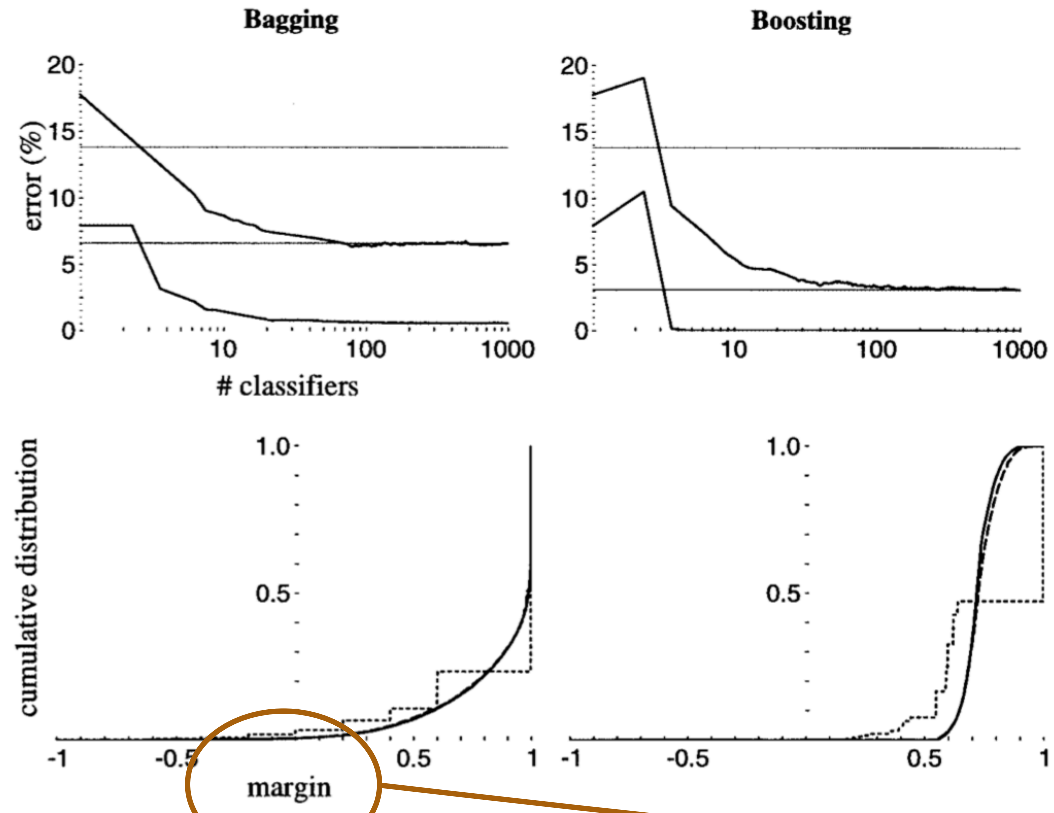
Phase Transitions of Margin Dynamics

Adaboost: resistant to overfitting



$$[f(x)]_y - \max_{\{j:j\neq y\}}[f(x)]_j$$

[Schapire et al. 1998]

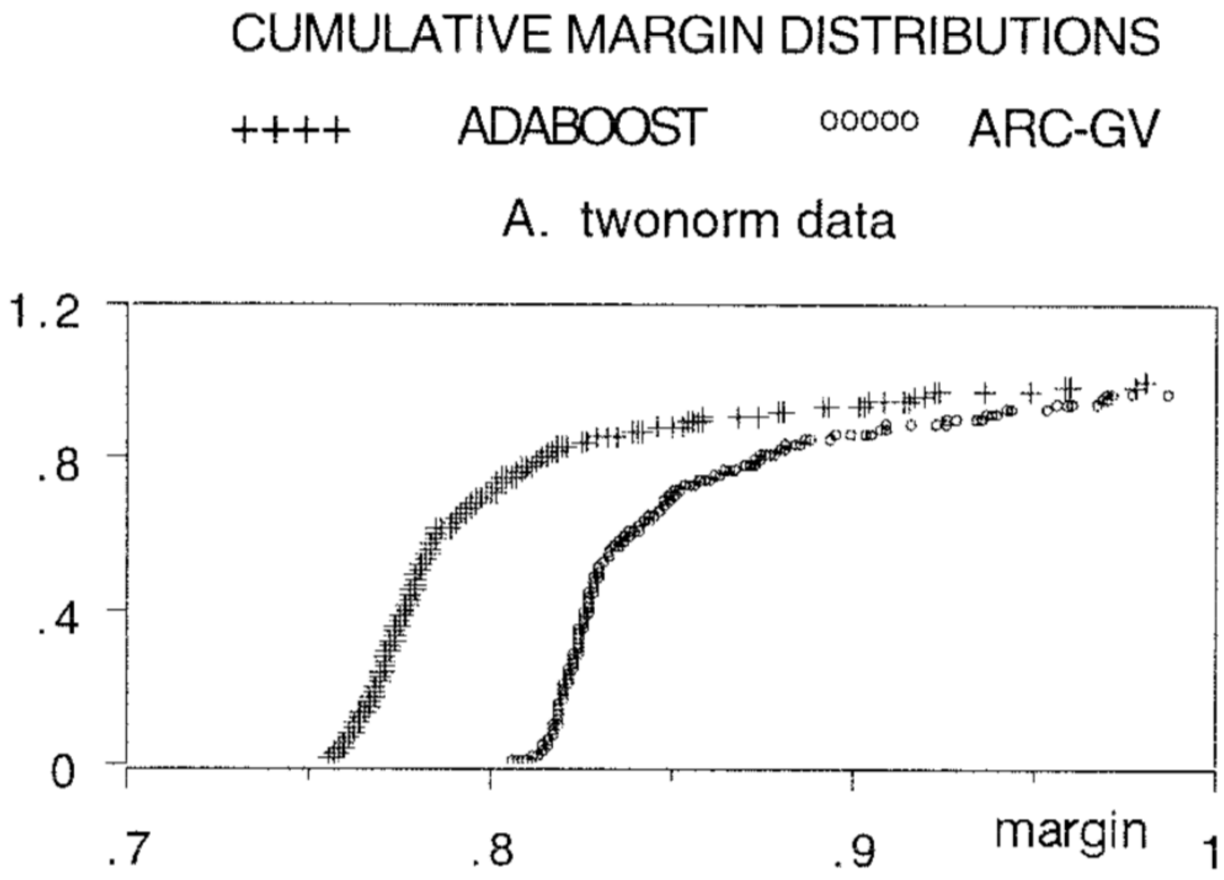
Generalization bound: margin explanation

