Stochastic Perceptron and Semiparametric Statistical Inference

Motoaki Kawanabe and Shun-ichi Amari

METR 93-04 March 1993

Stochastic Perceptron and Semiparametric Statistical Inference

Motoaki Kawanabe, Shun-ichi Amari,
University of Tokyo,[*]

August 30, 1993.

Abstract

It was reported (Kabashima and Shinomoto 1992) that estimators of a binary decision boundary show asymptotically strange behaviors when the probability model is ill-posed. We give a rigorous analysis of this phenomenon in a stochastic perceptron by using the estimating function method. A stochastic perceptron consists of a neuron which is excited depending on the weighted sum of inputs but its probability distribution form is unknown here. It is shown that there exists no $\sqrt{n}$-consistent estimator of the threshold value h, that is, no estimator $\hat{h}$ which converges to h in the order of $1/\sqrt{n}$ as the number n of observations increases. Therefore, the accuracy of estimation is much worse in this semiparametric case with an unspecified probability function than in the ordinary case. On the other hand, it is shown that there is a $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{w}}$ of the synaptic weight vector. These results elucidate strange behaviors of learning curves in a semiparametric statistical model.

[*] Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

1 Introduction

Learning of neural networks, especially of stochastic networks, can be regarded from the statistical point of view as sequential estimation of network parameters from randomly chosen examples. Neural networks give good non-linear models for analyzing non-linear multivariate statistical phenomena and conversely statistical analysis gives a good insight on the learning behaviors of neural networks. The present paper treats the simple stochastic perceptron (stochastic neuron) to study strange asymptotic behaviors of estimators of network parameters (synaptic weights and thresholds) in the semiparametric situation to be explained later, cf. Kabashima and Shinomoto (1992).

A simple stochastic perceptron (a stochastic neuron) classifies n-dimensional input signals $\boldsymbol{x}$ into two categories C_+ and C_- stochastically with probabilities depending on the inputs. More definitely, let $\boldsymbol{w}$ be the synaptic weight vector and let h be the threshold of a perceptron, where we assume $|\boldsymbol{w}| = 1$. Then, the signal space is divided into two parts by the separating hyperplane of the perceptron

$$H: \quad \boldsymbol{w} \cdot \boldsymbol{x} - h = 0. \tag{1.1}$$

The signed distance of a signal $\boldsymbol{x}$ from H is given by

$$d(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} - h, \tag{1.2}$$

where $d(\boldsymbol{x}) > 0$ when $\boldsymbol{x}$ is in the positive side of H and $d(\boldsymbol{x}) \leq 0$ otherwise. A stochastic perceptron emits an output $z = 1$ or -1 with a probability depending on the distance $d(\boldsymbol{x})$. We can write the probability

$$\begin{aligned} \mathrm{Prob}\{z = 1|\boldsymbol{x}\} &= 0.5 + \phi(\boldsymbol{w} \cdot \boldsymbol{x} - h), & (1.3) \\ \mathrm{Prob}\{z = -1|\boldsymbol{x}\} &= 0.5 - \phi(\boldsymbol{w} \cdot \boldsymbol{x} - h), & (1.4) \end{aligned}$$

or

$$\mathrm{Prob}\{z \,|\, \boldsymbol{x}\} = 0.5 + z\,\phi(\boldsymbol{w} \cdot \boldsymbol{x} - h), \tag{1.5}$$

where $\phi(u)$ is assumed to be a monotonically increasing smooth function with

$$\begin{aligned} \phi(0) &= 0\,, & (1.6) \\ |\,\phi(u)\,| &< 0.5\,. & (1.7) \end{aligned}$$

A specific sigmoidal function such as

$$\phi(u) = \frac{1}{2}\tanh u \tag{1.8}$$

is usually assumed for convenience, but the actual function ϕ is unknown in the real biological situation. In this case we need to estimate $\boldsymbol{w}$ and h under the condition that ϕ is unknown. The present paper proves that the asymptotic behaviors of the estimators $\hat{\boldsymbol{w}}$, $\hat{h}$ are quite different in the semiparametric situation. The same strange behaviors emerge also in learning of $\boldsymbol{w}$ and h from examples.

