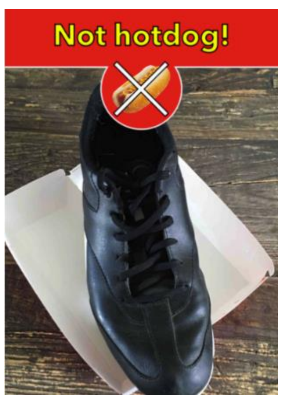
Adversarial ML: Old and New

Yaoliang Yu

Waterloo ML + Security + Verification Workshop

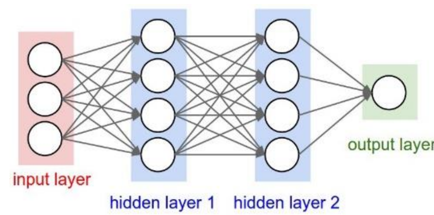
UNIVERSITY OF WATERLOO

ML in a nutshell

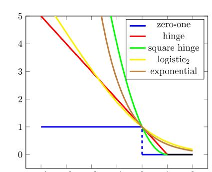


a training set of pairs of examples ($\mathbf{x}_i$, y_i)

a model for making predictions



a loss function for measuring progress



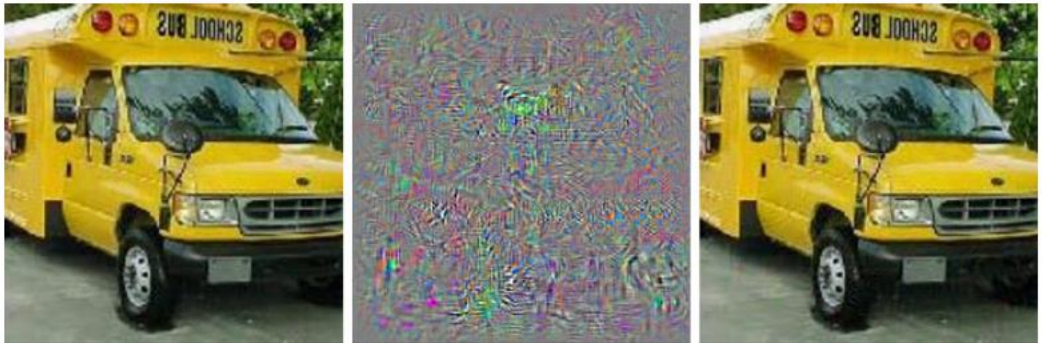
an algorithm for training

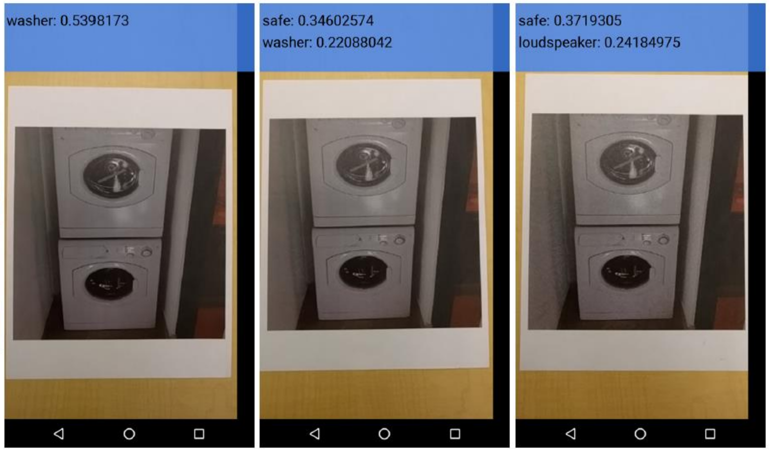
$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{\ell}(\mathbf{w})$$

$$\mathbf{w} \leftarrow \mathrm{hack}(\mathbf{w})$$

UNIVERSITY OF WATERLOO

And then the surprise…





C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus (2014). *Intriguing properties of neural networks*. ICLR.
A. Kurakin, I. Goodfellow and S. Bengio (2017). *Adversarial examples in the physical world*. ICLR workshop.
A. Athalye, L. Engstrom, A. Ilyas and K. Kwok (2018). *Synthesizing Robust Adversarial Examples*. ICML.

Why should we care?

I. Goodfellow, P. McDaniel and N. Papernot (2018). *Making machine learning robust against adversarial inputs*. CACM.

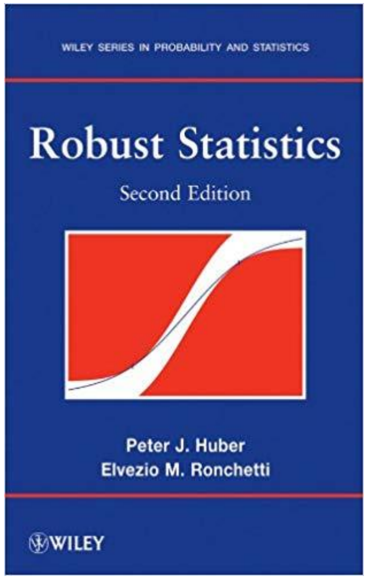
J. Gilmer, R. Adams, I. Goodfellow, D. Andersen and G. Dahl (2018). *Motivating the Rules of the Game for Adversarial Example Research*. CACM.

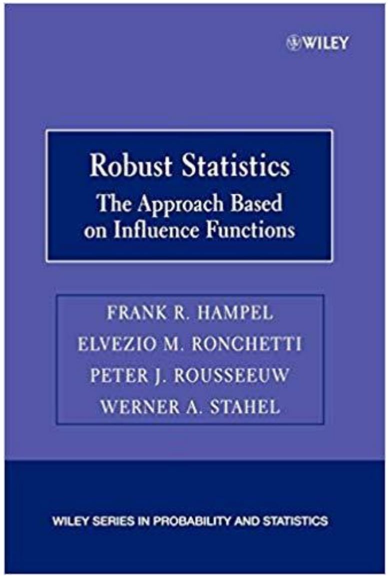
Robustness is *not* a new concern

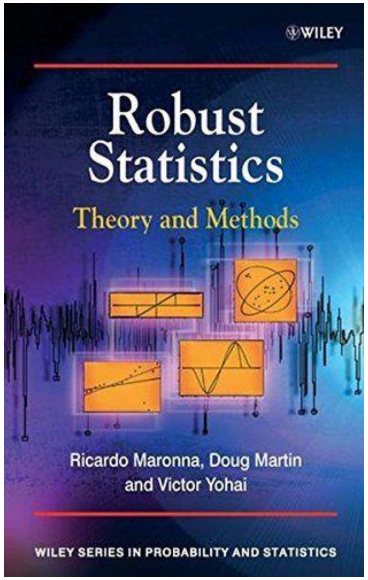
- Least-squares vs least absolute deviation

$$\|X\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{vs} \quad \|X\mathbf{w} - \mathbf{y}\|_1$$

