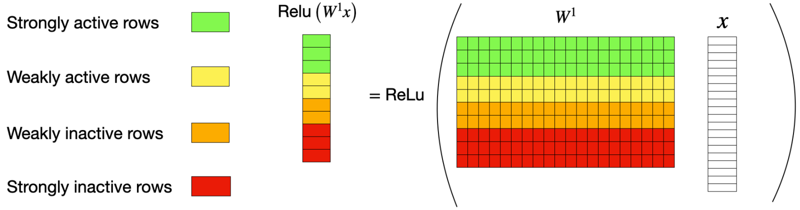
Figure: Distribution of neuron activity (size of $J^c(\S)$) in each layer c of a network trained on MNIST. At each layer only 40 percent of the neurons are activated.

Motivation - Effect of ReLu



Activation states are the result of interaction between rows of $^{-1}$ and input ^{2}.

Motivation - Effect of ReLu



For bounded perturbations, the strongly inactive rows remain inactive.

Sparse Local Lipschitz (SLL)

A representation map is ~dd °u}u§ _jq«§ if at 09H\ input ² 2 X and sparsity level 2 S, there exists[1]

- A stable inactive index set ^(²;) of size s for the representation (²)
- A sparse local radius function }_jq : X S ! R 0
- A sparse local Lipschitz scale function ℓ_jq : X S ! R 0

such that for any perturbation ,

$$k \; k_2 \quad \}_{\_jq}(^2; \;) \;\; Q) \begin{cases} k \;\; (^2 + \;\;) \quad (^2)k_2 \quad \ell_{\_jq}(^2; \;) \, k \;\; k_2 \\ \wedge(^2; \;) \text{ is inactive for } (^2 + \;\;): \end{cases}$$

[1] Thus we necessarily only talk of q k (²)k_0

Sparse Local Lipschitz (SLL)

A representation map is ~dd °u}u§ _jq«§ if at 09H\ input ² 2 X and sparsity level 2 S, there exists[1]

- A stable inactive index set ^(²;) of size s for the representation (²)
- A sparse local radius function $\}_{jq} : X$ S ! R ^{0}
- A sparse local Lipschitz scale function $\ell_{jq} : X$ S ! R ^{0}

such that for any perturbation ,

$$k \ k_2 \quad \}_{jq}(^2;\ )\ \mathrm{Q})\ \left(\begin{array}{l} k\ (^2 +\ ) \quad (^2)k_2 \quad \ell_{jq}(^2;\ )k\ \ k_2 \\ \text{^}(^2;\ ) \text{ is inactive for } (^2 +\ ): \end{array} \right.$$

SLL Q) local sensitivity to perturbation + invariance in representation sparsity pattern

[1] Thus we necessarily only talk of q k $(^2)k_0$

Feedforward Maps are SLL

Lemma

8 j ´ TOOLTn}°9}L h9qJ (²) := (¯ ² + @) _ ~dd °u}u§ _j q«§u

$$\text{^}(\,^2;\ \ ) := \underset{\substack{\text{^}\ \ I\,(\,^2);\\ j\text{^}j\,=}}{9\}X\,h\,9^2}\ \min_{\_2\text{^}} \frac{j\,°\,\_^{\,2} + @j}{k\,°\,\_k_2};$$

$$\}_{\_j\,q}(\,^2;\ \ ) := \min_{\_2\text{^}} \frac{j\,°\,\_^{\,2} + @j}{k\,°\,\_k_2};$$

$$e_{\_j\,q}(\,^2;\ \ ) := k\ ¯\ [\,`;:]k_2\,:$$

$` = (\text{^}(\,^2;\ \ ))^H$ _ §\0 Hn h qe0 h 0j § _jL0² 0§u

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

Sparse Local Radius at Layer c

For the feedforward map $^{(c)}$, the strongly inactive index set $^{\wedge c}$ $I^c(\S)$ is uniquely identified at layer input § and sparsity level $^{(c)}$.

To compute $^{\wedge c}$ we sort the normalized pre-activation vector $\mathbf{x}^c := \left(\frac{\circ^c \S + @^c}{k \circ^c_- k_2} \right)_{-=1}^{L^c}$:

Sparse Local Radius at Layer c

For the feedforward map $^{(c)}$, the strongly inactive index set $^{\wedge c}$ $I^{c}(\S)$ is uniquely identified at layer input § and sparsity level $^{(c)}$.

