On Mathematical Theories of Deep Learning: iii

Yuan YAO

HKUST

Generalization Ability

Over-parameterized models may generalize well without overfitting by maximizing margins

Generalization Error

- Consider the empirical risk minimization under i.i.d. samples

$$\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i;\theta)) + \mathcal{R}(\theta)$$

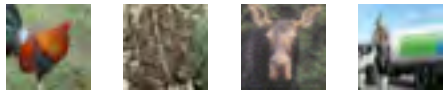
- The population risk with respect to unknown distribution

$$R(\theta) = \mathbf{E}_{x,y\sim P}\ell(y, f(x;\theta))$$

- Fundamental Theorem of Machine Learning (for 0-1 misclassification loss, called 'errors' below)

$$R(\theta) = \underbrace{\hat{R}_n(\theta)}_{\text{training loss/error}} + \underbrace{R(\theta) - \hat{R}_n(\theta)}_{\text{generalization loss/error}}$$

Why big models generalize well?



CIFAR10

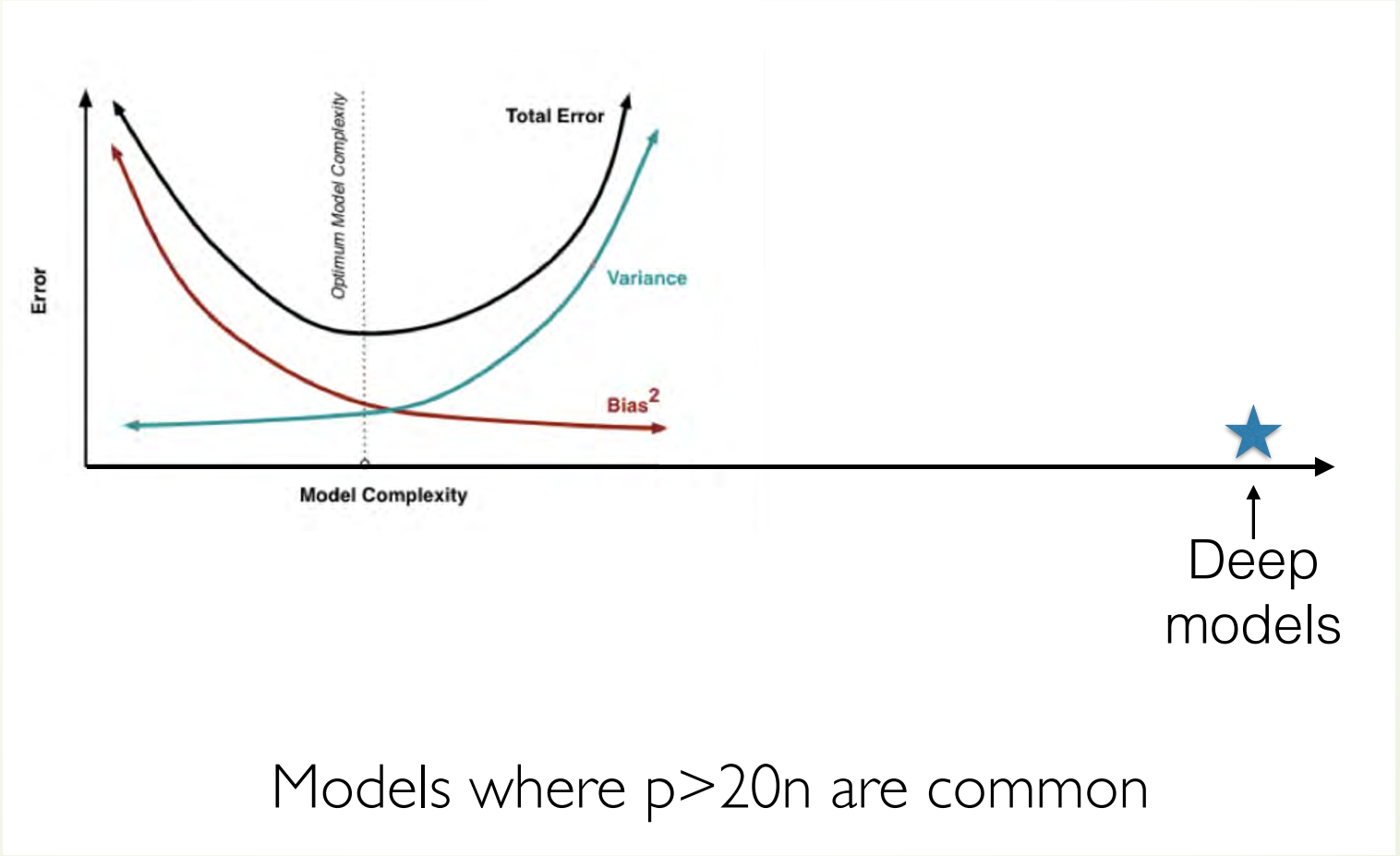
n=50,000
d=3,072
k=10

What happens when I turn off the regularizers?

Model	parameters	p/n	Train ***loss***	Test ***error***
CudaConvNet	145,578	2.9	0	23%
CudaConvNet (with regularization)	145,578	2.9	0.34	18%
MicroInception	1,649,402	33	0	14%
ResNet	2,401,440	48	0	13%

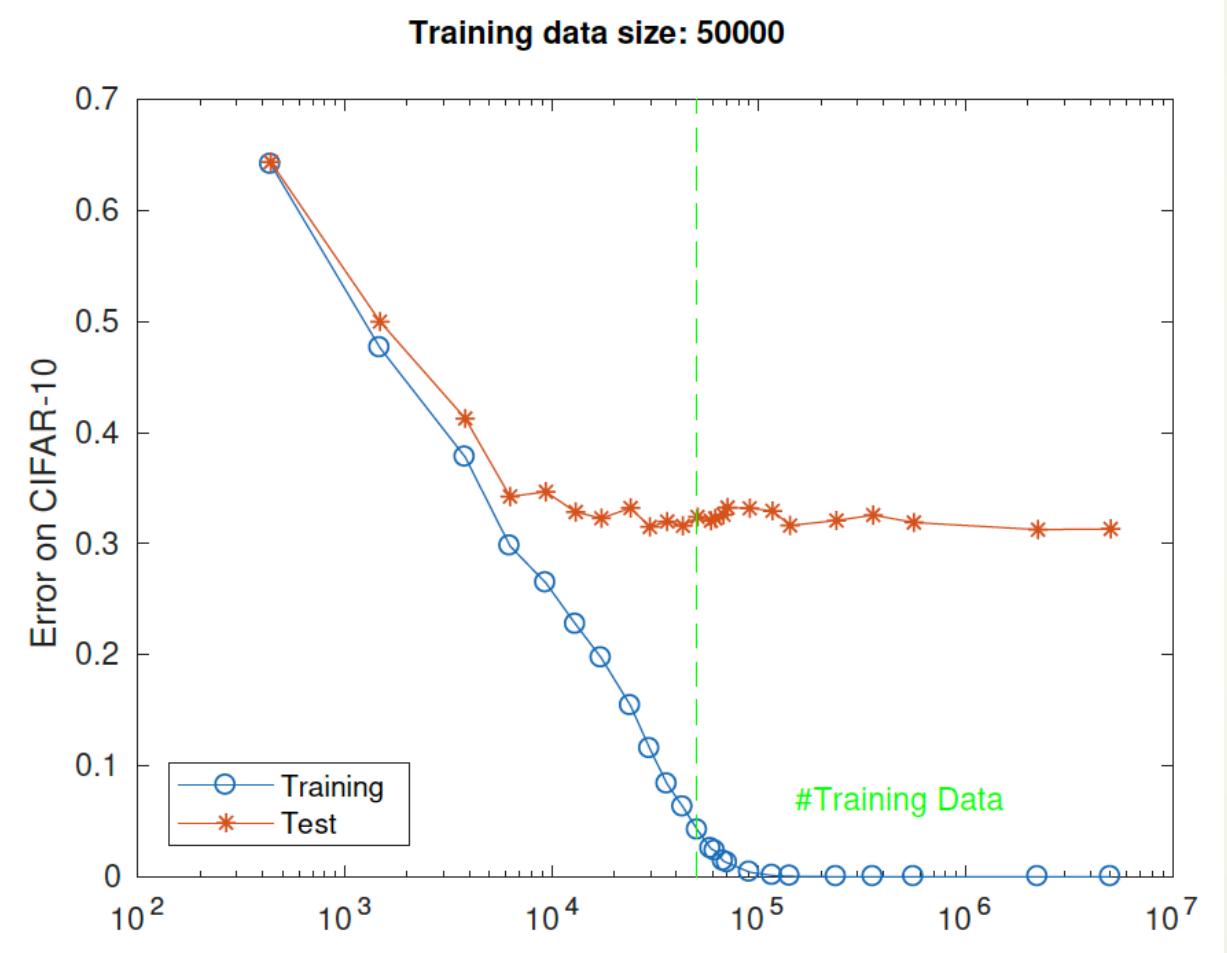
Ben Recht et al. 2016

The Bias-Variance Tradeoff?



