統計数理 (2000)
第48巻 第1号 197-212

[原著論文]

重み付き最尤推定量の情報量規準を用いた能動学習アルゴリズムの提案

総合研究大学院大学* 金森敬文
統計数理研究所 下平英寿

(受付 1999年4月5日；改訂 1999年11月1日)

要　旨

システムの入力と出力とのあいだに成立する条件付き確率分布を，入力分布を適切に選択することにより，学習データから推定する問題を議論する．観測者がシステムへの入力を選択できるような推定方式を能動学習という．本論文では，学習をおこなうときに設定したモデルは，一般には間違っているという現実的な仮定を採用する．モデルが間違っているときに能動学習をおこなうと，最尤推定量は一般に一致性をもたず，$O(1)$の大きさで，Kullback-Leibler divergenceの意味で最適なパラメータからずれてしまう．そこで，最適パラメータへの一致性を回復するために，適切な重み関数を用いた重み付き最尤推定量を採用して，能動学習アルゴリズムを構成する．つぎに，学習データ数が有限個という状況では，一致性が保証された重み付き最尤推定量よりも良い推定量が存在することを指摘する．これを考慮して，情報量規準を用いて適切な推定量を選択するというアルゴリズムを提案する．また，簡単な数値実験をおこない，提案したアルゴリズムの有効性について考察する．

キーワード：能動学習，重み付き最尤推定，情報量規準，統計的漸近論，リスク．

1. はじめに

学習理論では，学習機械がある入出力関係を実現するように，学習データをもちいて訓練するという状況を想定している．学習データの入力xと出力yとの関係は，条件付き確率密度関数$p(y|x)$によって規定されているとする．このとき$p(y|x)$をシステムとよぶ．一方，学習機械には調整可能なパラメータがあり，そのパラメータをさまざまな値に設定することにより多様な入出力関係を実現することができる．学習機械を統計的に解釈するとd次元ユークリッド空間$\boldsymbol{R}^d$の部分集合Θにパラメータをもつ統計モデル$M=\{p(y|x,\theta):\theta\in\Theta\subset\boldsymbol{R}^d\}$として表現することができる．したがって学習とは，多数の入出力のペアからなる学習データ$D_T=\{(x_1,y_1),\ldots,(x_T,y_T)\}$から，システムをよく近似する条件付き分布を統計モデルMを用いてパラメトリック推定する問題として定式化できる．このとき現実的には設定したモデルはシステムを含まない．そこで本論文では，必ずしも$p(y|x)\in M$ではないというmisspecificationの状況について考察する．

通常システムはある環境のなかに設置されている．ここでいう環境とは，入力xがしたがう確率分布を指している．本論文では環境を確率分布$q(x)$として，$q(x)$は既知と仮定する．シス

* 数物科学研究科 統計科学専攻：〒106-8569 東京都港区南麻布 4-6-7.

テムの振舞いを学習した学習機械は，その環境のなかで入力に対して適切な値を出力することが要求される．したがってシステムが設置されている環境のもとで学習機械がシステムをよく近似できるように学習をおこなう必要がある．つまりその環境下ではほとんど現れないような入力に対しては少々間違った出力をしてもよいが，よく現れる入力に対しては，学習機械がシステムを精度よく実現することが求められている．そこで損失関数を，$q(x)$ で平均した Kullback-Leibler divergence

$$KL(p, p_\theta|q) \equiv \int q(x)\, p(y|x) \log \frac{p(y|x)}{p(y|x, \theta)}\, dydx \tag{1.1}$$

とする．モデルMのなかでシステムをもっともよく近似する分布 $p(y|x, \theta^*)$ は

$$KL(p, p_{\theta^*}|q) = \min_{\theta \in \Theta} KL(p, p_\theta|q) \tag{1.2}$$

を満たす．ここで，モデルが必ずしもシステムを含まない状況を考えているので，(1.2) で定義されるパラメータ θ^* は $q(x)$ に依存することに注意する．パラメータ θ^* を最適パラメータとよぶ．

従来の学習理論の枠組みでは，学習データの入力はシステムが設置されている環境から生成されているという枠組みを扱うことが多い．このような学習系を本論文では受動学習とよぶ．これに対して本論文であつかう能動学習では，観測者はシステムへの入力がしたがう分布を集合 $Q = \{r(x|\xi) : \xi \in \Xi \subset \boldsymbol{R}^k\}$ のなかから指定できるという状況を想定している．環境が質問の集合に含まれている場合には，能動学習の立場では観測者は推定に有利な入力を選択することができるので，受動学習にくらべてより精度の高い学習が可能になると考えられる．

本論文であつかう問題の設定をまとめると，以下のようになる：

システム：$p(y|x)$,
モデル：$M = \{p(y|x, \theta) : \theta \in \Theta \subset \boldsymbol{R}^d\}$,
$p(y|x) \notin M$,
環境：$q(x)$ は既知,
質問の集合：$Q = \{r(x|\xi) : \xi \in \Xi \subset \boldsymbol{R}^k\}$.

このような状況のもとでリスク $\boldsymbol{E}_{D_T}\{KL(p, p_{\hat{\theta}}|q)\}$ をできるだけ小さくするような，能動学習による推定量 $\hat{\theta}$ を構成することが目的である．ここで，$\boldsymbol{E}_{D_T}$ は学習データの分布に関する期待値とする．

観測者が入力を設定できるという状況での推定問題は，数理統計学の分野では，最適実験計画として古くから研究がおこなわれている (Fedorov (1972), Silvey (1980))．とくに線形回帰モデルに関する凸解析などを用いた体系的な理論は，すでに確立している．最適実験計画においては，入力の適切さを表すさまざまな規準のあいだの関係を調べることが，重要な問題の一つとされている．それらの結果は Equivalence Theorem とよばれ，いろいろな状況のもとで証明が与えられている．また，モデルが非線形のときの実験計画では，局所最適性という規準のもとで最適な入力を求めるという研究がおこなわれている．

ニューラルネットワークにおける階層型パーセプトロンのような，非線形性の強いモデルに関しては，最適な入力を求めることは非常に困難である場合が多い．Mackay (1992) は Bayes 推定の観点から，非線形回帰モデルのパラメータの事後確率分布を正規近似することにより効率的な能動学習の方法を提案している．また，Fukumizu (1996) は，Fisher 情報量が退化している場合に重点をおいて，能動学習を一次漸近論の立場から考察している．

ここで問題になるのは統計モデルの妥当性である．従来の最適実験計画や能動学習の研究で

は設定した統計モデルが正しいと仮定していることが多い．しかし実際にはモデルが正しいということはまれである．モデルが間違っているときに従来の能動学習法を用いると，学習データ数が無限大の極限においても最尤推定量は最適パラメータに収束しない．したがって，モデルが間違っているような状況でも適用可能な能動学習の手法の開発は重要な課題である．そこで本論文では，重み付き最尤推定量による能動学習法を提案する．重み付き最尤推定量を用いることで，モデルが間違っている状況でも最適パラメータへ収束する能動学習法を構成することができる．

