CS 165: Foundations of Machine Learning **Winter 2019**

Lecture 11: 02/19/2019

Lecturer: Anima Anandkumar *Scribes: Sophie Dai, Eshan Govil*

11.1 Review

Last time, we covered error decomposition, bias-variance decomposition, and generalization bounds. We can classify errors as approximation error, estimation error, and optimization error. Approximation error is caused by limitations of the function class that is chosen, causing the model to not be able to perfectly approximate the function. In this case, even if we draw an infinite sample IID, we would always have some baseline approximation error. Estimation error is error, or excess risk, caused by fitting to a finite training sample size. We aim to minimize empirical risk with our choice of loss functions. Total error is the sum of approximation and estimation error, and we can write this error also in terms of bias (which comes from approximation error) and variance (which comes from estimation error).

Optimization error is due to the optimization algorithm used and is the difference between the local and global optimum. This can be difficult to characterize, because we are unsure about how much worse the local optimum is than the global difference, and we don't know what the difference between the two is in most cases. In theory, this error should blow up when due to non-convexity of the optimization landscape. However, in practice optimization error often goes to 0 even for non-convex problems, showing a gap between theory and practice. There is a resurgence of new theory in this area and a lot of potential for research, so we will look to build new assumptions and intuitions around this topic.

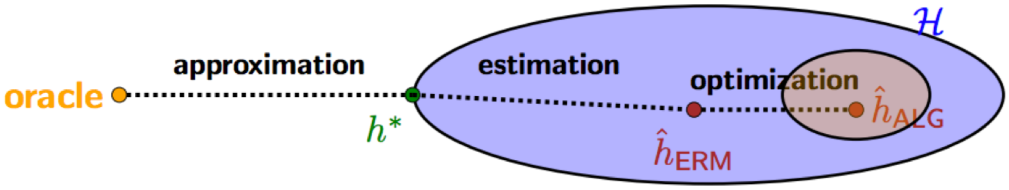


Figure 11.1: Decomposition of Errors

11.2 Classical Theory

Classical theory states that a good class of models should have a balance between bias and variance, or a balance between approximation error and estimation error [Figure 11.2]. In models with too few parameters and limited variance (ie. linear model), there is high approximation error but low estimation error, leading to underfitting. In complex models with too many parameters, we see low approximation error but high estimation error. This can be seen through the Radamacher/VC bounds of the model.

Over-parameterized regimes typically have more parameters in the model than number of samples. Classical theory suggests that high capacity models don't generalize well, and that we should never perfectly fit training data (and get a training error of 0) since optimizing to high precision also harms generalization and can lead to high out-of-sample error. To overcome these issues, the classical regime advises allowing some in-sample error, and using explicit methods of regularization and early stopping when implementing complex models as these also have sharper theoretical bounds. Additionally, conventional wisdom holds that non-convex optimization is hard.

However, we will see that none of these principles do not hold in practice. We don't want something that seems to be working but we don't understand why, so it's necessary to develop some new theoretical frameworks for neural networks that we can use to explain the findings and make deep learning results replicable and analyzable.

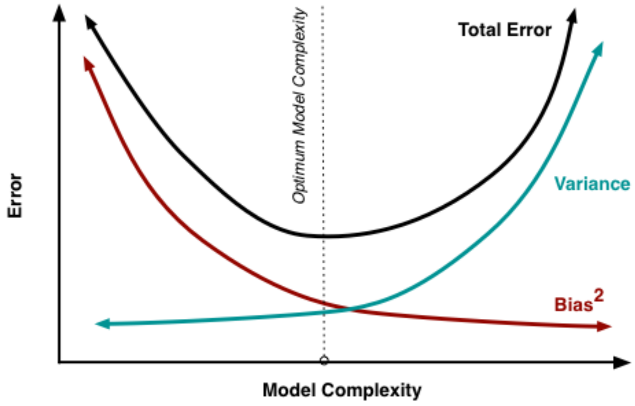


Figure 11.2: Diagram of Bias-Variance Tradeoff

11.3 Modern Neural Network Theory

11.3.1 Deep Neural Networks on CIFAR10

CIFAR10 is a simple beginner dataset for empirical researchers to establish results before moving to ImageNet. In a study comparing the test error of different models on CIFAR10, following results were observed:

Model	p/n	Train Loss	Test Error
Cuda ConvNet	2.9	0	23%
Cuda ConvNet (w/ regularization)	2.9	0.34	18%
MicroInception	33	0	14%
ResNet	48	0	13%

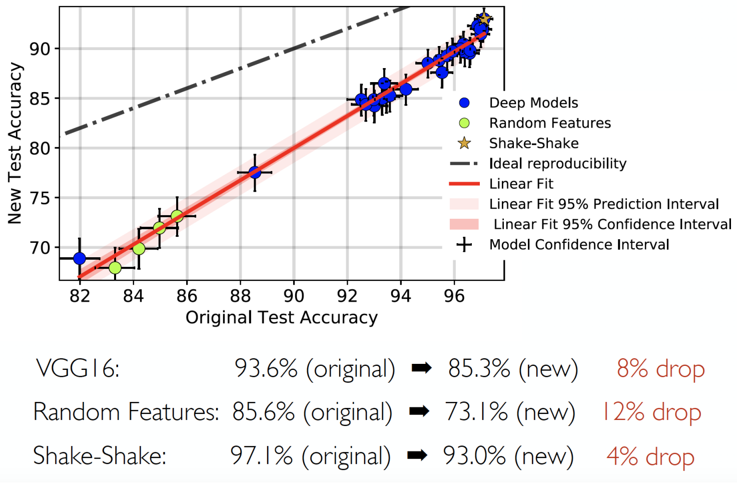
Table 11.1: Table of various models training loss and test error on the CIFAR10 dataset. p/n is the ratio of model capacity to number of data points, or the level of over-parameterization.

Classical wisdom says that p/n should be less than 1; as it goes higher, generalization worsens. We can see that the unregularized deep neural networks MicroInception and ResNet have a significantly larger number of parameters compared to the amount of training data and achieve 0 training loss, so by classical theory, they should have very high test error due to high variance. However, this is not the case. As the ratios increase down the table, test accuracy strictly improves for the bigger deep networks. In most of these networks, explicit regularization is also turned off. This contradicts theory and raises the question: is this overfitting? Why is test error not increasing as in the bias-variance U-curve? The results suggest that the bigger the model, the better we will do, so we need explanations for this phenomenon. With our metrics of success, things seem to be improving over the years without any downside (ignoring the increase compute needed for bigger models).

11.3.2 A Closer Look at CIFAR10

First thing we want to verify is if this test set is truly a test set, as classical theory assumes that training and test sets are IID. Who is giving us the IID samples, and is this set a truly independent sample from the same distribution? A study by Ben Recht showed that in the CIFAR10 dataset, 8% of the test set has a near duplicate in the training set, which may cause some bias in the measured test error. So, it is plausible that the networks were overfitting to CIFAR10 but still getting low test error due to the overlap between the training and testing sets. This represents a broader issue in literature where many researchers will cheat and fit hyperparameters to the test set, so it's critical to carefully read papers and challenge overly strong results.

As a solution, Ben Recht's group created a new "honest" test set for CIFAR10 by collecting new IID subsamples from the larger Tiny Images dataset (which CIFAR10 is a subset of), thereby eliminating near duplicates. What they found when testing the models on the new dataset was that all models perform worse, but the drop in performance from to the line of ideal reproducibility was not uniform: those with previously lower test accuracy dropped more, so the relative ranking of models was mostly preserved under the new test set [Figure 11.3]. Kernel methods had larger drops, whie Shake-Shake had the smallest. Thus, larger networks which were achieving higher accuracy on both the old and new test set are doing better and still not overfitting. (Recht et al., 2018) This is again contradictory to classical theory.



