Ann. Inst. Statist. Math.
Vol. 54, No. 3, 459–475 (2002)

STATISTICAL ASYMPTOTIC THEORY OF ACTIVE LEARNING

TAKAFUMI KANAMORI

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan, e-mail: kanamori@is.titech.ac.jp

(Received August 26, 1998; revised January 19, 2001)

Abstract. We study a parametric estimation problem. Our aim is to estimate or to identify the conditional probability which is called the system. We suppose that we can select appropriate inputs to the system when we gather the training data. This kind of estimation is called *active learning* in the context of the artificial neural networks. In this paper we suggest new active learning algorithms and evaluate the risk of the algorithms by using statistical asymptotic theory. The algorithms are regarded as a version of the experimental design with two-stage sampling. We verify the efficiency of the active learning by simple computer simulations.

Key words and phrases: Active learning, Kullback-Leibler divergence, risk, optimal experimental design.

1. Introduction

In this paper we consider a parametric estimation problem. Our main interest is to estimate the conditional probability $p(y \mid x)$ by using the d dimensional parametric model $M = \{p(y \mid x, \theta) : \theta = (\theta^1, \ldots, \theta^d) \in \Theta \subset \mathrm{R}^d\}$. We call $p(y \mid x)$ system. The system describes the relation between the input x and output y. Here we consider the specified case, that is, we suppose $p(y \mid x) \in M$. When we estimate the system, we can use the training data $D_T = \{(x_1, y_1), \ldots, (x_T, y_T)\}$. To measure the goodness of the fit of the estimated probability we adopt Kullback-Leibler divergence:

$$KL(p, p_\theta \mid q) = \int q(x)p(y \mid x) \log \frac{p(y \mid x)}{p(y \mid x, \theta)} dydx$$

as a loss function, where $q(x)$ is a probability density of the input x which is fixed and is unknown. To achieve good estimation we need to approximate the system well under the frequently observed inputs with respect to $q(x)$. Under this loss function the optimal parameter θ^* is defined as

$$KL(p, p_{\theta^*} \mid q) = \min_{\theta \in \Theta} KL(p, p_\theta \mid q). \tag{1.1}$$

Here $p(y \mid x, \theta^*) = p(y \mid x)$ holds because the model includes the system.

We suppose that we can choose input probability among the set of the probabilities $Q = \{r(x \mid \xi) : \xi \subset \Xi \in \mathrm{R}^k\}$ to collect the training data. That is, we can control the distribution of the inputs $\{x_1, \ldots, x_T\}$ to the system. Here ξ is the k dimensional parameter of the input probability which we can use.

The optimal experimental design treats the selection of the appropriate input distribution (Fedorov (1972)). In the field of the neural networks the term *active learning*

is used. On the other hand we can consider the situation that all inputs of the training data are randomly generated from $q(x)$. We call this kind of estimation *passive learning* in this paper. It is expected that we can estimate the system by the active learning more precisely than by the passive learning because we can select the input probability which may have advantage to estimate the system.

To construct the active learning method we adopt two-stage sampling as follows. At first the inputs of training data are drawn from $q(x)$. Based on these training data we select the appropriate input probability to the system among Q and observe the data by using the selected input probability.

We need to compare the goodness of the estimation methods. In this paper the goodness of an estimation method is measured by *risk* which is defined as $\mathrm{E}_{D_T}\{KL(p, p_{\hat{\theta}(D_T)} \mid q)\}$, where $\hat{\theta}(D_T)$ is an estimator of the optimal parameter when the training data D_T are given and E_{D_T} is the mean by the distribution of the training data. We are interested in constructing the algorithm for active learning and comparing the accuracy between the active learning and passive learning.

In the field of statistics several researchers are studying the optimal experimental design (Fedorov (1972), Silvey (1980) and Pukelsheim (1993)). The optimal experimental design treats the methods for estimation in the case that we can select the input points. In particular the optimal designs of the linear regression models are deeply investigated using the convex analysis. Several criteria to determine the optimal input points are suggested. Equivalence theorems explain the relations between several criteria that measure the goodness of inputs. In nonlinear regression problems several models have been considered (Ford *et al.* (1989)). Bayes methods are often used to the general nonlinear models (Chaloner and Verdinelli (1995)).

When the statistical model is complicated it is often difficult to calculate the optimal inputs exactly. MacKay suggested the effective methods of active learning that the posterior of the parameter in the nonlinear regression models is approximated by the normal distribution (MacKay (1992)). Fukumizu studied the active learning from the viewpoint of the statistical asymptotic theory and proposed the active learning algorithm including the model selection when the Fisher information matrix is degenerated (Fukumizu (1996)). Belue *et al.* studied active learning for multiple output multilayer perceptrons and applied the method to the real data (Belue *et al.* (1997)). Watkin studied the active learning method for simple perceptrons using statistical dynamics (Watkin and Rau (1992)).

In this paper we consider the active learning algorithm from the standpoint of the statistical asymptotic theory. The algorithm can be applied to nonlinear models satisfying some regularity conditions when the calculation of optimization is not difficult. In Section 2 we show the optimal distribution of inputs for the active learning. In Section 3 we describe the active learning algorithm based on the result of Section 2 and evaluate the risk using the statistical asymptotic theory. Numerical examples of nonlinear regression model and polynomial regression model are given in Section 4. We show some numerical examples that the active leaning algorithms work better than the passive learning.

2. Optimal distribution of the inputs

In this section we focus on the estimation of the optimal parameter θ^* satisfying (1.1) by using the training data $D_T = \{(x_1, y_1), \ldots, (x_T, y_T)\}$ which are independently

distributed. We use the maximum likelihood estimator (mle) $\hat{\theta}$ which is defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{t=1}^{T} \log p(y_t \mid x_t, \theta).$$

We consider following situation. The first wT training data $\{(x_1, y_1), \ldots, (x_{wT}, y_{wT})\}$ are identically distributed from $p(y \mid x)q(x)$ and the next $(1-w)T$ training data $\{(x_{wT+1}, y_{wT+1}), \ldots, (x_T, y_T)\}$ are identically distributed from $p(y \mid x)r(x \mid \xi)$, where w is an real value in $[0, 1]$ and $r(x \mid \xi)$ is an probability in the set Q.

LEMMA 2.1. *We use the mle to estimate the optimal parameter at the situation showed above. Then the risk is asymptotically evaluated as*

$$\text{(2.1)}\quad \mathrm{E}_{D_T}\{KL(p, p_{\hat{\theta}} \mid q)\} = \frac{1}{2T}\mathrm{Tr}G(\theta^*, q)\{wG(\theta^*, q) + (1-w)H(\theta^*, \xi)\}^{-1} + O\left(\frac{1}{T^2}\right),$$

where $G(\theta^*, q)$ and $H(\theta^*, \xi)$ are $d \times d$ matrixes, elements of which are defined as

$$\text{(2.2)}\qquad G(\theta^*, q)_{ij} := -\int p(y \mid x, \theta^*)q(x)\frac{\partial^2}{\partial\theta^i\partial j}\log p(y \mid x, \theta^*)dydx,$$

$$\text{(2.3)}\qquad H(\theta^*, \xi)_{ij} := -\int p(y \mid x, \theta^*)r(x \mid \xi)\frac{\partial^2}{\partial\theta^i\partial j}\log p(y \mid x, \theta^*)dydx$$

respectively. These are the Fisher information matrixes. We suppose that these matrixes are non-singular.

