# 線形変数変換に不変な自乗値ペナルティ項の構成法

斉藤和巳[†] 中野良平[††]

オリジナルスケールと変数変換を施した双方のデータで，2つのニューラルネットの学習を行うとき，その学習が不変とは，適用した変数変換に応じてニューラルネット重みを変換すれば，つねに同値なニューラルネットが学習結果として得られることを意味する．ペナルティ付き学習の場合には，ニューラルネット重みだけでなく，ペナルティ変数群の学習に対する不変性も考慮しなければならない．本稿では，これらの問題設定に基づき，線形変数変換に不変な自乗値ペナルティ項の構成法に関する理論的枠組みを提案し，その正当性を証明した．指数が整数に限定されない多項式型法則をニューラルネット学習で発見する問題への適用実験では，データが不要な説明変数を含み，かつ，ある程度のノイズを含む場合でも，提案したペナルティ項を用いれば，線形変数変換に不変となることを確認した．

# Squared Penalty Consistent with Linear Transformations of Variables

Kazumi Saito[†] and Ryohei Nakano[††]

If we train one neural network using original data and another network using data whose variables are transformed from the original, then consistency requires that we should obtain equivalent networks whose sets of weights can be changed to each other through the corresponding transformations. When networks are trained so as to minimize a penalized objective function, we must consider such a set of penalty factors that guarantees the consistency. For this purpose, we propose a theoretical framework for constructing squared penalty terms consistent with linear transformations of variables, and prove its plausibility. In our experiments concerning a numeric law discovery problem formulated as learning in neural networks, it was confirmed that such a penalty term can be consistent with a linear transformation of variables.

## 1. はじめに

変数値の範囲やスケールが計測単位などに依存して大きく異なる実データ☆では，ニューラルネットなどを用いた学習（非線形回帰）に先立ち，データに対して変数変換（正規化）する前処理がなされることが多い．実際，適切な変数変換を施さなければ，望ましい学習（非線形最適化）結果が得られないケースもある[2]．ゆえに，そうした変数変換が学習結果にどのような影響を与えるかは，非線形回帰問題[11]での重要な研究課題である．その際に考慮すべき最も重要な性質は変数変換に対するニューラルネット学習の不変性である．ここで不変とは，適用した変数変換に応じてニューラルネット重みを変換すれば，つねに同値なニューラルネットが学習結果として得られることを意味する[1]．その不変性が保証されないと，変数変換の前後で実質的に異なるニューラルネットが得られることになる．

一方，汎化性能をいかに向上させるかは，ニューラルネット学習におけるもう1つの重要な研究課題である．ここで汎化とは，未学習データに対する性能を意味する．学習の目的関数として，訓練データに関する通常の誤差項だけでなく，重みの値が大きくなることを抑制するペナルティ（正則化）項を付加することにより，多くの場合，汎化能力の高いニューラルネットを得られることが知られている[1),6]．我々の実験[9]では，自乗値ペナルティ項と2次学習アルゴリズムBPQ[8]の組合せは，他の組合せと比較して，収束性能が大幅に向上するとともに，高い汎化性能を示した．よって，本稿ではこの組合せを対象として，変数変換に対するニューラルネット学習の不変性について考える．

ペナルティ項を考慮した不変性に関する，基本的な性質はすでにいくつか知られている[1),4),10]．まず，すべてのニューラルネット重みに単一係数でペナルティを施す場合には，一般の線形変数変換に対して不変でなくなるが，すべての入力変数に対して同一の線形変

† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
†† 名古屋工業大学知能情報システム学科
Nagoya Institute of Technology

☆ ここでは，多数の変数群とその具体値を与えるサンプル群を列と行とする2次元配列の多変量データを想定し，単にデータと呼ぶ．

換を施す場合には，ニューラルネット学習を不変とするペナルティ項は容易に構成できることが示されている[1]．一方，実数指数を許容する一般化多項式（ニューラルネット）を用いた法則発見問題[5),7)]では，各変数ごとに異なるスケーリングを施しても，一部の重みを除いて構成するペナルティ項を用いれば，単一ペナルティ係数でもその学習は不変となることが示されている[4]．しかるに，一般の線形変換に対して，ニューラルネット学習を不変とするペナルティ項の構成法，および，ペナルティ係数群も学習する枠組みでの不変性などについては，あまり多くの考察がなされていない．

本稿では，線形変数変換に不変な自乗値ペナルティ項を構成する枠組みとともに，ニューラルネットを用いた法則発見問題への応用について述べる．2章では，ニューラルネット学習の基本的な枠組みを説明した後，最適ニューラルネット重みとペナルティ項に関する性質について考察する．3章では，学習不変性について基本的な定義を与え，線形変数変換に不変な自乗値ペナルティ項の構成法を提案するとともに，ニューラルネット学習の不変性を検証する．4章では，法則発見問題への適用実験により，ニューラルネット学習の不変性を検証する．