VGG16: 93.6% (original) ➡ 85.3% (new) 8% drop

Random Features: 85.6% (original) ➡ 73.1% (new) 12% drop

Shake-Shake: 97.1% (original) ➡ 93.0% (new) 4% drop

Figure 11.3: Results from Ben Recht's study with CIFAR10. We can see that models achieving higher original test accuracy (deep models) also had higher accuracy on the new test.

11.3.3 A Closer Look at ImageNet

The next step was to look at test distribution and results on ImageNet, a much larger dataset with 1000 imbalanced classes. Even though this dataset is the standard, there are issues with it, and in another study by Ben Recht, the effect of annotation bias in test data on test error of these models was observed with this dataset (Recht et al., 2019). In creating new subsamples of the same distribution as ImageNet, annotators were asked to go through a random set of images from the web and identify objects in the images using ImageNet labels. In this example, they were asked if an image contains a ”bow.” The annotation process creates subtle biases in the data due to annotator noise. For example, annotators are more likely to label images that are very obviously a bow in order to get annotations done quickly for pay, leading to a bias towards easy-to-recognize objects.

To analyze this, Ben's group created different test sets consisting of images based on a threshold of percentage of annotators who labeled that image (ie. test set of top images where all annotators labeled, where 70% and above of annotators labeled, and all annotated images). It was found that models performed better on images that 100% of annotators labeled (the most obvious images), and as the threshold lowered (harder examples), all models had lower accuracy because there aren't enough training instances of such images [Figure 11.4]. However, as with the CIFAR10 tests, the relative ranking of models was preserved, so deeper networks were still achieving relatively high accuracy. Clearly we haven't successfully generalized on ImageNet yet, but again there is no indication that overfitting is occurring. In conclusion, visual question-answering is hard and leads to all sorts of weird biases, so it's necessary to be careful about the datasets.

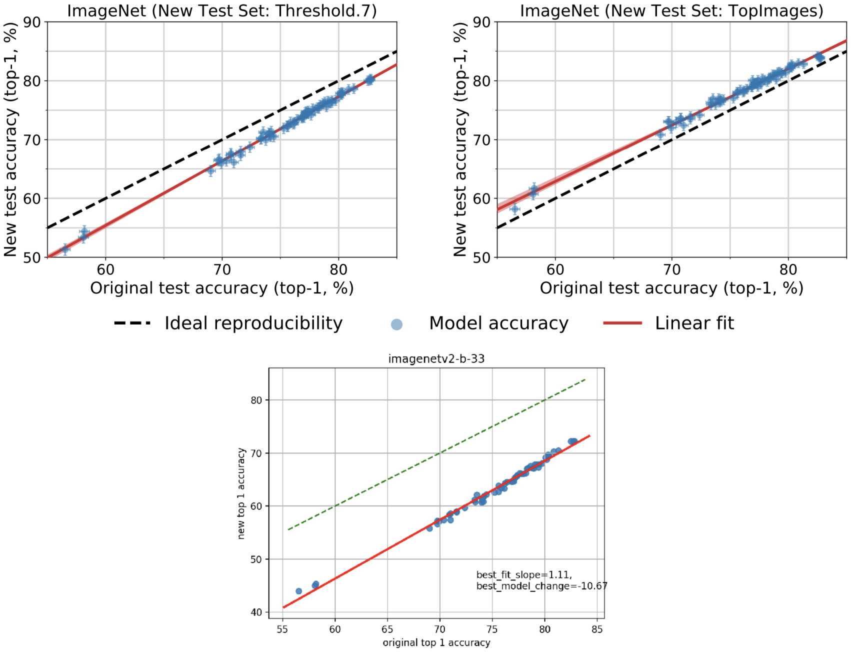
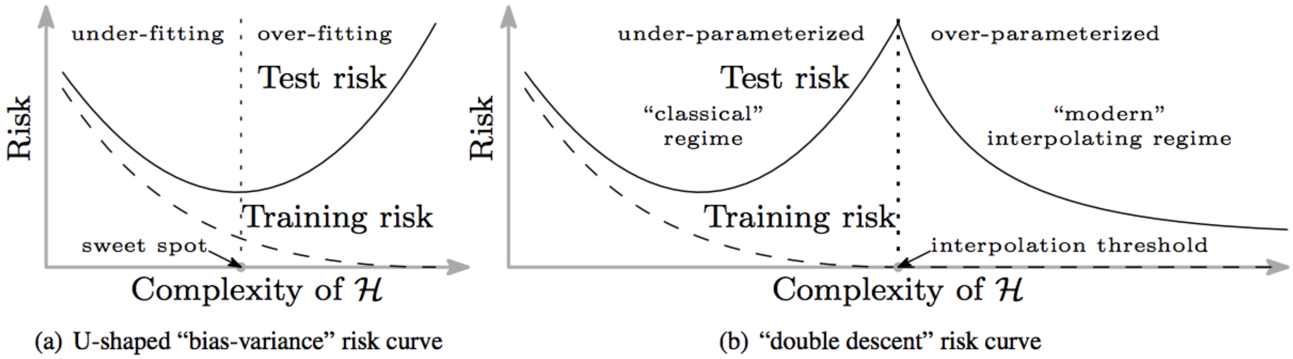


