Neurocomputing 140 (2014) 339–344

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

Singularities in the identification of dynamic systems

Junsheng Zhao [a,b], Haikun Wei [a], Weili Guo [a], Kanjian Zhang [a]

[a] Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
[b] School of Mathematical Science,. Liaocheng University, Liaocheng 252059, China

ARTICLE INFO

Article history:
Received 10 January 2014
Received in revised form
25 February 2014
Accepted 11 March 2014
Communicated by Guang Wu Zheng
Available online 8 April 2014

Keywords:
Singularity
Plateau phenomenon
Parameter identification
Gradient descent method

ABSTRACT

As is well known, the parameter spaces of hierarchical systems such as multilayer perceptrons include singularities and the plateau phenomenon is ubiquitous in the process of learning. In the singular regions, the Fisher information matrix degenerates and the loss function is almost unchanged when the parameters arrive in the singular regions, which is called the plateau phenomenon. We wonder about whether the singularities and the plateau phenomenon exist in the parameter identification process of the linear and the ordinary nonlinear systems. In this paper, we can see that in some of the parameter identification of the nonlinear systems, the Fisher information matrix degenerates, the singularities exist and we can see the plateau phenomenon in the learning curves. A simulation example is provided to demonstrate the theoretical analysis in Section 3.

1. Introduction

Parameter identification is a very active field of research in applied science, with a fast growing bibliography. During past decades, a wide variety of effective techniques have been introduced for the parameter identification, such as gradient decent method and least squares method [1–3].

Systems, either linear systems or nonlinear systems, receive input signals and transform them into output signals. Learning takes place by modifying the parameters of the system. The process of parameter identification is always called a process of learning. Learning is to imitate the stochastic behavior of the true model by modifying the parameters of the systems. Since systems are specified by a set of these parameters, we may regard the whole set of systems as a high-dimensional space or manifold whose coordinate system is given by these modifiable parameters. In the practical engineer, the behavior of a system is disturbed by a noise, so the system that receives input signal $\mathbf{x}$ emits output y stochastically. This stochastic behavior is determined by the parameters. Let us also assume that a pair $(\mathbf{x}, y)$ of input $\mathbf{x}$ and corresponding answer y is given from the outside by a teacher. A number of examples are generated by an unknown probability distribution $p_0(\mathbf{x}; y)$ of the teacher. Let us denote the set of examples as $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$. A system learns from the examples to imitate the stochastic behavior of the teacher best.

The behavior of a learning model with noise is described by a conditional probability distribution, which is the probability of output given the input, so that it can be regarded as a statistical model including a number of unknown parameters. Learning, is a kind of estimation where the parameters are modified sequentially by using examples one by one. The parameters change by learning, forming a trajectory in the manifold, and therefore we need to study the geometrical features to elucidate the behavior of learning.

However, in the neuromanifolds of neural networks, such as multilayer perceptrons, RBF networks, singular points always exist, where the Fisher information matrix degenerates. The identifiability of parameters is lost in such singular positions. It has been shown that once the parameters are attracted to singular points, the parameters are very slow to move away from them. This is a new area of research attracting much attention. Hagiwara et al. noticed this problem first [4], they used AIC to determine the size of perceptrons and found that did not work well. They found that this was because of the singular structure of the hierarchical models and investigated ways to overcome this difficulty. To accelerate the dynamics of learning, Amari proposed the natural or Riemannian gradient method of learning [5], which took into account the geometrical structure of the neuromanifold. Work done regarding this aspect includes that of Fukumizu and Amari [6], as well as the statistical–mechanical approaches taken by Saad and Solla [7], Rattray et al. [8], Rattray and Saad [9], Wei and Amari [10,11], Le Roux et al. [12], Amari [13], Pascanu and Bengio [14].

Recently, Watanabe [15] defines a widely applicable Bayesian information criterion (WBIC) by the average log likelihood function over the posterior distribution with the inverse temperature $1/\log n$. This gives us a method to establish the order of the regression models. What we wonder is that: in the identification of the linear and the ordinary nonlinear systems, except for neural

E-mail address: zhaojunshshao@163.com (J. Zhao).

http://dx.doi.org/10.1016/j.neucom.2014.03.004
0925-2312/

networks, it is useful to know whether the singularities exist. If exist, how they affect the dynamics of learning. Akaike [16–18], Hamparsum [19] and Amari et al. [20] all noticed the problem, but they did not give a detail explanation for this problem. In the present paper we intend to give an overview concerning with the above problem. We will give the singular regions of the nonlinear systems and analyzes the dynamics of them and this is the first time for which the singular regions are given for ordinary nonlinear systems.

The rest of the paper is organized as follows. In Section 2, we describe the learning paradigm. Section 3, we will discuss the singularities of the linear and the ordinary nonlinear systems. In Section 4, we give two numerical examples to demonstrate the theoretical analysis in Section 3. Section 5 is the conclusions and discussions of the paper.

2. Learning paradigm and the Fisher information matrix

2.1. Learning paradigm and gradient learning method

In fact, in the identification problems of the system models, it is required to imitate the functions specified by the teacher

$$y = f_0(\mathbf{x};\theta). \tag{1}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is an n-dimensional column vector.