S. PORTNOY AND R. KOENKER

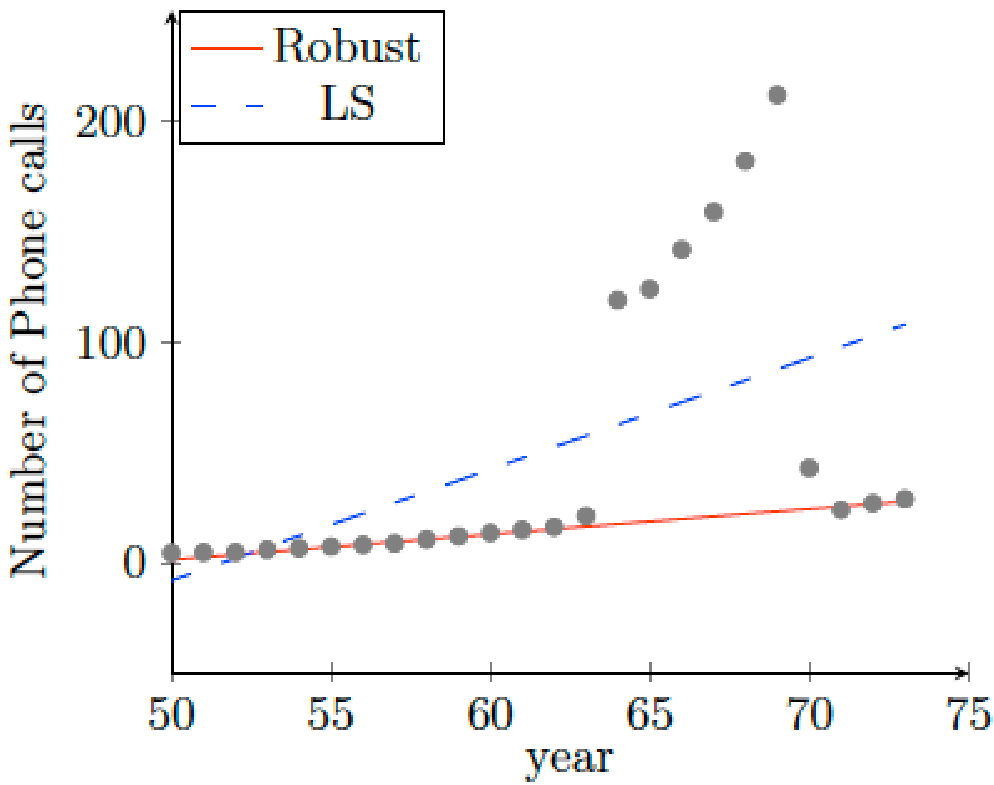






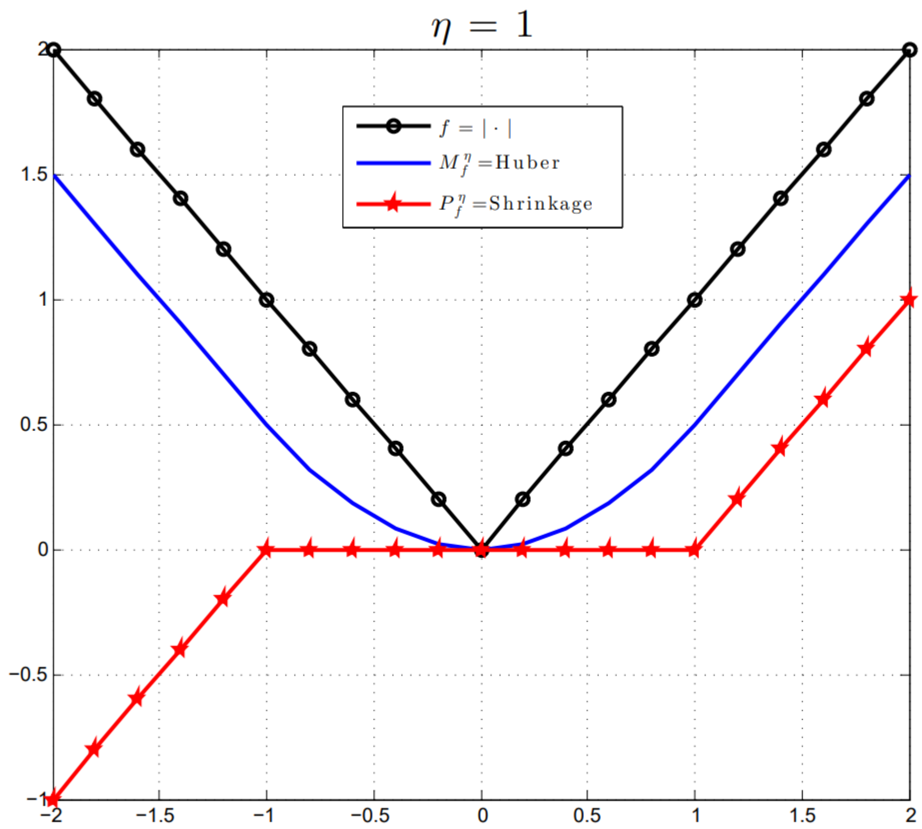
Classic Minimax Analysis

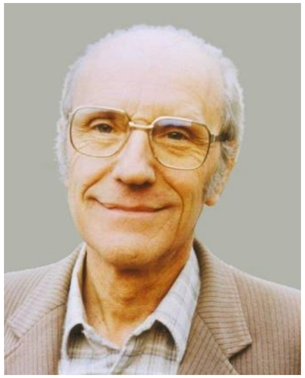
“Adversarial” example in old days



Peter J. Rousseeuw and Annick M. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley.

Huber's loss is Moreau's envelope





$$\mathsf{M}_{|\cdot|}^{\eta}(t) = \min_s \tfrac{1}{2}(s-t)^2 + \eta|s|$$
$$= \begin{cases} \frac{1}{2}t^2, & |t| \le \eta \\ \eta|t| - \frac{1}{2}\eta^2, & |t| \ge \eta \end{cases}$$

Jean J. Moreau (1962). *Fonctions convexes duales et points proximaux dans un espace hilbertien*. C.R.A.S.
Peter J. Huber (1964). *Robust estimation of a location parameter*. Annals of Statistics.

Epsilon-contamination

$$F = (1 - \epsilon)G + \epsilon H$$

$$\{X_1, \dots, X_n\} \stackrel{i.i.d.}{\sim} F$$

Threat model

- G: nominal; H: contamination; eps: size
- M-estimator: $T_n = T_n(X_1, \dots, X_n) = \operatorname*{argmin}_t \sum_{i=1}^{n} \rho(X_i - t)$
- Z-estimator: $\sum_{i=1}^{n} \psi(X_i - T_n) = 0, \quad \psi = \rho'$

Minimax optimality

$$F = (1-\epsilon)G + \epsilon H$$
$$\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} F$$

- Fix bias: $\mathbb{E}\psi(X_1) = 0, \quad i.e., \quad T_n \overset{n\to\infty}{\longrightarrow} 0$
- Derive asymptotic variance:

$$\sigma_\epsilon^2(T_n; \psi, H) \overset{n\to\infty}{\longrightarrow} \frac{\mathbb{E}\psi^2(X_1)}{\mathbb{E}^2\psi'(X_1)} =: \sigma_\epsilon^2(\psi, H)$$

- What loss to use? $\min_{\psi} \max_{H:\ \mathbb{E}\psi(X_1)=0} \sigma_\epsilon^2(\psi, H)$

- Under mild conditions, can derive psi explicitly!
  - G=Phi, eta = k(eps) → rho = Huber's loss

Peter J. Huber (1964). *Robust estimation of a location parameter*. Annals of Statistics.

Threat model digested

$$F = (1-\epsilon)G + \epsilon H$$

$$\Updownarrow$$

$$G \in \mathcal{R}_\epsilon(F) := \left\{ Q : Q \ll F, \frac{\mathrm{d}Q}{\mathrm{d}F} \leq \frac{1}{1-\epsilon} \right\}$$

$$\Downarrow$$

$$\boxed{\|F - G\|_{\mathrm{TV}} \leq \epsilon} \quad \left\{ \begin{array}{l} \mathsf{W}(F, G) \leq \epsilon \\ \\ \|\delta_x - \delta_{x'}\|_p \leq \epsilon \end{array} \right.$$

$$\Downarrow$$

$$G = (1-\epsilon)F + \epsilon H'$$

I. Cascos and M. Lopez-Diaz (2008). *Consistency of the a-trimming of a probability: Applications to central regions*. Bernoulli.

