正則化法によるニューラルネットの学習について

On the learning of neural networks based on the regularization method

赤穂 昭太郎

Shotaro AKAHO

電子技術総合研究所

Electrotechnical Laboratory

1. はじめに

本報告では正則化法に基づくニューラルネットの学習法を提案し, その評価を行なう.

ニューラルネットは確率的に発生する経験データをもとに自己組織的な学習を行なうシステムである.

ここで, 学習というのは**期待損失** (expected risk)

$$I(\alpha) = \int Q(\boldsymbol{z}, \alpha) P(\boldsymbol{z}) d\boldsymbol{z} \tag{1}$$

を最小とするようなパラメータ α を経験データをもとに推定する問題であると言うことができる [2][12]. ここで $P(\boldsymbol{z})$ はデータ $\boldsymbol{z}$ の従う未知の確率分布であり, $Q(\boldsymbol{z}, \alpha)$ は損失関数 (またはポテンシャル関数) とよばれ, 例えば $\boldsymbol{z} = (y, \boldsymbol{x})$ として $Q(y, \boldsymbol{x}) = (y - q(\boldsymbol{x}))^2$ というようなものを考える ($q(\boldsymbol{x})$ は $\boldsymbol{x}$ の関数).

一方, ニューラルネットは十分多くの素子を使えば任意の関数を近似する能力を持っていることが知られていて, 目的の関数は十分大きなサイズのネットワークを使えばいくらでも高い精度で実現可能である [6].

しかし実際には $P(\boldsymbol{z})$ は未知だから $I(\alpha)$ を直接評価することはできず, 有限個の経験データだけを使って評価しなければならない. そこで通常は (1) 式の代わりに n 個のデータ $\underset{(1)}{\boldsymbol{Z}}, \ldots, \underset{(n)}{\boldsymbol{Z}}$ が得られたときの**経験損失** (empirical risk)

$$I_{\mathrm{emp}}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} Q(\underset{(i)}{\boldsymbol{Z}}, \alpha) \tag{2}$$

を最小化することになる. ところが一般に経験損失の最小化と期待損失の最小化は一致せず, ネットワークの能力が高ければ高いほど経験データだけにオーバーフィットした不安定な関数が得られてしまい, 未知データに対する能力は逆に低下してしまうという問題が生じる (**過学習** あるいは **汎化性の欠如**). この問題を避けるための方法として, ネットワークの大きさや構造を限定して, 実現できる関数の族を限定することによって汎化性を高めるという, いわゆる統計的なモデル選択法がある.

本報告では, 経験損失を最小化するのではなく, $P(\boldsymbol{z})$ をデータから推定することによって期待損失を最小にする問題を考える. 一般に有限個のデータから確率密度関数を決める問題はいわ

ゆる**不良設定問題** (ill-posed problem) であり, 解が安定ではない. そのため本手法では**正則化** (regularization) を行なうことによって解を安定化させる. 直観的に言えば, 正則化法とは, 少ないデータしか得られない場合でもデータ点のまわりでぼかしを行なうことによって確率密度関数を推定するという方法である.

2節ではまずネットワークのモデル選択の従来手法を概観し, その性質を明らかにする. 3節では正則化法を用いた確率密度関数の近似法について述べ, 4節でそれを用いた学習法を具体的に構成し, その評価を行なう.

評価法としてはパラメトリックに行なう (真の関数がモデルの中に入っていると仮定する) か, ノンパラメトリックに行なうかという選択が考えられる. 複雑な現実問題を解く場合にはモデルを立てるのが困難なことが少なくないという考えから, 本報告では主としてノンパラメトリックな立場をとることにする.

2. ネットワークのモデル選択

本節ではネットワークの大きさや構造を限定することによって汎化性を持たせる従来方法について概観する.

この立場は経験データに最もよく当てはまるようなネットワークのうちできるだけ能力の低いもの, つまり高い汎化性をもつものを選ぶことによって最適なネットワークのサイズを決定しようとするものであり, 「けちの原理」として説明される.

種々の評価基準が統計などの立場から提案されているが, 代表的なものとしては

- 情報量基準に基づく AIC(Akaike's Information Criterion) や MDL(Minimum Discription Length principle) を用いるもの [9][14][13]
- VC 次元を用いて汎化性を評価するもの [12] [4][3]

などがある.

情報量基準に基づくモデル選択法の特徴は, パラメトリックな立場から最尤推定を行なうところにある. AIC と MDL との差は, 問題設定の違いである. 和田・川人は必ずしも最尤推定量が得られないという状況のもとでの新しい情報量基準を導出し, resampling 手法と組み合わせることによるモデル選択法を提案している [14].

VC 次元は Baum, Haussler によるネットワークのモデル選択法の提案 [3] 以来注目されている方法である. Vapnik らはノンパラメトリックな立場で, 経験損失が期待損失にどれぐらい近いかを非常に一般的な場合に対して評価した [12]. このとき出てきた概念が VC 次元である (Appendix A 参照). しかしながらこの評価は相当ゆるく, 直接モデル推定に用いるのは現実的ではないと考えられる.

このほか甘利, 藤田, 篠本はパラメトリックな立場で, 平均誤り率を種々の問題設定に対して評価した [1].

いずれの手法も現時点でのモデル選択法は, 基本的に

- いくつかのネットワークについて学習を行ない, 評価値を計算する.
- 最も良い評価値を得るネットワークを選択する.

という 2 段階の構成になっており, 選択された以外のネットワークについての学習は無駄になってしまう. また Back propagation などの学習では通常, モデルの構造が複雑になっていて, パラメータの個数の増減によって学習結果が大きく異なることが多く, それぞれのネットワークについての学習は独立して行なわなければならないのが普通である.

パラメータ数の削減を学習の中に組み込む方法もいくつか提案されているが ([7], [8] など), 一般的な評価をするのが難しく, また上記のようにパラメータ数によって学習結果が大きく変化する問題に対しては有効性が必ずしも明らかでないなど問題も多く残されている.

3. 正則化法による確率密度関数の近似

本節では, 正則化法による確率密度関数の近似手法について述べる.

3.1. 正則化法

最初に述べたように確率密度関数の推定問題は不良設定問題である. 不良設定問題を解くのに古くから用いられているノンパラメトリックな手法の一つが**正則化法** (regularization method) である [11]. 正則化法を確率密度関数 $f(t)$ の推定に即して書けば次のようになる. 積分方程式

$$Af(t) \equiv \int_{-\infty}^{\infty} u(x-t)f(t)dt = F(x) \tag{3}$$

を n 個の経験データから解くことを考える ($u(x)$ は unit step function). 確率分布関数は経験データから安定に近似できるが, それを直接上の積分方程式に入れて解いても不安定な解しか得られない. そこで次の汎関数の最小化問題を考える.

$$R_\gamma(\hat{f}, F) = \rho^2(A\hat{f}, F) + \gamma\Omega(\hat{f}) \tag{4}$$

ここで ρ は関数距離, Ω はいくつかの条件を満たす汎関数で stabilizer といい, γ は正の実定数で正則化パラメータと呼ばれる. この問題は安定に解くことができて F の近似の度合いに応じて γ の値を十分ゆっくり小さくして行けば漸近的に $\hat{f}$ は $Af = F$ の真の解に収束することが知られている.

