CS540 Course Project: Benign Overfitting in Neural Networks

Shivram Gowtham Yashaswini Murthy Nikhil Sheoran

1. Introduction

Benign overfitting occurs when a model is trained such that the empirical risk is vanishingly small, while obtaining minimal generalization error (BLLT20; Bel21; MSA+22). This is in contrast with classical machine learning results, which suggest overfitting models do not generalize well. Even when a certain fraction of the labels are flipped, the training error can be made arbitrarily small with good generalization.

There have been various works on benign overfitting in recent literature. The most fleshed out formulation on benign overfitting can be found in (BLLT20), where the authors show that in high dimensions, a wide array of datasets exhibits a large rank for a submatrix corresponding to low eigenvalues. They further show that this abundance of low-variance directions in the data causes overfitting.

This set off the study of benign overfitting in many other contexts. (Sha22) characterized the greater brittleness exhibited by linear regression in comparison to classification tasks. Due to asymptotic inconsistency of min norm estimators trained using square loss, linear regression overfitting generally holds true in a very narrow range of data properties, which is observed to not be the case with classification. (Sha22) studies benign overfitting in classification tasks using hinge loss, even though (MNS+20) have previously shown that optimal min norm estimators converge to the max margin estimator using square loss in the high dimensional overparametrization regime.

Another result by (HMX20) indicates that in high dimensions, all training samples behave as SVM vectors. This brings forth an important perspective different from generalization results in low dimensions (which relies on the ratio of SVM vectors to number of samples vanishing as the number of samples tend to infinity). (HMX20) provide sufficient and weakly necessary conditions for each sample to behave as an SVM vector in high dimensions. Although (MNS+20) identify regimes where such overfitting generalizes well, its counterpart in neural networks is unclear.

(CL20) consider logistic regression and provide finite sample bounds for benign overfitting in the presence of misclassification noise. (CLB21) consider benign overfitting in the context of linear neural networks along similar lines as (BLLT20) and (AMN+21). (CL22) show the deep linear neural netowrks generalise similar to their shallow counterparts. As far as we are aware, in the context of nonlinear neural networks, (FCB22)'s result on benign overfitting in the high dimensional, non-NTK regime with well separable data and logistic loss function is one of the very firsts. Their analysis is inspired by (CL20) and is the focus of our report.

2. Benign Overfitting in Shallow NNs

Sample Generation: Sample clean label $\tilde{y} \sim \mathsf{Uniform}(\{+1,-1\})$. Sample $z \sim \mathsf{N}(0,\Sigma)$, where $\|\Sigma\|_2 \leq 1$ and $\|\Sigma\|_2 \leq 1/\kappa$. Such a distribution can be shown to be $\lambda-$ strongly log-concave distribution over $\mathbb{R}^p$ for some $\lambda > 0$. Generate $x = z + \mu\tilde{y}$. Then introduce noise by flipping each label $\widetilde{y}$ independently with probability η.

Assumptions: Number of samples $n \geq C\log(1/\delta)$, Dimension $p \geq C\max\left\{n\|\mu\|^2, n^2\log(n/\delta)\right\}$, Norm of the mean $\|\mu\|^2 \geq C\log(n/\delta)$, Noise rate $\eta \leq 1/C$, Step size $\alpha \leq \left(C\max\left\{1, \frac{H}{\sqrt{m}}\right\}p^2\right)^{-1}$, where ϕ is H-smooth and Initialization variance satisfies $\omega_{\text{init}}\sqrt{mp} \leq \alpha$.

Theorem 2.1. *For any γ-leaky, H- smooth activation ϕ, for all $\kappa \in (0,1)$, $\lambda > 0$, there is a $C > 1$ such that the assumptions above are satisfied and the following holds.*

For any $0 < \epsilon < 1/2n$, by running gradient descent for $T \geq C\hat{L}(W^{(0)})/(\|\mu\|^2\alpha\epsilon^2)$ iterations, where $\hat{L}(W^{(i)})$ is the empirical loss at the ith step of gradient descent, with probability at least $1-2\delta$ over the random initialization and the draws of the samples, the following holds:

1. *All training points are classified correctly and the training loss satisfies $\hat{L}(W^{(T)}) \leq \epsilon$.*
2. *The test error satisfies:*

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

2.1. Salient features of the result

The proof hinges on the following key observations.

- Data sampling ensures that $\forall i \neq j, k \in [n], |\langle x_i, x_j\rangle| \leq C_1\left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) \leq \|x_k\|^2$. Consequently if $f(x;W)$ is the output of the neural network for input x, then $\forall\, i \neq j, k \in [n], |\langle\nabla f(x_i;W), \nabla f(x_j;W)\rangle| \leq C_1\left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) \leq \|\nabla f(x_k;W)\|^2$. This

implies that the gradients are roughly orthogonal. Consequently, the noisy samples do not affect learning in other directions. This is analogous to the effect of noise being buried in low variance directions in (BLLT20).

