Derek Lee Deep Learning Fall 2020

Professor Curro Quiz #8

**1. Make an argument about super-convergence as it relates to epoch-wise double descent.**

Double descent explains the contradiction between the ML theory that bigger models are better and the statistical theory concerning the bias-variance trade-off. The proposed explanation is that at the interpolation threshold (once the model reaches 0 training error), there is "effectively only one model that fits the train data… there are no 'good models' which both interpolate the train set and perform well on the test set."[1] However, once the model is over-parameterized, many models can fit the train set, with some of those models being the previously 'good models'. SGD has an implicit bias, which leads it to those 'good models'.

This phenomenon does not only appear when varying the size of the model; it also appears when varying the number of training epochs. The authors state that this is consistent with the generalized double descent hypothesis, because increasing the training time also increases the Effective Model Complexity (EMC). Therefore, a "sufficiently large model transitions from under- to over-parameterized over the course of training."

A model with super-convergence can reach the interpolation threshold significantly faster (in a fewer number of epochs) than a model without super-convergence. Once it gets past the interpolation threshold, it can eventually reach a lower test error than would be possible before the interpolation threshold. In addition, it would also potentially reach a lower test error faster than a model without super-convergence. Based on Figure 9 in "Double Descent", it takes around two orders of magnitude of additional training epochs to reach a test error lower than the minimum achieved before the interpolation threshold. Super-convergence can significantly speed-up the process, making training for epoch-wise double descent much more viable.

Another (potentially related) possibility is that super-convergence works symbiotically with epoch-wise double descent. In Figure 6 in "Super-convergence", the validation accuracies appear to show a phenomenon similar to double descent. Performance initially gets worse, before

[1] https://openai.com/blog/deep-double-descent/

getting better. However, it should be noted that the spike in accuracy near the end does not seem characteristic of double descent. Double descent has a more gradual increase, over many epochs, in contrast to the phenomenon depicted by the figure, which has a sharp increase. However, this may be an effect of combining super-convergence with double descent.