Models where p>20n are common

Over-parameterized models

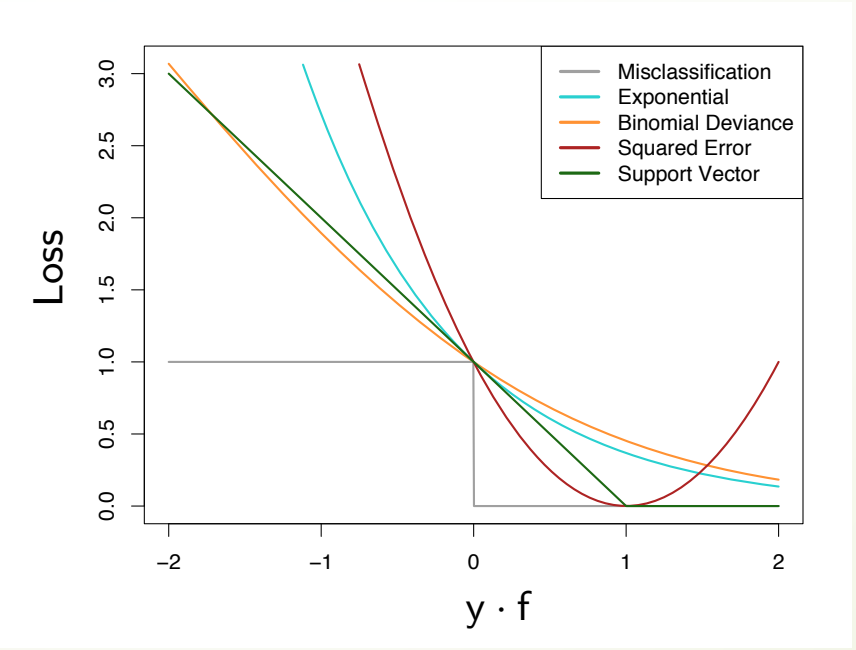


As model complexity grows (p>n), training error goes down to zero, but test error does not increase. Why overparameterized models do not overfit here? -- Tommy Poggio, 2018

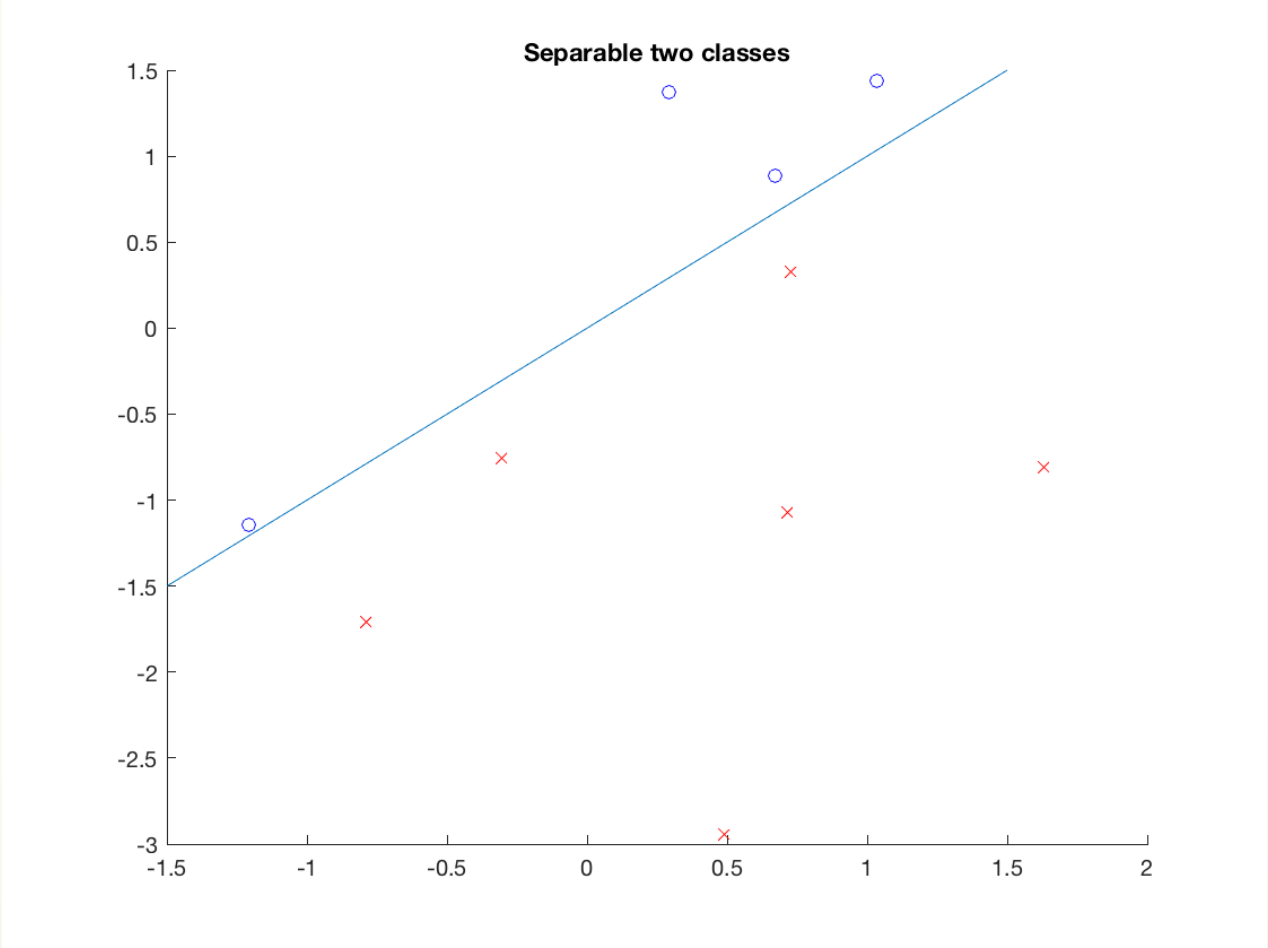
Binary Classification Problem

Consider a dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, with $\mathbf{x}_n \in \mathbb{R}^d$ and binary labels $y_n \in \{-1, 1\}$. We analyze learning by minimizing an empirical loss of the form

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \ell\left(y_n \mathbf{w}^\top \mathbf{x}_n\right) . \tag{1}$$