We can prove this lemma by the standard technique of the statistical asymptotic theory (White (1982)).

We define w^* and ξ^* as

$$\text{(2.4)}\qquad (w^*, \xi^*) = \arg\min_{w\in[0,1],\xi\in\Xi} \mathrm{Tr}G(\theta^*, q)\left\{wG(\theta^*, q) + (1-w)H(\theta^*, \xi)\right\}^{-1},$$

that is, the parameter w^* and ξ^* are the minimizer of the risk in the order of $1/T$. From Lemma 2.1 the sampling plan using the parameter w^* and ξ^* is asymptotically optimal in the meaning of the risk. Generally w^* and ξ^* depend on θ^* and $q(x)$ both of which we do not know. Then we need to estimate w^* and ξ^* in the active learning algorithm, which we construct in the next section.

On the passive learning the input probability to the system is fixed to $q(x)$. The risk of passive learning is $d/2T + O(1/T^2)$, where d is the dimension of the parameter space Θ. If w and ξ satisfy

$$\text{(2.5)}\qquad d > \mathrm{Tr}G(\theta^*, q)\left\{wG(\theta^*, q) + (1-w)H(\theta^*, \xi)\right\}^{-1},$$

the optimal parameter can be estimated more precisely by using active learning than by using passive learning.

3. Active learning algorithm

In this section we construct active leaning algorithms (ALA) based on the result in the previous section and we evaluate the risk of the active learning algorithm.

We call the following algorithm ALA(t), where t is a parameter of the algorithm which is the number of initial training data. We suppose that the total number of the training data is T. In the following algorithm the input probability of first t training data is $q(x)$. The probability of the inputs at the next stage are determined from the first t training data.

ALA(t)

Step 1. Gather t inputs from $q(x)$ and gather t training data $\{(x_s, y_s) \mid s = 1, \ldots, t\}$, where t satisfies $t = o(T)$ and $\lim_{T\to\infty} t = \infty$.

Step 2. Calculate the maximum likelihood estimator $\hat{\theta}_0$ based on t training data

$$\hat{\theta}_0 = \arg\max_{\theta\in\Theta} \sum_{s=1}^{t} \log p(y_s \mid x_s, \theta), \tag{3.1}$$

and then obtain $(\hat{w}, \hat{\xi})$ which satisfies

$$(\hat{w}, \hat{\xi}) = \arg \min_{w\in[0,1],\xi\in\Xi} \mathrm{Tr}\hat{G}(w\hat{G} + (1-w)H(\hat{\theta}_0, \xi))^{-1}, \tag{3.2}$$

where $\hat{G}$ is defined as

$$\hat{G}_{ij} = -\frac{1}{t}\sum_{s=1}^{t} \int p(y \mid x_s, \hat{\theta}_0) \frac{\partial^2}{\partial\theta^i \partial\theta^j} \log p(y \mid x_s, \hat{\theta}_0) dy. \tag{3.3}$$

Step 3. If $\hat{w} < \frac{t}{T}$ then Case 1 else Case 2.

Case 1. Gather $T - t$ training data $\{(x_s, y_s) \mid s = t+1, \ldots, T\}$, where their inputs are identically distributed from $r(x \mid \hat{\xi})$.

Case 2. Gather $\hat{w}T - t$ training data $\{(x_s, y_s) \mid s = t+1, \ldots, \hat{w}T\}$, where their inputs are identically distributed from $q(x)$. Next gather $(1-\hat{w})T$ training data $\{(x_s, y_s) \mid s = \hat{w}T + 1, \ldots, T\}$, where their inputs are identically distributed from $r(x \mid \hat{\xi})$.

Step 4. Calculate the maximum likelihood estimator $\hat{\theta}$ based on all obtained training data $\{(x_s, y_s) \mid s = 1, \ldots, T\}$:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{s=1}^{T} \log p(y_s \mid x_s, \theta). \tag{3.4}$$

It is important to verify that the risk of ALA(t) is asymptotically equal to

$$\frac{1}{2T}\mathrm{Tr}G(\theta^*, q)\{w^*G(\theta^*, q) + (1-w^*)H(\theta^*, \xi^*)\}^{-1}, \tag{3.5}$$

where w^* and ξ^* are defined by (2.4). The following theorem shows that the above assertion is correct and that moreover we can decide the optimal parameter t_{op} of ALA(t) by calculating the higher order term of the risk of ALA(t).

THEOREM 3.1. *We suppose that there exist a value* $\epsilon \in (0,1)$ *which satisfies* $\lim_{T\to\infty} T^{\epsilon}/t = 0$. *When* w^* *is equal to* 0, *the risk of* ALA(t) *is*

$$\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1} + \frac{1}{Tt}A + \frac{t}{T^2}B + o\left(\frac{1}{Tt},\frac{t}{T^2}\right), \tag{3.6}$$

where A *and* B *are positive numbers defined as*

$$A = \frac{1}{4}\sum_{i,j}\alpha^{ij}\frac{\partial^2}{\partial\xi^i\partial\xi^j}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1},$$
$$\alpha^{ij} = \lim_{t\to\infty} t\cdot \mathrm{E}_{(x^{(t)},y^{(t)})}\{(\hat{\xi}-\xi^*)^i(\hat{\xi}-\xi^*)^j\},$$
$$B = \frac{1}{2}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1}\{H(\theta^*,\xi^*)-G(\theta^*,q)\}H(\theta^*,\xi^*)^{-1},$$

where $\mathrm{E}_{(x^{(t)},y^{(t)})}$ *is the mean by the distribution of* $\{(x_s,y_s) \mid s = 1,\ldots,t\}$. *We can calculate* α^{ij} *explicitly (see the Appendix).*

When w^* *is a value in the open interval* $(0,1)$, *the risk of* ALA(t) *is*

$$\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)\{w^*G(\theta^*,q)+(1-w^*)H(\theta^*,\xi^*)\}^{-1} + \frac{1}{Tt}C + O\left(\frac{\sqrt{t}}{T^2},\frac{t\sqrt{t}}{T^2\sqrt{T}},\frac{1}{Tt\sqrt{t}}\right), \tag{3.7}$$

where C *is a positive number defined as*

$$C = \frac{1}{4}\sum_{i,j}\beta^{ij}\frac{\partial^2}{\partial\delta^i\partial\delta^j}\mathrm{Tr}G(\theta^*,q)\{w^*G(\theta^*,q)+(1-w^*)H(\theta^*,\xi^*)\}^{-1}.$$

Here β^{ij} *and* δ *are defined as*

$$\beta^{ij} = \lim_{t\to\infty} t\cdot \mathrm{E}_{(x^{(t)},y^{(t)})}\{(\hat{\delta}-\delta^*)^i(\hat{\delta}-\delta^*)^j\}$$
$$\hat{\delta} = (\hat{w},\hat{\xi}), \qquad \delta^* = (w^*,\xi^*),$$

respectively.

When w^* *is equal to* 1, *the risk is*

$$\frac{d}{2T} + O\left(\frac{1}{T^2}\right), \tag{3.8}$$

where d *is the dimension of the parameter space* Θ.