Distributional Robustness

Uncertain ERM

- Empirical Risk Minimization $$\min_{\mathbf{w}} \ \mathbb{E}\underbrace{\ell(\mathbf{w}^\top \mathbf{X})}_{\rho_{\mathbf{w}}(X)}, \quad \mathbf{X} \sim F$$
- What if *F* is contaminated? $$\min_{\mathbf{w}} \max_{G \in \mathcal{F}_\epsilon} \ \mathbb{E}\ell(\mathbf{w}^\top \mathbf{X}), \quad \mathbf{X} \sim G$$
- Uncertainty set $\mathcal{F}_\epsilon$
- Lots of work on constructing *tractable* uncertainty sets

Cooper, Dantzig, Dupacova, Prékopa, Bertsemas, Ben-Tal, Calafiore, Campi, Delage, El Ghaoui, Mannor, Nemirovski, Ruszczynski, Shapiro, Sim, Xu, Zhang…

UNIVERSITY OF WATERLOO

A simple uncertainty set

$$\boxed{\mathcal{F} = \{G : \mathbf{m}(F) = \mathbf{m}(G), \Sigma(F) = \Sigma(G)\}}$$

$$\min_{\mathbf{w}} \max_{G \in (\mathbf{m}, \Sigma)} \mathbb{E}\ell(\mathbf{w}^\top \mathbf{X}), \quad \mathbf{X} \sim G$$

$$\Downarrow$$

Theorem The linear projection $(\mu, \Sigma) \overset{A}{\mapsto} (A^\top \mu, A\Sigma A^\top)$ is ONTO.

$$\Downarrow$$

$$\min_{\mathbf{w}} \underbrace{\max_{H \in (\mathbf{w}^\top \mathbf{m}, \mathbf{w}^\top \Sigma \mathbf{w})} \mathbb{E}\ell(Z)}_{r(\mathbf{w})}, \quad Z \sim H$$

I. Popescu (2007). *Robust mean-covariance solutions for stochastic optimization*. Operations Research.
Y. Yu, Y. Li, D. Schuurmans and C. Szepesvari (2008). *A general projection property of distribution families*. NeurIPS.
L. Chen, S. He and S. Zhang (2011). *Tight Bounds for Some Risk Measures, with Applications to Robust Portfolio Selection*. Operations Research.

New uncertainty sets

$$\boxed{\mathcal{F}_\epsilon = \{G : \mathsf{W}(F, G) \leq \epsilon\}}$$

$$\min_{\mathbf{w}} \underbrace{\max_{G \in \mathcal{F}_\epsilon} \ \mathbb{E}\ell(\mathbf{w}^\top \mathbf{X}),}_{} \quad \mathbf{X} \sim G$$

$$r(\mathbf{w}) \qquad \boxed{\min_{\gamma \geq 0} r\epsilon^p - \mathbb{E}\mathsf{M}^{\gamma}_{-\rho}(\mathbf{X}; \mathbf{w}), \quad \rho(\mathbf{x}) = \ell(\mathbf{w}^\top \mathbf{x}), \ \mathbf{X} \sim F}$$

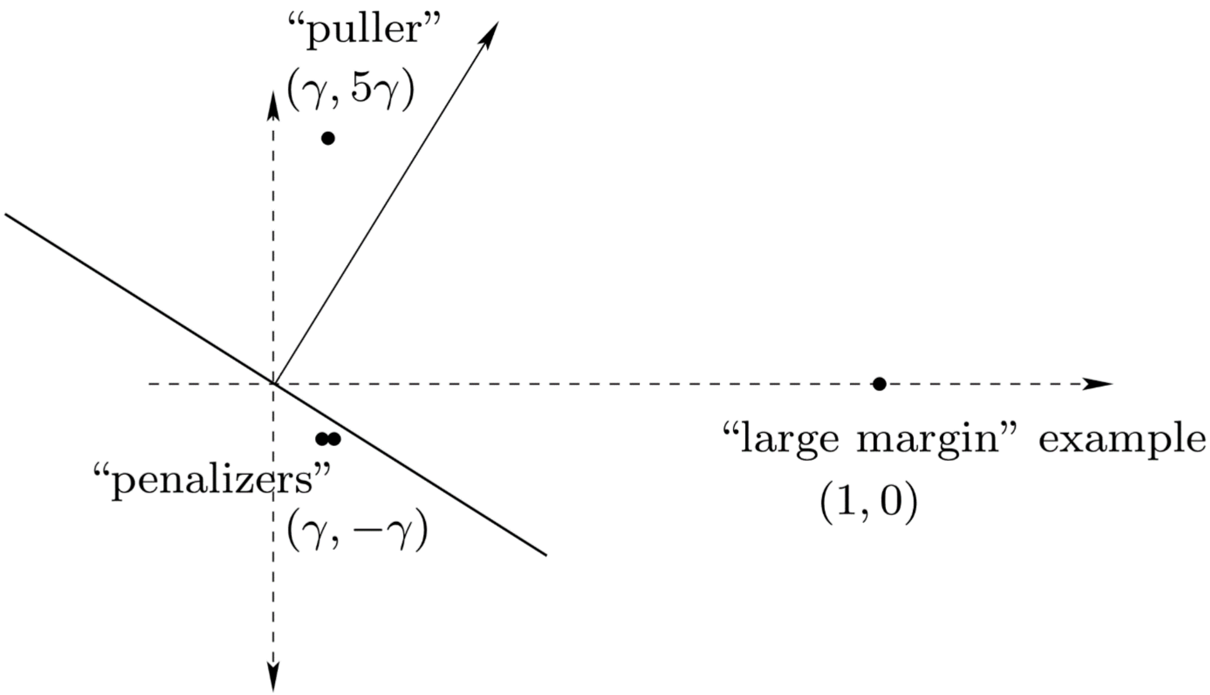
$$\leq$$

$$\mathbb{E}\ell(\mathbf{w}^\top \mathbf{X}) + \epsilon \mathsf{Lip}(\ell_{\mathbf{w}}), \quad \mathbf{X} \sim F$$

P. Esfahani and D. Kuhn (2018). *Data-driven distributionally robust optimization using the Wasserstein metric*. Mathematical Programming.
A. Sinha, H. Namkoong, J. Duchi (2018). *Certifying some distributional robustness with principled adversarial training*. ICLR.
Z. Cranko, S. Kornblith, Z. Shi and R. Nock (2018). *Lipschitz Networks and Distributional Robustness*. arXiv:1809.01129.

Robust Loss for Adversarial Training

Boosting



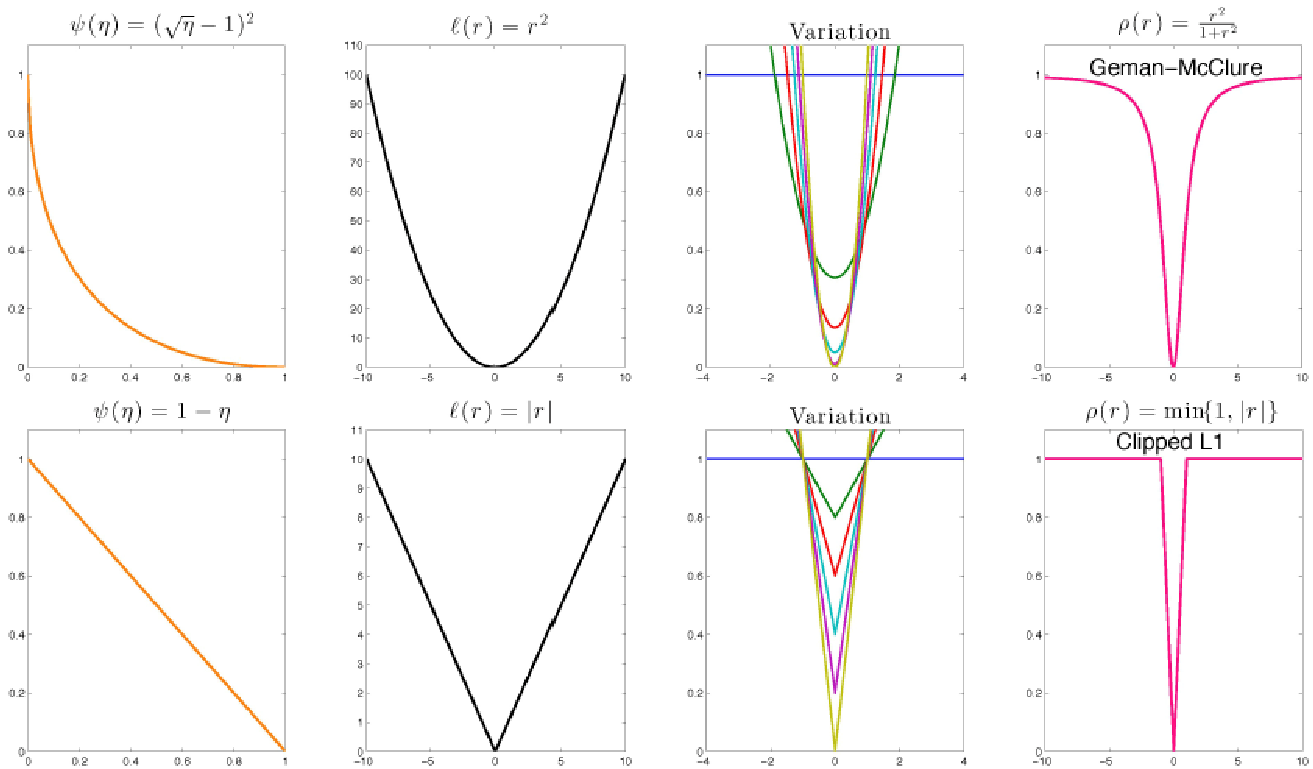
Theorem

All convex potential function based boosters can not tolerate random classification noise at rate $\eta \in (0, 1/2)$.