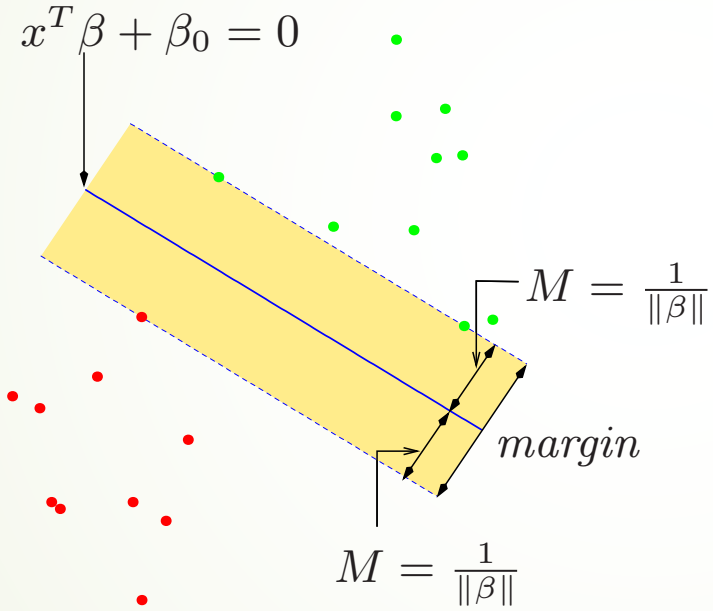
Separable Classification



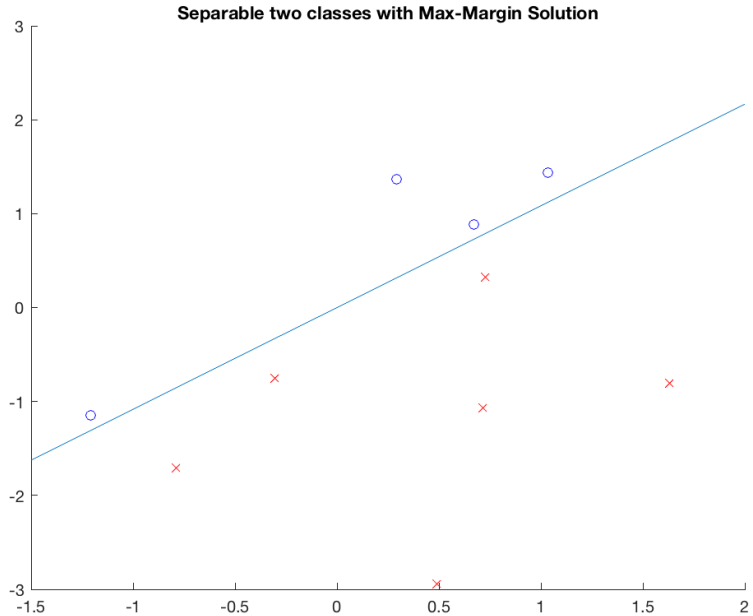
Max-Margin Classifier (SVM)

$$\text{minimize}_{\beta_0,\beta_1,\ldots,\beta_p} \|\beta\|^2 := \sum_j \beta_j^2$$

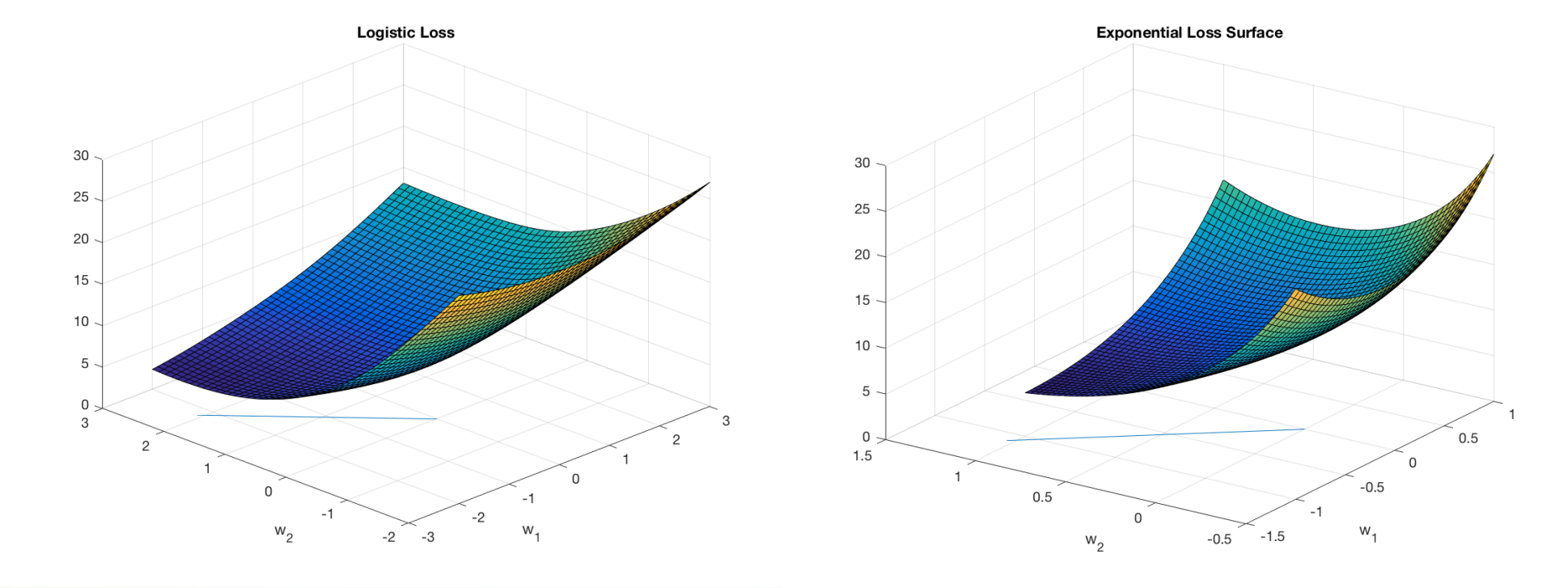
$$\text{subject to } y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq 1 \text{ for all } i$$



Vladmir Vapnik, 1994



Landscape of Logistic/Exponential Loss



The minimizers are at infinity, asymptotically in the direction of max-margin classifier

[Soudry, Hoffer, Nacson, Gunasekar, Srebro, 2017]

Theorem 3 *For any dataset which is linearly separable (Assumption 1), any β-smooth decreasing loss function (Assumption 2) with an exponential tail (Assumption 3), any stepsize $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$, the gradient descent iterates (as in eq. 2) will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log t + \boldsymbol{\rho}(t) , \tag{3}$$

where $\hat{\mathbf{w}}$ is the L_2 max margin vector (the solution to the hard margin SVM):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1, \tag{4}$$

and the residual grows at most as $\|\boldsymbol{\rho}(t)\| = O(\log\log(t))$, and so

$$\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}.$$

Furthermore, for almost all data sets (all except measure zero), the residual $\rho(t)$ is bounded.

Telgarsky 2013; Srebro et al. 2018;

Assumptions on General Loss Functions

Assumption 1 *The dataset is linearly separable:* $\exists \mathbf{w}_*$ *such that* $\forall n: \mathbf{w}_*^\top \mathbf{x}_n > 0$.

Assumption 2 $\ell(u)$ *is a positive, differentiable, monotonically decreasing to zero*[1], *(so* $\forall u: \ell(u) > 0, \ell'(u) < 0$ *and* $\lim_{u\to\infty} \ell(u) = \lim_{u\to\infty} \ell'(u) = 0$*) and a* β*-smooth function,* i.e. *its derivative is* β*-Lipshitz.*

Assumption 2 includes many common loss functions, including the logistic, exp-loss[2], probit and sigmoidal losses. Assumption 2 implies that $\mathcal{L}(\mathbf{w})$ is a $\beta\sigma_{\max}^2(\mathbf{X})$-smooth function, where $\sigma_{\max}(\mathbf{X})$ is the maximal singular value of the data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$.

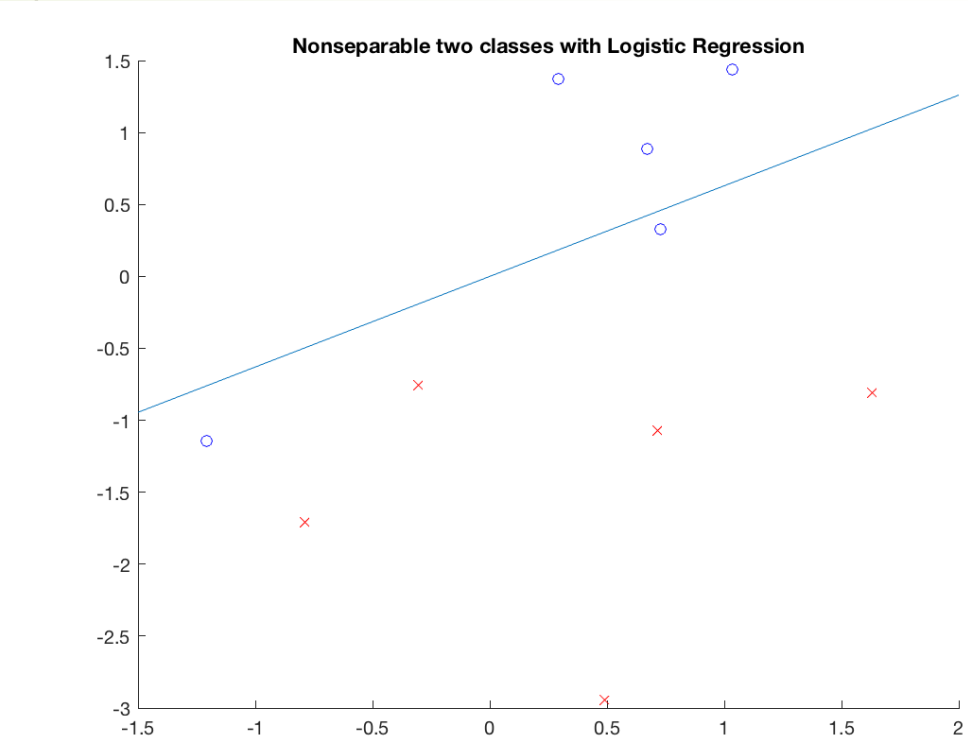
Definition 2 *A function* $f(u)$ *has a "tight exponential tail", if there exist positive constants* c, a, μ_+, μ_-, u_+ *and* u_- *such that*

