The Pitfalls of Simplicity Bias in Neural Networks

Shah *et al*, NeurIPS 2020

Presented by: {Richard Zhu, Abishek Sridhar, Simon Seo, Rebecca Yu, Selina Carter}
(Group 1)
ML-715 (Fall 2022)

Carnegie Mellon University

Their claim: NNs lead to extreme simplicity bias.

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:
  - ✓ Simpler models generalize well: good results on test data

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:
  - ✓ Simpler models generalize well: good results on test data
    - ➡ simplicity = good 😇

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:
  - ✓ Simpler models generalize well: good results on test data
    - ➡ simplicity = good 😇
  - × But, NNs are **not "robust":** poor results on noisy data or outliers[1]

1: Paschali et al, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging", 2018 (https://arxiv.org/pdf/1804.00504.pdf)

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:
  - ✓ Simpler models generalize well: good results on test data
    - ➡ simplicity = good 😇
  - × But, NNs are **not "robust":** poor results on noisy data or outliers[1]
    - ➡ noisy data = bad 👿

1: Paschali et al, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging", 2018 (https://arxiv.org/pdf/1804.00504.pdf)

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:

🤔

✓ Simpler models generalize well: good results on test data

➡ simplicity = good 😇

× But, NNs are **not "robust":** poor results on noisy data or outliers[1]

➡ noisy data = bad 😈

1: Paschali et al, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging", 2018 (https://arxiv.org/pdf/1804.00504.pdf)

Their claim: NNs lead to extreme simplicity bias.

- Conundrum:

  ✓ Simpler models generalize well: good results on test data

  ➡ simplicity = good

  × But, NNs are **not "robust":** poor results on noisy data or outliers[1]

  ➡ noisy data = bad

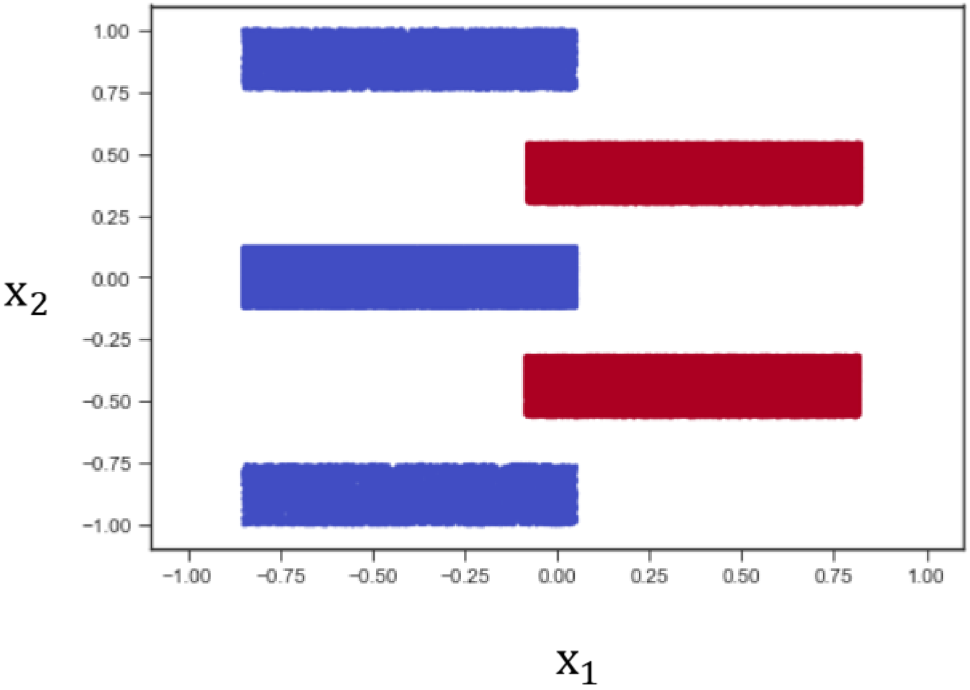
- The reason: NNs lead to "**extreme** simplicity bias," i.e., reliance *exclusively* on simple features, even when more complex features are better predictors.

1: Paschali et al, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging", 2018 (https://arxiv.org/pdf/1804.00504.pdf)

Example: How neural nets *over*-simplify

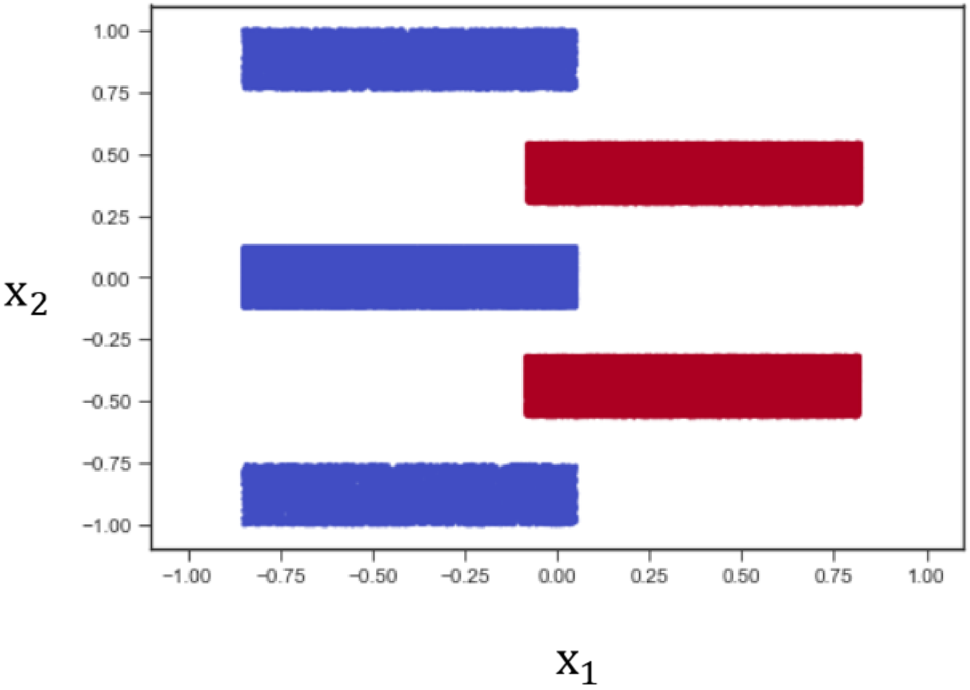
Example: How neural nets *over*-simplify

- Suppose we have $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^2$, $y_i \in \{\mathbf{0}, \mathbf{1}\}$