3.2. 正則化法と Parzen 法

確率密度関数の推定に関して, 正則化法は次の Parzen 法 (または Kernel type density estimation method) と同値であることが知られている [12]. Parzen 法とは確率密度関数 $f(x)$ に従って現れる n 個の経験データ $\underset{(1)}{X}, \ldots, \underset{(n)}{X}$ をもとに,

$$f_n(x, a_n) = \frac{a_n}{n} \sum_i K(a_n(x - \underset{(i)}{X})) \tag{5}$$

という関数で $f(x)$ を推定する (1 次元の場合). ここで $K(x)$ は適当な条件を満たす関数であり, a_n は定数である. 正則化法との対応でいえば $K(x)$ は stabilizer に対応し, a_n は正則化パラメータに対応する (a_n は n とともに増加する値である).

大ざっぱに言えば $K(x)$ によってサンプル点のまわりでぼかしを行なう手法である. この手法ではぼかしのパラメータを決めることや真の確率密度関数への近さの評価をすることが重要なポイントとなる. これをうまく行なわないと結局, 過学習やぼかし過ぎの問題が生じる.

Parzen 法では漸近的に最適な a_n を経験データから決定することができる [10]. その詳細な結果は Appendix B に示すが, その要点は

- p 次元の確率密度関数が s 階連続偏微分可能であるという仮定をおく ((18) 式).
- $K(x)$ として適当な条件を満たすものをとる ((21) 式, Theorem 3).

- すると真の関数と近似した関数との自乗誤差の期待値が評価できる (Theorem 2).
- それをもとに漸近的に最適なパラメータ a_n を決めることができる ((36) 式).

ということである. Parzen 法を使った簡単な実験結果を図 1に示す. 実験では確率密度

$$\frac{N(-2,1)+N(2,1)}{2}$$

に従うデータを発生させた. ただし $N(\mu,\sigma^2)$ は平均 μ, 分散 σ^2 の正規分布をあらわす. また核関数 $K(x)$ としては正規分布関数を用いた.

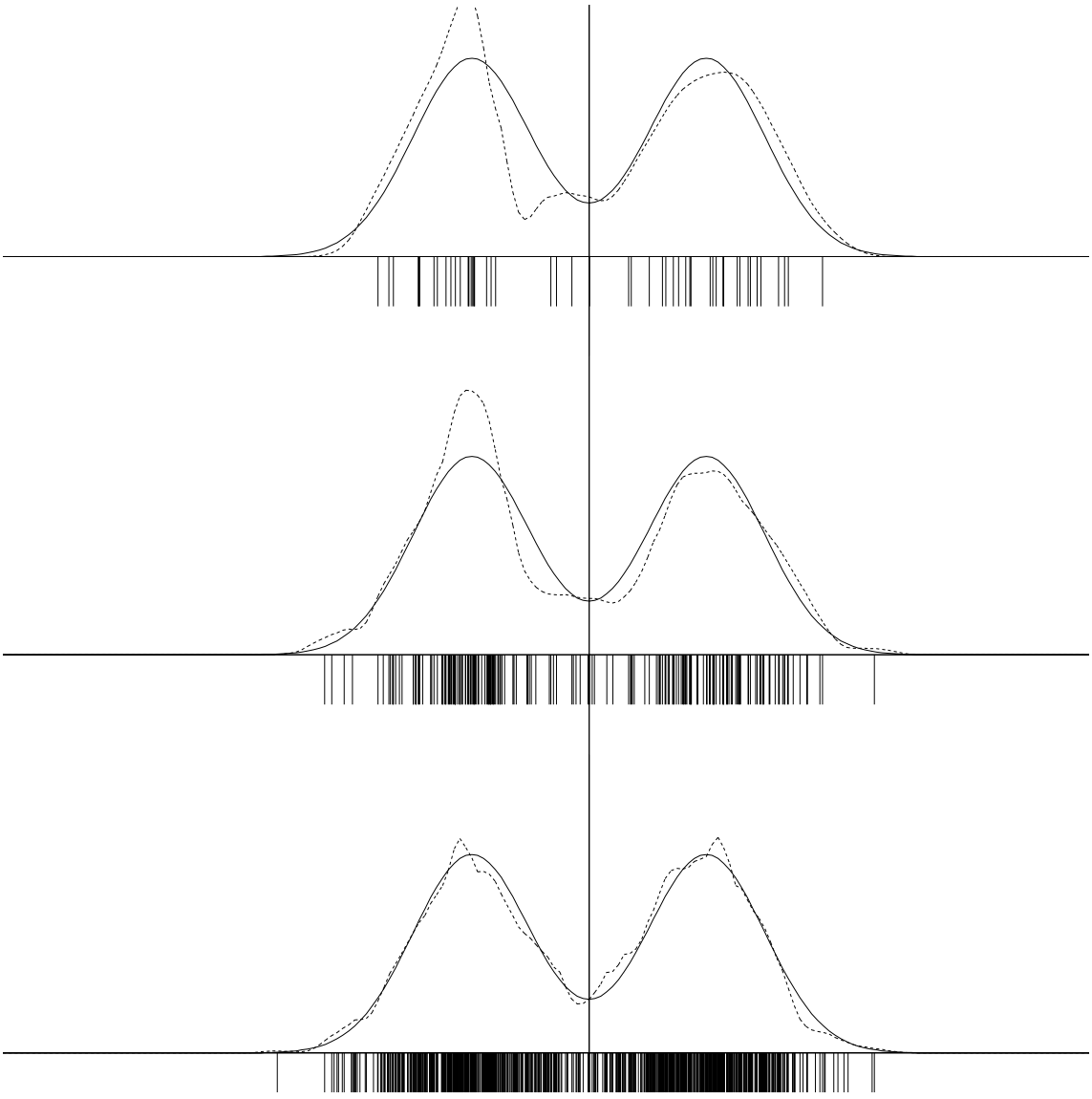

図 1. Parzen 法の実験結果:　上から順にデータ数 50, 200, 800 のグラフ. 実線: 真の確率密度; 点線: Parzen 法による近似; 横軸の下の短い線分: 各データ点の位置

4. 正則化法を用いた学習法

本節では近似された確率密度関数を用いたニューラルネットの学習とその評価について考えてみる.

