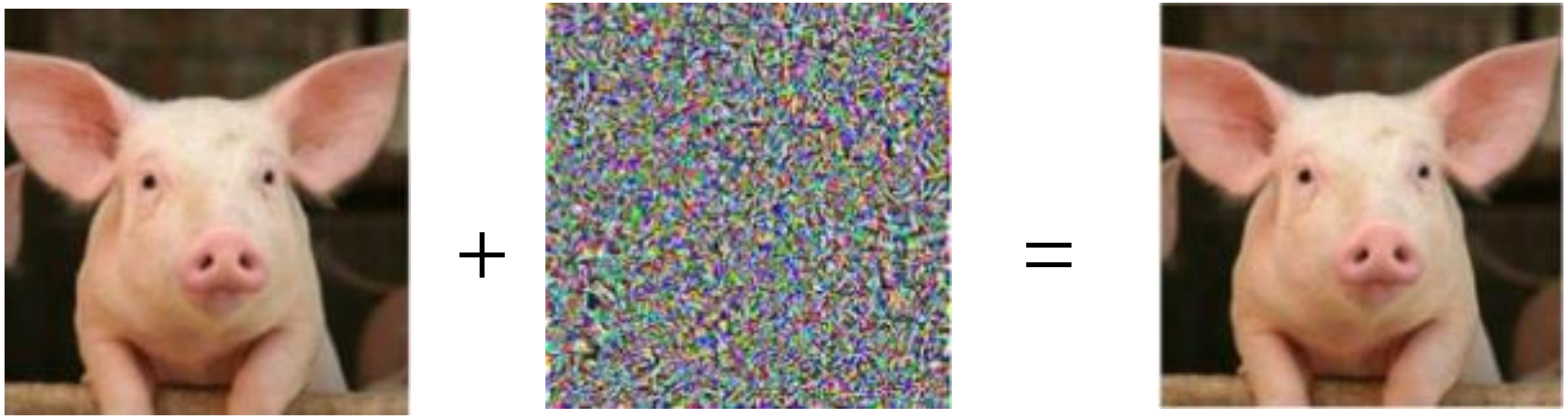
Towards Understanding the Regularization of Adversarial Robustness on Neural Network

Yuxin Wen*, Shuai Li*, Kui Jia

South China University of Technology (SCUT)

