日本神経回路学会誌　Vol. 16, No. 2（2009），81–92

研究論文

バグサイズを可変とするバグ外推定による汎化能力向上[*1]

黒　木　秀　一

九州工業大学大学院 工学研究院[*2]

Improving Generalization Performance Via Out-of-Bag Estimate Using Variable Size of Bags

Shuichi Kurogi

Faculty of Engineering Kyushu Institute of Technology[*2]

This paper describes a method for improving the generalization performance by means of the out-of-bag estimate for the generalization error in regression problems. We analyze the effect of the size of bags from the viewpoint of piecewise linear prediction achieved by the CAN2 (competitive associative net). Here, the CAN2 basically is a neural net for learning efficient piecewise linear approximation of nonlinear functions. We also examine and validate the effectiveness via numerical experiments.

1. はじめに

本論文では，バグサイズを可変とするバギングによる予測誤差の性質を検討し，汎化誤差のバグ外推定により汎化能力の高い予測を行うモデル選択（パラメタの選択）が行えることを示す．アンサンブル学習のひとつであるバギング[1]は，単一の学習機械による予測のバラツキを減少させ，より安定な予測を実現するために有効な手法である．さらに汎化誤差のバグ外推定（out-of-bag estimate）はバギングの汎化誤差を小さくするモデル選択に利用することができる[2~6]．バグ外推定では，汎化誤差を見積もるためにしばしば用いられる K-fold クロスバリデーション（K-fold CV）のように汎化誤差推定のための新たな学習を行う必要がない．また K-fold CV による推定値は，K が小さいとバラツキは小さいがバイアスが大きくなり，K が大きいとバイアスは小さくなるがバラツキが大きくなるが[18]，バグ外推定ではバイアスは無いと考えられている[6]．バグ外推定は分類問題および回帰問題に適用可能であるが[2~6]，回帰問題については，予測誤差のバイアス–バリアンス分解[4]や，種々のデータセットを用いた数値実験[4~6]によりその有効性が示されているが，バグサイズが可変な場合の解析等は行われていない．

本稿では，バグサイズを可変とするバギングにおけるバグ外推定について検討する．通常はバグサイズ，すなわちバグに含まれるデータ数は与えられた訓練データ数と同じにするが，これを可変にして，汎化誤差のバグ外推定を行うとより汎化能力の高いバギングが行えることを示す．この解析のため，本稿では，競合連想ネット CAN2（Competitive Associative Net 2）にバギングを適用し，回帰問題における汎化誤差のバグ外推定を区分的線形近似の見地から検討する．ここで競合連想ネットとは競合ネット[7]と連想ネット[8]の機能を統合して構成した非線形関数を学習し区分的線形関数として近似するネットであり，その有効性は関数近似[9~11]，制御[9,12]，降水量推定[13]，時系列予測[14]などへの応用で示されている．特に国際会議 NIPS2004 の Evaluating Predictive Uncertainty Challenge においては，この手法を用いた結果，我々は regression winner に選ばれた[15]．このネットの特徴のひとつは，非線形かつ連続な基底関数を用いる多くのニューラルネットと異なり，区分的に線形な近似を行うので，線形関数に対する従来の知見や手法を非線形関数に適用することが容易になり，制御[9,12]や距離画像からの平面抽出[16]などへ応用されている．

以下，次節で CAN2 とバギング CAN2 について説明し，**3** で予測誤差の解析を区分的線形近似に基づいて行った後，**4** で数値実験結果を示し，提案手法の有

[*1] 2008年12月8日受付，2009年2月12日受理

[*2] 〒 804–8550　福岡県北九州市戸畑区仙水町 1–1

効性を示す．

2. CAN2 とバギング CAN2

2.1 データ

まず n 個の入出力データ $(\boldsymbol{x}_i, y_i)$ からなる訓練データセット $D^n \triangleq \{(\boldsymbol{x}_i, y_i)|i \in I^n\}$ が与えられるとする．ここで $I^n \triangleq \{1, 2, \cdots, n\}$ はデータを区別するための添字集合であり，k 次元入力ベクトル $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})^T$ と 1 次元出力 y_i は

$$y_i \triangleq r_i + e_i = r(\boldsymbol{x}_i) + e_i, \tag{1}$$

の関係を満たすとする．ここで $r_i \triangleq r(\boldsymbol{x}_i)$ は入力 $\boldsymbol{x}_i$ に対する出力の真値あるいは下述の学習の目標値であり，$r(\cdot)$ は非線形関数（以下では目的関数ともいう）である．e_i は平均 0，分散 σ_e^2 の観測ノイズを表わす．さらに，$\boldsymbol{x}_i$ $(i \in I^n)$ は，ある母集団データ $\boldsymbol{x}_j$ $(j \in I^{\mathrm{pop}})$ からサンプルされたものであり，独立同一分布 (i.i.d.; independently and identically distributed) に従うとする．

2.2 CAN2

CAN2 は N 個 $(N \geq 1)$ のユニットをもち，第 l ユニット（$l \in I^N = \{1, 2, \cdots, N\}$）は荷重（列）ベクトル $\boldsymbol{w}_l \triangleq (w_{l1}, \cdots, w_{lk})^T \in \mathbb{R}^{k\times 1}$ と連想行列（行ベクトル）$\boldsymbol{M}_l \triangleq (M_{l0}, M_{l1}, \cdots, M_{lk}) \in \mathbb{R}^{1\times(k+1)}$ をもち，上述の目的関数 $r(\boldsymbol{x})$ を

$$\hat{y} \triangleq \hat{y}_c \triangleq \boldsymbol{M}_c \widetilde{\boldsymbol{x}}, \tag{2}$$

により近似する（Fig. 1 参照）．ここで $\widetilde{\boldsymbol{x}} \triangleq (1, \boldsymbol{x}^T)^T$ は線形近似のバイアス項を生成するために $\boldsymbol{x}$ に 1 を付加したベクトルであり，$\boldsymbol{M}_c$ は入力ベクトル $\boldsymbol{x}$ に最近隣の荷重ベクトル

$$\boldsymbol{w}_c \triangleq \underset{\{\boldsymbol{w}_l|l\in I^N\}}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_l\|. \tag{3}$$

をもつ第 c ユニットの連想行列である．

以上の関数近似は入力空間 $V = \mathbb{R}^k$ を N 個のボロノイ領域（またはディリクレ領域），すなわち

$$V_l \triangleq \{\boldsymbol{x} \mid l = \underset{j\in I^N}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_j\|\} \tag{4}$$

に分割し，関数 $r(\boldsymbol{x})$ の区分的線形近似を行うことを意味する．またこの CAN2 の学習法として開発したバッチ学習法[11]の性能の高さは種々の応用で示されており[11, 12, 14, 15]，本稿の後述の実験でもこの学習法を用いた．

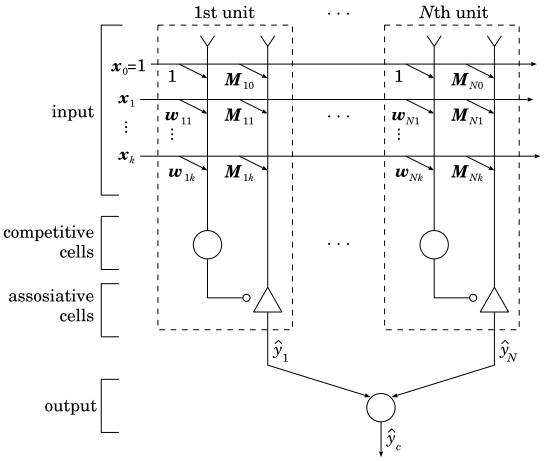


Fig. 1 Schematic diagram of the CAN2

2.3 バギング CAN2

与えられた訓練データセット D^n からの復元抽出（resampling with replacement）により生成した要素数 $n\alpha$ 個のバグ（多重集合またはブートストラップ標本集合ともいう）を $D^{n\alpha^\sharp, j} = \{(\boldsymbol{x}_i, y_i)|i \in I^{n\alpha^\sharp, j})\}$ と記す．ここで，α (> 0) はバグサイズ比，すなわちバグと訓練データセットとの要素数の比を表し，j はバグの番号を表わす．バグの個数は全部で b 個とし，$j \in J^{\mathrm{bg}} \triangleq \{1, 2, \cdots, b\}$ とする．ここでバグ $D^{n\alpha^\sharp, j}$ 内の重複しない要素のみで構成した集合 $D^{n\alpha, j}$ の要素数は約 $n(1 - e^{-\alpha})$ となることを注意しておく．これは D^n のある要素は確率 $(1 - 1/n)^{n\alpha} \simeq e^{-\alpha}$ $(n \gg 1)$ でバグ外（out-of-bag）にある（すなわちバグ $D^{n\alpha^\sharp, j}$ 内には存在しない）ことから導かれる．例えば，$\alpha = 1$ のとき $0.632n$ となり，これは従来の.632 予測器[17, 19]やバギング法[1, 5]などで通常用いられているが，本稿では可変の α を考える．

バギングによる予測は，$j \in J^{\mathrm{bg}}$ に対するバグ $D^{n\alpha^\sharp, j}$ を，それぞれある学習機械で学習した後，入力 $\boldsymbol{x}_i$ に対する目的関数値 $r_i = r(\boldsymbol{x}_i)$ を

$$\hat{y}_i^{\mathrm{bg}} \triangleq \frac{1}{b} \sum_{j\in J^{\mathrm{bg}}} \hat{y}_i^j \equiv \left\langle \hat{y}_i^j \right\rangle_{j^{\mathrm{bg}}} \tag{5}$$

により予測するものである．ここで $\hat{y}_i^j \triangleq \hat{y}^j(\boldsymbol{x}_i)$ は $D^{n\alpha^\sharp, j}$ を学習後の学習機械の入力 $\boldsymbol{x}_i$ に対する出力である．さらに，記号 $\langle\cdot\rangle$ は平均を表わし，その右下添字は平均をとる範囲，および j^{bg} は $j \in J^{\mathrm{bg}}$ を表す．

3. 予測誤差の解析

3.1 予測誤差

まず，各 $j(\in J^{\mathrm{bg}})$ に対して，ユニット数 N をもつ学習機械（CAN2）を用意し，バグ $D^{n\alpha^\sharp,j}$ 内のデータに対する二乗平均予測誤差（以下，訓練誤差という）

$$L^{\mathrm{tr},j} \triangleq \left\langle \left(\hat{y}_i^j - y_i\right)^2 \right\rangle_{i^{n\alpha^\sharp,j}} \tag{6}$$