THEOREM 2. *Let $\mathscr{D}$ be a distribution over $X \times \{-1, 1\}$, and let S be a sample of m examples chosen independently at random according to $\mathscr{D}$. Suppose the base-classifier space $\mathscr{H}$ has VC-dimension d, and let $\delta > 0$. Assume that $m \geq d \geq 1$. Then with probability at least $1 - \delta$ over the random choice of the training set S, every weighted average function $f \in \mathscr{C}$ satisfies the following bound for all $\theta > 0$:*

$$\mathbf{P}_{\mathscr{D}}[yf(x) \leq 0] \leq \mathbf{P}_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{d \log^2(m/d)}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

[Schapire et al. 1997]

Critic: margin is NOT universal



Data Set	Test Set Error	
	arc-gv	Adaboost
Twonorm		
$k = 8$	5.3	4.9
$k = 16$	6.0	4.9

[Breiman 1999]

Arcing algorithms

Definition 2. *The prediction game is a two-player zero-sum matrix game. Player I chooses* $\mathbf{z}_n \in T$. *Player II chooses* $\{c_m\}$. *Player I wins the amount* $er(\mathbf{z}_n, \mathbf{c})$.

$$\phi^* = \inf_{\mathbf{c}} \sup_{Q} E_Q er(\mathbf{z}, \mathbf{c}) = \sup_{Q} \inf_{\mathbf{c}} E_Q er(\mathbf{z}, \mathbf{c})$$

pure strategy

mix strategy

NEGATIVE related to margin

$$\text{arc-gv:} \quad \lim_{k \to \infty} \sup_{Q} \mathbb{E}_Q \left[er(z, c_k) \right] = \phi^*$$

State-of-Art Result in Neural Networks

Theorem 1.1. *Let nonlinearities* $(\sigma_1, \ldots, \sigma_L)$ *and reference matrices* $(M_1, \ldots, M_L)$ *be given as above (i.e.,* σ_i *is* ρ_i*-Lipschitz and* $\sigma_i(0) = 0$*). Then for* $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ *drawn iid from any probability distribution over* $\mathbb{R}^d \times \{1, \ldots, k\}$*, with probability at least* $1 - \delta$ *over* $((x_i, y_i))_{i=1}^n$*, every margin* $\gamma > 0$ *and network* $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$ *with weight matrices* $\mathcal{A} = (A_1, \ldots, A_L)$ *satisfy*

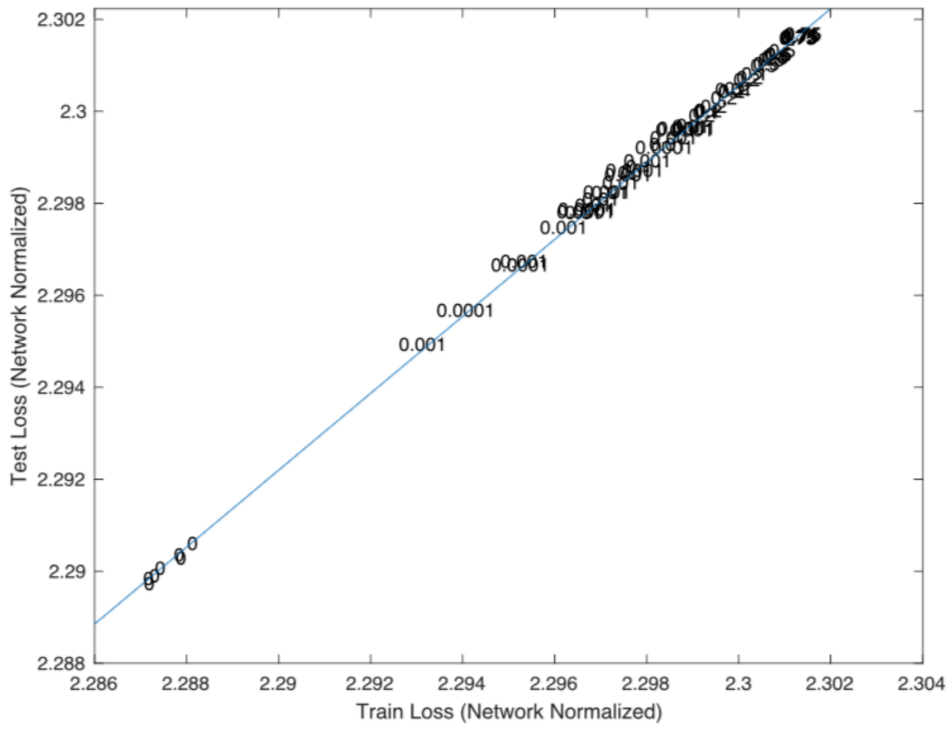
$$\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_2 R_{\mathcal{A}}}{\gamma n}\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right),$$