To compute $^{\wedge c}$ we sort the normalized pre-activation vector $\mathbf{x}^{c} := \left(\frac{\circ^{c} \S + @^{c}}{k \circ^{c}_{\_} k_{2}} \right)_{\_ = 1}^{L^{c}}$:

Figure: Illustration of the radius $\}^{(c)}_{\_\mathrm{j\,q}}(\S;\ ^{(c)})$ for the intermediate feedforward representation $^{(c)}$, given the (sorted) values of the normalized pre-activations.

AKT +iQ7 aGG M HvbBb

> 2`2 i?2 BM/2t b2? 2 bi`QM;Hv BM +iBp2 BM/2t b2iX
h?2 bi #BHBiv Q7 K?2 +i2bi i?2 `2T`2b2Mi iBQM +QKTmi2
H v2`bX

JQiBp iBQM @ 1z2+i Q7 _2Gm

G2C = (A)* X 6Q` T2`im`# iBQMb rBi?BM i?2 (tT)-bi2?2HQ+
`2T`2b2Mi iBQM +QKTmi2/ Bb 2[mBp H2Mi iQ `Q`rb/BM+2
q M/A +QHmKMb qXBM

>2M+2 i?2 b2MbBiBpBiv BM }`bi Q`/2` [v2:]b)T`QT ; i2b b

6Q` i?2 b2+QM/ H v2` i?2`2 Bb bT `bBiv T ii2`M BM i?2
T ii2`M BM i?2 Q`B;BM H H v2` BMTmiX q2 + M T`QT ;

* Q K T Q b B i B Q M Q 7 a G G K T b B b a G G

* Q M b B E/B 2 M i 2 ` K 2 / B i 2 H v 2 ` ` 2 T ` 2 b 2 M i B M ; i? 2 + ?
` 2 i ? 2 M + Q K T Q b 2 / i Q - Q # i B M

$$[E](t) := (E) \quad (E \; 1) \quad (1) \; (t):$$

G 2 K K

b b m K 2 2 (F) + ? B b a G G r X ` X i X B M T m i b H X B i ? `
h ? 2 + Q K T Q b 2 / K T b U m T i Q ` 2 H v 2 ` F a G G r B i ? ` M / B m b `
G B T b + ? B i X b B p 2 M # v

$$\{F\}_{BMT}\; t;\; b^{[F]} := \underset{1 \;\; M \;\; F}{KBM} \frac{(M)_{BMT}\;[M\;1]\;(t);\; b^{(M)}}{H^{[M\;1]}_{BMT}\;\; t;\; b^{[M\;1]}}$$

$$H^{[F]}_{BMT}\; t;\; b^{[F]} := \prod_{M=1}^{F} H^{(M)}_{BMT}\;[M\;1]\;(t);\; b^{(M)} :$$

6 Q ` M v T 2 ` i m ` # B i B Q M ( B M t; b ) - B M / 2 t b ; 2 ; A A ` 2 K B M
B M + i B p 2 X

---

k > 2 ` ( 2 b ; b ; : : : ; b ) ` 2 b T ` b B i v H 2 p 2 H b 7 Q ` 2 + ? b M = 2 ` K 2 / B i 2 b K i ? 2
H v 2 ` @ r B b 2 B M T m i @ Q m i T m b b : = ( ` b ; b) v H 2 p 2 H b m K M H i B p 2 B M T m i @ Q m i

_2/m+2/ .BK2MbBQM HBiv Q7 aGG T`2/B+iQ`b

- h?2 `2T`2b2MiBM; +QQKTmi2/ #v 2/7Q`r `/ H v2`b Bb aGG ` /Bmb M/ HQ+ H GBTb+?Bix b+ H2

_2/m+2/ .BK2MbBQM HBiv Q7 aGG T`2/B+iQ`b

- h?2 `2T`2b2MiiBQM Q7 HBM2i# v2/7Q`r `/ H v2`b Bb aGG ` /Bmb$_{BMT}$ M/ HQ+ H GBTb+?BixB i x b+ H2
- 6 2 2/7Q`r `/ M2m` H M2irQ`Fb 2t?B#Bi i?2 `2/m+2/ / T2`im`# iBQM b i?2 HQ+ H ` /Bmb-

$$?(t) = q^{E} \quad q^{E\ 1} \quad q^{1}\ t + \#^{1} + \#^{E\ 1} + \#^{E}$$

$$= \ {}_{`2/} \ q^{E}_{`2/} \quad q^{E\ 1}_{`2/} \quad q^{1}_{`2/}\ t + \#^{1}_{`2/} + \#^{E\ 1}_{`2/} + \#^{E}_{`2/}$$

$$=: \ ?_{`2/}(t)$$

r?2`2 $q^{F}_{`2/} := q^{F}\,[C^{F}; C^{F\ 1}]\ 2\ R^{(/^{F}\ \ b^{F})\ (/^{F\ 1}\ \ b^{F\ 1})}$