When ϕ is known, it is an ordinary statistical problem to estimate $\boldsymbol{w}$ and h from the training set

$$D_n = \{\, (\boldsymbol{x}_1, z_1), \cdots, (\boldsymbol{x}_n, z_n)\, \} \tag{1.9}$$

of n randomly chosen examples. When n is large, the maximum likelihood estimator is asymptotically best, and its expected square error is easily calculated from the Fisher information matrix F of the underlying statistical model. The estimation error decreases in proportion inversely to the square root of the number n of examples,

$$\hat{\boldsymbol{w}} - \boldsymbol{w} = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{1.10}$$

$$\hat{h} - h = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.11}$$

We call such an estimator a $\sqrt{n}$-consistent estimator. This shows that the squared error decrease in the oder of $1/n$, as is the case with learning curves of generalization error [Haussler et. al. (1988), Amari and Murata (1993)].

When ϕ is unknown, the situation changes drastically. Because, in addition to the parameters $\boldsymbol{w}$ and h of interest, the statistical model includes the parameter ϕ which has infinite- or function-degrees of freedom. This additional parameter is called a nuisance parameter, because we do not have any interest in estimating it. Such a model is called a semiparametric statistical model and statisticians have been searching for effective methods of inference in such a model [see, for example, Bickel et. al. (1993), Amari (1993), etc.].

We show that there is a severe difficulty in estimating the threshold h rather than in estimating the weights $\boldsymbol{w}$. To show this, we study the following models separately:

1. Threshold model (Dose-response model):

In this case, an input x is a scalar and $w = 1$, so that $d(x) = x - h$. Here, the threshold h is the only parameter to be estimated, and the probability model is

$$\mathrm{Prob}\{\, z \,|\, x \,\} = 0.5 + z\,\phi(x - h). \tag{1.12}$$

This model is also called the (semiparametric) dose-response model. In this case, an amount x of dose is applied to a subject and we observe success ($z = 1$) and failure ($z = -1$) of the result of the treatment. The probability of the success ($z = 1$) increases with the amount x of the dose as in equation 1.12, and we want to estimate the amount $x = h$ of the dose where the success probability balances with the failure probability.

Kabashima and Shinomoto (1992) studied this model, and found that the maximum likelihood estimator does not work in this case. They proposed an estimator $\hat{h}$ whose squared error converges to 0 with the oder of $n^{-2/3}$. In other words, the estimator is consistent with the stochastic order $n^{-1/3}$.

$$\hat{h} - h = O_p\left(n^{-1/3}\right) \tag{1.13}$$

This convergence is slower than $n^{-1/2}$ of the ordinary one. Their estimator is known as the maximum score estimator [see Manski (1975), Kim and Pollard (1990)]. We prove in the present paper that no estimators which converge to the true value with the stochastic order $n^{-1/2}$ exists. This shows that the learning curve has a slower characteristic in this case. We then prove that there exists an estimator better than the maximum score estimator which converges in order of $n^{-p/(2p+1)}$ where p is an arbitrary integer. This result is also known by statisticians (Nawata 1989), but the estimator which we newly propose here is much simpler and easier to construct.

2. Synaptic weights model with $h = 0$ (Orientation detection):

Here, $\boldsymbol{x}$ is m-dimensional ($m \geq 0$), $\boldsymbol{w}$ is to be estimated and $h = 0$. When the separating hyperplane passes through the origin, the probability is determined depending on $\boldsymbol{w} \cdot \boldsymbol{x}$,

$$\mathrm{Prob}\{\, z \,|\, \boldsymbol{x} \,\} = 0.5 + z\,\phi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{1.14}$$

We show that there exists an estimator $\hat{\boldsymbol{w}}$ converging to the true value $\boldsymbol{w}$ in stochastic order $n^{-1/2}$. We also construct the estimator explicitly.

3. The general case:

Combining the above two results, we propose an estimator $\hat{\boldsymbol{w}}$ and $\hat{h}$ applicable to the general case. Here $\hat{\boldsymbol{w}}$ is $\sqrt{n}$-consistent but $\hat{h}$ is $n^{-p/(2p+1)}$-consistent.

Our method is based on the geometrical consideration on the estimating functions (Amari and Kumon 1988, Amari 1993). An estimating function gives a consistent estimator of order $n^{-1/2}$ when it exists, and the estimating equation is very simple. However, we show that no estimating functions exist in the one-dimensional case. We instead propose an asymptotic estimating function which balances the bias and the variance term [cf. Geman and Bienenstock (1992)], giving a consistent estimator of order $n^{-p/(2p+1)}$.