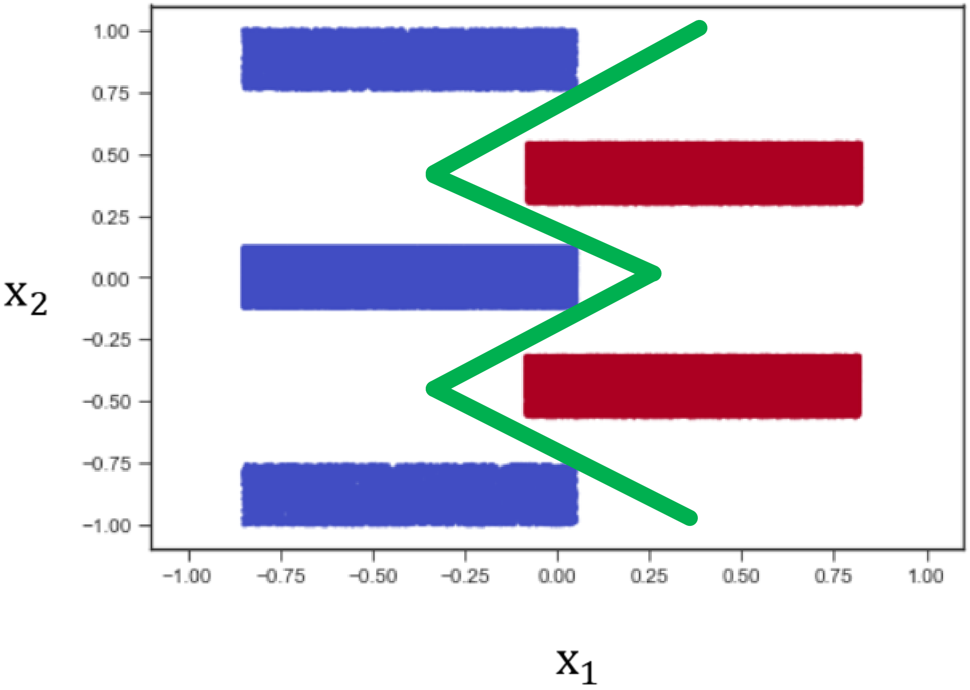
Example: How neural nets *over*-simplify

- Suppose we have $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^2$, $y_i \in \{0, 1\}$



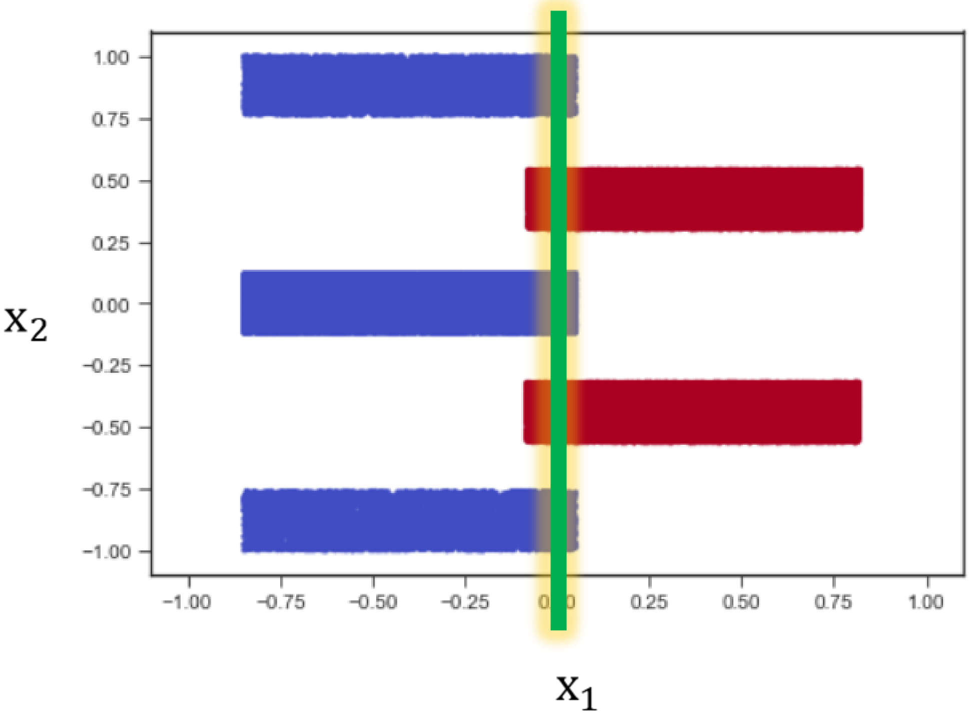
<u>Example</u>: How neural nets *over*-simplify

- Suppose we have $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^2$, $y_i \in \{0, 1\}$



Example: How neural nets *over*-simplify

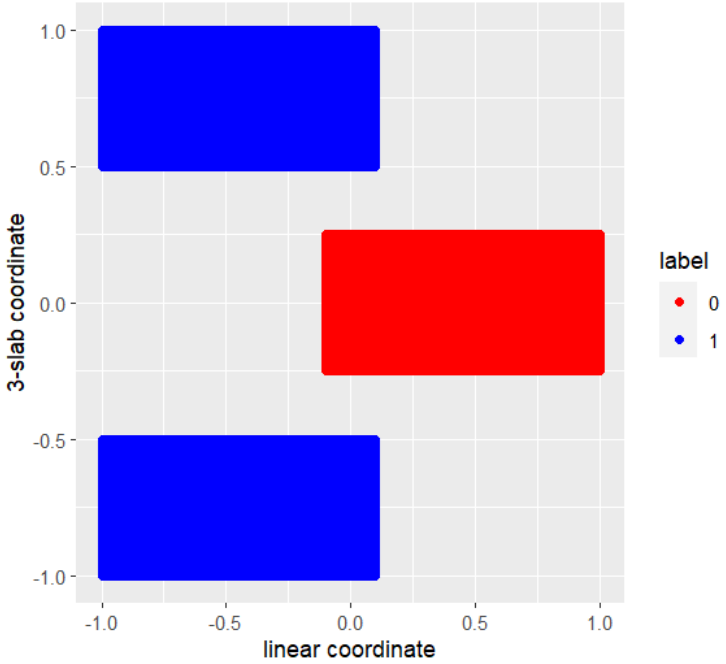
- Suppose we have $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^2$, $y_i \in \{0, 1\}$



Datasets used:

