Benign Overfitting without Linearity

Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data

Yujun Kim

Jaeryeong Kim

Contents

Introduction

The Benign Overfitting

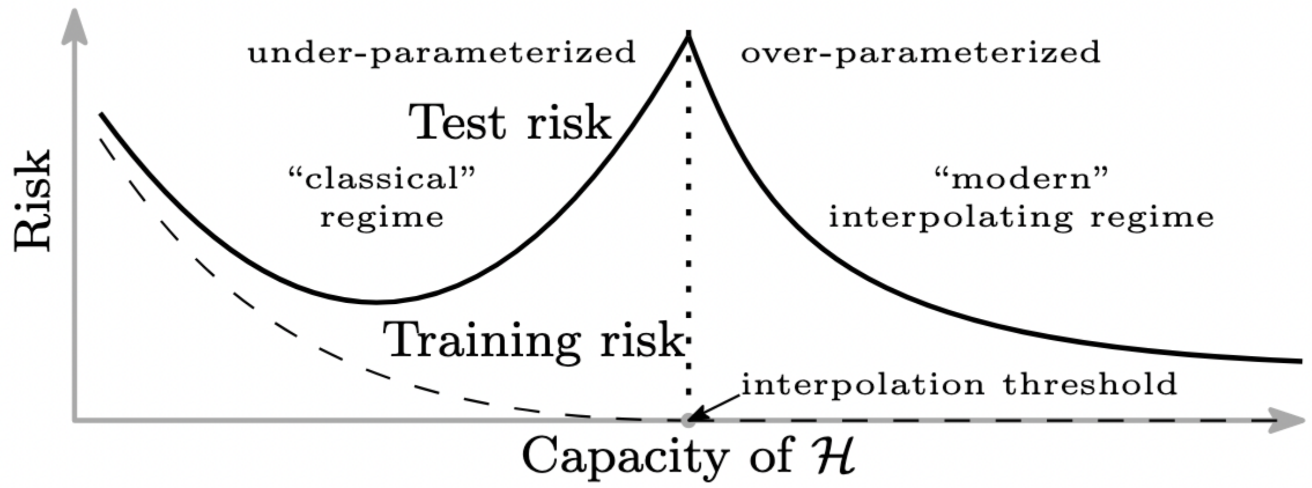
The Overparametrized Regime

- The Excess Risk
  - $R(f_A) - R(f^*) = \boxed{R(f_A) \; - R(f_F^*)} + \boxed{R(f_F^*) - R(f^*)}$

    Estimation error　　Approximation Error
- ERM
  - Recall. $P_S\left[R\left(f_S^{ERM}\right) - \inf_{f \in F} R(f) \leq 2\boldsymbol{R_m}(F) + 2\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}}\right] \geq 1 - \delta$
  - *c.f.* d *−dim linear classifier, cosine kernel classifier*

The Overparametrized Regime

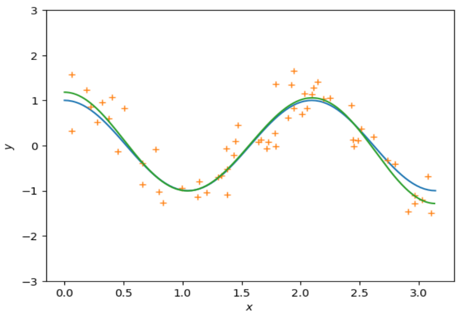
- Traditional Point of View
  - Overparameterization poorly generalize
  - $argmin_{f \in F} R(f)$ vs f_S^{ERM}
- Modern AI
  - Even overparameterized model generalize well