本論文の構成を以下に示す．2節で重み付き最尤推定量を用いた能動学習について説明する．3節では，重みを情報量規準により適応的に選択する推定方式について解説する．4節において，2，3節で構成した能動学習アルゴリズムの有効性を数値実験によって検証する．5節で，本論文で提案したアルゴリズムについての考察をおこなう．

2. 重み付き最尤推定量を用いた能動学習

T個の独立な学習データ D_T から，(1.2) を満たす最適パラメータ θ^* を推定することを考える．(1.1) において，モデルのパラメータに依存する部分は

$$-\int p(y|x)\, q(x) \log p(y|x, \theta)\, dydx \tag{2.1}$$

の項である．ここで推定量として最尤推定量

$$\hat{\theta}_{mle} = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \log p(y_t|x_t, \theta) \tag{2.2}$$

を用いるとする．学習データ D_T の入力 $x_1, \ldots, x_T$ が，環境 $q(x)$ の分布にしたがって独立に生成されているときには，大数の法則から，

$$\frac{1}{T} \sum_{t=1}^{T} \log p(y_t|x_t, \theta) \longrightarrow \int p(y|x)\, q(x) \log p(y|x, \theta)\, dydx, \qquad (T \longrightarrow \infty)$$

となる．したがって $\hat{\theta}_{mle}$ は最適パラメータに確率収束することがわかる．しかし学習データ D_T の入力が質問分布 $r(x)$ にしたがっているときには

$$\frac{1}{T} \sum_{t=1}^{T} \log p(y_t|x_t, \theta) \longrightarrow \int p(y|x)\, r(x) \log p(y|x, \theta)\, dydx, \qquad (T \longrightarrow \infty)$$

となる．したがってモデルが間違っている状況では必ずしも $\hat{\theta}_{mle}$ は θ^* に確率収束せず，最尤推定量は良い推定量とは言えない．能動学習では，学習データを生成するための入力分布は一般に環境 $q(x)$ とは異なる分布を使用する．このため能動学習においては最尤推定量を補正する必要がある．そこで以下のような重み付き最尤推定量

$$\hat{\theta}_w = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} w(x_t) \log p(y_t|x_t, \theta) \tag{2.3}$$

を考える．ここで，重み $w(x)$ が任意の $\theta \in \Theta$ に対して

$$\int p(y|x)\, r(x)\, w(x) \log p(y|x, \theta)\, dydx = \int p(y|x)\, q(x) \log p(y|x, \theta)\, dydx \tag{2.4}$$

を満たしているなら，入力分布 $r(x)$ を用いて生成した学習データに対して，重み付き最尤推定

量 $\hat{\theta}_w$ は最適パラメータ θ^* に確率収束する．ここで重みを $w(x)=w_1(x)\equiv q(x)/r(x)$ とすれば (2.4) を満たす．以上の結果，環境とは異なる入力分布 $r(x)$ を用いて学習データを生成したときには，重みを $w_1(x)$ とすれば良いことがわかる．

つぎに重み付き推定量のリスクを計算する．入力分布を $r(x|\xi)\in Q$ として，$q(x)/r(x|\xi)$ を重みとする重み付き最尤推定量を用いて推定を行うときのリスクは漸近的に

$$\boldsymbol{E}_{D_T}\{KL(p,\ p_{\hat{\theta}(D_T)}|q)\}=KL(p,\ p_{\theta*}|q)+\frac{1}{2T}\mathrm{Tr}H^{-1}K(\xi)+O\left(\frac{1}{T\sqrt{T}}\right) \tag{2.5}$$

となる (付録 A 参照)．ここで H，$K(\xi)$ は $d\times d$ の行列であり，それらの要素は分布 $p(y|x)\cdot q(x)$ による期待値 $\boldsymbol{E}_{pq}$ を用いて

$$H_{ij}=-\boldsymbol{E}_{pq}\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y|x,\ \theta^*)\right\} \tag{2.6}$$

$$K(\xi)_{ij}=\boldsymbol{E}_{pq}\left\{\frac{q(x)}{r(x|\xi)}\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta^*)\frac{\partial}{\partial\theta_j}\log p(y|x,\ \theta^*)\right\} \tag{2.7}$$

と定義される．以上の結果からリスクを漸近的にもっとも小さくする入力分布は

$$\mathrm{Tr}H^{-1}K(\xi^*)=\min_{\xi\in\Xi}\mathrm{Tr}H^{-1}K(\xi) \tag{2.8}$$

を満たす $r(x|\xi^*)$ であることがわかる．

実際に能動学習をおこなうときには，ξ^* を推定する必要がある．これを考慮して，つぎの能動学習アルゴリズム (Active Learning Algorithm) $ALA\ (s)$ を提案する．

$ALA\ (s)$

Step 1. 環境 $q(x)$ を入力として，s 個の学習データを収集する．ただし，$s=o(T)$, $\lim_{T\to\infty}s=\infty$ とする．このとき収集するデータを $D_0=\{(x_1,\ y_1),\ \ldots,\ (x_s,\ y_s)\}$ とおく．推定量 $\hat{\theta}_{D_0}$ を

$$\hat{\theta}_{D_0}=\arg\max_{\theta\in\Theta}\sum_{t=1}^{s}\log p(y_t|x_t,\ \theta)$$

とする．さらに，

$$\hat{H}_{ij}=-\frac{1}{s}\sum_{t=1}^{s}\frac{\partial^2\log p(y_t|x_t,\ \hat{\theta}_{D_0})}{\partial\theta_i\partial\theta_j},$$

$$\hat{K}(\xi)_{ij}=\frac{1}{s}\sum_{t=1}^{s}\frac{q(x_t)}{r(x_t|\xi)}\cdot\frac{\partial\log p(y_t|x_t,\ \hat{\theta}_{D_0})}{\partial\theta_i}\frac{\partial\log p(y_t|x_t,\ \hat{\theta}_{D_0})}{\partial\theta_j}$$

とおいて，

$$\hat{\xi}=\arg\min_{\xi\in\Xi}\mathrm{Tr}\hat{H}^{-1}\hat{K}(\xi)$$

を求める．

Step 2. 入力分布を質問 $r(x|\hat{\xi})$ として，$T-s$ 個の学習データを生成する．このとき得られたデータを $D_1=\{(x_{s+1},\ y_{s+1}),\ \ldots,\ (x_T,\ y_T)\}$ とおく．

Step 3. 観測されたすべてのデータ $D_T=D_0\cup D_1$ から

$$\hat{\theta}_{act}=\arg\max_{\theta\in\Theta}\left\{\sum_{i=1}^{s}\log p(y_i|x_i,\ \theta)+\sum_{j=s+1}^{T}\frac{q(x_j)}{r(x_j|\hat{\xi})}\log p(y_j|x_j,\ \theta)\right\} \tag{2.9}$$