2 Estimating Function

We briefly explain estimating functions. Given a statistical model $\{p(x, \theta)\}$ where probability distribution $p(x, \theta)$ of random variable x is specified by a scalar parameter θ, a function $g(x, \theta)$ is said to be an estimating function when it satisfies

$$1) \qquad E_\theta \left[\, g(x, \theta) \,\right] = 0, \tag{2.1}$$

$$2) \qquad E_\theta \left[\, \partial_\theta g(x, \theta) \,\right] \neq 0, \tag{2.2}$$

where E_θ is the expectation with respect to $p(x, \theta)$ and $\partial_\theta = d/d\theta$. Given n independent observations $x_1, \cdots, x_n$, $\sum g(x_i, \theta)$ is the empirical substitute to the expectation $E_\theta \left[\, g(x, \theta) \,\right]$, so that it is plausible that

$$\sum_{i=1}^{n} g(x_i, \hat{\theta}) = 0 \tag{2.3}$$

gives a good estimate $\hat{\theta}$. The score function $u(x, \theta) = \partial_\theta \log p(x, \theta)$ (that is the derivative of the log likelihood) satisfies the above conditions 1), 2) and the estimating function

$$u(x, \theta) = \partial_\theta \log p(x, \theta) \tag{2.4}$$

gives the maximum likelihood estimator.

The idea of the estimating function was introduced by Godambe (1960) as a generalization of the maximum likelihood method. It is known that the estimating function equation 2.3 gives an asymptotically normally distributed $\sqrt{n}$-consistent estimator $\hat{\theta}$.

The estimating function method can be applicable to the semiparametric model $\{\, p(x, \theta, \phi) \,\}$ where the probability distribution is specified by a parameter θ which is to be estimated and also by a nuisance parameter ϕ

of infinite dimensions in which we do not have any interest. It is in general very difficult to estimate ϕ from a finite number of observations. However, if there exists a function $g(x, \theta)$, not depending on the nuisance parameter ϕ, such that

$$1) \qquad E_{\theta,\phi}\left[\, g(x, \theta) \,\right] = 0, \tag{2.5}$$

$$2) \qquad E_{\theta,\phi}\left[\, \partial_\theta g(x, \theta) \,\right] \neq 0, \tag{2.6}$$

hold for any ϕ, where $E_{\theta,\phi}$ denotes the expectation with respect to $p(x, \theta, \phi)$, we can avoid the tedious procedure of estimating ϕ and a good estimator is obtained by the simple estimating equation

$$\sum_{i=1}^{n} g(x_i, \hat{\theta}) = 0. \tag{2.7}$$

Moreover, the estimating function method directly leads to a learning procedure. The stochastic approximation method suggests the following learning algorithm,

$$\hat{\theta}_{n+1} = \hat{\theta}_n - c_n\, g(x_{n+1}, \hat{\theta}_n), \tag{2.8}$$

where $\hat{\theta}_n$ is the estimator obtained from n previous data $x_1, \cdots, x_n$, $\hat{\theta}_{n+1}$ is the new estimator to be obtained from $\hat{\theta}_n$ and the new data x_{n+1}, and c_n is a constant satisfying

$$\sum c_n = \infty, \qquad \sum c_n^2 < \infty. \tag{2.9}$$

It is possible to study the accuracy of learning in a similar way as in the present statistical analysis [see Kabashima and Shinomoto (1993)].

However, it is in general not easy to find an estimating function in a semiparametric case. The score function does not in general satisfy the condition 1) and the maximum likelihood estimator is not necessarily unbiased. It is even not certain if an estimating function ever exists or not. Amari and Kumon (1988) analyzed this problem by generalizing the dual information geometry (Amari 1985), and gave a definite answer to this problem. Amari (1993) extended the results to be applicable to general semiparametric models.

The result becomes simpler if the probability distribution is linear in ϕ as is in the present case. In this case, the projected score or the effective score denoted by $u^E(x, \theta, \phi_0)$ gives an estimating function for any ϕ_0 even if ϕ_0 does not coincide with the true ϕ. We will explain about the projected score $u^E(x, \theta, \phi_0)$.

Let us construct a curve

$$\phi_t = \phi_0 + t\,\xi \tag{2.10}$$

showing a change of the nuisance function ϕ in the direction of ξ. We then have a parametric model

$$p(x,t) = p(x,\theta,\phi_t) \tag{2.11}$$

parameterized by t (θ being fixed). The score function of the nuisance parameter in the direction ξ is given by

$$v(x,\theta,\phi_0,\xi) = \frac{d}{d\,t} \log p(x,\theta,\phi_t)\bigg|_{t=0}. \tag{2.12}$$

Let T^N be the closure of linear space spanned by random variables v in all the directions ξ of the change in the nuisance function. We call T^N the nuisance subspace.