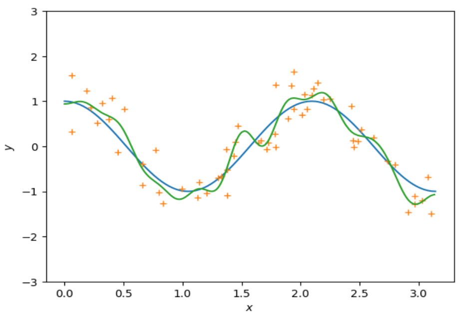
Benign Overfitting in Regression

- Kernel Regression with $\{\frac{\cos(it)}{i}\}_{i=1}^{k}$

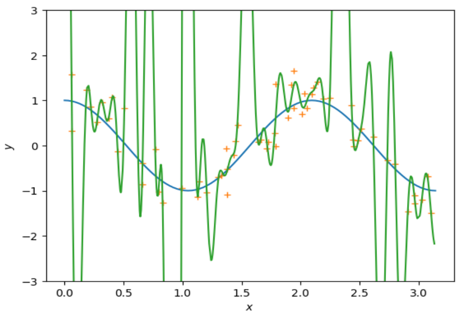
$k = 5$



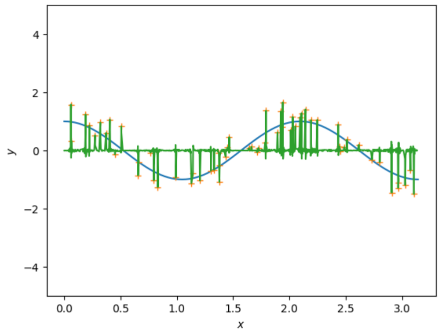
$k = 20$

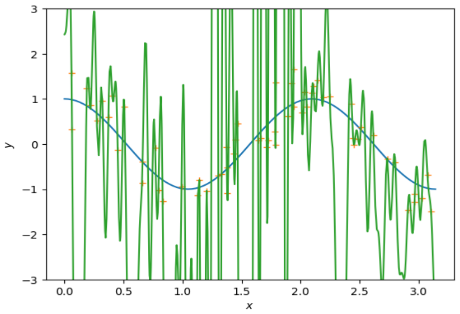


$k = 50$

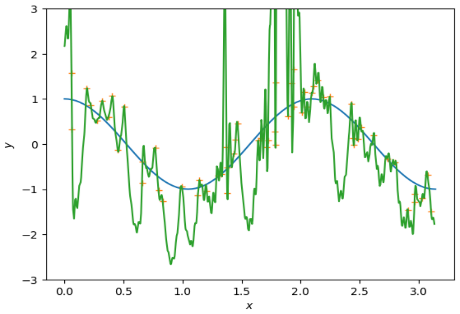


$\{\cos(it)\}_{i=1}^{k}$
$k = 2000$

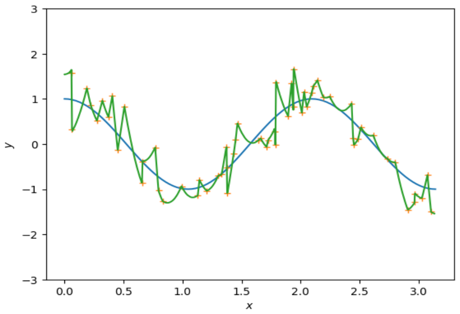




$k = 100$



$k = 200$

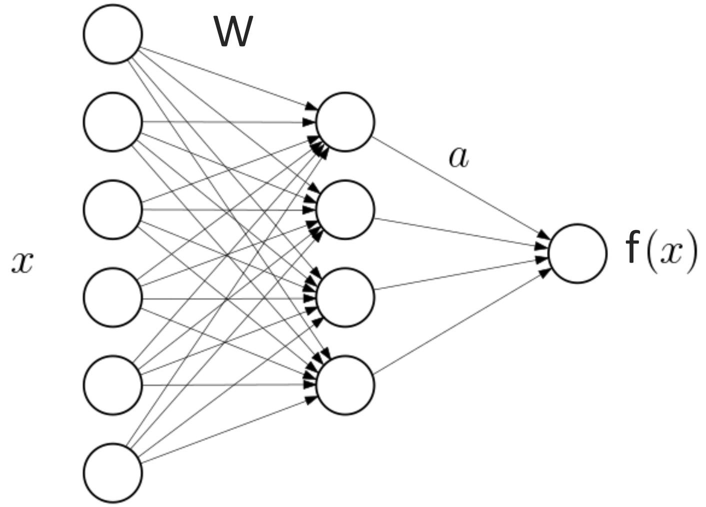


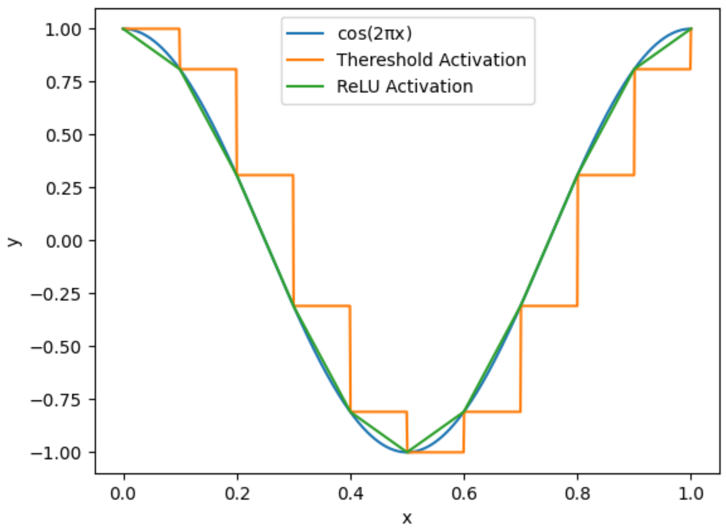
$k = 2000$

The Approximation Theorem

- Representation of 2-Layer NN
- Continuous function h in a compact domain can be approximated by 2-layer ReLU network in $||\cdot||_\infty$

$$f(x; W, a, b) = \sum_i a_i \phi(W_i^T x - b_i)$$

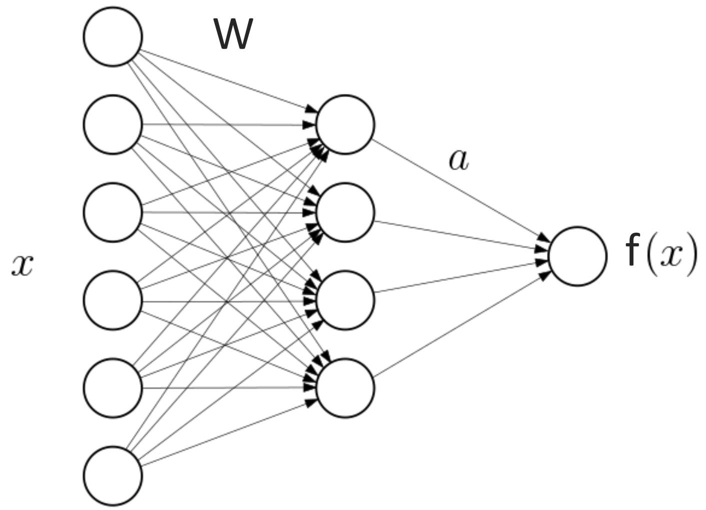


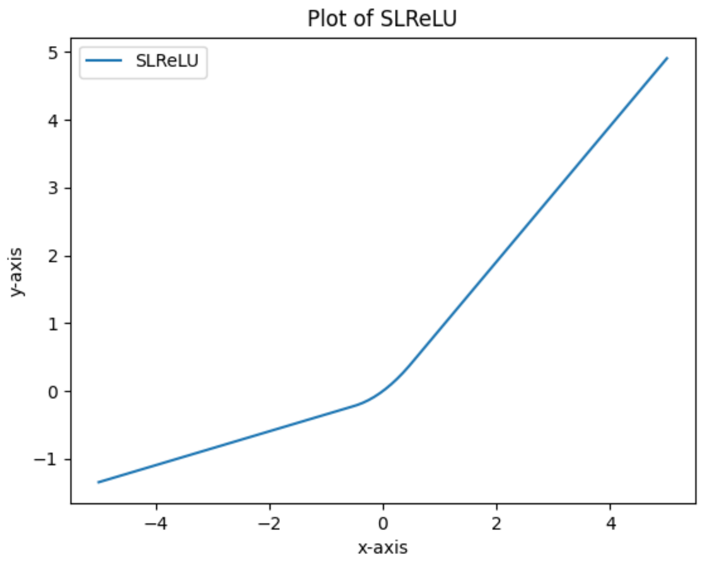


Our Simplified Architecture

- $W \in \mathbb{R}^{m \times p}$ trainable
- $\forall a_i \in \{-1/\sqrt{m}, 1/\sqrt{m}\}$ uniformly random

$$f(x; W, a) = \sum_i a_i \phi(W_i^T x)$$





Meaning of the Main Theorem

Theoretical Guarantee of Benign Overfitting

Theorem 3.1

- When a **neural network** is trained under certain assumptions, it will exhibit **benign overfitting**.