4.1. 損失最小化とその評価

本節では, 期待損失 $I(\alpha)$ を最小化する立場から, データから推定された確率密度関数 $\hat{P}(\boldsymbol{z})$ を用いた近似関数

$$\hat{I}(\alpha) = \int Q(\boldsymbol{z}, \alpha)\hat{P}(\boldsymbol{z})d\boldsymbol{z} \tag{6}$$

を最小化する方法を考える. $\hat{I}(\alpha)$ はデータだけから決まる値だからこの最小化は原理的には可能である. この最小化ができたとして, 真の期待損失にどれだけ近いかを評価しよう. これに関しては Appendix B の Theorem 3より容易に次の定理が導かれる (証明および詳細は Appendix C を参照).

Theorem 1. p 次元の確率密度関数 $P(\boldsymbol{z})$ はすべての変数に関して s 階連続微分可能であるとし, 有界な定義域を持つとする. このとき適当な条件のもとで, 十分大きい n に対して,

$$\mathrm{E}[\hat{I}(\alpha) - I(\alpha)]^2 \leq Dn^{-2s/(2s+p)} \tag{7}$$

ただし E はサンプルに関する期待値

$$\mathrm{E}[\,\cdot\,] = \int \cdot \prod_j P(\underset{(j)}{\boldsymbol{Z}})d\underset{(j)}{\boldsymbol{Z}}$$

をとる. また D は $Q(\boldsymbol{z}), P(\boldsymbol{z})$ に依存して決まる定数である.

さてここで Vapnik と類似の基準を用いることにして, 適当な δ に対し,

$$\mathrm{E}[I(\hat{\alpha}_0) - I(\alpha_0)]^2 < \delta \tag{8}$$

を達成することが汎化性の基準であると考える. ただし $\hat{\alpha}_0$ は $\hat{I}(\alpha)$ を最小にする α であり, α_0 は $I(\alpha)$ を最小にする α である. するとその十分条件として

$$\mathrm{E}[\hat{I}(\alpha_0) - I(\alpha_0)]^2 + \mathrm{E}[\hat{I}(\hat{\alpha}_0) - I(\hat{\alpha}_0)]^2 < \frac{\delta}{2} \tag{9}$$

が得られる. この式の左辺は Theorem 1 の結果から評価できる.

$\hat{I}(\alpha)$ の最小化を反復的なニューラルネットの学習で行なうには, 例えば $\hat{P}(\boldsymbol{z})$ に従うサンプルを疑似的に増加させればよい.

具体的なアルゴリズムとしては次のようなものが考えられる (簡単のために 1 次元の記法で書く).

[アルゴリズムの例]

1: サンプルが n 個あるとする $(\underset{(1)}{X}, \ldots, \underset{(n)}{X})$.

2: 以下の I),II), III) を繰り返す:

I) ランダムにサンプル $\underset{(i)}{X}$ をとる.

II) $a_n K(a_n(x - \underset{(i)}{X}))$ に従う疑似サンプルを 1 つまたは数個作る.

III) それらをもとに (少し) 学習する.

ここで注意すべき点を列挙しておく.

- まず $a_n K(a_n(x - \underset{(i)}{X}))$ は $K(x)$ の選び方によって必ずしも確率密度関数とはならないことである. この場合には疑似サンプルの生成は近似的に行なうことになる.

- また III) の学習においてはローカルミニマムに落ちてしまうなどの問題を生じるためあまり学習し過ぎないことが望ましい．しかし定量的には各学習アルゴリズムの特性に依存するのでここでは深入りしない．

- 学習の際に作り出す疑似的なサンプルの数を含めた数はネットワークの表現能力などに依存した値になるはずである．これに関する評価はモデル選択における結果を使って議論することができるであろう．

以上のことから, この方法が有効であるのは,

" 学習は高速に行なえるけれども, 実際のサンプルをとるのが困難であったり, 非常に時間のかかる場合 "

であると考えられる．

上記のアルゴリズムに従って行なった実験結果を示す (図 2–4).

サンプル 入力 x および出力 y はともに 1 次元で, x は $[0, 1]$ の一様分布に従い, 定数関数

$$y = 0.5$$

に分散 0.01 の正規ノイズを加えたものを出力サンプルとしてとった．

サンプルサイズ 上記のデータを 5 個生成した．

ネットワークの大きさ 入力層, 中間層, 出力層をそれぞれ 1,10,1 にとった．

学習法 正則化法による学習ではサンプルから 1000 個の疑似サンプルを生成してそれにたいして一括型の Back propagation 学習を 10 ステップ行なうということを "1 学習ステップ " として, 100 学習ステップ行なった．

1 学習ステップに対応して, サンプルだけからの学習を $2000 = 1000 \times 10/5$ 回, 一括型 Back propagation で行なった．

Back propagation 学習は加速を全く行なわない, 最も単純なものを使った．最急降下の係数は 0.5 とした．

評価 サンプルに対する平均自乗誤差および 10000 個のテストデータに対する平均自乗誤差で評価を行なった．

結果 いろいろな乱数初期値に対して上記の実験を行なって, 各評価値をプロットしたのが図のグラフである．横軸が学習ステップ数をあらわし, 縦軸は値を対数目盛でとったものである．

曲線はそれぞれ,
· サンプル学習のサンプルに対する誤差 (下の実線)
· 正則化学習のサンプルに対する誤差 (点線)
· 正則化学習のテストデータに対する誤差 (鎖線)
· サンプル学習のテストデータに対する誤差 (上の実線)
をあらわす．

データには分散 0.01 のノイズを加えているので, 最適な関数に収束したとしてもテストデータに対しては平均 0.01 程度の自乗誤差が存在するはずである．

いずれの場合も, サンプル学習ではデータにオーバーフィットして汎化が行なわれていないことがわかる．正則化学習ではサンプルに対する誤差もテストデータに対する誤差も 0.01 付近で振動していて, 汎化されていることがわかる．振動の原因は新しい疑似サンプルを次々と作り出しているからであると考えられる．