The proof is deferred to the Appendix.

COROLLARY 3.1. *Let us define* t_{op} *as the minimizer of the risk of* ALA(t). *We can asymptotically calculate the optimal parameter* t_{op} *of* ALA(t) *from* (3.6) *and* (3.7). *When* w^* *is equal to* 0 *and* B *which is defined in Theorem* 3.1 *is not equal to* 0,

$$t_{op} = \sqrt{T}\sqrt{\frac{A}{B}}, \tag{3.9}$$

and the risk is evaluated as

$$\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1}+O\left(\frac{1}{T\sqrt{T}}\right). \tag{3.10}$$

When w^* is a value in $(0,1)$,

$$t_{op}=O(T^{3/5}), \tag{3.11}$$

and the risk is

$$\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)\{w^*G(\theta^*,q)+(1-w^*)H(\theta^*,\xi^*)\}^{-1}+O\left(\frac{1}{T^{8/5}}\right). \tag{3.12}$$

The proposition of this corollary can be verified from (3.6) and (3.7).

Knowing the result of Corollary 3.1 we can improve ALA(t). If we use the algorithm ALA(t), the optimal order of t_{op} depends on the value of w^*. It is incompatible since we cannot estimate whether w^* is equal to 0 or not before observing the training data. The improved algorithm is ALA2(t_1,t_2) as follows, where t_1 is $O(\sqrt{T})$ and t_2 is $O(T^{3/5})$ respectively.

ALA2(t_1,t_2)

Step 1. Gather t_1 training data $\{(x_s,y_s) \mid s=1,\ldots,t_1\}$, where their inputs are distributed from $q(x)$.

Step 2. Calculate the values $(\hat{w}_0,\hat{\xi}_0)$ from t_1 training data in the same way as ALA(t).

Step 3. If $\hat{w}_0<\frac{t_1}{T}$ then Case 1 else Case 2.

Case 1. Gather $T-t_1$ training data $\{(x_s,y_s) \mid s=t_1+1,\ldots,T\}$, where their inputs are distributed from $r(x \mid \hat{\xi}_0)$.

Case 2. Gather t_2-t_1 training data $\{(x_s,y_s) \mid s=t_1+1,\ldots,t_2\}$, where their inputs are distributed from $q(x)$ and calculate $(\hat{w}_1,\hat{\xi}_1)$ from t_2 training data $\{(x_s,y_s) \mid s=1,\ldots,t_2\}$ in the same way as ALA(t). If $\hat{w}_1<\frac{t_2}{T}$ then Case a else Case b.

Case a. Gather $T-t_2$ training data $\{(x_s,y_s) \mid s=t_2+1,\ldots,T\}$, where their inputs are distributed from $r(x \mid \hat{\xi}_1)$.

Case b. Gather $\hat{w}_1T-t_2$ training data $\{(x_s,y_s) \mid s=t_2+1,\ldots,\hat{w}_1T\}$, where their inputs are distributed from $q(x)$. Next gather $(1-\hat{w}_1)T$ training data $\{(x_s,y_s) \mid s=\hat{w}_1T+1,\ldots,T\}$, where their inputs are distributed from $r(x \mid \hat{\xi}_1)$.

Step 4. Calculate the maximum likelihood estimator $\hat{\theta}$ from all obtained training data $\{(x_s,y_s) \mid s=1,\ldots,T\}$.

We can calculate the risk of ALA2(t_1,t_2) in the same manner of the Theorem 3.1. When we use the algorithm ALA2(t_1,t_2), the risk is equal to (3.10), (3.12) or (3.8) according to the value of w^*. That is, the algorithm ALA2(t_1,t_2) is adaptive with respect to w^* in this sense.

4. Numerical experiment

In this section we show the simple computer simulations of active learning.

Example 1. (Two-layer perceptron models) The two-layer perceptron models is often used in the field of neural networks (Bishop (1995), Chapter 4). The two-layer perceptron models M_m is defined as

$$M_m = \left\{ \sum_{i=1}^{m} a_i u(xw_i + b_i) | a_i, w_i, b_i \in \mathrm{R} \right\}, \tag{4.1}$$

where $u(x) = 1/(1 + \exp(-x))$. When the parameters $a_i, w_i, b_i, i = 1, \ldots, m$ is fixed, a function from the input $x \in \mathrm{R}$ to the output $y \in \mathrm{R}$ is specified. We consider the regression problem, that is, the estimation of the mean values of output y corresponding to input x by using the two-layer perceptron model.

We suppose that the range of the inputs is restricted on the interval $[-1, 1]$. The probability $q(x)$ is the linear transformed beta distribution, that is, the input x is expressed as $x = 2x_0 - 1$, where x_0 is distributed from the beta distribution whose density function is proportional to $x_0^2(1 - x_0)^3$. When the input x is given, output y is written as

$$y = 2.0\ u(5.0x) + 5.0\ u(-x - 2.0) - 1.5\ u(3.0x - 1.0) + \epsilon, \tag{4.2}$$

where ϵ is according to $N(0, 0.1^2)$ which means the normal distribution with the mean 0.0 and the standard deviation 0.1. The regression curve of (4.2) is shown in Fig. 1. We use M_3 as the parametric model. Thus the system is included in the model. We define the set of probabilities as histogram on the interval $[-1, 1]$. More precisely let us define as

$$\delta_i(x) = \begin{cases} 1 & x \in \left[\dfrac{i-1}{5} \dfrac{i}{5} \right) \\ 0 & \text{otherwise} \end{cases},$$

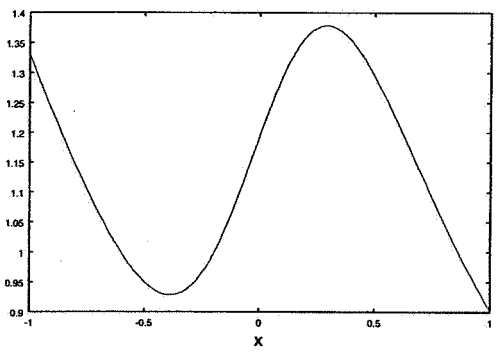

Fig. 1. The regression curve of the two-layer perceptron (4.2).

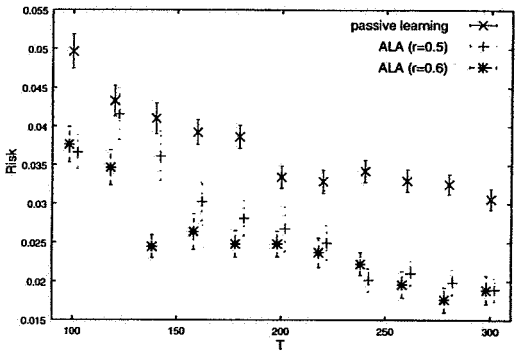

Fig. 2. The risk of ALA($4T^r$) and that of passive learning are plotted as functions of the total number of training data T when the model is two-layer perceptron. The parameter r takes 0.5 and 0.6 respectively.

then the set of the probabilities Q is defined as

$$Q = \left\{ r(x \mid \xi) = \sum_{i=1}^{5} \xi_i \, \delta_i \left(\frac{x+1}{2} \right) \;\middle|\; \sum_{i=1}^{5} \xi_i = 5, \xi_i \geq 0 \text{ for all } i \right\}.$$

The passive learning method and the active learning algorithm ALA(t) are used to estimate the parameter of the two-layer perceptron model. The risk is approximated by the mean of Kullback Leibler divergence between the true parameter and the estimated one.