  1. **Interpolation**: Achieves arbitrary small **training loss**, $\widehat{L}(W) < \epsilon$
  2. **Generalization**: Achieves test error close to the **noise rate**,

$$\mathbb{P}_{(x,y)\sim \mathrm{P}}\left[y \neq \operatorname{sgn}\left(f\left(x; W^{(T)}\right)\right)\right] \leq \eta + 2\exp\left(-\frac{n\|\mu\|^4}{Cp}\right)$$

Assumptions on the **Training Objectives**

Trained with Full-batch Gradient Descent on Empirical Loss, $\hat{L}(W)$

$$W^{(t+1)} = W^{(t)} - \alpha \Delta \hat{L}(W^{(t)})$$

Empirical Loss, $\hat{L}(W)$

$$\hat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} l\left(y_i f(x_i; W)\right)$$

$$where\ l(z) = \log(1 + \exp(-z))$$

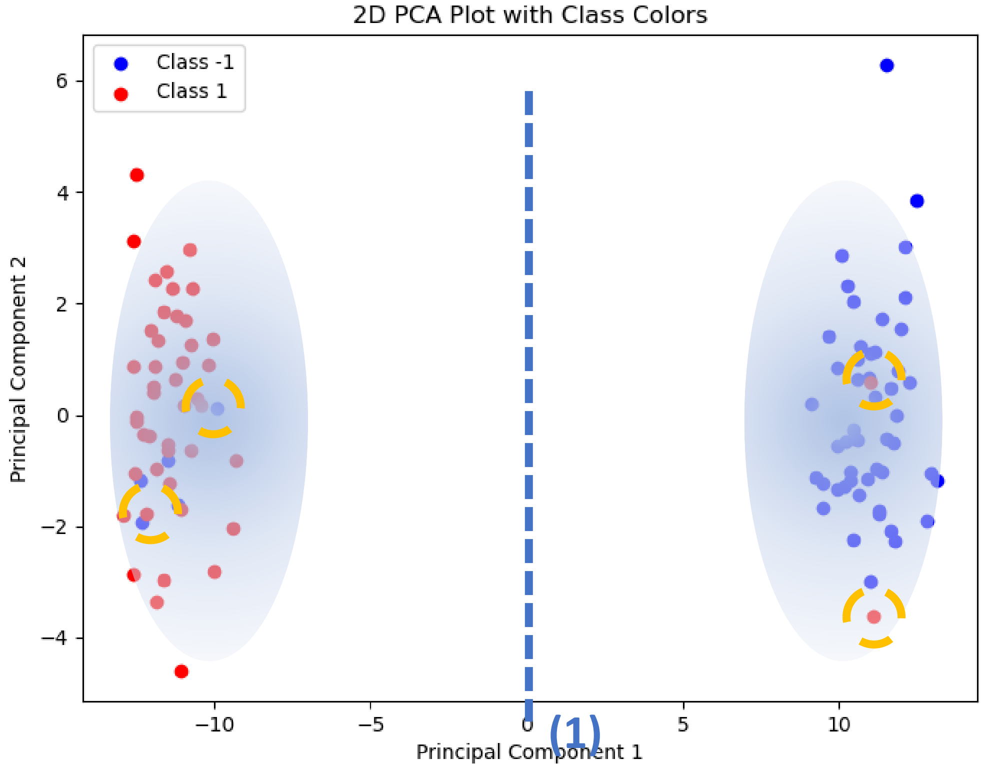
Margin, $y_i f(x_i; W)$

(+) when $y_i = \operatorname{sgn}\left(f(x_i; W)\right)$

(-) when $y_i = -\operatorname{sgn}\left(f(x_i; W)\right)$

Assumptions on the **Generated Dataset**

- A joint distribution P over $(x, y) \in \mathbb{R}^p \times \{\pm 1\}$, where p is very **large**.



(1) Linearly separable gaussian distribution dataset,

(2) Invert labels with probability η

Assumptions on the **Parameters**

- Six Assumptions (A1)-(A6)

(A1) Number of samples $n \geq C \log(1/\delta)$.

(A2) Dimension $p \geq C \max\{n\|\mu\|^2, n^2 \log(n/\delta)\}$.

(A3) Norm of the mean satisfies $\|\mu\|^2 \geq C \log(n/\delta)$.

From Chatterji and Long [CL21b],
which studied the same sample model setup as here
analyzing maximum margin linear classifier

(A4) Noise rate $\eta \leq 1/C$.

(A5) Step-size $\alpha \leq \left(C \max\left\{1, \frac{H}{\sqrt{m}}\right\} p^2\right)^{-1}$, where ϕ is H-smooth.

(A6) Initialization variance satisfies $\omega_{\text{init}}\sqrt{mp} \leq \alpha$.

Significance of Theorem 3.1

- Largely **generalizes** the condition when “benign overfitting” occurs
  1. Using **richer class**, the two-layered classification NN.
     - (Chatterji and Long [CL21b] proved for maximum margin linear classifier)
  2. Using **loose assumptions** compared to other theoretical analyses of NNs.
     - Allows networks of an arbitrary width, m, (m hidden neurons)
     - Arbitrary small initialization variance, ω_{init}
     - Arbitrary long training time, T
- Does not require the Neural Tangent Kernel approximation (NTK)
  - NTK is conducted in the infinite width limit
  - NTK fails to capture several aspects of NN, such as **the ability to learn features**

Sketch of a proof of Thm 3.1.

