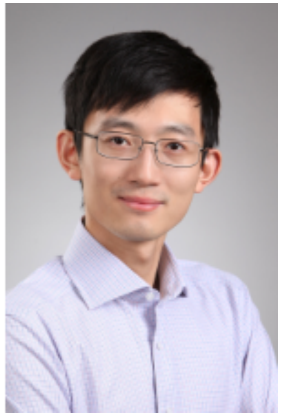
Data-Dependent Sample Complexities for Deep Neural Networks

Colin Wei Tengyu Ma

Stanford University



How do we design principled regularizers for deep models?

How do we design principled regularizers for deep models?

- Many regularizers are designed ad-hoc

How do we design principled regularizers for deep models?

- Many regularizers are designed ad-hoc
- A principled approach:
  - Theoretically prove upper bounds on generalization error

How do we design principled regularizers for deep models?

- Many regularizers are designed ad-hoc
- A principled approach:
  - Theoretically prove upper bounds on generalization error
  - Empirically regularize the upper bounds

How do we design principled regularizers for deep models?

- Many regularizers are designed ad-hoc
- A principled approach:
  - Theoretically prove upper bounds on generalization error
  - Empirically regularize the upper bounds
- Bottleneck in prior work:
  - Mostly considers norm of weights

[Bartlett et. al’17, Neyshabur et. al’17, Nagarajan and Kolter’19]

How do we design principled regularizers for deep models?

- Many regularizers are designed ad-hoc
- A principled approach:
  - Theoretically prove upper bounds on generalization error
  - Empirically regularize the upper bounds
- Bottleneck in prior work:
  - Mostly considers norm of weights
  - $\Rightarrow$ Loose/pessimistic bounds (e.g., exponential in depth)

[Bartlett et. al’17, Neyshabur et. al’17, Nagarajan and Kolter’19]

Data-Dependent Generalization Bounds

Data-Dependent Generalization Bounds

$$\text{generalization} \leq g(\text{weights}, \text{training data})$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Data-Dependent Generalization Bounds

$$\text{generalization} \leq g(\text{weights}, \text{training data})$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

Data-Dependent Generalization Bounds

$$\text{generalization} \leq g(\text{weights}, \text{training data})$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Jacobian norm = max norm of the Jacobian of model w.r.t hidden layers on training data

Data-Dependent Generalization Bounds

$$\text{generalization} \leq g(\text{weights, training data})$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Jacobian norm = max norm of the Jacobian of model w.r.t hidden layers on training data
- Hidden layer norm = max norm of hidden activation layer on training data

Data-Dependent Generalization Bounds

$$\text{generalization} \leq g(\text{weights, training data})$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Jacobian norm = max norm of the Jacobian of model w.r.t hidden layers on training data
- Hidden layer norm = max norm of hidden activation layer on training data
- Margin = largest logit – second largest logit

Data-Dependent Generalization Bounds

generalization ≤ g(weights, training data)

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Measures stability/Lipschitzness of the network around training examples

Data-Dependent Generalization Bounds

$$\text{generalization} \leq \text{g(weights, training data)}$$

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Measures stability/Lipschitzness of the network around training examples
  - Prior works consider worst-case stability over all inputs ⇒ exponential depth dependency [Bartlett et. al'17, Neyshabur et. al'17, etc.]

[Bartlett et. al'17, Neyshabur et. al'17, etc.]

Data-Dependent Generalization Bounds

generalization ≤ g(weights, training data)

- Add $g(\cdot)$ to the loss as an explicit regularizer

Theorem (informal):

$$g(\cdot) = \frac{\text{jacobian norm} \cdot \text{hidden layer norm}}{\text{margin} \sqrt{\text{train set size}}} + \text{low-order terms}$$

- Measures stability/Lipschitzness of the network around training examples
  - Prior works consider worst-case stability over all inputs ⇒ exponential depth dependency [Bartlett et. al'17, Neyshabur et. al'17, etc.]
  - Noise stability also studied in [Arora et. al'19, Nagarajan and Kolter'19] with looser bounds

[Bartlett et. al'17, Neyshabur et. al'17, etc.]