P. Long and R. Servedio (2010). *Random classification noise defeats all convex potential boosters*. Machine Learning.

Variational Loss

$$r(t) = \min_{0 \leq \eta \leq 1} \eta \ell(t) + \psi(\eta)$$



M. Black and A. Rangarajan (1996). *On the unification of line processes, outlier rejection, and robust statistics with applications in early vision*. IJCV.

Y. Yu, O. Aslan and D. Schuurmans (2012). *A Polynomial-time Form of Robust Regression*. NeurIPS.

Dilemma

	bounded	convex	output 0		yak!
Properties			true or false		
M-estimator	1	1	1	0	1
Consistency	1	1	0	1	1
Robustness	1	0	1	1	1
Tractability	0	1	1	1	1
Achievable?	✓	✓	✓	?	✗

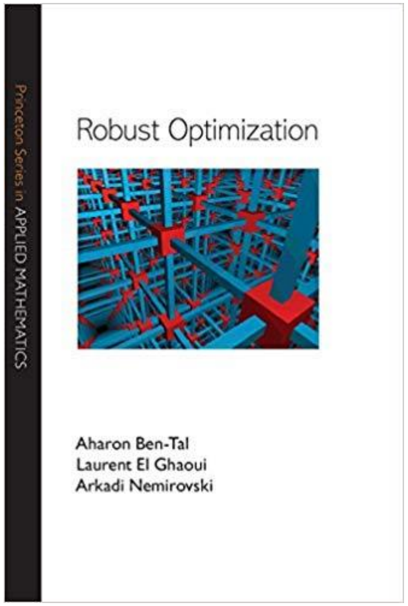
Robustness: $\epsilon \to 0, \quad H = \delta_{\mathbf{x}} \to \infty$

Y. Yu, O. Aslan and D. Schuurmans (2012). *A Polynomial-time Form of Robust Regression*. NeurIPS.

UNIVERSITY OF WATERLOO

Adversarial training

$$\min_{\mathbf{w}} \quad \mathbf{E}\left[\underbrace{\max_{\|\Delta\mathbf{x}\| \le \epsilon} \ell(\mathbf{x} + \Delta\mathbf{x}; \mathbf{w})}_{\bar{\ell}(\mathbf{x};\mathbf{w})} \right]$$



- One of the best defensive mechanisms
  - though inner maximization may be hard

- But, it basically amounts to changing the loss in a different way

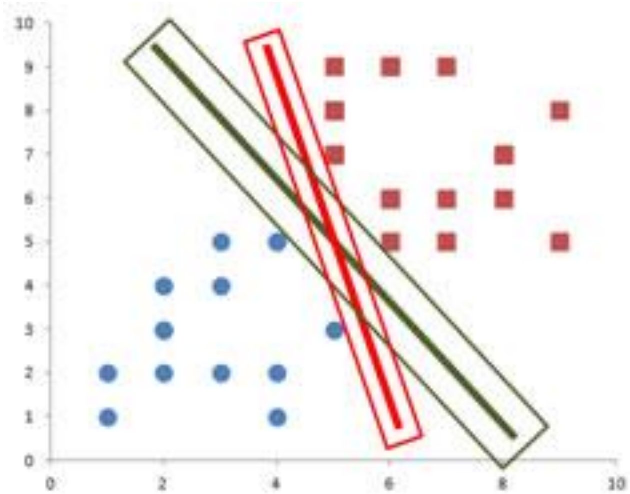
A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu (2018). *Towards deep learning models resistant to adversarial attacks*. ICLR.

Average Margin vs Minimal Margin

Margin

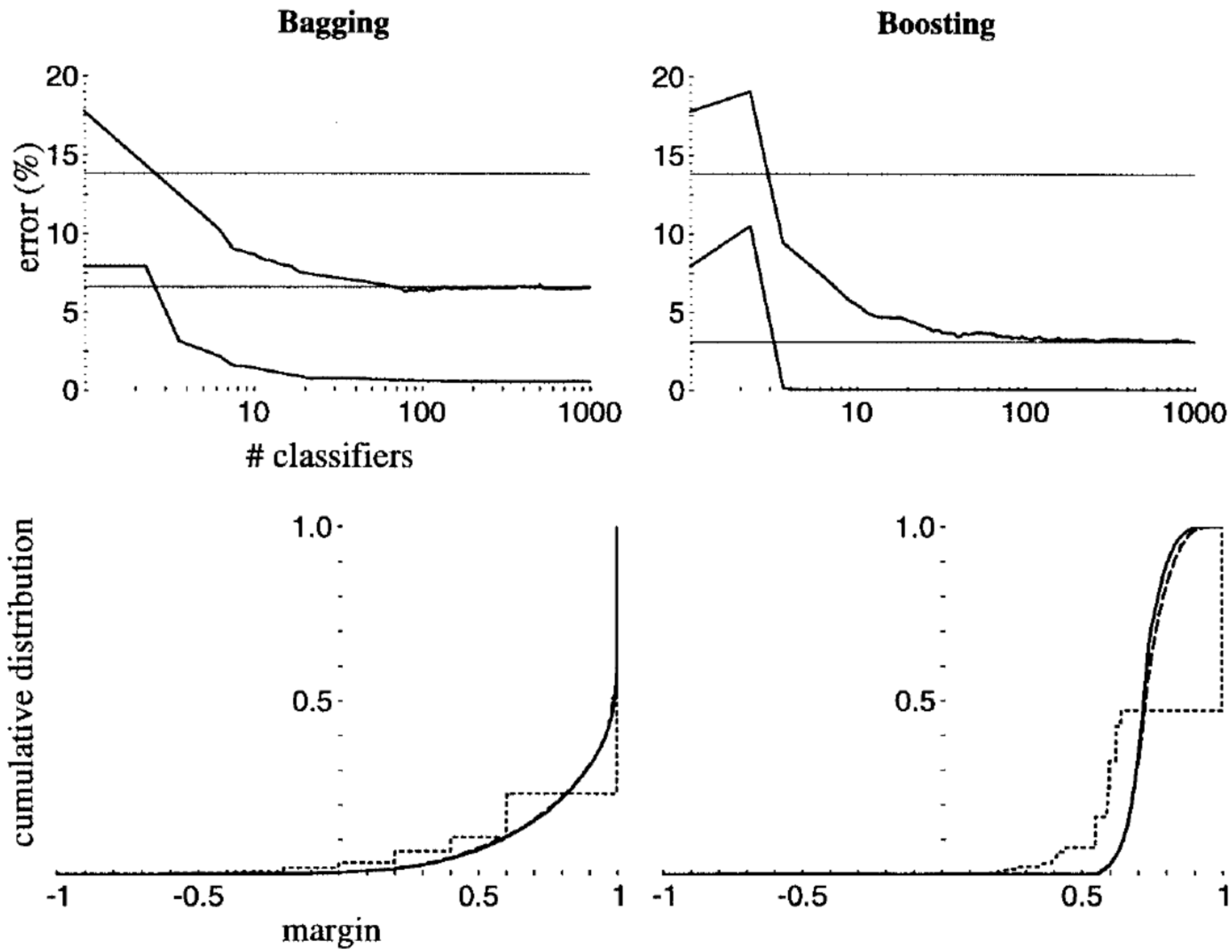
$$\mathsf{m}(\mathbf{x}, y; \{\mathsf{F}_k\}) := \mathrm{sign}(\hat{y}(\mathbf{x}), y) \cdot \mathsf{d}(\mathbf{x}, \mathrm{bd}\, \mathsf{F}_{\hat{y}})$$

- Minimum margin vs Average margin
- Well-known that SVM maximizes minimum margin
- Abundant work relating margin with generalization