The total number T of training data takes various values from 100 to 300. In each value of T, 100 trainings are performed by using each learning algorithm and estimate the corresponding values of the risk with the standard errors. The results are shown in Fig. 2. The simulation is done when the parameters of ALA(t) are $t = 4T^{0.5}$ and $t = 4T^{0.6}$. When the parametric model is non-linear with respect to the parameter such as two-layer perceptron model, it is difficult to calculate which is better between $t = O(T^{0.5})$ and $t = O(T^{0.6})$. But in both cases the active learning is superior to the passive learning. In this simple simulation we do not need to give much care to choosing the parameter t.

Example 2. (Linear regression models) We suppose that a system $p(y \mid x)$ is

$$y = 3x(2x-1)(5x-4) + \epsilon, \tag{4.3}$$

where ϵ is a random variable which has the normal distribution $N(0, 0.3^2)$. We suppose that we know the distribution of ϵ. The input region is restricted to the interval $[0, 1]$. The parametric model is

$$M = \{\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon \mid \theta_1, \theta_2, \theta_3 \in \mathrm{R}\}. \tag{4.4}$$

The parametric model M includes the system (4.3). The system (4.3) is realized when θ_0, θ_1, θ_2 and θ_3 are equal to $0, 12, -39$ and 30 respectively.

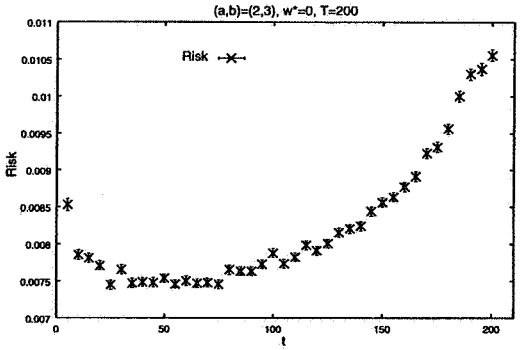

Fig. 3. The simulation of ALA(t) in the case of $(a, b) = (2, 3)$, that is, w^* is equal to 0. The total number of training data is 200. The values of the risk are plotted as a function of t with the standard-error bands.

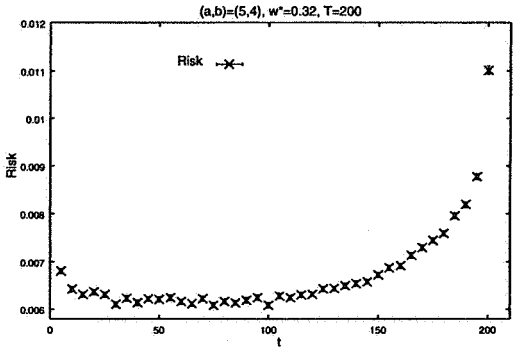

Fig. 4. The simulation of ALA(t) in the case of $(a, b) = (5, 4)$, that is, w^* is not equal to 0. The total number of training data is 200. The values of the risk are plotted as a function of t with the standard-error bands.

The probability $q(x)$ is the beta distribution on the interval $[0,1]$, that is, $q(x) \propto x^{(a-1)}(1-x)^{(b-1)}$ on the interval. We have the simulations in the cases of $(a,b) = (2,3)$ and $(a,b) = (5,4)$. Moreover the set of the probabilities Q is defined as

$$Q = \left\{ r(x \mid \xi) = \sum_{i=1}^{5} \xi_i \delta_i(x) \;\middle|\; \sum_{i=1}^{5} \xi_i = 5, \xi_i \geq 0 \text{ for all } i \right\},$$

where $\delta_i(x)$ is defined in the Example 1.

When (a,b) is equal to $(2,3)$ the optimal ratio w^* is 0. When (a,b) is equal to $(5,4)$ the optimal ratio w^* is 0.32. Figures 3 and 4 show the results of the estimation of ALA(t). The total number of training data is $T = 200$. Figures 3 and 4 correspond to the case of $(a,b) = (2,3)$ and the case of $(a,b) = (5,4)$ respectively. The risk of the algorithm ALA(t) is numerically calculated as a function of t. The risk is approximated by the mean of Kullback Leibler divergence between the true parameter and the estimated one. In each value of t, 10000 trainings are performed by using ALA(t). The calculated values of the risk are shown with the standard-error bands. In both cases of $(a,b) = (2,3)$ and $(a,b) = (5,4)$ there exist optimal size t which minimize the risk of ALA(t). The second dominant term of the risk among the order of $o(1/T)$ is different between the case of $w^* = 0$ and that of $w^* \neq 0$. This difference appears the shape of the graph of the risk as a function of t. When (a,b) is equal to $(5,4)$ the graph of the risk as a function of t is nearly flat until $t = 150$ and is rising urgently at the range of $t \geq 150$. On the other hand when (a,b) is equal to $(2,3)$ the graph of the risk is not flat.

In Figs. 5 and 6 we compare the active learning algorithms to the passive learning and D-optimal design. We use ALA(t) and ALA2(t_1, t_2) as the active learning algorithms. These figures show the risk as functions of the total number of training data. In each value of training data, 10000 training are performed by using each learning algorithm. The parameter of ALA(t) is $t = 2T^r$ and the parameters of ALA2(t_1, t_2) are $t_1 = 2T^{0.5}$ and $t_2 = 2T^{0.6}$. D-optimal design does not depend on the probability $q(x)$. In this case

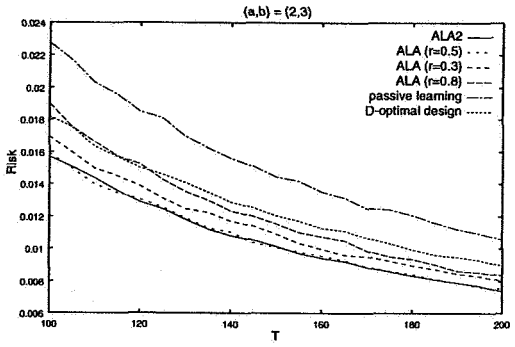

Fig. 5. The risk of ALA($2T^r$), that of ALA2($2T^{0.5}, 2T^{0.6}$), that of passive learning and that of D-optimal design are plotted as functions of the total number of training data T in the case of $(a,b) = (2,3)$. The parameter r takes 0.3, 0.5 and 0.8 respectively.

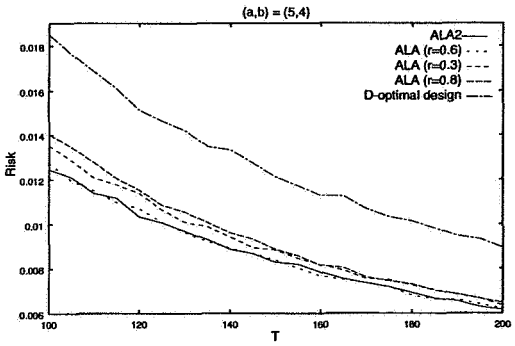

Fig. 6. The risk of ALA($2T^r$), that of ALA2($2T^{0.5}, 2T^{0.6}$) and that of D-optimal design are plotted as functions of the total number of training data T in the case of $(a,b) = (5,4)$. The parameter r takes 0.3, 0.6 and 0.8 respectively.

the D-optimal design is as follows (Fedorov (1972)):

$$\mathrm{Prob}\{x=0\}=\mathrm{Prob}\{x=1\}=\mathrm{Prob}\left\{x=\frac{5-\sqrt{5}}{10}\right\}=\mathrm{Prob}\left\{x=\frac{5+\sqrt{5}}{10}\right\}=\frac{1}{4},$$

that is, the D-optimal design is realized by four points which are distributed uniformly.