が最小になるように学習させる．ここで $i^{n\alpha^\sharp,j}$ は $i \in I^{n\alpha^\sharp,j}$ を表し，以下の解析ではこの学習は最適に行われるとする．さらに，すべての j $(\in J^{\mathrm{bg}})$ に対する訓練誤差を

$$L^{\mathrm{tr}} \triangleq \left\langle L^{\mathrm{tr}(\theta),j} \right\rangle_{j^{\mathrm{bg}}} \tag{7}$$

により定義する．次に，バグ外データに対する二乗平均予測誤差（以下，バグ外誤差という）を

$$L^{\mathrm{ob}} \triangleq \left\langle \left(\hat{y}_i^{\mathrm{ob}} - y_i\right)^2 \right\rangle_{i^n} \tag{8}$$

により定義する．ここで i^n は $i \in I^n$ を表し，

$$\hat{y}_i^{\mathrm{ob}} \triangleq \left\langle \hat{y}_i^j \right\rangle_{j_i^{\mathrm{ob}}} \tag{9}$$

はバグ外予測，すなわち $\boldsymbol{x}_i$ を含まないバグを学習した学習機械による $y_i = r(\boldsymbol{x}_i)$ の予測 $\hat{y}_i^j$ の平均を表す．ここで j_i^{ob} は $j \in J_i^{\mathrm{ob}}$ を表し，$J_i^{\mathrm{ob}} \triangleq \{j | (\boldsymbol{x}_i, y_i) \notin D^{n\alpha^\sharp,j}, j \in J^{\mathrm{bg}}\}$ は $\boldsymbol{x}_i$ を含まないバグの添字集合を表す．バグ数 b が大きいとき，J_i^{ob} の要素数 b_i^{ob} はすべての i についてほとんど同じ値 $b_i^{\mathrm{ob}} = |J_i^{\mathrm{ob}}| \simeq b^{\mathrm{ob}} = be^{-\alpha}$ となる．

バギング予測の性能は，テストデータセット $D^{\mathrm{tst}} \triangleq \{(\boldsymbol{x}_i, y_i) \mid i \in I^{\mathrm{tst}}\}$ に対する二乗平均予測誤差

$$L^{\mathrm{tst}} \triangleq \left\langle \left(\hat{y}_i^{\mathrm{bg}} - y_i\right)^2 \right\rangle_{i^{\mathrm{tst}}} \tag{10}$$

により評価し，この L^{tst} を汎化誤差という．ここで i^{tst} は $i \in I^{\mathrm{tst}}$ であり，I^{tst} はテストデータセットに含まれるデータの添字集合である．また，テストデータセットの要素数を $n^{\mathrm{tst}} \triangleq |I^{\mathrm{tst}}|$ とする．

3.2 予測誤差のバイアス−バリアンス分解

まずバグ $D^{n\alpha^\sharp,j}$ を学習した学習機械による入力ベクトル $\boldsymbol{x}_i$ に対する訓練データ予測，バグ外予測，およびバギング予測をそれぞれ

$$\hat{y}_i^j = \begin{cases} \mu_i^{\mathrm{tr}} + \epsilon_i^{\mathrm{tr},j}, & \text{for } j \in J^{\mathrm{bg}},\ i \in I^{n\alpha^\sharp,j} \\ \mu_i^{\mathrm{ob}} + \epsilon_i^{\mathrm{ob},j}, & \text{for } j \in J_i^{\mathrm{ob}},\ i \in I^n \\ \mu_i^{\mathrm{tst}} + \epsilon_i^{\mathrm{tst},j}, & \text{for } j \in J^{\mathrm{bg}},\ i \in I^{\mathrm{tst}} \end{cases} \tag{11}$$

と表わす．ここで，μ_i^{tr}，μ_i^{ob}，μ_i^{tst} は，バグ数，すなわち J^{bg} と J_i^{ob} の要素数をそれぞれ無限大としたときの $\hat{y}_i^j$ の平均とし，それぞれの変動成分を $\epsilon_i^{\mathrm{tr},j} \triangleq \hat{y}_i^j - \mu_i^{\mathrm{tr}}$，$\epsilon_i^{\mathrm{ob},j} \triangleq \hat{y}_i^j - \mu_i^{\mathrm{ob}}$，$\epsilon_i^{\mathrm{tst},j} \triangleq \hat{y}_i^j - \mu_i^{\mathrm{tst}}$ とする．すると，訓練誤差は

$$\begin{aligned} L^{\mathrm{tr}} &= \left\langle \left(\left(\mu_i^{\mathrm{tr}} + \epsilon_i^{\mathrm{tr},j}\right) - (r_i + e_i)\right)^2 \right\rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} \\ &\simeq \left\langle (\beta_i^{\mathrm{tr}})^2 \right\rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} + \left\langle e_i^2 \right\rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} \\ &\quad + \left\langle (\epsilon_i^{\mathrm{tr},j})^2 \right\rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} - 2\left\langle \beta_i^{\mathrm{tr}} e_i \right\rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} \end{aligned} \tag{12}$$

と表すことができる．ここで $\beta_i^{\mathrm{tr}} \triangleq \mu_i^{\mathrm{tr}} - r_i$ は訓練データ予測のバイアスであり，上式の近似は変動項 e_i と $\epsilon_i^{\mathrm{tr},j}$ が統計的に独立であるとして，それらの和の項を省略して得た．学習機械はこの誤差を最小化する学習を行うので，$\langle \beta_i^{\mathrm{tr}} e_i \rangle_{i^{n\alpha^\sharp,j},j^{\mathrm{bg}}} > 0$，すなわち β_i^{tr} と e_i は正の相関をもつ可能性がある．換言すると，訓練データの予測 $\hat{y}_i^j = \beta_i^{\mathrm{tr}} + r_i + \epsilon_i^{\mathrm{tr},j}$ は e_i に過適合（overfit）する可能性がある．

一方，バグ外予測のバイアス $\beta_i^{\mathrm{ob}} \triangleq \mu_i^{\mathrm{ob}} - r_i$ と e_i は無相関，すなわち $\left\langle \beta_i^{\mathrm{ob}} e_i \right\rangle_{i^n} \simeq 0$ となり，バグ外誤差は

$$L^{\mathrm{ob}} \simeq \left\langle (\beta_i^{\mathrm{ob}})^2 \right\rangle_{i^n} + \left\langle e_i^2 \right\rangle_{i^n} + \frac{1}{b^{\mathrm{ob}}} \left\langle (\epsilon_i^{\mathrm{ob},j})^2 \right\rangle_{i^n,j_i^{\mathrm{ob}}} \tag{13}$$

と近似できる．同様に汎化誤差も

$$L^{\mathrm{tst}} \simeq \left\langle \left(\beta_i^{\mathrm{tst}}\right)^2 \right\rangle_{i^{\mathrm{tst}}} + \left\langle e_i^2 \right\rangle_{i^{\mathrm{tst}}} + \frac{1}{b} \left\langle (\epsilon_i^{\mathrm{tst},j})^2 \right\rangle_{i^{\mathrm{tst}},j_i^{\mathrm{tst}}} \tag{14}$$

と近似できる．

以上の解析は，従来のバイアス−バリアンス分解に基づいた一般の学習機械で成立する議論であるが，以下，CAN2 の区分線形近似に基づいて解析する．

3.3 区分的線形近似における予測誤差の近似

ユニット数 N の CAN2 にバグ $D^{n\alpha^\sharp,j}$ を学習させるとき，入力空間は N 個の区分領域に分割され，各区分領域 v_i $(i \in I^N)$ についての訓練誤差

$$L_{v_i}^{\mathrm{tr},j} = \frac{1}{n_{v_i}^{\mathrm{tr}}} \left\| \left(\boldsymbol{M}_{v_i} \boldsymbol{X}_{v_i}^j - \boldsymbol{Y}_{v_i}^j\right) \boldsymbol{P}_{v_i}^j \right\|^2 \tag{15}$$

を最小化する連想行列 $\boldsymbol{M}_{v_i} = \boldsymbol{M}_{v_i}^j$ として，

$$\boldsymbol{M}_{v_i}^j = \boldsymbol{Y}_{v_i}^j \boldsymbol{P}_{v_i}^j \left(\boldsymbol{X}_{v_i}^j \boldsymbol{P}_{v_i}^j\right)^+ \tag{16}$$

が求められる．ここで，v_i 内の重複しないバグ内デー

タ数を $n_{v_i}^{\rm tr} \simeq n(1-e^{-\alpha})/N$，バグ内データ $(\boldsymbol{x}_l, y_l)$ の重複数を m_l^j $(l \in I_{v_i}^{\rm tr} = \{1,2,\cdots,n_{v_i}^{\rm tr}\})$ とし，$p_l^j \triangleq m_l^j / \langle m_l^j \rangle_{l_{v_i}^{\rm tr}}$ および $\boldsymbol{P}_{v_i}^j \triangleq {\rm diag}(\cdots, p_l^j, \cdots)$ とした．また，$(\cdot)^+$ は $(\cdot)$ の一般化逆行列を表し，$\boldsymbol{X}_{v_i}^j$ と $\boldsymbol{Y}_{v_i}^j$ は，それぞれ，v_i 内の重複しないバグ内データ $(\boldsymbol{x}_l, y_l) \in D_{v_i}^{n\alpha,j} = \{(\boldsymbol{x}_l, y_l) \in D^{n\alpha,j} \mid \boldsymbol{x}_l \in v_i\}$ に対する $\tilde{\boldsymbol{x}}_l = (1, \boldsymbol{x}_l^T)^T$ と y_l を横に並べた行列である．なお，厳密には，v_i および $n_{v_i}^{\rm tr}$ はバグ $D^{n\alpha^\sharp,j}$ の j にも依存するが，ここではその変動はわずかであるとして無視する．また，m_l^j $(l=1,2,\cdots,n)$ は二項分布 $B(n\alpha, 1/n)$ に従うので平均 $E(m_l^j) = \alpha$，分散 $V(m_l^j) = \alpha(1-1/n)$ となることから，$m_l^j \neq 0$ に対する p_l^j $(l \in I_{v_i}^{\rm tr})$ の平均と分散は，

$$\mu_p \triangleq E(p_l^j) = 1 \tag{17}$$

$$\sigma_p^2 \triangleq V(p_l^j) = \frac{1}{\alpha}\left(1-\frac{1}{n}\right)\left(1-e^{-\alpha}\right) - e^{-\alpha} \tag{18}$$