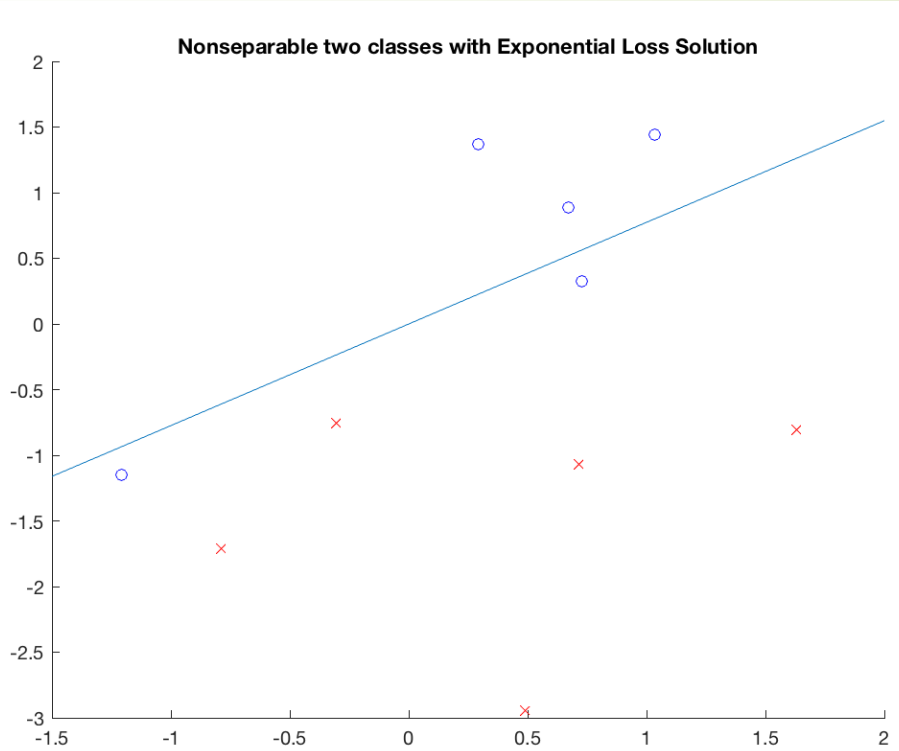
$$\forall u > u_+ : f(u) \le c\left(1 + \exp\left(-\mu_+ u\right)\right) e^{-au}$$
$$\forall u > u_- : f(u) \ge c\left(1 - \exp\left(-\mu_- u\right)\right) e^{-au}.$$

Assumption 3 *The negative loss derivative* $-\ell'(u)$ *has a tight exponential tail (Definition 2).*

For example, the exponential loss $\ell(u) = e^{-u}$ and the commonly used logistic loss $\ell(u) = \log\left(1 + e^{-u}\right)$ both follow this assumption with $a = c = 1$. We will assume $a = c = 1$ — without loss of generality, since these constants can be always absorbed by re-scaling $\mathbf{x}_n$ and η.

Nonseparable classification?

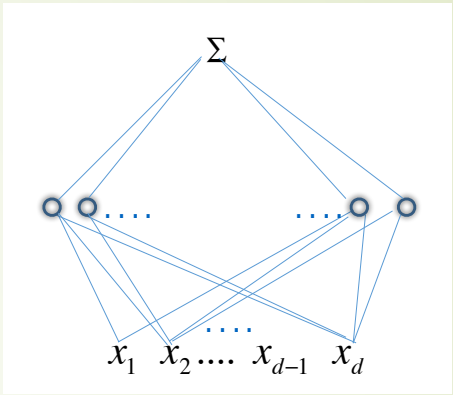




Left: GD solution for logistic loss

Right: GD solution for exponential loss

Deep Networks makes it separable

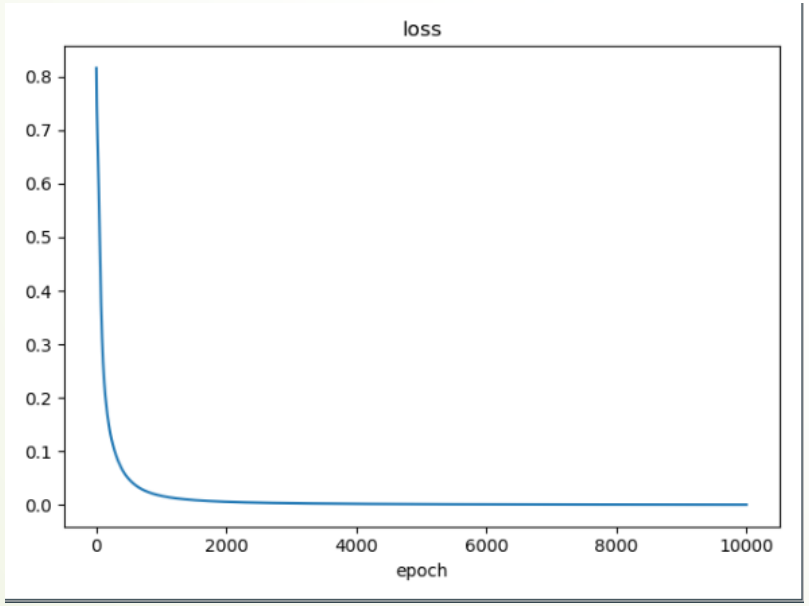


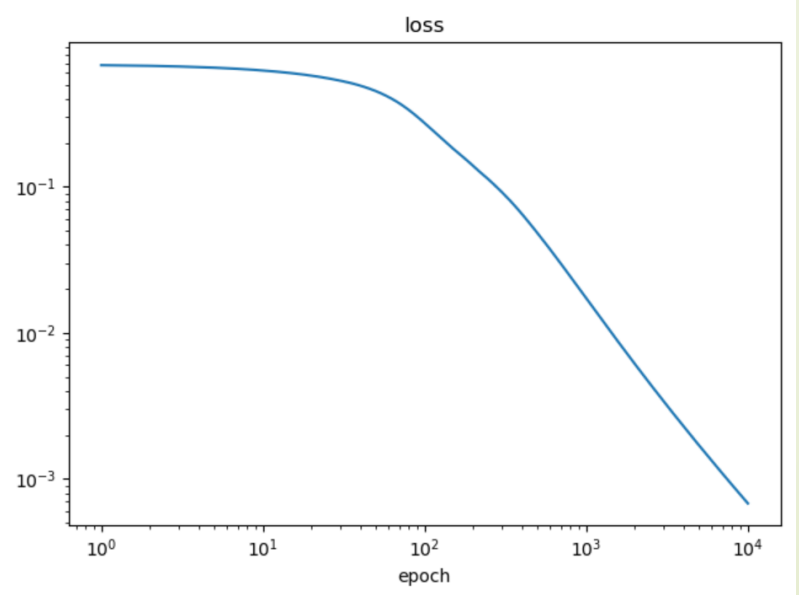
2-layer neural network:

$$f(x) = W_2\sigma(W_1x)$$

where $\sigma(u) = \max(0, u)$ is ReLU, $W_1 \in R^{d\times q}$, and $W_2 \in R^{q\times 1}$