Figure 11.4: Results from Ben Recht's study with ImageNet. Upper left: threshold = 1, upper right: threshold = 0.7, bottom: all annotated images. We can see that the accuracy of all models drops as the threshold for fraction of annotators who labelled that image decreases, but the relative ranking of models is preserved.

11.3.4 Theory for Modern Neural Networks

Looking for a theoretical explanation for these results leads us to a theory for modern neural networks. This theory extends the bias-variance tradeoff curve by adding a ”modern” interpolating regime past the ”classical” regime where the test error continuously decreases as model complexity increases [Figure 11.5]. This is known as the ”double descent” risk curve (Belkin et al., 2018). After a certain threshold, the large number of parameters is no longer overparameterization; rather, the model becomes implicitly regularized by running SGD since the model tries to interpolate between points as smoothly as possible during the local search process. We see this result when looking at the model norm using Random Fourier Features, Random Relu, and a fully connected neural network [Figure 11.6]. In classical theory, empirical risk minimization is achieved through smoothness as well, only through the use of less complex models.



(a) U-shaped “bias-variance” risk curve

(b) “double descent” risk curve

Figure 11.5: Classical theory (left) vs modern theory (right)

We see this phenomenon when training error is 0, so we should force the training error to be as low as possible to see this. According to minimum-norm theory, we should perfectly fit the samples but be as smooth as possible during interpolation, because a higher norm means higher complexity. In the special case of neural networks, this minimizes test error. This holds up theoretically in the noiseless setting (see Theorem 1 on slide 25), but is in need of further exploration to compare to the classical setting and doesn’t account for label noise.