A bit of Details

Generalization on Noisy Data

CLAIM #1: Trained Network achieves test error close to the noise rate

$$\mathbb{P}_{(x,y)\sim\mathrm{P}}\left[y \neq \operatorname{sgn}\left(f\left(x; W^{(T)}\right)\right)\right] \leq \eta + 2\exp\left(-\frac{n\|\mu\|^4}{Cp}\right)$$

Lemma 4.1

- Establish an upper bound for the test error

  in terms of the **expected normalized margin** on clean points

Lemma 4.1. *Suppose that* $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathrm{P}}}[\tilde{y} f(x; W)] \geq 0$. *Then there exists a universal constant* $c > 0$ *such that*

$$\underbrace{\boxed{\mathbb{P}_{(x,y)\sim\mathrm{P}}\left(y \neq \operatorname{sgn}(f(x; W))\right)}}_{\text{Test error}} \leq \eta + 2\exp\left(-c\lambda \underbrace{\boxed{\left(\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathrm{P}}}[\tilde{y} f(x; W)]}{\|W\|_F}\right)^2}}_{\text{Expected Normalized Margin}}\right).$$

Key technical Lemmas for Proof

Lemma 4.8

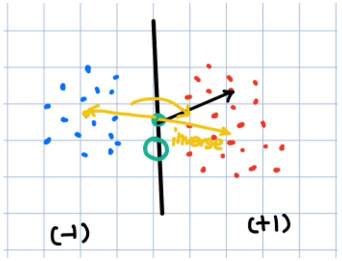
- Lower bound for **the change in unnormalized margin** every update,

$$y\left[f\left(x; W^{(t+1)}\right) - f\left(x; W^{(t)}\right)\right]$$

$$\geq \frac{\alpha}{n}\sum_{i=1}^{n} \boxed{g_i^{(t)}} \left[\xi_i \boxed{\langle y_i x_i, yx\rangle} - \frac{HC_1 p\|x\|^2 \alpha}{2\sqrt{m}}\right]$$

surrogate loss: $\boldsymbol{g_i^{(t)}} \coloneqq -\boldsymbol{l'(y_i f(x_i; W^{(t)})}$

- For a Clean data sample, $(x, \tilde{y}) \sim P$
  - $< y_i x_i, \tilde{y}x >$ is positive, when $(x_i, y_i) \sim \mathcal{C}$
  - $< y_i x_i, \tilde{y}x >$ is negative, when $(x_i, y_i) \sim \mathcal{N}$
- We should guarantee the **g losses to be balanced**



Key technical Lemmas for Proof

Lemma 4.9. (Loss Ratio Bound)

- Ensures that **the noisy points cannot have an outsized influence**

Lemma 4.9. *For a γ-leaky, H-smooth activation ϕ, there is an absolute constant $C_r = 16C_1^2/\gamma^2$ such that on a good run, provided $C > 1$ is sufficiently large, we have for all $t \geq 0$,*

$$\max_{i,j \in [n]} \frac{g_i^{(t)}}{g_j^{(t)}} \leq C_r.$$

Lemma 4.11. (Lower Bound on the Expected Normalized Margin)

Lemma 4.11. *For a γ-leaky, H-smooth activation ϕ, and for all $C > 1$ sufficiently large, on a good run, for any $t \geq 1$,*

$$\boxed{\frac{\mathbb{E}_{(x,\tilde{y}) \sim \tilde{\mathsf{P}}}[\tilde{y} f(x; W^{(t)})]}{\|W^{(t)}\|_F}} \geq \frac{\gamma^2 \|\mu\|^2 \sqrt{n}}{8 \max(\sqrt{C_1}, C_2)\sqrt{p}},$$

Interpolation on Noisy training dataset

CLAIM #2: Trained Network achieves arbitrary small training loss, when $T \geq C\hat{L}(W^{(0)})/\|\mu\|^2\alpha\epsilon^2$ holds.

$$\hat{L}(W) < \epsilon$$

Lemma 4.12. (Upper Bound on the Empirical Loss respect to T)

Lemma 4.12. *For a γ-leaky, H-smooth activation ϕ, provided $C > 1$ is sufficiently large, then on a good run we have for all $t \geq 0$,*

$$\|\nabla \hat{L}(W^{(t)})\|_F$$

(1) Upper bound the Surrogate Loss, $\hat{G}(W^{(T)})$ in terms of T

Moreover, any $T \in \mathbb{N}$,

$$\frac{1}{n}\sum_{i=1}^{n} 1(y_i \neq \mathrm{sgn}(f(x_i; W^{(T-1)}))) \boxed{\leq 2\hat{G}(W^{(T-1)}) \leq 2\left(\frac{32\hat{L}(W^{(0)})}{\gamma^2\|\mu\|^2\alpha T}\right)^{1/2}}$$

In particular, for $T \geq 128\hat{L}(W^{(0)})/\left(\gamma^2\|\mu\|^2\alpha\varepsilon^2\right)$,

(2) Using the condition on T, bound $\hat{G}(W^{(T)})$ in terms of ϵ

Follow-up Research

Using a Different Approach

Idea

1. Gradient flow approximates gradient descent
2. Gradient flow on *logistic loss* converges to KKT point of margin maximization problem
3. Our data generation implies orthogonality
4. Training with orthogonal data,
   KKT point is almost a uniform average of inputs
5. Solution that is almost a uniform average of inputs exhibits
   benign overfitting

1. The Gradient Flow

- Discrete v.s. Continuous Dynamic

Gradient descent w_i

$$W_{i+1} = W_i - \alpha \nabla L(W_i)$$

Gradient flow $w(t)$

$$\frac{dW}{dt} = -\nabla L\big(W(t)\big)$$

- Unification via a theorem

$$\sup_{t \in [0,T]} \left| W(t) - W_{\left\lfloor \frac{t}{\alpha} \right\rfloor} \right| \to 0 \text{ as } \alpha \to 0$$

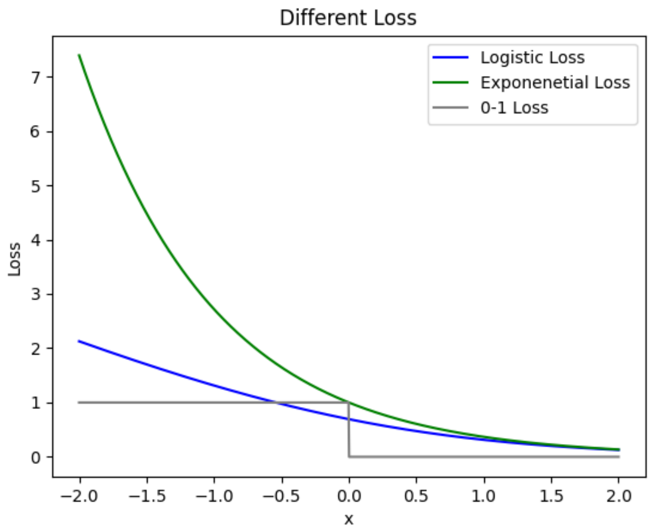
2. Gradient Flow on Logistic Loss

[Ji and Telgarsky]

$$L(W) = \frac{1}{n}\sum_i l(y_i f(x_i; W))$$

- $l(q) = \log(1 + \exp(-q))$ or $l(q) = \exp(-q)$
- $L\big(W(0)\big) < \log(2)\,/n$
- W^*: KKT point

