Understanding Adversarial Robustness: The Trade-off between Minimum and Average Margin

Kaiwen Wu, Yaoliang Yu
{kaiwen.wu, yaoliang.yu}@uwaterloo.ca

Summary

- Deep neural networks are vulnerable to adversarial attacks.
- Standard training typically maximizes the minimum margin for high accuracy, either explicitly or implicitly.
- Meanwhile, average margin is sacrificed, hence hurting adversarial robustness.
- We propose an average margin regularizer to explicitly maximize average margin and improve adversarial robustness.

A Motivating Example

Margin definition: $\mathsf{m}(\mathbf{x}, y; \{\mathsf{F}_k\}) := \operatorname{sign}(\hat{y}(\mathbf{x}), y) \cdot \mathsf{d}(\mathbf{x}, \operatorname{bd} \mathsf{F}_y)$.

- Standard training maximizes the minimum margin **implicitly**.

Theorem 1 (Soudry et al. 2018 [1]) For almost all linearly separable binary datasets and any smooth decreasing loss with an exponential tail, **gradient descent** with small constant step size and any starting point $\mathbf{w}_0$ converges to the (unique) solution $\widehat{\mathbf{w}}$ of hard-margin SVM, *i.e.*

$$\lim_{t\to\infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} = \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|}. \qquad \blacksquare$$

- We train a binary logistic regression on MNIST, classifying 0 and 1.

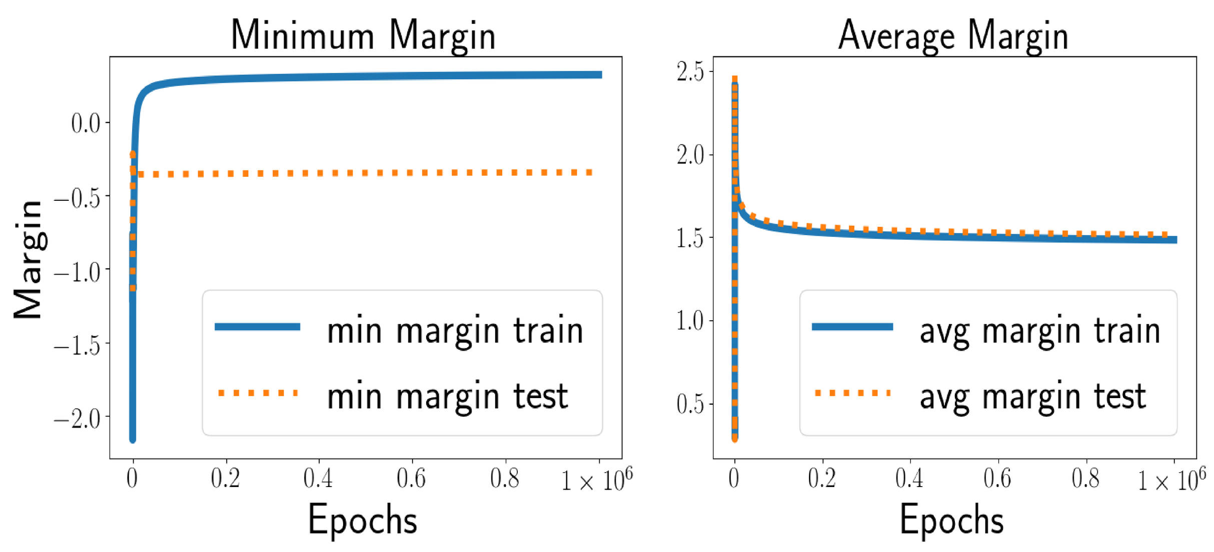


Figure: x-axis: epochs

- Indeed, minimum margin is maximized, as predicted in Theorem 1.
- Meanwhile, average margin is sacrificed (decreased).