- Splitting of generalization error between clean and noisy samples yields the following:

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}(y \neq \operatorname{sgn}(f(x;W))) \leq \eta + 2\exp\left(-c\lambda\left(\mathbb{E}_{(x,\widetilde{y})\sim\widetilde{\mathsf{P}}}[\widetilde{y}f(x;W)]/\|W\|_F\right)^2\right)$$

In order to characterize the generalization with respect to clean samples, consider their margin evolution: there exists a $\xi \in [\gamma^2, 1]$ such that,

$$\widetilde{y}\left[f(x;W^{(t+1)}) - f(x;W^{(t)})\right] \geq \frac{\alpha}{n}\sum_{i=1}^{n} g_i^{(t)}\left[\xi_i\langle y_i x_i, \widetilde{y}x\rangle - \frac{HC_1^2p^2\alpha}{2\sqrt{m}}\right], \quad (1)$$

where $g_i^{(t)} = 1/(1+\exp(y_i f(x_i;W^{(t)})))$.

For good generalization performance it is necessary for this margin to grow every iteration. This is ensured if every term in the summand is positive. However, when a label y_i is noisy, the inner product $\langle y_i x_i, \widetilde{y}x\rangle$, will be negative. If there is a bound on the value of $g_i^{(t)}$, it is possible to still obtain positive increase in margin, as long as the number of noisy samples is proportionately small. Hence, the key contribution of the paper is in bounding this function to obtain an overall improvement in generalization that is, provided $C > 1$ is sufficiently large, there exists an absolute constant $C_r = 16C_1^2/\gamma^2$ such that for all $t \geq 0$,

$$\max_{i,j\in[n]} g_i^{(t)}/g_j^{(t)} \leq C_r. \quad (2)$$

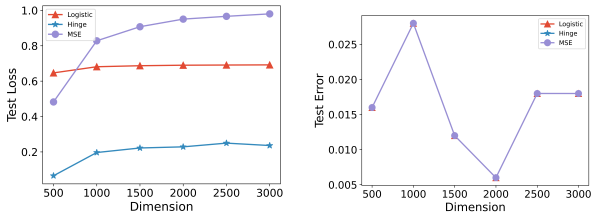
Note that the bound in (2) is also crucial to show that the training error goes to zero. Consequently, the following lower bound on the generalization error for clean data can be obtained:

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

It is also worth noting that these results are independent of neural network width and heavily rely on $p \gg n$ and μ, where μ represents the (large) data separation.

3. Experiments

To better understand the behavior of benign overfitting in shallow neural networks we performed a small set of experiments. The setup of our experiments is similar to the one used by (FCB22). We utilized the code provided by (CL22) and build on top of it to perform these experiments.



(a) Test Loss (b) Test Error (All 3 overlap)

Figure 1. Variation of the loss function. Test loss is the raw value of the loss. Test error is the classification error with $\hat{y} = sgn(f(x))$.

Setup. The data and the labels are generated using the sample generation process described in Section 2. The trained model is a shallow LeakyReLU network with 1 hidden layer i.e. $f(x) = \sum_i a_i * \sigma_r(w^T x)$ where $\sigma_r = \max(0,x) + 0.01 * \min(0,x)$. The weights of the final layer (i.e. a_i) are fixed at initialization whereas the weights of the hidden layer are trained.[1] We experimented with different loss functions namely Logistic Loss, MSE Loss and Hinge Loss. Figure 1 shows the result for the different loss functions. A value of $C = 10, n = 100, \|\mu\| = 8$, $\delta = 0.05$, $\eta = 0.02$, $m = 50$ was used for this experiment. Surprisingly, all the three loss functions — although the behavior of losses across dimensions was slightly different — had the same test errors (on the given dataset). This is counter-intuitive, but one possible explanation could be the low value of n (100) and the significantly low noise ($\eta = 0.02$) when evaluating the test error, thus each loss function seemingly performing well.

We also tried experiments with varying values of $C \in \{10, 25, 50\}$, keeping other parameters fixed, affecting $\|\mu\|$ — but observed similar behavior. We will further investigate these experiments to reason about this behavior.

4. Future Work and Open Questions

An interesting direction to pursue would be to check for SVM vector proliferation with increasing dimensions in neural networks as previously shown in (HMX20). Since (MNS+20) characterize generalization performance in different regimes for SVM, it would be interesting to check for similar parallels in the context of nonlinear neural networks. Another interesting line of future work is to consider such a setting in the context non-logistic loss. Since the boundedness of derivative ratios for the logistic loss is incredibly crucial to the proof, it would be interesting to determine analogous conditions for losses such as square loss (which have been shown to also induce overfitting in binary classification). Since min norm interpolation using square loss has been shown to converge to max margin predictor, it would be interesting to compare the square-loss analysis with the margin maximising analysis in an asymptotic regime.

[1] Code at https://github.com/nikhil96sher/cs540-project

References

[AMN+21] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. *CoRR*, abs/2102.09769, 2021.

[Bel21] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation, 2021.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gá bor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, apr 2020.

[CL20] Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime, 2020.

[CL22] Niladri S. Chatterji and Philip M. Long. Deep linear networks can benignly overfit when shallow ones do, 2022.

[CLB21] Niladri S. Chatterji, Philip M. Long, and Peter L. Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks, 2021.

[FCB22] Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data, 2022.

[HMX20] Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions, 2020.

[MNS+20] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? 2020.

[MSA+22] Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting, 2022.

[Sha22] Ohad Shamir. The implicit bias of benign overfitting. *CoRR*, abs/2201.11489, 2022.