Regularizing our Bound

Regularizing our Bound

- Penalize squared Jacobian norm in loss

Regularizing our Bound

- Penalize squared Jacobian norm in loss
  - Hidden layer controlled by normalization layers (BatchNorm, LayerNorm)

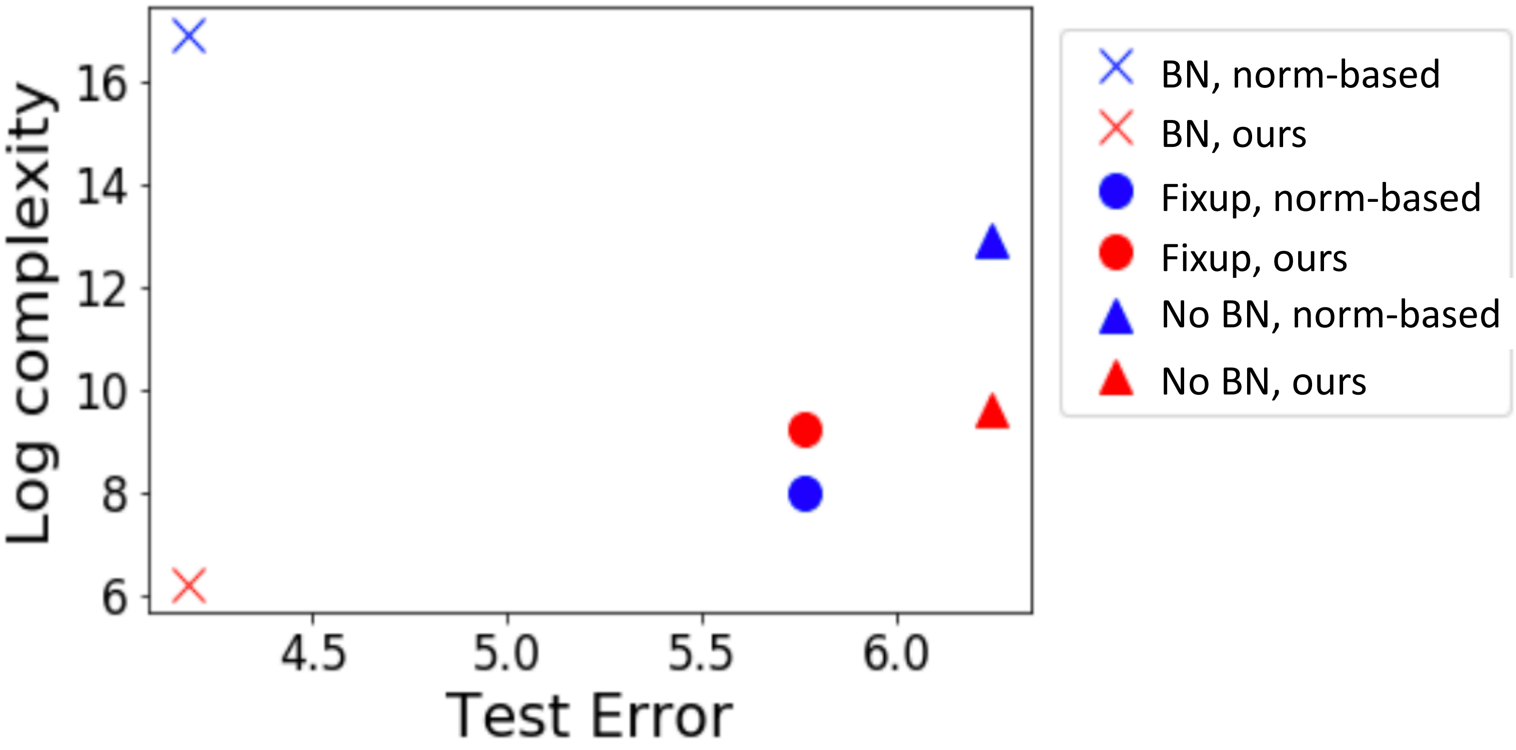
Regularizing our Bound

- Penalize squared Jacobian norm in loss
  - Hidden layer controlled by normalization layers (BatchNorm, LayerNorm)
- Helps in variety of settings which lack regularization compared to baseline

Setting	Normalization	Jacobian Reg	Test Error
Low learning rate (0.01)	BatchNorm	×	5.98%
		✓	**5.46%**
No data augmentation	BatchNorm	×	10.44%
		✓	**8.25%**
No BatchNorm	None	×	6.65%
	LayerNorm [Ba et al., 2016]	×	6.20%
		✓	**5.57%**