* indiates equal contribution

Adversarial examples



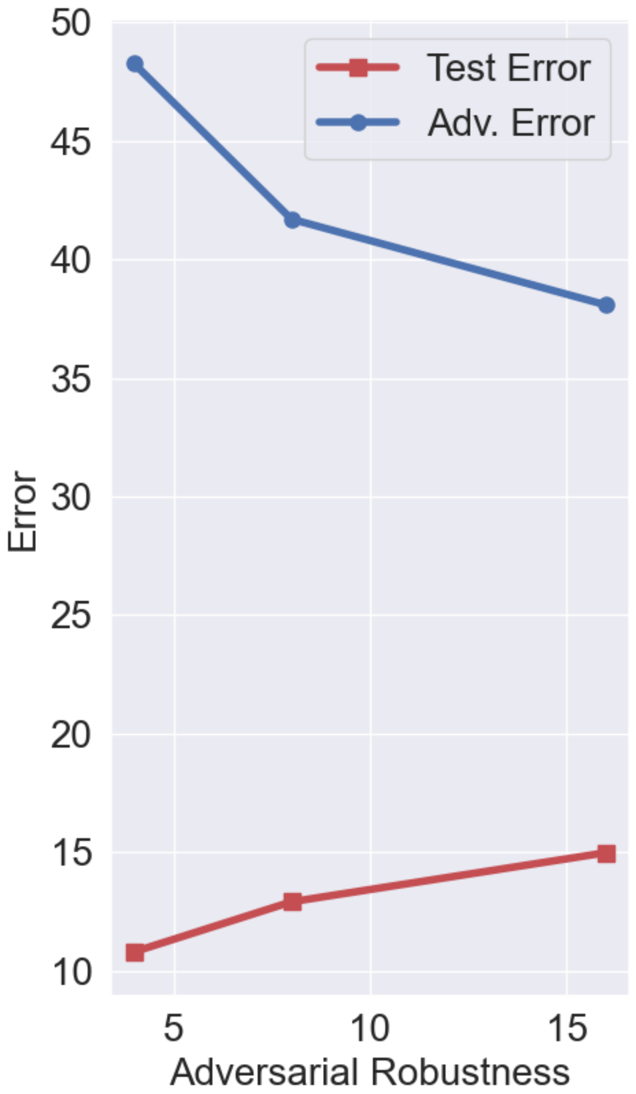
pig texture changes airliner

Adversarial training

adversarial examples generated via
multi-step method

used for training

Puzzling oberservations

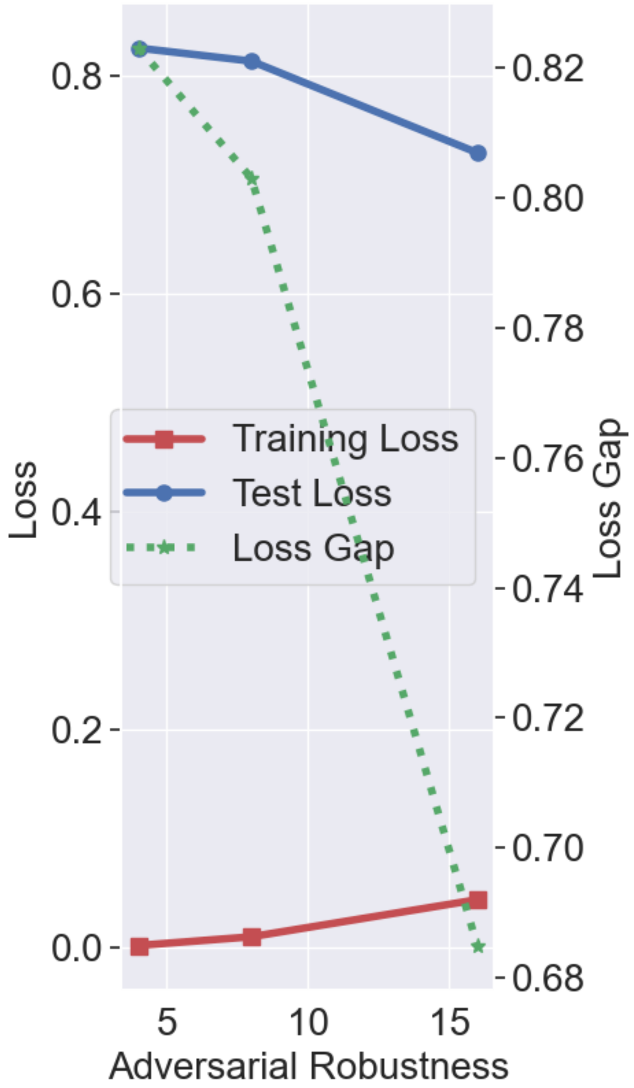


There exists a trade-off between adversarial robustness and accuracy

Puzzling oberservations

Adversarial training exacerbate overfitting (the error gap is enlarged) ?

Puzzling oberservations



Or an effective regularization (loss gap is reduced) ?

We study why effective regularization of adversarial robustness leads to poorer performance.

Key results

- Establish a generalization bound that characterizes the generalization errors through the **margin**, **adversarial robustness radius** and **singular values of weight matrices** of neural networks.
- Empirical results:
  - For NNs with **high adversarial robustness**, the **singular values of weight matrices** has **low variance**.
  - The reduced variance of singular values of weight matrices results in **concentration of examples** around decision boundaries.
- The concentration of examples around decision boundaries smoothens **sudden changes** induced by perturbations, but also **increases indecisive misclassifications**.

The generalization bound

Theorem 3.1. *Let T denote a NN with ReLU and MaxPooling nonlinear activation functions (the definition is put at eq. (6) for readers' convenience), l_γ the ramp loss defined at definition 4, and $\mathcal{Z}$ the instance space assumed in section 3. Assume that $\mathcal{Z}$ is a k-dimensional regular manifold that accepts an ϵ-covering with covering number $(\frac{C_\mathcal{X}}{\epsilon})^k$, and assumption assumption 3.1 holds. If T is ϵ_0-adversarially robust (defined at definition 2), $\epsilon \leq \epsilon_0$, and denote $v_{\min}$ the smallest IM margin in the covering balls that contain training examples (defined at definition 6), $\sigma^i_{\min}$ the smallest singular values of weight matrices $\mathbf{W}_i, i = 1, \dots, L-1$ of a NN, $\{w_i\}_{i=1,\dots,|\mathcal{Y}|}$ the set of vectors made up with ith rows of $\mathbf{W}_L$ (the last layer's weight matrix), then given an i.i.d. training sample $S_m = \{z_i = (x_i, y_i)\}_{i=1}^m$ drawn from $\mathcal{Z}$, its generalization error $GE(l \circ T)$ (defined at eq. (1)) satisfies that, for any $\eta > 0$, with probability at least $1 - \eta$*