  (…of what?)



$$\frac{W(t)}{||W(t)||} \to \frac{W^*}{||W^*||}$$

2. KKT Point of Margin Maximization

- Margin maximization problem

$$\min_{W \in R^{m \times d}} ||W||_F^2 \text{ such that } y_i f(x_i; W) \geq 1$$

  - c.f. Max margin SVM

  Gradient descent converges to max margin SVM

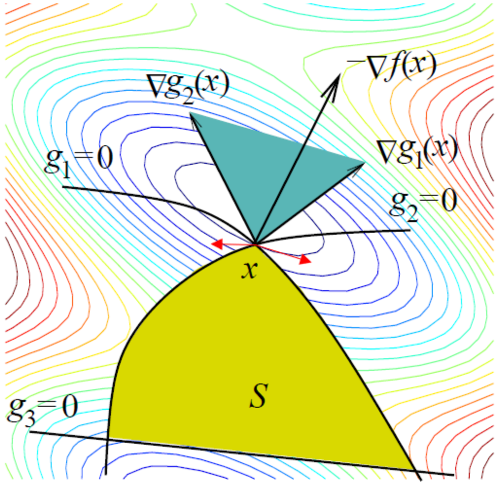


Figure Reference

3. Orthogonality and Uniformity

- Set of data n points are p-orthogonal

$$R_{min}^2 \geq pR^2 n \max_{i \neq j} |\langle x_i, x_j \rangle|$$

  - $R_{min}^2 = \min||x_i||^2, R_{max}^2 = \max||x_i||^2, R^2 = R_{max}^2 / R_{min}^2$
  - Depends on $\{(x_i, y_i)\}$

- $w \in R^d$ is τ-uniform if

$$w = \sum s_i y_i x_i$$

  - $\{s_i\}_{i=1}^n$ positive and $\frac{\max s_i}{\min s_i} \leq \tau$
  - But we don't have w…?

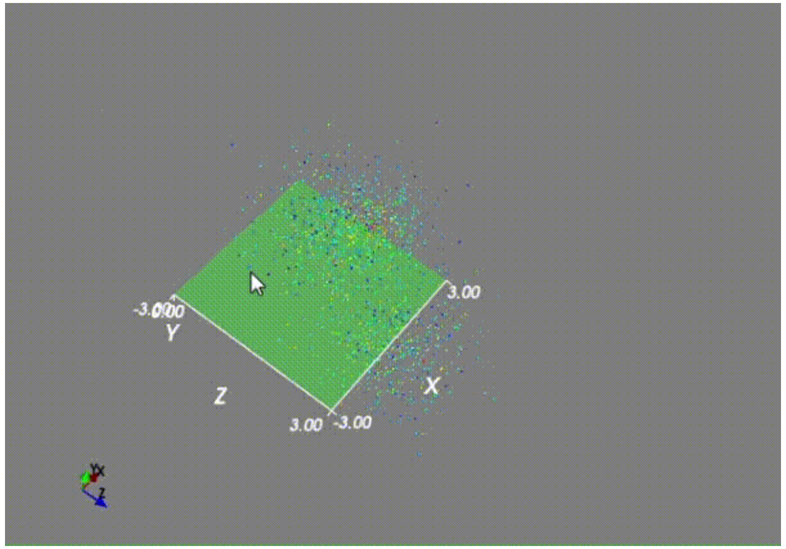
3. Data Generation implies Orthogonality

- NN Architecture
  - *Leaky ReLU* activation

- Data Generation
  - From k clusters with mean μ^i for $i = 1, \cdots, k$
  - Cluster means are nearly orthogonal

$$\min_i ||\mu||^2 \geq Ck \max_{i \neq j} |\langle \mu^i, \mu^j \rangle|$$

  - Each cluster assigned for label $\{\pm 1\}$
  - Implies p-orthogonality *with high probability*



$k = 3$

4. Orthogonality implies Uniformity

- Orthogonality implies uniformity
- If W^* is KKT point of margin maximization,

$\exists w$ such that

$$sign\left(f(\cdot; W^*)\right) = sign(\langle w, \cdot \rangle)$$

and w is τ-uniform

$$\tau = \frac{R^2}{\gamma^2}\left(1 + \frac{2}{\gamma p R^2 - 2}\right)$$

5. Theorem

- *With high probability,* τ-uniform $w \in R^d$

$$y_k = sign(\langle w, x_k \rangle) \text{ for all } k \in [n]$$

$$\eta \leq P_{(x,y)}[y \neq sign(\langle w, x \rangle)] \leq \eta + \exp(-\frac{n \min \left|\left|\mu^i\right|\right|^4}{C' k^2 d})$$

- **Benign overfitting** if $n \min \left|\left|\mu^i\right|\right|^4 = \omega(k^2 d)$

Corollary

- KKT point W^* of the Margin maximization problem satisfies followings with high probability

$$y_k = sign(f(x_k; W^*)) \text{ for all } k \in [n]$$

$$\eta \leq P_{(x,y)}[y \neq sign(f(x; W^*))] \leq \eta + \exp(-\frac{n \min \left|\left|\mu^i\right|\right|^4}{C' k^2 d})$$

Recall. W_i $W(t)$ W^*

Recap!