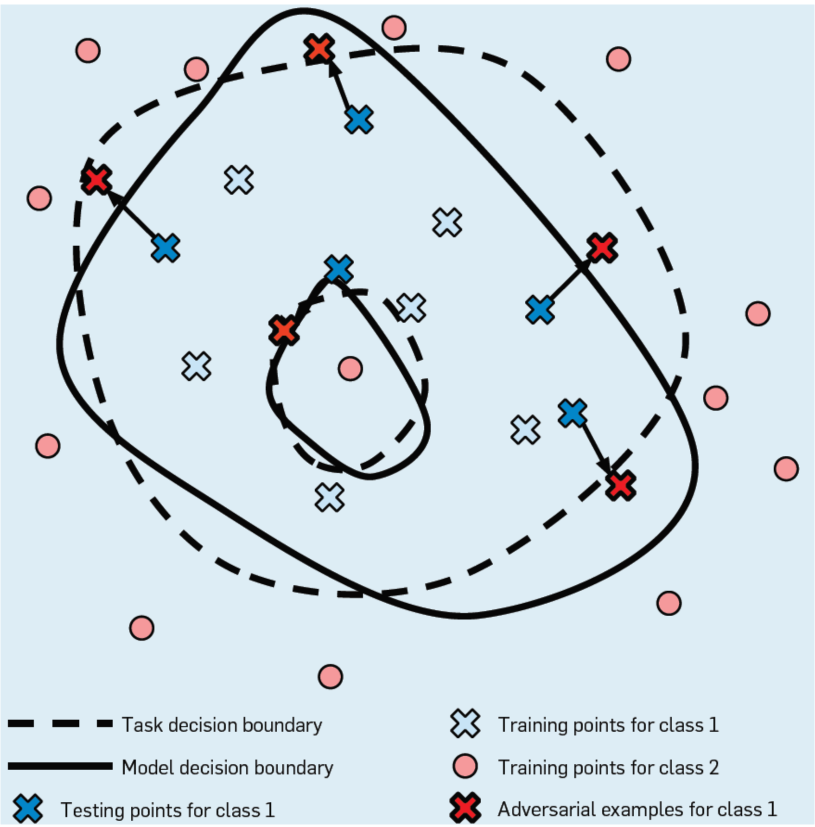
UNIVERSITY OF WATERLOO

Why Boosting does not overfit?



R. Schapire, Y. Fruend, P. Bartlett and W. Lee (1998). *Boosting the margin- a new explanation for the effectiveness of voting methods*. Annals of Statistics.

The new challenge?



Deep Learning

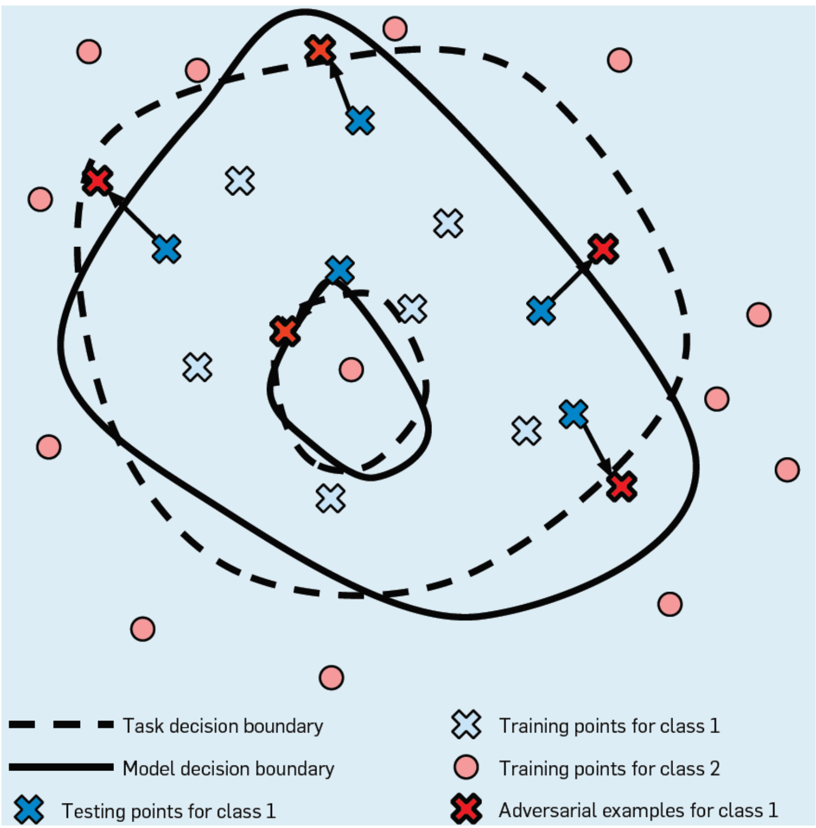
$$\mathbf{x} \xrightarrow{\varphi} \varphi(\mathbf{x}) \xrightarrow{\mathbf{w}} \mathbf{w}^\top \varphi(\mathbf{x}) =: \hat{y}$$

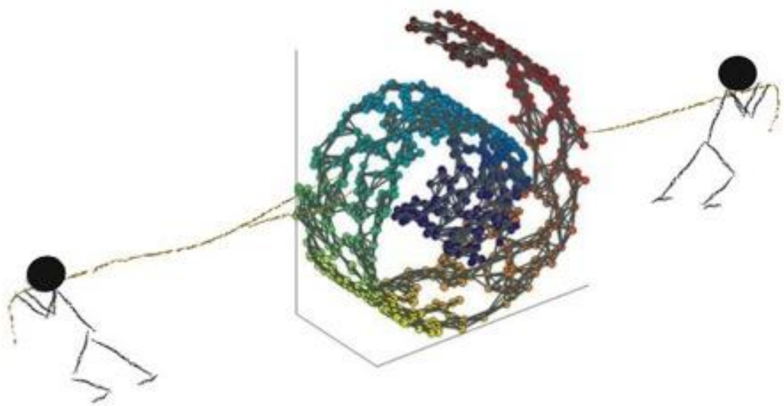
$$\min_{\mathbf{w},\varphi} \quad \ell(y, \hat{y})$$

- nonlinear transformation φ
- linear classifier w
- trained jointly by SGD

I. Goodfellow, P. McDaniel and N. Papernot (2018). *Making machine learning robust against adversarial inputs*. CACM.

A possible explanation





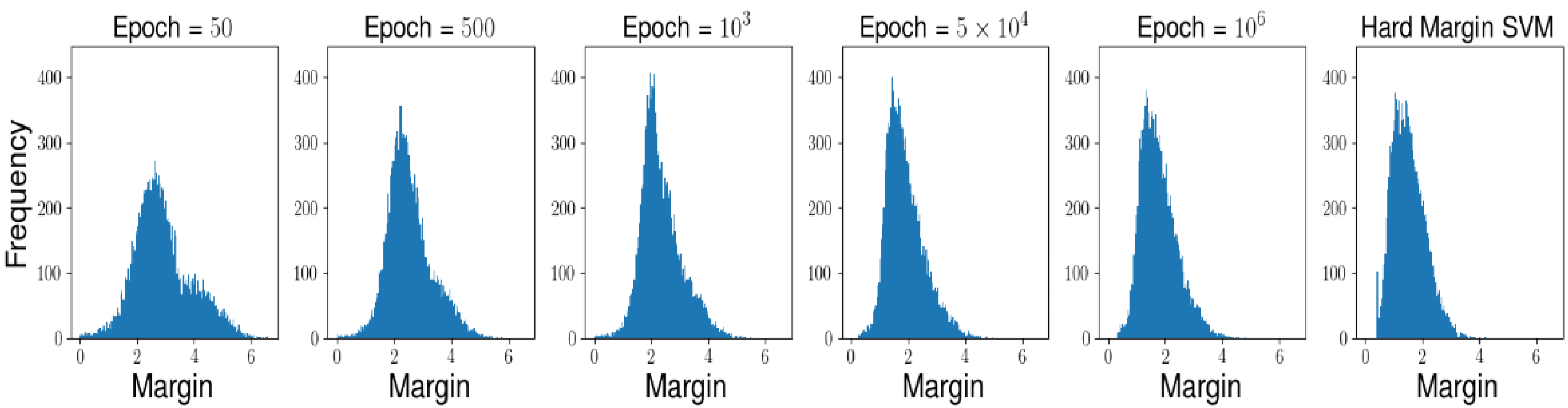
Maximum variance unfolding

I. Goodfellow, P. McDaniel and N. Papernot (2018). *Making machine learning robust against adversarial inputs*. CACM.
K. Weinberger and L. Saul (2006). *Unsupervised Learning of Image Manifolds by Semidefinite Programming*. IJCV.