Correlation of our Bound with Test Error

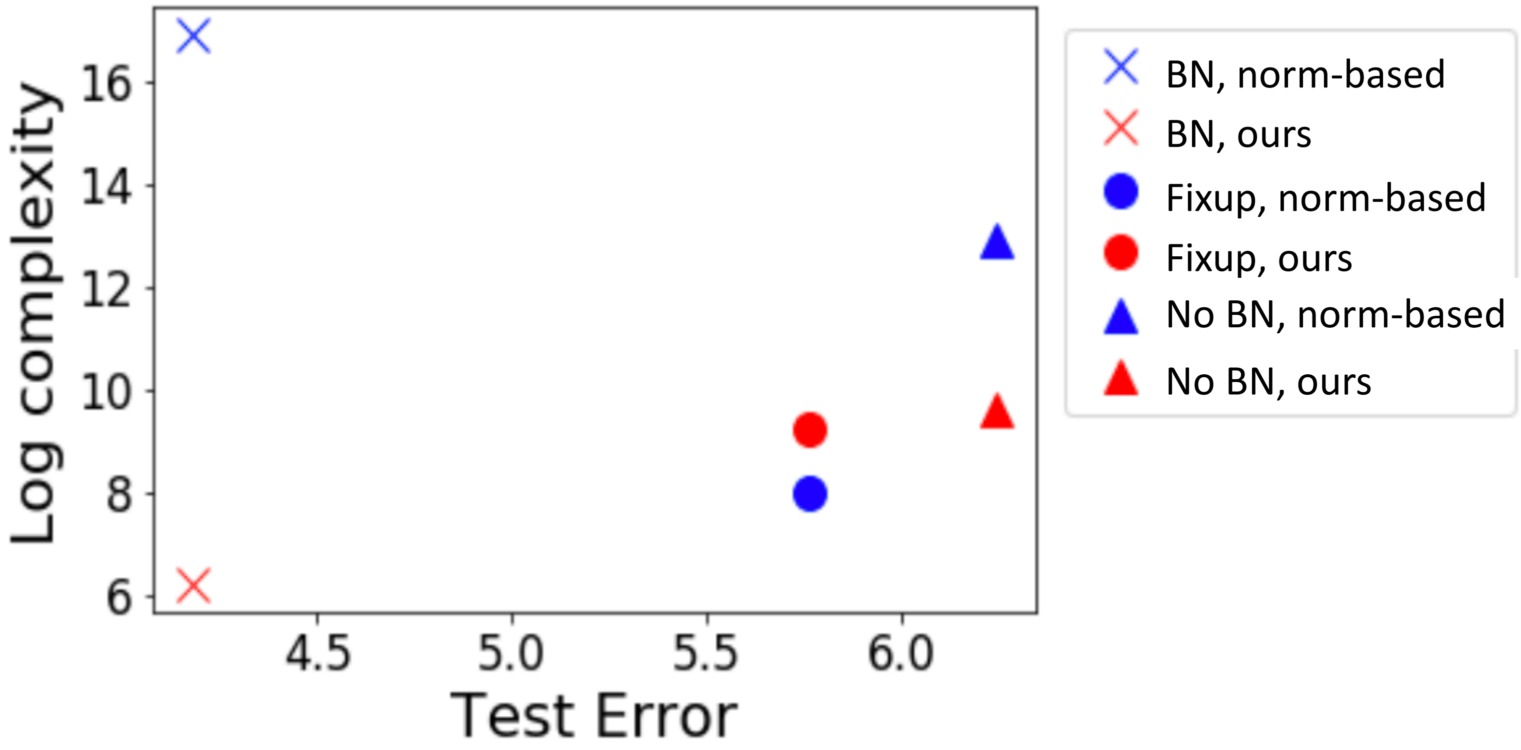
- Ours (red) vs. norm-based bound (blue) [Bartlett et. al'17]



[Fixup: Zhang et. al'19]

Correlation of our Bound with Test Error

- Ours (red) vs. norm-based bound (blue) [Bartlett et. al'17]