1. Gradient descent
2. Gradient flow
3. KKT point of margin maximization problem
4. p-orthogonality
5. Uniform average
6. **benign overfitting**

Comparison to Previous Work

	FCB 22	FVBS 23
Analysis	Discrete	Continuous
Data Generation	Negative two Cluster	Orthogonal k cluster
Linearly Separable Assumption	Implicit(w.h.p.)	Implicit(w.h.p.)
Architecture	Smoothed Leaky ReLU	Leaky ReLU

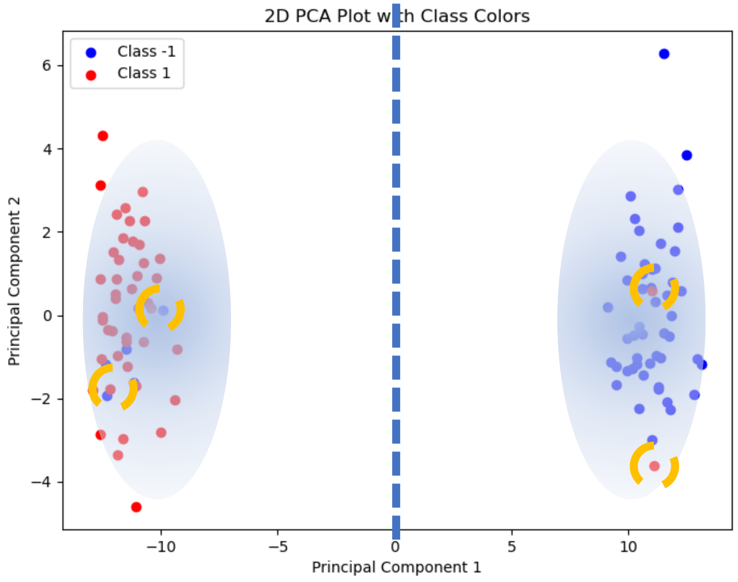
Empirical Analysis

So, does it really happen?

Generating Samples from Data Distribution

- Generated $N = 100$ samples using Gaussian distribution

Example 2.1. *If* $\mathsf{P}_{\mathsf{clust}} = \mathsf{N}(0, \Sigma)$*, where* $\|\Sigma\|_2 \leq 1$ *and* $\|\Sigma^{-1}\| \leq 1/\kappa$*, and each of the labels are flipped independently with probability* η*, then all the properties listed above are satisfied.*

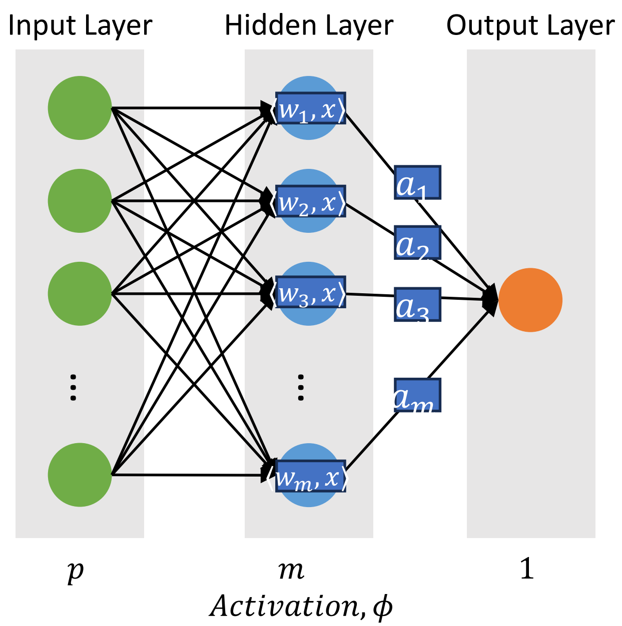


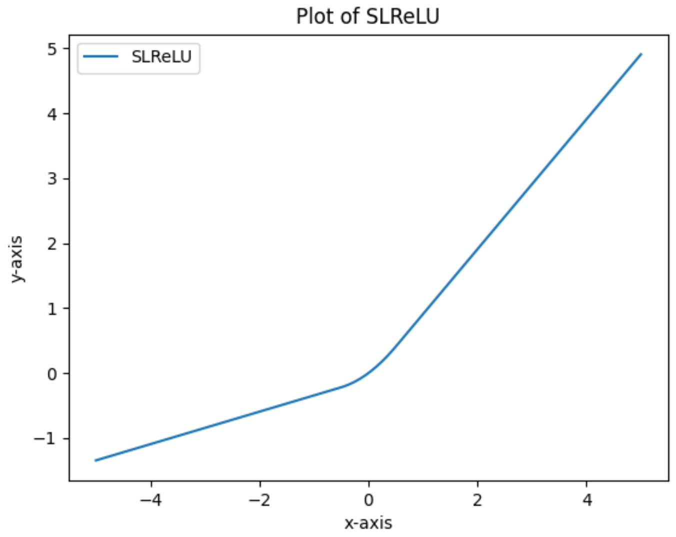
Model Training

- Full-batch Gradient Descent

$$W^{(t+1)} = W^{(t)} - \alpha \Delta \hat{L}(W^{(t)})$$

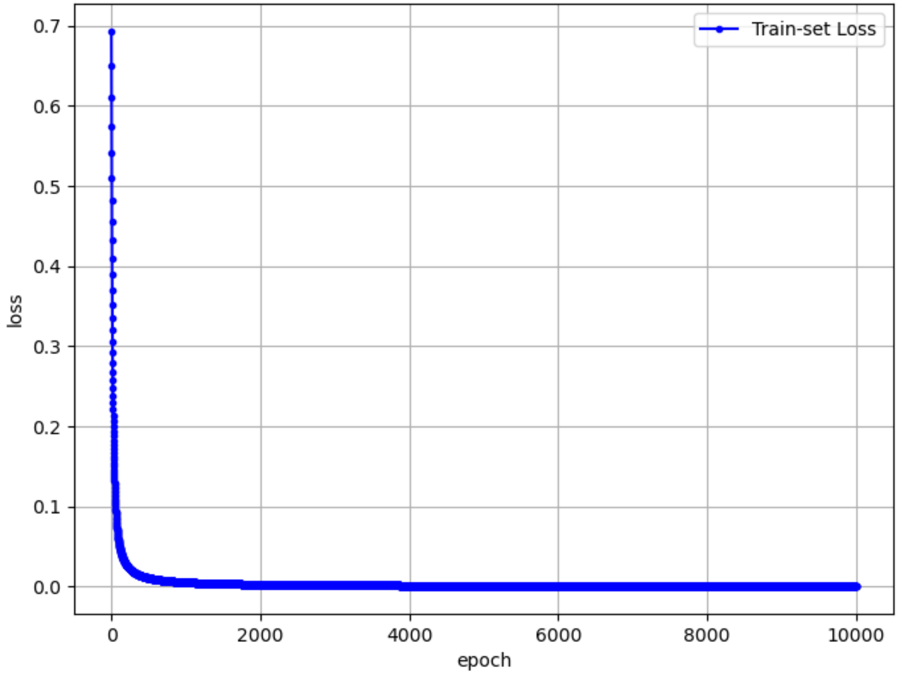
  - With Epoch = 10000, **α(learning_rate) = 0.001**
    - α violates the assumption (A5), but theoretical upper bound for α is too small
    - Furthermore, the epoch is bounded in terms of α, so it violates the assumption too.