となることが導かれる．さらに，$m_l^j \neq 0$ となる p_l^j のみを用いるので ${\rm rank}\left(\boldsymbol{P}_{v_i}^j\right) = n_{v_i}^{\rm tr}$ とする．

今，式 (10) の汎化誤差 $L^{\rm tst}$ を最小化する全バグ予測による区分的線形近似 $\hat{y}_i = \boldsymbol{M}_{v_i}\tilde{\boldsymbol{x}}_i$ はパラメータが $\theta = \theta^*$，区分領域が $v_i^* = v_i(\theta^*)$，連想行列が $\boldsymbol{M}_{v_i} = \boldsymbol{M}_{v_i^*}^*$ のとき達成されるとし，そのときの補正 $h_i^* \triangleq y_i - \boldsymbol{M}_{v_i^*}^*\tilde{\boldsymbol{x}}_i$ を用いて与えられた出力データを $y_i = \boldsymbol{M}_{v_i^*}^*\tilde{\boldsymbol{x}}_i + h_i^*$ と表すと，式 (16) より

$$\boldsymbol{M}_{v_i}^j = \boldsymbol{M}_{v_i^*}^* - \boldsymbol{H}_{v_i}^*\boldsymbol{P}_{v_i}^j\left(\boldsymbol{X}_{v_i}^j\boldsymbol{P}_{v_i}^j\right)^+ \tag{19}$$

が得られる．ここで $\boldsymbol{H}_{v_i}^*$ は v_i 内の重複しないバグ内データ $(\boldsymbol{x}_l, y_l) \in D_{v_i}^{n\alpha,j}$ の補正 $h_l^* \triangleq y_l - \boldsymbol{M}_{v_i^*}^*\tilde{\boldsymbol{x}}_l$ を横に並べた行列（行ベクトル）である．さらに，v_i 内のテストデータ $(\boldsymbol{x}_l, y_l) \in D^{\rm tst}$ に対する $\tilde{\boldsymbol{x}}_l$ と y_l を，それぞれ，横に並べた行列を $\boldsymbol{X}_{v_i}^{\rm tst}$ と $\boldsymbol{Y}_{v_i}^{\rm tst}$ とすると，$\boldsymbol{M}_{v_i} = \boldsymbol{M}_{v_i}^*$ は $v_i = v_i^*$ における汎化誤差 $L_{v_i}^{\rm tst} \triangleq (1/n_{v_i}^{\rm tst})\left\|\boldsymbol{M}_{v_i^*}\boldsymbol{X}_{v_i}^{\rm tst} - \boldsymbol{Y}_{v_i}^{\rm tst}\right\|^2$ を最小にする解に近いと考えられるので，$\partial L_{v_i^*}^{\rm tst}/\partial \boldsymbol{M}_{v_i^*} \simeq 0$ となり，さらに $\boldsymbol{X}_{v_i}^{\rm tst}$ の第 1 行がすべて 1 となることを用いて，

$$\langle h_l^* \rangle_{l_{v_i}^{\rm tst}} \simeq 0 \quad (v_i = v_i^* \text{ のとき}) \tag{20}$$

が得られる．さらに訓練データとテストデータが独立同一分布であり，v_i^* 内の訓練データ数 $n_{v_i^*}^{\rm tr}$ が $n_{v_i^*}^{\rm tst}$ と同程度に十分大きいならば

$$\langle h_l^* \rangle_{l_{v_i}^{\rm tr}} \simeq 0 \quad (v_i = v_i^* \text{ のとき}) \tag{21}$$

となる．

(1) バグ内データ数が小さい領域での予測誤差

区分領域 v_i 内の重複しないバグ内データ数 $n_{v_i}^{\rm tr} \simeq n(1-e^{-\alpha})/N$ が小さく，${\rm rank}\left(\boldsymbol{P}_{v_i}^j\right) = n_{v_i}^{\rm tr} \leq {\rm rank}\left(\boldsymbol{X}_{v_i}^j\right) \leq k+1$ とすると，式 (19) より

$$\boldsymbol{M}_{v_i}^j\left(\boldsymbol{X}_{v_i}^j\boldsymbol{P}_{v_i}^j\right) = \boldsymbol{M}_{v_i^*}^*\boldsymbol{X}_{v_i}^j\boldsymbol{P}_{v_i}^j + \boldsymbol{H}_{v_i^*}^*\boldsymbol{P}_{v_i}^j \tag{22}$$

となるので，両辺の各要素から，$\hat{y}_l^j = \boldsymbol{M}_{v_i}^j\tilde{\boldsymbol{x}}_l = \boldsymbol{M}_{v_i^*}^*\tilde{\boldsymbol{x}}_l + h_l^* = y_l$ となり，予測 $\hat{y}_l^j$ はバグ内データ y_l に完全に過適合する．すると，

$$L_{v_i}^{\rm tr} \simeq 0 \tag{23}$$

が得られる．このとき訓練データにおける目標出力 r_i に対する予測 $\hat{y}_i$ のバイアスは $\beta_i^{\rm tr} = \hat{y}_i - r_i = e_i$ であり，式 (13) と (14) で $\left\langle(\beta_i^{\rm tr})^2\right\rangle \simeq \left\langle(\beta_i^{\rm ob})^2\right\rangle \simeq \left\langle(\beta_i^{\rm tst})^2\right\rangle \simeq \sigma_e^2$ とすると，式 (13) と (14) より

$$L_{v_i}^{\rm ob} \simeq 2\sigma_e^2 \tag{24}$$

$$L_{v_i}^{\rm tst} \simeq 2\sigma_e^2 \tag{25}$$

となる．

(2) バグ内データ数が大きい領域での予測誤差

区分領域 v_i 内のバグ内データ数 $n_{v_i}^{\rm tr}$ が十分大きく，${\rm rank}\left(\boldsymbol{X}_{v_i}^j\right) > k+1$ となると，厳密な求解は困難になる．近似解のひとつとして，区分領域 v_i が小さい場合，$\boldsymbol{X}_{v_i}^j \simeq (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_i, \cdots)$ と置くと，式 (19) より

$$\boldsymbol{M}_{v_i}^j \simeq \boldsymbol{M}_{v_i^*}^* + \left\langle h_l^* q_l^j \right\rangle_{l_{v_i}^{\rm tr}} \frac{\tilde{\boldsymbol{x}}_i^T}{\|\tilde{\boldsymbol{x}}_i\|^2} \tag{26}$$

が得られる．ここで $q_l^j \triangleq (p_l^j)^2 / \left\langle \left(p_t^j\right)^2 \right\rangle_{t_{v_i}^{\rm tr}}$ であり，$n_{v_i}^{\rm tr}$ が十分大きいとき，その平均 $\mu_q \triangleq E(q_l^j)$ は 1 となり，分散 $\sigma_q^2 \triangleq V(q_l^j)$ は二項分布 (m_l^j) の 4 次モーメントまで[20]を用いて，

$$\sigma_q^2 \simeq \frac{(1+7\alpha+6\alpha^2+\alpha^3)}{n\alpha(1+\alpha)^2} - \frac{1}{n(1-e^{-\alpha})} \tag{27}$$

が導かれる．上式 (26) により，ある入力 $\boldsymbol{x}_s$ $(\in v_i)$ に対するこの第 j バグによる予測 $\hat{y}_s^j = \boldsymbol{M}_{v_i}^j\boldsymbol{x}_s$ は

$$\hat{y}_s^j \simeq \boldsymbol{M}_{v_i^*}^*\boldsymbol{x}_s + \left\langle h_l^* q_l^j \right\rangle_{l_{v_i}^{\rm tr}} \tag{28}$$

となる．ここで，$\hat{y}_s^j$ をすべての j についての平均 $\mu_s^\theta = \left\langle y_s^j \right\rangle_{j^\infty}$ とその変動成分 $\epsilon_s^{\theta,j} = \hat{y}_s^j - \mu_s^\theta$ に分解すると $\hat{y}_s^j = \mu_s^\theta + \epsilon_s^{\theta,j}$ と書け，

$$\mu_s^\theta \simeq \boldsymbol{M}_{v_i^*}^*\tilde{\boldsymbol{x}}_s + \langle h_l^* \rangle_{l_{v_i}^{\rm tr}} \tag{29}$$

$$\epsilon_s^{\theta,j} \simeq \left\langle h_l^* \Delta q_l^j \right\rangle_{l_{v_i}^{\rm tr}} \tag{30}$$

となる．ここで $\Delta q_l^j = q_l^j - 1$ である．さて式 (29) において，$\boldsymbol{x}_s$ がバグ内入力のときは s が $I_{v_i}^{\mathrm{tr}}$ の中のひとつ，すなわち $s \in I_{v_i}^{\mathrm{tr}}$ であり，$\boldsymbol{x}_s$ がバグ外入力またはテスト入力の場合は $s \notin I_{v_i}^{\mathrm{tr}}$ であると解釈できる．すると，入力 $\boldsymbol{x}_s$ に対する出力を $y_s = \boldsymbol{M}_{v_i^*}^* \tilde{\boldsymbol{x}}_s + h_s^*$ として，区分領域 v_i における訓練誤差 $L_{v_i}^{\mathrm{tr}} \triangleq \left\langle \|\hat{y}_s^j - y_s\|^2 \right\rangle_{s_{v_i}^{n\alpha^\sharp, j, j^{\mathrm{bg}}}}$，バグ外誤差 $L_{v_i}^{\mathrm{ob}} \triangleq \left\langle \|\hat{y}_s^{\mathrm{ob}} - y_s\|^2 \right\rangle_{s_{v_i}^n}$ および汎化誤差 $L_{v_i}^{\mathrm{tst}} \triangleq \left\langle \|\hat{y}_s^{\mathrm{bg}} - y_s\|^2 \right\rangle_{s_{v_i}^{\mathrm{tst}}}$ の期待値を計算して，

$$L_{v_i}^{\mathrm{tr}} \simeq c_{v_i,0}^{\mathrm{tr}} - \frac{c_{v_i,1}^{\mathrm{tr}}}{n_{v_i}^{\mathrm{tr}}} + \left(\rho_{h*}^{\mathrm{tr}} \sigma_q \lambda_{h*q}^{\mathrm{bg}}\right)^2 \tag{31}$$

$$L_{v_i}^{\mathrm{ob}} \simeq c_{v_i,0}^{\mathrm{ob}} + \frac{c_{v_i,1}^{\mathrm{ob}}}{n_{v_i}^{\mathrm{tr}}} + \frac{\left(\rho_{h*}^{\mathrm{tr}} \sigma_q \lambda_{h*q}^{\mathrm{ob}}\right)^2}{b^{\mathrm{ob}}} \tag{32}$$

$$L_{v_i}^{\mathrm{tst}} \simeq c_{v_i,0}^{\mathrm{tst}} + \frac{c_{v_i,1}^{\mathrm{tst}}}{n_{v_i}^{\mathrm{tr}}} + \frac{\left(\rho_{h*}^{\mathrm{tr}} \sigma_q \lambda_{h*q}^{\mathrm{bg}}\right)^2}{b} \tag{33}$$