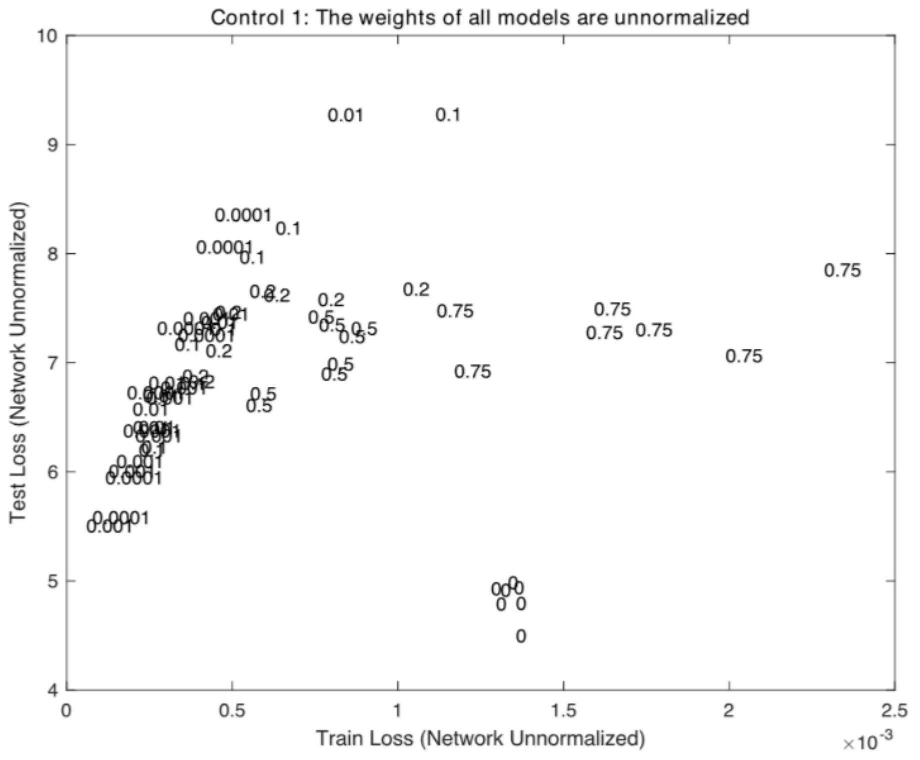
where $\widehat{\mathcal{R}}_\gamma(f) \leq n^{-1}\sum_i \mathbb{1}\left[f(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} f(x_i)_j\right]$ *and* $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

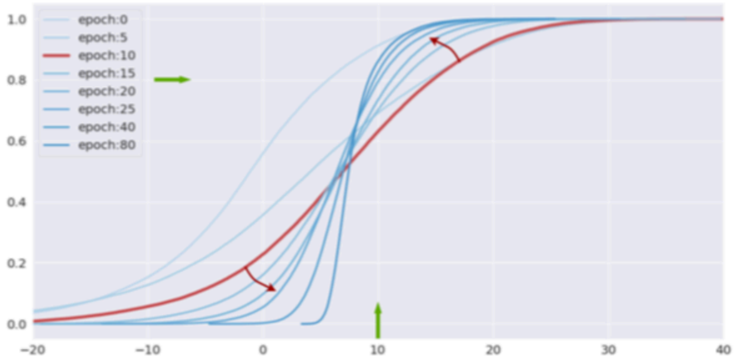
$$R_{\mathcal{A}} := \left(\prod_{i=1}^L \rho_i \|A_i\|_\sigma\right)\left(\sum_{i=1}^L \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\right)^{3/2}$$

[Bartlett et al. 2017; Koltchinskii et al. 2002]