により，最適パラメータの推定値 $\hat{\theta}_{act}$ を求める．□

$ALA(s)$ の推定量 $\hat{\theta}_{act}$ のリスクの高次項を統計的漸近論にしたがって計算することにより，前半の学習データのバッチサイズ s を最適化することが可能である．その結果 $s=O(\sqrt{T})$ とすればリスクの高次項のオーダーがもっとも小さくなることを示すことができる（金森(1999)）．本節で構成した能動学習のアルゴリズムは，学習データ数が無限大の極限において推定量が最適パラメータへ収束することが保証されている．

3. 重み付き最尤推定量の情報量規準を用いた能動学習

前節において，入力分布が $r(x)$ のとき重みとして $q(x)/r(x)$ を用いて推定量を構成すれば推定量は最適パラメータに収束することを利用して，能動学習のアルゴリズムを構成した．しかし実際には学習データ数は有限なので，$q(x)/r(x|\xi)$ という重みよりもリスクを小さくするような重みが存在することがある．したがって重みを適切に選択することにより，より精度の高い推定量を構成することができる．本節において，情報量規準を用いて適切な重みを選択する方法について解説する．

任意の重み関数のなかから適切な重みを選択することは，学習データ D_T の入力 $(x_1, ..., x_T)$ に対して，$(w(x_1), ..., w(x_T)) \in \boldsymbol{R}^T$ の値を選択することと同値である．しかし，T 個の学習データから $(w(x_1), ..., w(x_T))$ を選択するとばらつきが大きくなってしまいリスクの意味で性能が悪くなってしまう．また，計算量の観点からみても現実的ではない．

そこで，重みに一次元のパラメータ λ を導入して，$w_\lambda(x) \equiv (q(x)/r(x))^\lambda$ と定義する．重み $w_\lambda(x)$ を用いた重み付き最尤推定量を $\hat{\theta}_\lambda$ とおく．ここで $\hat{\theta}_0$ は最尤推定量であり，システムがモデル M のなかに含まれている場合には精度が高い推定量であることが知られている．また $\hat{\theta}_1$ は一致性が保証された重み付き推定量である．したがって，重みが1の場合と $q(x)/r(x)$ の場合をパラメータ λ でつなぐことには意味がある．しかし $w_\lambda(x)$ のようなつなぎ方には理論的な根拠はない．

このような推定量の集合をあらかじめ用意しておいて，システムがモデルに近いときには λ を0に近い値に設定し，システムがモデルから離れているときには λ を1に近い値に設定するようにすれば，リスクが小さくなると期待できる．そこで，情報量規準を用いて学習データから最適な λ を選択するという推定方式を考える．

ここで情報量規準について解説する．モデルと推定量を指定すると，その推定量のリスクを計算することができる．このリスクの推定量を情報量規準とよぶ．重み付き最尤推定量 $\hat{\theta}_\lambda$ の情報量規準 IC_λ は

$$\boldsymbol{E}_{D_T}\{IC_\lambda\} = -\boldsymbol{E}_{D_T}\{\boldsymbol{E}_{pq}\log p(y|x,\ \hat{\theta}_\lambda)\} + o\left(\frac{1}{T}\right) \tag{3.1}$$

をみたすように構成される．具体的に

$$IC_\lambda = -\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t)}\log p(y_t|x_t,\ \hat{\theta}_\lambda) + \frac{1}{T}\mathrm{Tr}\hat{H}_{\lambda,r}(\hat{\theta}_\lambda)^{-1}\hat{K}_{\lambda,r}(\hat{\theta}_\lambda) \tag{3.2}$$

と定義すれば (3.1) をみたす（記号の定義や導出は付録B参照）．情報量規準により，さまざまな推定量のあいだのリスクの大小関係を比較することができる．そこで，情報量規準 IC_λ を推定量を選択するための規準として採用する．学習データ D_T が $p(y|x)\,r(x)$ にしたがって生成されているとき

$$IC_{\hat{\lambda}} = \min_{\lambda \in \boldsymbol{R}} IC_{\lambda} \tag{3.3}$$

を満たす $\hat{\lambda}$ を求め，推定量として $\hat{\theta}_{\hat{\lambda}}$ を用いる．

前節で構成した $ALA(s)$ に，さらに情報量規準 IC_{λ} を適用した以下の能動学習アルゴリズム $IC\text{-}ALA(s)$ を提案する．

$IC\text{-}ALA(s)$

Step 1，*Step* 2 は $ALA(s)$ と同じ．

Step 3. 観測されたすべてのデータ $D_T = D_0 \cup D_1$ を用いた推定量 $\hat{\theta}_{act,\lambda}$ を

$$\hat{\theta}_{act,\lambda} = \arg\max_{\theta \in \boldsymbol{\Theta}} \left\{ \sum_{i=1}^{s} \log p(y_i|x_i,\ \theta) + \sum_{j=s+1}^{T} \left(\frac{q(x_j)}{r(x_j|\hat{\xi})} \right)^{\lambda} \log p(y_j|x_j,\ \theta) \right\} \tag{3.4}$$

と定義する．行列 $\hat{H}_{\lambda,\hat{\xi}}$ と $\hat{K}_{\lambda,\hat{\xi}}$ をそれぞれ

$$(\hat{H}_{\lambda,\hat{\xi}})_{ij} = -\frac{1}{T} \sum_{t=s+1}^{T} \left(\frac{q(x_t)}{r(x_t|\hat{\xi})} \right)^{\lambda} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(y_t|x_t,\ \hat{\theta}_{act,\lambda}), \tag{3.5}$$

$$(\hat{K}_{\lambda,\hat{\xi}})_{ij} = \frac{1}{T} \sum_{t=s+1}^{T} \left(\frac{q(x_t)}{r(x_t|\hat{\xi})} \right)^{\lambda+1} \frac{\partial}{\partial\theta_i} \log p(y_t|x_t,\ \hat{\theta}_{act,\lambda}) \frac{\partial}{\partial\theta_j} \log p(y_t|x_t,\ \hat{\theta}_{act,\lambda}) \tag{3.6}$$

として，情報量規準

$$IC_{act,\lambda} = -\sum_{i=1}^{s} \log p(y_i|x_i,\ \hat{\theta}_{act,\lambda}) - \sum_{j=s+1}^{T} \frac{q(x_j)}{r(x_j|\hat{\xi})} \log p(y_j|x_j,\ \hat{\theta}_{act,\lambda}) + \mathrm{Tr}\hat{H}_{\lambda,\hat{\xi}}^{-1} \hat{K}_{\lambda,\hat{\xi}} \tag{3.7}$$

を定義する．$IC_{act,\lambda}$ を最小にするような λ を $\hat{\lambda}$ とおいて，$\hat{\theta}_{act,\hat{\lambda}}$ を最適パラメータの推定値とする．