が得られる（**付録 A** 参照）．ここで，各誤差の第 1 項は定数項，第 2 項は $n_{v_i}^{\mathrm{tr}}$ に反比例する項，第 3 項はバグの変動成分の影響を受ける項である．以下，各誤差の第 2 項を詳しく検討する．まず，**付録 A** より，

$$\frac{c_{v_i,1}^{\mathrm{tr}}}{n_{v_i}^{\mathrm{tr}}} \simeq \frac{c_{v_i,1}^{\mathrm{ob}}}{n_{v_i}^{\mathrm{tr}}} \simeq \frac{c_{v_i,1}^{\mathrm{tst}}}{n_{v_i}^{\mathrm{tr}}} \simeq L_{v_i}^{\mathrm{fit}} \triangleq \langle h_s^* \rangle_{s_{v_i}^{\mathrm{tr}}}^2 \tag{34}$$

と置くことができ，h_s^* が異なる $s \in I_{v_i}^{\mathrm{tr}}$ に対して統計的に独立であるとすると

$$L_{v_i}^{\mathrm{fit}} = \langle h_s^* \rangle_{s_{v_i}^{\mathrm{tr}}}^2 \simeq \frac{(\rho_{h*}^{\mathrm{tr}})^2}{n_{v_i}^{\mathrm{tr}}} \quad (n_1 \le n_{v_i}^{\mathrm{tr}} \le n_2 \text{のとき}) \tag{35}$$

となる．ここで上式が成立する下限 n_1 は $n_1 \gtrsim k+1$，上限 n_2 は以下の検討より $n_2 \lesssim n_{v_i^*}^{\mathrm{tr}}$ を満たすと考えられる．なお上式の $L_{v_i}^{\mathrm{fit}}$ は**付録 A** の導出過程より，予測 y_s^j の訓練データ y^j への過適合度を表しているとみることができる．

つぎに，h_s^* $(s \in I_{v_i}^{\mathrm{tr}})$ の統計的独立性は，式 (21) より，$v_i = v_i^*$ あるいは $\theta = \theta^*$ または $n_{v_i}^{\mathrm{tr}} \simeq n(1-e^{-\alpha})/N = n_{v_i^*}^{\mathrm{tr}}$ のとき消失して

$$L_{v_i}^{\mathrm{fit}} = L_{v_i^*}^{\mathrm{fit}} = \langle h_s^* \rangle_{s_{v_i^*}^{\mathrm{tr}}}^2 \simeq 0 \ (n_{v_i}^{\mathrm{tr}} = n_{v_i^*}^{\mathrm{tr}} \text{ のとき}) \tag{36}$$

となる．さらに $n_{v_i}^{\mathrm{tr}} > n_{v_i^*}^{\mathrm{tr}}$ のときは，

$$L_{v_i}^{\mathrm{fit}} \ge L_{v_i^*}^{\mathrm{fit}} (\simeq 0) \quad (n_{v_i}^{\mathrm{tr}} > n_{v_i^*}^{\mathrm{tr}} \text{ のとき}) \tag{37}$$

となるが，次の 2 つの場合が考えられる．まず，目的関数 $r(\boldsymbol{x})$ が N^* 個の区分的線形関数で表せる場合，$N = N^*$ と固定して $\alpha > \alpha^*$ あるいは $n_{v_i}^{\mathrm{tr}} > n_{v_i^*}^{\mathrm{tr}}$ とすると，各区分領域には線形の訓練データが必要数以上あるので，$L_{v_i}^{\mathrm{fit}} \simeq L_{v_i^*}^{\mathrm{fit}}$ が維持されると考えられる．一方，$r(\boldsymbol{x})$ が $N = N^*$ 個の区分的線形関数で表せない場合，$N = N^*$ として $n_{v_i}^{\mathrm{tr}} > n_{v_i^*}^{\mathrm{tr}}$ とすると，各区分領域内で非線形な訓練データ数が多くなり区分的線形近似による予測誤差は大きくなり，$L_{v_i}^{\mathrm{fit}} > L_{v_i^*}^{\mathrm{fit}}$ になると考えられる．

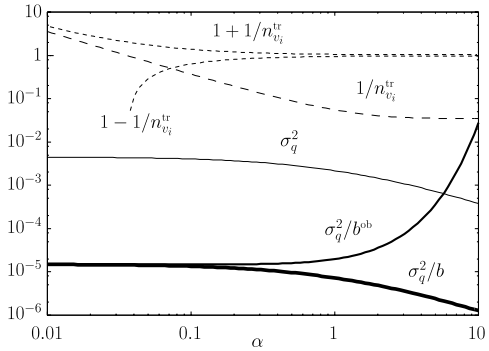


Fig. 2　Component values of the regional losses for $b = 300$, $n = 1000$, $N = 35$, $n_{v_i}^{\mathrm{tr}} = n(1-e^{-\alpha})/N$ and $b^{\mathrm{ob}} = be^{-\alpha}$.

ここで参考のため後述の実験でのパラメタ $n = 1000$, $N = 35$, $b = 300$ における α の変化に対する各誤差に含まれる変数の変化の様子を Fig. 2 に示す．この図からまず $\sigma_q^2 < 10^{-2} \ll 1$ であることが分かる．次に式 (31)〜(33) で $c_{v_i,0}^{\mathrm{tr}} \simeq c_{v_i,0}^{\mathrm{ob}} \simeq c_{v_i,0}^{\mathrm{tst}} \simeq c_{v_i,1}^{\mathrm{tr}} \simeq c_{v_i,1}^{\mathrm{ob}} \simeq c_{v_i,1}^{\mathrm{tst}} \simeq (\rho_{h*}^{\mathrm{tr}})^2$ とすると，各式の右辺第 1 項と第 2 項の和は $L_{v_i}^{\mathrm{tr}} \simeq (1 - 1/n_{v_i}^{\mathrm{tr}})(\rho_{h*}^{\mathrm{tr}})^2$ および $L_{v_i}^{\mathrm{ob}} \simeq L_{v_i}^{\mathrm{tst}} \simeq (1 + 1/n_{v_i}^{\mathrm{tr}})(\rho_{h*}^{\mathrm{tr}})^2$ となり，右辺第 3 項は無視できることが分かる．

3.4　バグサイズ比 α の増加に対する予測誤差

以下，L^{tst} を最小化するパラメタ $\theta = (N, \alpha)$ を $\theta^* = (N^*, \alpha^*)$ とし，$N = N^*$ は固定し，α を 0 から増加していく場合を考える．まず，α が非常に小さいとき，ほとんどの区分領域において，重複しないバグ内データ数 $n_{v_i}^{\mathrm{tr}} \simeq n(1-e^{-\alpha})/N^*$ は $k+1$ より小さいので，式 (23) より，$L_{v_i}^{\mathrm{tr}} \simeq 0$ となり，$L^{\mathrm{tr}} \simeq 0$ となる．このとき，いくつかの区分領域では学習すべきバグ内データ数が 0 となるが，何も学習していない場合の予測を $\hat{y}_i^j = 0$ とすると，バグ外誤差と汎化誤差は与えられたデータの二乗平均，すなわち $L^{\mathrm{ob}} \simeq \left\langle (\hat{y}_i)^2 \right\rangle_{i^n}$ と $L^{\mathrm{tst}} \simeq \left\langle (\hat{y}_i)^2 \right\rangle_{i^{\mathrm{tst}}}$ となる．

次に，α を大きくしていくと，式 (31)〜(33) が成立する領域数が増加する．これら式の右辺第 2 項は，$n_{v_i}^{\mathrm{tr}} \simeq n(1-e^{-\alpha})/N^*$ とすると，$L_{v_i}^{\mathrm{tr}}$ については増加

し，$L_{v_i}^{\rm ob}$ と $L_{v_i}^{\rm tst}$ については減少する．この変化は，前節の検討より，$k+1 \lesssim n_1 \le n_{v_i}^{\rm tr} \simeq n(1-e^{-\alpha})/N^* \le n_2 \lesssim n_{v_i^*}^{\rm tr}$ となる α の範囲で成立し，さらに α が大きくなると式 (37) が成立するようになる．この式 (37) で $L_{v_i}^{\rm fit} > L_{v_i^*}^{\rm fit}$ となるとき $L_{v_i}^{\rm fit}$ は $\alpha = \alpha^*$ で極小値をとる．α がさらに大きくなると $b^{\rm ob} \simeq be^{-\alpha}$ が小さくなるので，式 (32) の右辺第 3 項が大きくなり，$L^{\rm ob}$ は $L^{\rm tst}$ よりも大きくなる．

以上の説明の妥当性を検討するため，各誤差が区分領域と全領域で，それぞれ，ほぼ同様に変化すると考え，後述の実験結果（Fig. 4 の実線）を式 (31)〜(35) により回帰すると Fig. 4 の破線が得られた．Fig. 4 の各図から，各誤差は $\alpha = \alpha^*$(または $\alpha = \alpha^{\rm ob}$) の近傍でうまく近似されており，上述の説明が妥当であることが分かる．

3.5 ユニット数 N の増加に対する予測誤差

バグサイズ比 $\alpha = \alpha^*$ を固定し，区分領域数 N を 1 から増加していく場合を考える．ここで n および $n^{\rm tst}$ は十分大きく，$N^* > 1$ とする．まず，$N = 1$ のときの CAN2 は線形予測器になり，n および $n^{\rm tst}$ が十分大きいならば $L_{v_i}^{\rm tr} \simeq L_{v_i}^{\rm ob} \simeq L_{v_i}^{\rm tst}$ となる．