Let

$$u(x,\theta,\phi_0) = \frac{d}{d\,\theta} \log p(x,\theta,\phi_0) \tag{2.13}$$

be the score function of θ, which includes information how the log likelihood changes as θ changes. However, this change might have common directions to a change in the nuisance parameter ϕ. These directions cannot be used for estimating θ because they can be produced by changing ϕ. Therefore, we project u to the space orthogonal to the nuisance subspace T^N. Here, the orthogonal projection is defined by the inner product of two random variables $a(x)$ and $b(x)$ given by

$$\langle a,b\rangle = E_{\theta,\phi_0}\left[\,a(x)\,b(x)\,\right]. \tag{2.14}$$

The projected score u^E is this orthogonal component of the score $u(x,\theta,\phi_0)$.

The main results of Amari and Kumon (1988) and Amari (1993) are summarized in the following theorem.

Theorem 1

1. *An estimating function exists when T^N does not include u, that is u^E is not null.*

2. *For any ϕ',*
$$g(x, \theta) = u^E(x, \theta, \phi') \tag{2.15}$$
is an estimating function satisfying
$$E_{\theta,\phi}\left[u^E(x, \theta, \phi')\right] = 0 \tag{2.16}$$
for any ϕ.

3. *The estimator $\hat{\theta}$ is asymptotically normally distributed and is $\sqrt{n}$-consistent, with the asymptotic variance*
$$\lim_{n\to\infty} nE(\hat{\theta} - \theta)^2 = \frac{E_{\theta,\phi_0}\left[\{g(x, \theta)\}^2\right]}{\{E_{\theta,\phi_0}\left[\partial_\theta g(x, \theta)\right]\}^2}, \tag{2.17}$$
where ϕ_0 is the true nuisance parameter.

If we can choose the true ϕ_0 or one close to it, the estimator $\hat{\theta}$ is asymptotically the best. However, The point is that even if we misspecify ϕ and use a wrong ϕ, the estimator $\hat{\theta}$ is still unbiased and $\sqrt{n}$-consistent. This is a very attractive point of the estimating function method.

The result can easily be extended to the vector parameter case where $\boldsymbol{g}(x, \boldsymbol{\theta})$ is a vector function having the same dimensions as $\boldsymbol{\theta}$. Estimating functions satisfy

$$1) \qquad E_{,\phi}\ [\,\boldsymbol{g}(x, \boldsymbol{\theta})\,] = \mathbf{0}, \tag{2.18}$$
$$2) \qquad |\,E_{,\phi}\ [\mathrm{grad}\ \boldsymbol{g}(x, \boldsymbol{\theta})\,]\,| \neq 0, \tag{2.19}$$

where $|\cdot|$ means the determinant of a matrix. In the present case, the random variable is a pair $(\boldsymbol{x}, z)$ and the parameter are $(\boldsymbol{w}, h)$.

3 Dose-response Curve

We first show that no estimating functions exist for estimating h. We assume that signal x is uniformly distributed in the interval $[0, 1]$. Then, the joint probability distribution of x and z is given by
$$p(x, z; h, \phi) = 0.5 + z\phi(x - h). \tag{3.1}$$

Therefore, the score function is

$$u(x,z;h,\phi)=\frac{z\,\phi'(x-h)}{0.5+z\,\phi(x-h)}. \tag{3.2}$$

On the other hand, a change in $\phi(x)$ is written as

$$\phi_t(x)=\phi_0(x)+t\,\xi(x) \tag{3.3}$$

where $\xi(x)$ is an arbitrary smooth function satisfying

$$\xi(0)=0, \tag{3.4}$$

because of

$$\phi_t(0)=0. \tag{3.5}$$

The score in the direction ξ is given by

$$v(x,z;h,\phi,\xi)=\frac{z\,\xi(x-h)}{0.5+z\,\phi(x-h)}. \tag{3.6}$$

Because of $\phi'(0)>0$, the score u is not represented as a linear combination of v's. However, the u is proved to be included in T^N which is the closure of the set spanned by v's. This together with Begun et. al. (1983) gives the following result.

Theorem 2 *No estimating functions exist in the semiparametric dose-response model, nor exist any $\sqrt{n}$-consistent estimators.*

Remark When the curve ϕ is an odd function satisfying

$$\phi(-u)=-\phi(u), \tag{3.7}$$

in an interval $[-u_0,u_0]$, there exist estimating functions. So, there exists a $\sqrt{n}$-consistent estimator.