ここで，能動学習における情報量規準 $IC_{act,\lambda}$ は，正確には推定量 $\hat{\theta}_{act,\lambda}$ の情報量規準にはなっていないことに注意する．これは，前半の学習データ D_0 と後半の学習データ D_1 との間には，能動学習による相関が生じるので，全学習データが独立にならないためである．この相関を考慮して，正確な情報量規準を計算することは可能であるが，本論文では，上記の $IC_{act,\lambda}$ を用いることにする．

例 3.1. 多項式回帰モデルの場合の推定量と情報量規準を求める．モデルを

$$y = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \epsilon, \qquad \epsilon \sim N(0,\ \sigma^2) \tag{3.8}$$

とする．$N(\mu,\ \sigma^2)$ は平均 μ，分散 σ^2 の正規分布とする．パラメータは $\theta = (\beta_0,\ \beta_1,\ \ldots,\ \beta_d,\ \sigma^2)$ である．$y = (y_0,\ y_1)$ を目的変数の T 次元観測ベクトルとして，$X' = (X_0',\ X_1')$ を説明変数の観測行列とする．X' は行列 X の転置を意味する．また y_0，y_1 と X_0，X_1 はそれぞれ能動学習における学習データ D_0，D_1 に対応している．さらに $((q(x_{s+1})/r(x_{s+1}))^{\lambda}, \ldots, (q(x_T)/r(x_T))^{\lambda})$ を対角要素にもつ対角行列を $W_1^{(\lambda)}$ とする．このとき (3.4) は

$$\hat{\beta}_{\lambda} = (X_0'X_0 + X_1'W_1^{(\lambda)}X_1)^{-1}(X_0'y_0 + X_1'W_1^{(\lambda)}y_1) \tag{3.9}$$

$$\hat{\sigma}_\lambda^2 = \frac{1}{s + \mathrm{Tr}\, W_1^{(\lambda)}} \{(y_0 - X_0\hat{\beta}_\lambda)'(y_0 - X_0\hat{\beta}_\lambda) + (y_1 - X_1\hat{\beta}_\lambda)' W_1^{(\lambda)} (y_1 - X_1\hat{\beta}_\lambda)\} \tag{3.10}$$

となる．$e_t, t = 1, \ldots, T$ を残差ベクトル，h_t を重み付き最尤推定量に用いられるハット行列の対角要素とすると $IC\text{-}ALA(s)$ における情報量規準は

$$T \cdot IC_{act,\lambda} = \frac{1}{2\hat{\sigma}_\lambda^2} \sum_{D_0} e_t^2 + \frac{1}{2\hat{\sigma}_\lambda^2} \sum_{D_1} \frac{q(x_t)}{r(x_t|\hat{\xi})} e_t^2 + \frac{s + \mathrm{Tr}\, W_1^{(\lambda)}}{2} \log 2\pi\hat{\sigma}_\lambda^2 \tag{3.11}$$

$$+ \sum_{D_1} \frac{q(x_t)}{r(x_t|\hat{\xi})} \frac{e_t^2}{\hat{\sigma}_\lambda^2} h_t + \frac{1}{2} \sum_{D_1} \frac{(q(x_t)/r(x_t|\hat{\xi}))^\lambda}{\mathrm{Tr}\, W_1^{(\lambda)}} \frac{q(x_t)}{r(x_t|\hat{\xi})} \left(\frac{e_t^2}{\hat{\sigma}_\lambda^2} - 1 \right)^2 \tag{3.12}$$

となる．(3.11) が重み付き尤度の項であり (3.12) が補正項である．

4. 数値実験

本節では，2節と3節で提案した能動学習アルゴリズムの計算機実験をおこなう．実験の設定を以下に示す．$\phi(x)$ は標準正規分布の密度関数をあらわすとする．

例 4.1. 最初に能動学習の受動学習に対する有効性を示すために，環境 $q(x)$ が質問分布の集合 Q に含まれる場合について計算機実験をおこなう．設定を以下に示す．

システム：$p(y|x) = \frac{1}{0.3}\phi\left(\frac{y - s_\delta(x)}{0.3}\right)$，　$s_\delta(x) = 1 - x + x^2 + \delta x^3$,

環境：$q(x) = \frac{1}{0.4}\phi\left(\frac{x - 0.2}{0.4}\right)$,

モデル：$p(y|x, \theta) = \frac{1}{\sigma}\phi\left(\frac{y - \theta_1 - \theta_2 x - \theta_3 x^2}{\sigma}\right)$，　パラメータは $\theta = (\theta_1, \theta_2, \theta_3, \sigma)$,

質問の集合：$r(x|\xi) = \frac{1}{\xi}\phi\left(\frac{x - 0.2}{\xi}\right)$，　$\xi > 0$.

環境 $q(x)$ が質問分布の集合に含まれているときに，能動学習のほうが受動学習よりもリス

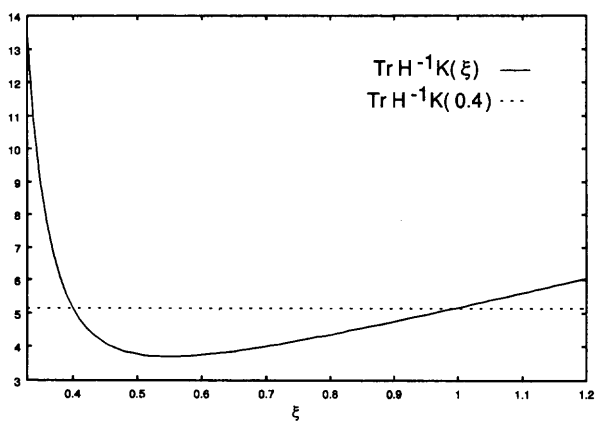

図1．　$\mathrm{Tr}H^{-1}K(\xi)$ のプロット．

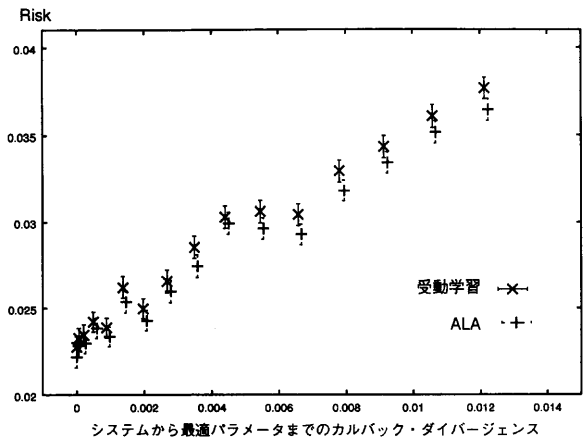

図2. 受動学習と能動学習のリスクの比較. 学習データ数は100. 能動学習には $ALA(30)$ をもちいた. エラーバーは標準誤差を表している.

クが小さくなる場合があることは, $\mathrm{Tr}H^{-1}K(\xi)$ を計算することからわかる. この例の場合には図1のようになる. 最適な質問分布は $\xi^* \doteqdot 0.55$ となり, $q(x)$ とは異なることが数値的に確認できる. 質問分布のパラメータが $\xi = 0.4$ のときには $q(x) = r(x|0.4)$ であるから, 能動学習と受動学習とは同じ入力分布を用いていることになり, リスクは同じである. また質問分布のパラメータ ξ がおよそ $0.4 \leq \xi \leq 1.0$ の範囲にあるときには, そのような質問分布を用いた能動学習は受動学習よりもリスクが小さくなっている.

