Paper Review: Reconciling modern machine learning practice and the classical bias-variance trade-off

Brief Summary:

A common practice when utilizing large neural networks recently has been to choose a network architecture that is large enough to obtain zero empirical (training) loss. Contrary to the classical bias-variance trade-off method that limits the number of parameters for the network based on a trade-off between over and underfitting, these large neural networks result in optimal prediction results on the test data. This precise contradiction has been the motivation for the authors in reconciling the classical and modern machine learning practices through a *double-descent* unified performance curve.

The authors are able to connect the true/test performance of the model with model capacity, provide empirical evidence with real and synthetic data testing and suggest underlying reasons for the unified *double-descent* performance curve. The point of interpolation is defined as one where the model achieves perfect fits on the training data, and when the functional class capacity (ex: number of model parameters in a neural network) is below this interpolation threshold, the learned predictors exhibit the classical U-shaped curve displaying the bias-variance trade-off. The authors show that increasing the functional class capacity beyond the interpolation threshold results in decreasing risks (even below that depicted by the bottom of the classical U-shaped curve). For all the empirical learning problems considered by the authors, considering a large function class leads to interpolating functions with smaller functional norms which are appropriate estimates for the inductive bias of the problem and thus explain the improved true performance for models with large functional class capacity.

Initially, the authors consider the example of random Fourier features (RFF) models and show that as the number of parameters (N) of the model tends to infinity, the function class becomes a closer approximation to the reproducing kernel Hilbert space (RKHS) corresponding to the Gaussian Kernel. They find the predictor using Empirical Risk Minimization with squared loss and use the l_2 norm to choose a unique minimizer as it is the best approximation for the RKHS norm (inductive bias estimate). Using a subset of the MNIST dataset, the authors are able to obtain the *double-descent* test risk curve. They show that the interpolation threshold occurs when the functional capacity (N) equals the number of data points (n). Increasing N leads to better approximations to the smallest norm function and therefore better inductive bias approximations as N crosses the interpolation threshold. For the case of neural networks where the final layer parameters are not fixed, the authors observe similar results to that of RFF but warn that in the case of underparameterized models ($N \ll n$), the non-convexity of the ERM problem could cause the SGD to be highly sensitive to initialization making it difficult to observe the *double-descent* curve and therefore increasing the number of parameters may not result in reduced training risk. Also, for large datasets such as ImageNet, $\sim 10^9$ parameters could be required to achieve interpolation and therefore simply using the classical U-curve would be a more computationally feasible option. For the case of decision trees, the authors use the observation that the size of a decision tree can be used directly to parametrize the function class capacity and consider ensembles of multiple trees to further enlarge the functional class. Similar to the previous examples, they observe the *double-descent* curve for better generalization.

Paper Contributions:

An important contribution of the paper has been the introduction of the *double-descent* performance curve to combine the results observed through the classical and modern models. Along with showing empirical evidence of reducing test risks beyond the interpolation threshold, the authors are also able to explain the observed reduction through the minimum functional norm solution.