The values of r are $0.3, 0.5$ and 0.8 when (a,b) is equal to $(2,3)$. Figure 5 shows that the risk in the case of $r=0.5$ is better than the others. Moreover it can be verified that the risk of ALA($2T^{0.5}$) and that of ALA2($2T^{0.5}, 2T^{0.6}$) are almost equal. Similarly when (a,b) is equal to $(5,4)$ the risk in the case of $r=0.6$ is better than the others. Moreover the risk of ALA($2T^{0.6}$) and that of ALA2($2T^{0.5}, 2T^{0.6}$) are almost equal. These results of simulations can be explained by the result of Corollary 3.1. In the case of $(a,b)=(5,4)$ the risk of passive learning is too large to draw in the same figure. Then the graph of the passive learning is not drawn. In both cases of $(a,b)=(2,3)$ and $(a,b)=(5,4)$ D-optimal design does not perform well than the active learning. The reason is that the probability $q(x)$ is not taken into account in D-optimal design.

5. Conclusion

We propose active learning algorithms ALA(t) and ALA2(t_1, t_2) and evaluate the risk of these algorithms. We theoretically conclude that the active learning can be more efficient than passive learning if we choose the probabilities of the inputs appropriately. When we gather T training data, the optimal order of the parameter in ALA(t) is $t_{op}=O(T^{0.5})$ or $t_{op}=O(T^{0.6})$. It depends on the true parameter θ^* and the probability $q(x)$. Because of the lack of the knowledge about the true parameter and $q(x)$ before observing the training data, the optimal order of the parameter t cannot be determined in advance. To resolve this problem we construct ALA2(t_1, t_2). If we choose t_1 as $O(T^{0.5})$ and t_2 as $O(T^{0.6})$ respectively, the higher order of risk of the algorithm ALA2(t_1, t_2) is asymptotically equal to the optimal case of ALA(t). This result can be a guide to decide how many training data we should gather at the first stage of the active learning.

In this paper we suppose that the model $M=\{p(y \mid x,\theta) \mid \theta\in\Theta\}$ includes the system $p(y \mid x)$. But actually the model may not include the true probability. Then we need to extend the active learning algorithm to misspecification case.

Acknowledgements

I am indebted to Prof. S. Eguchi, Assist. Prof. F. Komaki, H. Shimodaira and the referees for their help in making this research possible.

Appendix: Proof of Theorem 3.1

In this appendix let us define $\|\cdot\|$ as Euclidean norm and ∇f as $\nabla f=(\partial_1 f,\ldots,\partial_d f)$ and the matrix $\nabla\nabla f$ as $(\nabla\nabla f)_{ij}=\partial_i\partial_j f(i,j=1,\ldots,d)$ respectively, where ∂_i means $\frac{\partial}{\partial\theta^i}$. We suppose that we get T training data $\{(x_s,y_s) \mid s=1,\ldots,T\}$ according to the active learning algorithm ALA(t). Let D_0 and D_1 be $D_0=\{(x_s,y_s) \mid s=1,\ldots,t\}$ and $D_1=\{(x_s,y_s) \mid s=t+1,\ldots,T\}$ respectively. First the risk of active learning is calculated on the condition that we get the training data D_0. Next we calculate the risk with respect to the distribution of the training data D_0.

Let us define $\tilde{w}$ as

$$\tilde{w} = \begin{cases} \frac{t}{T} & \hat{w} = 0 \\ \hat{w} & \hat{w} \neq 0. \end{cases} \tag{A.1}$$

Let us define $L(\theta)$ and $v(\theta)$ as

$$L(\theta) = -\frac{1}{T}\sum_{s=t+1}^{T} \log p(y_s \mid x_s, \theta),$$

$$v(\theta) = -\frac{1}{t}\sum_{s=1}^{t} \log p(y_s \mid x_s, \theta)$$

respectively. The estimator $\hat{\theta}$ satisfies

$$\nabla L(\hat{\theta}) + \frac{t}{T}\nabla v(\hat{\theta}) = 0. \tag{A.2}$$

Let us define $\tilde{\theta}$ as

$$\nabla L(\tilde{\theta}) = 0. \tag{A.3}$$

Considering the Taylor expansion of (A.2) around $\tilde{\theta}$, we obtain

$$\hat{\theta} - \tilde{\theta} = -\frac{t}{T}(\nabla\nabla L(\tilde{\theta}))^{-1}\nabla v(\tilde{\theta}) + O_p\left(\|\hat{\theta} - \tilde{\theta}\|^2, \frac{t}{T}\|\nabla\nabla v(\tilde{\theta})(\hat{\theta} - \tilde{\theta})\|\right). \tag{A.4}$$

Furthermore considering the Taylor expansion of (A.4) around the optimal parameter θ^*, we obtain

$$\begin{aligned} \hat{\theta} - \tilde{\theta} = & -\frac{t}{T}(\nabla\nabla L)^{-1}\nabla v - \frac{t}{T}(\nabla\nabla L)^{-1}\nabla\nabla v \cdot (\tilde{\theta} - \theta^*) \\ & + \frac{t}{T}\sum_{i=1}^{d}(\tilde{\theta} - \theta^*)^i(\nabla\nabla L)^{-1}\partial_i\nabla\nabla L(\nabla\nabla L)^{-1}\nabla v \\ & + O_p\left(\frac{t}{T^2}, \frac{t^2}{T^2}\|\nabla v\|^2, \frac{t^2}{T^2}\nabla v\right), \end{aligned} \tag{A.5}$$

where $\partial_i\nabla\nabla L$ is a matrix, kl element of which is $\partial_i\partial_k\partial_l L$. The law of large numbers gives

$$\begin{aligned} & \nabla\nabla L \\ & = \frac{\tilde{w}T - t}{T}\left(G(\theta^*, q) + O_p\left(\frac{1}{\sqrt{T}}\right)\right) + (1 - \tilde{w})\left(H(\theta^*, \hat{\xi}) + O_p\left(\frac{1}{\sqrt{T}}\right)\right). \end{aligned} \tag{A.6}$$

By substituting (A.6) to (A.5) we obtain

$$\begin{aligned} \hat{\theta} - \tilde{\theta} = & -\frac{t}{T}\{\tilde{w}G(\theta^*, q) + (1 - \tilde{w})H(\theta^*, \hat{\xi})\}^{-1}\nabla v \\ & -\frac{t}{T}\{\tilde{w}G(\theta^*, q) + (1 - \tilde{w})H(\theta^*, \hat{\xi})\}^{-1}\nabla\nabla v \cdot (\tilde{\theta} - \theta^*) \\ & + O_p\left(\frac{t}{T\sqrt{T}}\|\nabla v\|, \frac{t^2}{T^2}\|\nabla v\|, \frac{t}{T^2}\partial_i\partial_j v, \frac{t^2}{T^2\sqrt{T}}\partial_i\partial_j v, \frac{t}{T^2}, \frac{t^2}{T^2}\|\nabla v\|^2\right). \end{aligned} \tag{A.7}$$