学習データ数は $T = 100$ とする. 実験では, システムを指定するパラメータ δ をひとつ固定して, 学習を1000回おこないリスクを推定する. 能動学習をおこなうときには $ALA(30)$ をもちいた.

最初の30個の学習データは受動学習と能動学習で同じデータをもちいた. これを, δ を少しずつ変えていったときのグラフを図2に示す. 図2では横軸をシステムから最適パラメータまでのカルバック・ダイバージェンス $KL(p_\delta, p_{\theta^*}|q)$ としている. 実験の結果から能動学習は受動学習を常にリスクの意味で改良していることがわかる.

さらに, 情報量規準を用いる能動学習について数値実験をおこなう. ここでは, $ALA(30)$ と $IC\text{-}ALA(30)$ を比較する. また, 最尤推定量を用いた能動学習アルゴリズムとも比較する. 最尤推定量を用いたアルゴリズムとは, $IC\text{-}ALA$ の Step 3 において, $\lambda = 0$ として最尤推定量による推定をおこなう方式のことである. ALA と $IC\text{-}ALA$ のどちらのアルゴリズムを用いても得られる学習データは同じであることに注意する. 得られた学習データに対して, ALA と $IC\text{-}ALA$ のそれぞれを用いてパラメータの推定をおこなう. ALA で得られた推定値の損失から $IC\text{-}ALA$ で得られた推定値の損失を引いて, 両者のアルゴリズムの損失の差を測る. これを各 δ に対して1000回繰り返して, その結果を標準誤差とともにプロットする. また同様にして, ALA で得られた推定値の損失から, 最尤推定量を用いたアルゴリズムから得られた推定値の損失を引いた量を, 両者のアルゴリズムの差としてプロットする (図3). さらに $IC\text{-}ALA$ で推定をおこなったときの重み関数のパラメータ λ の分布をプロットする (図4).

まず図3をみると, システムがモデルに近い状況では, $IC\text{-}ALA$ は ALA をリスクの意味で改良していることがわかる. また, システムがモデルから離れている状況では, 両者のアルゴリズムにはほとんど差がないことがわかる. これより情報量規準をもちいた能動学習のアルゴリズムは, システムとモデルとの位置関係に対して適応的に振舞うことが示唆される. また,

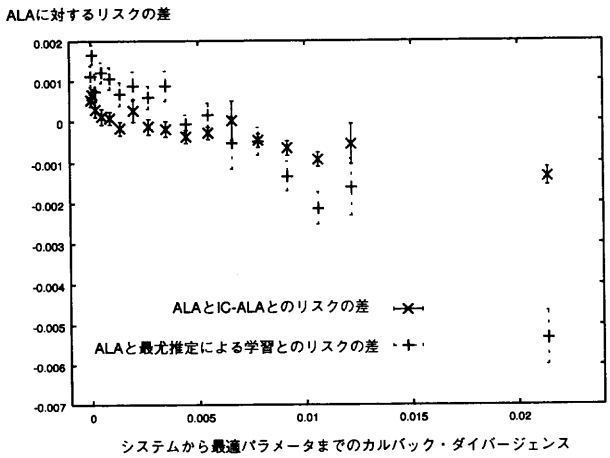

図3. リスクの差のプロット：*ALA* のリスクから *IC-ALA* のリスクを引いた量と，*ALA* のリスクから最尤推定量による学習のリスクを引いた量をそれぞれプロットした．横軸はシステムから最適パラメータまでのカルバック・ダイバージェンス．エラーバーは標準誤差を表している．

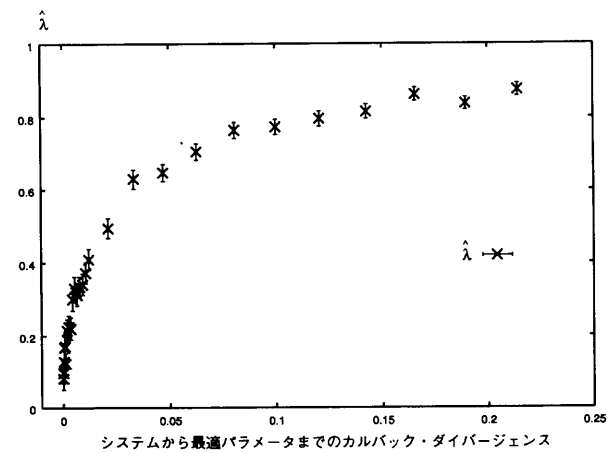

図4. *IC-ALA* の Step 3 における $\hat{\lambda}$ の期待値をプロットしたグラフ．横軸はシステムから最適パラメータまでのカルバック・ダイバージェンス．エラーバーは標準誤差を表している．

最尤推定量を用いた方法は，システムがモデルから乖離するにしたがって，*ALA* に対する性能が悪くなっていく．これは，最尤推定量は最適パラメータに収束しないことを示している．この結果から，重み関数で尤度を補正した推定量はモデルが間違っているという現実的な状況で有効であると考えられる．

また図4から，システムがモデルから離れるにしたがって，$\hat{\lambda}$ の値は1に近づいていく様子がわかる．

例 4.2. 環境 $q(x)$ にしたがう学習データを利用するのに非常にコストがかかる場合を想定して，質問の集合に環境が含まれない設定での能動学習について実験をおこなう．ここでは，例 4.1 で質問の集合を

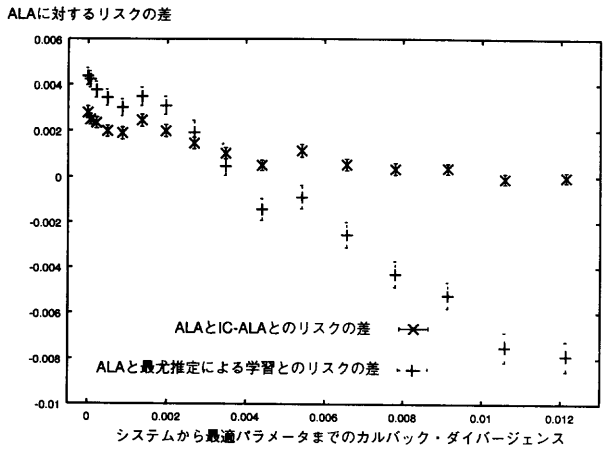

図5. リスクの差のプロット：*ALA* のリスクから *IC-ALA* のリスクを引いた量と，*ALA* のリスクから最尤推定量による学習のリスクを引いた量をそれぞれプロットした．横軸はシステムから最適パラメータまでのカルバック・ダイバージェンス．エラーバーは標準誤差を表している．