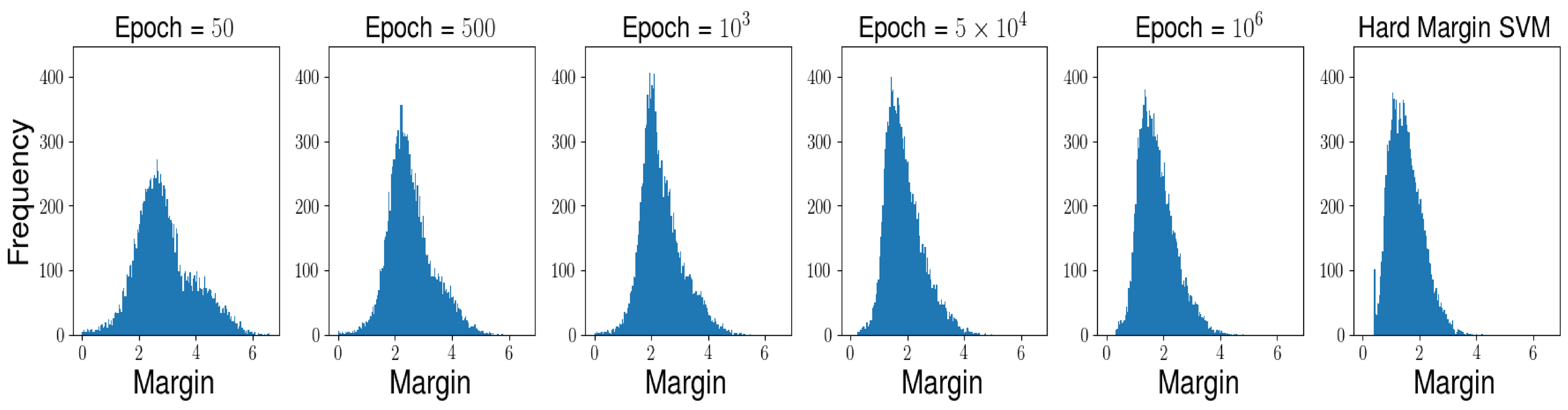


Figure: Margin histograms on the training set.

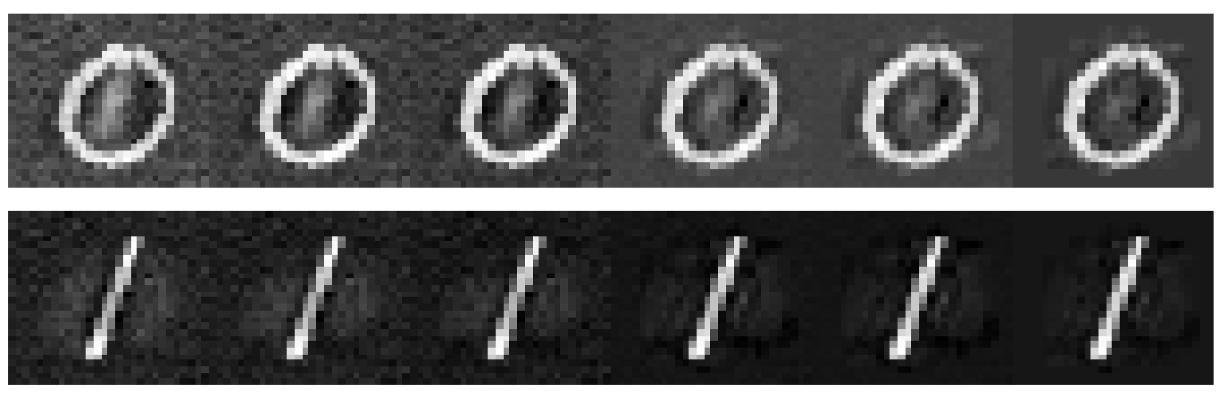


Figure: Visualization of adversarial examples at different epochs during training.

- Similar phenomenon is also observed
  (a) for deep models; (b) on different datasets.
- The trade-off cannot be an artifact of overfitting:
  (a) the training error and test error never increase;
  (b) we observe similar phenomenon on **both** training and test set.

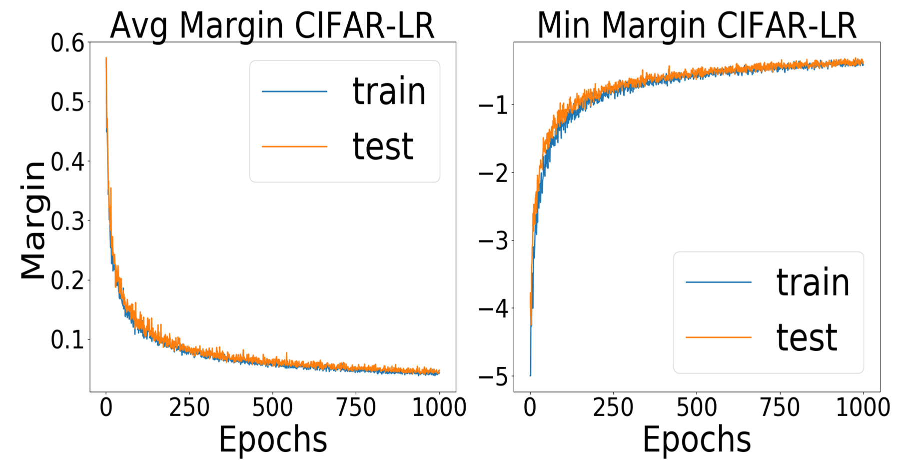
References

[1] D. Soudry, E. Hoffer, and N. Srebro. The Implicit Bias of Gradient Descent on Separable Data. In *International Conference on Learning Representations*, 2018.