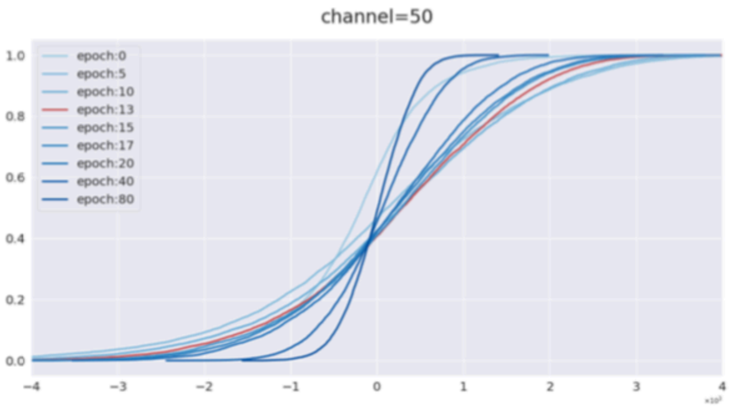
Normalizing Networks



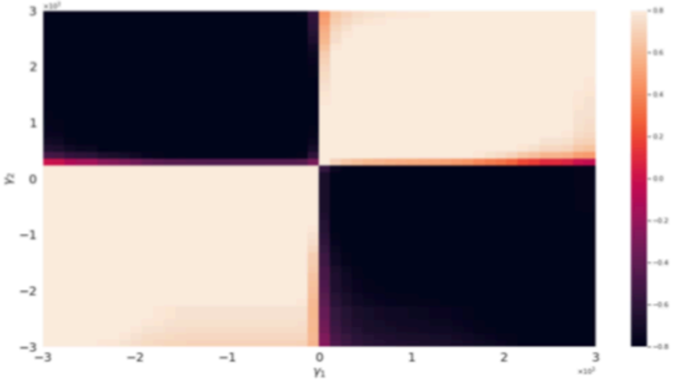




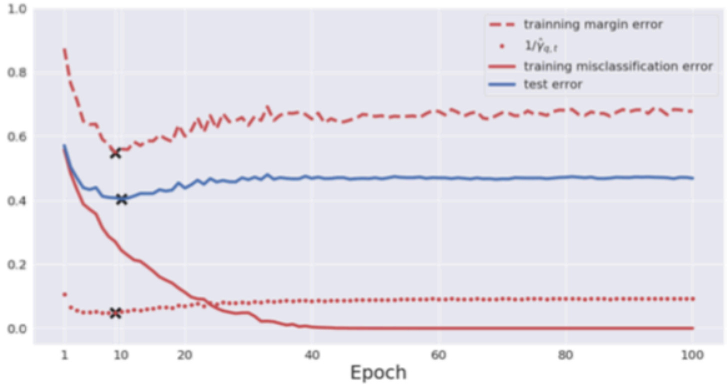
(a) Training Margin Distributions



(b) Test Margin Distributions

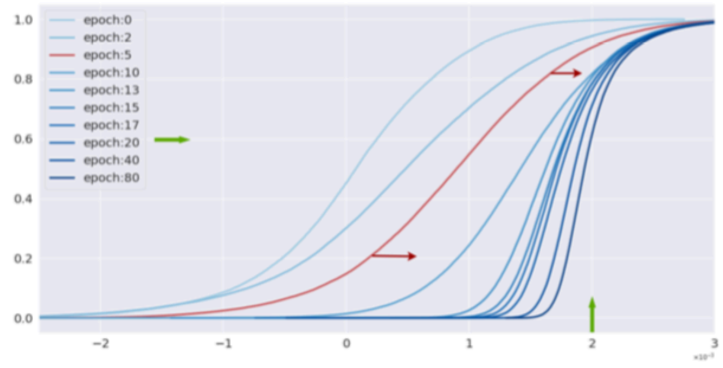


(c) Rank correlations (Spearman-ρ)

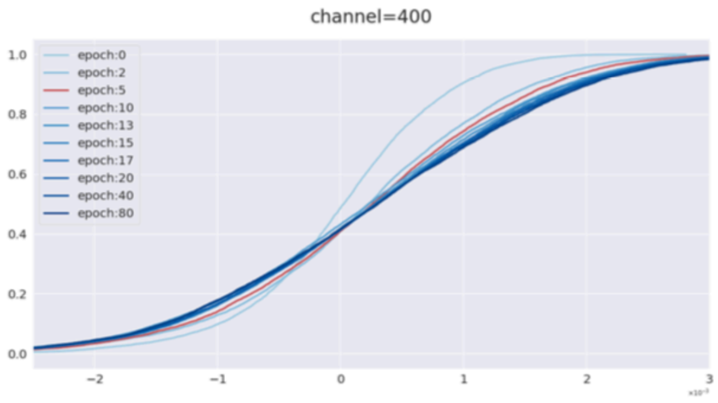


(d) test error prediction

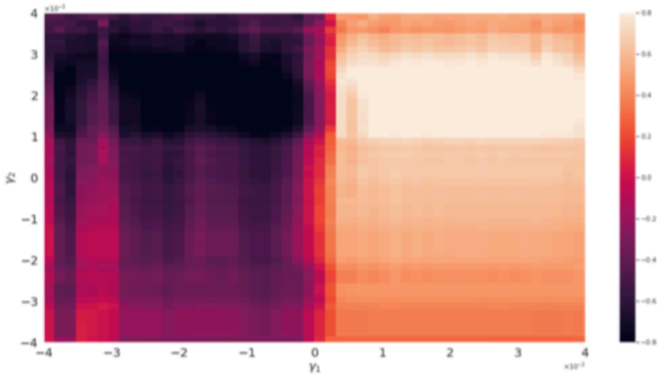
Data: CIFAR10
Network: CNN (50)



