CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 90 **January 3, 2019**

Theory III: Dynamics and Generalization in Deep Networks[1]

Tomaso Poggio[1], Qianli Liao[1], Brando Miranda[1], Sasha Rakhlin[1], Andrzej Banburski[1], Lorenzo Rosasco[1], Kenji Kawaguchi[1], Jack Hidary[2]

[1]Center for Brains, Minds, and Machines, MIT
[2]Alphabet (Google) X

Abstract

The general features of the optimization problem for the case of overparametrized nonlinear networks have been clear for a while: SGD selects with high probability global minima vs local minima. In the overparametrized case, the key question is not optimization of the empirical risk but optimization with a generalization guarantee. In fact, a main puzzle of deep neural networks (DNNs) revolves around the apparent absence of "overfitting", defined as follows: the expected error does not get worse when increasing the number of neurons or of iterations of gradient descent. This is superficially surprising because of the large capacity demonstrated by DNNs to fit randomly labeled data and the absence of explicit regularization. Several recent efforts, including our previous versions of this technical report, strongly suggest that good test performance of deep networks depend on constraining the norm of their weights. Here we prove that

- the loss functions of deep RELU networks under square loss and logistic loss on a compact domain are invex functions;
- for such loss functions any equilibrium point is a global minimum;
- convergence is fast, the minima are close to the origin;
- the global minima have in general degenerate Hessians for which there is no direct control of the norm, apart from initialization close to the origin;
- a simple variation of gradient descent techniques called *norm-minimizing (NM) gradient descent* guarantees minimum norm minimizers under both the square loss and the exponential loss, independently of initial conditions.

A convenient norm for a deep network is the product of the Frobenius norms of the weight matrices. Control of the norm by NM ensures generalization for regression (because of the associated control of the Rademacher complexity). Margin bounds ensure control of classification error by maximization of the margin of $\tilde{f}$ – the classifier with normalized Frobenius norms – obtained by the minimization of an exponential-type loss by NM iterations.

[1]This replaces previous versions of Theory IIIa and TheoryIIIb updating several vague or incorrect statements.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.