次に，N を増加させると，各区分領域 v_i の大きさが小さくなり，より複雑な関数近似が可能になるので，各誤差は減少する．しかし，さらに N を大きくしていくと，バグ内データ数 $n_{v_i}^{\rm tr} \simeq n(1-e^{-\alpha})/N$ が減少するので，過適合の式 (31)〜(33) が成立するようになり，N の増加に対し，$L_{v_i}^{\rm tr}$ は減少し，$L_{v_i}^{\rm ob}$ と $L_{v_i}^{\rm tst}$ は増加する．ここで $N = N^*$ で $L_{v_i}^{\rm tst}$ が最小値をとるので，N^* は式 (33) が成立する最も小さな N であると解釈できる．N がさらに大きくなると，$n_{v_i}^{\rm tr} < k+1$ となる領域がでてきて，最も過適合した式 (23)〜(25) が成立するようになる．

4. 数値実験と考察

4.1 目的関数とデータセット

目的関数 $r(\boldsymbol{x})$ として，視覚的に理解しやすい 1 入力 1 出力の関数（文献[21]および Fig. 3 参照）

$$r(x) = \sin(5x) + \sin(15x) + \sin(25x) \qquad (38)$$

を用いて，バグ外推定による最適パラメータの探索を行い，その有効性と前節の解析の妥当性を検討する．まず，区間 $[0,1]$ の一様乱数により入力 $x = x_i$ を生成して上式の $r(x)$ に適用し，平均 0，分散 $\sigma_e^2 = 0.04$ のガウス性ノイズを式 (1) の e_i として加え，種々のデータ数 n に対する訓練データセット $D^n = \{(x_i, y_i) | i \in I^n\}$ を生成した．同様に，入力を $x = x_i = i/1000$ ($i = 0, 1, 2, \cdots, 999$) としてデータ数 $n^{\rm tst} = 1000$ のテストデータセット $D^{\rm tst}$ を生成した．またバグ数は十分大きな数として $b = 300$ とした．

Fig. 3 に目的関数 $r(x)$，$n = 1000, 100, 10$ に対する訓練データ $(\boldsymbol{x}_i, y_i)$，および最適パラメタ $\theta^* = (N^*, \alpha^*)$ とバグ外推定パラメタ $\theta^{\rm ob} = (N^{\rm ob}, \alpha^{\rm ob})$ を用いた学習により得られた予測 $\hat{y}_i^{\rm bs}$ を示す．なおパラメタ $\theta = (N, \alpha)$ の探索は N は 1 刻み，α は 0.1 刻みで行った．この図より $n = 1000$ と 100 については θ^* と $\theta^{\rm ob}$ による予測の違いはほとんど無いと考えられる．$n = 10$ については各予測の違いは大きいが，ノイズの分散 $\sigma_e^2 = 0.04$ などの事前情報が得られない状況では，$\theta^{\rm ob}$ による予測も θ^* による予測と同じく妥当ではないかと考えられる．

4.2 パラメタ $\boldsymbol{\theta = (N, \alpha)}$ の変化に対する誤差

Fig. 4 に，$n = 1000, 100, 50, 10$ に対して $N = N^*$ および $N = N^{\rm ob}$ としたときの，α の変化に対する $L^{\rm tr}$，$L^{\rm ob}$，$L^{\rm tst}$ の変化を示す．また，式 (31)〜(35) の定数を最小二乗法により求めて再現した回帰誤差を同図の破線 $\hat{L}^{\rm tr}$，$\hat{L}^{\rm ob}$，$\hat{L}^{\rm tst}$ で表す．さらに Fig. 5 に，$n = 1000, 100, 50, 10$ に対して $N = N^*$ および $N = N^{\rm ob}$ としたときの，N の変化に対する $L^{\rm tr}$，$L^{\rm ob}$，$L^{\rm tst}$ の変化を示す．これらの図より，各誤差は **3.4** および **3.5** で説明した変化をしていることが分かる．なお，Fig. 4 で α が小さいとき，$\hat{L}^{\rm tr}$，$\hat{L}^{\rm ob}$，$\hat{L}^{\rm tst}$ とそれぞれの実験値との差が大きいのは，式 (31)〜(35) は α が十分大きくないと成立しないからである．

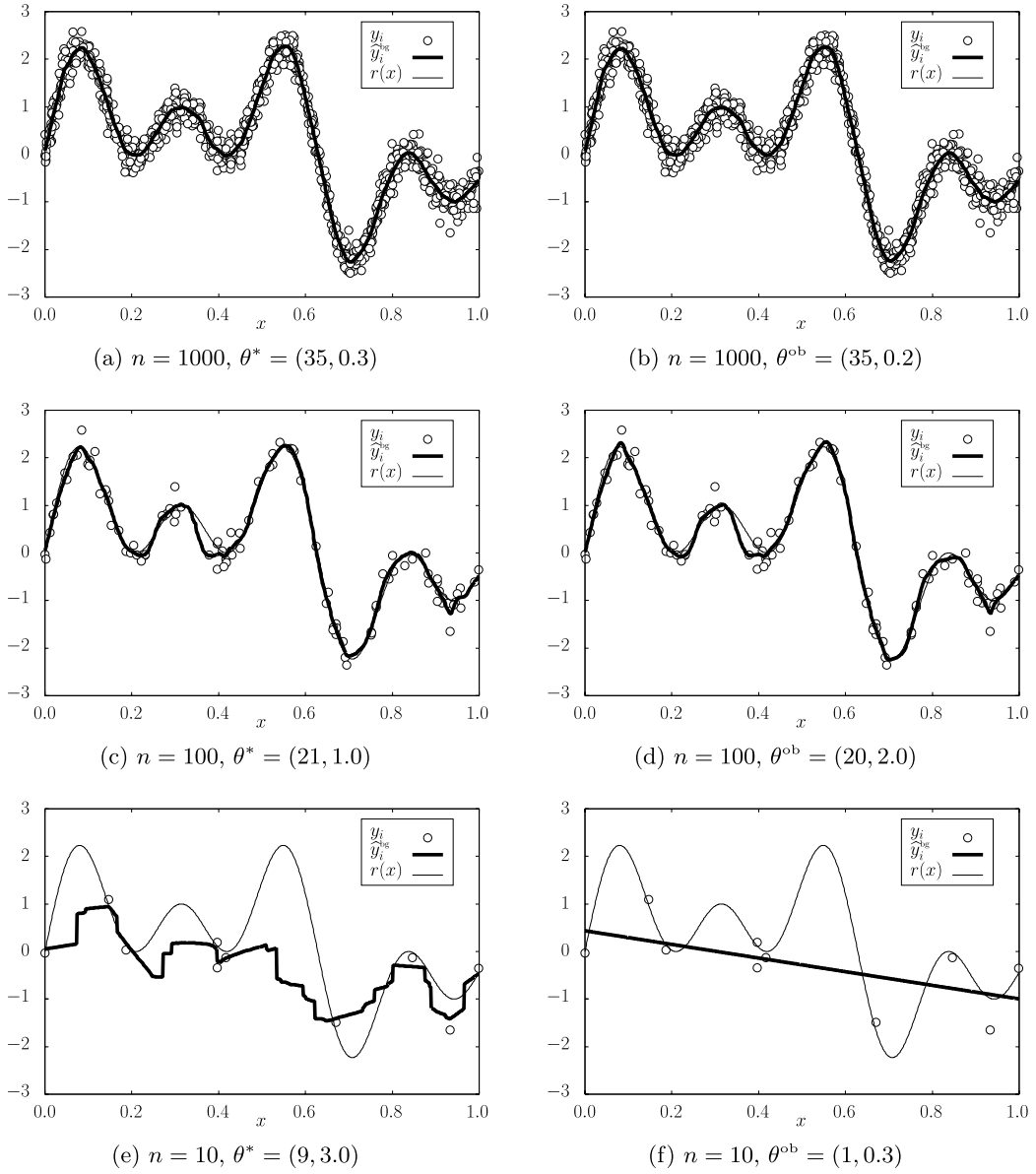
以下，バグ外推定により得られる $\theta^{\rm ob}$，θ^*，$L^{\rm ob}$，$L^{\rm tst}$ について，さらに詳細に検討する．

4.3 バグ外推定と汎化能力向上

パラメタ $\theta = (N, \alpha)$ が θ^*，$\theta^{\rm ob}$ および $\theta^{\rm ob1}$ であるときの誤差 $L^{\rm tr}, L^{\rm ob}, L^{\rm tst}$ を，異なるデータ数 n に対して求めた結果を Table 1 に示す．ここで，$\theta^{\rm ob1} = (N^{\rm ob1}, 1)$ は，通常のバグ外推定，すなわちバグ外誤差 $L^{\rm ob}$ を最小化するパラメータを求めるとき α は 1 に固定して N だけを最適化する手法により求めたものである．また各誤差を最小にする N と α の探索は，N については 1 刻み，α については 0.1 刻みで離散的に求めた．以下，この表の結果を用いて，種々の見地から検討する．

4.3.1 データ数 n に対する汎化誤差 $L^{\rm tst}$

Table 1 より，θ^*，$\theta^{\rm ob}$，$\theta^{\rm ob1}$ により得られた汎化誤差 $L^{\rm tst}$ はデータ数 n の増加に対してそれぞれ減少していることが分かる．これは，各パラメタ探索法は，（優劣は別にして）それぞれ，各 n に対する訓練データセットに対して，汎化能力を向上するように機能したことを示唆する．ただし，θ^* と $\theta^{\rm ob}$ については $n = 50$ の



(a) $n = 1000$, $\theta^* = (35, 0.3)$ (b) $n = 1000$, $\theta^{\text{ob}} = (35, 0.2)$

(c) $n = 100$, $\theta^* = (21, 1.0)$ (d) $n = 100$, $\theta^{\text{ob}} = (20, 2.0)$

(e) $n = 10$, $\theta^* = (9, 3.0)$ (f) $n = 10$, $\theta^{\text{ob}} = (1, 0.3)$

Fig. 3 Prediction by the CAN2s with θ^* and θ^{ob}. The open circles represent the given data $(x_i, y_i) \in D^n$, the thick lines the bagging prediction $\hat{y}_i^{\text{bg}}$, and the thin lines the underlying function $r(x)$.