_2/m+2/ .BK2MbBQM HBiv Q7 aGG T`2/B+iQ`b

- h?2 `2T`2b2Mi[E] i+BQKMmi2E 7#22/7Q`r `/ H v2`b Bb aGG
  ` /Bm[E]b_{BMT} M/ HQ+ H GBTb+?[E]_{BMT}XBix b+ H2

- 622/7Q`r `/ M2m` H M2irQ`Fb 2t?B#Bi i?2 `2/m+2/ /
  T2`im`# iBrQBMb?B M i?2 HQ+ H ` /Bmb-

$$?(t) = q^{E} q^{E\ 1} q^{1}\ t + \#^{1} + \#^{E\ 1} + \#^{E}$$

$$= \ _{`2/} q^{E}_{`2/} q^{E\ 1}_{`2/} q^{1}_{`2/}\ t + \#^{1}_{`2/} + \#^{E\ 1}_{`2/} + \#^{E}_{`2/}$$

$$=: ?_{`2/}(t)$$

r?2`2 $q^{F}_{`2/} := q^{F}[C^{F}; C^{F\ 1}] 2 R^{(/^{F}\ b^{F})\ (/^{F\ 1}\ b^{F\ 1})}$

- M Bp2 2biBK i2 Q7 i?2 ;HQ# H GBTb+?Bix +QMbi $Q_{F=1}^{E+1} q^{F}\ ?_{2}$ X

- _2/m+2/ HQ+ H /BK2MbBQM HBv+B2Mi iQ /B`2+iHv GBTb+?Bix +QMbi MiQ7 i?2 7mHH Q`B;BM H M2irQ rBi? /2Ti? $Q_{F=1}^{E+1}$ b $q^{F}_{`2/\ 2}$ BX2X $Q_{F=1}^{E+1} q^{F}[C^{F}; C^{F\ 1}]_{2}$ X

*2`iB}2/ _Q#mbiM2bb 7Q` aGG T`2/B+iQ`b

h?2Q`2K

G2i (t) := (t) #2 T`2/B+iQ` bm+? i? i i?2 `2T`2b2MiiB

rBi? ` /Bmb 7mM+iBQM `BMT b+?Bix b+ HX 7mM+iBQM H

h?2 T`2/B+i2/ ( t) # 2 K BMb mM+? M;2/ B7 M /p2`

+Q``mTiBQM Bb rBi?BM i?2 +2`iB}2/ ` /Bmb `

$$`_{aGG}(t; ) := KBM`_{BMT}(t;); \frac{(t)}{2k\ k_2\ H_{BMT}(t;)} :$$

>2`2(-t) Bb i?2 +H bbB}+ iBQM K `;BMX

*2`iB}2/ _Q#mbiM2bb 7Q` aGG T`2/B+iQ`b

h?2Q`2K

G2i (t) := (t) #2 T`2/B+iQ` bm+? i? i i?2 `2T`2b2MiiiB rBi? ` /Bmb 7mM+iBQM $_{BMT}$ ?Bix b+ $_{BMT}$ HX 7mM+iBQM H h?2 T`2/B+i2(t) #2 BMi`Q/m+2K BMb mM+? M;2/ B7 M /p2` +Q``mTiBQM Bb rBi?BM i?2 $_{+2}$(t;) iB}2/ ` /Bmb `

$$`_{aGG}(t; ) := KBM `_{BMT}(t;); \frac{(t)}{2k \; k_2 \; H_{BMT}(t;)} \; :$$

>2`2(-t) Bb i?2 +H bbB}+ iBQM K `;BMX

- 6Q` /2T2M/2Mi 7Q`r `/ M2irQ`Fb- i?2 HQ+ H $_{BMT}$ ` /Bm ... M/ i?2 HQ+ H GBTb+?Bix +QMbi Mi H_{BMT} H2 7mM+iBQM Bb
- GQ+ H f:HQ# H GBTb+?Bix +QMbi Mi Bb +Q``2bTQM/ i
- q2 + M QTiBKBx2 Qp2` bT `bBiv H2p2Hb iQ ;2i i?2 #

h`m2 +2`iB}2/ ` /Bmb

^(t) := G #2H T`2/B?+Q2M #v
BM TtmXi

`$_{+2}$(`$_i$t) := K BkMk$_2$

bi:^(t +) 6 ^(t)