$$GE(l_\gamma \circ T) \leq \max\left\{0, 1 - \frac{u_{\min}}{\gamma}\right\} + \sqrt{\frac{2\log(2) C_X^k}{\varepsilon^k m} + \frac{2\log(1/\eta)}{m}}$$

where $u_{\min} = \min_{y \neq \hat{y}} \left\| w_y - w_{\hat{y}} \right\|_2 \prod_{i=1}^{L-1} \sigma^i_{\min} v_{\min}$

The generalization bound

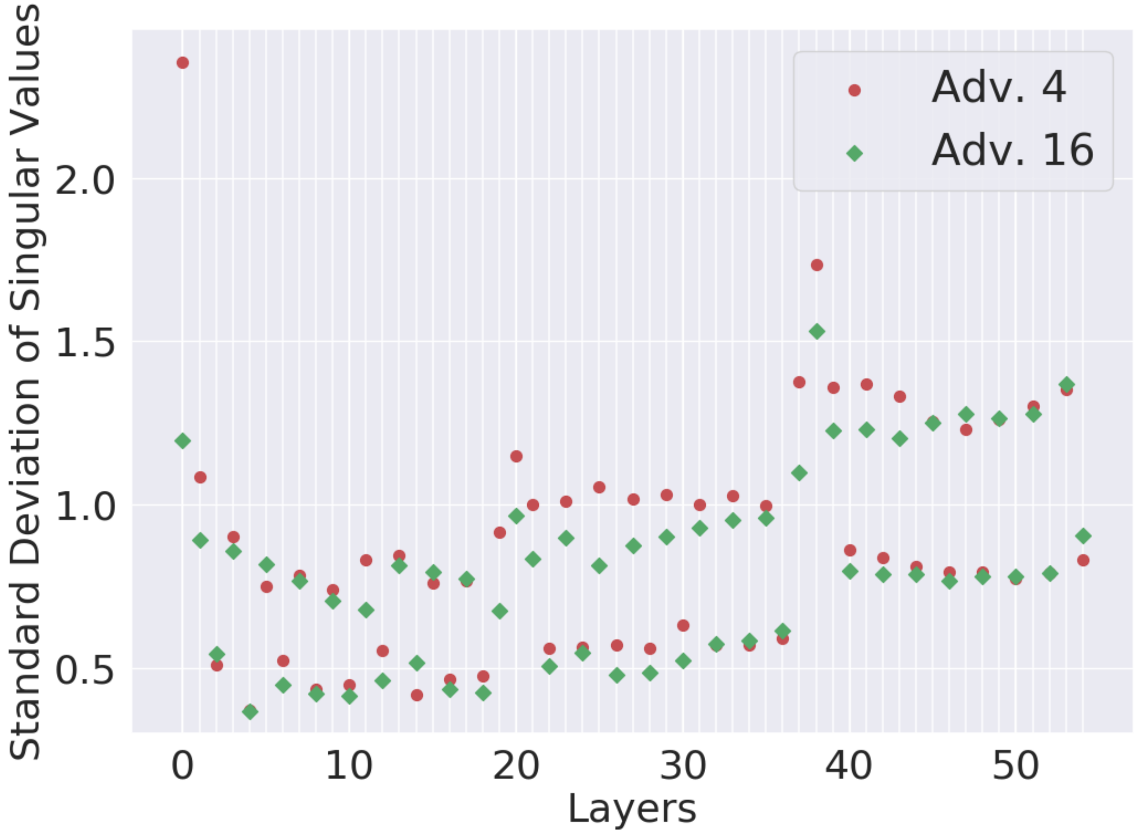
smallest margin on feature space: $$u_{\min} = \min_{y\neq\hat{y}} \overbrace{\left\|w_y - w_{\hat{y}}\right\|_2}^{\text{last layers' weight matrix}} \prod_{i=1}^{L-1} \overbrace{\sigma_{\min}^{i}}^{\text{smallest singular values of neural networks' weight matrices}} \underbrace{v_{\min}}_{\text{smallest instance-space margin, depends on the adversarial robustness radius}}$$