## 2. 基本的枠組み

### 2.1 問題設定

各サンプルの変数群を $(x_1,\cdots,x_K,y)$，あるいは $(\boldsymbol{x},y)$ とする．ただし，$x_k$ は説明変数（explanatory variable），$y$ は基準変数（target variable）を表す．本稿では，ニューラルネット $f(\boldsymbol{x};\boldsymbol{\Theta})$ のクラスとして，以下のような，中間ユニットの活性化関数を $g(u)$ とする3層パーセプトロンを考える．

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = w_0 + \sum_{j=1}^{J} w_j g\left(w_{j0} + \sum_{k=1}^{K} w_{jk}x_k\right) = w_0 + \sum_{j=1}^{J} w_j g\left(w_{j0} + \boldsymbol{w}_j'\boldsymbol{x}\right). \quad (1)$$

ただし，各ニューラルネット重み（パラメータ）$w_j$ や $w_{jk}$ は未知の実数，$J$ は中間ユニット数を表し，$\boldsymbol{\Theta}$ は $M$ 次元縦ベクトルで，重み $w_j, j=0,\cdots,J$ と $w_{jk}, j=1,\cdots,J,\ k=0,\cdots,K$ を並べて構成される（$M=1+J(K+2)$）．また，$\boldsymbol{w}_j'$ は $\boldsymbol{w}_j$ の転置を表す．なお，活性化関数 $g(u)$ は非線形な単調関数とする．

訓練サンプル集合を $D=\{(\boldsymbol{x}^\mu,y^\mu),\mu=1,\cdots,N\}$ とする．ただし，$N$ はサンプル数を表し，各サンプル $(\boldsymbol{x}^\mu,y^\mu)$ は独立に同一分布から得られるとする．また，訓練サンプル集合 $D$ と独立なテストサンプルを $T=(\boldsymbol{x}^\nu,y^\nu)$ とする．いま，重み $\boldsymbol{\Theta}$ に対する各サンプルの自乗誤差を以下で定義する．

$$\mathcal{E}^\mu(\boldsymbol{\Theta}) = \left(y^\mu - f(\boldsymbol{x}^\mu;\boldsymbol{\Theta})\right)^2. \quad (2)$$

本稿では，ニューラルネット学習の最終的な目標として，サンプル集合 $D$ を固定し，以下で定義する期待値（汎化誤差）を最小化する真の最適重み $\boldsymbol{\Theta}^*$ を求める問題について考える．

$$\mathcal{G}(\boldsymbol{\Theta}) = E_T[\mathcal{E}^\nu(\boldsymbol{\Theta})]. \quad (3)$$

### 2.2 一般的な正則化の枠組み

$\widehat{\boldsymbol{\Theta}}$ で表記される $\boldsymbol{\Theta}^*$ の最小自乗推定（least-squares estimate）は，訓練サンプル集合に対する以下の自乗誤差を最小化して得られる．

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{\mu=1}^{N}\mathcal{E}^\mu(\boldsymbol{\Theta}). \quad (4)$$

以下では，式 (4) を目的関数とするニューラルネット学習をペナルティなし重み学習と呼ぶ．ところが，この目的関数を単純に採用すれば，学習結果は訓練サンプルのノイズにオーバフィットしてしまう傾向があり，式 (3) で定義した汎化性能の観点では，一般に良い解を得ることができない[1]．

すでに述べたように，式 (4) に重みの値が大きくなることを抑制するペナルティ項を付加することにより，多くの場合，汎化能力の高いニューラルネットを得られることが知られている．本稿では，以下のような二次形式のペナルティ項を考える．

$$\Omega(\boldsymbol{\Theta};\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\Theta}'\boldsymbol{\Lambda}(\boldsymbol{\lambda})\boldsymbol{\Theta}. \quad (5)$$

ただし，ペナルティ行列 $\boldsymbol{\Lambda}$ は $M\times M$ 非負定値行列で，ペナルティ変数☆（パラメータ）$\boldsymbol{\lambda}$ は $M_0$ 次元ベクトルとし，行列 $\boldsymbol{\Lambda}$ はペナルティ変数 $\boldsymbol{\lambda}$ の連続微分可能な関数群として規定されるとする．このとき，学習結果の重みは以下の目的関数を最小化して得られる．

$$\mathcal{F}(\boldsymbol{\Theta};\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\Theta}) + \Omega(\boldsymbol{\Theta};\boldsymbol{\lambda}). \quad (6)$$