6Q` HH T2`im`# iBQMb rBi?BM
`$_{+2}$(`$_i$t) - i?2 H #2H `2K BMb mM@
+? M;2/X

/p2`b `B H mTT2` #QmM/

6Q` Mv /p2`b `B H ii +F- TB+F
i?2 2t KTH2 rBi? H2 bi 2M2`;vX

$$
`_{/p}(t) := \underset{/p \ ii \ +Fb}{KBM} \|k\|_2
$$

$$
bi \colon \hat{\ } V(t + \ ) \leq \hat{\ } V(t)
$$

lTT2` #QmM/ bBM+2 S:. /Q2bMöi
T`Qp #Hv +QMp2`;2 iQ QTiBK H
T2`im`# iBQMX

:HQ# H GBTb+?Bix +2`iB}+ i2

G2i$_{BMT}$ G# 2 i?2 ;HQ# H GBTb+?Bix
+QMbi MiX 6Q` Mv T2`im`# iBQM-

$$k ?(t +) ?(t)k_2 \quad G_{BMT} k k_2 :$$

h?2 ;HQ# H +2`iB}2/ ` /Bmb

$$`_{;HQ\# H}(t) := \frac{(t)}{G_{BMT}} ;$$

2Mbm`2b T2`im`# iBQMb /QMöi
+`Qbb /2+BbBQM #QmM/ `B2bX

A7 Bb HQ+ H GBTb+?Bix-

$k \; k_2 \quad `_{BM}(T t)$

$4) \; k \; ?(t+ \;) \quad ?(t) k_2 \quad k \quad k_2 \; H_{BM}(T t):$

h?2 HQ+ H +2`iB}2/ ` /Bmb Bb

$$`_{HQ+H}(t) := KBM\left(`_{BM}(T t); \frac{(t)}{2k \quad k_2 \; H_{BM}(T t)}\right)$$

2Mbm`2b T2`im`# iBQMb /QMöi 2t@ +22/ HQ+ H GBTb+?Bix ` /Bmb Q` K `;BM BM QmiTmi bT +2X

aT `b2 GQ+ H GBTb+?Bix *2`iB}+ i2

h?2 bT `b2 +2`iB}+ i2 Bb-

$`_{aGG}(t;)$

$$:= \mathrm{KBM} \, `_{BM}(Tt;); \frac{(t)}{2k \quad k_2 \quad H_{BM}(Tt;)} \quad :$$

1[mBp H2Mi iQ HQ+ H GBTb+?Bix
M Hvb Bb07XQ`

PTiBKBx2 Qp2` bT `bBiv H2p2Hb 7Q` #2bi +2`iB}+ i2-

$`_{bT}\ `_{b2}(t)$

$$:= K_b\ t`_{aG}(G_t;\ b)$$

$$= K_b\ tKBM`_{BM}(T_t;\ b);\ \frac{(\ t)}{2k\ \ k_2\ \ H_{BM}(T_t;\ b)}\ :$$

i 2 +?- i?2 QTiBK H bT `bBiv H2p2Hb bT2+B}+ `2/m+2/ M2irQ`FX

_2/m+2/ .BK2MbBQM HBiv

U V >BbiQ;` K Q7 `2/m+2/ rB/i?b i H U #V >BbiQ;` K Q7 `2/m+2/ rB/i?b i

6B;m`2 , 6Q` M Qz@i?2@b?2H7 I- i?2 1- M/ U#V `2T`2b2Mi i?2 /Bb rB/i?b Q7 i?2 T `iB+mH ` `2/m+2/ M +QMi`Q mbBM; i?2 `2/m+2/ rB/i?b 2 +? H v2` +Q``2bTQM/ iQ i?2 K t BK H `2/m+2/ rB/i? 2p2H

