Adversarial Robustness of Sparse Local Lipschitz Predictors

Ramchandran Muthukumar

January 17, 2023
Adversarial Robustness

\[ x + \varepsilon \text{sign}(\nabla_x J(\theta, x, y)) = x + \varepsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

"panda" 57.7% confidence

+ .007 \times

sign(\nabla_x J(\theta, x, y))

"nematode" 8.2% confidence

= 

"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 $x$.

- **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: $\mathcal{X} := \{x \in \mathbb{R}^d, \|x\|_2 \leq 1\}$

- Output space: $\mathcal{Y} := \{1, \ldots, C\}$.

- Perturbation Space: $\mathcal{B}_\nu := \{\delta \in \mathbb{R}^d, \|\delta\|_2 \leq \nu\}$

- Data Distribution: $\mathcal{D}_Z := \mathcal{D}_X \times \mathcal{D}_Y$ on $Z := \mathcal{X} \times \mathcal{Y}$.

- Training sample (i.i.d): $S_T := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$

- Hypothesis class: $\mathcal{H} : \mathcal{X} \to \mathbb{R}^C$ with embedded norm $\|\cdot\|_{\mathcal{H}}$. 
We only consider *representation-linear hypothesis classes*.

\[ \mathcal{H} := \{ h_{A,W}(x) := A \Phi_W(x), \ \forall (A, W) \in \mathcal{A} \times \mathcal{W}\} \]

Here, \( \Phi_W \) is a representation map and \( A \) is a classification weight.
We only consider representation-linear hypothesis classes.

\[ \mathcal{H} := \{ h_{A,W}(x) := A\Phi_W(x), \forall (A, W) \in \mathcal{A} \times \mathcal{W}\} . \]

Here, \( \Phi_W \) is a representation map and \( A \) is a classification weight.

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

\[ \Phi^{[K]}(x) := \sigma (W^K \sigma (W^{K-1} \ldots \sigma (W^1 x + b^1) \ldots + b^{K-1}) + b^K) . \]
Sensitivity

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

  $$\| h(\tilde{x}) - h(x) \|_2 \leq L_{\text{inp}} \| \tilde{x} - x \|_2$$

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

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

- If there is a structural property at a predictor output $h(x)$, within what radius can we guarantee that $h(\tilde{x})$ 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 $\Phi^{(k)}(t) := \sigma(W^k t)$.
Motivation - Feedforward layers

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

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

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

Figure: Distribution of neuron activity (size of $J^k(t)$) 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 $W^1$ and input $x$. 
Motivation - Effect of ReLu

For bounded perturbations, the strongly inactive rows remain inactive.
Sparse Local Lipschitz (SLL)

A representation map $\Phi$ is \textit{SLL w.r.t inputs} if at each input $x \in \mathcal{X}$ and sparsity level $s \in \mathcal{S}$, there exists\(^1\)

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

such that for any perturbation $\delta$,

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

\(^1\)Thus we necessarily only talk of $s \leq p - \|\Phi(x)\|_0$
Sparse Local Lipschitz (SLL)

A representation map $\Phi$ is \textit{SLL} \textit{w.r.t inputs} if at each input $x \in \mathcal{X}$ and sparsity level $s \in \mathcal{S}$, there exists

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

such that for any perturbation $\delta$,

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

\text{SLL} \implies \text{local sensitivity to perturbation } + \text{ invariance in representation sparsity pattern}

---

\textsuperscript{1} Thus we necessarily only talk of $s \leq p - \|\Phi(x)\|_0$
Feedforward Maps are SLL

Lemma

Any feedforward map, \( \Phi(x) := \sigma(Wx + b) \) is SLL w.r.t input.

\[
I(x,s) := \arg\max_{I \subseteq I(x), \|I\|=s} \min_{i \in I} \frac{|w_i x + b_i|}{\|w_i\|_2},
\]

\[
r_{\text{inp}}(x,s) := \min_{i \in I} \frac{|w_i x + b_i|}{\|w_i\|_2},
\]

\[
l_{\text{inp}}(x,s) := \|W[J,:]\|_2.
\]

\( J = (I(x,s))^c \) 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 $I^k \subset I^k(t)$ is uniquely identified at layer input $t$ and sparsity level $s^{(k)}$.