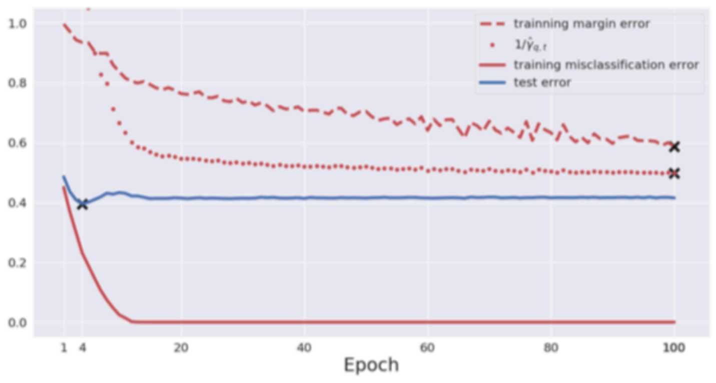
(a) Training Margin Distributions



(b) Test Margin Distributions



(c) Rank correlations (Spearman-ρ)



(d) Overfitting

Data: CIFAR10
Network: CNN (400)

Notations [1/3]

Network

Define $\mathcal{F}$ to be the space of functions $f : \mathcal{X} \to \mathbb{R}^K$ represented by neural networks,

$$\begin{cases} x_0 &= x, \\ x_i &= \sigma_i(W_i x_{i-1} + b_i), \qquad i = 1, \ldots, l-1, \\ f(x) &= W_l x_{l-1} + b_l, \end{cases} \tag{1}$$

Lipschitz semi-norm

$$\|f\|_{\mathcal{F}} := \sup_{x \neq x\prime} \frac{\|f(x) - f(x')\|_2}{\|x - x'\|_2} \leq L_\sigma \prod_{i=1}^{l} \|W_i\|_\sigma := L_f, \tag{2}$$

Notations [2/3]

Hypothesis space

$$\mathcal{H} = \{h(x) = [f(x)]_y : \mathcal{X} \to \mathbb{R}, f \in \mathcal{F}, y \in \mathcal{Y}\},$$

Restricted hypothesis space

$$\mathcal{H}_L = \{h(x) = [f(x)]_y : \mathcal{X} \to \mathbb{R}, h(x) = [f(x)]_y \in \mathcal{H} \text{ with } \|f\|_{\mathcal{F}} \leq L, y \in \mathcal{Y}\}.$$

Notations [3/3]

Margin

$$\zeta(f(x), y) = [f(x)]_y - \max_{\{j:j \neq y\}} [f(x)]_j$$

Margin Error

$$e_\gamma(f(x), y) = \begin{cases} 1 & \zeta(f(x), y) \leq \gamma \\ 0 & \zeta(f(x), y) > \gamma \end{cases} .$$

Classic result

Lemma 2.1. *Given a $\gamma_0 > 0$, then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for any $f \in \mathcal{F}$ with $\|f\|_{\mathcal{F}} \leq L$,*

$$\mathbb{E}[\ell_{\gamma_0}(f(x), y)] \leq \frac{1}{n}\sum_{i=1}^{n}[\ell_{\gamma_0}(f(x_i), y_i)] + \frac{4K}{\gamma_0}\mathcal{R}_n(\mathcal{H}_L) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{6}$$

where

$$\mathcal{R}_n(\mathcal{H}_L) = \mathbb{E}_{x_i,\varepsilon_i} \sup_{h \in \mathcal{H}_L} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(x_i) \tag{7}$$

is the Rademacher complexity of function class $\mathcal{H}_L$ with respect to n samples, and the expectation is taken over x_i, ε_i, $i = 1, ..., n$.

[Koltchinskii et al. 2002; Cortes et al. 2013; Kuznetsov et al. 2015]

Necessity of Normalizing Network

Proposition 1. *Consider the networks with activation functions σ, where we assume σ is Lipschitz continuous and there exists x_0 such that $\sigma'(x_0) \neq 0$ and $\sigma''(x_0)$ exists. Then for any $L > 0$, there holds,*

$$\mathcal{R}_n(\mathcal{H}_L) \geq CL\mathbb{E}_S[\sqrt{x_1^2 + \ldots + x_n^2}] \tag{8}$$

where $C > 0$ is a constant that does not depend on S.

[Zhu, Huang and Yao 2018; Bartlett et al. 2017]

Normalized margin error

Theorem 1. *Given γ_1 and γ_2 such that $\gamma_2 > \gamma_1 \geq 0$ and $\Delta := \gamma_2 - \gamma_1 \geq 0$, for any $\delta > 0$, with probability at least $1-\delta$, along the training epoch $t = 1, \ldots, T$, the following holds for each f_t,*

$$\mathbb{P}[\zeta(\tilde{f}_t(x), y) < \gamma_1] \leq \mathbb{P}_n 1[\zeta(\tilde{f}_t(x), y) < \gamma_2] + \frac{C_{\mathcal{H}}}{\Delta} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{9}$$

where $C_{\mathcal{H}} = 4K\mathcal{R}_n(\mathcal{H}_1)$.