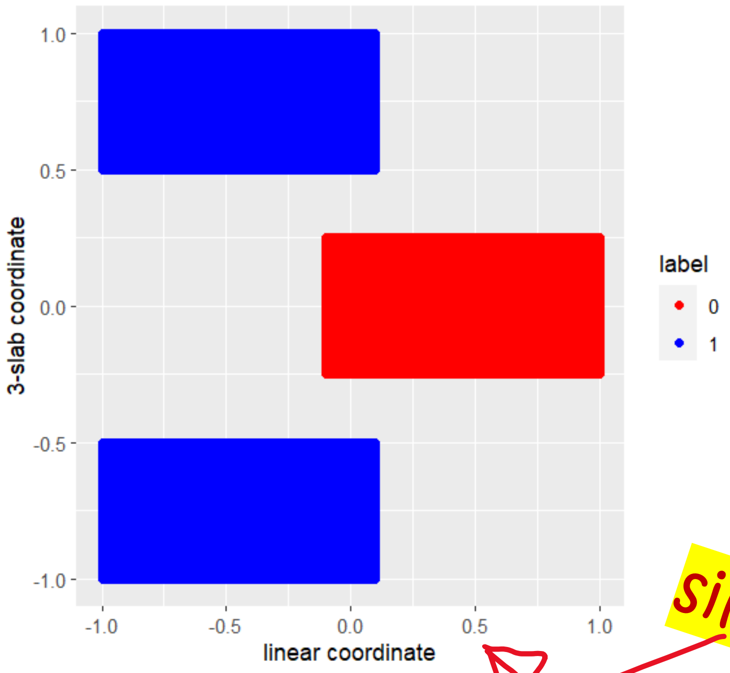
Datasets used:

LMS: Overlapped linear and "slab" features



Datasets used:

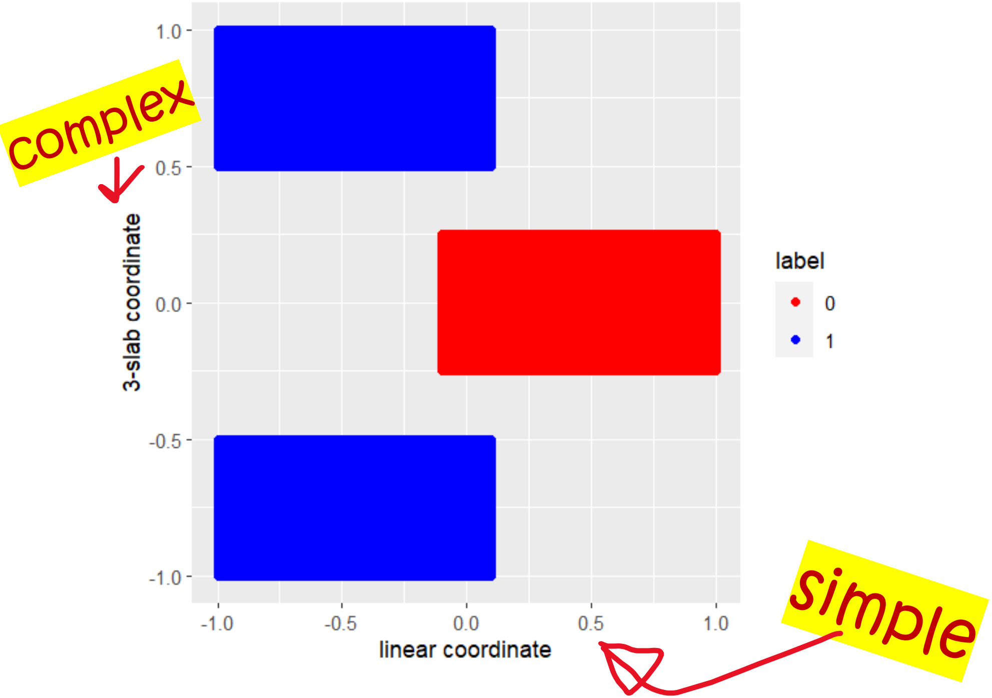
LMS: Overlapped linear and "slab" features



simple

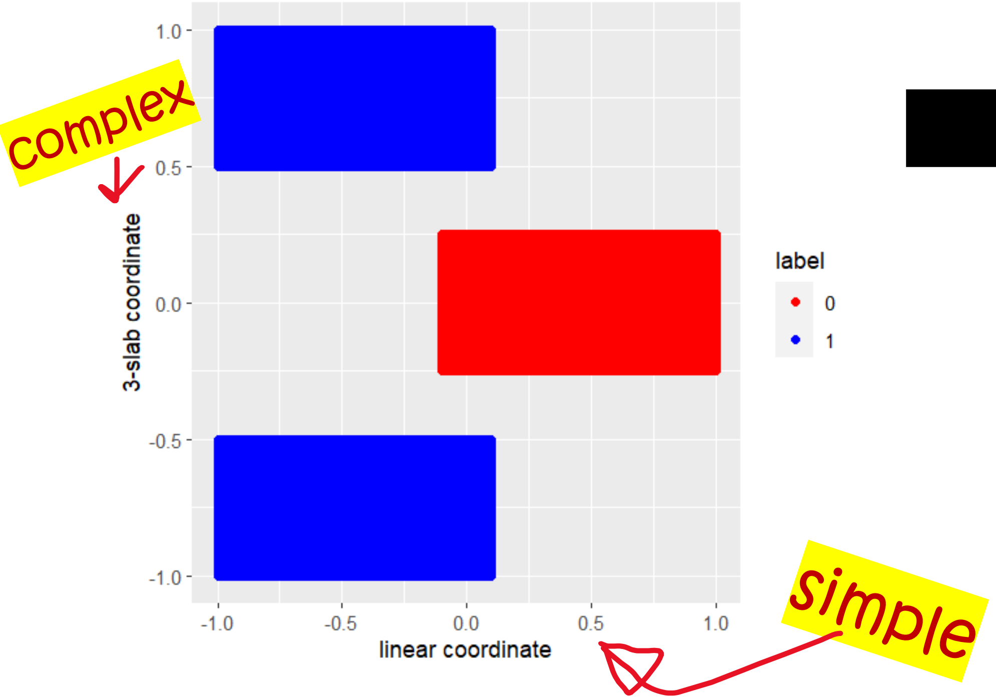
Datasets used:

LMS: Overlapped linear and "slab" features



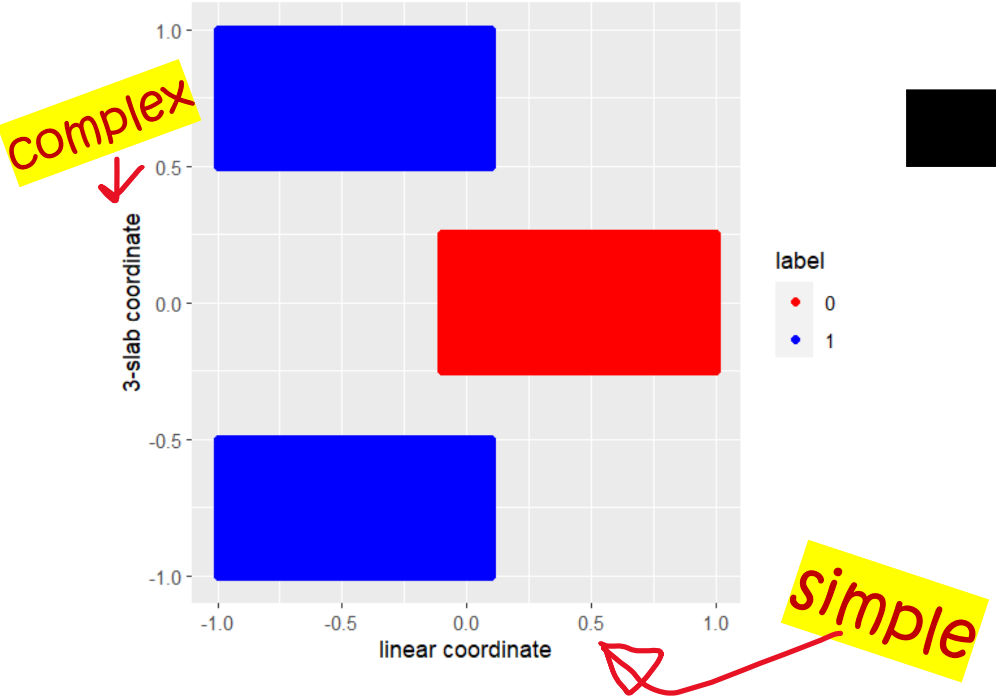
Datasets used:

LMS: Overlapped linear and "slab" features

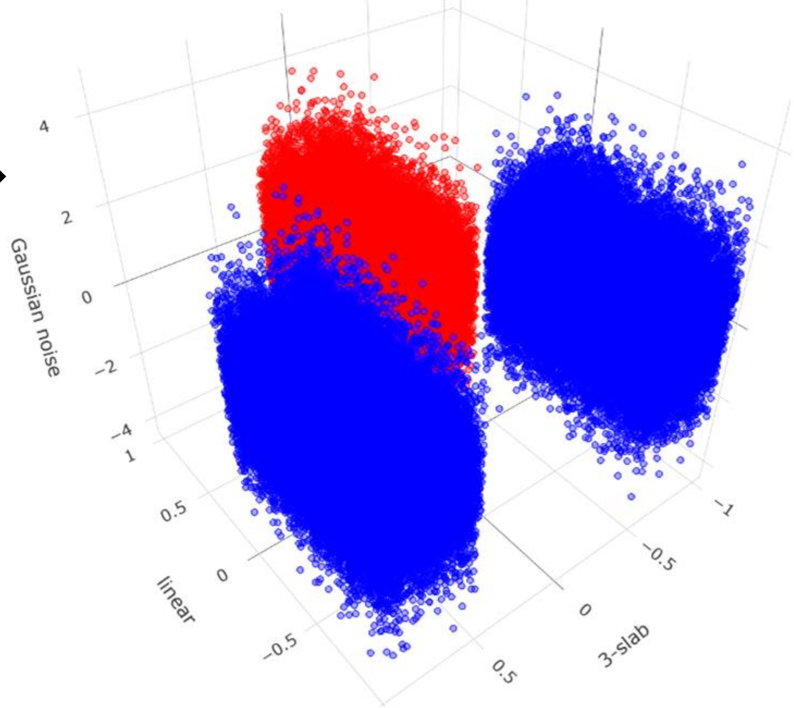


Datasets used:

LMS: Overlapped linear and "slab" features



LSN: Gaussian noise in remaining d-2 dimensions.



Theorem: Small NNs put more weight on the “simpler” feature

Theorem: Small NNs put more weight on the “simpler” feature

Assumptions

- One hidden-layer NN, k hidden neurons
- ReLU activation
- Dimension $d > \Omega(\sqrt{k} \log k)$ (dimension is **big**)
- Hinge loss
- SGD
- LSN data
- $O(1)$ iterations: single pass over training data

Theorem: Small NNs put more weight on the “simpler” feature

Assumptions

- One hidden-layer NN, k hidden neurons
- ReLU activation
- Dimension $d > \Omega\left(\sqrt{k}\log k\right)$ (dimension is **big**)
- Hinge loss
- SGD
- LSN data
- $O(1)$ iterations: single pass over training data

Theorem: Small NNs put more weight on the "simpler" feature

Assumptions

- One hidden-layer NN, k hidden neurons
- ReLU activation
- Dimension $d > \Omega\left(\sqrt{k}\log k\right)$ (dimension is **big**)
- Hinge loss
- SGD
- LSN data
- $O(1)$ iterations: single pass over training data

$\Rightarrow$ the learned weights are:

Theorem: Small NNs put more weight on the "simpler" feature

Assumptions

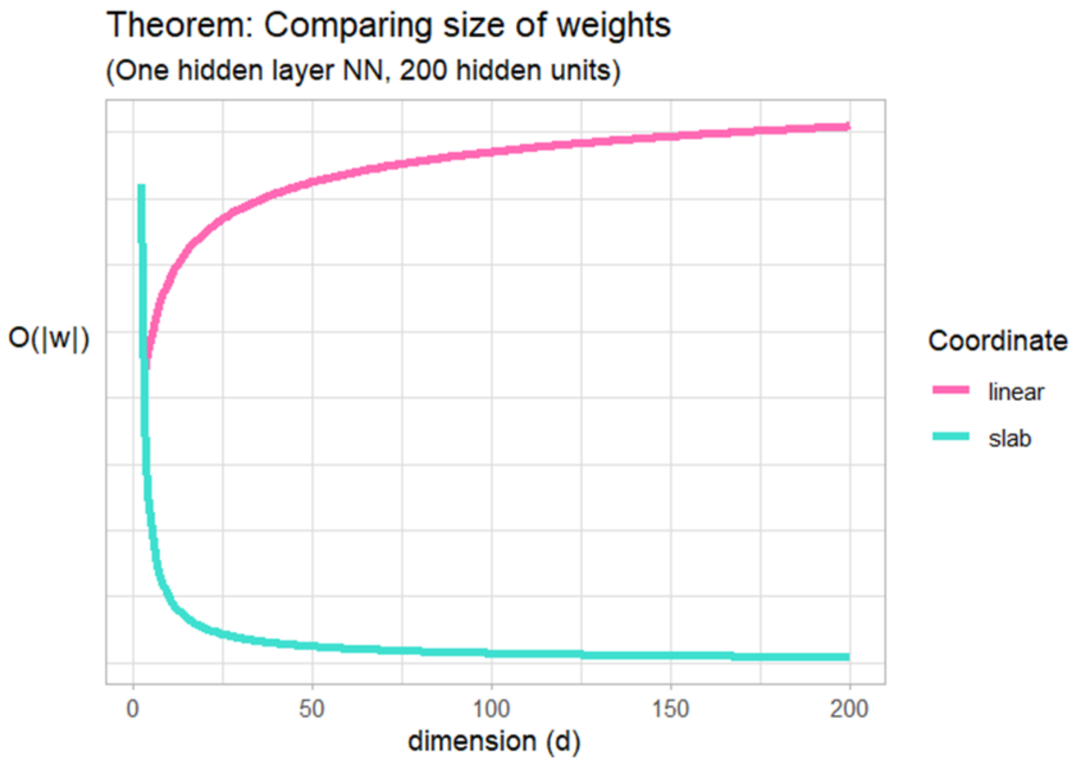
- One hidden-layer NN, k hidden neurons
- ReLU activation
- Dimension $d > \Omega\left(\sqrt{k}\log k\right)$ (dimension is **big**)
- Hinge loss
- SGD
- LSN data
- $O(1)$ iterations: single pass over training data