Even though no $\sqrt{n}$-estimator exists, we can obtain a consistent estimator with slower convergence. To obtain such one, we go back to the original idea of the estimating equation,

$$\sum_{i=1}^{n} f(x_i,z_i;\hat{h})=0 \tag{3.8}$$

and analyze the behavior of $\hat{h}$ carefully. When the solution $\hat{h}$ of this equation is close to the true value h, we have by expansion

$$\sum_{i=1}^{n} f(x_i, z_i; h) + \sum_{i=1}^{n} f'(x_i, z_i; h)\ (\hat{h} - h) + O\left(|\hat{h} - h|^2\right) = 0. \tag{3.9}$$

By neglecting the term of $|\hat{h} - h|^2$, we have

$$\hat{h} - h = -\frac{\sum f(x_i, z_i; h)}{\sum f'(x_i, z_i; h)}. \tag{3.10}$$

By the law of large numbers, the denominator converges to na, where

$$a = E\left[\, f'(x, z; h)\,\right]. \tag{3.11}$$

Let us put

$$b = E\left[\, f(x, z; h)\,\right] \tag{3.12}$$

which is 0 when $f(x, z; h)$ is an estimating function. It is not 0 because no estimating functions exist in the present case. We put

$$v^2 = V\left[\, f(x, z; h)\,\right], \tag{3.13}$$

where V denotes the variance. When b is small, the central limit theorem shows that the numerator is asymptotically normally distributed random variable which can be approximated by

$$n\, b + \sqrt{n}\, v\, \epsilon, \tag{3.14}$$

where ϵ is the standard normal random variable subject to $N(0, 1)$.

This simple but rough analysis shows that $\hat{h} - h$ is distributed approximately as

$$\hat{h} - h = -\frac{b}{a} - \frac{v}{\sqrt{n}\, a}\, \epsilon, \tag{3.15}$$

and the expected square error is

$$E\left[\, (\hat{h} - h)^2\,\right] = \frac{v^2}{n\, a^2} + \frac{b^2}{a^2}. \tag{3.16}$$

For large n, the first variance term goes to 0 but the second bias term remains finite. Therefore, we cannot obtain a consistent estimator.

In order to minimize the square error, we need to minimize both the bias and variance terms. To this end, we consider the following window function

$$f_\tau(x, z; h) = \frac{z}{\tau}\, w\left(\frac{x-h}{\tau}\right), \tag{3.17}$$

where w is a smooth rapidly decreasing function satisfying the normalization condition

$$\int w(x)\, dx = 1. \tag{3.18}$$

When τ is small, the function $\tau^{-1} w\{(x-h)/\tau\}$ is almost 0 outside a small neighborhood of $x = h$. On the other hand, the probability of $z = 1$ and $z = -1$ is fifty-fifty at $x = h$ whatever ϕ is chosen, and is almost so in a small neighborhood of $x = h$. Hence, by choosing the above f_τ, the bias term b becomes small. However, the variance term v becomes large, because this estimator takes only those data in a neighborhood at $x = h$ into account and discard all the other data outside the small window. This is the dilemma explained in Geman and Bienenstock (1991). The problem is how to compromise the bias term and the variance term by choosing a good window function. It is also important how small τ we choose depending on the number n of observations.

When a window function $w(s)$ satisfies

$$\int w(s)\, s^k ds \;=\; 0, \qquad k = 1, \cdots, p-1 \tag{3.19}$$

$$\int w(s)\, s^p ds \;=\; b \quad (\neq 0\,), \tag{3.20}$$

it is called a p-th order function.

Theorem 3 *There exists an estimator $\hat{h}$ which converges to the true value in the order of $n^{-p/(2p+1)}$. This estimator is given by using a p-th order window function.*

Proof We first calculate the necessary terms.

1. bias term:

$$b_\tau \;=\; E\,[\, f_\tau(x, z; h)\,] \tag{3.21}$$

$$\;=\; 2\int \frac{1}{\tau}\, w\left(\frac{x-h}{\tau}\right) \phi(x-h)\, dx. \tag{3.22}$$

Now we expand ϕ as

$$\phi(x) = \sum_{k=1}^{\infty} \alpha_k x^k. \tag{3.23}$$

and use a p-th order window function. Then the bias is

$$b_\tau = 2\, b\, \alpha_p \tau^p + O\left(\tau^{p+1}\right), \tag{3.24}$$

converging to 0 in the order of τ^p when τ is small.

2. variance term:

$$v_\tau^2 = E[f_\tau^2] - \{E[f_\tau]\}^2 \approx \frac{v^2}{\tau} \tag{3.25}$$

where

$$v^2 = \int \{w(s)\}^2 \, ds. \tag{3.26}$$

This shows that the variance term diverges to ∞ in the order of τ^{-1}.

3. information term:

$$a = E[f_\tau'] \;=\; 2\alpha_1 \int w(s)\, ds + O(\tau) \tag{3.27}$$

$$\approx\; 2\phi'(0) > 0 \tag{3.28}$$

This term does not depend on τ asymptotically.