(an example of margin : $f_y(x) - \max_{i\neq y} f_i(x)$)

$$GE(l_\gamma \circ T) \leq \max\left\{0, 1 - \frac{\overbrace{u_{\min}}}{\underbrace{\gamma}_{\text{constant value}}}\right\} + \underbrace{\sqrt{\frac{2\log(2)C_X^k}{\varepsilon^k m} + \frac{2\log(1/\eta)}{m}}}_{\text{standard term}}$$

Reduced variance of singular values

(a) Results from CIFAR10

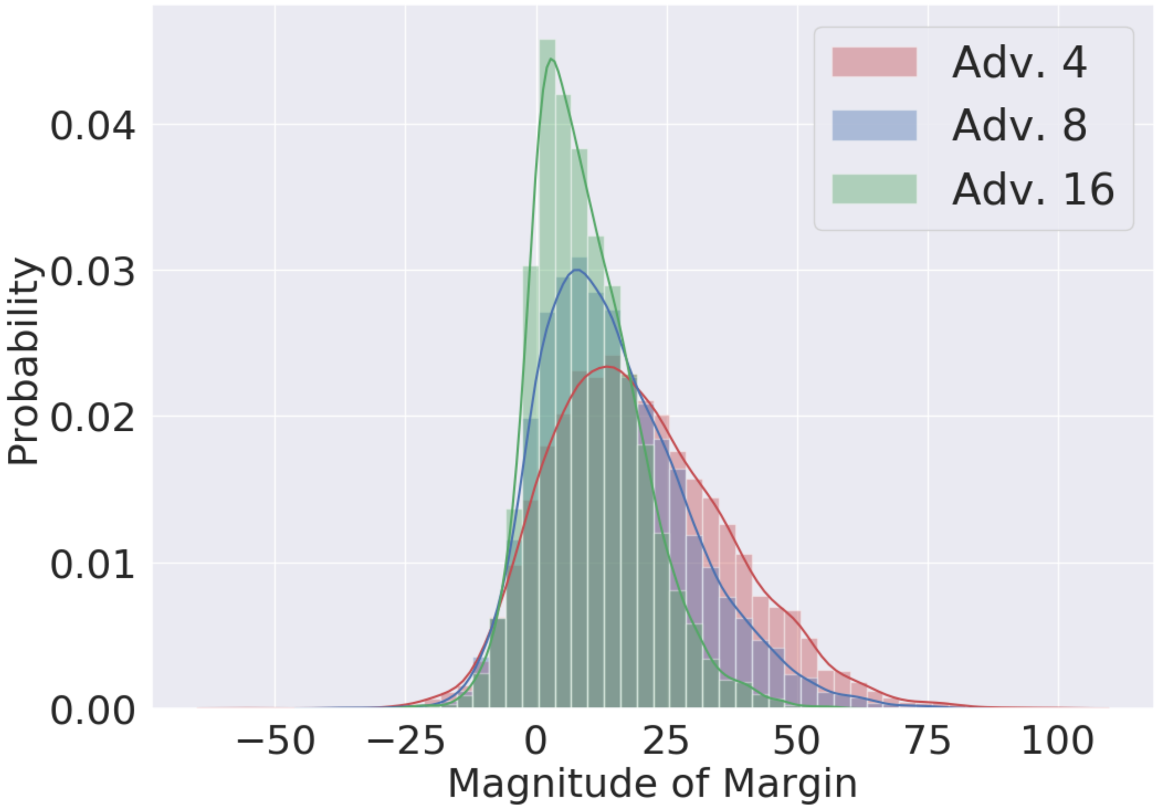


$$u_{\min} = \min_{y \neq \hat{y}} \|w_y - w_{\hat{y}}\|_2 \prod_{i=1}^{L-1} \sigma_{\min}^{i} \nu_{\min}$$

Take away message: stronger adversarial robustness reduces variance of singular values; and the receded variance results in reduced variance of the norms of the activation outputs.

