Complexity of Functions Versus Complexity of Predictions

Here’s an illustration of how a complex model — a neural network with 100 hidden units — can produce simple predictions when there is little data, despite being capable of representing more complex functions that fit the data better. The data is from the function $x^3 + 2 \exp(-6(x-0.3)^2)$, with Gaussian noise with standard deviation 0.2 added. The plots show 20 functions from the posterior distributions given 10 data points (left) and 100 data points (right):

The bold line is the average of these functions, the Bayesian prediction for minimizing squared error, which is smoother than some individual functions.