The overall error term is now written as

$$E\left[(\hat{h} - h)^2\right] = \frac{1}{4a}\left[\frac{v^2}{n\,\tau} + 4b^2\alpha_p^2\,\tau^{2p}\right], \tag{3.29}$$

when we use a p-th order function w. It is easily seen that, as τ becomes small, the variance term increases while the bias term decreases. The best compromise is to choose τ depending on n such that

$$\frac{1}{n\,\tau} \approx \tau^{2p} \tag{3.30}$$

or

$$\tau \sim n^{-1/(2p+1)}. \tag{3.31}$$

The overall error is then given by

$$E\left[(\hat{h} - h)^2\right] = c\, n^{-2p/(2p+1)}, \tag{3.32}$$

proving that there exists a $n^{p/(2p+1)}$-consistent estimator. □

The last but important problem is to construct a p-th order function $w_p(x)$ explicitly. To this end, we use the Hermite polynomials defined by

$$h_p(x) = \frac{(-1)^p}{\sqrt{p!}} e^{x^2/2} \frac{d^p}{d^p x} e^{-x^2/2}. \tag{3.33}$$

They form an orthonormal system,

$$\int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} h_p(x) h_q(x) dx = \delta_{pq} \tag{3.34}$$

and $h_p(x)$ is a polynomial of degree p.

By expanding $w(s)$ in the form of

$$w(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} \left\{ 1 + \sum c_i h_i(s) \right\} \tag{3.35}$$

and by taking the conditions (3.19) into account, we have

$$w_p(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} \left\{ \frac{h_{p-1}(s)}{s} \right\} \tag{3.36}$$

when p is even. See Fig.1. It is possible to obtain a similar one for an odd p but it is not useful.

4 Orientation Detection

In the m-dimensional case where the separating hyperplane passes through the origin, we assume that $\boldsymbol{x}$ is isotropically distributed. More definitely, we assume that it is normally distributed with mean $\mathbf{0}$ and with the unit covariance matrix,

$$q(\boldsymbol{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^m} \exp\left\{ -\frac{1}{2} \boldsymbol{x} \cdot \boldsymbol{x} \right\}. \tag{4.1}$$

The probability distribution is

$$p(\boldsymbol{x}, z; \boldsymbol{w}) = q(\boldsymbol{x}) \left\{ 0.5 + z\, \phi(\boldsymbol{w} \cdot \boldsymbol{x}) \right\}. \tag{4.2}$$

We now calculate the score function,

$$u = \frac{z\, \boldsymbol{x}\, \phi'(\boldsymbol{w} \cdot \boldsymbol{x})}{0.5 + z\, \phi(\boldsymbol{w} \cdot \boldsymbol{x})}. \tag{4.3}$$

The score function of the nuisance ϕ in the direction of $\xi(\boldsymbol{w}\cdot\boldsymbol{x})$ is similarly given by

$$v[\xi] = \frac{z\,\xi(\boldsymbol{w}\cdot\boldsymbol{x})}{0.5 + z\,\phi(\boldsymbol{w}\cdot\boldsymbol{x})}. \tag{4.4}$$

Now let us put

$$s = \boldsymbol{w}\cdot\boldsymbol{x}. \tag{4.5}$$

Then $v[\xi]$ is a function of $s = \boldsymbol{w}\cdot\boldsymbol{x}$ and z. Since $\xi(s)$ is an arbitrary smooth function satisfying $\xi(0) = 0$, the nuisance tangent space T^N is included in the set of random variables expressed as a function of s and z. The projection of a random variable u to this space is given by the conditional expectation

$$E[\,u\,|\,s, z\,]. \tag{4.6}$$

The projection of the score function u is then given by

$$E[\,u\,|\,s, z\,] = \frac{z\,\boldsymbol{m}(s)\,\phi'(\boldsymbol{w}\cdot\boldsymbol{x})}{0.5 + z\,\phi(\boldsymbol{w}\cdot\boldsymbol{x})}, \tag{4.7}$$

where ϕ' is the derivative of ϕ and

$$\boldsymbol{m}(s) = E[\,\boldsymbol{x}\,|\,\boldsymbol{w}\cdot\boldsymbol{x} = s\,]. \tag{4.8}$$

Since this is included in T^N, the effective score u^E is given by

$$u^E = \frac{z\,\{\,\boldsymbol{x} - \boldsymbol{m}(\boldsymbol{w}\cdot\boldsymbol{x})\,\}\,\phi'(\boldsymbol{w}\cdot\boldsymbol{x})}{0.5 + z\,\phi(\boldsymbol{w}\cdot\boldsymbol{x})}. \tag{4.9}$$