Concentration of margin

(a) Results from CIFAR10



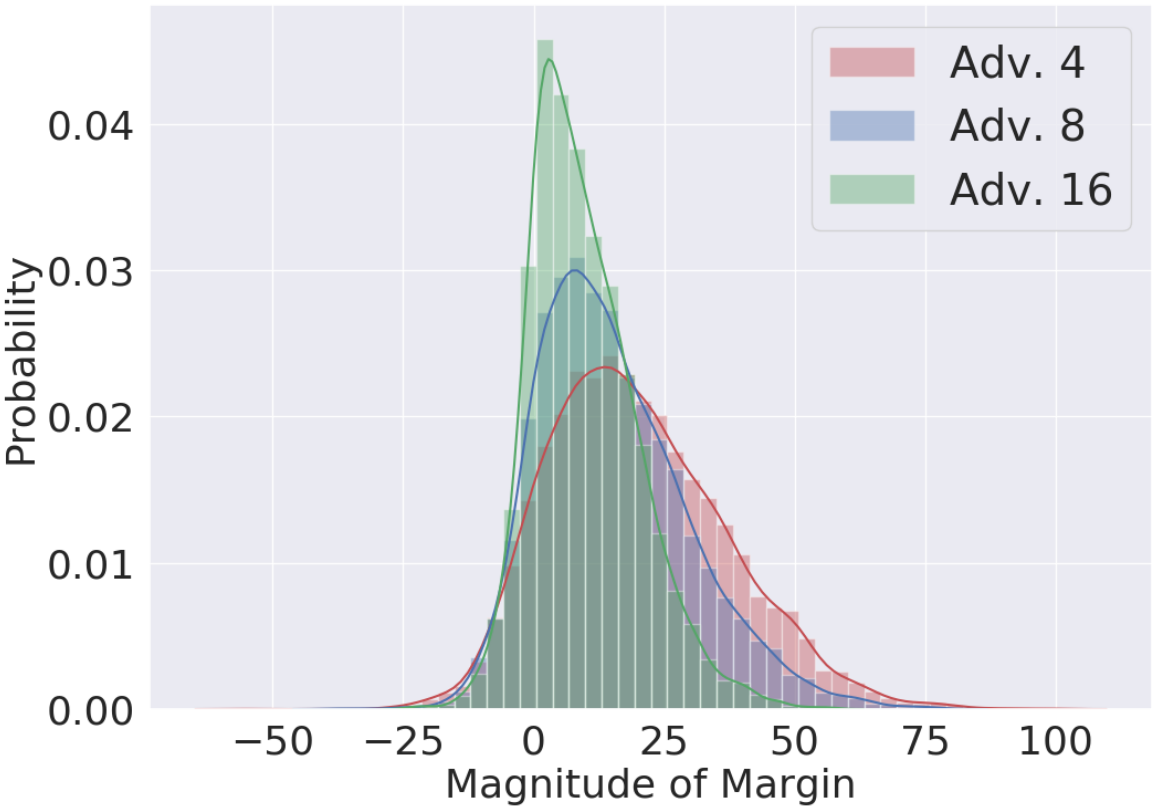
$$GE(l_\gamma \circ T) \leq \max\left\{0, 1 - \frac{u_{\min}}{\gamma}\right\}$$