To compute $I^k$ we sort the normalized pre-activation vector $q^k := \left[ \frac{w^k_i + b^k_i}{\|w^k_i\|_2} \right]_{i=1}^{d^k}$. 
Sparse Local Radius at Layer $k$

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

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

**Figure:** Illustration of the radius $r_{\text{inp}}^{(k)}(t, s^{(k)})$ 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 $J^1 = (l^1)^C$. For perturbations within the sparse local radius $r_{\text{inp}}(x, s)$, the representation computed is equivalent to a reduced network without $l^1$ rows in $W^1$ and $l^1$ columns in $W^2$.

Hence the sensitivity in first layers propagates as $\|W^1[J^1, :]\|_2$.
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 $\Phi^{[K]}$,

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

Lemma

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

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

$$r^{[k]}_{\text{inp}}(x, s^{[k]}) := \min_{1 \leq n \leq k} \frac{r^{(n)}_{\text{inp}}(\Phi^{[n-1]}(x), s^{(n)})}{l^{[n-1]}_{\text{inp}}(x, s^{[n-1]})}$$

$$l^{[k]}_{\text{inp}}(x, s^{[k]}) := \prod_{n=1}^{k} l^{(n)}_{\text{inp}}(\Phi^{[n-1]}(x), s^{(n)}).$$

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

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

- A naive estimate of the global Lipschitz constant $\prod_{K=1}^{K+1} W_{K}^{2}$.

- Reduced local dimensionality is inefficient to directly compute the Lipschitz constant of the full original network. The local sensitivity scales with depth as $\prod_{K=1}^{K+1} W_{K}^{2}$, i.e., $\prod_{K=1}^{K+1} W_{K \left[ J_{K} \right]}^{2}$.
Reduced Dimensionality of SLL predictors

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

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

$$h(\tilde{x}) = A\sigma \left( W^K \sigma \left( W^{K-1} \cdots \sigma \left( W^1 \tilde{x} + b^1 \right) \cdots + b^{K-1} \right) + b^K \right)$$

$$= A_{\text{red}} \sigma \left( W^K_{\text{red}} \sigma \left( W^{K-1}_{\text{red}} \cdots \sigma \left( W^1_{\text{red}} \tilde{x} + b^1_{\text{red}} \right) \cdots + b^{K-1}_{\text{red}} \right) + b^K_{\text{red}} \right)$$

$$=: h_{\text{red}}(\tilde{x})$$

where $W^k_{\text{red}} := W^k[J^k, J^{k-1}] \in \mathbb{R}^{(d^k-s^k) \times (d^{k-1} - s^{k-1})}$
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 $\tilde{x}$ within the local radius,

$$h(\tilde{x}) = A \sigma \left( W^K \sigma \left( W^{K-1} \cdots \sigma (W^1 \tilde{x} + b^1) \cdots + b^{K-1} \right) + b^K \right)$$

$$= A_{\text{red}} \sigma \left( W_{\text{red}}^K \sigma \left( W_{\text{red}}^{K-1} \cdots \sigma (W_{\text{red}}^1 \tilde{x} + b_{\text{red}}^1) \cdots + b_{\text{red}}^{K-1} \right) + b_{\text{red}}^K \right)$$

$$=: h_{\text{red}}(\tilde{x})$$

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

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

- 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 $\prod_{k=1}^{K+1} \| W_{\text{red}}^k \|_2$ i.e. $\prod_{k=1}^{K+1} \| W^k [J^k, J^{k-1}] \|_2$. 
Certified Robustness for SLL predictors

Theorem

Let \( h(x) := A\Phi(x) \) be a predictor such that the representation map \( \Phi \) is SLL with radius function \( r_{\text{inp}} \) and Lipschitz scale function \( l_{\text{inp}} \).

The predicted label \( \hat{y}(x) \) at input \( x \) remains unchanged if an adversarial corruption is within the certified radius \( r_{\text{cert}}(x, s) \),

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

Here, \( \rho(x) \) is the classification margin.
Certified Robustness for SLL predictors

**Theorem**

Let \( h(x) := A\Phi(x) \) be a predictor such that the representation map \( \Phi \) is SLL with radius function \( r_{\text{inp}} \) and Lipschitz scale function \( l_{\text{inp}} \).

