Critical points for least-squares problems involving certain analytic functions, with applications to sigmoidal nets

Eduardo D. Sontag*

Department of Mathematics, Rutgers University, New Brunswick, NJ 089003, USA
E-mail: sontag@hilbert.rutgers.edu

This paper deals with nonlinear least-squares problems involving the fitting to data of parameterized analytic functions. For generic regression data, a general result establishes the countability, and under stronger assumptions finiteness, of the set of functions giving rise to critical points of the quadratic loss function. In the special case of what are usually called “single-hidden layer neural networks”, which are built upon the standard sigmoidal activation \( \tanh(x) \) (or equivalently \( (1 + e^{-x})^{-1} \)), a rough upper bound for this cardinality is provided as well.

**Keywords:** Analytic geometry, theory of exponentials, neural networks, regression.

1. Introduction

A very typical problem concerning function approximation and regression with so-called artificial neural networks, especially in applications dealing with learning and pattern recognition, is as follows. There is given a specification of a wiring diagram (a labeled graph) that stipulates how information flows from node to node (nodes being typically called “neurons”), and, for each such node, there is a rule that restricts the particular type of combination (linear, polynomial, and so forth) of the incoming signals that will be used as input to the node. These signals arrive from other nodes as well as from external sources. In addition, a transfer function (“activation”) is specified for each node; this function indicates what computation is performed by that node on its input in order to produce the output computed by the respective node. One of the nodes acts as a designated “output node”, and its output represents the response of the whole network to the external inputs. Once such an architecture has been defined, it remains to set the numerical values of the constants appearing as “weights” or “parameters”

*Supported in part by US Air Force Grant AFOSR-94-0293.

© J.C. Baltzer AG, Science Publishers
(such as the coefficients of linear combinations or polynomials); for each choice of these parameters, a particular function of inputs is computed. The values of parameters are often obtained by minimization of a quadratic loss function which measures the goodness of fit to a given set of numerical data.

By far the most common model in experimental work is that in which affine combinations are performed at the input of each internal node, each of which then computes an application of the "standard sigmoid" \( \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \) or equivalently, if a range of \( \{0, 1\} \) is preferred, then up to a rescaling and translation \((1 + e^{-x})^{-1}\). The output node then performs a final affine combination of the outputs of the internal nodes. These are "single hidden layer networks", which compute functions of the following type:

\[
\beta(x, u) = c_0 + \sum_{i=1}^{K} c_i \tanh(A_i u + b_i).
\]

The inputs \( u \) are vectors in \( \mathbb{R}^m \). The integer \( K \) (the "number of hidden units" in neural network terminology) is assumed to be fixed. The \( K(m + 2) + 1 \) parameters of the network (summarized by the vector "\( x \)"), namely the scalars \( c_0, \ldots, c_K \) and \( b_1, \ldots, b_K \), and the \( m \)-row vectors \( A_1, \ldots, A_K \), are thought of as variables that will be tuned so as to make \( \beta(x, u_i) \approx y_i \) when given a set of inputs and corresponding target outputs. There are portions of the parameter space that give rise to degeneracies. For instance, if one coefficient \( c_i \) (\( i \neq 0 \)) vanishes, then the loss function \( \beta \) is independent of the values of the corresponding \( A_i \) and \( b_j \). If some \( A_i = 0 \) then the corresponding term is constant and can be absorbed into \( c_0 \). If for some pair \( i \neq j \) it is the case that \( A_i = A_j \) and \( b_i = b_j \), then the terms corresponding to \( i \) and \( j \) can be combined, and only the sum \( c_i + c_j \) is relevant, resulting also in a loss of dimensionality. Similarly, since \( \tanh \) is an odd function, if \((A_i, b_i) = -(A_j, b_j)\) then terms can be combined as well. Thus a natural parameter space is the set \( X \) consisting of all the \( b_i \)'s, \( c_i \)'s, and \( A_j \)'s for which these exceptional situations do not occur.

Assume given a training or regression data set ("labeled sample")

\[
(u, y) = ((u_1, \ldots, u_N), (y_1, \ldots, y_N)),
\]

where we interpret the \( u_i \)'s as input vectors ("regressors" in statistical terms) and the scalars \( y_i \) as targets or response vectors desired for the respective \( u_i \)'s. The regression problem is that of minimizing (typically by means of steepest descent or other local search techniques) the quadratic loss

\[
E^{(u,y)}(x) := \frac{1}{2} \sum_{i=1}^{N} (\beta(x, u_i) - y_i)^2
\]

over \( X \). It has been often remarked that, even for extremely simple cases (such as \( K=1 \) and supposing that the inputs are binary vectors) there arise critical points associated to non-global local minima, and thus the study of the set of critical points of \( E^{(u,y)} \) has been frequently put forward as a research topic; see