Remark. *In particular, when we take $\gamma_1 = 0$ and $\gamma_2 = \gamma > 0$, the bound above becomes,*

$$\mathbb{P}[\zeta(f_t(x), y) < 0] \leq \mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma] + \frac{C_{\mathcal{H}}}{\gamma} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{10}$$

[Zhu, Huang and Yao 2018]

Dual perspective: quantile margin

$$\hat{\gamma}_{q,f} = \inf\left\{\gamma : \mathbb{P}_n 1[\zeta(f(x_i), y_i) \leq \gamma] \geq q\right\}. \tag{11}$$

Theorem 2. *Assume the input space is bounded by $M > 0$, that is $\|x\|_2 \leq M$, $\forall x \in \mathcal{X}$. Given a quantile $q \in [0,1]$, for any $\delta \in (0,1)$ and $\tau > 0$, the following holds with probability at least $1 - \delta$ for all f_t satisfying $\hat{\gamma}_{q,\tilde{f}_t} > \tau$,*

$$\mathbb{P}[\zeta(f_t(x), y) < 0] \leq C_q + \frac{C_{\mathcal{H}}}{\hat{\gamma}_{q,\tilde{f}_t}} \tag{12}$$

$$C_q = q + \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{\log\log_2(4(M+l)/\tau)}{n}} \text{ and } C_{\mathcal{H}} = 8K\mathcal{R}_n(\mathcal{H}_1).$$

Remark. *We simply denote $\gamma_{q,t}$ for $\gamma_{q,\tilde{f}_t}$ when there is no confusion.*

[Zhu, Huang and Yao 2018]

Example

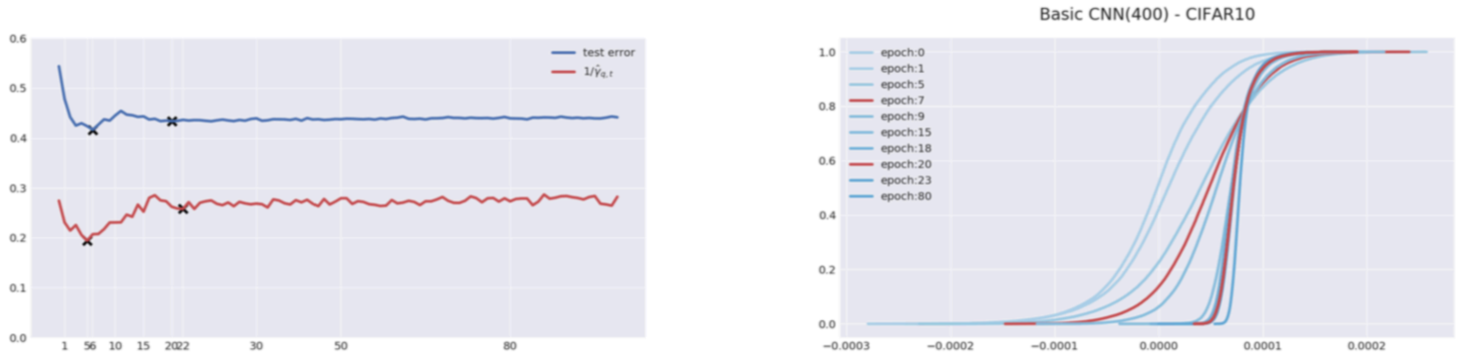
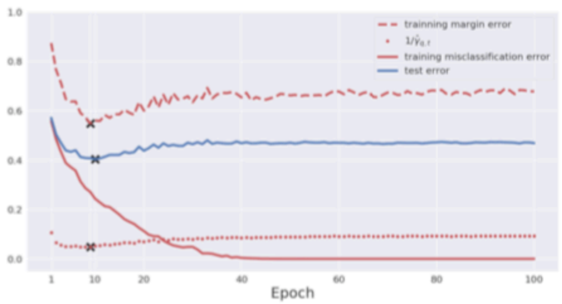
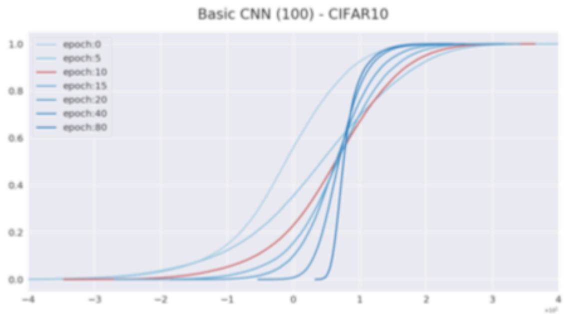
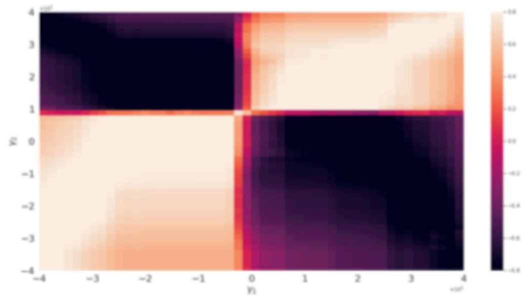


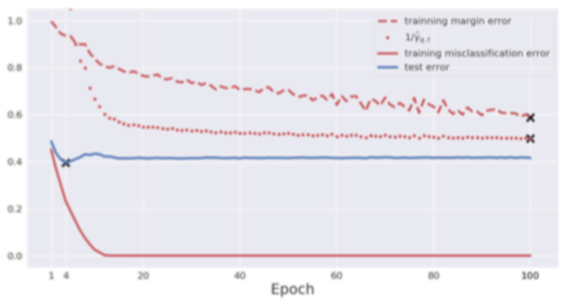
Figure 6: Inverse quantile margin. Net structure: CNN(400). Dataset: CIFAR10 with 10 percents label corrupted. Left: the dynamics of test error (blue) and inverse quantile margin with $q = 0.95$ (red). Two local minima are marked by "x" in each curve. Right: dynamics of training margin distributions, where two distributions in red color correspond to when the two local minima occur. The inverse quantile margin successfully captures two local minima of test error.

