Carlo Lucibello;

Entropic gradient descent algorithms and wide flat minima

The properties of flat minima in the empirical risk landscape of neural networks have been debated for some time. Increasing evidence suggests they possess better generalization capabilities with respect to sharp ones. First, I'll discuss Gaussian mixture classification models and show analytically that there exist Bayes optimal pointwise estimators which correspond to minimizers belonging to wide flat regions. These estimators can be found by applying maximum flatness algorithms either directly on the classifier (which is norm independent) or on the differentiable loss function used in learning. Next, we extend the analysis to the deep learning scenario by extensive numerical validations. Using two algorithms, Entropy-SGD and Replicated-SGD, that explicitly include in the optimization objective a non-local flatness measure known as local entropy, we consistently improve the generalization error for common architectures (e.g. ResNet, EfficientNet). An easy to compute flatness measure shows a clear correlation with test accuracy.