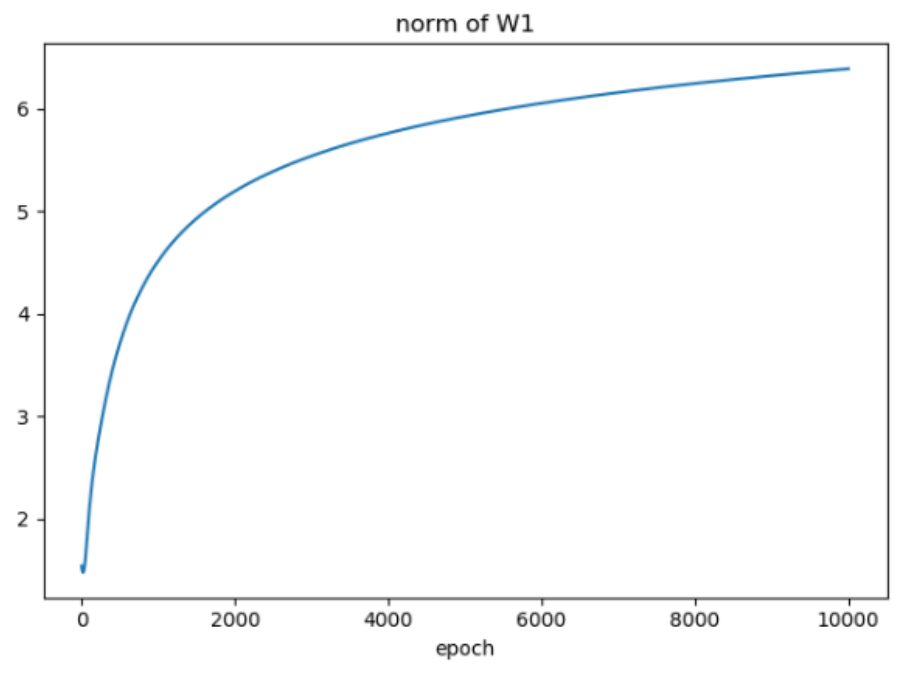
For large *q*, e.g. *q*=5, it becomes separable: logistic loss drops down at ~1/k

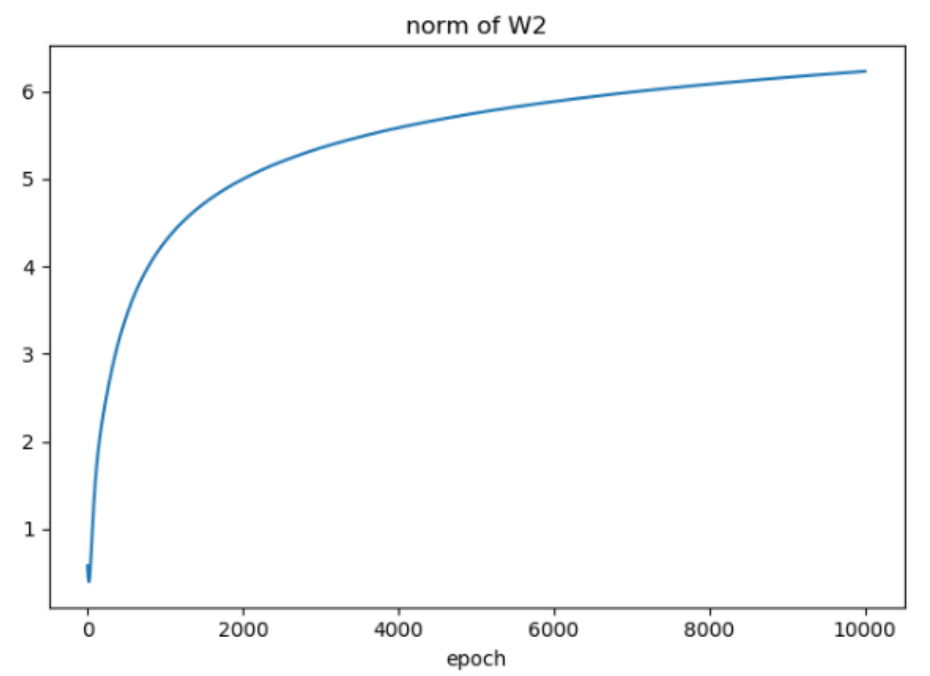




By *Yifei HUANG*, HKUST

Both W_1 and W_2 grows to infinity (log k)!





By *Yifei HUANG*

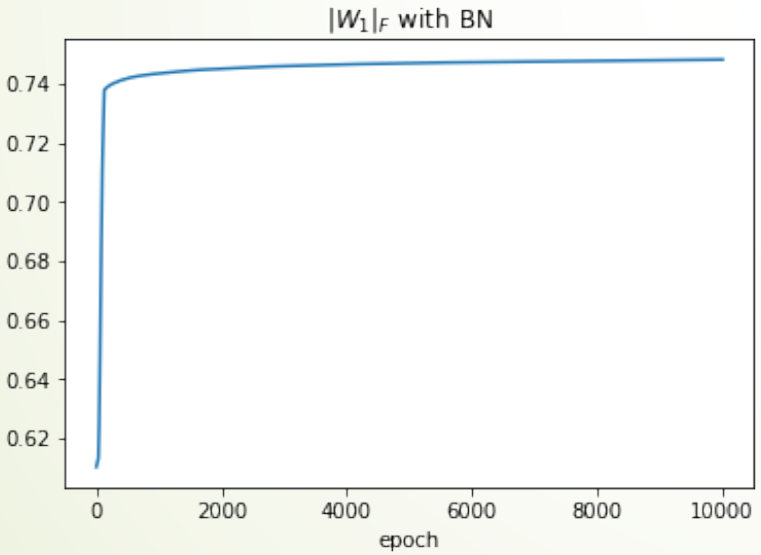
Batch Normalization

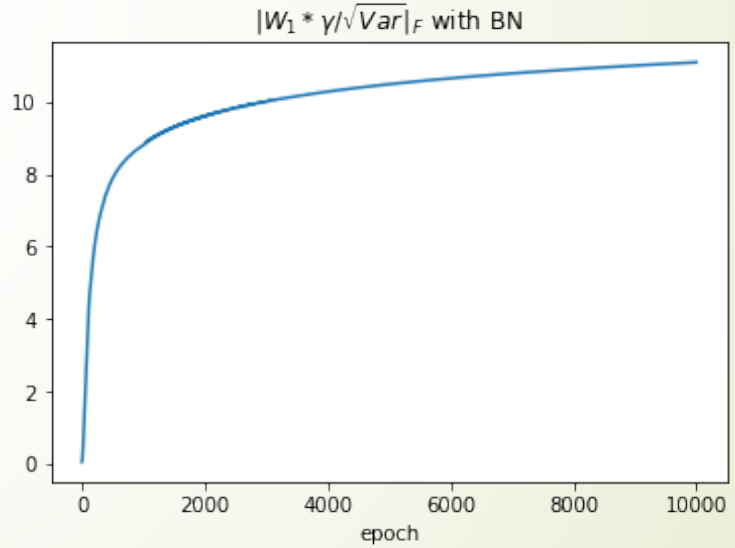
Batch Normalization:

$$\tilde{x}_p = \frac{x_p - \mu_t(x_p)}{\sqrt{\sigma_t^2(x_p) + \epsilon}}\gamma + \beta, \quad \epsilon = 10^{-5}.$$

So a linear layer $x_{l+1} = W_l x_l + b_l$ followed by Batch Normalization has its weight matrix rescaled by