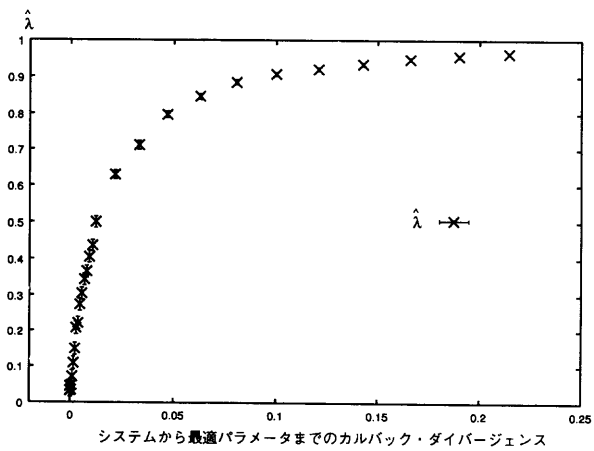

図6. *IC-ALA* の Step 3 における $\hat{\lambda}$ の期待値をプロットしたグラフ．横軸はシステムから最適パラメータまでのカルバック・ダイバージェンス．エラーバーは標準誤差を表している．

$$r(x|\xi)=\frac{1}{\xi}\phi\left(\frac{x}{\xi}\right),\qquad \xi>0,$$

とする．この場合には $q(x)\notin Q$ である．

ALA(30)，*IC-ALA*(30)，最尤推定量を用いた能動学習アルゴリズムを例 4.1 と同じように比較する．この場合にも，*IC-ALA* が *ALA* を改良していることがわかる（図5）．また図6から，システムがモデルから離れるにしたがって $\hat{\lambda}$ の値が1に近づいていくことも確認できる．また，$\delta=0.5$ のときの，ある学習データから計算された $IC_{act,\lambda}$ を図7に示す．図7の重み付き尤度は $\lambda=1$ で最小値をとるが，バイアス補正のために $\hat{\lambda}\doteqdot 0.76$ となっている．

図8では，従来の実験計画法と比較している．数値実験では多項式回帰モデルをあつかって

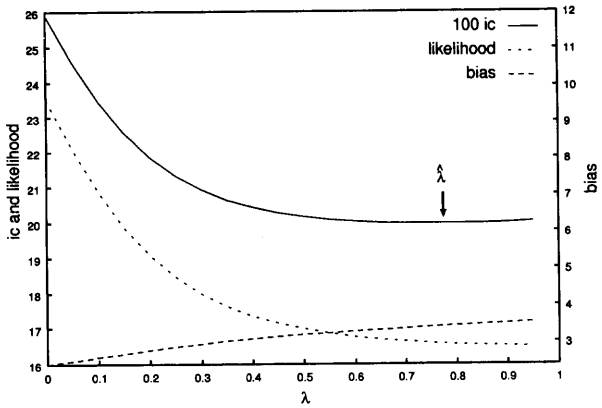

図7.　$100 \times IC_{act,\lambda}$ のグラフ．likelihood は (3.11) のプロットであり，bias は (3.12) のプロットである．

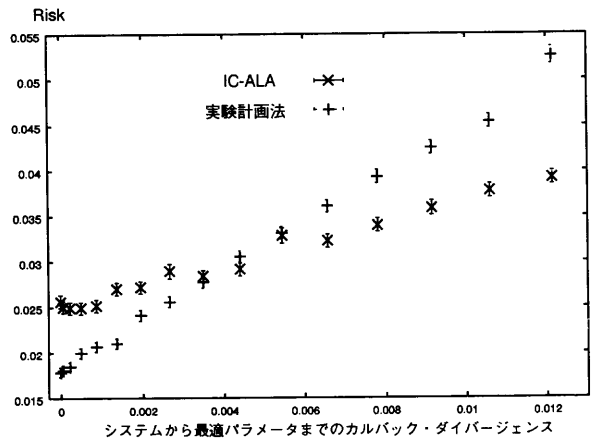

図8.　リスクのプロット：*IC-ALA*，通常の実験計画法のそれぞれのリスクを標準誤差とともにプロットしたグラフ．横軸はシステムから最適パラメータまでのカルバック・ダイバージェンス．エラーバーは標準誤差を表している．

いるので，モデルが正しい場合の最適な入力を学習データを得る前に求めることができる．これにより，質問分布の標準偏差 ξ が大きいほど，モデルが正しいときの推定精度が良くなることがわかる．図8では，ξ として $\xi^* = 0.54$ を用いたときの従来の実験計画法のリスクをプロットした．δ が小さくシステムがモデルに近いときには，実験計画法にしたがって最尤推定量を用いた方法がリスクを小さくしている．しかし，システムがモデルから離れるにしたがって最尤推定量を用いた実験計画法はリスクが急激に増加していく様子が観察できる．重み付き最尤推定量を用いた方法はシステムがモデルから離れていても実験計画法ほどリスクは大きくならず，ロバストであるといえる．

5. むすび

本論文では，学習理論における能動学習について解説した．モデルが間違っているときには，

最尤推定量を用いた能動学習では，一般には最適なパラメータへの一致性が保証されない．そこで，適切な重み関数で最尤推定量を補正することにより，一致性が保証されることを示して，重み付き最尤推定量を用いた能動学習のアルゴリズム *ALA* を構成した．

ここで実際の状況を考えてみる．一致性が保証されている推定量は，データ数が無限大のときには好ましいが，データ数が有限という現実的な状況では必ずしも最良の重み付き推定量というわけではない．そのことを考慮して，学習データから適切な重み関数を決定するために，重み付き最尤推定量の情報量規準を構成して，その情報量規準を用いた能動学習アルゴリズム *IC-ALA* を構成した．能動学習のアルゴリズム *IC-ALA* は，*ALA* の推定量を学習データに対して推定量を適応的に選択できるように改良したものである．

最適実験計画の分野では実験の設計のためのさまざまな基準が提案されているが，ここでは学習理論におけるリスクを実験設計の基準として用いた．それにより学習理論や推定論で発展しているさまざまな手法を導入することが可能になる．本論文で提案した情報量規準を用いる方法も，一般的な推定論や学習理論の方法を実験の設計に応用したものである．一般的に，情報量規準はモデル選択という文脈で現れることが多いが，ここでは推定量を選択するための規準として用いている．本論文で示したような情報量規準の使い方も，学習に柔軟性を持たせる意味で非常に有効であると考えられる．

謝　　辞

本論文の作成にあたり，統計数理研究所の江口真透教授，査読者の方々から，大変有用かつ貴重なご意見をいただきました．ここに記して，謝意を表します．

付　　録

A. (2.5) の証明

学習データ $D_T = \{(x_1, y_1), \ldots, (x_T, y_T)\}$ が $p(y|x)\, r(x|\xi)$ から独立に生成されているとする．$q(x)/r(x|\xi)$ を重みとする重み付き最尤推定量 $\hat{\theta}_w$ の推定方程式は

$$\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t|\xi)}\frac{\partial}{\partial\theta_i}\log p(y_t|x_t,\ \hat{\theta}_w) = 0, \qquad i = 1, \ldots, d \tag{A.1}$$