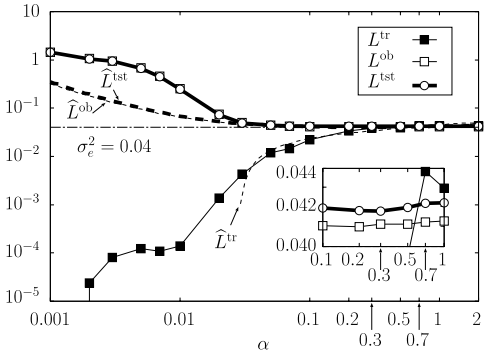
ときは上記性質の例外となっているが，与えられた訓練データセットまたはバギング学習に含まれる確率的変動による影響ではないかと考えられる．

4.3.2 データ数 n に対する N^*，N^{ob}，N^{ob1}

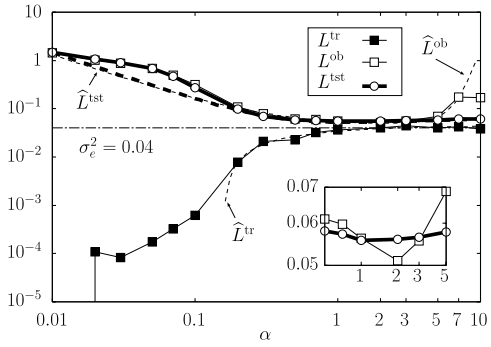
Table 1 より，n の減少に対して，最適なユニット数 N^*，N^{ob}，N^{ob1} はそれぞれ減少する傾向があることが分かる．これは，n が小さいならば，過適合を抑えるために N を小さくすべきこと，逆に n が十分大きいならば，より大きな N はより複雑な関数近似を行えること，から理解できる．

4.3.3 データ数 n に対する α^* と α^{ob}

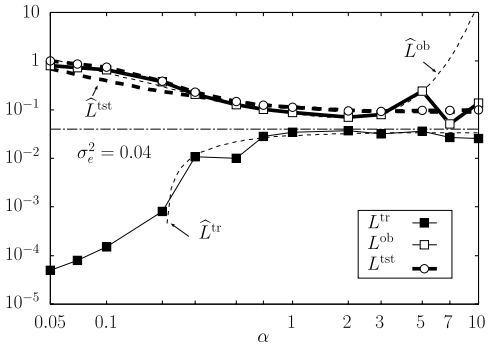
Table 1 より，$n \geq 150$ のとき，n の増加に対して，バグサイズ比 α^* と α^{ob} は減少していることが分かる．これは次の理由によるのではないかと考えられる．ま



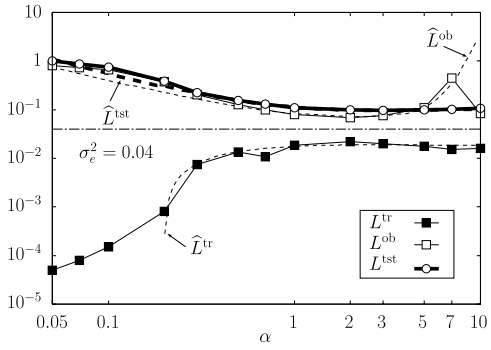
(a) $n = 1000$, $N = 35$ ($\alpha^* = 0.3$, $\alpha^{\rm ob} = 0.2$)



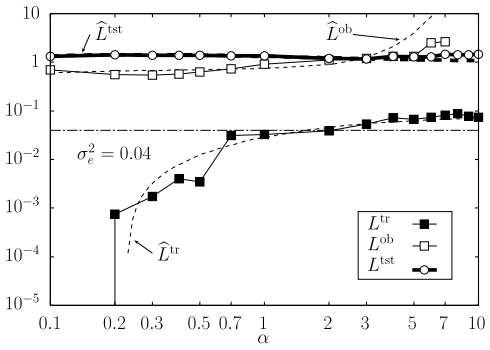
(b) $n = 100$, $N = 20$ ($\alpha^* = 1.0$, $\alpha^{\rm ob} = 2.0$)



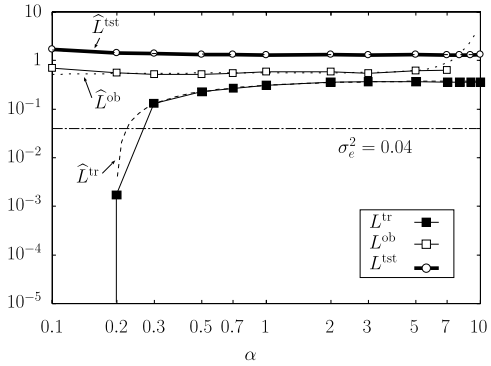
(c) $n = 50$, $N = 18$ ($\alpha^* = 3.0$)



(d) $n = 50$, $N = 21$ ($\alpha^{\rm ob} = 2.0$)



(e) $n = 10$, $N = 9$ ($\alpha^* = 3.0$)



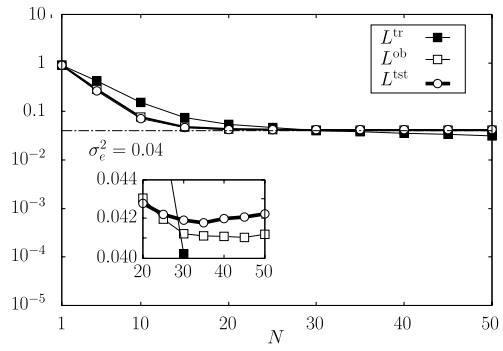
(f) $n = 10$, $N = 1$ ($\alpha^{\rm ob} = 0.3$)

Fig. 4 Losses $L^{\rm tr}$, $L^{\rm ob}$, $L^{\rm tst}$ and the regression losses $\hat{L}^{\rm tr}$, $\hat{L}^{\rm ob}$, $\hat{L}^{\rm tst}$ vs. α

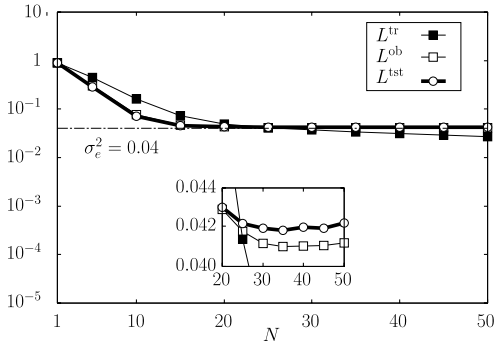
ず N あるいは v_i を固定して訓練データ数 $n^{\rm tr}_{v_i}$ を十分大きくすると，過適合も十分小さくなり，$L^{\rm tr}_{v_i}$ はある値に収束していくと考えられる．そこで式 (31) と式 (33) で過適合の項を $c^{\rm tr}_{v_i,1} \simeq c^{\rm tst}_{v_i,1} \simeq 0$ とし，さらに定数項 $c^{\rm tr}_{v_i,0} \triangleq \left\langle (h^*_l)^2 \right\rangle_{l^{\rm tr}_{v_i}}$ と $c^{\rm tr}_{v_i,0} \triangleq \left\langle (h^*_l)^2 \right\rangle_{l^{\rm tst}_{v_i}}$ については $n^{\rm tr}_{v_i}$ と $n^{\rm tst}_{v_i}$ が十分大きいとして $c^{\rm tr}_{v_i,0} \simeq c^{\rm tst}_{v_i,0}$ とすると

$$L^{\rm tst}_{v_i} \simeq L^{\rm tr}_{v_i} + (\rho^{\rm tr}_{h*}\sigma_q\lambda^{\rm bg}_{h*q})^2 \left(\frac{1}{b} - 1\right) \tag{39}$$

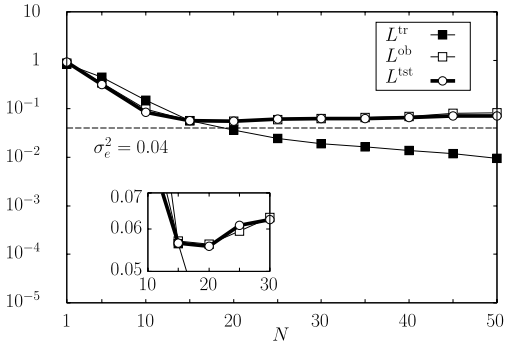
が得られる．ここで α の増加に対して，$L^{\rm tr}_{v_i}$, $\rho^{\rm tr}_{h*}$, $\lambda^{\rm bg}_{h*q}$



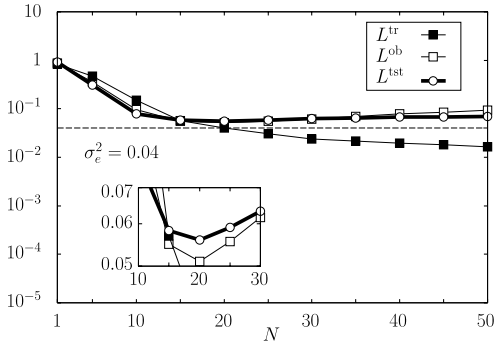
(a) $n = 1000$, $\alpha = 0.3$ ($N^* = 35$)



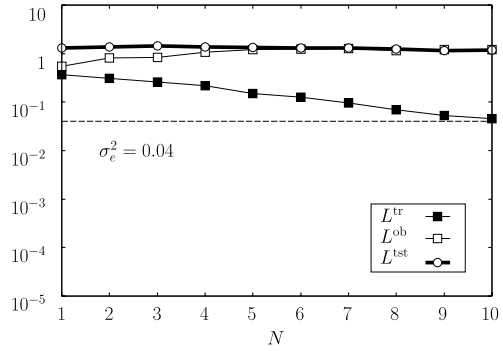
(b) $n = 1000$, $\alpha = 0.2$ ($N^{\rm ob} = 35$)



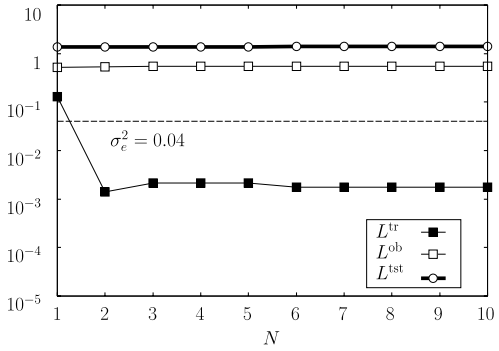
(c) $n = 100$, $\alpha = 1.0$ ($N^* = 20$)