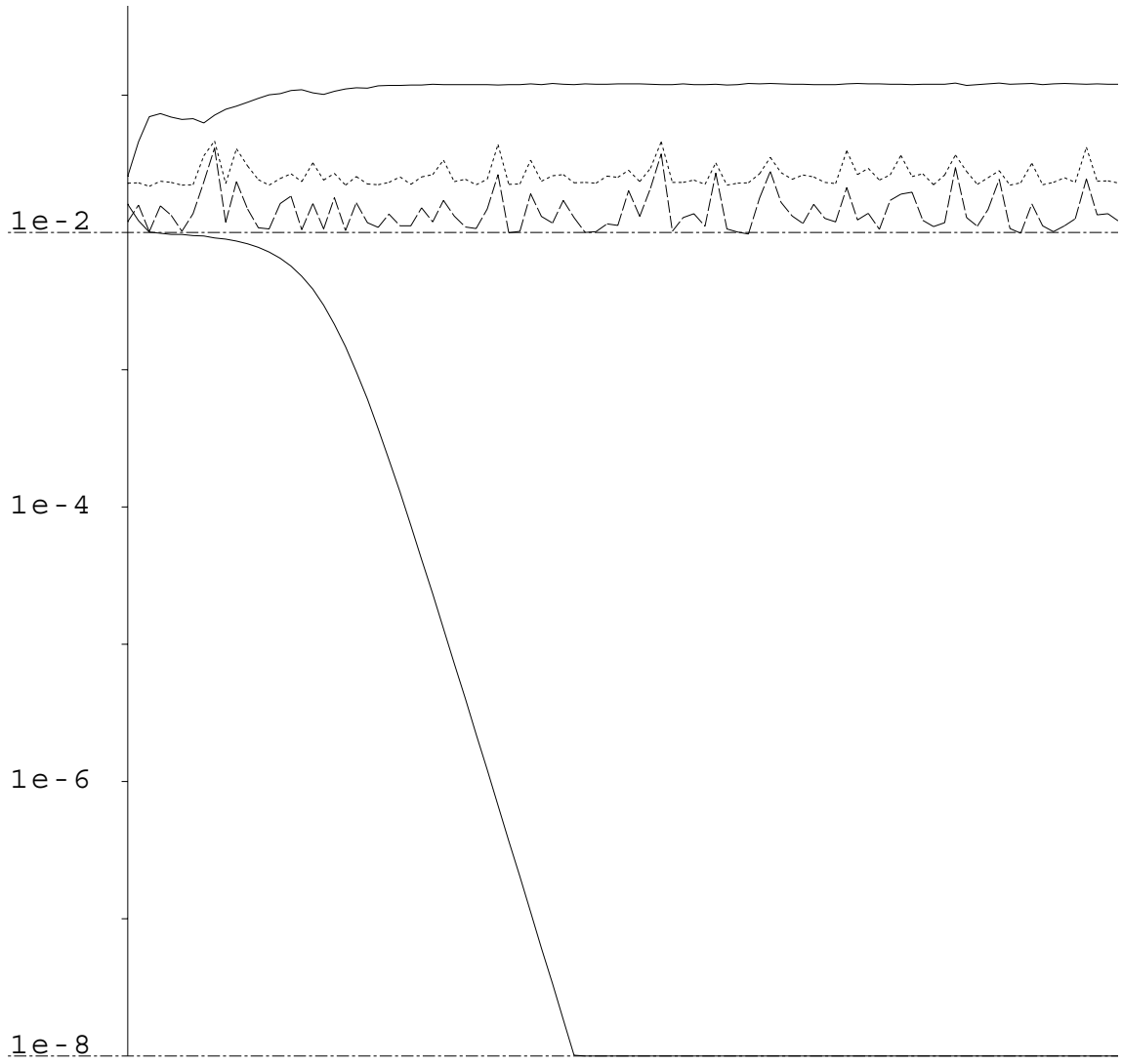

図 2. 期待損失最小化アルゴリズムの実験結果 (1): 最も典型的な収束例. 問題が簡単なのでサンプル学習は比較的速く収束することが多い.

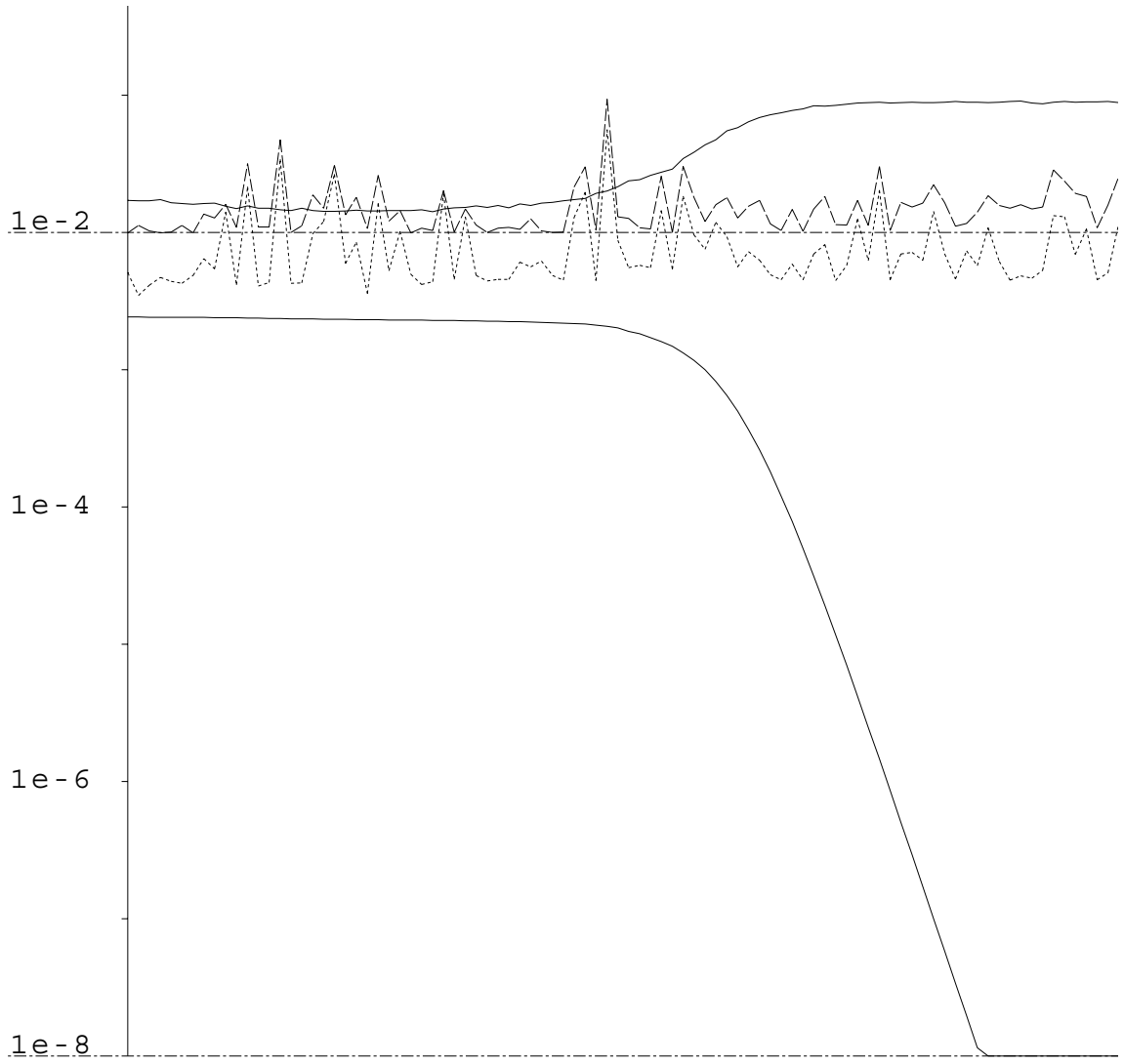

図 3. 期待損失最小化アルゴリズムの実験結果 (2):　サンプル学習が途中から急に学習が進んでそれとともに汎化能力が落ちている.