Then we obtain $\hat{\theta}-\theta^*=(\hat{\theta}-\tilde{\theta})+(\tilde{\theta}-\theta^*)$ as follows:

$$\begin{aligned}(\text{A.8})\quad \hat{\theta}-\theta^* &= \left[I-\frac{t}{T}\{\tilde{w}G(\theta^*,q)+(1-\tilde{w})H(\theta^*,\hat{\xi})\}^{-1}\nabla\nabla v\right](\tilde{\theta}-\theta^*)\\ &\quad -\frac{\sqrt{t}}{T}\{\tilde{w}G(\theta^*,q)+(1-\tilde{w})H(\theta^*,\hat{\xi})\}^{-1}\sqrt{t}\nabla v\\ &\quad +O_p\left(\frac{t}{T\sqrt{T}}\|\nabla v\|,\frac{t^2}{T^2}\|\nabla v\|,\frac{t}{T^2}\partial_i\partial_j v,\frac{t^2}{T^2\sqrt{T}}\partial_i\partial_j v,\frac{t}{T^2},\frac{t^2}{T^2}\|\nabla v\|^2\right),\end{aligned}$$

where I is d dimensional identity matrix. It is noted that

$$\mathrm{E}_{D_1|D_0}\{\tilde{\theta}-\theta^*\}=O\left(\frac{1}{T}\right),$$

$$\mathrm{E}_{D_1|D_0}\{(\tilde{\theta}-\theta^*)(\tilde{\theta}-\theta^*)'\}=\{(\tilde{w}T-t)\ G(\theta^*,q)+(1-\tilde{w})T\ H(\theta^*,\hat{\xi})\}^{-1}+O\left(\frac{1}{T^2}\right)$$

where $\mathrm{E}_{D_1|D_0}$ is the expectation by the distribution of training data D_1 on the condition of training data D_0 and $\cdot'$ is transposition. Let us define $K(w,\xi)$ as

$$K(w,\xi)=wG(\theta^*,q)+(1-w)H(\theta^*,\xi). \tag{A.9}$$

We can calculate the asymptotic variance of $\hat{\theta}-\theta^*$ on the condition of training data D_0 by substituting (A.8) as follows:

$$\begin{aligned}(\text{A.10})\ &\mathrm{E}_{D_1|D_0}\{(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)'\}\\ &=\frac{1}{T}K(\tilde{w},\hat{\xi})^{-1}-\frac{2t}{T^2}K(\tilde{w},\hat{\xi})^{-1}\nabla\nabla vK(\tilde{w},\hat{\xi})^{-1}\\ &\quad +\frac{t}{T^2}K(\tilde{w},\hat{\xi})^{-1}G(\theta^*,q)K(\tilde{w},\hat{\xi})^{-1}+\frac{t}{T^2}K(\tilde{w},\hat{\xi})^{-1}\sqrt{t}\nabla v\sqrt{t}\nabla vK(\tilde{w},\hat{\xi})^{-1}\\ &\quad +O\left(\frac{1}{T^2},\frac{t}{T^2}\|\nabla v\|,\frac{t^2}{T^2\sqrt{T}}\|\nabla v\|,\frac{t}{T^2\sqrt{T}}\partial_i\partial_j v,\frac{t^2}{T^3}\partial_i\partial_j v,\frac{t}{T^2\sqrt{T}},\right.\\ &\qquad \frac{t^2}{T^2\sqrt{T}}\|\nabla v\|^2,\frac{t^3}{T^3}\|\nabla v\|^2,\frac{t^2}{T^3}\|\nabla\nabla v\nabla v\|,\frac{t^3}{T^3\sqrt{T}}\|\nabla\nabla v\nabla v\|,\\ &\qquad \left.\frac{t^2}{T^3}\|\nabla v\|,\frac{t^3}{T^3}\|\nabla v\|^3\right).\end{aligned}$$

It is noted that

$$\nabla v\sim N\left(0,\frac{1}{t}G(\theta^*,q)\right),\qquad \mathrm{E}_{D_0}\{\nabla\nabla v\}=G(\theta^*,q),\qquad \text{and}$$

$$\mathrm{E}_{D_0}\{\partial_{i_1}\partial_{i_2}\cdots\partial_{i_j}v\}=O(1),\qquad (j\geq 3). \tag{A.11}$$

Thus we obtain the asymptotic expansion of the risk:

$$\begin{aligned}(\text{A.12})\quad &\frac{1}{2}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{\mathrm{E}_{D_1|D_0}\{(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)'\}\}\\ &\quad =\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{K(\tilde{w},\hat{\xi})^{-1}\}\end{aligned}$$

$$
\begin{aligned}
&-\frac{2t}{T^2}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{K(\tilde{w},\hat{\xi})^{-1}\nabla\nabla v K(\tilde{w},\hat{\xi})^{-1}\}\\
&+\frac{t}{T^2}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{K(\tilde{w},\hat{\xi})^{-1}G(\theta^*,q)K(\tilde{w},\hat{\xi})^{-1}\}\\
&+\frac{t}{T^2}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{K(\tilde{w},\hat{\xi})^{-1}\sqrt{t}\nabla v\sqrt{t}\nabla v' K(\tilde{w},\hat{\xi})^{-1}\}\\
&+O\left(\frac{\sqrt{t}}{T^2},\frac{t\sqrt{t}}{T^2\sqrt{T}}\right).
\end{aligned}
$$

Here we write $K(\delta)$ as $K(w,\xi)$ where $\delta=(w,\xi)\in \mathrm{R}^{k+1}$. Let us define δ^* and $\hat{\delta}$ as (w^*,ξ^*) and $(\hat{w},\hat{\xi})$ respectively. First we consider the case of $0<w^*<1$. Expanding $K(\tilde{w},\hat{\xi})$ around (w^*,ξ^*) and neglecting the higher orders we obtain

$$
\begin{aligned}
\text{(A.13)}\quad K(\hat{\delta}) &= K(\delta^*)+\sum_{i=1}^{d}\frac{\partial}{\partial\delta^i}K(\delta^*)(\hat{\delta}-\delta^*)^i+\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial\delta^i\partial\delta^j}K(\delta^*)(\hat{\delta}-\delta^*)^i(\hat{\delta}-\delta^*)^j\\
&\quad +O_p(\|\hat{\delta}-\delta^*\|^3).
\end{aligned}
$$