以下では，式 (6) を目的関数とするニューラルネット学習をペナルティ付き重み学習と呼ぶ．なお，広く採用される weight decay は $\boldsymbol{\Lambda}=\lambda\boldsymbol{I}$ のケースであり，$\boldsymbol{I}$ は $M\times M$ 単位行列で，ペナルティ変数（係数）$\lambda$ はスカラーとなる（$M_0=1$）．一方，重みごとにペナルティを施すケースでは，たとえば $\boldsymbol{\Lambda}=\boldsymbol{diag}(\boldsymbol{\lambda})$ のようになり，$\boldsymbol{diag}(\boldsymbol{\lambda})$ は対角行列，その各対角要素は $M$ 次元ペナルティ変数（係数）$\boldsymbol{\lambda}$ の各要素で規定さ

☆ 重みの自乗に対する係数の場合だけでなく，$\boldsymbol{\Lambda}$ を規定する任意の関数の引数として $\boldsymbol{\lambda}$ を扱うので，本稿ではペナルティ変数と呼んでいる．

れる（$M_0 = M$）．本稿では，これらのケースを含む一般形としてペナルティ行列を考える．

ペナルティ付き学習では，一般に，重みだけでなくペナルティ変数もデータから学習しなければならない．本稿では，ペナルティ変数の学習に交差検証法（cross-validation）[12] から導かれる目的関数を採用する[10]．交差検証法では，与えられたデータ $D$ をランダムに $S$ 個のセグメントに分割する（$G_s : s = 1, \cdots, S$）．そして，$S-1$ 個のセグメントを学習に使用し，残りの 1 個を（汎化の）テストに使用する．この手順を $S$ 回繰り返し，最終的には以下の自乗誤差 $\mathcal{CV}$ を求める．

$$\mathcal{CV} = \frac{1}{2}\sum_{s=1}^{S}\sum_{\mu\in G_s}\mathcal{E}^{\mu}(\widehat{\boldsymbol{\Theta}}_s). \tag{7}$$

ただし，$\widehat{\boldsymbol{\Theta}}_s$ は次の目的関数を最小化する重みである．

$$\mathcal{F}_s(\boldsymbol{\Theta}_s;\boldsymbol{\lambda}) = \frac{1}{2}\sum_{\mu\notin G_s}\mathcal{E}^{\mu}(\boldsymbol{\Theta}_s) + \Omega(\boldsymbol{\Theta}_s;\boldsymbol{\lambda}). \tag{8}$$

なお，セグメント数とサンプル数が等しい場合（$S = N$）は，特に，1 つ抜き法（leave-one-out）と呼ばれ，少数サンプルの問題に対して頻繁に適用される．

陰関数定理（implicit function theorem）[3] より，式 (8) を最小化する重み $\widehat{\boldsymbol{\Theta}}_s$ は，$\boldsymbol{\lambda}$ の陰関数を要素とするベクトルとして定義できる．つまり，式 (7) はペナルティ変数に対する連続な目的関数として定義できるので，式 (7) を最小化する $\widehat{\boldsymbol{\lambda}}$ を効率良く計算できる[10]．よって，式 (6) に $\widehat{\boldsymbol{\lambda}}$ を代入すれば，学習結果の重みは，式 (6) を最小化する $\widehat{\boldsymbol{\Theta}}$ として得られる．以下では，式 (7) を目的関数とするニューラルネット学習をペナルティ変数学習と呼ぶ．

### 2.3 ペナルティ変数の次元

最適なニューラルネット重み $\boldsymbol{\Theta}^*$ は，式 (3) で定義した汎化誤差を最小化するので，極値に関する以下の必要条件を満たす．

$$\frac{\partial\mathcal{G}(\boldsymbol{\Theta}^*)}{\partial\boldsymbol{\Theta}} = E_T\left[\frac{\partial\mathcal{E}^{\nu}(\boldsymbol{\Theta}^*)}{\partial\boldsymbol{\Theta}}\right] = \mathbf{0}. \tag{9}$$

一方，ペナルティ付き学習の目的関数 $\mathcal{F}$ を採用すれば，学習結果の重み $\widehat{\boldsymbol{\Theta}}$ は，式 (6) を最小化するので，以下の必要条件を満たす．

$$\frac{\partial\mathcal{F}(\widehat{\boldsymbol{\Theta}};\boldsymbol{\lambda})}{\partial\boldsymbol{\Theta}} = \frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}} + \boldsymbol{\Lambda}(\boldsymbol{\lambda})\widehat{\boldsymbol{\Theta}} = \mathbf{0}. \tag{10}$$