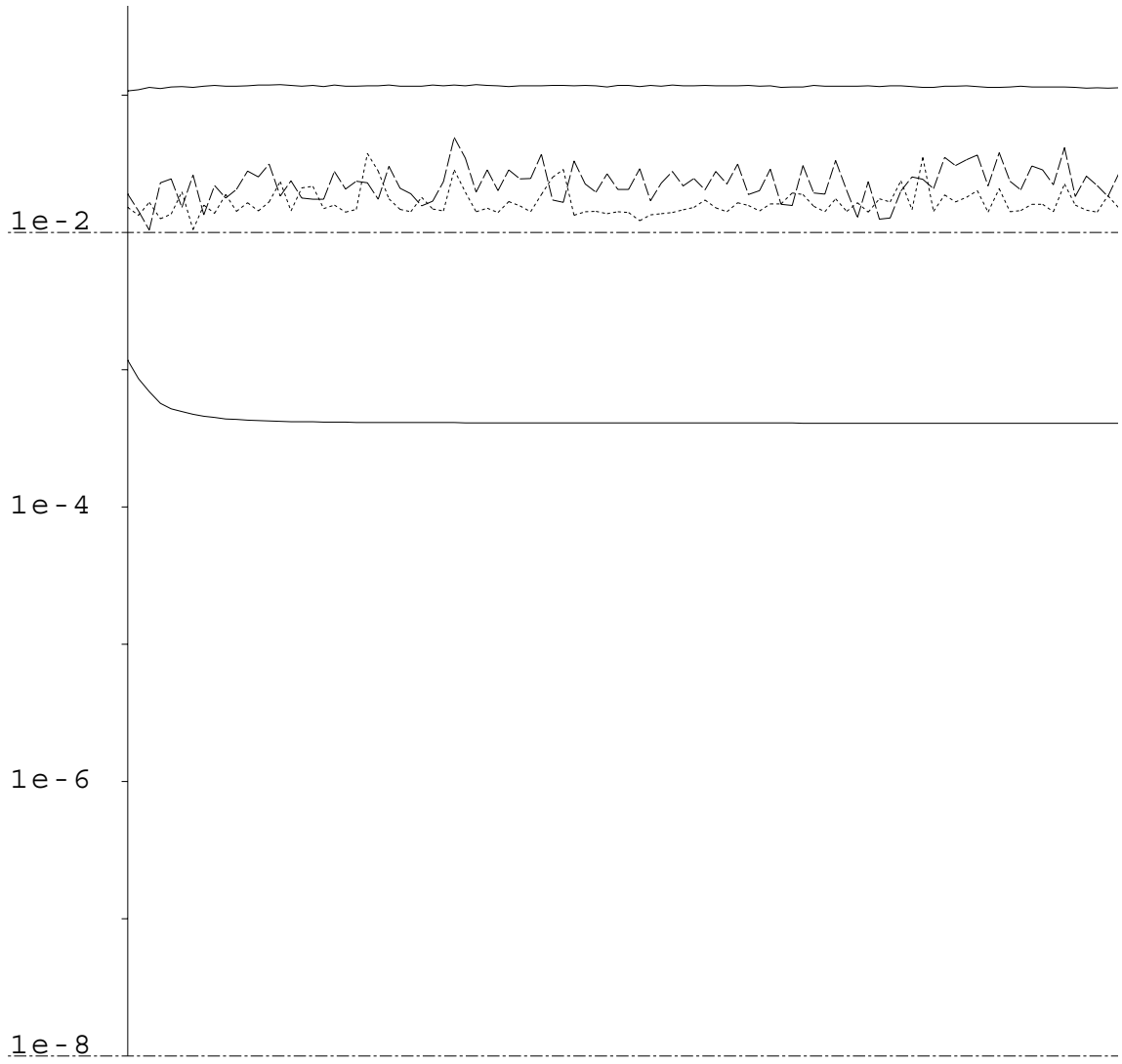

図 4. 期待損失最小化アルゴリズムの実験結果 (3): サンプル学習がローカルミニマムに落ちている

4.2. Parzen 法をシミュレートする回路網

前節とは異なるアプローチとして出力を確率密度関数と考えて, Parzen 法をシミュレートするようなネットワークを考えることができる (図 5).