(d) $n = 100$, $\alpha = 2.0$ ($N^{\rm ob} = 20$)



(g) $n = 10$, $\alpha = 3.0$ ($N^* = 9$)



(h) $n = 10$, $\alpha = .03$ ($N^{\rm ob} = 1$)

Fig. 5　Losses vs. N

がそれぞれ一定であるとすると，$1/b - 1 < 0$ であるのでより大きな σ_q^2 が $L_{v_i}^{\rm tst}$ を小さくすることになる．これはバギングなどアンサンブルの手法においては，より大きな多様性（diversity）がより高い汎化能力を与えるという知見[22]に対応している．この σ_q^2 は α の増加に対して減少するので（式 (27) あるいは Fig. 2 参照），$L_{v_i}^{\rm tr}$ がほぼ一定の範囲においてより小さな α^* が得られる．さらに N を固定して α を増加させると，n が大きいならば，小さな α でもバグ内データ数 $n_{v_i}^{\rm tr} \simeq n(1-e^{-\alpha})/N$ は大きくなり，$L_{v_i}^{\rm tr}$ がほぼ一定になる．以上により，n の増加に対して α^* が減少すると考えられる．

式 (39) と同様に，バグ外誤差の近似式

Table 1 Losses L^{tr}, L^{ob} and L^{tst} for $\theta=\theta^*$, θ^{ob} and θ^{ob1}. The unit of the losses is 10^{-2}.

	optimal (θ^*)					out-of-bag (θ^{ob})					conventional out-of-bag (θ^{ob1})			
n	N^*	α^*	L^{tr}	L^{ob}	L^{tst}	N^{ob}	α^{ob}	L^{tr}	L^{ob}	L^{tst}	N^{ob1}	L^{tr}	L^{ob}	L^{tst}
1000	35	0.3	3.819	4.110	*4.177*	35	0.2	3.397	4.097	**4.179**	40	3.989	4.124	4.225
500	30	0.4	3.795	4.573	*4.328*	25	0.3	3.937	4.558	**4.372**	35	4.100	4.602	4.378
400	30	0.5	3.379	4.354	*4.350*	25	0.3	3.489	4.284	**4.390**	26	3.979	4.368	4.391
200	26	0.6	2.637	4.447	*4.625*	26	0.5	2.470	4.300	**4.654**	30	3.130	4.343	4.724
160	22	0.7	3.295	4.490	*4.819*	19	0.6	3.521	4.425	**4.851**	29	2.573	4.400	4.872
150	26	0.8	2.442	4.827	*4.916*	20	0.7	3.660	4.540	**5.012**	20	4.337	4.782	5.022
100	20	1.0	3.629	5.617	*5.566*	20	2.0	3.497	5.106	5.593	20	3.629	5.617	**5.566**
90	16	1.1	5.107	5.569	*5.693*	17	1.0	3.787	5.326	**5.877**	17	3.787	5.326	**5.877**
80	17	1.0	3.350	5.811	*5.927*	17	1.0	3.350	5.811	**5.927**	17	3.350	5.811	**5.927**
70	15	1.4	4.337	5.498	*6.869*	15	1.5	4.575	5.372	**7.112**	16	4.212	5.796	7.146
60	14	1.3	4.347	7.252	*8.117*	19	1.3	2.369	6.859	**8.645**	20	2.427	7.743	8.679
50	14	3.0	3.464	8.067	*9.034*	18	2.0	2.024	6.479	**9.286**	21	1.618	7.816	11.19
40	13	3.4	4.507	11.48	*8.303*	16	3.0	3.026	10.76	**9.082**	30	0.629	12.73	13.67
30	15	3.0	5.872	21.36	*14.28*	15	3.0	5.872	21.36	**14.28**	27	0.491	24.09	17.15
20	18	3.0	1.341	50.65	*36.37*	19	1.0	0.576	44.66	**45.06**	19	0.576	44.66	**45.06**
10	9	3.0	5.328	119.5	*116.4*	1	0.3	13.00	52.02	138.0	1	30.79	58.36	**130.5**

$$L_{v_i}^{\mathrm{ob}} \simeq L_{v_i}^{\mathrm{tr}} + (\rho_{h*}^{\mathrm{tr}}\sigma_q\lambda_{h*q}^{\mathrm{ob}})^2\left(\frac{1}{b^{\mathrm{ob}}}-1\right) \tag{40}$$

が得られ，$b^{\mathrm{ob}} \simeq be^{-\alpha}$ とし，b が十分大きいとすると，$1/b^{\mathrm{ob}}-1$ は負でその絶対値は α の増加とともに減少する．従って，α^* と同様に，n の増加に対して α^{ob} が減少することが導かれる．

4.3.4 提案手法と通常手法の比較

4.3.3 の検討は，固定の $\alpha=1$ を使用する通常手法は n がある程度大きいとき有効でないことを示唆する．実際には，Table 1 より，$N=N^{\mathrm{ob}}$ と $\alpha=\alpha^{\mathrm{ob}}$ を探索する提案手法は，$\alpha=1$ として $N=N^{\mathrm{ob1}}$ だけを探索する通常手法よりも，ほとんどの場合，小さな L^{tst} を探索できたことが分かる．この結果をより詳細に検討するため，N と α の機能について考える．まず，最適な $N=N^*$ および $\alpha=\alpha^*$ から N を大きくあるいは α を小さくすることは共に，$n_{v_i}^{\mathrm{tr}} \simeq n(1-e^{-\alpha})/N$ を小さくして，過適合を大きくし，L^{tst} を増大させることになる．一方，区分領域数 N を $N=N^*$ より小さくすることは，関数近似能力を減少させ L^{tst} を増大させることになる．これに対してバグサイズ比 α を $\alpha=\alpha^*$ より大きくすることは，式 (39) において σ_q^2 を小さくし，$L_{v_i}^{\mathrm{tst}}$ を増大させることになる．これは N と α の機能が異なることを意味し，L^{tst} を最小化する解 $\theta^*=(N^*,\alpha^*)$ は唯一であることを示唆する．従って，α を 1 に固定する通常手法により，N と α を最適する解 θ^* に近い解を求めることは $\alpha^* \simeq 1$ 以外の場合では困難であることになる．

5. 結　　論

本稿では，バグサイズを可変としたバギングにおける訓練誤差，バグ外誤差および汎化誤差の性質を区分的線形予測の見地から検討し，バグ外推定によるバグサイズ調整の有効性を示した．さらに，区分的線形予測を行う CAN2 を用いた数値実験により，バグサイズを訓練データセットのサイズと同じにする通常のバギングよりも，バグサイズを調整するバギングの方が，多くの場合，小さな汎化誤差を実現できることを示した．特に訓練データ数が大きい場合は，バグサイズ比を小さくした方が汎化誤差が小さくなるという性質は，一般の学習機械におけるバギングでも成立するのではないかと考えられるが，詳しい検討は今後の課題である．

参 考 文 献

1) Breiman, L. (1996): Bagging predictors, Machine Learning, Vol. 26, No. 2, pp.123–140
2) Breiman, L. (1996): Out-of-bag estimation, Technical Report, Berkley, California: Department of Statistics, University of California
3) Tibshirani, R. (1996): Bias, variance and prediction error for classification rules, Technical

Report, Statistics Department, University of Toronto
4) Wolpert, D.H. and Macready, W.G. (1997): An efficient method to estimate Bagging's generalization error, Machine Learning, Vol. 35, No. 1, pp.41–55
5) Carney, J.G. and Cunningham, P. (1999): The NeuralBAG algorithm: Optimizing generalization performance in bagged neural networks, Proceedings of ESANN'1999, pp.21–23
6) Breiman, L. (2001): Random forests, Machine Learning, Vol. 45, No. 1, pp.5–32
7) Ahalt, A.C., Krishnamurthy, A.K., Chen, P. and Melton, D.E. (1990): Competitive learning algorithms for vector quantization, Neural Networks, Vol. 3, pp.277–290
8) Kohonen, T. (1977): Associative Memory. Springer Verlag
9) Kurogi, S. and Ren, S. (1997): Competitive associative network for function approximation and control of plants, Proc. NOLTA'97, pp.775–778
10) 黒木秀一 (2003)：競合連想ネットの漸近最適性と比線型関数の逐次学習への応用, 電子情報通信学会論文誌 (D-II), Vol. J86-D-II, No. 2, pp.184–194
11) 黒木秀一, 西田　健, 渕川康裕 (2006)：バッチ学習型競合連想ネットとその性質, 計測自動制御学会論文集, Vol. 42, No. 8, pp.916–925
12) Kurogi, S., Araki, N., Miyamoto, H., Fuchikawa, Y. and Nishida, T. (2004): Temperature control of RCA cleaning solutions using batch learning competitive associative net, Proc. SCI2004 Vol. 5, pp.18–23
13) 黒木秀一, 藤　　誠, 寺田精一 (2001)：競合連想ネットを用いる降水量推定, 2001 年電子情報通信学会総合大会講演論文集, Vol. SD-1, pp.260–261
14) Kurogi, S., Ueno, T. and Sawa, M. (2004): Batch learning competitive associative net and its application to time series prediction, Proc. of IJCNN 2004, in CD-ROM
15) `http://predict.kyb.tuebingen.mpg.de/pages/home.php`
16) 黒木秀一, 西田　健 (2007)：競合連想ネットによる距離画像からの平面抽出, 日本神経回路学会誌, Vol. 70, No. 1, pp.273–281
17) Efron, B. (1983): Estimating the error rate of a prediction rule: improvement on cross-validation, Journal of the American Statistical Association, Vol. 78, No. 382, pp.316–331
18) Kohavi, R. (1995): A study of cross-validation and bootstrap for accuracy estimation and model selection, Proc. of the Fourteenth International Conference 18 on Artificial Intelligence (IJCAI), pp.1137–1143
19) Efron, B. and Tbshirani, R. (1997): Improvements on cross-validation: the .632+ bootstrap method, Journal of the American Statistical Association, Vol. 92, pp.548–560
20) Johnson, N.L., Kotz, S. and Kemp, A.W. (1992): Univariate Discrete Distributions, second edition, John Wiley, New York
21) Lendasse, A., Werz, V., Simon, G. and Verleysen, M. (2004): Fast bootstrap applied to LS-SVM for long term prediction of time series, Proc. of IJCNN2004, pp.705–710
22) Carney, J. and Cunningham, P. (2000): Tuning diversity in bagged ensembles, International Journal of Neural Systems, Vol. 10, No. 4, pp.267–279