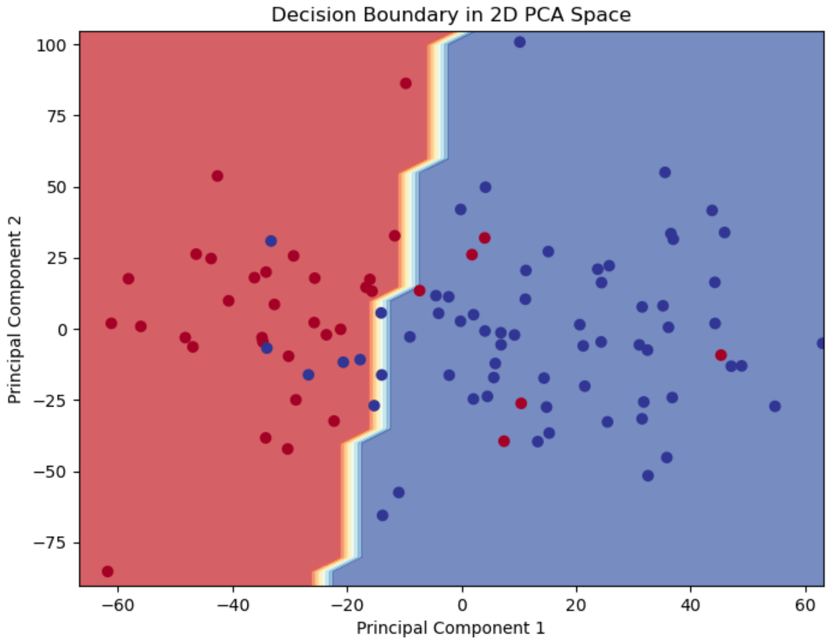
It really works!

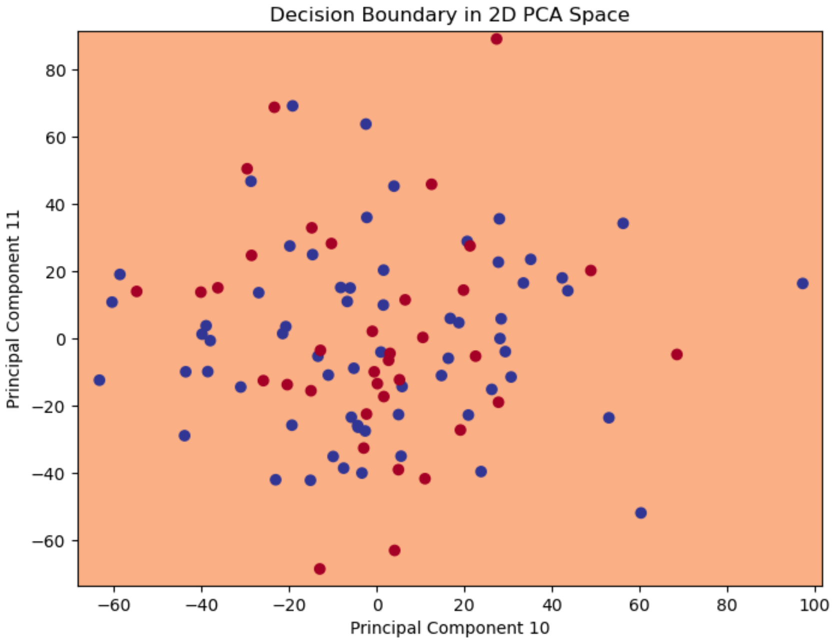
- Convergence of Train loss occurs
- Test error is 0



Does it generalize well?

- Decision Boundary of hyperplane
  - Set $||\mu||^2 = 128.0, p = 80000, \eta = 0.1$
  - Boundary is **perpendicular** to the first principal component direction
  - Boundary is almost **parallel** to the (10, 11)-principal hyperplane.