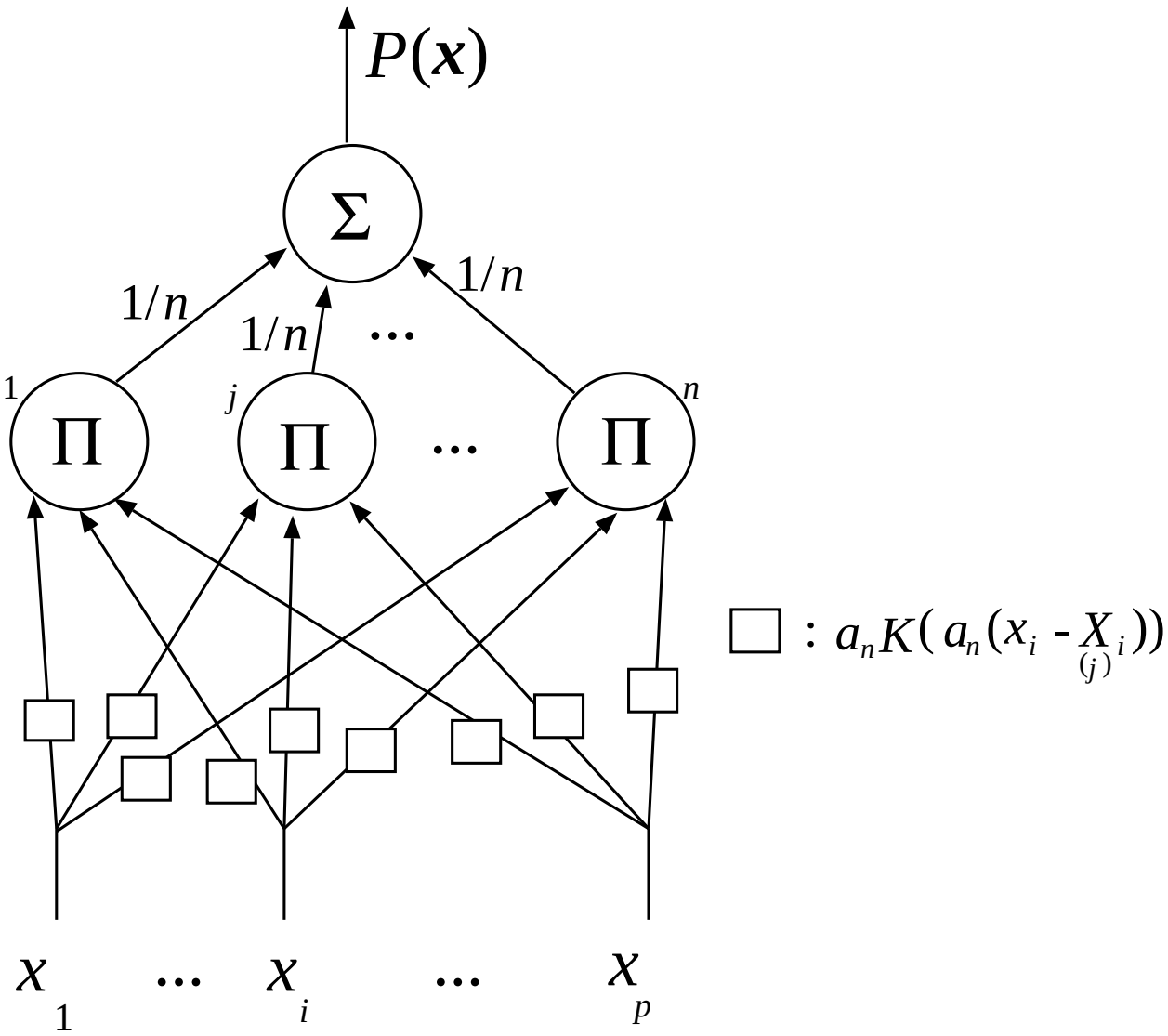

図 5. Parzen 法をシミュレートする回路網

1 つの Π 素子が 1 サンプルに対応していて, i 番目の入力から j 番目の Π 素子には $a_n K(a_n(x_i - X_i^{(j)}))$ $(i = 1, \cdots, p)$ なる入力が入ってくるとする. Parzen 法に対しては Theorem 3 によって評価できる.

この回路そのものは, いわゆるニューラルネットの枠組の外にあり, 例題の数に比例した多くのハードウェアを必要とするなどの欠点を持つ一方で,

- モデル選択が不要である
- 反復的学習が不要である

などの利点を持つ.

従ってこの方法が有効であるのは

“ 問題が難しくて通常の反復学習では非常に時間がかかる一方, 得られる例題の数は少ないような場合 ”

であると考えられる.

4.3. 教師あり学習

さて, 前節では入力をデータそのものとし, 出力をそれに対する確率密度関数と仮定した. いわゆる教師なし学習の場合はまさにこの枠組に属している. 一方通常の教師あり学習ではデータの一部 (例えば $\boldsymbol{z} = (\boldsymbol{x}, y)$ のうちの $\boldsymbol{x}$) が与えられた時に y を出力するようなシステムを求めたいわけであるが, 前節で述べた回路は直接的にはそのような形をしていないため, 若干の工夫をする必要がある. 本節ではそれについて述べる.

4.3.1. パターン識別問題の場合

パターン識別の問題の場合は, データがクラスという属性を持つが, それは離散的な値をとるので Parzen 法のようにぼかしを行なうことが適当でないと考えられる. この場合はクラスの属性とデータ点の属性を分離して考えなければならない.

前節で述べた回路をパターン識別問題に応用することを考えてみよう. Bayes 的な決定則を考え, $P(k|\boldsymbol{x})$ を最大にする k を出力することにすれば, 与えられた $\boldsymbol{x}$ に対して,

$$P(k, \boldsymbol{x}) = P(k)P(\boldsymbol{x}|k) \tag{10}$$

を最大にするクラス k を出力するような回路を構成ればよい. $P(k)$ としては例えば経験頻度 $\hat{P}(k)$ を使い, $P(\boldsymbol{x}|k)$ を計算するために各 k について前節に述べた回路を用意する. 最後に各回路の出力に重み $\hat{P}(k)$ をかけて最大値をとればよい.

4.3.2. 回帰の問題の場合

回帰の問題を非常に一般的に言えば, $\boldsymbol{x}, y$ が連続値をとる場合に, 入力 $\boldsymbol{x}$ に対する出力 y を推定する問題ということができる. y は確率変数だから平均値 $\mathrm{E}\, y$ を出力することにする.

すると, 求めるものは

$$\mathrm{E}\, y = \int y P(y|\boldsymbol{x}) dy = \frac{\int y P(\boldsymbol{x}, y) dy}{P(\boldsymbol{x})} \tag{11}$$

と書け, $P(\boldsymbol{x}, y), P(\boldsymbol{x})$ としては Parzen 法から近似される密度関数を用いることにすれば,

$$\mathrm{E}\, y \simeq \hat{y} = \frac{\int y \hat{P}(\boldsymbol{x}, y) dy}{\hat{P}(\boldsymbol{x})} \tag{12}$$

となる. 具体的に書き下せば ($\boldsymbol{x}$ が 1 次元の場合),

$$\begin{aligned} \hat{y} &= \frac{\int y (\frac{a_n}{n})^2 \sum_i K(a_n(x - \underset{(i)}{X})) K(a_n(y - \underset{(i)}{Y})) dy}{\frac{a'_n}{n} \sum_i K(a'_n(x - \underset{(i)}{X}))} \\ &= \frac{a_n \sum_i \underset{(i)}{Y} K(a_n(x - \underset{(i)}{X}))}{a'_n \sum_i K(a'_n(x - \underset{(i)}{X}))} \end{aligned} \tag{13}$$

となる. ただし a_n は 2 次元データ x, y に対するパラメータ値, a'_n は 1 次元データ x に対するパラメータ値であり, いずれもこの場合定数として計算した.

5. 考 察

- 本手法と関係の深い手法として smoothing spline の分野がある [5]. 本手法が確率密度の推定を中心としているのに対し, smoothing spline は回帰関数の推定が中心であるという点で若干問題設定が異なってくるが, どちらも正則化法の枠組でとらえることができる. Smoothing spline では通常 generalized cross-validation と呼ばれる resampling 手法を用いた精密な評価によってパラメータを決定する. 本報告では, できるだけ容易にパラメータを決定するためにデータにほとんど依存しないパラメータ決定法を選んだが, より精密なパラメータ決定が必要な場合には resampling 手法をとり入れることも考えられる.

- Theorem 1 および Theorem 3の評価はデータの次元が上がれば上がるほど悪くなる. 従って本手法では生データから有効な低次元特徴量を抽出することが重要な問題となる.

6. む す び

本報告では通常のモデル選択法と異なるアプローチから汎化を行なう手法として正則化法に基づく方法を提案した.

より精密な評価や手法の改良, 実際の問題に対する有効性を示すことなどは今後の課題として残されている.

謝 辞

研究の機会を与えていただいた 田村浩一郎 情報科学部長に感謝します. また日頃より御討論いただく 大津展之 知能情報部長, 麻生英樹 · 栗田多喜夫 研究官をはじめとする情報数理研究室の皆様に感謝します.