$\Longrightarrow$ the learned weights are:

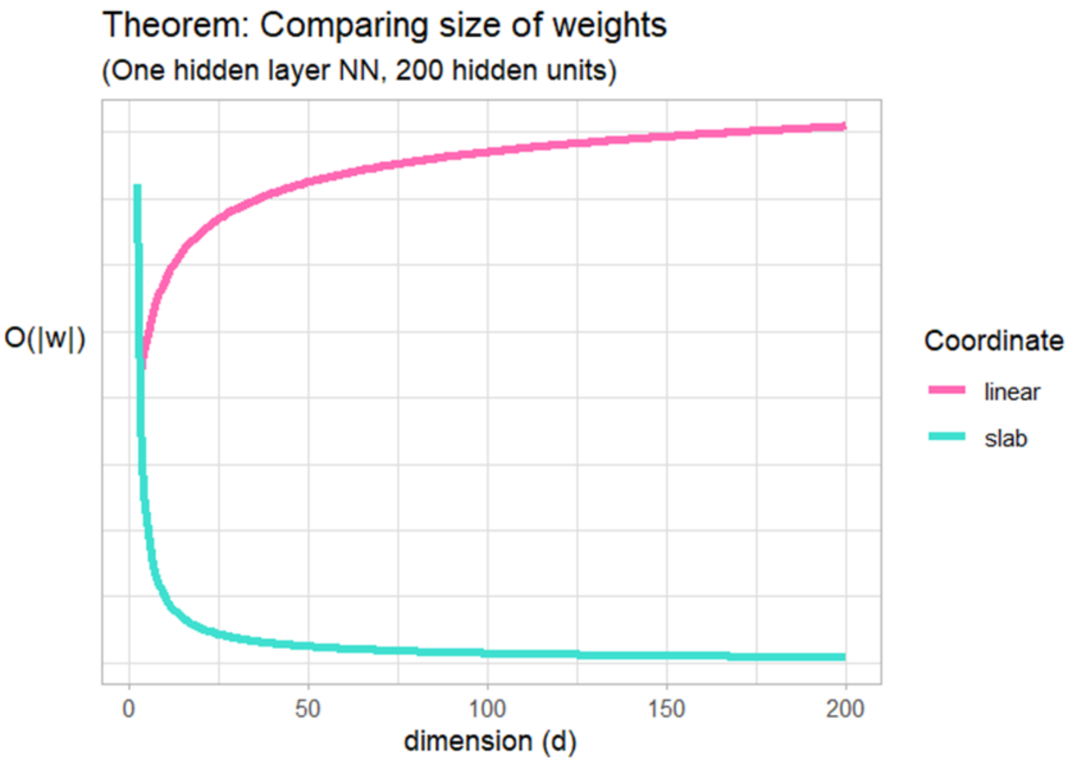
$$\underbrace{|w_{1j}| = \frac{2}{\sqrt{k}}\left(1 - \frac{c}{\sqrt{\log d}}\right) + O\left(\frac{1}{\sqrt{dk}\log d}\right)}_{\textit{Linear Coordinate}}, \quad \underbrace{|w_{2,j}| = O\left(\frac{1}{\sqrt{dk}\log d}\right)}_{\textit{Slab Coordinate}}, \quad \underbrace{\|w_{3:d,j}\| = O\left(\frac{1}{\sqrt{k}\log d}\right)}_{d-2\ \textit{Noise Coordinates}}$$

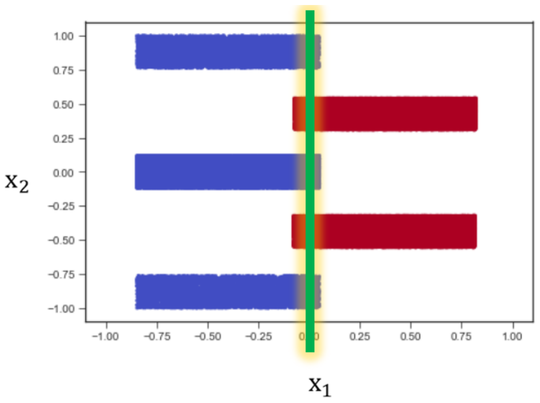
Theorem: Small NNs put more weight on the “simpler” feature

Theorem: Small NNs put more weight on the "simpler" feature



Theorem: Small NNs put more weight on the "simpler" feature





Dissection

❶ **Missing**: *No* experiments using LSN and real-world datasets

1: Ansuini et. al. "Intrinsic dimension of data representations in deep neural networks." NeurIPS 2019.

❶ **Missing**: *No* experiments using LSN and real-world datasets

- <u>The paper's goal</u>: bridge the gap between theory and practice. 🏆

❶ **Missing**: *No* experiments using LSN and real-world datasets

- The paper's goal: bridge the gap between theory and practice. 🏆
- But, the results might not hold for real-world datasets. 🤔
  - Real-world data have highly **correlated** features and noise.[1]
  - But datasets in the paper have many **independent** feature coordinates.

1: Ansuini et. al. "Intrinsic dimension of data representations in deep neural networks." NeurIPS 2019.

❶ **Missing**: *No* experiments using LSN and real-world datasets

- The paper's goal: bridge the gap between theory and practice.
- But, the results might not hold for real-world datasets.
  - Real-world data have highly **correlated** features and noise.[1]
  - But datasets in the paper have many **independent** feature coordinates.
- Their theorem uses LSN, but their experiments do not.
  - Paper leaves out certain failing cases.

1: Ansuini et. al. "Intrinsic dimension of data representations in deep neural networks." NeurIPS 2019.