The predicted label \( \hat{y}(x) \) at input \( x \) remains unchanged if an adversarial corruption is within the certified radius \( r_{\text{cert}}(x, s) \),

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

Here, \( \rho(x) \) 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

\[ \hat{y}(x) := \text{Label predicted by } h \text{ on input } x. \]

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

For all perturbations within \( r_{\text{cert}}(x) \), the label remains unchanged.
Adversarial upper bound

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

$$ r_{adv}(x) := \min_{\text{adv attacks}} \|\delta\|_2 $$

subject to $$ \hat{y}(x + \delta) \neq \hat{y}(x) $$

Upper bound since PGD doesn’t provably converge to optimal perturbation.
Global Lipschitz certificate

Let $L_{\text{inp}}$ be the global Lipschitz constant. For any perturbation,

$$\| h(x + \delta) - h(x) \|_2 \leq L_{\text{inp}} \| \delta \|_2 .$$

The global certified radius

$$r_{\text{global}}(x) := \frac{\rho(x)}{L_{\text{inp}}} ,$$

ensures perturbations don’t cross decision boundaries.
Local Lipschitz certificate

If \( \Phi \) is local Lipschitz,
\[
\| \delta \|_2 \leq r_{\text{inp}}(x) \\
\implies \| h(x + \delta) - h(x) \|_2 \leq \| A \|_2 l_{\text{inp}}(x).
\]

The local certified radius is
\[
r_{\text{local}}(x) := \min \left\{ r_{\text{inp}}(x), \frac{\rho(x)}{2 \| A \|_2 l_{\text{inp}}(x)} \right\}
\]

ensures perturbations don’t exceed local Lipschitz radius or margin in output space.
Sparse Local Lipschitz Certificate

The sparse certificate is,

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

Equivalent to local Lipschitz analysis for \( s = 0 \).
Optimize over sparsity levels for best certificate,

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

At each \( x \), 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 $h_{\text{red}}$ at each input $x$. The reduced widths at each layer correspond to the choice of optimal sparsity level.
Figure: Histogram of optimal sparse local Lipschitz scale across inputs. At each input, the size of the reduced network corresponds to $s^*(x)$. 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 $\prod_{k=1}^{K+1} \| W^k \|_2$.
- SLL certificate with local Lipschitz constant $\prod_{k=1}^{K+1} \| W^k[J^k, J^{k-1}] \|_2$.
- Heuristic upper bound from common adversarial attacks.

\textbf{Figure:} Security curves for feed-forward neural networks on MNIST.
- 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 $\tilde{x}$ and perturbed weight $\tilde{W}$ 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,

\[
R_{\text{rob}}(h) - \hat{R}_{\text{rob}}(h) \leq \tilde{O}\left(\frac{b\sqrt{\ln\left(\mathcal{N}\left(\frac{1}{m(K+1)}, \mathcal{H}^{K+1}\right)\right)} + \ln\left(\frac{2}{\alpha}\right)}{2m} + \frac{L_{\text{loss}}(1 + \nu)}{m} \prod_{k=1}^{K+1} \|\mathbf{W}^k\|_{2,\infty} \sqrt{1 + \mu_{s^k,s_{k-1}}(\mathbf{W}^k)}\right)
\]

Here, $s = (s^1, \ldots, s^K)$ 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 $\ell_2$ norm of a row in $\mathbf{W}$. 
Thank you for attending my talk :)

Certified Robustness for Feed-forward Neural Networks

Corollary

Consider a trained depth-$K + 1$ feed-forward neural network $h$. Let $s = (s^1, \ldots, s^K)$ be a choice of sparsity levels at each layer. Let $v^{(k)} := (s^{k-1}, s^k)$ be the corresponding layer-wise input-output sparsity levels.

The predicted label remains unchanged, whenever $\|\delta\|_2 \leq r_{\text{cert}}(x, s)$, where

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

Here, $r_{\text{inp}}^{(k)}$ is the local radius for the feedforward map at layer $k$, and $P_{j_k, j_k-1}(\mathbf{W}^k)$ 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 histogram of the size of a particular reduced network $h_{\text{red}}$ at each input $x$. The reduced widths at each layer correspond to the choice of optimal sparsity level.
Reduced Lipschitz constant for regularized networks

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

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

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 \((d_2 - s_2)\).

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 \( 0 \leq s_2 \leq d_2 - 1 \).