_2/m+2/ GBTb+?Bix +QMbi Mi

6B;m`2 BbiQ;` K Q7 QTiBK H bT `b2 HQ+ H GBTb+?Bix b+ H2
i?2 bBx2 Q7 i?2 `2/m+2/ M2irQ`F (`?Q`BxQMi H `2/ TQH/BM2 K `Fb i?2 M
2biBK i2 Q7 ;HQ# H GBTb+?Bix +QMbi MiX

*2`iB}2/ _Q#mbiM2bb 7Q` 622/@7Q`r `/ L2m` H

q2 THQi +2`iB}2/ +Q` +v BM2/ T`2/B+iQ` mbBM;-

- L Bp2 +2`iB}+ i2 rBi? ;HQ# H GBTb+?Bix +QMbi Mi
- aGG +2`iB}+ i2 rBi? HQ+ H GBTb+?Bix +QMbi Mi
- >2m`BbiB+ mTT2` #QmM/ 7`QK +QKKQM /p2`b `B H

U Pz@i?2@b?2H7 U#_2;mH `Bx2/

6B;m`2 +m`Bv +m`p2b 7Q` 722/@7Q`r `/ M2m` H M2irQ

aT `b2 GQ+ H GBTb+?Bix rX`Xi S ` K2i2`b M/ A

- M HvbBb + M #2 2ti2M/2/ iQ T2`im`# iBQMb iQ #Qi
- aT `b2 HQ+ H ` /Bmb ; BM [m MiB}2b bi #BHBiv Q7
- aBKBH ` `2/m+2/ /BK2MbBQM HBiv 2ti2+M/7Q` T2`i
  T2`im`#2/ r2 Br;?ii?BM HQ+ H ` /BmbX

Robust Generalization Bound for Feedforward Neural Networks

Theorem

¯_§\ q}n@9@_e_§´ 9§ e09 § (1) n®0} §\0 H\n_H0 nT _u_uL §}9_j_jX 9 h qe0 S¦ 9jL «je9@0e0L L9§9 S«J Tn} 9j ´ h «e§_]e9´0}0L j0«}9e j0§°n}c q}0L_H§n} \ 2 H^{b+1} °_§\ q9}9 h 0§0} f ¯ ^{c}g §\0 }n@« § §nH\9 §_H }_ c _ @n«jL0L 9 J

$$| _{}n@} (\backslash) \quad \uparrow_{}n@} (\backslash)$$

$$O \quad @ \sqrt{\frac{\ln \quad N \quad \frac{1}{h(b+1)}; H^{b+1} \quad + \ln(\frac{2}{})}{2h}}$$

$$+ \frac{\mathsf{L}_{en} \quad (1 + \quad)}{h} \prod_{c=1}^{b+1} - c \quad {}_{2;\,1} \quad \mathrm{q} \overline{1 + \quad {}_{c;\,c} \quad {}_1(\bar{} \ {}^c)} \quad !$$

[0}0J = (^{1}; : : : ; ^{b}) _ 9j nq§_h 9e q9} _§´ e0®0e H\n 0j @9 0L nj S¦ 9jL S«u ${}_{c;\,c} \quad {}_1(\bar{} \ {}^c)$ _ 9 }0L«H0L @9@0e T«jH§_nj 9jL k ¯ $k_{2;\,1}$ _ §\0 h 9²_h 9e `$_2$ jn}h nT 9 }n° _j ¯ u

Thank you for attending my talk :)

Corollary

Gnj _L0} 9 §}9_j0L L0q§\]b + 1 T00L]Tn}°9}L j0«}9e j0§°n}c \u

d0§ = (^{1}; : : : ; ^{b}) @0 9 H\n_H0 nT q9} _§´ e0®0e 9§ 09H\ e9´0}

d0§ ®$^{(c)}$:= (^{c} ^{1}; ^{c}) @0 §\0 Hn}}0 qnjL_jX e9´0}]°_ 0 _jq«§]n«§q«§ q9} _§´ e0®0e u

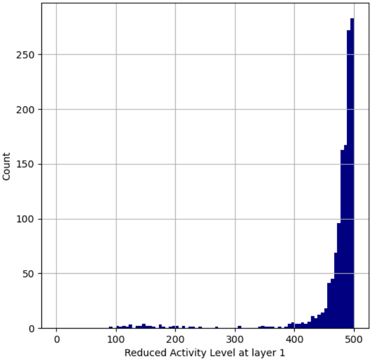
¦ \0 q}0L_H§0L e9@0e }0h9_j «jH\9jX0LJ °\0j0®0} k k$_2$ }HO}§(²;)J °\0}0

$$}_{HO}§(²;) := \min \left\{ \min_{: 1 \quad c \quad b} \frac{}^{(c)}_{\_jq}(\; ^{[c \quad 1]}(²); ®^{(c)})}{\prod_{j=1}^{c} P_{`j;`j \quad 1}(- \; ^j)_2} ; \frac{(²)}{2k8k_2 \prod_{c=1}^{b} P_{`c;`c \quad 1}(- \; ^c)_2} \right\};$$