Margin maximization

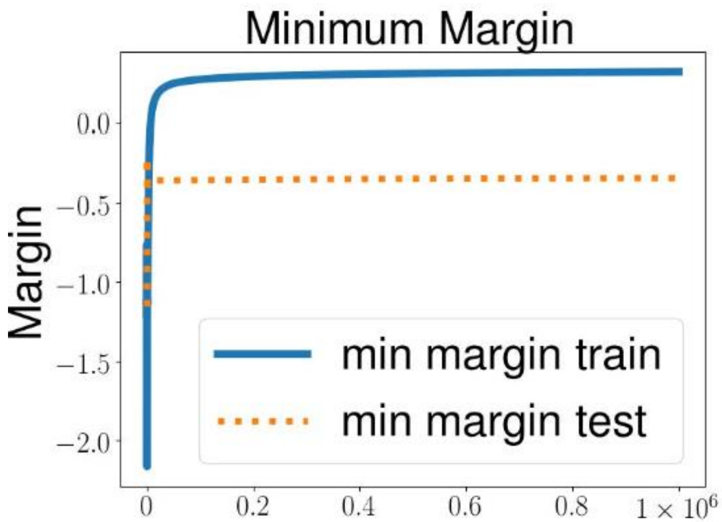
Theorem 1 (Soudry et al. 2018 [3]) For almost all linearly separable binary datasets and any smooth decreasing loss with an exponential tail, **gradient descent** with small constant step size and any starting point $\mathbf{w}_0$ converges to the (unique) solution $\widehat{\mathbf{w}}$ of hard-margin SVM, *i.e.*

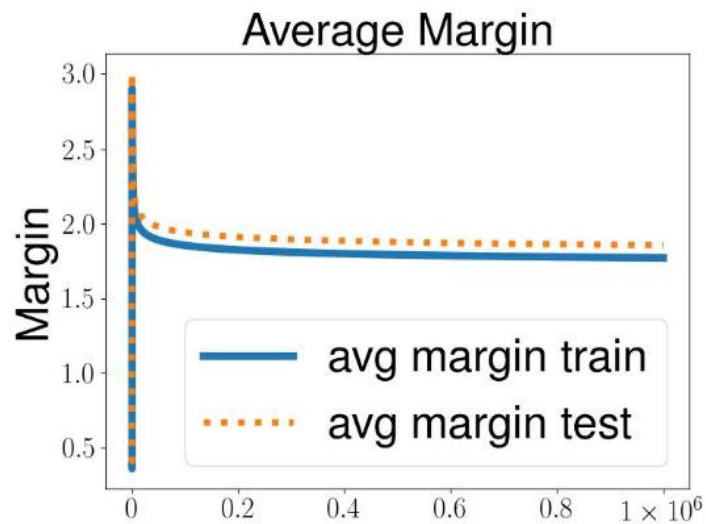
$$\lim_{t\to\infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} = \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|}. \qquad \blacksquare$$

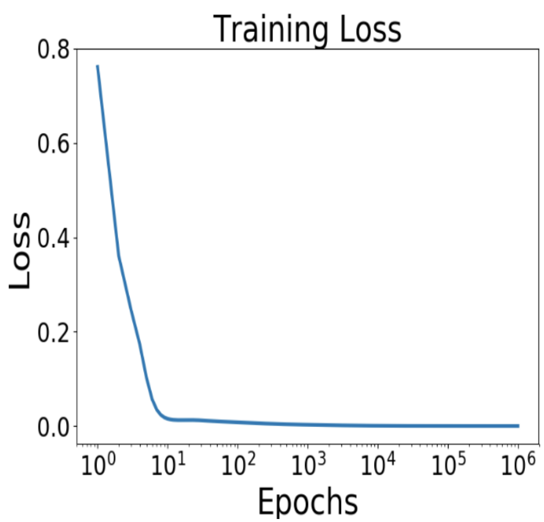


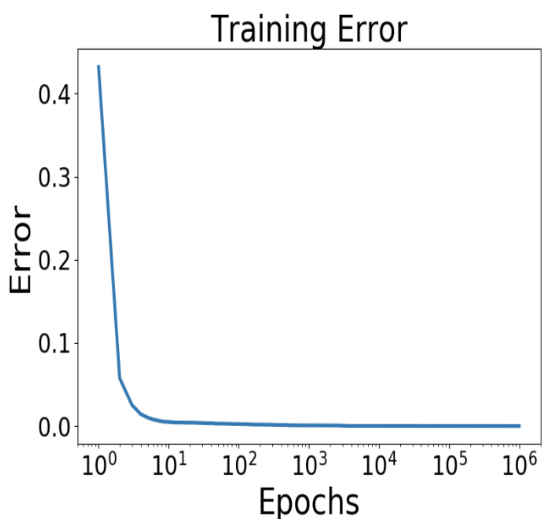
D. Soudry, E. Hoffer and N. Srebro (2018). *The Implicit Bias of Gradient Descent on Separable Data*. ICLR.

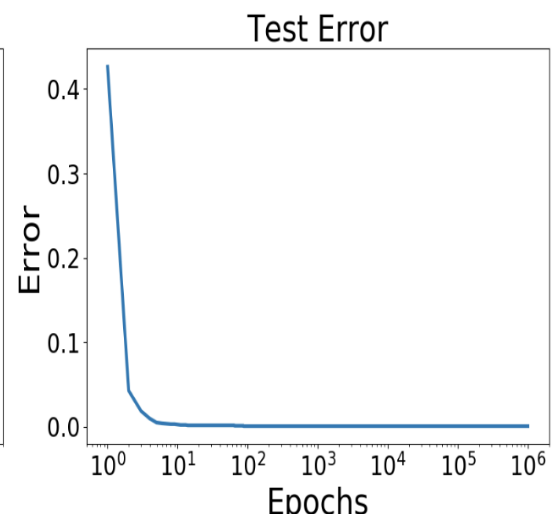
Min vs Avg


Minimum Margin
Margin
0.0
−0.5
−1.0
−1.5
−2.0
0
0.2
0.4
0.6
0.8
1×10^6
min margin train
min margin test


Average Margin
Margin
3.0
2.5
2.0
1.5
1.0
0.5
0
0.2
0.4
0.6
0.8
1×10^6
avg margin train
avg margin test