❷ **Limitation**: Extreme SB occurs only in *large dimensions*

❷ **Limitation**: Extreme SB occurs only in *large dimensions*

Extreme SB is *not universal.*

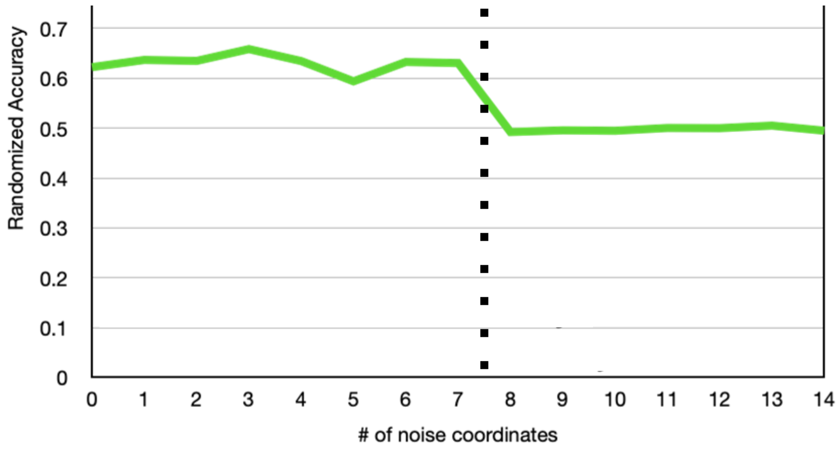
➡ **Observation**: extreme SB does *not* occur
for datasets with *small* dimensions.

❷ **Limitation**: Extreme SB occurs only in *large dimensions*

Extreme SB is *not universal.*

➡ **Observation**: extreme SB does *not* occur for datasets with *small* dimensions.

Accuracy on data with randomized linearly separable feature

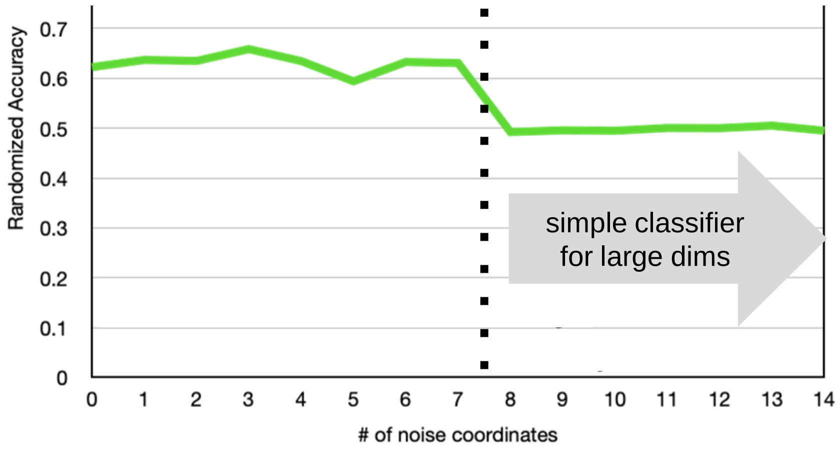


❷ **Limitation**: Extreme SB occurs only in *large dimensions*

Extreme SB is *not universal.*

➡ **Observation**: extreme SB does *not* occur for datasets with *small* dimensions.

Accuracy on data with randomized linearly separable feature

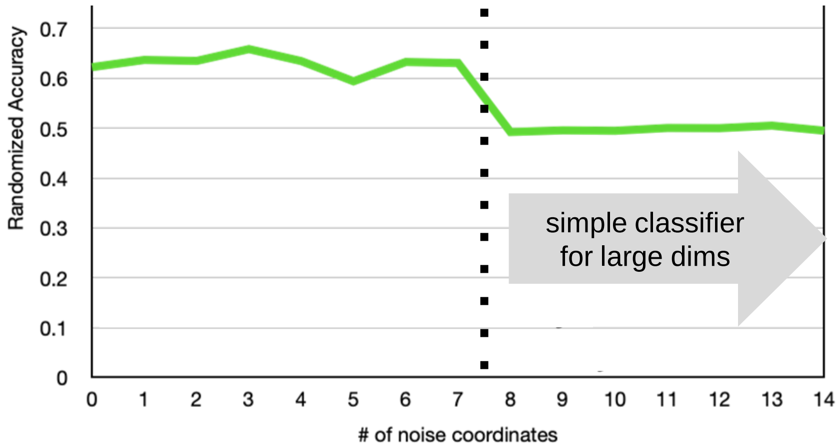


❷ **Limitation**: Extreme SB occurs only in *large dimensions*

Extreme SB is *not universal.*

➡ **Observation**: extreme SB does *not* occur for datasets with *small* dimensions.

Accuracy on data with randomized linearly separable feature



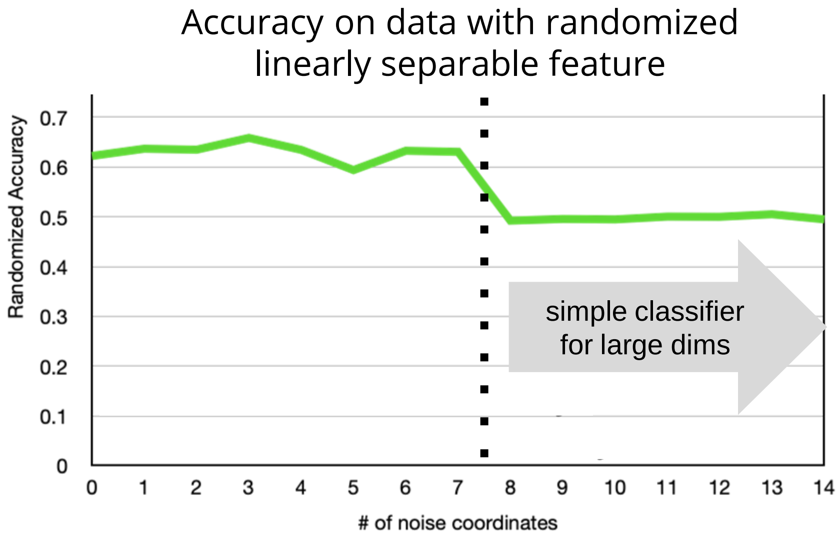
Why?

- ✶ NN used in paper was too small.
- ✶ Toy dataset had uncorrelated noise.

❷ **Limitation**: Extreme SB occurs only in *large dimensions*

Extreme SB is *not universal.*

➡ **Observation**: extreme SB does *not* occur for datasets with *small* dimensions.



Why? ✶ NN used in paper was too small.

✶ Toy dataset had uncorrelated noise.

Going forward… ➡ Explore theorems explaining extreme SB for smaller dimensions.