すでに述べたように，$\widehat{\boldsymbol{\Theta}}$ は $\boldsymbol{\lambda}$ の陰関数を要素とするベクトルとして定義できるので，$\boldsymbol{\lambda}$ を適切に設定すれば，最適重み $\boldsymbol{\Theta}^*$ に近い学習結果の重み $\widehat{\boldsymbol{\Theta}}$ を得ることが期待できる．そこで，$\widehat{\boldsymbol{\Theta}}$ を $\boldsymbol{\lambda}$ の関数と見なし，式 (3) に代入すれば，汎化誤差も $\boldsymbol{\lambda}$ の関数と見なせるので，式 (3) を最小化するための $\boldsymbol{\lambda}$ に対する必要条件が以下のように得られる．

$$\begin{aligned}\frac{\partial\mathcal{G}(\widehat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\lambda}} &= \left(\frac{\partial\widehat{\boldsymbol{\Theta}}}{\partial\boldsymbol{\lambda}'}\right)'\frac{\partial\mathcal{G}(\widehat{\boldsymbol{\Theta}})}{\partial\widehat{\boldsymbol{\Theta}}}\\ &= \left(\frac{\partial\widehat{\boldsymbol{\Theta}}}{\partial\boldsymbol{\lambda}'}\right)' E_T\left[\frac{\partial\mathcal{E}^{\nu}(\widehat{\boldsymbol{\Theta}})}{\partial\widehat{\boldsymbol{\Theta}}}\right] = \mathbf{0}.\end{aligned} \tag{11}$$

ここで，最適重み $\boldsymbol{\Theta}^*$ と学習結果の重み $\widehat{\boldsymbol{\Theta}}$ との関係について考える．ペナルティ変数 $\boldsymbol{\lambda}$ の次元は $M_0$ なので，$\partial\widehat{\boldsymbol{\Theta}}/\partial\boldsymbol{\lambda}'$ は $M\times M_0$ 行列となる（この行列の計算は付録 A.1 参照）．ペナルティ変数と重みの次元が等しい場合（$M_0 = M$），行列 $\partial\widehat{\boldsymbol{\Theta}}/\partial\boldsymbol{\lambda}'$ が $M\times M$ 正則（nonsingular）となれば，式 (9) と (11) より，$\boldsymbol{\Theta}^*$ と $\widehat{\boldsymbol{\Theta}}$ が満たすべき必要条件は一致することが分かる．一方，ペナルティ変数の次元が重みの次元よりも小さい場合（$M_0 < M$），明らかに，$\partial\mathcal{G}(\widehat{\boldsymbol{\Theta}})/\partial\widehat{\boldsymbol{\Theta}} \neq \mathbf{0}$ となる重みでも，式 (11) の必要条件を満たすことになる．よって，ペナルティ変数と重みの次元を等しく設定すれば，学習結果の重み $\widehat{\boldsymbol{\Theta}}$ を最適重み $\boldsymbol{\Theta}^*$ に，より近づけることが期待できる．なお，現実の学習問題において，式 (3) で定義した汎化誤差は計算できず，目的関数として採用できないが，その十分妥当な近似として，式 (7) で定義した交差検証誤差を目的関数として採用することができる．

### 2.4 提案枠組みの位置づけ

本稿で提案した枠組みは，ペナルティ行列に対して非負定値性のみ仮定するので，特定の重み集合群にはペナルティを施さないケースも包含する．たとえば，3 層パーセプトロンにおいて，各ユニットのバイアスにはペナルティを施さず，重みについては各層間ごとに異なるペナルティ係数を施す提案[1] もこの枠組みに入る．

一方，ペナルティ変数学習では，交差検証誤差を目的関数として最小化することのみ用いている．また，法則発見などにおいて不要な重みを除去するために提案された MCV（minimum cross-validation）正則化法[10] では，各重みごとに個別のペナルティ係数を施すので対角ペナルティ行列を用いている．よって，本稿で提案する正則化法は MCV 正則化法をも包含する一般的なものある．

## 3. 線形変数変換に対する不変性

### 3.1 変数変換不変性の定義と条件

以下の線形変数変換について考える．

$$\widetilde{x}_k = a_k x_k + b_k, \widetilde{y} = cy + d. \tag{12}$$

ただし，$a_k \neq 0$，$c \neq 0$ とする．また，式 (12)

で定義したデータに対する変数変換を $\varphi$ で表記し，$(\widetilde{\boldsymbol{x}}, \widetilde{y}) = \varphi(\boldsymbol{x}, y)$ の意味で用いる．データの変数変換に対してニューラルネット学習が不変とは，その変換に応じて重みを変換すれば，つねに同値なニューラルネットが学習結果として得られることと定義される[1)]．本稿では，中間ユニット数が等しい 2 つのニューラルネットの重みを $\boldsymbol{\Theta}_1$ と $\boldsymbol{\Theta}_2$ とするとき，中間ユニットの並び順を変更するだけで $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$ となるとき同値と定義する．なお，式 (1) で定義した各層メッシュ結合のニューラルネットでは，その構造を規定するパラメータは中間ユニット数だけであり，その数が決まれば，重みの次元も一意に定まる．逆に，重みの次元が等しければ中間ユニット数，すなわち構造も一致する．以下では，与えられたデータの変数変換 $\varphi$ に対して，ニューラルネット学習が不変となる条件を導く．