Knowing that

$$
\text{(A.14)}\qquad \frac{\partial}{\partial\delta^i}\mathrm{Tr}G(\theta^*,q)K(\delta^*)^{-1}=0
$$

for all i from definition of (w^*,ξ^*), we can write (A.12) as follows:

$$
\begin{aligned}
\text{(A.15)}\quad &\frac{1}{2}\mathrm{Tr}G(\theta^*,q)\mathrm{E}_{D_0}\{\mathrm{E}_{D_1|D_0}\{(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)'\}\}\\
&\quad=\frac{1}{2T}\mathrm{Tr}G(\theta^*,q)\{w^*G(\theta^*,q)+(1-w^*)H(\theta^*,\xi^*)\}^{-1}\\
&\qquad+\frac{1}{4T}\sum_{i,j=1}^{d}\mathrm{E}_{D_0}\{(\hat{\delta}^i-\delta^{*i})(\hat{\delta}^j-\delta^{*j})\}\frac{\partial^2}{\partial\delta^i\partial\delta^j}\mathrm{Tr}G(\theta^*,q)K(\delta^*)^{-1}\\
&\qquad+O\left(\frac{\sqrt{t}}{T^2},\frac{t\sqrt{t}}{T^2\sqrt{T}},\frac{1}{Tt\sqrt{t}}\right),
\end{aligned}
$$

where we assume that the third order moment of $\hat{\delta}-\delta^*$ is $O(\frac{1}{t\sqrt{t}})$. Let β^{ij} be

$$
\text{(A.16)}\qquad \beta^{ij}=\lim_{t\to\infty}t\cdot\mathrm{E}_{D_0}\{(\hat{\delta}^i-\delta^{*i})(\hat{\delta}^j-\delta^{*j})\}.
$$

We can rewrite (A.15) as follows

$$
\begin{aligned}
\text{(A.17)}\quad \mathrm{E}_{D_T}\{KL(p,p_{\hat{\theta}(D_T)}\mid q)\} &= \frac{1}{2T}\mathrm{Tr}G(\theta^*,q)\{w^*G(\theta^*,q)+(1-w^*)H(\theta^*,\xi^*)\}^{-1}\\
&\quad+\frac{1}{4Tt}\sum_{i,j=1}^{d}\beta^{ij}\frac{\partial^2}{\partial\delta^i\partial\delta^j}\mathrm{Tr}G(\theta^*,q)K(\delta^*)^{-1}\\
&\quad+O\left(\frac{\sqrt{t}}{T^2},\frac{t\sqrt{t}}{T^2\sqrt{T}}\right).
\end{aligned}
$$

Next we suppose that w^* is equal to 0 or 1. We can prove that the probability of $\hat{w}\neq w^*$ decreases exponentially from the assumption of Theorem 3.1. It is a consequence of

large deviation theory. The exponentially decreasing term does not affect the asymptotic expansion. Then we can put $\hat{w} = w^*$ when w^* is equal to 0 or 1.

When w^* is equal to 1, the probability density of all training data is $p(y \mid x)q(x)$. Then we can use the result of ordinary learning theory and we can obtain

$$\mathrm{E}_{D_T}\{KL(p, p_{\hat{\theta}(D_T)} \mid q)\} = \frac{d}{2T} + O\left(\frac{1}{T^2}\right). \tag{A.18}$$

When w^* is equal to 0, we obtain the following expression from (A.12)

$$\begin{aligned}\mathrm{E}_{D_T}&\{KL(p, p_{\hat{\theta}(D_T)} \mid q)\} \\ &= \frac{1}{2T}\mathrm{Tr}G(\theta^*, q)H(\theta^*, \xi^*)^{-1} + \frac{1}{4Tt}\sum_{i,j=1}^{d}\alpha^{ij}\frac{\partial^2}{\partial\xi^i\partial\xi^j}\mathrm{Tr}G(\theta^*, q)H(\theta^*, \xi^*)^{-1} \\ &\quad + \frac{t}{2T^2}\mathrm{Tr}G(\theta^*, q)H(\theta^*, \xi^*)^{-1}\{H(\theta^*, \xi^*) - G(\theta^*, q)\}H(\theta^*, \xi^*)^{-1} \\ &\quad + O\left(\frac{\sqrt{t}}{T^2}, \frac{t\sqrt{t}}{T^2\sqrt{T}}\right),\end{aligned} \tag{A.19}$$

where α^{ij} is defined in Theorem 3.1.

Next we calculate α^{ij} when w^* is equal to 0. From the definition $\hat{\xi}$ satisfies

$$\frac{\partial}{\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0, \hat{\xi})^{-1} = 0, \tag{A.20}$$

where $\hat{G}$ is defined in ALA(t). Expanding (A.20) around ξ^* and neglecting the higher order terms then we obtain

$$\frac{\partial}{\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0, \xi^*)^{-1} + \sum_{s=1}^{k}\frac{\partial^2}{\partial\xi^s\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0, \xi^*)^{-1}(\hat{\xi}^s - \xi^{*s}) = 0. \tag{A.21}$$

Because of the law of large numbers and the consistency of the estimator $\hat{\theta}_0$

$$\frac{\partial^2}{\partial\xi^s\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0, \xi^*)^{-1} \to \frac{\partial^2}{\partial\xi^s\partial\xi^u}\mathrm{Tr}G(\theta^*, q)H(\theta^*, \xi^*)^{-1}, \quad (t \to \infty) \tag{A.22}$$

holds. Next we calculate the asymptotic variance of $\frac{\partial}{\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0, \xi^*)^{-1}$. Let $G(\theta; x)$ be

$$G(\theta; x) = -\int p(y \mid x, \theta)\nabla\nabla p(y \mid x, \theta)dy \tag{A.23}$$

then we obtain

$$\hat{G} = \frac{1}{t}\sum_{s=1}^{t}G(\hat{\theta}_0; x_s). \tag{A.24}$$

Let us define the matrix E as

$$E = \frac{1}{t}\sum_{s=1}^{t}G(\theta^*; x_s) - G(\theta^*, q). \tag{A.25}$$

Noting that the probability density of the input data $x_1, \ldots, x_t$ is $q(x)$, we obtain the asymptotic distribution of the matrix E as follows. When the matrix E is rearranged as column vector, the distribution of the vector is asymptotically multinomial normal distribution with mean 0. Let E_{ab} be the element of the matrix E. The covariance between E_{ab} and E_{cd} is calculated as follows:

$$\mathrm{E}_{D_0}\{E_{ab}E_{cd}\} = \frac{1}{t}\sigma_{ab,cd} + o\left(\frac{1}{t}\right), \tag{A.26}$$

where $\sigma_{ab,cd}$ is defined as

$$\sigma_{ab,cd} = \int q(x)\{G(\theta^*;x)_{ab} - G(\theta^*,q)_{ab}\}\{G(\theta^*;x)_{cd} - G(\theta^*,q)_{cd}\}dx. \tag{A.27}$$

We expand $\hat{G}$ as follows

$$\hat{G} = G(\theta^*,q) + E + \sum_{i=1}^{d}(\hat{\theta}_0^i - \theta^{*i})\frac{\partial}{\partial\theta^i}G(\theta^*,q) + o_p\left(\frac{1}{\sqrt{t}}\right). \tag{A.28}$$

Moreover we expand $H(\hat{\theta}_0,\xi^*)^{-1}$ as follows:

$$\begin{aligned}\frac{\partial}{\partial\xi^u}&H(\hat{\theta}_0,\xi^*)^{-1} \\ &= \frac{\partial}{\partial\xi^u}H(\theta^*,\xi^*)^{-1} + \sum_{i=1}^{d}\frac{\partial^2}{\partial\theta^i\partial\xi^u}H(\theta^*,\xi^*)^{-1}(\hat{\theta}_0^i - \theta^{*i}) + o_p\left(\frac{1}{\sqrt{t}}\right).\end{aligned} \tag{A.29}$$