Training Loss
Loss
Epochs

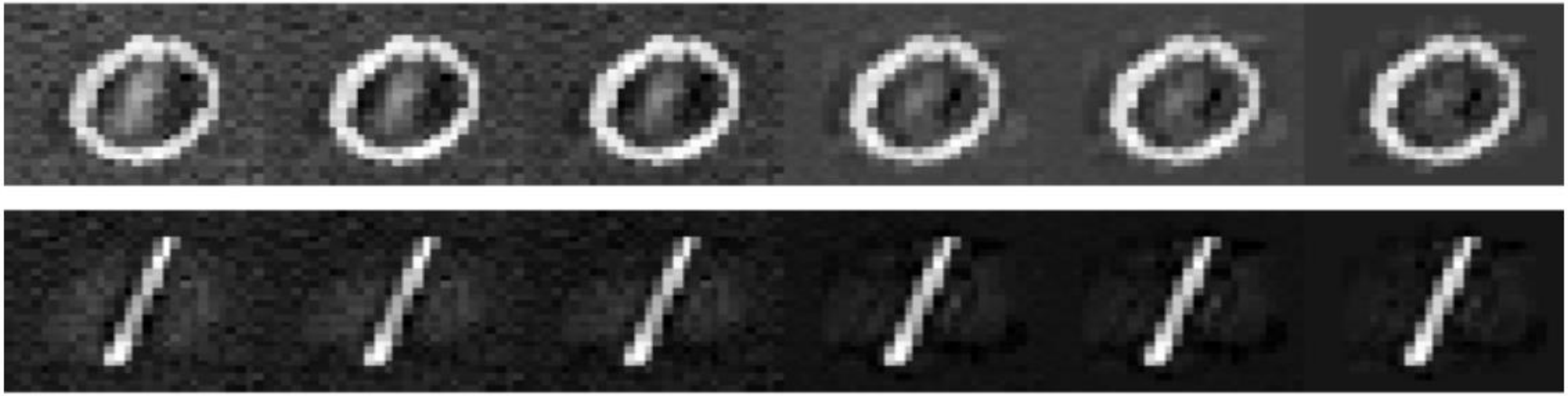

Training Error
Error
Epochs


Test Error
Error
Epochs

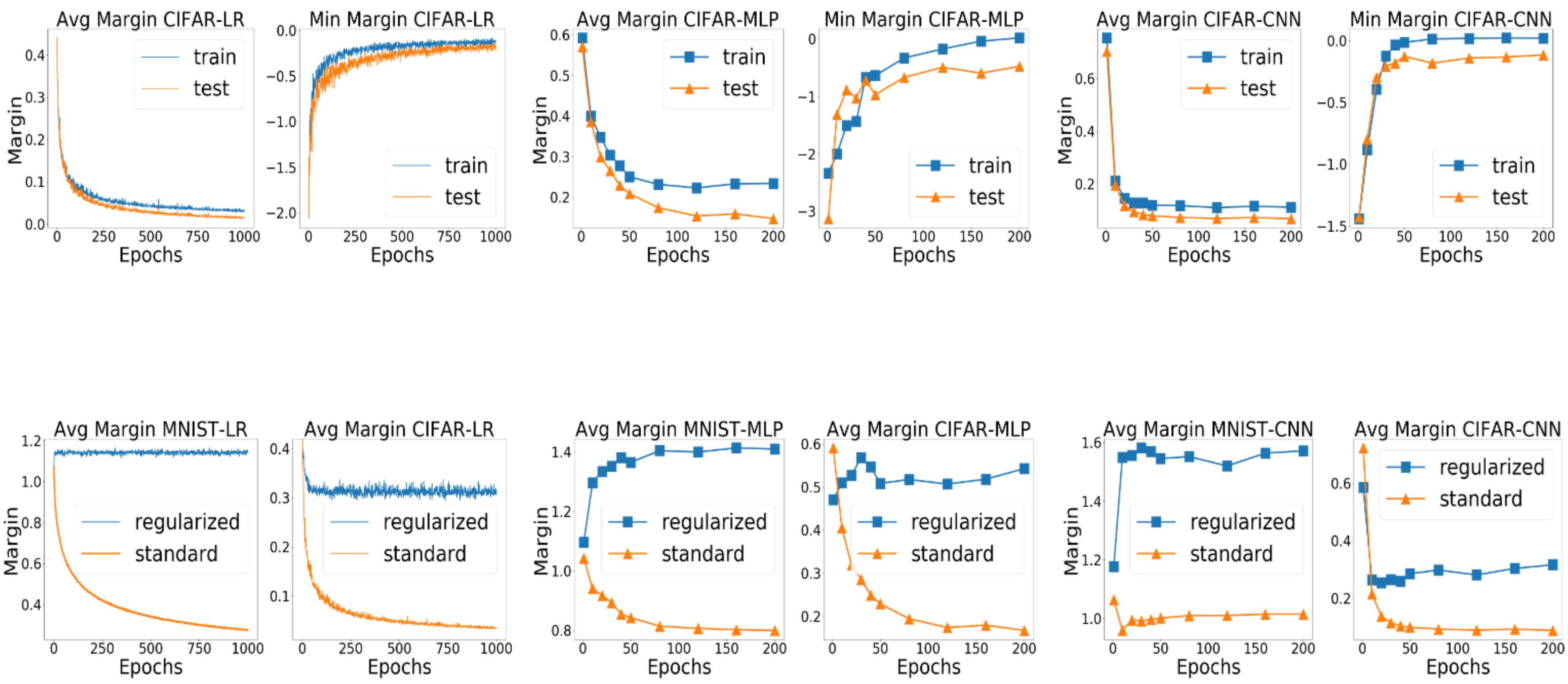
Adversarial Examples

- For linear classifiers, have closed form:

$$\mathbf{x}^{adv} = \mathbf{x} - \frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{w}\|^2}\mathbf{w}$$