❸ **Questionable assumption**: Noise *must* be Gaussian

❸ **Questionable assumption**: Noise *must* be Gaussian

Observation:

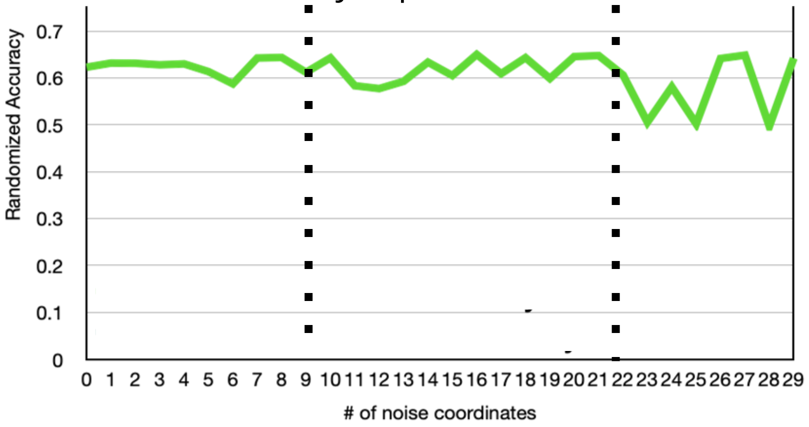
Changing noise distribution can reduce extreme SB.

❸ **Questionable assumption**: Noise *must* be Gaussian

Observation:

Changing noise distribution can reduce extreme SB.

Accuracy on data with randomized linearly separable feature

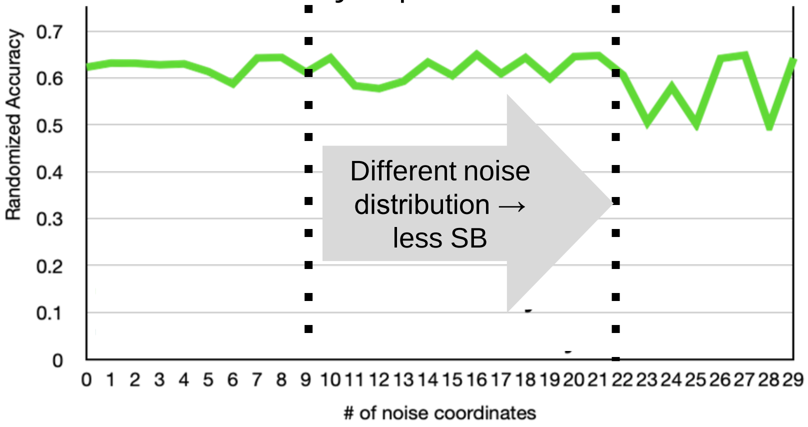


❸ **Questionable assumption**: Noise *must* be Gaussian

Observation:

Changing noise distribution can reduce extreme SB.

Accuracy on data with randomized linearly separable feature

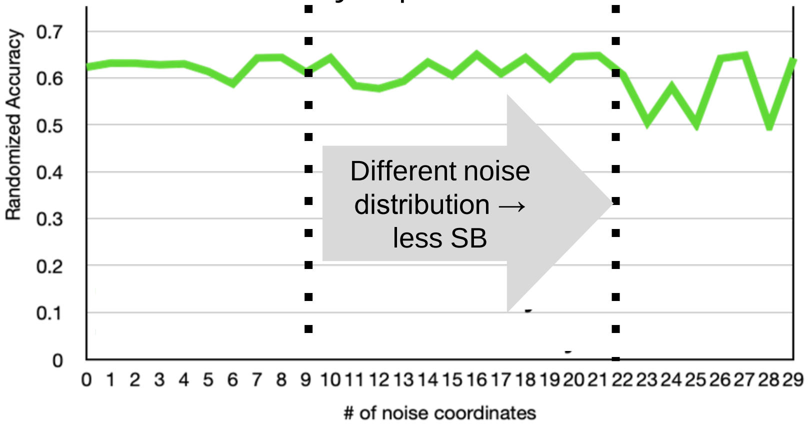


❸ **Questionable assumption**: Noise *must* be Gaussian

Observation:

Changing noise distribution can reduce extreme SB.

Accuracy on data with randomized linearly separable feature



Why?

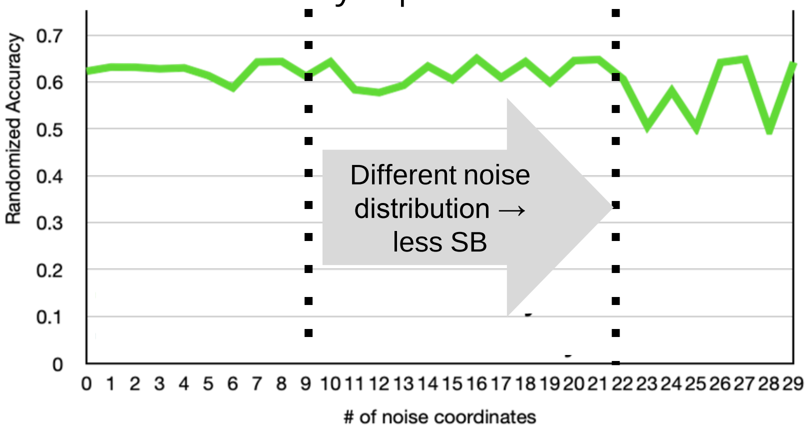
- ✶ Maybe due to high variance of noise.
- ✶ GD methods are sensitive to noise distribution for convergence.

❸ **Questionable assumption**: Noise *must* be Gaussian

Observation:

Changing noise distribution can reduce extreme SB.

Accuracy on data with randomized linearly separable feature



Why?

- ✶ Maybe due to high variance of noise.
- ✶ GD methods are sensitive to noise distribution for convergence.

Going forward…

- ➡ Explore effect of real-world noise distribution on extreme SB.
- ➡ Explore Gaussian noise removal techniques (like smoothening).

❹ **Questionable assumption**: NN must be *small*

❹ **Questionable assumption**: NN must be *small*

- Theorem: assumes 1 hidden layer NN.