$$\tilde{W}_l = \frac{\gamma}{\sigma_t} W_l.$$

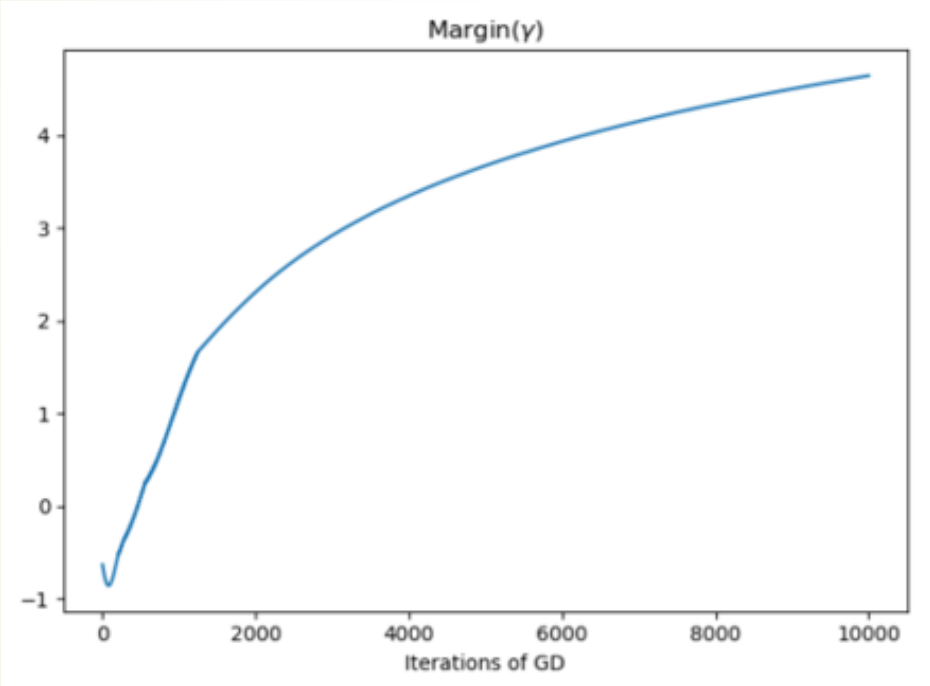


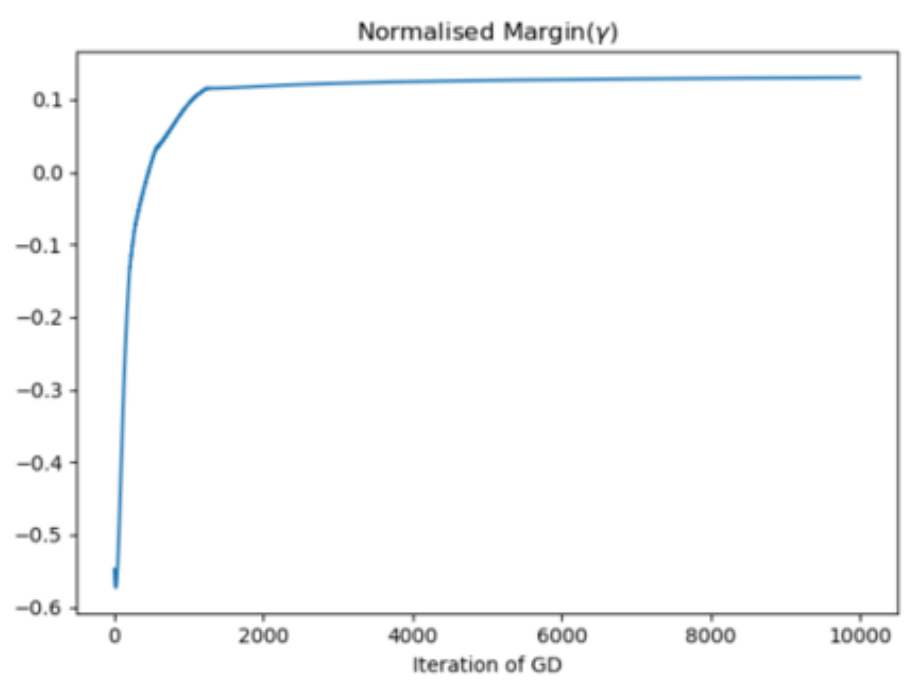


Normalized Margin is scale invariant and keeps on improving in training

$$\gamma := \min_i y_i f(x_i)$$

$$\gamma_n := \frac{\gamma}{\prod_{i=1}^{2} \|W_i\|}$$





After about 1000 epochs, it correctly classifies all training examples and continues to improve the margin.

By *Yifei HUANG*.

Spectrally-Normalized Margin Bounds [Bartlett-Foster-Telgarsky'2017]

$$F_{\mathcal{A}}(x) := \sigma_L(A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_1(A_1x)\cdots)). \tag{1.1}$$

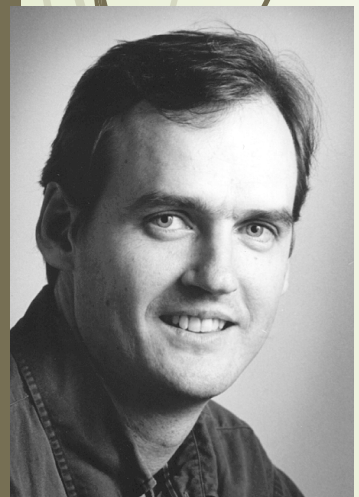
$\mathcal{A} = (A_1, \ldots, A_L)$ *reference matrices* $(M_1, \ldots, M_L)$ with the same dimensions as $A_1, \ldots, A_L$

$$R_{\mathcal{A}} := \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\right)\left(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\right)^{3/2}. \tag{1.2}$$

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* $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