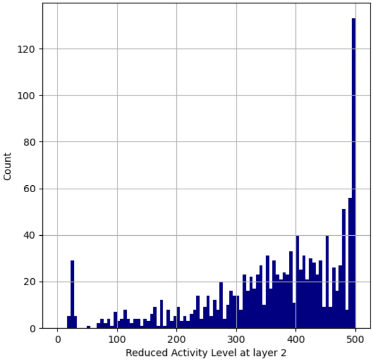
[0}0J }$^{(c)}_{\_jq}$ _ §\0 enH9e }9L_« Tn} §\0 T00LTn}°9}L h9q 9§ e9´0} cu

9jL P$_{`c;`c \quad 1}$($-$ ^{c}) _ §\0 9H§_®9§0L °0_X\§ 9§ e9´0} cu

Reduced Widths for regularized networks



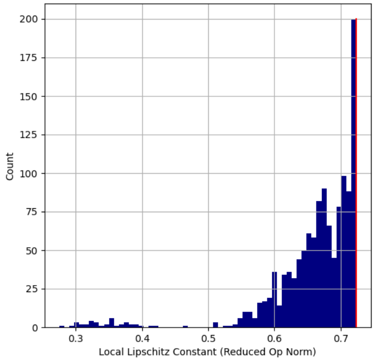
(a) Histogram of reduced widths at layer 1



(b) Histogram of reduced widths at layer 2

Figure: For an original regularized trained network \, this plot is a histrogram of the size of a particular reduced network $\backslash_{\}0L}$ at each input ². 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 $(\ ^{2})$.

Definition

For any matrix $^{-}\ 2\ \mathrm{R}^{L_1 \quad L_2}$, we define the reduced babel function at row sparsity level $_1\ 2\ f0; : : : ; L_1 \quad 1g$ and column sparsity level $_2\ 2\ f0; : : : ; L_2 \quad 1g$ as,

$$_{1;\ 2}(\ ^{-}\ ) := \underset{\substack{`_1 \quad [L_1]; \\ j`_1 j = L_1 \quad _1}}{\max} \ \underset{a\, 2\, `_1}{\max} \ "\ \underset{\substack{\_2\, `_1; \\ \_6=a}}{X} \ \underset{\substack{`_2 \quad [L_2] \\ j`_2 j = L_2 \quad _2}}{\max} \frac{jP_{`_2}(\ ^{\circ}\ \_)P_{`_2}(\ ^{\circ}\ _a)^{'} j}{kP_{`_2}(\ ^{\circ}\ \_)k_2\ kP_{`_2}(\ ^{\circ}\ _a)k_2}\ \#\ ;$$

the maximum cumulative mutual coherence between a reference row in $`_1$ of size $(L_1 \quad _1)$ and any other row in $`_1$, each restricted to any subset of columns $`_2$ of size[3] $(L_2 \quad _2)$.

Lemma

Sn} 9j´ h9§}_² $^{-}\ 2\ \mathrm{R}^{L_1 \quad L_2}$ J §\0 nq0}9§n} jn}h nT 9j´ jnj]§}_®_9e[4] *«@] h9§}_² _jL0²0L @´ 0§ $`_1 \quad [L_1]$ nT _¶0 $(L_1 \quad _1)$ 9jL $`_2 \quad [L_2]$ nT _¶0 $(L_2 \quad _2)$ H9j @0 @n«jL0L 9*

$$kP_{`_1; `_2}(\ ^{-}\ )k_2 \quad \mathrm{p}\overline{1 + \ _{1;\ 2}(\ ^{-}\ )}\ \ k\ ^{-}\ k_{2;\ 1}\ :$$

[3]When $_1 = L_1 \quad 1; j`_1 j = 1$, we simply define $_{(\ 1,\ 2)}(\ ^{-}\ ) := 0$.

[4]That is $0 \quad _1 \quad L_1 \quad 1$ and $0 \quad _2 \quad L_2 \quad 1$.