This gives an estimating function whatever ϕ is chosen. In practical situation, we can employ adaptive method, that is, at first we compute a rough estimator $\hat{\phi}$ of the true nuisance parameter ϕ_0, and then use the effective score u^E at $\hat{\phi}$ to derive estimator $\hat{\boldsymbol{w}}$. The conditional expectation is explicitly written as

$$E[\,\boldsymbol{x}\,|\,\boldsymbol{w}\cdot\boldsymbol{x} = s\,] = s\,\boldsymbol{w} = (\boldsymbol{w}\cdot\boldsymbol{x})\,\boldsymbol{w}. \tag{4.10}$$

Therefore, the optimal estimating function is

$$u^E(x, z; \boldsymbol{w}) = \frac{z\,\{\,\boldsymbol{x} - (\boldsymbol{w}\cdot\boldsymbol{x})\,\boldsymbol{w}\,\}\,\phi_0'(\boldsymbol{w}\cdot\boldsymbol{x})}{0.5 + z\,\phi_0(\boldsymbol{w}\cdot\boldsymbol{x})}. \tag{4.11}$$

The asymptotic variance of the estimator $\hat{\boldsymbol{w}}$ derived by this estimating function is

$$\lim_{n\to\infty} nE\left[\,(\,\hat{\boldsymbol{w}} - \boldsymbol{w}\,)\,(\,\hat{\boldsymbol{w}} - \boldsymbol{w}\,)^{\mathrm{T}}\,\right] = E\left[\frac{\{\phi_0'(s)\}^2}{0.25 - \{\phi_0(s)\}^2}\,\mathrm{Cov}\,[\,\boldsymbol{x}\,|\,\boldsymbol{x}\cdot\boldsymbol{w} = s\,]\right] \tag{4.12}$$

showing that the $\hat{\boldsymbol{w}}$ is a $\sqrt{n}$-consistent estimator.

5 General Case

In the general case of a stochastic neuron, no estimating function exist because of the threshold term. However, we can construct a good asymptotic estimating function by combining the above two results. More definitely, we can use the combined function

$$\boldsymbol{f}_n(\boldsymbol{x}, z; h, \boldsymbol{w}) = \Big(f_\tau(\boldsymbol{x}, z; h, \boldsymbol{w}),\ u^E(\boldsymbol{x}, z; h, \boldsymbol{w}) \Big) \tag{5.1}$$

having two components where

$$f_\tau(x, z; h, \boldsymbol{w}) = \frac{z}{\tau}\, w \left(\frac{\boldsymbol{x} \cdot \boldsymbol{w} - h}{\tau} \right) \tag{5.2}$$

is an asymptotic estimating function constructed with an p-th order window function, and u^E is the effective score function of the direction $\boldsymbol{w}$ defined by (4.11).

In the same way as previous sections, we can derive asymptotic properties of the estimator $(\, \hat{h},\ \hat{\boldsymbol{w}}\,)$ which is defined as a solution of the estimating equation of $\boldsymbol{f}_n$. It can be shown that $(\, \hat{h},\ \hat{\boldsymbol{w}}\,)$ is a consistent estimator of the stochastic order

$$\hat{h} - h \;=\; O_p\left(n^{-p/(2p+1)}\right), \tag{5.3}$$

$$\hat{\boldsymbol{w}} - \boldsymbol{w} \;=\; O_p\left(n^{-1/2}\right). \tag{5.4}$$

We remark that the convergence speed of $\hat{h}$ is slower than $n^{-1/2}$, but this doesn't influence that of $\hat{\boldsymbol{w}}$.

6 Conclusion

In this paper, we discussed estimation of parameters of stochastic neurons in the semiparametric situation. It is shown that there is no estimating function for threshold parameter h, while there exist estimating functions for orientation parameter $\boldsymbol{w}$. We defined the p-th order asymptotic estimating functions of window type for threshold h. We also proposed an estimator $(\, \hat{h},\ \hat{\boldsymbol{w}}\,)$ derived from the combination of the asymptotic estimating function for h and the effective score function for $\boldsymbol{w}$. The estimator $\hat{h}$ of threshold is consistent with the order $n^{-p/(2p+1)}$ where p is an arbitrary integer. The convergence speed of the orientation estimator $\hat{\boldsymbol{w}}$ is of stochastic order $n^{-1/2}$.

Although we dealt with estimation of 50% point and hyperplane in this paper, it is easy to extend these results to estimation of $(\alpha \times 100)\%$ point and hyperplane.