付　　録

A. 予測誤差の期待値

以下，訓練データ数 $n_{v_i}^{\rm tr}$，バグ数 b，バグ外データ数 $b^{\rm ob}$ が十分大きいときの予測誤差の期待値を導く．まず訓練誤差 $L_{v_i}^{\rm tr} \triangleq \left\langle \|\hat{y}_s^j - y_s\|^2 \right\rangle_{s_{v_i}^{n\alpha^\sharp},j,j^{\rm bg}}$ は，

$$\begin{aligned} L_{v_i}^{\rm tr} &= \left\langle p_s^j \| \langle h_l^* \rangle_{l_{v_i}^{\rm tr}} + \left\langle h_l^* \Delta q_l^j \right\rangle_{l_{v_i}^{\rm tr}} - h_s^* \|^2 \right\rangle_{s_{v_i}^{\rm tr},j^{\rm bg}} \\ &= \left\langle p_s^j \right\| \left(\frac{1}{n_{v_i}^{\rm tr}} - 1 \right) h_s^* + \left(\langle h_l^* \rangle_{l_{v_i}^{\rm tr}}^2 - \frac{h_s^*}{n_{v_i}^{\rm tr}} \right) \\ &\quad + \left\langle h_l^* \Delta q_l^j \right\rangle_{l_{v_i}^{\rm tr}} \Big\|^2 \Big\rangle_{s_{v_i}^{\rm tr},j^{\rm bg}} \end{aligned} \tag{A.1}$$

となる．ここで $p_s^j = \Delta p_s^j + 1$ とおき，$\left\langle \Delta p_s^j \right\rangle_{s_{v_i}^{\rm tr}} \simeq 0$，および Δp_s^j と h_s には相関があり，Δp_s^j と h_l $(l \neq s)$ には相関がないとすると，

$$\begin{aligned} L_{v_i}^{\rm tr} \simeq & \left(1 - \frac{1}{n_{v_i}^{\rm tr}} \right)^2 \left\langle \Delta p_s^j (h_s^*)^2 \right\rangle_{s_{v_i}^{\rm tr},j^{\rm bg}} \\ & + \left(1 - \frac{1}{n_{v_i}^{\rm tr}} \right)^2 \left\langle (h_s^*)^2 \right\rangle_{s_{v_i}^{\rm tr}} \\ & + \Big(\frac{1}{n_{v_i}^{\rm tr}} \sum_{\substack{l \in I_{v_i}^{\rm tr} \\ l \neq s}} h_l^* \Big)^2 \\ & + \left\langle \left\langle h_l^* \Delta q_l^j \right\rangle_{l_{v_i}^{\rm tr}}^2 \right\rangle_{j^{\rm bg}} \end{aligned} \tag{A.2}$$

となる．ここで第 1 項は，$\left\langle \Delta p_s^j \right\rangle_{j^{\rm bg}} \simeq 0$ とすると

$$\left(1 - \frac{1}{n_{v_i}^{\rm tr}} \right)^2 \left\langle \left\langle \Delta p_s^j \right\rangle_{j^{\rm bg}} (h_s^*)^2 \right\rangle_{s_{v_i}^{\rm tr}} \simeq 0 \tag{A.3}$$

となる．また第 2 項と第 3 項の和は h_l^* が異なる $l(\in$

$I^{\mathrm{tr}}_{v_i}$) に対して相関が無いとすると，$(1-1/n^{\mathrm{tr}}_{v_i})(\rho^{\mathrm{tr}}_{h*})^2$ となる．ここで $(\rho^{\mathrm{tr}}_{h*})^2$ は訓練データの h^*_l の二乗平均 $(\rho^{\mathrm{tr}}_{h*})^2 \triangleq \left\langle (h^*_l)^2 \right\rangle_{l^{\mathrm{tr}}_{v_i}}$ である．さらに第4項は Cauchy-Schwarz の不等式より，

$$\begin{aligned}
&\left\langle \left\langle h^*_l \Delta q^j_l \right\rangle^2_{l^{\mathrm{tr}}_{v_i}} \right\rangle_{j^{\mathrm{bg}}} \\
&\leq \frac{1}{b} \sum_{j \in j^{\mathrm{bg}}} \left(\frac{1}{(n^{\mathrm{ib}}_{v_i})^2} \sum_{l \in I^{\mathrm{tr}}_{v_i}} (h^*_l)^2 \sum_{l \in I^{\mathrm{tr}}_{v_i}} (\Delta q^j_l)^2 \right) \\
&\simeq (\rho^{\mathrm{tr}}_{h*})^2 \sigma_q^2 \qquad \text{(A.4)}
\end{aligned}$$

となる．ここで σ_q^2 は q^j_l の分散であり式 (27) で与えられる．よって $L^{\mathrm{tr}}_{v_i}$ は

$$L^{\mathrm{tr}}_{v_i} \simeq (\rho^{\mathrm{tr}}_{h*})^2 - \frac{(\rho^{\mathrm{tr}}_{h*})^2}{n^{\mathrm{tr}}_{v_i}} + (\rho^{\mathrm{tr}}_{h*} \sigma_q \lambda^{\mathrm{bg}}_{h*q})^2 \qquad \text{(A.5)}$$

と表せる．ここで

$$(\lambda^{\mathrm{bg}}_{h*q})^2 \triangleq \frac{\left\langle \left\langle h^*_l \Delta q^j_l \right\rangle^2_{l^{\mathrm{tr}}_{v_i}} \right\rangle_{j^{\mathrm{bg}}}}{(\rho^{\mathrm{tr}}_{h*} \sigma_q)^2} \qquad \text{(A.6)}$$

であり，$0 \leq \lambda^{\mathrm{bg}}_{h*q} \leq 1$ を満たす．また $v_i \neq v^*_i$ のとき，h^*_l と Δq^j_l は無関係であり，$\lambda^{\mathrm{bg}}_{h*q}$ はほぼ一定値をとると考えられる．

次にバグ外誤差 $L^{\mathrm{ob}}_{v_i} \triangleq \left\langle \|\hat{y}^{\mathrm{ob}}_s - y_s\|^2 \right\rangle_{s^n_{v_i}}$ は

$$\begin{aligned}
L^{\mathrm{ob}}_{v_i} &= \left\langle \| \langle h^*_l \rangle_{l^{\mathrm{tr}}_{v_i}} + \left\langle h^*_l \Delta q^j_l \right\rangle_{l^{\mathrm{tr}}_{v_i}, j^{\mathrm{ob}}_s} - h^*_s \|^2 \right\rangle_{s^{\mathrm{tr}}_{v_i}} \\
&= \left\langle (h^*_s)^2 \right\rangle_{s^{\mathrm{tr}}_{v_i}} + \langle h^*_l \rangle^2_{l^{\mathrm{tr}}_{v_i}} + \left\langle h^*_l \Delta q^j_l \right\rangle^2_{l^{\mathrm{tr}}_{v_i}, j^{\mathrm{ob}}_s} \\
&\simeq (\rho^{\mathrm{tr}}_{h*})^2 + \frac{(\rho^{\mathrm{tr}}_{h*})^2}{n^{\mathrm{tr}}_{v_i}} + \frac{(\rho^{\mathrm{tr}}_{h*} \sigma_q \lambda^{\mathrm{ob}}_{h*q})^2}{b^{\mathrm{ob}}} \qquad \text{(A.7)}
\end{aligned}$$

となる．ここで

$$(\lambda^{\mathrm{ob}}_{h*q})^2 \triangleq \frac{\left\langle \left\langle h^*_l \Delta q^j_l \right\rangle^2_{l^{\mathrm{tr}}_{v_i}} \right\rangle_{j^{\mathrm{ob}}}}{(\rho^{\mathrm{tr}}_{h*} \sigma_q)^2} \qquad \text{(A.8)}$$

であり，$0 \leq \lambda^{\mathrm{ob}}_{h*q} \leq 1$ を満たす．同様に，汎化誤差 $L^{\mathrm{tst}}_{v_i} \triangleq \left\langle \|\hat{y}^{\mathrm{bg}}_s - y_s\|^2 \right\rangle_{s^{\mathrm{tst}}_{v_i}}$ は

$$\begin{aligned}
L^{\mathrm{tst}}_{v_i} &= \left\langle \| \langle h^*_l \rangle_{l^{\mathrm{tr}}_{v_i}} + \left\langle h^*_l \Delta q^j_l \right\rangle_{l^{\mathrm{tr}}_{v_i}, j^{\mathrm{bg}}} - h^*_s \|^2 \right\rangle_{s^{\mathrm{tst}}_{v_i}} \\
&= \left\langle (h^*_s)^2 \right\rangle_{s^{\mathrm{tst}}_{v_i}} + \langle h^*_l \rangle^2_{l^{\mathrm{tr}}_{v_i}} + \left\langle h^*_l \Delta q^j_l \right\rangle^2_{l^{\mathrm{tr}}_{v_i}, j^{\mathrm{bg}}} \\
&\simeq (\rho^{\mathrm{tst}}_{h*})^2 + \frac{(\rho^{\mathrm{tr}}_{h*})^2}{n^{\mathrm{tr}}_{v_i}} + \frac{(\rho^{\mathrm{tr}}_{h*} \sigma_q \lambda^{\mathrm{bg}}_{h*q})^2}{b} \qquad \text{(A.9)}
\end{aligned}$$

となる．ここで $(\rho^{\mathrm{tst}}_{h*})^2 \triangleq \left\langle (h^*_l)^2 \right\rangle_{l^{\mathrm{tst}}_{v_i}}$ である．なお，以上は平均化するデータ数が多いとして導出したものであり，$L^{\mathrm{tr}}_{v_i}$ と $L^{\mathrm{ob}}_{v_i}$ の第1項は同じ $(\rho^{\mathrm{tr}}_{h*})$ となり，3つの誤差の第2項の絶対値も同じ $(\rho^{\mathrm{tr}}_{h*})^2/n^{\mathrm{tr}}_{v_i}$ になるが，データ数が少ないときは異なる値になるので，より一般的に異なる変数を用いて式 (31)〜(33) のように表した．