まず，ペナルティなし重み学習について考える．変数変換後の重みを $\widetilde{\boldsymbol{\Theta}}$ とすれば，式 (4) で定義した変換前の目的関数 $\mathcal{L}$ に対して，変換後の目的関数は以下のように導ける．

$$\widetilde{\mathcal{E}}^{\mu}(\widetilde{\boldsymbol{\Theta}}) = (\widetilde{y}^{\mu} - f(\widetilde{\boldsymbol{x}}^{\mu}; \widetilde{\boldsymbol{\Theta}}))^2, \tag{13}$$

$$\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_{\mu=1}^{N} \widetilde{\mathcal{E}}^{\mu}(\widetilde{\boldsymbol{\Theta}}). \tag{14}$$

よって，ペナルティなし重み学習における不変性の条件は以下のように定義できる．

$$(\exists\psi)(\forall\boldsymbol{\Theta})(\exists\widetilde{\boldsymbol{\Theta}}) \left( \frac{\partial \mathcal{L}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \boldsymbol{0} \ \Rightarrow \left( \frac{\partial \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}})}{\partial \widetilde{\boldsymbol{\Theta}}} = \boldsymbol{0} \wedge \boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}}; \varphi) \right) \right). \tag{15}$$

ただし，重み変換関数 $\psi$ は，変数変換 $\varphi$ のみであらかじめ定まるものとする．$\forall$ は全称限量子（universal quantifier），$\exists$ は存在限量子（existential quantifier）であり，対象とする目的関数 $\mathcal{L}$ や $\widetilde{\mathcal{L}}$ には一般に複数の局所解が存在するため上記の表現となる．目的関数の一階微分がゼロとなる極値条件は，変数変換前後の重みが学習結果として得られることを意味する．

次に，ペナルティ付き重み学習について考える．変数変換後のペナルティ項を $\widetilde{\Omega}$ とすれば，式 (6) で定義した変換前の目的関数 $\mathcal{F}$ に対して，変換後の目的関数は以下のように導ける．

$$\widetilde{\mathcal{F}}(\widetilde{\boldsymbol{\Theta}}; \boldsymbol{\lambda}) = \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}}) + \widetilde{\Omega}(\widetilde{\boldsymbol{\Theta}}; \boldsymbol{\lambda}). \tag{16}$$

このとき，ペナルティ項付き重み学習における不変性は以下のように定義できる．

$$(\exists\psi)(\exists\widetilde{\Omega})(\forall\boldsymbol{\lambda})(\forall\boldsymbol{\Theta})(\exists\widetilde{\boldsymbol{\Theta}}) \left( \frac{\partial \mathcal{F}(\boldsymbol{\Theta}; \boldsymbol{\lambda})}{\partial \boldsymbol{\Theta}} = \boldsymbol{0} \ \Rightarrow \left( \frac{\partial \widetilde{\mathcal{F}}(\widetilde{\boldsymbol{\Theta}}; \boldsymbol{\lambda})}{\partial \widetilde{\boldsymbol{\Theta}}} = \boldsymbol{0} \wedge \boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}}; \varphi) \right) \right). \tag{17}$$

ただし，変換後のペナルティ項 $\widetilde{\Omega}$ も，データの変数変換 $\varphi$ のみであらかじめ定まるものとする☆．

最後に，ペナルティ変数学習について考える．式 (7) で定義した変換前の目的関数 $\mathcal{CV}$ に対して，変換後の目的関数は以下のように導ける．

$$\widetilde{\mathcal{CV}}(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{s=1}^{S} \sum_{\mu \in G_s} \widetilde{\mathcal{E}}^{\mu}(\widehat{\widetilde{\boldsymbol{\Theta}}}_s). \tag{18}$$

ただし，$\widehat{\widetilde{\boldsymbol{\Theta}}}_s$ は次の目的関数を最小化する重みである．

$$\widetilde{\mathcal{F}}_s(\widetilde{\boldsymbol{\Theta}}_s; \boldsymbol{\lambda}) = \frac{1}{2}\sum_{\mu \notin G_s} \widetilde{\mathcal{E}}^{\mu}(\widetilde{\boldsymbol{\Theta}}_s) + \widetilde{\Omega}(\widetilde{\boldsymbol{\Theta}}_s; \boldsymbol{\lambda}). \tag{19}$$

このとき，ペナルティ変数学習における不変性は以下のように定義できる．

$$(\exists\psi)(\exists\widetilde{\Omega})(\forall\boldsymbol{\lambda})(\forall\boldsymbol{\Theta})(\exists\widetilde{\boldsymbol{\Theta}}) \left( \left( \frac{\partial \mathcal{CV}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{0} \wedge \frac{\partial \mathcal{F}(\boldsymbol{\Theta}; \boldsymbol{\lambda})}{\partial \boldsymbol{\Theta}} = \boldsymbol{0} \right) \Rightarrow \left( \frac{\partial \widetilde{\mathcal{CV}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{0} \wedge \frac{\partial \widetilde{\mathcal{F}}(\widetilde{\boldsymbol{\Theta}}; \boldsymbol{\lambda})}{\partial \widetilde{\boldsymbol{\Theta}}} = \boldsymbol{0} \wedge \boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}}; \varphi) \right) \right). \tag{20}$$