となる．(A.1) を θ^* のまわりで展開して高次項を無視すると，

$$\begin{aligned} &-\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t|\xi)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y_t|x_t,\ \theta^*)(\hat{\theta}_w - \theta^*)_j \\ &\quad = \frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t|\xi)}\frac{\partial}{\partial\theta_i}\log p(y_t|x_t,\ \theta^*) + O_p\left(\frac{1}{\sqrt{T}}\right) \end{aligned} \tag{A.2}$$

と評価できる．ここで，パラメータ θ の添え字に関しては，アインシュタインの総和の規約を適用している．さらに

$$-\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t|\xi)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y_t|x_t,\ \theta^*) = H_{ij} + O_p\left(\frac{1}{\sqrt{T}}\right) \tag{A.3}$$

と近似できる．また (A.2) の右辺は漸近的に平均 0，分散 $K(\xi)/T$ の正規分布にしたがう．よって，推定量 $\hat{\theta}_w$ の分散は漸近的に

(A.4)　$$\boldsymbol{E}_{D_T}\{(\hat{\theta}_w-\theta^*)_i(\hat{\theta}_w-\theta^*)_j\}=\frac{1}{T}(H^{-1}K(\xi)H^{-1})_{ij}+O\left(\frac{1}{T\sqrt{T}}\right)$$

となる．さらに，損失 $KL(p,\ p_{\hat{\theta}_w}|q)$ を θ^* のまわりで展開すると，リスクは以下のように書ける：

(A.5)　$$\begin{aligned}&\boldsymbol{E}_{D_T}\{KL(p,\ p_{\hat{\theta}_w}|q)\}\\&\quad=KL(p,\ p_{\theta^*}|q)+\frac{1}{2}\mathrm{Tr}H\boldsymbol{E}_{D_T}\{(\hat{\theta}_w-\theta^*)(\hat{\theta}_w-\theta^*)'\}+O\left(\frac{1}{T\sqrt{T}}\right).\end{aligned}$$

ここで $(\hat{\theta}_w-\theta^*)'$ は，縦ベクトル $(\hat{\theta}_w-\theta^*)$ の転置を意味する．(A.4) を (A.5) に代入すると，(2.5) がえられる．

B. (3.2) の証明

学習データ数が十分多ければ，推定量 $\hat{\theta}_\lambda$ は

(B.1)　$$\theta_\lambda^*=\arg\max_{\theta\in\Theta}\boldsymbol{E}_{pq}\left\{\left(\frac{q(x)}{r(x)}\right)^{\lambda-1}\log p(y|x,\ \theta)\right\}$$

を満たす θ_λ^* に十分近い．そこで，$\boldsymbol{E}_{D_T}\boldsymbol{E}_{pq}\{-\log p(y|x,\ \hat{\theta}_\lambda(D_T))\}$ を θ_λ^* のまわりで展開して，高次項を無視すると

(B.2)　$$\begin{aligned}&\boldsymbol{E}_{D_T}\boldsymbol{E}_{pq}\{-\log p(y|x,\ \hat{\theta}_\lambda(D_T))\}\\&\quad=-\boldsymbol{E}_{pq}\{\log p(y|x,\ \theta_\lambda^*)\}-\boldsymbol{E}_{pq}\left\{\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\right\}\boldsymbol{E}_{D_T}\{(\hat{\theta}_\lambda-\theta_\lambda^*)_i\}\\&\qquad+\frac{1}{2}\boldsymbol{E}_{pq}\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y|x,\ \theta_\lambda^*)\right\}\boldsymbol{E}_{D_T}\{(\hat{\theta}_\lambda-\theta_\lambda^*)_i(\hat{\theta}_\lambda-\theta_\lambda^*)_j\}+o\left(\frac{1}{T}\right)\end{aligned}$$

となる．入力 $r(x)$ を用いて，推定量 $\hat{\theta}_\lambda$ で推定をおこなう状況を考える．このとき (B.2) の左辺の推定量としてもっとも単純なものは

(B.3)　$$L(\hat{\theta}_\lambda,\ D_T)\equiv-\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t)}\log p(y_t|x_t,\ \hat{\theta}_\lambda)$$

であろう．だが，$L(\hat{\theta}_\lambda, D_T)$ は $O(1/T)$ のバイアスをもつので，(3.1) をみたさない．そこで，$L(\hat{\theta}_\lambda,\ D_T)$ のバイアスを補正して情報量規準を構成する．推定量 $\hat{\theta}_\lambda$ を θ_λ^* のまわりで展開して $\boldsymbol{E}_{D_T}\{L(\hat{\theta}_\lambda,\ D_T)\}$ を計算すると

(B.4)　$$\begin{aligned}&\boldsymbol{E}_{D_T}\{L(\hat{\theta}_\lambda,\ D_T)\}\\&\quad=-\boldsymbol{E}_{pq}\{\log p(y|x,\ \theta_\lambda^*)\}+\boldsymbol{E}_{D_T}\left\{\frac{\partial}{\partial\theta_i}L(\theta_\lambda^*,\ D_T)(\hat{\theta}_\lambda-\theta_\lambda^*)_i\right\}\\&\qquad+\frac{1}{2}\boldsymbol{E}_{D_T}\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\theta_\lambda^*,\ D_T)(\hat{\theta}_\lambda-\theta_\lambda^*)_i(\hat{\theta}_\lambda-\theta_\lambda^*)_j\right\}+o\left(\frac{1}{T}\right)\end{aligned}$$

となる．さらに

(B.5)　$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\theta_\lambda^*,\ D_T)=\boldsymbol{E}_{pq}\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y|x,\ \theta_\lambda^*)\right\}+O_p\left(\frac{1}{\sqrt{T}}\right)$$

を用いると

(B.6)
$$\begin{aligned}&\boldsymbol{E}_{D_T}\boldsymbol{E}_{pq}\{-\log p(y|x,\ \hat{\theta}_\lambda(D_T))\}\\&=\boldsymbol{E}_{D_T}\{L(\hat{\theta}_\lambda,\ D_T)\}+\boldsymbol{E}_{D_T}\left\{\left(\boldsymbol{E}_{pq}\left\{-\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\right\}\right.\right.\\&\left.\left.-\frac{\partial}{\partial\theta_i}L(\theta_\lambda^*,\ D_T)\right)(\hat{\theta}_\lambda-\theta_\lambda^*)_i\right\}+o\left(\frac{1}{T}\right)\end{aligned}$$

と評価できる．(B.6) の右辺の第二項をさらに計算すると

(B.7)
$$\begin{aligned}&\boldsymbol{E}_{D_T}\boldsymbol{E}_{pq}\{-\log p(y|x,\ \hat{\theta}_\lambda(D_T))\}\\&=\boldsymbol{E}_{D_T}\{L(\hat{\theta}_\lambda,\ D_T)\}+\frac{1}{T}\mathrm{Tr}H_{\lambda,r}(\theta_\lambda^*)^{-1}K_{\lambda,r}(\theta_\lambda^*)+o\left(\frac{1}{T}\right)\end{aligned}$$