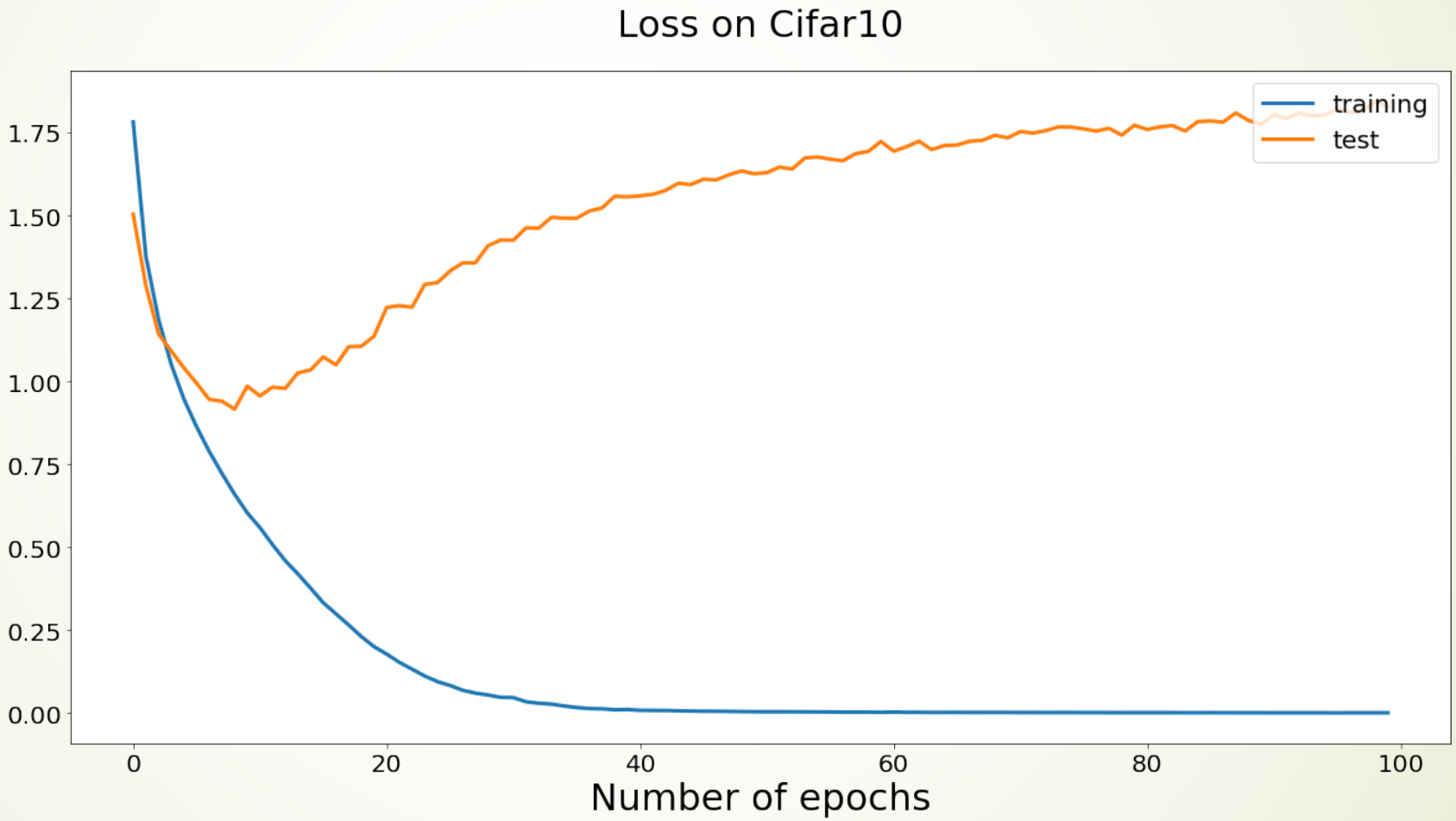


N. Srebro et al. ICLR 2018 has another PAC-Bayes generalization error bound.

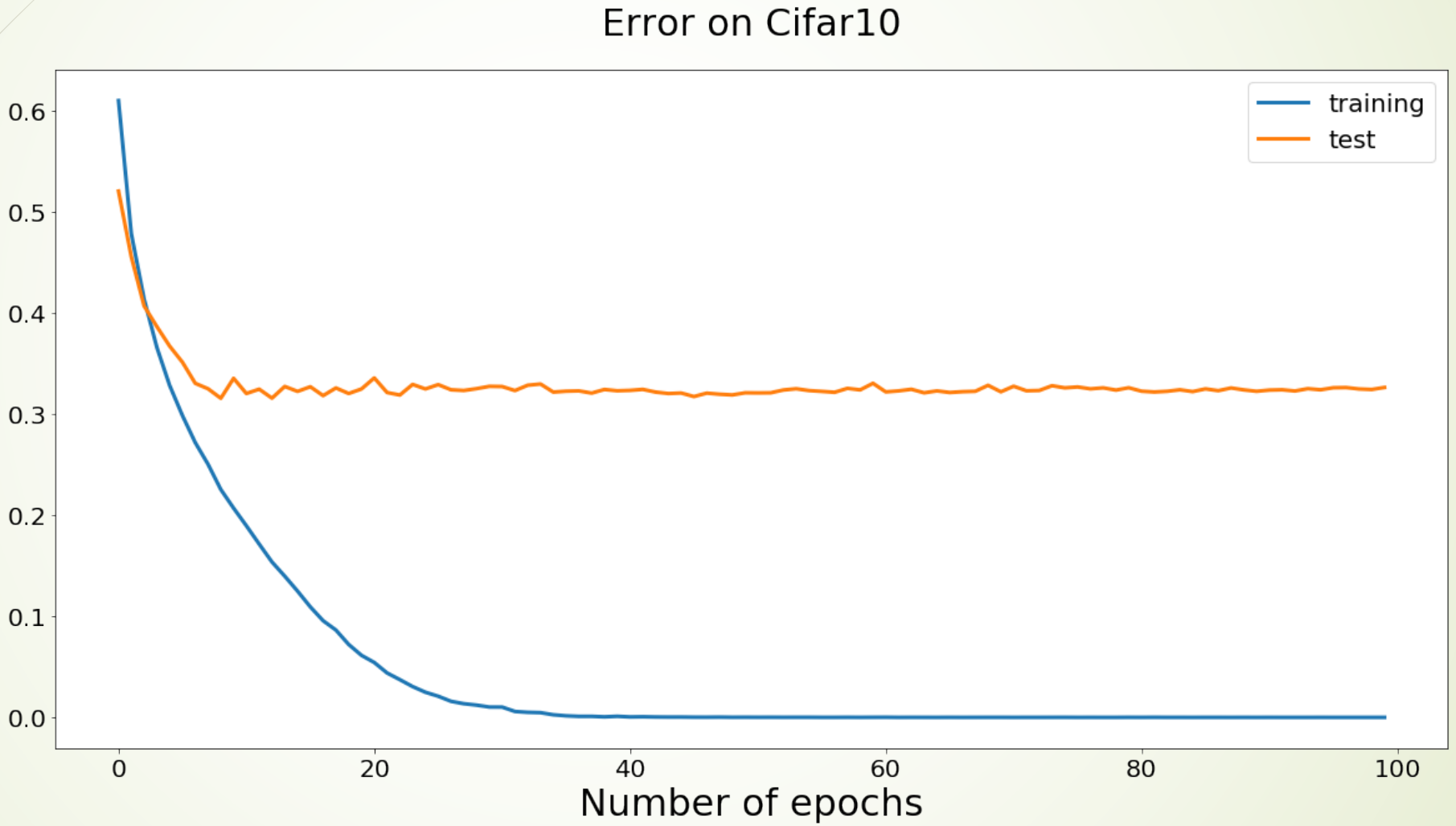
Train 5-layer CNN on Cifar10

- 5 convolutional layers:
  - 100 channels,
  - kernel 3x3,
  - stride 2,
  - padding 1
- ReLU, and/or
- Batch-normalization (batch_size=100)
- Last layer is fully-connected classifier with Cross-Entropy loss

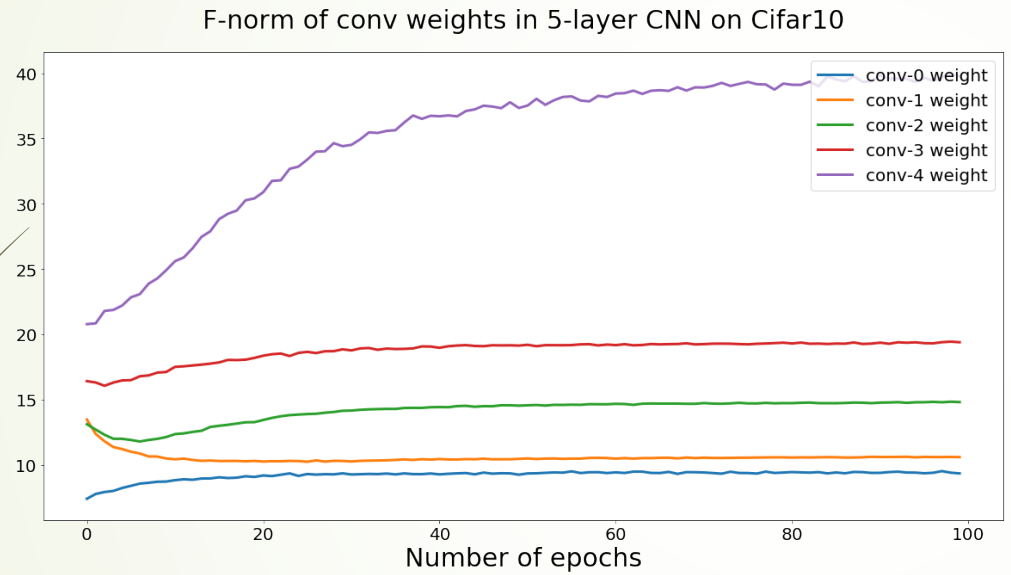
Cross-Entropy Loss overfits!



Yet, Test Error does not overfit!