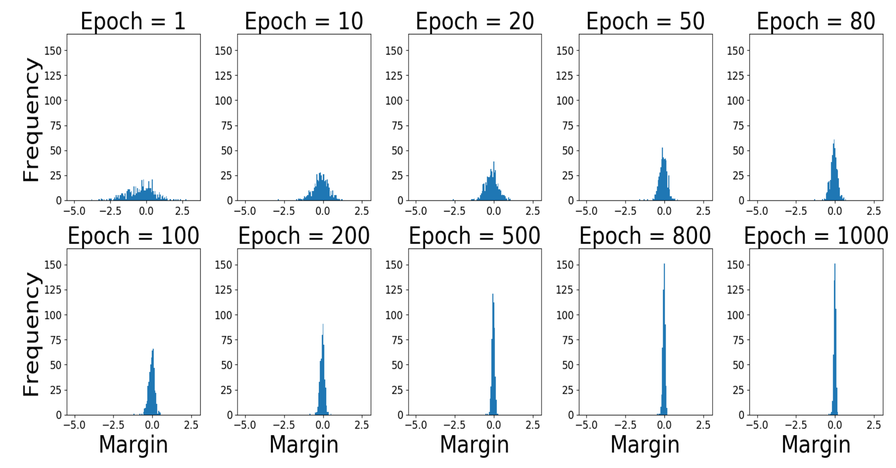
Minimum-average Margin Trade-off on Real Datasets

- For deep models, use the following approximation of margins:

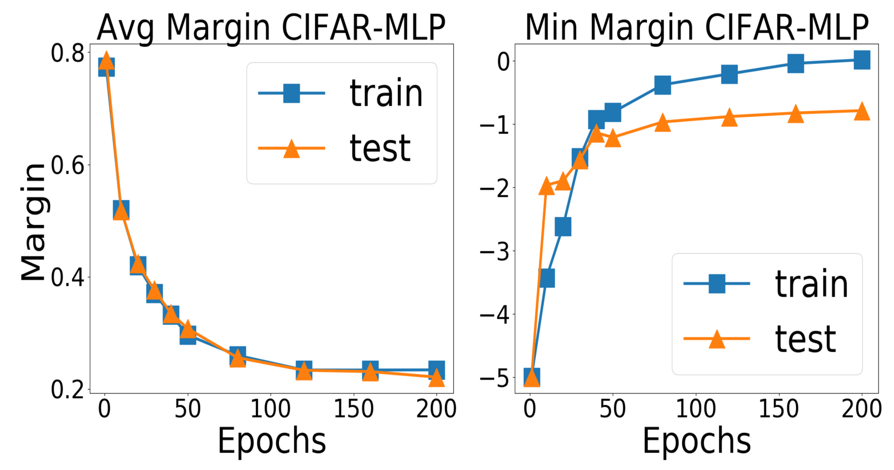
$$\left[\min_{k\neq y} \frac{f_y(\mathbf{x}) - f_k(\mathbf{x})}{L_{\mathbf{x}}^k}\right]_{-r}^{r}.$$



