Theory of Deep Learning III: explaining the non-overfitting puzzle

A very brief Summary

Aim

The aim is to show that gradient descent minimization of nonlinear networks is topologically equivalent to linear gradient system based on an almost degenerate Hessian.[AO1]

Slide 2

AO1 The Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field. It describes the local curvature of a function of many variables. (wikipedia)
Alex Obinikpo, 10/1/2018

Approach

- The approach here is to use the classical ODE(Ordinary Differential Equation) since this basically stops a puzzle about generalization in deep learning with elementary properties of gradient optimization techniques.

The main method

- The gradient dynamical system corresponding to training a deep network with gradient descent with a loss function L is

$$\dot{W} = -\nabla_W L(W) = -F(W).$$

- The main interest here is the behavior of the dynamical system near stable equilibrium points W^*, where $f(W^*)= 0$.
- Linearizing using H of L at W^* $(\mathbf{H}L)_{ij} = \frac{\partial^2 L}{\partial w_i \partial w_j}$ gives $\dot{W} = -HW,$
- The matrix H, has only real eigenvalues (since it is symmetric), which defines 2 main subspaces:
  - the stable subspace spanned by eigenvectors corresponding to negative eigenvalues
  - the center subspace corresponding to zero eigenvalues.

- By the center manifold existence theorem(CME) there is a neighborhood of W* such that
  - all solutions from the neighborhood tend exponentially fast to a solution in the center manifold. That is, the properties of the solution in the center manifolds depends on the non linear parts of the Jacobian matrix F or the Hessian H.
- Thus the following result were obtained by mathematical proofs in the work:
  - Polynomial deep networks can approximate arbitrarily well a standard Deep Network with ReLU activations.
- Now, if the GD of a MultiLayer over parametrized network converges to a minimum with zero, the Hessian has one or more zero Eigen values.
- Again, if W* is a stable equilibrium of the Gradient dynamics, and W* is a zero minimizer
  - then by implication each of the W*s found by the GD, is locally well approximated by a quadratic degenerate minimum. Thus showing that the dynamics of gradient descent for a deep network near such a minimum is topologically equivalent to the dynamics of the corresponding linear network.

Slide 5

AO2 CME states that if F has r derivatives (as in the case of deep polynomial networks)
then at every equilibrium ,W*, there is a Cr stable manifold and a C-r-1 center manifold which is sometimes called slow manifold.

Alex Obinikpo, 10/1/2018