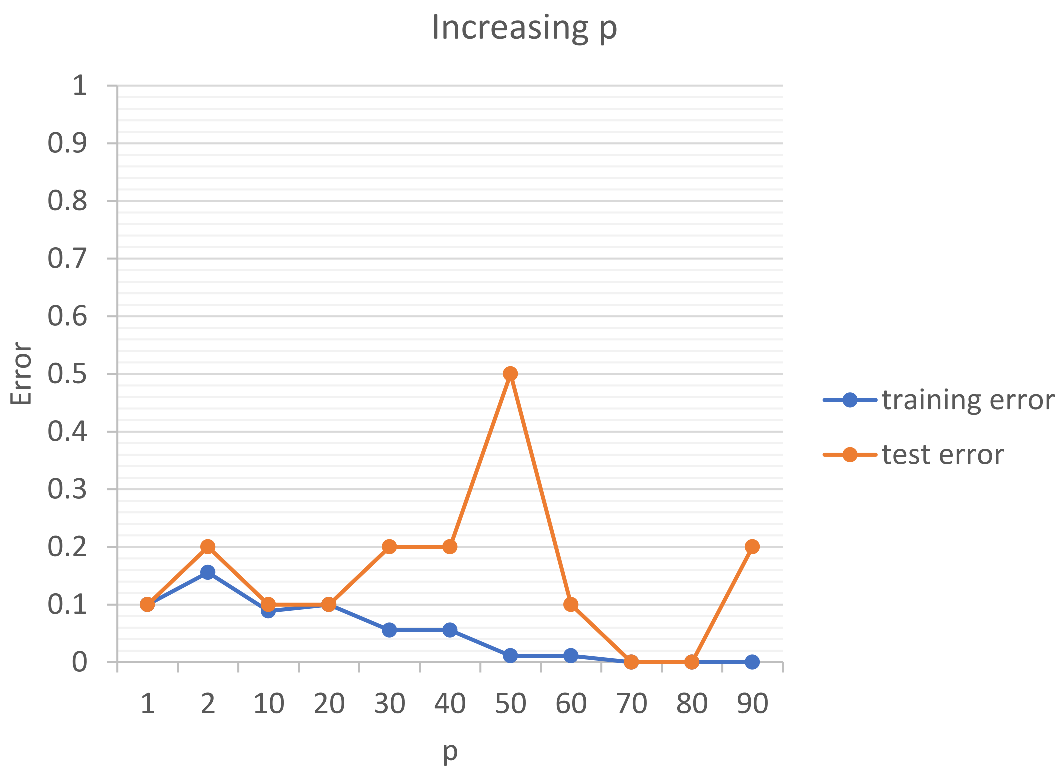


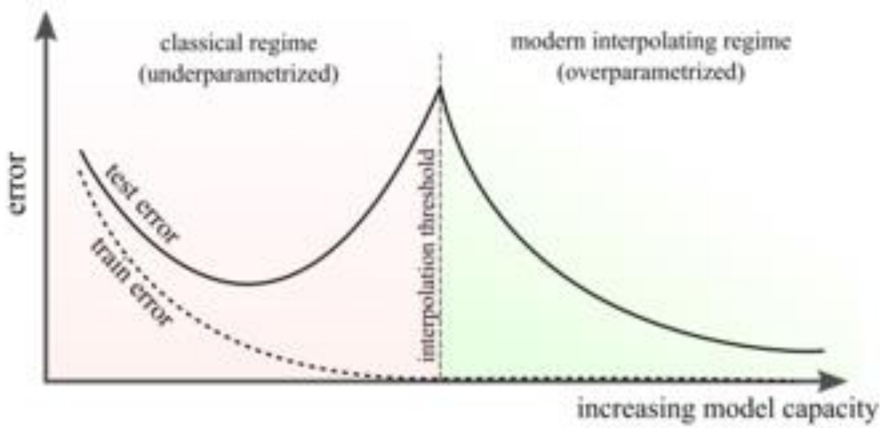
Is high p required?

- Measure the training loss and the accuracy by changing p.
  - when $n = 100, m = 64, \eta = 0.1,\ \|\mu\|^2 = 128.0$, 100000 Epoch
- Throughout the whole paper $p \geq n$ was required. → Is it required?
- We've successfully reproduced the heuristic result of the train/test error tendency respect to model capacity

p	10	50	80	80000
Train Loss	0.281	0.251	0.165	5.178e-04
Train Accuracy	0.911	0.867	0.922	1.0
Test Accuracy	0.9	0.7	1.0	1.0

Table 1: Growing p



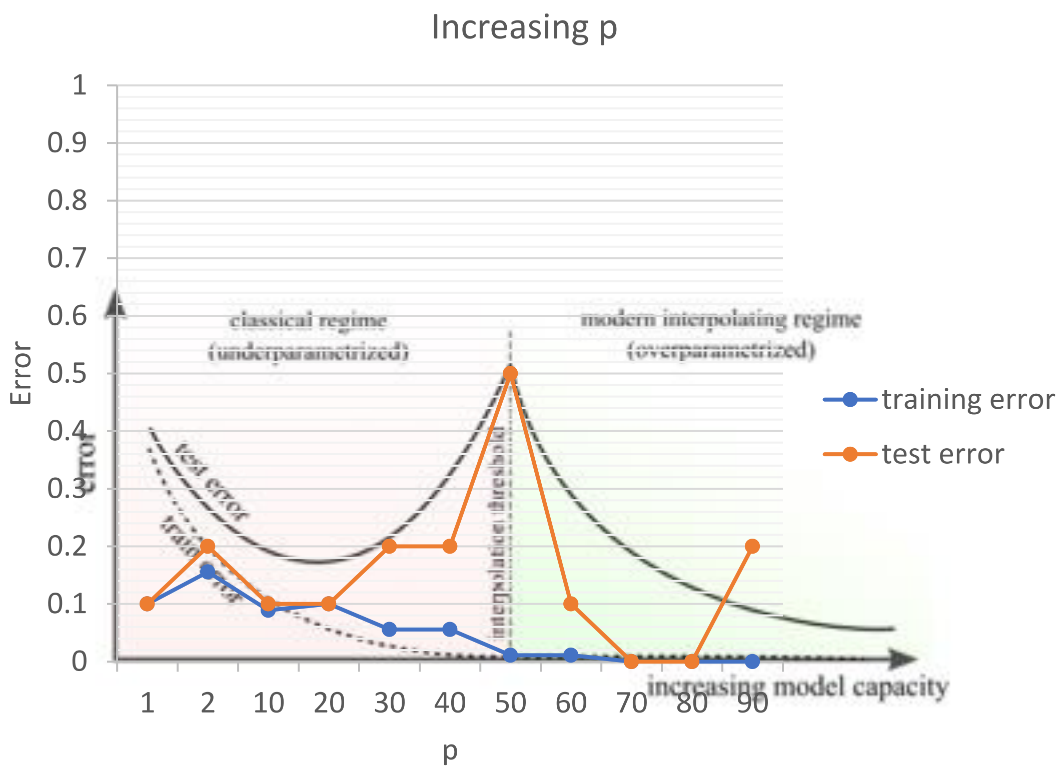


Is high p required?

- Measure the training loss and the accuracy by changing p.
  - when $n = 100, m = 64, \eta = 0.1,\ ||\mu||^2 = 128.0$, 100000 Epoch
- Throughout the whole paper $p \geq n$ was required. → Is it required?
- We've successfully reproduced the heuristic result of the train/test error tendency respect to model capacity

p	10	50	80	80000
Train Loss	0.281	0.251	0.165	5.178e-04
Train Accuracy	0.911	0.867	0.922	1.0
Test Accuracy	0.9	0.7	1.0	1.0

Table 1: Growing p



How tight is the test error bound?

- Set $\eta = 0.1$
- As $||\mu||^2$ changes, the test error bound gets tighter and the real error is bounded

$\lVert\mu\rVert^2$	8.0	16.0	32.0	64.0	128.0	256.0
η	0.1	0.1	0.1	0.1	0.1	0.1
$2\exp\left(-\frac{n\lVert\mu\rVert^4}{Cp}\right)$	1.929	1.732	1.124	0.200	1.988E-4	1.955E-16
Test Error Bound	2.029	1.832	1.224	0.3	0.1	0.1
Real Test Error	0.4	0.2	0.1	0.1	0	0

Table 2: Growing $||\mu||^2$

Question and Discussion

Thank you☺