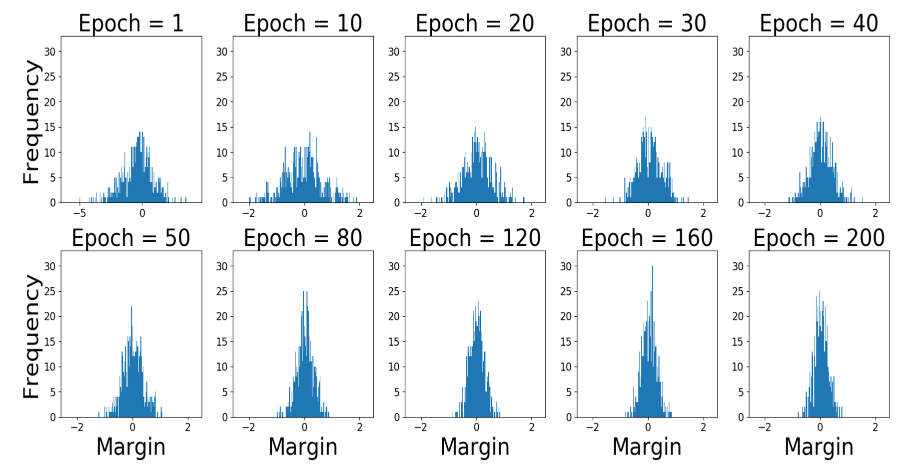
Avg vs Min CIFAR-LR



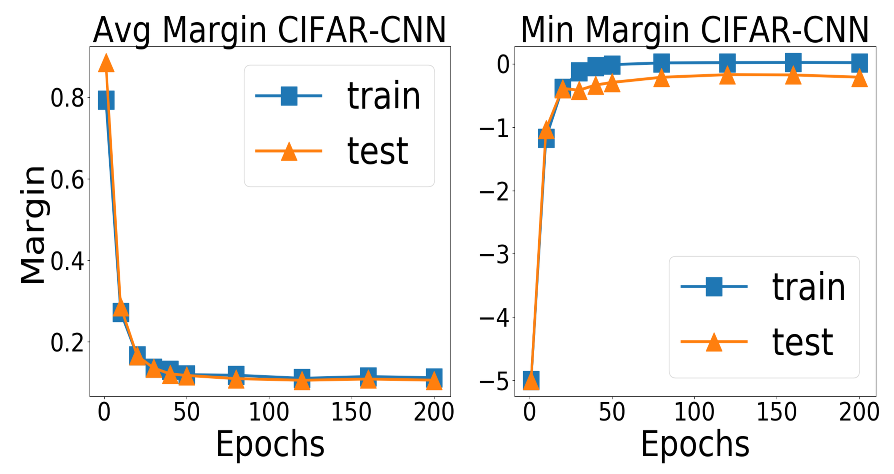
Margin Histograms on CIFAR-LR



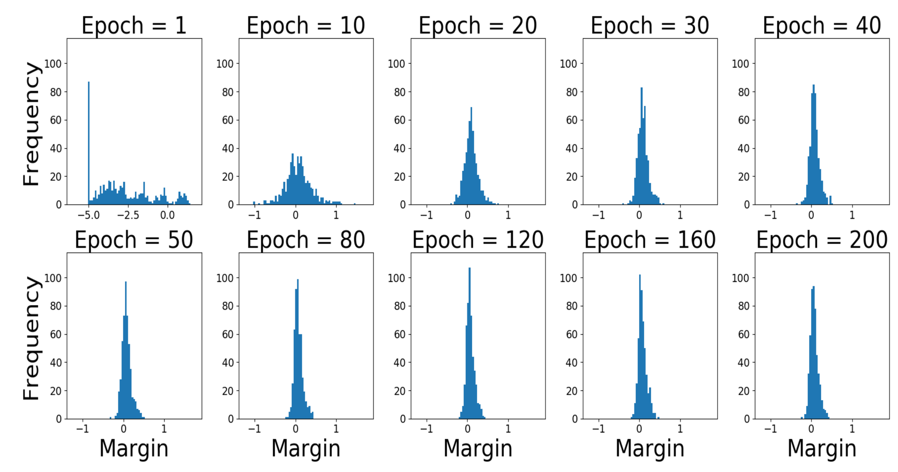
Avg vs Min CIFAR-MLP



Margin Histograms on CIFAR-MLP



Avg vs Min CIFAR-CNN



Margin Histograms on CIFAR-CNN

- The minimum margin continues increasing while at the same time the average margin keeps decreasing.
- The majority of data points is pushed closer to the boundary.

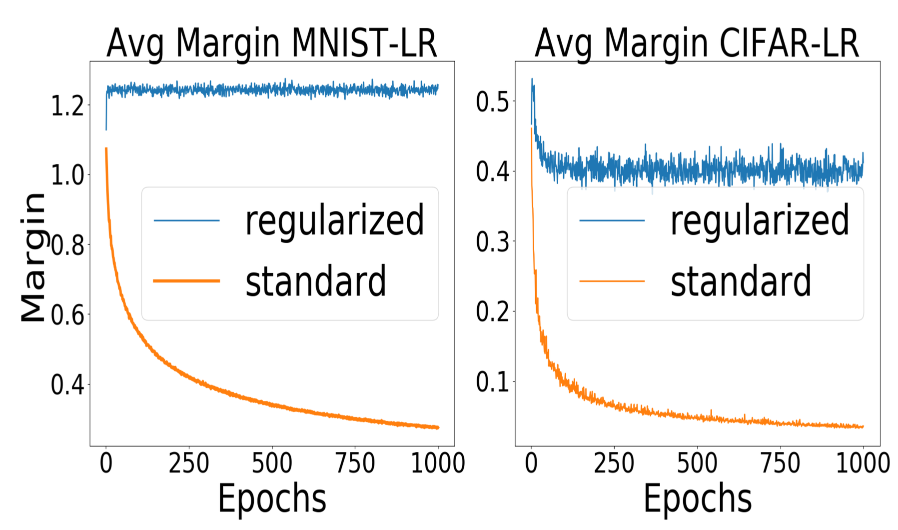
An Average Margin Regularizer

- Maximizing the input space margin for nonlinear classifiers is intractable.
- Deep network: a linear classifier after a nonlinear feature transformation Φ.
- The feature space margin provides a lower bound of the input space margin:

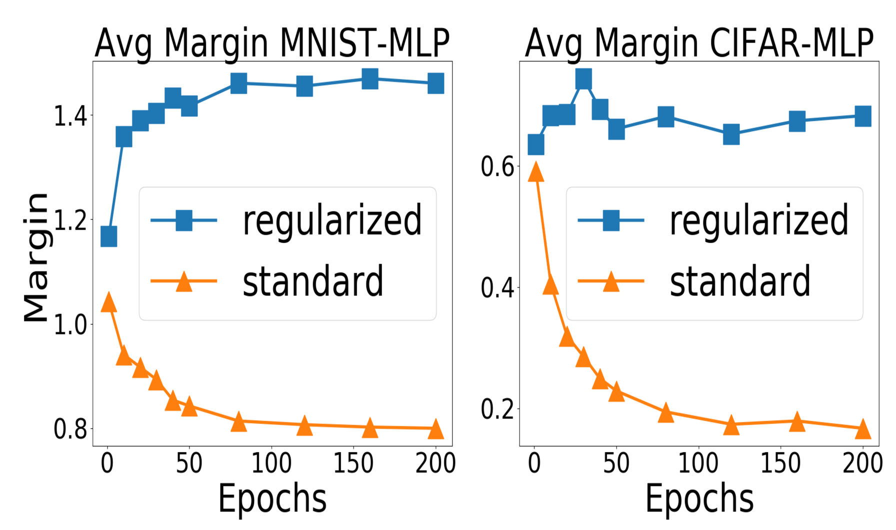
$$\|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\| \leq Lip(\Phi)\|\mathbf{x}_1 - \mathbf{x}_2\|$$

$\Rightarrow$ control the Lipschitz constant and maximize the feature space margin.

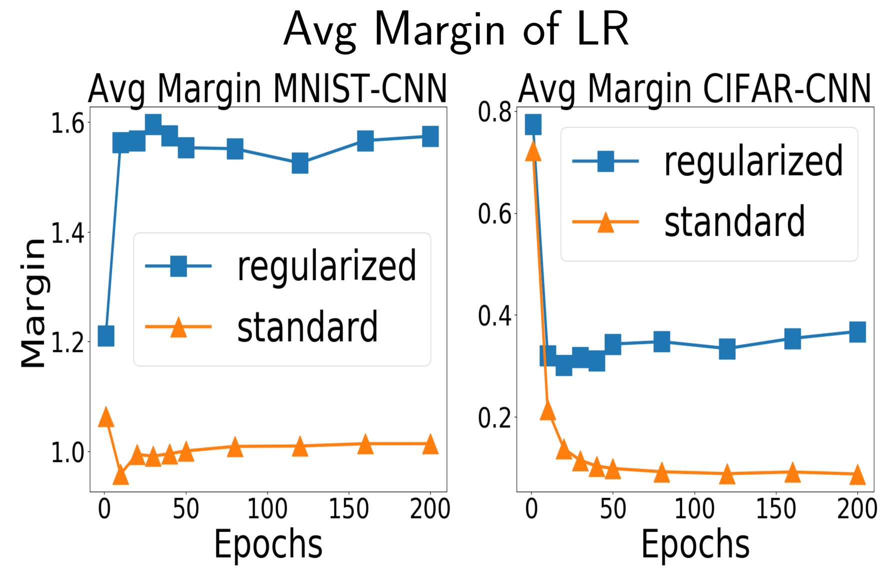
$$\sum_{i=1}^{n} \phi\left(y_i, \mathbf{f}(\mathbf{x}_i)\right) - \lambda \left[\min_{k\neq y_i} (\mathbf{w}_{y_i} - \mathbf{w}_k)^\top \Phi(\mathbf{x}_i)\right]_0^{\tau} + \beta \sum_{1\leq l\leq L} \|W_l W_l^\top - I\|_{\mathrm{F}}^2.$$



Avg Margin of LR



Avg Margin of MLP



Avg Margin CIFAR-CNN

- The regularized models no longer sacrifice the average margin.
- Significantly improves robustness.