となる．ここで，$H_{\lambda,r}(\theta_\lambda^*)$，$K_{\lambda,r}(\theta_\lambda^*)$ は $d\times d$ の行列であり，その ij 成分はそれぞれ

(B.8)
$$(H_{\lambda,r}(\theta_\lambda^*))_{ij}=-\boldsymbol{E}_{pq}\left\{\left(\frac{q(x)}{r(x)}\right)^{\lambda-1}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(y|x,\ \theta_\lambda^*)\right\}$$

(B.9)
$$(K_{\lambda,r}(\theta_\lambda^*))_{ij}=\boldsymbol{E}_{pq}\left\{\left(\frac{q(x)}{r(x)}\right)^{\lambda}\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\frac{\partial}{\partial\theta_j}\log p(y|x,\ \theta_\lambda^*)\right\}$$

と定義される．(B.7) を示すためには，

(B.10)
$$\boldsymbol{E}_{D_T}\left\{\left(\boldsymbol{E}_{pq}\left\{-\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\right\}-\frac{\partial}{\partial\theta_i}L(\theta_\lambda^*,\ D_T)\right)(\hat{\theta}_\lambda-\theta_\lambda^*)_i\right\}$$

(B.11)
$$=\frac{1}{T}\mathrm{Tr}H_{\lambda,r}(\theta_\lambda^*)^{-1}K_{\lambda,r}(\theta_\lambda^*)+o\left(\frac{1}{T}\right)$$

を証明すればよい．推定量 $\hat{\theta}_\lambda$ の推定方程式

(B.12)
$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{q(x_t)}{r(x_t)}\right)^{\lambda}\log p(y_t|x_t,\ \hat{\theta}_\lambda)=0$$

をパラメータ θ_λ^* のまわりで展開して，

(B.13)
$$-\frac{1}{T}\sum_{t=1}^{T}\left(\frac{q(x_T)}{r(x_t)}\right)^{\lambda}\log p(y_t|x_t,\ \theta_\lambda^*)=H_{\lambda,r}(\theta_\lambda^*)+O_p\left(\frac{1}{\sqrt{T}}\right)$$

を用いると，

(B.14)
$$\hat{\theta}_\lambda-\theta_\lambda^*=\tilde{H}_{\lambda,r}(\theta_\lambda^*)^{ij}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{q(x_t)}{r(x_t)}\right)^{\lambda}\frac{\partial}{\partial\theta_j}\log p(y_t|x_t,\ \theta_\lambda^*)+O_p\left(\frac{1}{T}\right)$$

となる．ここで，$\tilde{H}_{\lambda,r}(\theta_\lambda^*)^{ij}$ は $H_{\lambda,r}(\theta_\lambda^*)$ の逆行列の ij 成分をあらわしている．(B.10) に (B.14) を代入すると，

(B.15)
$$\begin{aligned}&\boldsymbol{E}_{D_T}\left\{\left(\boldsymbol{E}_{pq}\left\{-\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\right\}-\frac{\partial}{\partial\theta_i}L(\theta_\lambda^*,\ D_T)\right)(\hat{\theta}_\lambda-\theta_\lambda^*)_i\right\}\\&=\tilde{H}_{\lambda,r}(\theta_\lambda^*)^{ij}\boldsymbol{E}_{D_T}\left\{\left(\boldsymbol{E}_{pq}\left\{-\frac{\partial}{\partial\theta_i}\log p(y|x,\ \theta_\lambda^*)\right\}-\frac{\partial}{\partial\theta_i}L(\theta_\lambda^*,\ D_T)\right)\right.\\&\left.\times\frac{1}{T}\sum_{t=1}^{T}\left(\frac{q(x_t)}{r(x_t)}\right)^{\lambda}\frac{\partial}{\partial\theta_j}\log p(y_t|x_t,\ \theta_\lambda^*)\right\}+o\left(\frac{1}{T}\right)\\&=\tilde{H}_{\lambda,r}(\theta_\lambda^*)^{ij}K_{\lambda,r}(\theta_\lambda^*)_{ij}+o\left(\frac{1}{T}\right)\end{aligned}$$

となる．よって，IC_λ を

$$IC_\lambda = -\frac{1}{T}\sum_{t=1}^{T}\frac{q(x_t)}{r(x_t)}\log p(y_t|x_t,\ \hat{\theta}_\lambda) + \frac{1}{T}\mathrm{Tr}\hat{H}_{\lambda,r}(\hat{\theta}_\lambda)^{-1}\hat{K}_{\lambda,r}(\hat{\theta}_\lambda) \tag{B.16}$$

と定義すれば (3.1) をみたす．ここで $\hat{H}_{\lambda,r}$, $\hat{K}_{\lambda,r}$ は $H_{\lambda,r}$, $K_{\lambda,r}$ の適当な一致推定量 ── たとえば，$\boldsymbol{E}_{pq}$ をサンプル平均に，θ^*_λ を $\hat{\theta}_\lambda$ にそれぞれ置き換えたもの ── とする．

参 考 文 献

安倍直樹，中村篤祥 (1997)．特集 能動学習，情報処理，**38**(7)，558-588．

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

Fukumizu, K. (1996). Active learning in multilayer perceptrons, *Advances in Neural Information Processing Systems*, **8**, 295-301.

金森敬文 (1999)．重み付き最尤推定量を用いた能動学習のアルゴリズム，信学技報，**98**，No. 674，NC 98-175．

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

Shimodaira, H. (1998). Improving predictive inference under covariate shift by weighting the log-likelihood function, Research Memo., No. 712, The Institute of Statistical Mathematics, Tokyo.

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

An Active Learning Algorithm Using an Information Criterion for the Maximum Weighted Log-likelihood Estimator

Takafumi Kanamori
(The Graduate University for Advanced Studies)

Hidetoshi Shimodaira
(The Institute of Statistical Mathematics)

We suppose that outputs of a system are generated from a random mechanism having the conditional distribution given inputs. In this paper we study the estimation or identification of the system when the observer can select appropriate inputs to the system. We call such method of estimation *Active Learning*. We suppose that the statistical model does not correctly specify the system. When the statistical model is not correct, the active learning using the maximum likelihood estimator does not enjoy consistency. Here consistency means the convergence to the optimal parameter fitted by the Kullback-Leibler divergence. Hence we suggest an active learning algorithm using maximum weighted log -likelihood estimator (mwle). The algorithm is shown to enjoy of consistency. Moreover we point out that there are estimators which are better than the consistent estimators when the number of the data is finite. Considering such result we construct another active learning algorithm which selects an appropriate mwle using our information criterion. We give some experiments by computer simulation and explore the effect of the proposed algorithms.

Key words: Active learning, maximum weighted log-likelihood estimator, information criterion, statistical asymptotic theory, risk.