参考文献

[1] S. Amari, N. Fujita, and S. Shinomoto: Four types of learning curves. *preprint.*

[2] 甘利 俊一: 神経回路網の数理. 産業図書, 1978.

[3] E.B. Baum and D. Haussler: What size net gives valid generalization? *Neural Computation*, Vol. 1, pp. 151–160, 1989.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth: Learnability and the Vapnik-Chervonenkis dimension. *J. of the Assoc. for Comp. Machinery*, pp. 929–965, 1989.

[5] P. Craven and G. Wahba: Smoothing noisy data with spline functions. *Numerische Mathematik*, Vol. 31, pp. 377–403, 1979.

[6] K. Hornik, M. Stinchcombe, and H. White: Multilayer feedforward networks are universal approximators. *Neural Networks*, Vol. 2, pp. 359–366, 1989.

[7] 石川 真澄: コネクショニストモデルの忘却を用いた構造化学習. 信学技法 **MBE**88–144, 1989.

[8] J.K. Kruschke: Improving generalization in back-propagation networks with distributed bottlenecks. In *Proc. of IJCNN '89*, volume I, pp. 443–447, 1989.

[9] 栗田 多喜夫: ニューラルネットにおけるモデル選択の試み. 信学技法 **PRU**89–16, 1989.

[10] E.A. Nadaraya: *Nonparametric estimation of probability densities and regression curves.* Kluwer Academic Publishers, 1989.

[11] A.N. Tikhonov and V.Ya. Arsenin: *Solutions of Ill-posed Problems.* Winston, Washington, 1977.

[12] V.A. Vapnik: *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, 1984.

[13] K. Yamanishi: Learning non-parametric-densities using finite-dimensional parametric hypotheses. In *Proc. of ALT '91*, pp. 175–186, 1991.

[14] 和田、川人: 新しい情報量基準と cross validation による汎化能力の推定. 信学会論文誌, Vol. J74-D-II, No. 7, pp. 955–965, 1991.

Appendix A. Vapnik による汎化性の評価

Vapnik は期待損失 $I(\alpha)$ を経験データをもとに最小化する問題を扱った.

$I(\alpha)$ の最小化の代わりに, n 個のデータが得られたときの経験損失 $I_{\rm emp}(\alpha)$ を最小化する. このとき $I_{\rm emp}(\alpha)$ を小さくすると同時に, κ, δ を適当に決めて

$$\Pr[I(\alpha_{\rm emp}) - I(\alpha_0) > \kappa] < \delta \tag{14}$$

にしたいというのが Vapnik による汎化性の基準である. ここで $\alpha_{\rm emp}$ は経験損失 $I_{\rm emp}(\alpha)$ を最小にする α であり, α_0 は期待損失 $I(\alpha)$ を最小にする α である.

Vapnik は以下に述べるような方法で (14) 式の左辺を上から押さえている.

- (14) 式の十分条件である.

$$\Pr[\sup_{\alpha} |I(\alpha) - I_{\rm emp}(\alpha)| > \frac{\kappa}{2}] < \delta \tag{15}$$

  を (14) 式のかわりに評価する.

- 上式の左辺は

$$2\Pr[\sup_{\alpha} |I_{\rm emp1}(\alpha) - I_{\rm emp2}(\alpha)| > \frac{\kappa}{4}] \tag{16}$$

  で上から押さえられる (ただし emp1,emp2 は互いに独立な実験でそれぞれ n 個の経験データからなる). これによって $P(\boldsymbol{z})$ に関する積分の評価を避けている.

- 一般に変数 x に関して

$$\sup x < \sum_{i \in A} x_i \quad (A \text{ は } x \text{ のすべての場合}) \tag{17}$$

  が成り立つから (16) 式の sup をすべての場合の数に関する総和でおきかえる. (この部分が最もゆるい評価になっていると考えられる)

最後の評価における「場合の数」をさらに上から評価するために導入されたものが VC 次元であり, 定性的には表現力の高い関数族ほど場合の数は大きくなるから期待損失と経験損失との差は大きくなる (通常は VC 次元は 2 分割問題にしか適用できないような言い方がされるがもっと一般的に定義されている). モデル選択の立場からみれば, 経験損失そのものの大きさと, 経験損失と期待損失との差との間のトレードオフによってモデルのサイズを最適に決めることができるように思えるが現時点では評価が非常にゆるい評価によってしか与えられていないので現実的なモデル選択には適用できないと考えられる.

しかし, ノンパラメトリックな枠組での評価としては他には見られない結果であり, 計算量のオーダー評価など理論的な応用は十分にある.

Appendix B. Parzen 法におけるパラメータ決定法

Parzen 法では漸近的に最適な a_n を経験データから決定することができる [10]. 以下ではこの結果を述べる.

ここでは p 次元の確率変数 $\boldsymbol{x} = (x_1, \ldots, x_p)$ の確率密度関数を考える. 確率密度関数に対する仮定として, $f(\boldsymbol{x})$ は

$$G_s: \quad \text{すべての } x_i \text{ に関して } C^{(s)} \text{ 級関数の集合} \tag{18}$$

に属するとする. $f(\boldsymbol{x})$ を n 個のサンプルデータ $\underset{(1)}{\boldsymbol{X}},\ldots,\underset{(n)}{\boldsymbol{X}}$ を使って

$$f_n(\boldsymbol{x},a_n)=\frac{1}{n}\sum_{j=1}^{n}\prod_{i=1}^{p}a_nK(a_n(x_i-\underset{(j)}{X_i})) \tag{19}$$

で近似するのが (多次元の) Parzen 法である.

評価関数として

$$U(\boldsymbol{x},a_n)=\mathrm{E}[f_n(\boldsymbol{x},a_n)-f(\boldsymbol{x})]^2 \tag{20}$$

をとる. ここで E はサンプルに関する期待値

$$\mathrm{E}[\,\cdot\,]=\int\ \cdot\ \prod_j f(\underset{(j)}{\boldsymbol{X}})d\underset{(j)}{\boldsymbol{X}}$$

をとる.

また核関数のクラスとして

$$H_s:\quad (22)\text{–}(26)\text{ 式の条件を満たす関数 }K(x)\text{ の集合} \tag{21}$$

を考える. たとえば正規分布の分布関数は H_2 に属している.

$$K(x)=K(-x),\qquad \int K(x)dx=1, \tag{22}$$

$$\sup|K(x)|<\infty, \tag{23}$$

$$\int x^iK(x)=0\quad (i=1,\cdots,s-1), \tag{24}$$

$$\alpha=\int x^sK(x)dx\neq 0, \tag{25}$$

$$\int x^s|K(x)|dx<\infty. \tag{26}$$

このとき次の定理が成り立つ.