$$+\sqrt{\frac{2\log(2)C_X^k}{\varepsilon^k m} + \frac{2\log(1/\eta)}{m}}$$

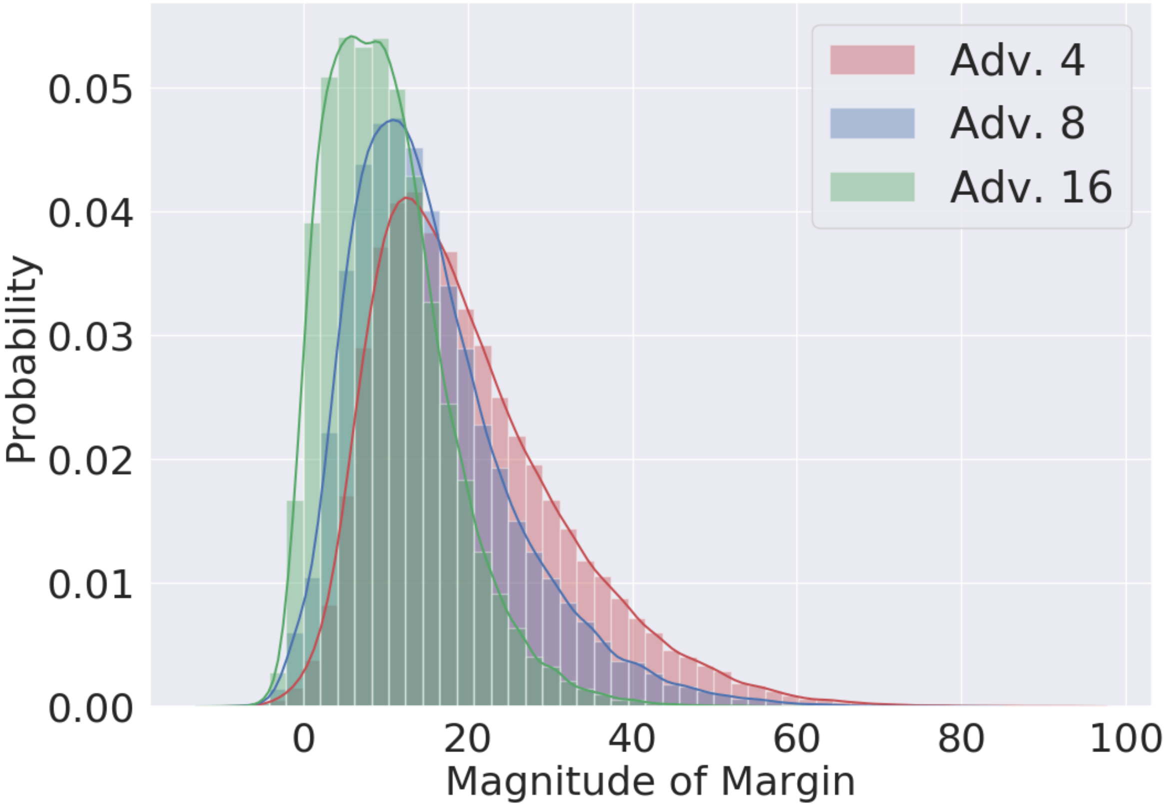
Take away message: reduced variance of the norms of the activation outputs results in concentration of examples; and the concentration results in diffident output of model.

Concentration of margin

(a) Results from CIFAR10 test set



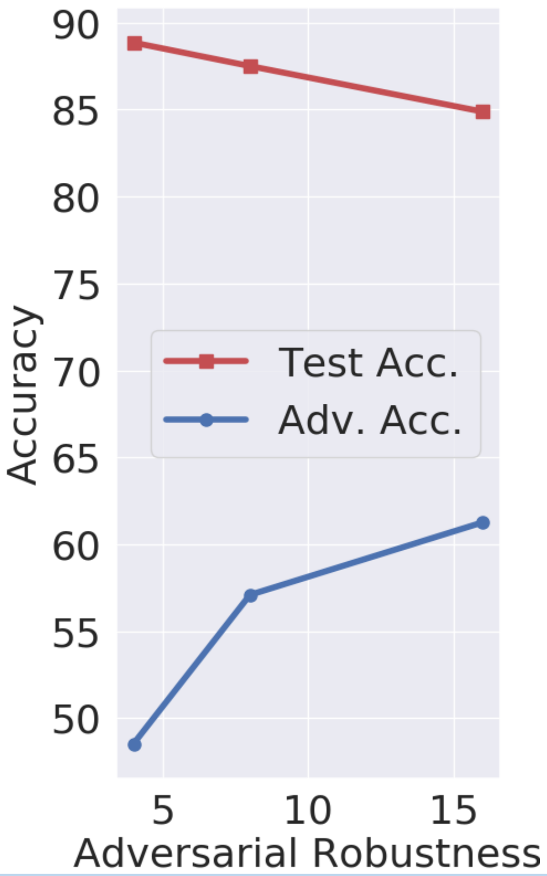
(b) Results from CIFAR10 training set



Take away message: the concentration also results in reduced loss/GE gap.

The overall effect

(a) Results from CIFAR10



Take away message: the sample concentration around decision boundaries smoothens sudden changes induced perturbations, but also increases indecisive misclassification

Take away message

Adversarial training indeed regularizes NNs, however, it does so by hurting the capacity of the NN hypothesis space.

Future works

- Study the possible hypothesis: The concentration phenomena in NNs induced by AR suggests that to reduce the effects of adversarial noise, a NN might sacrifice its ability to distinguish inter-class difference.

Thanks for your attention