When Do Neural Networks Outperform Kernel Methods?

For a certain scaling of the initialization of stochastic gradient descent (SGD), wide neural networks (NN) have been shown to be well approximated by reproducing kernel Hilbert space (RKHS) methods. Recent empirical work showed that, for some classification tasks, RKHS methods can replace NNs without a large loss in performance. On the other hand, two-layers NNs are known to encode richer smoothness classes than RKHS and we know of special examples for which SGD-trained NN provably outperform RKHS. This is true also in the wide network limit, for a different scaling of the initialization.

How can we reconcile the above claims? For which tasks do NNs outperform RKHS? If feature vectors are nearly isotropic, RKHS methods suffer from the curse of dimensionality, while NNs can overcome it by learning the best low-dimensional representation. Here we show that this curse of dimensionality becomes milder if the feature vectors display the same low-dimensional structure as the target function, and we precisely characterize this tradeoff. Building on these results, we present a model that can capture in a unified framework both behaviors observed in earlier work. We hypothesize that such a latent low-dimensional structure is present in image classification. We test numerically this hypothesis by showing that specific perturbations of the training distribution degrade the performances of RKHS methods much more significantly than NNs.