### 3.2 重み変換関数

データに対する変数変換として式 (12) を考えるとき，重み変換関数 $\psi$ は以下のように導ける．まず，変数変換後の重みとして

$$\begin{cases} \widetilde{w}_0 &= cw_0 + d, \\ \widetilde{w}_j &= cw_j, \\ \widetilde{w}_{j0} &= w_{j0} - \sum_{k=1}^{K} a_k^{-1} b_k w_{jk}, \\ \widetilde{w}_{jk} &= a_k^{-1} w_{jk}. \end{cases} \tag{21}$$

を用いれば，変数変換前後でペアとなる中間ユニットの出力値は以下のように一致する．

$$\begin{aligned} & g(\widetilde{w}_{j0} + \widetilde{\boldsymbol{w}}_j' \widetilde{\boldsymbol{x}}) \\ &= g\left( w_{j0} - \sum_{k=1}^{K} \frac{b_k w_{jk}}{a_k} + \sum_{k=1}^{K} \frac{w_{jk}(a_k x_k + b_k)}{a_k} \right) \\ &= g(w_{j0} + \boldsymbol{w}_j' \boldsymbol{x}). \end{aligned} \tag{22}$$

また，出力ユニットでは以下が導ける．

☆ すでに述べたように，重みの自乗に対する係数の場合だけでなく，$\boldsymbol{\Lambda}$ を規定する任意の関数の引数として $\boldsymbol{\lambda}$ を扱うため，変数変換後のペナルティ変数 $\widetilde{\boldsymbol{\lambda}}$ ではなく，より一般に変換後のペナルティ項 $\widetilde{\Omega}$ を導入している．

$$f(\widetilde{\boldsymbol{x}};\widetilde{\boldsymbol{\Theta}}) = cw_0 + d + \sum_{j=1}^{J} cw_j g(w_{j0} + \boldsymbol{w}_j'\boldsymbol{x}) = cf(\boldsymbol{x};\boldsymbol{\Theta}) + d. \quad (23)$$

すなわち，式 (21) で定義した重みを用いれば，変数変換前後のニューラルネット出力は，基準変数 $y$ に施した変数変換と同じ関係になることが分かる．

一方，$a_k \neq 0$，$c \neq 0$ より式 (21) を逆に解けば，$\widetilde{\boldsymbol{\Theta}}$ から $\boldsymbol{\Theta}$ への重み変換関数が以下のように得られる．

$$\begin{cases} w_0 &= c^{-1}\widetilde{w}_0 - c^{-1}d, \\ w_j &= c^{-1}\widetilde{w}_j, \\ w_{j0} &= \widetilde{w}_{j0} + \sum_{k=1}^{K} b_k \widetilde{w}_{jk}, \\ w_{jk} &= a_k \widetilde{w}_{jk}. \end{cases} \quad (24)$$

よって，$M \times M$ 正則行列 $\boldsymbol{T}$ を用いれば，重み変換関数 $\psi$ は以下のように表現できる．

$$\boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}};\varphi) = \boldsymbol{T}\widetilde{\boldsymbol{\Theta}} + \boldsymbol{t}. \quad (25)$$

ただし，$\boldsymbol{t} = (-c^{-1}d, 0, \cdots, 0)'$ である．

### 3.3 不変性の検証

まず，ペナルティなし重み学習の不変性を検証する．式 (4) と (14) で定義した目的関数 $\mathcal{L}$ と $\widetilde{\mathcal{L}}$ に，式 (25) の重み変換関数 $\psi$ で対応する重みのペア $\boldsymbol{\Theta}$ と $\widetilde{\boldsymbol{\Theta}}$ をそれぞれ代入すれば，式 (23) より以下の関係が導ける．

$$\begin{aligned} \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}}) &= \frac{1}{2}\sum_{\mu=1}^{N} \widetilde{\mathcal{E}}^{\mu}(\widetilde{\boldsymbol{\Theta}}) \\ &= \frac{1}{2}\sum_{\mu=1}^{N} (cy^{\mu} + d - (cf(\boldsymbol{x}^{\mu};\boldsymbol{\Theta}) + d))^2 \\ &= c^2 \frac{1}{2}\sum_{\mu=1}^{N} \mathcal{E}^{\mu}(\boldsymbol{\Theta}) = c^2\mathcal{L}(\boldsymbol{\Theta}). \end{aligned} \quad (26)$$