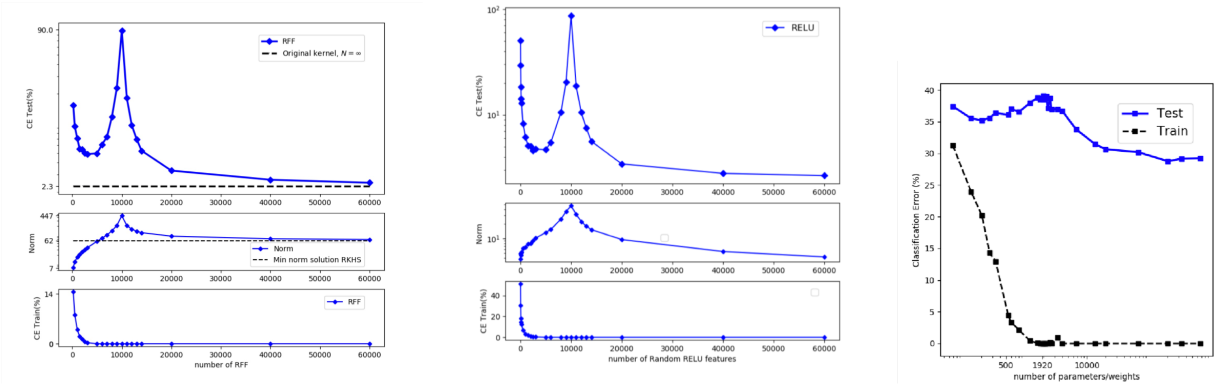


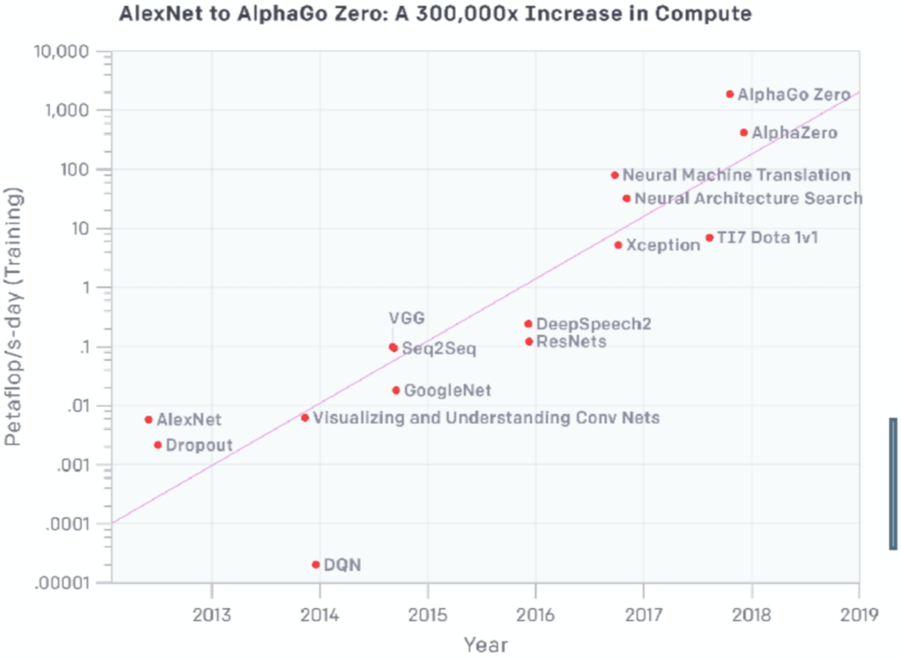
Figure 11.6: Error curves of models trained with varying parameters. Reflects the behavior of the double-U curve.

11.4 Optimization

In the case of fully-connected networks using Gradient Descent, we don't have an answer to the optimization error in theory but recent work explains to an extent why GD is able to find a globally optimal solution in neural networks (Du et al., 2018). Overall, it seems that if we force training loss to 0, we get good regularization and good generalization.

11.5 Compute Costs

The big model seems to be the solution to everything, but comes with massive costs in terms of compute during model training, especially when training a model from scratch.



G. Brockman

Figure 11.7: Cost exponentially increasing over time because models are increasing in capacity.

References

Belkin, M., D. Hsu, S. Ma, and S. Mandal
2018. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*.

Du, S. S., J. D. Lee, H. Li, L. Wang, and X. Zhai
2018. Gradient descent finds global minima of deep neural networks. *CoRR*, abs/1811.03804.

Recht, B., R. Roelofs, L. Schmidt, and V. Shankar
2018. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*.

Recht, B., R. Roelofs, L. Schmidt, and V. Shankar
2019. Do imagenet classifiers generalize to imagenet?