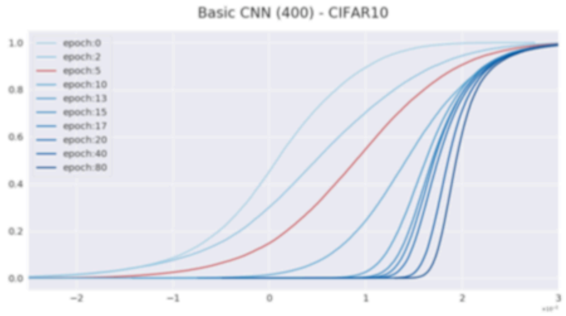
Success and Failure

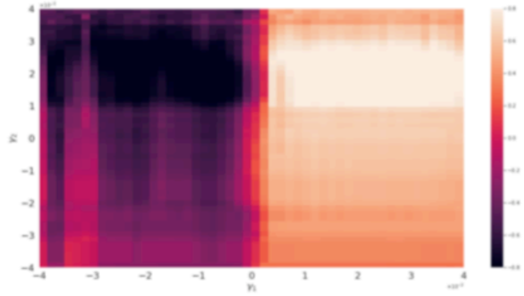


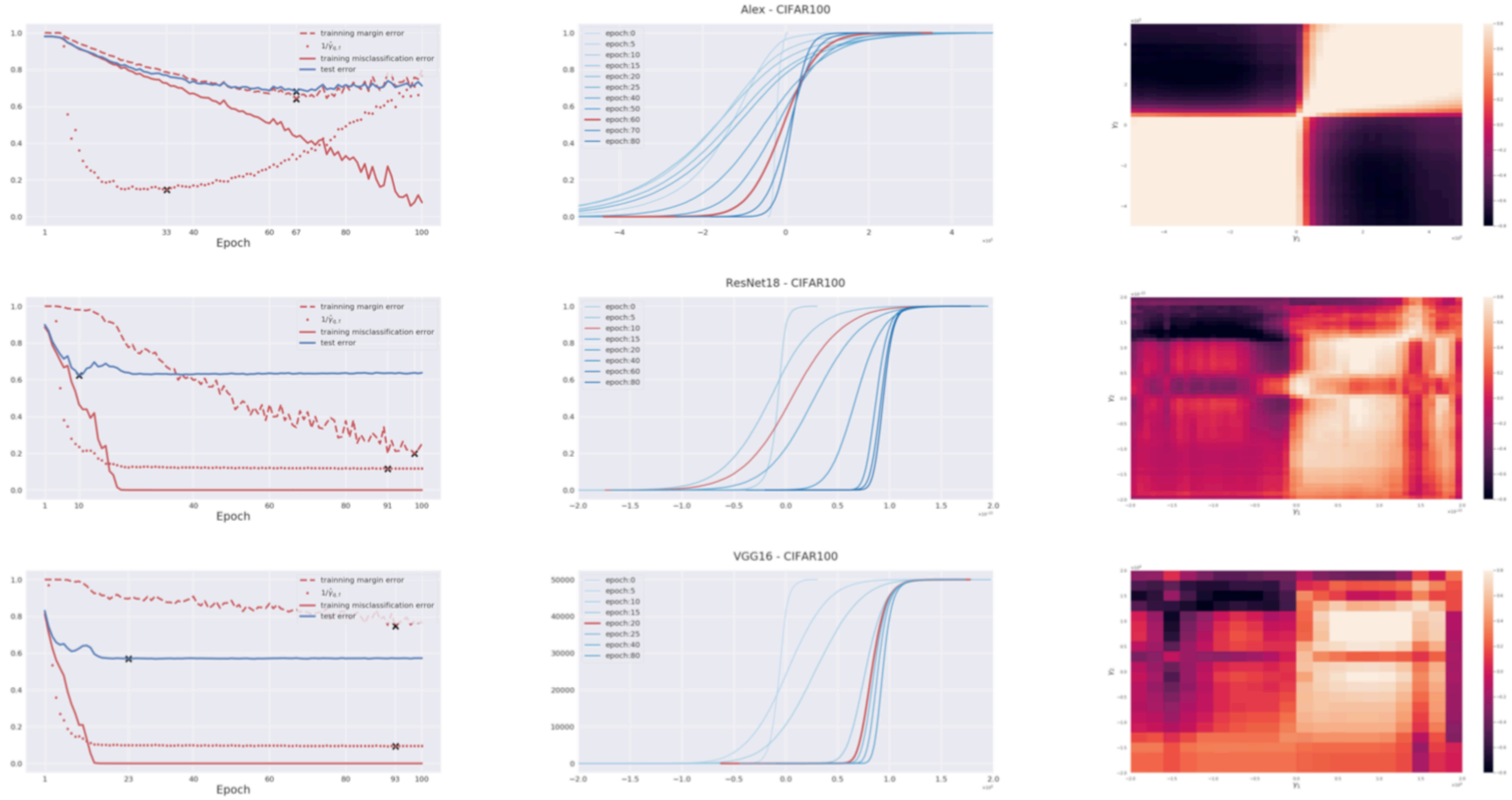


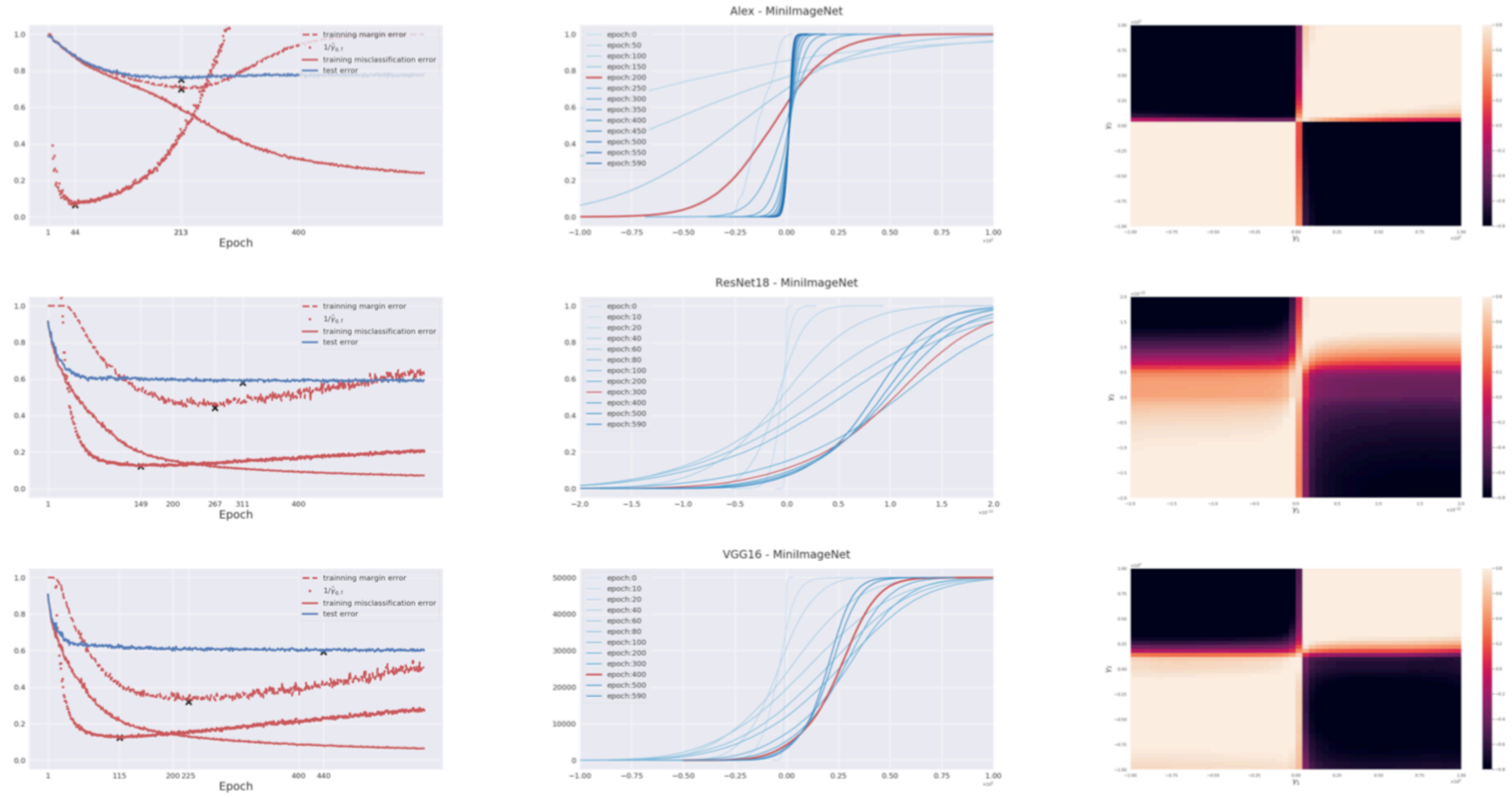









Alex - CIFAR100
ResNet18 - CIFAR100
VGG16 - CIFAR100
trainning margin error
training misclassification error
test error
Epoch
epoch:0
epoch:5
epoch:10
epoch:15
epoch:20
epoch:25
epoch:40
epoch:50
epoch:60
epoch:70
epoch:80


Alex - MiniImageNet
ResNet18 - MiniImageNet
VGG16 - MiniImageNet
trainning margin error
training misclassification error
test error
Epoch

Summary

1. Phase transitions of normalized margin dynamics shed light on model expressiveness against data complexity.

2. When model expressiveness is comparable to data complexity, such that training margins and test margins share similar phase transitions, one can predict test error using training margin dynamics by restricted Rademacher complexity bounds.

3. When model is over-expressive against data, such that training margins are monotonically improved in training, training margins will fail to predict test error.