Theorem 2. $f(x)\in G_s$, $K(x)\in H_s$ なら,

$$U(\boldsymbol{x},a_n)\simeq\frac{a_n^p}{n}\|K\|^{2p}f(\boldsymbol{x})+a_n^{-2s}\frac{\alpha^2}{(s!)^2}(\sum_{i=1}^{p}\partial_i^sf(\boldsymbol{x}))^2. \tag{27}$$

ただし

$$\partial_i^sf(\boldsymbol{x})=\frac{\partial^sf(\boldsymbol{x})}{\partial {x_i}^s}, \tag{28}$$

$$\|K\|^2=\int K(x)^2dx. \tag{29}$$

また $a\simeq b$ は $a/b\to 1$ $(n\to\infty)$ をあらわす.

この定理から $U(\boldsymbol{x},a_n)$ を最小にする $a_n(=a_n^0)$ が ($f(\boldsymbol{x})$ に依存して) 次のように決まる.

$$a_n^0=C(\boldsymbol{x})n^\gamma,\qquad \gamma=\frac{1}{2s+p} \tag{30}$$

ただし

$$C(\boldsymbol{x})=\left(\frac{2s\alpha^2}{(s!)^2p\|K\|^{2p}}\frac{(\sum_i\partial_i^sf(\boldsymbol{x}))^2}{f(\boldsymbol{x})}\right)^\gamma. \tag{31}$$

このとき $U(\boldsymbol{x}, a_n^0) = O(n^{-2s/(2s+p)})$ となる.

実際には $f(\boldsymbol{x})$ は未知だから $C(\boldsymbol{x})$ を何らかの意味で近似しなくてはならない. ここで用いる手法は $C(\boldsymbol{x})$ の中の $f(\boldsymbol{x})$ を $f_n(\boldsymbol{x})$ で近似するというものである. そのために数列 $\{\tau_{n,i}\}, \{b_n\}$ を考える.

$$\{\tau_{n,0}\} \ : \ O(n^{\gamma}) \tag{32}$$

$$\{\tau_{n,i}\} \ : \ O(n^{\gamma^2}) \quad (i = 1, \cdots, p), \tag{33}$$

$$\{b_n\} \ : \ b_n \to 0 \quad (n \to \infty), \quad nb_n \geq C > 0. \tag{34}$$

この数列を用いて $C(\boldsymbol{x})$ を次のように近似する.

$$\hat{C}(\boldsymbol{x}) = \left(\frac{2s\alpha^2}{(s!)^2 p \| K \|^{2p}} \frac{(\sum_i \partial_i^s f_n(\boldsymbol{x}, \tau_{n,i}))^2 + b_n}{|f_n(\boldsymbol{x}, \tau_{n,0})| + b_n} \right)^{\gamma}. \tag{35}$$

すると a_n^0 の近似として

$$\hat{a_n^0}(\boldsymbol{x}) = \hat{C}(\boldsymbol{x}) n^{\gamma}. \tag{36}$$

が得られ, 次の定理が成り立つ.

Theorem 3. Theorem 2 の条件に加えて, $K(x)$ は有界可積分な s 次導関数をもち

$$\int x^s K^{(s)}(x) dx < \infty, \tag{37}$$

$$K^*(x) \equiv p \prod_{j=1}^{p} K(x_j) + \sum_{i=1}^{p} (\prod_{\substack{j=1 \\ j \neq i}}^{p} K(x_j)) K^{(1)}(x_i) x_i \tag{38}$$

に対し, $[0, \infty)^p$ における非増加・可積分な優関数が存在するとする. このとき, (36) で定義される $\hat{a_n^0}$ に関して次式が成り立つ.

$$U(\hat{a_n^0}) \simeq U(a_n^0) = O(n^{-2s/(2s+p)}). \tag{39}$$

Appendix C. Theorem 1 の証明

$P(\boldsymbol{z})$ は有界な定義域を持つとする. つまり, 集合

$$\{\boldsymbol{z} \mid P(\boldsymbol{z}) > 0 \text{ または } \hat{P}(\boldsymbol{z}) > 0\} \tag{40}$$

が有界集合 X に属しているとする.

$$\begin{aligned} V &= \mathrm{E}[\hat{I}(\alpha) - I(\alpha)]^2 \\ &= \mathrm{E}[\int_X Q(\boldsymbol{z}, \alpha)(\hat{P}(\boldsymbol{z}) - P(\boldsymbol{z})) d\boldsymbol{z}]^2 d\boldsymbol{z} \end{aligned} \tag{41}$$

とおくと, Cauchy-Schwarz の不等式および積分の交換操作により,

$$V \leq \int_X Q^2(\boldsymbol{z}, \alpha) d\boldsymbol{z} \int_X \mathrm{E}[\hat{P}(\boldsymbol{z}) - P(\boldsymbol{z}))]^2 d\boldsymbol{z}. \tag{42}$$

右辺の第 2 因子は (30) 式, Theorem 2 および Theorem 3 により評価できる. 従って,

$$V \leq D n^{-2s/(2s+p)} \tag{43}$$

となる. ただし

$$D = (p+2s)\int_X Q(\boldsymbol{z},\alpha)^2 d\boldsymbol{z} \int_X (\frac{\|K\|^{2p}}{2s}P(\boldsymbol{z}))^{2s\gamma}(\frac{1}{s!\sqrt{p}}\sum_i \partial_i^s P(\boldsymbol{z}))^{2p\gamma} d\boldsymbol{z}. \quad (44)$$

D は $Q(\boldsymbol{z})$, $P(\boldsymbol{z})$ に依存し, n には依存しない. Q.E.D.

さらに (40) で定義される集合が有界ではない場合を考えてみる.

$$\int_{X^c} |\hat{P}(\boldsymbol{z})| + P(\boldsymbol{z}) d\boldsymbol{z} = o(n^{-2s/(2s+p)}) \quad (45)$$

を満たす X で有界なものをとることは常にできる. 一方このとき

$$\int_X Q(\boldsymbol{z},\alpha)^2 d\boldsymbol{z} \quad (46)$$

は一般に n の関数 $g(n)$ になるが, $g(n)$ ができるだけ小さくなるように X をとることにする. すると Q が X^c の上で有界であれば,

$$V \leq O(g(n)n^{-2s/(2s+p)}) \quad (47)$$

となる.

いずれにしても評価式には

$$\int Q(\boldsymbol{z},\alpha)^2 d\boldsymbol{z} \quad (48)$$

という項が含まれ, $Q = (y - q(\boldsymbol{x},\alpha))^2$ のとき, 関数 q の形に依存する項である. これの意味はあまり明確ではないが, モデル選択法の解析で出てくる, パラメータの数とか q の表現能力といったものとは無関係であるように思われる. その根拠はパラメータの数が少なかったり表現能力が小さかったりしても上の式を大きくするような場合があるからである.