つまり，単純な比例関係となる．また，これら目的関数の一階微分では，式 (26) より以下の関係が導ける．

$$\begin{aligned} \frac{\partial\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}})}{\partial\widetilde{\boldsymbol{\Theta}}} &= \frac{c^2\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\widetilde{\boldsymbol{\Theta}}} \\ &= c^2\left(\frac{\partial\boldsymbol{\Theta}}{\partial\widetilde{\boldsymbol{\Theta}}'}\right)' \frac{\mathcal{L}(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}} = c^2\boldsymbol{T}'\frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}. \end{aligned} \quad (27)$$

よって，$c \neq 0$ かつ $\boldsymbol{T}$ は正則行列より，$\boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}};\varphi)$ のとき，目的関数 $\mathcal{L}(\boldsymbol{\Theta})$ と $\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}})$ の極値条件が一致することが分かる．すなわち，目的関数 $\mathcal{L}$ から得られる任意の重み $\widehat{\boldsymbol{\Theta}}$ に対して，目的関数 $\widetilde{\mathcal{L}}$ から得られる重み $\widehat{\widetilde{\boldsymbol{\Theta}}}$ が存在し，式 (25) で定義した重み変換関数 $\psi$ で，これら重みが対応することを示している．ゆえに，式 (15) で定義した不変性の条件を満たしていることが分かる．

次に，ペナルティ付き重み学習の不変性を検証する．いま，変数変換後のペナルティ項を以下で構成する．

$$\widetilde{\Omega}(\widetilde{\boldsymbol{\Theta}};\boldsymbol{\lambda}) = \frac{c^2}{2}(\boldsymbol{T}\widetilde{\boldsymbol{\Theta}} + \boldsymbol{t})'\boldsymbol{\Lambda}(\boldsymbol{\lambda})(\boldsymbol{T}\widetilde{\boldsymbol{\Theta}} + \boldsymbol{t}). \quad (28)$$

このとき，式 (6) と (16) で定義した目的関数 $\mathcal{F}$ と $\widetilde{\mathcal{F}}$ に，重み変換関数 $\psi$ で対応する重みのペア $\boldsymbol{\Theta}$ と $\widetilde{\boldsymbol{\Theta}}$ をそれぞれ代入すれば，式 (26) と (28) より以下の関係が導ける．

$$\begin{aligned} &\widetilde{\mathcal{F}}(\widetilde{\boldsymbol{\Theta}};\boldsymbol{\lambda}) = \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Theta}}) + \widetilde{\Omega}(\widetilde{\boldsymbol{\Theta}};\boldsymbol{\lambda}) \\ &= c^2\mathcal{L}(\boldsymbol{\Theta}) + c^2\Omega(\boldsymbol{\Theta};\boldsymbol{\lambda}) = c^2\mathcal{F}(\boldsymbol{\Theta};\boldsymbol{\lambda}). \end{aligned} \quad (29)$$

つまり，この場合も単純な比例関係となる．よって，式 (27) と同様に一階微分を計算することにより，任意の $\boldsymbol{\lambda}$ に対して，$\boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}};\varphi)$ のとき，目的関数 $\mathcal{F}(\boldsymbol{\Theta};\boldsymbol{\lambda})$ と $\widetilde{\mathcal{F}}(\widetilde{\boldsymbol{\Theta}};\boldsymbol{\lambda})$ の極値条件は一致することが容易に分かる．ゆえに，ペナルティなし重み学習のときと同様にして，式 (17) で定義した不変性の条件を満たしていることが検証できる．

最後に，ペナルティ変数学習の不変性を検証する．ペナルティ付き重み学習で示したように，式 (8) と (19) を目的関数として，重み変換関数 $\psi$ で対応する重みのペアの集合 $\{(\widehat{\boldsymbol{\Theta}}_s, \widehat{\widetilde{\boldsymbol{\Theta}}}_s)\}$ が得られる．ゆえに，式 (7) と (18) で定義した目的関数 $\mathcal{CV}$ と $\widetilde{\mathcal{CV}}$ に，これら重みペアの集合 $\{(\widehat{\boldsymbol{\Theta}}_s, \widehat{\widetilde{\boldsymbol{\Theta}}}_s)\}$ をそれぞれ代入すれば，式 (26) より以下の関係が容易に導ける．

$$\begin{aligned} \widetilde{\mathcal{CV}}(\boldsymbol{\lambda}) &= \frac{1}{2}\sum_{s=1}^{S}\sum_{\mu\in G_s} \widetilde{\mathcal{E}}^{\mu}(\widehat{\widetilde{\boldsymbol{\Theta}}}_s) \\ &= \frac{1}{2}\sum_{s=1}^{S}\sum_{\mu\in G_s} (cy^{\mu} + d - (cf(\boldsymbol{x}^{\mu};\widehat{\boldsymbol{\Theta}}_s) + d))^2 \\ &= c^2\frac{1}{2}\sum_{s=1}^{S}\sum_{\mu\in G_s} \mathcal{E}^{\mu}(\widehat{\boldsymbol{\Theta}}_s) = c^2\mathcal{CV}(\boldsymbol{\lambda}). \end{aligned} \quad (30)$$