❹ **Questionable assumption:** NN must be *small*

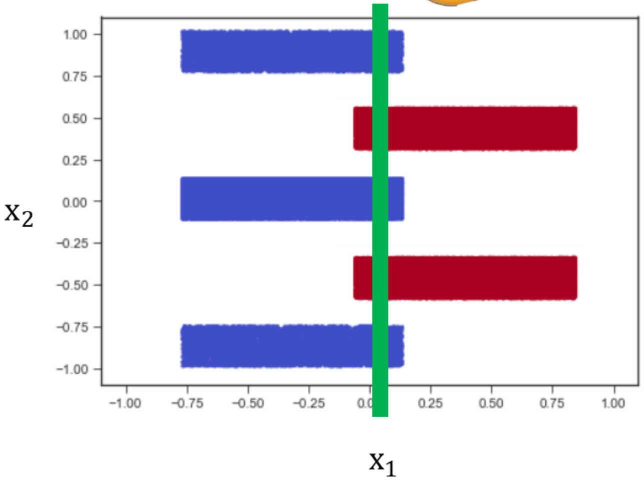
- Theorem: assumes 1 hidden layer NN.
- Experimental data, on LMS data:

❹ **Questionable assumption**: NN must be *small*

- Theorem: assumes 1 hidden layer NN.

- Experimental data, on LMS data:
  - **Authors**: (100,1)-FCN learns a simple classifier.

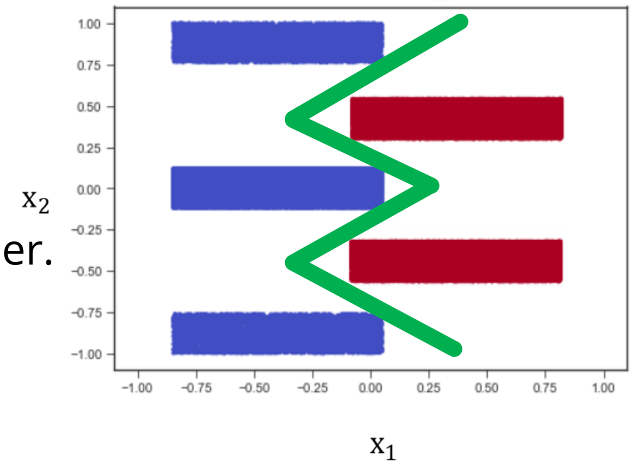
Theirs:



❹ **Questionable assumption**: NN must be *small*

- Theorem: assumes 1 hidden layer NN.
- Experimental data, on LMS data:
  - **Authors**: (100,1)-FCN learns a simple classifier.
  - **Ours**: (300, 2)-FCN learns a complex, perfect classifier.

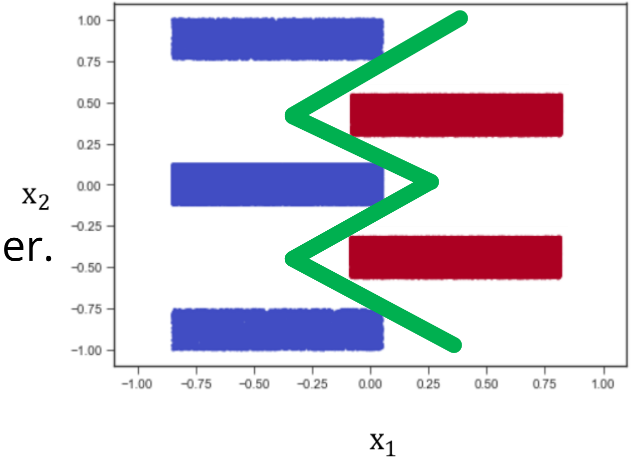
Ours:



❹ **Questionable assumption**: NN must be *small*

- Theorem: assumes 1 hidden layer NN.
- Experimental data, on LMS data:
  - **Authors**: (100,1)-FCN learns a simple classifier.
  - **Ours**: (300, 2)-FCN learns a complex, perfect classifier.

Ours:



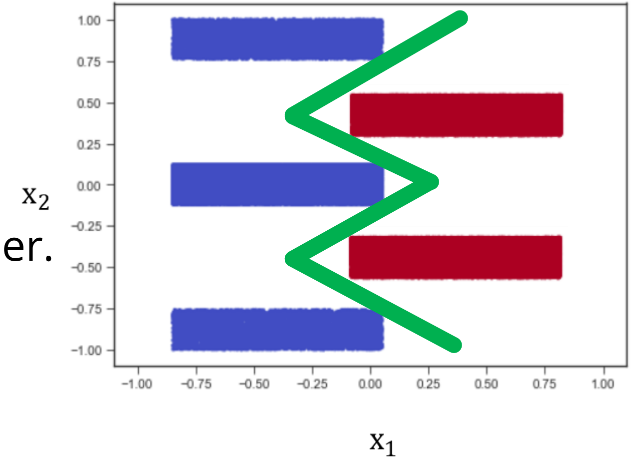
Why? ✶ "Complex" features become simpler to classify in deeper layers.[2]

2: Chakraborty et. al. "Constructing Energy-efficient Mixed-precision Neural Networks through Principal Component Analysis for Edge Intelligence." 2019.

❹ **Questionable assumption**: NN must be *small*

- Theorem: assumes 1 hidden layer NN.
- Experimental data, on LMS data:
  - **Authors**: (100,1)-FCN learns a simple classifier.
  - **Ours**: (300, 2)-FCN learns a complex, perfect classifier.

Ours:



Why? ✶ "Complex" features become simpler to classify in deeper layers.[2]

Going forward… ➡ Explore theorems explaining Extreme SB for larger NNs.

2: Chakraborty et. al. "Constructing Energy-efficient Mixed-precision Neural Networks through Principal Component Analysis for Edge Intelligence." 2019.

❺ **Unexplained**: Weights *after* the first training epoch

❺ **Unexplained**: Weights *after* the first training epoch

- Theorem shows results after a *single* epoch of training.

❺ **Unexplained**: Weights *after* the first training epoch

- Theorem shows results after a *single* epoch of training. 🤔

Why?

✶ Updates to certain feature weights can take precedence after several epochs (depending on loss surface).

❺ **Unexplained**: Weights *after* the first training epoch

- Theorem shows results after a *single* epoch of training.

Why?

✶ Updates to certain feature weights can take precedence after several epochs (depending on loss surface).

Going forward…

➡ Analyze what happens when NN is trained until convergence.

➡ Prove validity of theorem to subsequent stages of training.

Summary

	Critique	How and why this might happen	Future work and how to pursue this
1	They don't experiment with LSN data	✶ Mysterious	➡ Do experiment with LSN data
2	Result requires Gaussian noise	✶ High variance of noise ✶ GD fails to find optimal path	➡ Explore real-world noise distribution on extreme SB ➡ Gaussian noise removal techniques
3	Result requires large dimension	✶ Small model used in paper ✶ Toy dataset had uncorrelated noise	➡ Theorems explaining extreme SB for smaller dimensions
4	They assume a small NN	✶ "Complex" features become simpler to classify in deeper layers	➡ Explore theorems explaining extreme SB for larger NNs
5	Theorem assumed a single epoch	✶ Updates to certain feature weights can take precedence after several epochs	➡ Further training of NN ➡ Prove validity of theorem for more epochs