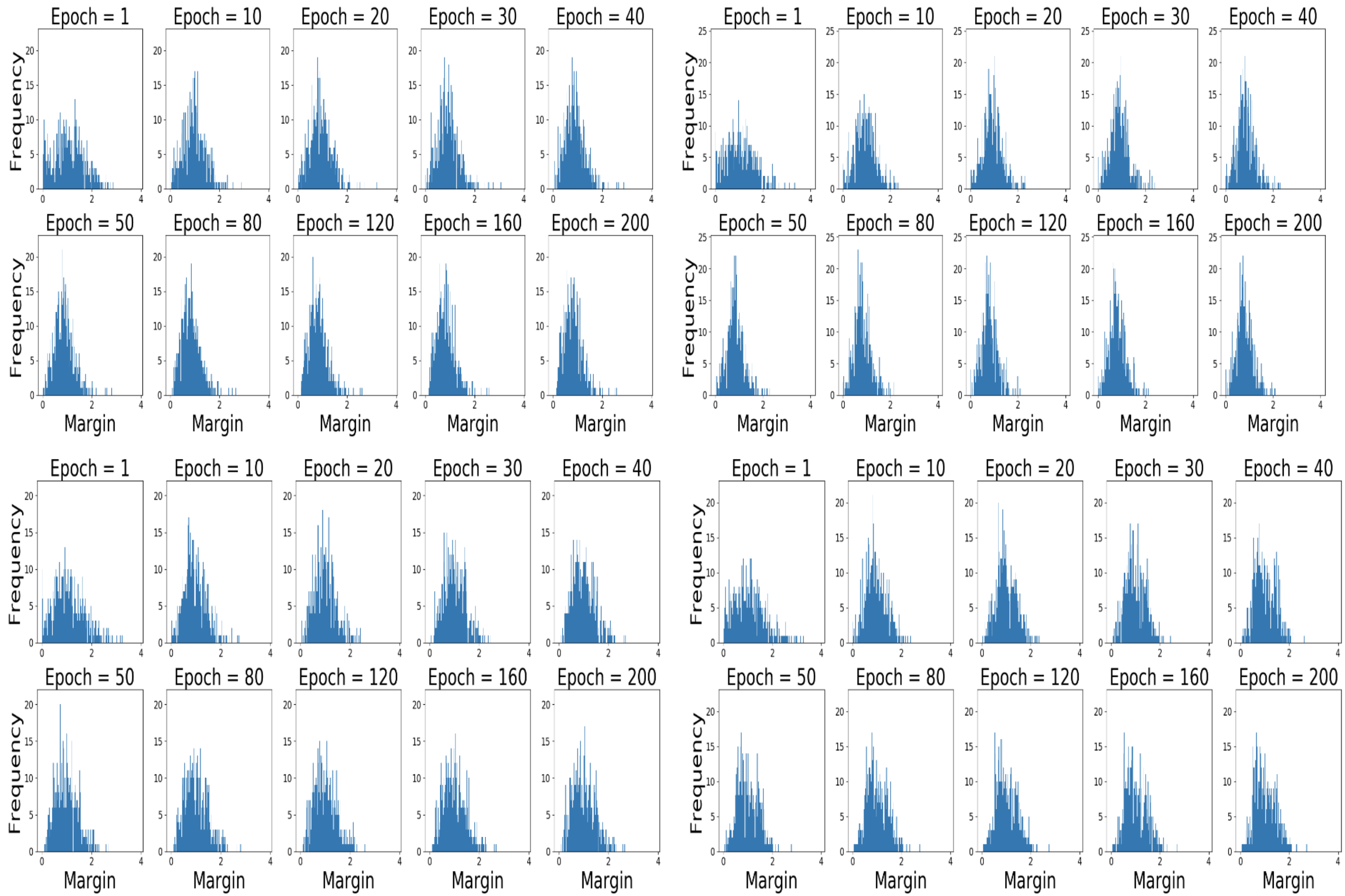


Same for deep models

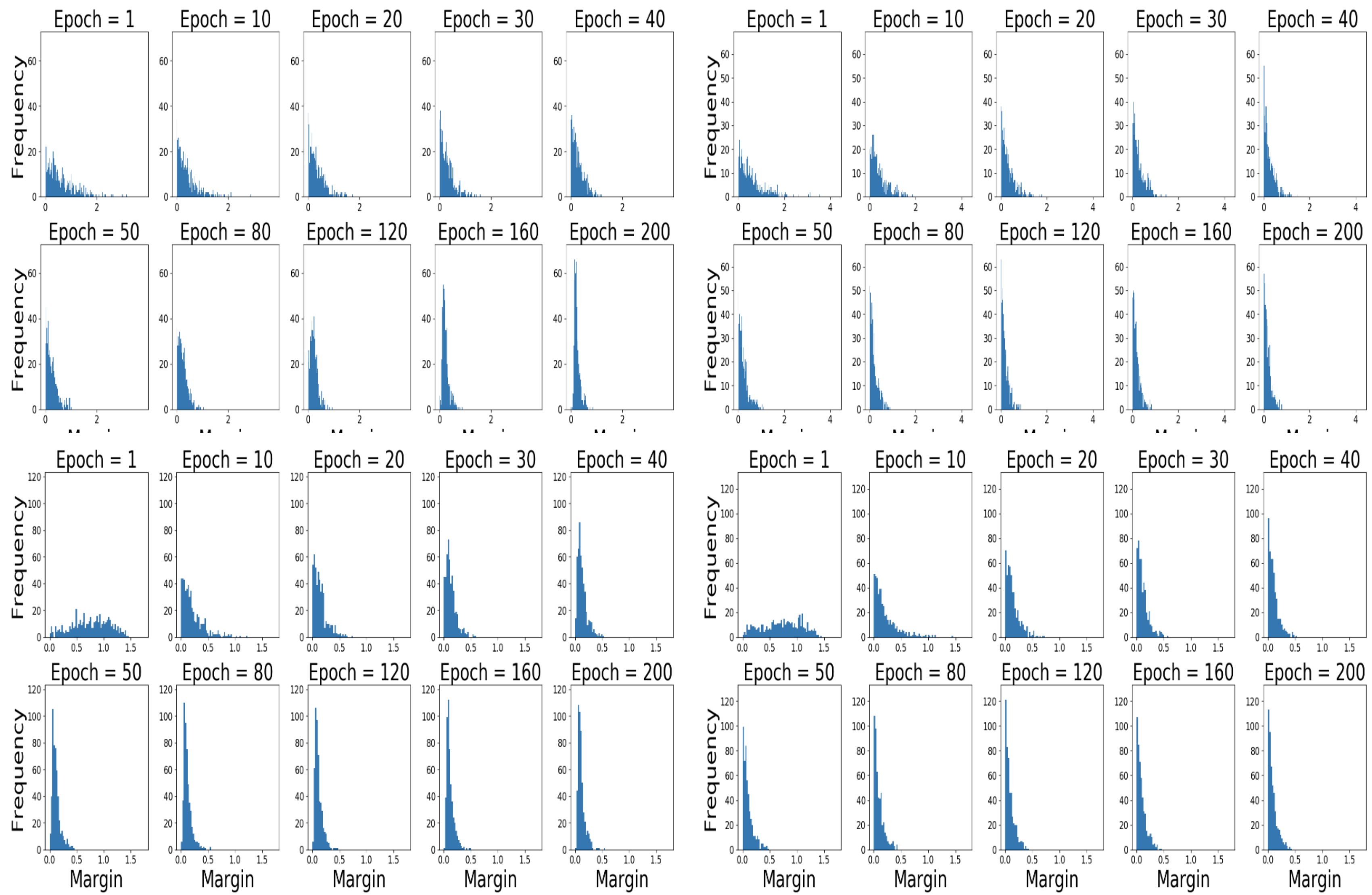


UNIVERSITY OF WATERLOO

Margin Histogram on MNIST



Margin Histogram on CFIAR10



Margin regularizer

$$\min_{\mathbf{f}:\mathcal{X}\to\mathbb{R}^c} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi(y_i,\mathbf{f}(\mathbf{x}_i))}_{\text{classifier}} - \lambda\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{y_i=\hat{y}_i}\cdot \mathsf{d}_\tau(\mathbf{x}_i,\mathrm{bd}\,\mathsf{F}_{y_i})}_{\text{regularizer}}$$

- For binary linear classifiers, can compute margin explicitly:

$$\min_{\|\mathbf{w}\|=1} \sum_{i=1}^{n}\phi(y_i\mathbf{w}^\top\mathbf{x}_i) - \lambda\cdot\sum_{i=1}^{n}[y_i\mathbf{w}^\top\mathbf{x}_i]_0^\tau$$

K. Wu and Y. Yu (2019). *Understanding Adversarial Robustness: The Trade-off between Minimum and Average*. arXiv:1907.11780.

Multi-class Extension

$$\mathbf{f}(\mathbf{x}) = \underbrace{W_L}_{\text{linear classifier}} \cdot \underbrace{\sigma(W_{L-1} \cdot \sigma(\cdots \sigma(W_1 \cdot \mathbf{x})))}_{\text{feature transform Phi}}$$

$$\underbrace{\|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|}_{\text{dist in feature space}} \leq \underbrace{Lip(\Phi)}_{\text{tightness}} \underbrace{\|\mathbf{x}_1 - \mathbf{x}_2\|}_{\text{dist in input space}}$$

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

Experiments

	Models	Method	Clean	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$	$\epsilon = 2.0$	Avg Margin
MNIST	MLP	Std	**98.24**	89.26	49.23	15.78	5.06	0.80
		LCR	95.99	91.12	80.62	60.56	34.07	1.36
		AMR	96.01	**91.18**	**81.01**	**62.93**	**38.44**	**1.41**
	CNN	Std	99.14	95.95	90.48	88.72	87.52	1.01
		LCR	**99.29**	97.21	87.83	58.30	26.96	1.34
		AMR	99.25	**97.90**	**97.60**	**97.51**	**97.33**	**1.40**
	Models	Method	Clean	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	Avg Margin
CIFAR	MLP	Std	**54.04**	39.77	26.67	16.77	9.95	0.17
		LCR	50.09	46.23	41.65	37.09	32.62	0.45
		AMR	50.36	**46.28**	**42.81**	**39.06**	**35.67**	**0.54**
	CNN	Std	78.77	54.32	39.81	37.14	36.27	0.09
		LCR	**80.54**	**70.84**	58.72	44.82	32.54	0.26
		AMR	78.12	69.59	**59.84**	**49.02**	**38.65**	**0.31**

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin and N. Usunier (2017). *Parseval networks: Improving robustness to adversarial examples.* ICML.
J. Lin, C. Gan and S. Han (2019). *Defensive Quantization: When Efficiency Meets Robustness*. ICLR.

Experiments cont'

		Method	Clean	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$	$\epsilon = 2.0$
MNIST	LeNet	Std	98.98	95.75	80.59	44.36	22.64
		DD	99.34	**97.64**	92.09	83.53	79.35
		Adv	**99.48**	97.45	91.99	88.88	87.51
		AMR	99.01	96.80	**94.03**	**93.61**	**93.12**
	LeNetSmall	MMR	97.43	89.90	58.86	23.95	6.80
		AMR	**97.83**	**91.94**	**73.70**	**32.56**	**8.79**
		Method	Clean	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
CIFAR	ConvNet	Std	76.61	56.47	37.37	30.17	28.89
		DD	**79.10**	**68.95**	59.51	**56.84**	**56.38**
		Adv	75.27	68.79	**61.79**	54.35	47.00
		AMR	76.77	68.00	57.87	52.38	50.03
	LeNetSmall	MMR	59.07	49.01	39.42	30.41	22.53
		AMR	**68.40**	**61.33**	**53.84**	**46.03**	**39.05**

A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu (2018). *Towards deep learning models resistant to adversarial attacks*. ICLR.
Z. Yan, Y. Guo and C. Zhang (2018). *Deep defense: Training DNNs with improved adversarial robustness*. NeurIPS.
F. Croce, M. Andriushchenko and M. Hein (2019). *Provable robustness of relu networks via maximization of linear regions.* AISTATS.

Conclusion

Summary

- Lots of work done before on robustness

- Lots of work being done now on robustness

- Connections can be drawn to substantiate understanding

- And enable further development

UNIVERSITY OF WATERLOO

Questions?