Instead of the analytical form of $f_0(\mathbf{x})$, the input–outputs $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ are given, where $y_i, i = 1, 2, \ldots, N$, are the noisy versions of the true outputs,

$$y_i = f_0(\mathbf{x}_i;\theta) + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{2}$$

and ϵ_i, $i = 1, 2, \ldots, N$, denote the additive noises. The model parameter θ is adjusted to fit the training examples. The distribution of training input is assumed to be uncorrelated with the noises ϵ_i, $i = 1, 2, \ldots, N$, where the latter are subject to zero mean Gaussian distributions. Generally, when the student model is $f(\mathbf{x};\theta)$, which is different from $f_0(\mathbf{x};\theta)$, we define the instantaneous loss function as follows:

$$l(y, \mathbf{x};\theta) = \tfrac{1}{2}(y - f(\mathbf{x};\theta))^2. \tag{3}$$

Here, we will adopt the batch mode learning, where all data in the training set are summarized, and we use the gradient descent algorithm to modify the parameters

$$\theta_{t+1} = \theta_t - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{l(y_i, \mathbf{x}_i;\theta_t)}{\partial \theta_t}, \tag{4}$$

where N is the number of examples in the training set. When the number N is large, (4) is equivalent as follows:

$$\theta_{t+1} = \theta_t - \eta \left\langle \frac{l(y, \mathbf{x};\theta_t)}{\partial \theta_t} \right\rangle. \tag{5}$$

where $\langle \cdot \rangle$ denote the expectation over $(\mathbf{x}, y)$.

2.2. The Fisher information matrix

The Fisher information matrix is defined by

$$G(\theta) = E_{\mathbf{x} \sim q(\mathbf{x})} \left[E_{y|\mathbf{x};\theta} \left[\frac{\partial l(y, \mathbf{x};\theta)}{\partial \theta} \frac{\partial l(y, \mathbf{x};\theta)}{\partial \theta}^T \right] \right], \tag{6}$$

where $l(y, x;\theta) = \log p(x, y;\theta)$ is a fundamental quantity in statistics. It is positive definite in a regular statistical model, and plays the role of the Riemannian metric of the parameter space, as is shown by the information geometry [21].

The Fisher information gives the average amount of information included in one pair $(y, \mathbf{x})$ of data which is used to estimate the parameter θ.

Cramër–Rao theorem. *Let $\widehat{\theta}$ be an unbiased estimator from n examples in a regular statistical model, then the error covariance of $\widehat{\theta}$ satisfies*

$$E[(\theta - \widehat{\theta})(\theta - \widehat{\theta})^T] \geq \frac{1}{n} G^{-1}(\theta). \tag{7}$$

However, in the singular regions, the Fisher information matrix (7) degenerates, and its inverse G^{-1} does not exist. The Cramër–Rao theorem is no longer valid at the singular regions. This makes it difficult to analyze the performance of estimation and learning when the true distribution is in the singular region or in the neighborhood.

3. Dynamical systems and their singularities

In parametric system models, the input–output description is expressed by a mathematical function determined by a finite number of parameters. In order to analyze the singularity appearing in the identification process, we consider two distinct classes of systems models: linear system models and nonlinear system models. All the cases can be represented in terms of the state vector of past outputs and present and past inputs

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-m}, u_t, \ldots, u_{t-n}, \theta), \tag{8}$$

where the time lags m and n are assumed to be fixed.

3.1. Linear system models and their singularities

As the section title suggests, this class of models includes all the parametric descriptions of linear systems whose outputs are linearly related to the parameters. Generally, these are linear series expansions of this form

$$y_k = \sum_{i=1}^{m} a_i y_{k-i} + \sum_{j=0}^{n} a_j u_{k-j} + \epsilon_k, \tag{9}$$

where $\epsilon_k \sim N(0, \sigma^2)$. Suppose that the input set $u_1, u_2, \ldots, u_N$ and the corresponding output set $y_1, y_2, \ldots, y_N$ are known, and we assume

$$Z_N = H_N \theta + E, \tag{10}$$

where

$$Y_N = [y_1, y_2, \ldots, y_N]^T, \tag{11}$$

$$E = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T, \tag{12}$$

$$\theta = [a_1, \ldots, a_m, b_1, \ldots, b_n]^T, \tag{13}$$

and

$$H_N = \begin{bmatrix} -y_0 & \cdots & -y_{1-m} & u_0 & \cdots & u_{1-n} \\ -y_1 & \cdots & -y_{2-m} & u_1 & \cdots & u_{2-n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -y_N & \cdots & -y_{N-m} & u_N & \cdots & u_{N-n} \end{bmatrix}, \tag{14}$$

so we have

$$Z_N \sim N(H_N \theta, \sigma^2 I), \tag{15}$$

and the probability density function (*pdf*) of Z_n in the condition of θ can be obtained as the following:

$$f(Z_N|\theta) = (2\pi)^{-N/2} (\sigma^2)^{-N/2} \exp\left(-\tfrac{1}{2}(Z_N - H_N\theta)^T (Z_N - H_N\theta)\right). \tag{16}$$

We adopt the negative log-likelihood as a loss function, i.e.

$$l(Z_N|\theta) = -\log f(Z_N|\theta)$$