There remain some important problems unsolved in this paper. The first is how we choose the optimal order p_n of the asymptotic estimating function for given sample size n. The second is how we can construct a consistent estimating function when the distribution of the input $\boldsymbol{x}$ is unknown. We can't use the effective score u^E in this case, because the conditional expectation of $\boldsymbol{x}$ for given $s = \boldsymbol{w} \cdot \boldsymbol{x}$ is included in it.

Although we treated only one stochastic neuron in this paper, we want to extend these results to networks of such neurons in the future.

References

[1] Amari,S. 1985. *Differential-Geometrical Method in Statistics.* Springer Lecture Note in Statistics, 28, Springer, New York.

[2] Amari,S., and Kumon,M. 1988. Estimation in the presence of infinitely many nuisance parameters — Geometry of estimating functions. *Ann.Statist.* **16**, 1044–1068.

[3] Amari,S. and Murata,N. 1993. Statistical Theory of Learning curves under entropic loss criterion. *Neural Computation*, **5**, 140–153.

[4] Amari,S., 1993. Efficient estimating functions in semiparametric statistical model. to appear.

[5] Begun,J.M., Hall,W.J., Hung,W.M., and Wellner,J.A. 1983 Information and asymptotic efficiency in parametric-nonparametric model. *Ann.Statist.* **11**, 432–452.

[6] Bickel,P.J., Klaassen,C.A.J., Ritov,V., and Wellner,J.A. 1993. *Efficient and Adaptive Estimation in Semiparametric Models.* Johns Hopkins University Press, Baltimore.

[7] Geman,S., Bienenstock,E., and Doursat,R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation.* **4**, 1–58.

[8] Godambe,V.P., and Kale,B.K. 1991. Estimating functions: an overview. In *Estimating Functions*, Godambe,V.P. ed., pp.3–20. Oxford Univ. Press, New York.

[9] Haussler,D., Littlestone,N., and Warmuth,K. 1988. Predicting $0,1$ functions on randomly drawn points. *Proc. COLT'88*, pp.280–295. Morgan Kaufmann, San Mateo, CA.

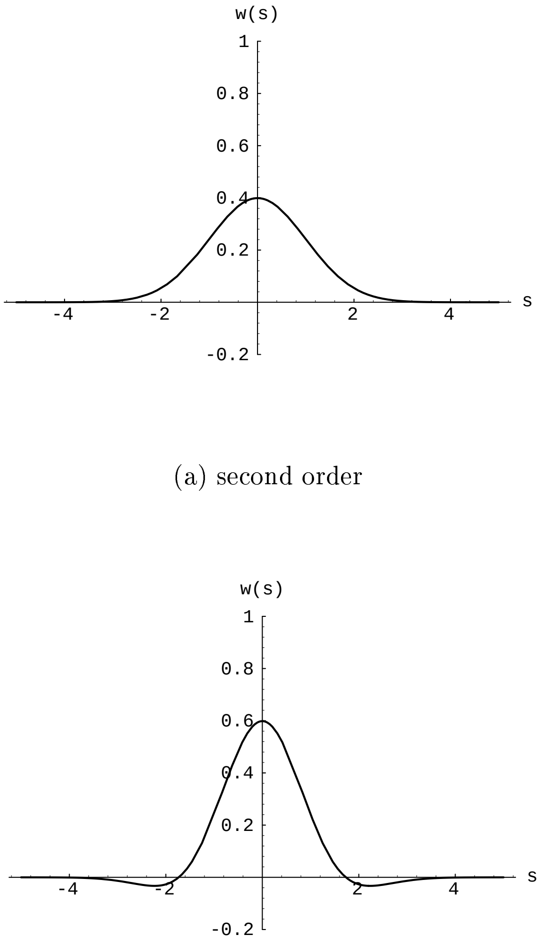
[10] Kabashima,Y., and Shinomoto,S. 1992. Learning curves for error minimum and maximum likelihood algorithms. *Neural Computation.* **4**, 712–719.

[11] Kabashima,Y., and Shinomoto,S. 1993. Learning a decision boundary from stochastic examples: Incremental algorithms with and without queries, to appear.

[12] Kim,J., and Pollard,D. 1990. Cube root asymptotics. *Ann.Statist.* **18**, 191–219.

[13] Manski,C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics.* **3**, 205–228.

[14] Nawata,K. 1989. Semiparametric estimation and efficiency bounds of binary choice models when the models contain one continuous variable. *Economics Letters.* **31**, 21–26.



(a) second order

(b) fourth order

Figure 1. window functions