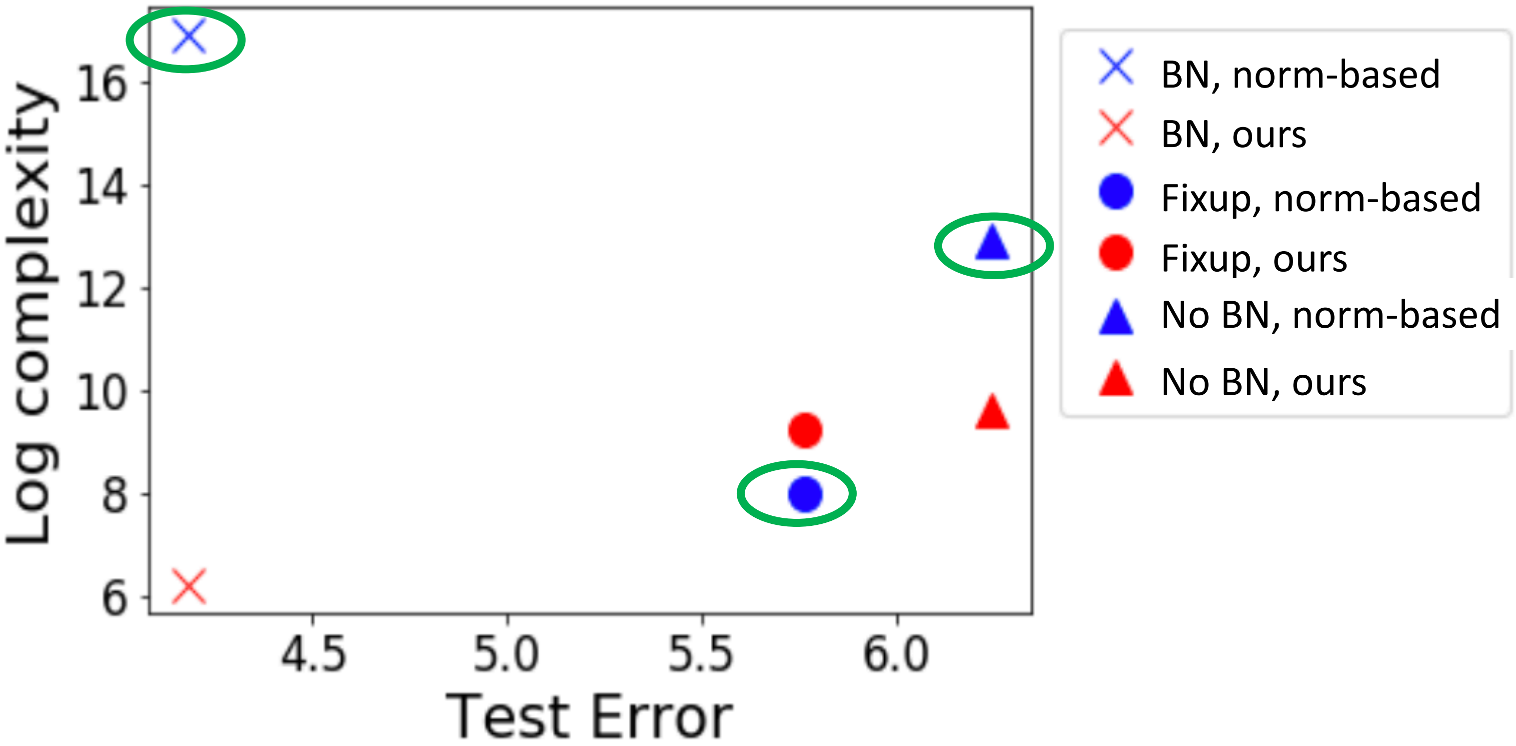


- Our bound correlates better with test error

[Fixup: Zhang et. al'19]

Correlation of our Bound with Test Error

- Ours (red) vs. norm-based bound (blue) [Bartlett et. al'17]