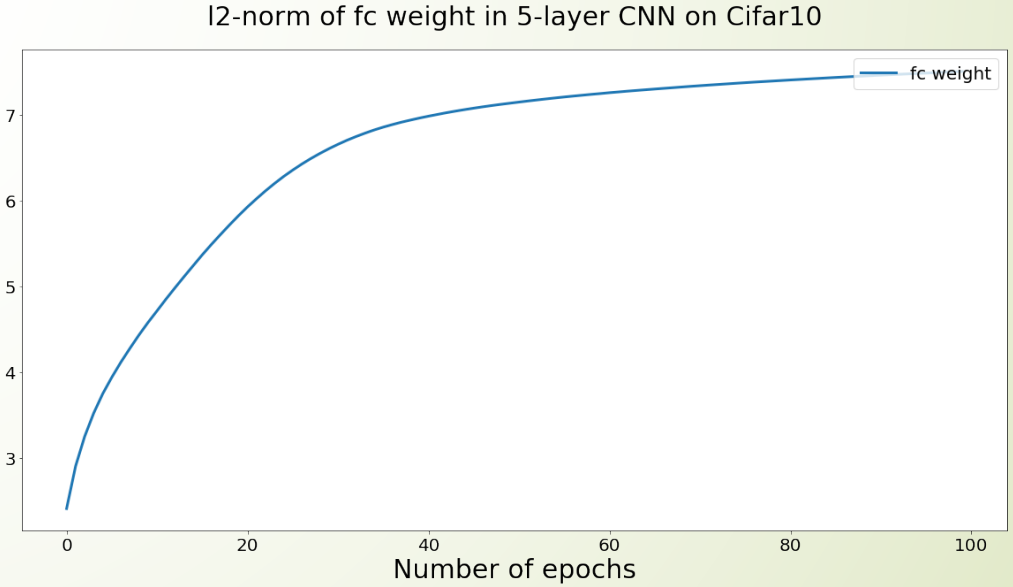


Rescaled weights are growing…

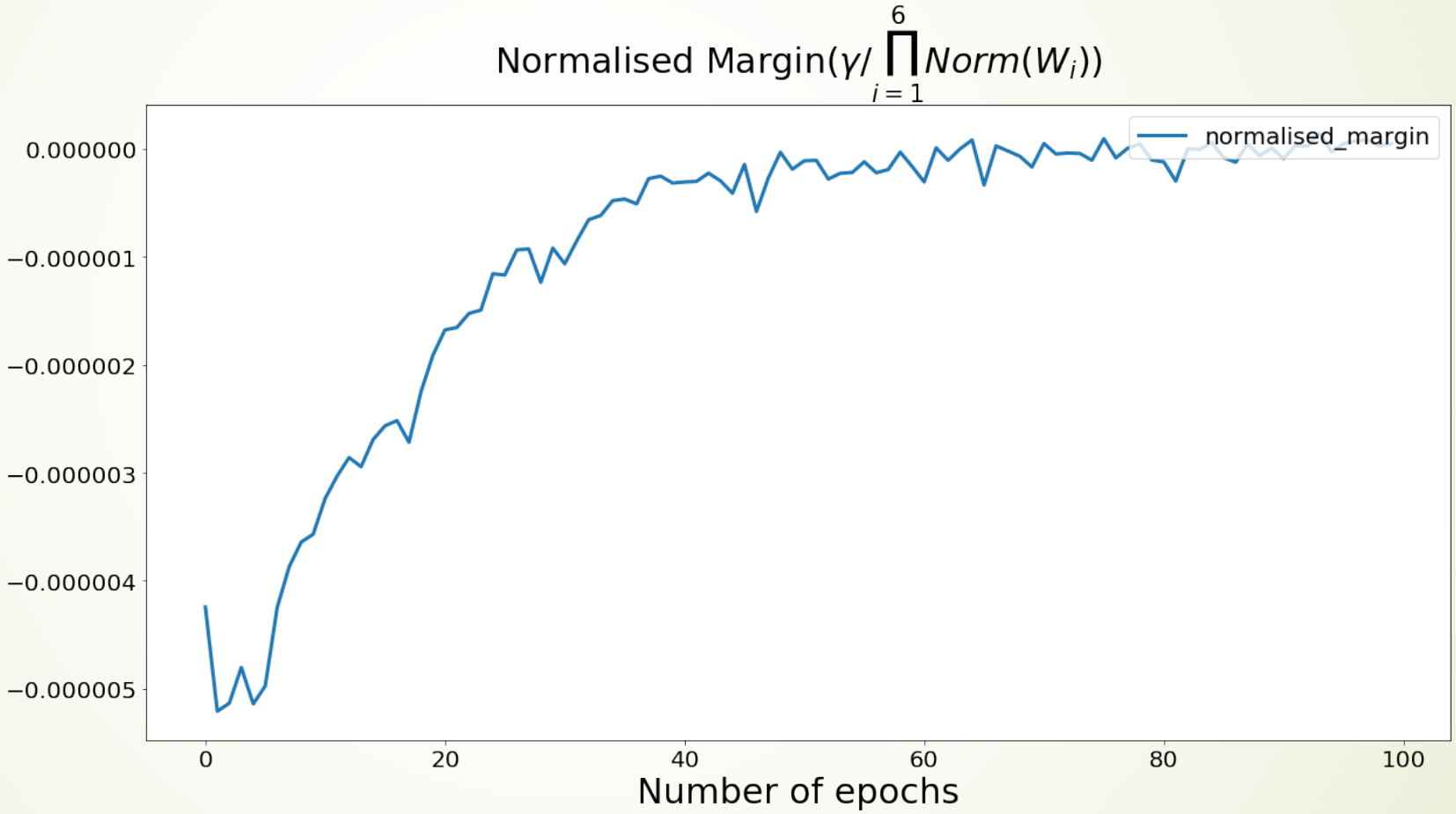


Late layer weights change fast while early ones change slowly.

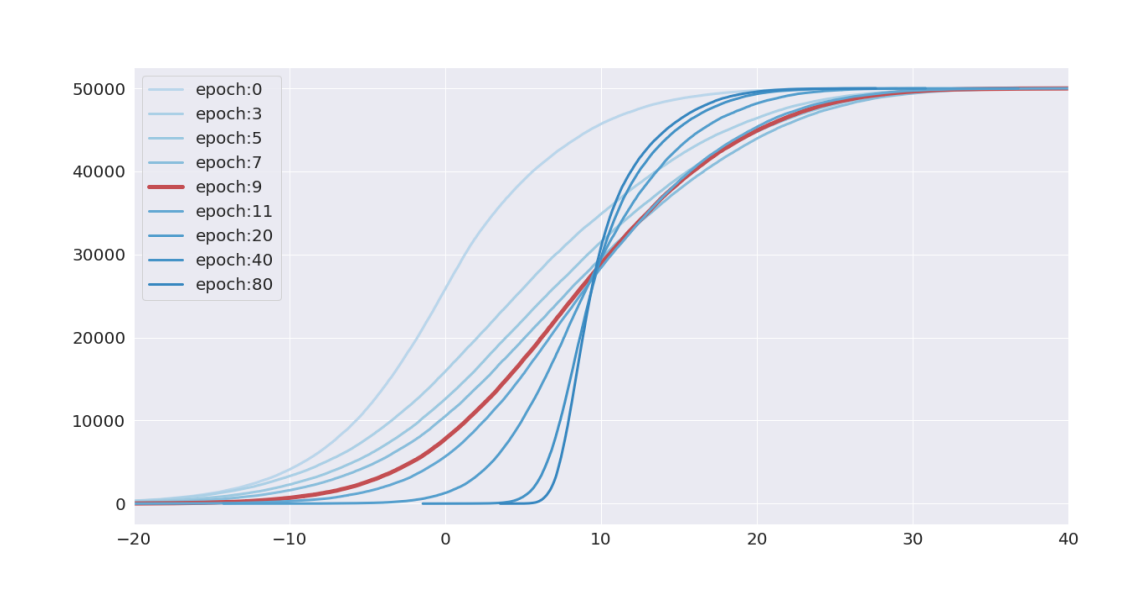
Classification layer weight grows

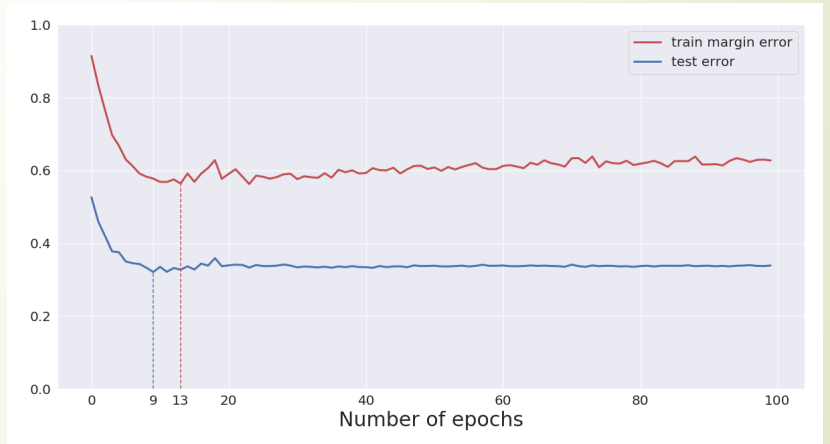


After rescaling, Normalized Margin keeps on improving with a flattened level eventually.



Dynamics of Margin Distributions



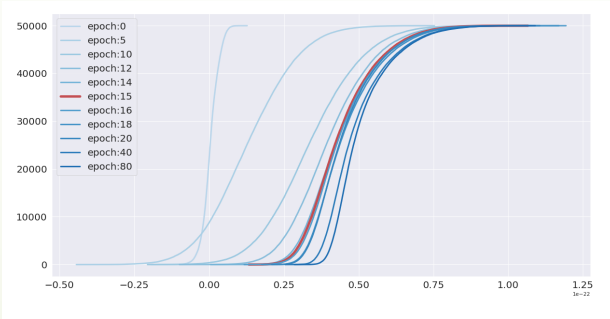


By Yifei HUANG et al.

Summary

- For separable classification, GD for logistic regression, cross entropy loss, and exponential loss, etc., converges at infinity to the maximal margin solution in direction
- For non-separable classification, over-parameterized deep networks may make it separable and GD converges toward improving margin at infinity
- Normalized margin may provide a data-dependent measure of generalization ability
- Yet, Leo Breiman's Quandary: margin may not reflect the generalization ability -- ``if we try to hard to improve margin then overfitting sets in..'' [Breiman 1998, Prediction Games and Arcing]…

Train margins of ResNet18 on Cifar10



Test margins of ResNet18