Then we obtain the asymptotic expansion of $\frac{\partial}{\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0,\xi^*)$ as follows:

$$\begin{aligned}\frac{\partial}{\partial\xi^u}&\mathrm{Tr}\hat{G}H(\hat{\theta}_0,\xi^*)^{-1} \\ &= \mathrm{Tr}\left\{G(\theta^*,q) + E + \sum_{i=1}^{d}\frac{\partial}{\partial\theta^i}G(\theta^*,q)(\hat{\theta}_0^i - \theta^{*i})\right\} \\ &\quad\times\left\{\frac{\partial}{\partial\xi^u}H(\theta^*,\xi^*)^{-1} + \sum_{j=1}^{d}\frac{\partial^2}{\partial\theta^j\partial\xi^u}H(\theta^*,\xi^*)^{-1}(\hat{\theta}_0^j - \theta^{*j})\right\} + o_p\left(\frac{1}{\sqrt{t}}\right) \\ &= \frac{\partial}{\partial\xi^u}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1} + \mathrm{Tr}E\frac{\partial}{\partial\xi^u}H(\theta^*,\xi^*)^{-1} \\ &\quad+\sum_{i=1}^{d}(\hat{\theta}_0^i - \theta^{*i})\frac{\partial^2}{\partial\theta^i\partial\xi^u}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1} + o_p\left(\frac{1}{\sqrt{t}}\right).\end{aligned} \tag{A.30}$$

Note that

$$\frac{\partial}{\partial\xi^u}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi)^{-1}|_{\xi=\xi^*} = 0$$

for $1 \le u \le k$ from the definition of ξ^*. From (A.30) we can calculate asymptotic variance of $\frac{\partial}{\partial\xi}\mathrm{Tr}\hat{G}H(\hat{\theta}_0,\xi^*)^{-1}$ as follows

$$\mathrm{E}_{D_0}\left\{\left(\frac{\partial}{\partial\xi^u}\mathrm{Tr}\hat{G}H(\hat{\theta}_0,\xi^*)^{-1} - \frac{\partial}{\partial\xi^u}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1}\right)\right. \tag{A.31}$$

$$\left(\frac{\partial}{\partial\xi^v}\mathrm{Tr}\hat{G}H(\hat{\theta}_0,\xi^*)^{-1} - \frac{\partial}{\partial\xi^v}\mathrm{Tr}G(\theta^*,q)H(\theta^*,\xi^*)^{-1}\right)\Bigg\}$$
$$= \frac{1}{t}\left\{\sum_{a,b,c,d=1}^{d}\sigma_{ab,cd}\frac{\partial\tilde{H}^{ab}}{\partial\xi^u}\frac{\partial\tilde{H}^{cd}}{\partial\xi^v} + \sum_{i,j=1}^{d}\tilde{G}^{ij}\Gamma_{iu}\Gamma_{jv}\right\} + o\left(\frac{1}{t}\right),$$

where let $\tilde{H}^{ab}$ and $\tilde{G}^{ij}$ be the elements of $H(\theta^*,\xi^*)^{-1}$ and $G(\theta^*,q)^{-1}$ respectively and let Γ_{iu} be

$$\Gamma_{iu} = \frac{\partial^2}{\partial\theta^i\partial\xi^u}\mathrm{Tr}G(\theta,q)H(\theta,\xi)^{-1}\Big|_{\theta=\theta^*,\xi=\xi^*}.$$

From (A.21), (A.22) and (A.31) we obtain α^{ij}.

To derive (A.31) we use $\mathrm{E}_{D_0}\{(\hat{\theta}_0^i - \theta^{*i})E_{ab}\} = o(1/t)$. Finally we prove it. From the definition of E_{ab}

(A.32) $\mathrm{E}_{D_0}\{(\hat{\theta}_0^i - \theta^{*i})E_{ab}\}$

$$= \frac{1}{t}\sum_{s=1}^{t}\mathrm{E}_{D_0}\{(\hat{\theta}_0^i(x_1,\ldots,x_t) - \theta^{*i})(G(\theta^*;x_s)_{ab} - G(\theta^*,q)_{ab})\}$$
$$= \mathrm{E}_{D_0}\{(\hat{\theta}_0^i(x_1,\ldots,x_t) - \theta^{*i})(G(\theta^*;x_1)_{ab} - G(\theta^*,q)_{ab})\}$$
$$= \mathrm{E}_{(x_1,y_1)}\{(G(\theta^*;x_1)_{ab} - G(\theta^*,q)_{ab})\mathrm{E}_{D_0\setminus(x_1,y_1)}\{\hat{\theta}_0^i(x_1,\ldots,x_t) - \theta^{*i}\}\}$$

where $\mathrm{E}_{(x_1,y_1)}$ and $\mathrm{E}_{D_0\setminus(x_1,y_1)}$ are the expectation with respect to the density $p(y_1 \mid x_1)q(x_1)$ and the density $\prod_{s=2}^{t}p(y_s \mid x_s)q(x_s)$ respectively. From brief calculation we obtain

$$\text{(A.33)}\quad \mathrm{E}_{D_0\setminus(x_1,y_1)}\{\hat{\theta}_0^i(x_1,\ldots,x_t) - \theta^{*i}\} = \frac{b^i}{t} + \frac{1}{t}\sum_{j=1}^{d}\tilde{G}^{ij}\frac{\partial}{\partial\theta^j}\log p(y_1 \mid x_1,\theta^*) + o\left(\frac{1}{t}\right)$$

where b^i is the bias term of the mle. By substituting (A.33) to (A.32) and using

$$\mathrm{E}_{(x_1,y_1)}\{(G(\theta^*;x_1)_{ab} - G(\theta^*,q)_{ab})b^i\} = 0,$$
$$\int p(y_1 \mid x_1,\theta^*)\frac{\partial}{\partial\theta}\log p(y_1 \mid x_1,\theta^*)dy_1 = 0$$

we find that the $O(1/t)$ term vanishes.

References

Belue, L. M., Bauer, K. W., Jr. and Ruck, D. W. (1997). Selecting optimal experiments for multiple output multilayer perceptrons, *Neural Computation*, **9**, 161–183.

Bishop C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, New York.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review, *Statist. Sci.*, **10**, 273–304.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*, Academic Press, New York.

Ford, I., Titterington, D. M. and Kitsos, C. P. (1989). Recent advances in nonlinear experimental designs, *Technometrics*, **31**, 49–60

Fukumizu, K. (1996). Active learning in multilayer perceptrons, *Advances in Neural Information Processing Systems* (ed. D. S. Touretzky, M. C Mozer and M. E. Hasselmo), **8**, 295–301, MIT Press, Cambridge, Massachusetts.

MacKay, D. (1992). Information-based objective function for active data selection, *Neural Computation*, **4**, 305–318

Pukelsheim F. (1993). *Optimal Design of Experiments*, Wiley, New York.

Silvey, D. S. (1980). *Optimal Design*, Monographs on Applied Probability and Statistics, Chapman and Hall, London.

Watkin, T. L. H. and Rau, A. (1992). Selecting examples for perceptrons, *Journal of Physics A: Mathematical and General*, **25**, 113–121.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.