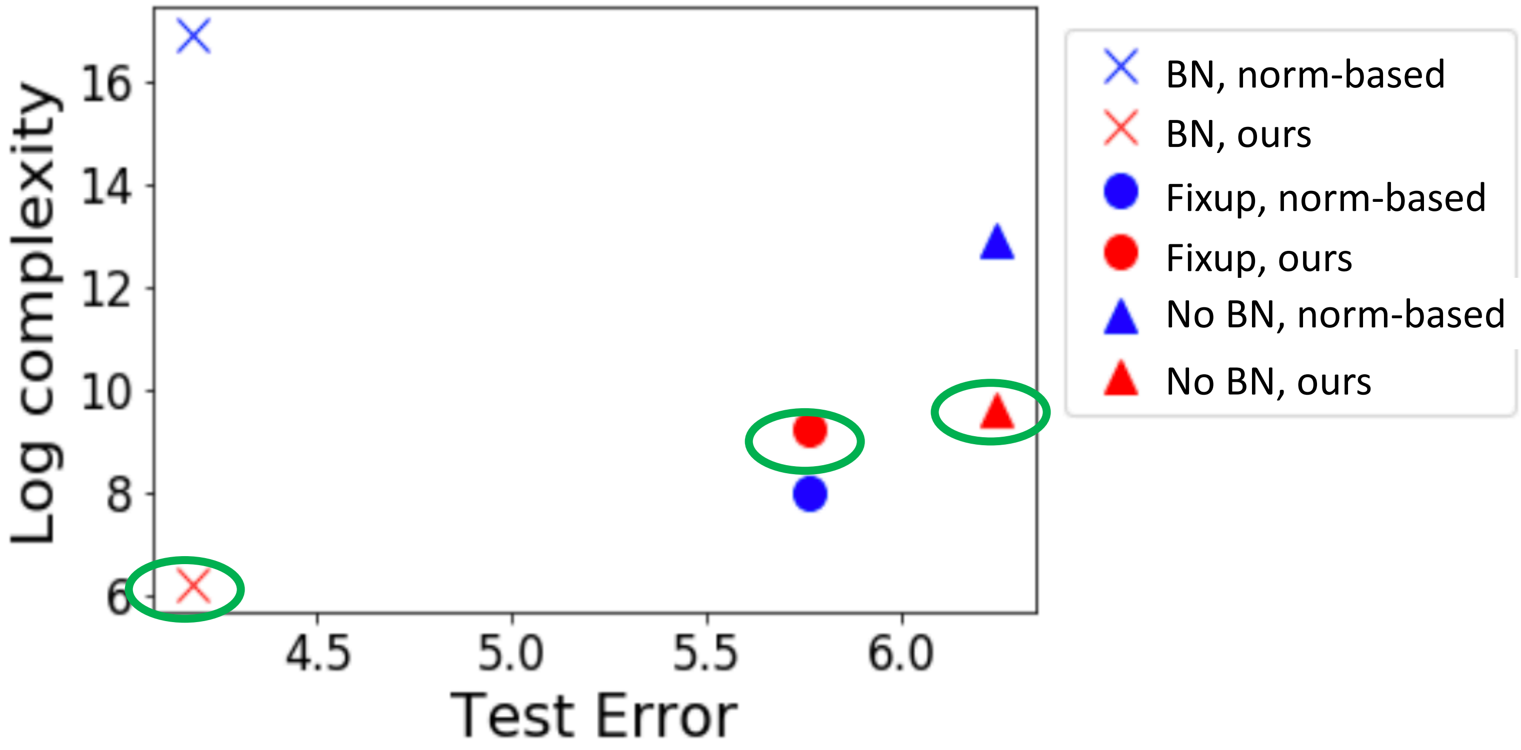


- Our bound correlates better with test error

[Fixup: Zhang et. al'19]

Correlation of our Bound with Test Error

- Ours (red) vs. norm-based bound (blue) [Bartlett et. al'17]



- Our bound correlates better with test error

[Fixup: Zhang et. al'19]

Conclusion

Conclusion

- Tighter bounds by considering data-dependent properties (stability on training data)

Conclusion

- Tighter bounds by considering data-dependent properties (stability on training data)
- Our bound avoids exponential dependencies on depth

Conclusion

- Tighter bounds by considering data-dependent properties (stability on training data)
- Our bound avoids exponential dependencies on depth
- Optimizing this bound improves empirical performance

Conclusion

- Tighter bounds by considering data-dependent properties (stability on training data)
- Our bound avoids exponential dependencies on depth
- Optimizing this bound improves empirical performance
- **Follow up work:** tighter bounds and empirical improvement over strong baselines
  - Works for both robust and clean accuracy

[Wei and Ma'19, "Improved Sample Complexities for Deep Networks and Robust Classification via an All-Layer Margin"]

Conclusion

- Tighter bounds by considering data-dependent properties (stability on training data)
- Our bound avoids exponential dependencies on depth
- Optimizing this bound improves empirical performance
- **Follow up work:** tighter bounds and empirical improvement over strong baselines
  - Works for both robust and clean accuracy

[Wei and Ma'19, "Improved Sample Complexities for Deep Networks and Robust Classification via an All-Layer Margin"]

Come find our poster: 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #220!