つまり，この場合も単純な比例関係となる．しかるに，これら目的関数は重み $\{\widehat{\boldsymbol{\Theta}}_s\}$ と $\{\widehat{\widetilde{\boldsymbol{\Theta}}}_s\}$ を介在して定義されるので，式 (27) と同様には，それらの一階微分を計算できない．実際，それら目的関数の一階微分では，正則行列 $\boldsymbol{T}$ が現れず，以下の関係となる（付録 A.2 参照）．

$$\frac{\partial\widetilde{\mathcal{CV}}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} = c^2\frac{\partial\mathcal{CV}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}. \quad (31)$$

ゆえに，$\boldsymbol{\Theta} = \psi(\widetilde{\boldsymbol{\Theta}};\varphi)$ のとき，目的関数 $\mathcal{CV}(\boldsymbol{\lambda})$ と $\widetilde{\mathcal{CV}}(\boldsymbol{\lambda})$ の極値条件は一致するので，これまでと同じ議論により，式 (20) で定義した不変性の条件を満たしていることは容易に検証できる．

## 4. ペナルティ付き学習の法則発見への応用

### 4.1 法則発見ネットでの不変性

数法則 $f(\boldsymbol{x};\boldsymbol{\Theta})$ のクラスとして，以下のような，指数が整数に限定されない一般化多項式を考える．

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = w_0 + \sum_{j=1}^{J} w_j \prod_{k=1}^{K} x_k^{w_{jk}} = w_0 + \sum_{j=1}^{J} w_j \exp\left(\boldsymbol{w}_j' \ln \boldsymbol{x}\right). \quad (32)$$

式 (1) で定義したニューラルネットと比較すれば，中間ユニットの活性化関数が指数関数で，入力ベクトルに対数変換が施され，そして，中間ユニットのバイアス項 $w_{j0}$ が現れないことが特徴となる．よって，これまでの議論を単純に適用できない．そこで，以下のように重み $w_j$ の変換を考える．

$$w_{j0} = \ln|w_j|, \quad j = 1, \cdots, J. \quad (33)$$

このとき，式 (32) は以下のように変形できる．

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = w_0 + \sum_{j=1}^{J} \varsigma(w_j) \exp(w_{j0} + \boldsymbol{w}_j' \ln \boldsymbol{x}). \quad (34)$$

ただし，$\varsigma(w_j)$ は $w_j$ の符号を返す関数を表す．なお，$\varsigma(w_j)$ をあらかじめ知ることはできないが，ニューラルネット学習では，中間ユニットの順は任意であるので，正負の符号個数のみ考えれば，全組合せが学習可能となる．つまり，$J+1$ 通りの組合せを調べれば十分である．明らかに，式 (6) で定義した目的関数をペナルティ付き重み学習に用い，式 (7) で定義した目的関数をペナルティ変数学習に用いれば，これまでの議論を直接適用することができる．

ここでは，以下の説明変数に対する変換のみを考える．

$$\ln \widetilde{x}_k = a_k \ln x_k + b_k. \quad (35)$$

なぜなら，基準変数 $y$ に定数 $d$ を加える変換は，式 (26) で示したように相殺され，$c$ 倍する変換は，目的関数や一階微分を単純に定数倍するからである．よって，本実験では，基準変数に対する変換を省略する．なお，変数変換後の学習結果の重みは，変数変換前で対応する重みへ，以下のようにして変換できる．

$$\begin{cases} w_0 &= \widetilde{w}_0, \\ w_j &= \varsigma(\widetilde{w}_j) \exp\left(\widetilde{w}_{j0} + \sum_{k=1}^{K} b_k \widetilde{w}_{jk}\right), \\ w_{jk} &= a_k \widetilde{w}_{jk}. \end{cases} \quad (36)$$

2.3 節の議論より，$\boldsymbol{\lambda}$ を $M$ 次元ベクトルとして $\boldsymbol{\Lambda}(\boldsymbol{\lambda})$ を具体的に考える．一般に，$\mathcal{F}(\widehat{\boldsymbol{\Theta}};\boldsymbol{\lambda})$ の $\boldsymbol{\lambda}$ に対する 2 階微分（Hesse）行列は正定値となることが期待できるので，行列 $\partial\widehat{\boldsymbol{\Theta}}/\partial\boldsymbol{\lambda}'$ の正則性が期待できる最も単純なペナルティ行列として以下を採用する．

$$\boldsymbol{\Lambda}(\boldsymbol{\lambda}) = \begin{pmatrix} \exp(\lambda_1) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \exp(\lambda_M) \end{pmatrix}. \quad (37)$$

なお，各ペナルティ変数に指数関数を施すのは，ペナルティ行列の非負定値性を保証するためである．

### 4.2 人工データ

以下の人工法則（関数）の発見問題[5),7)] を考える．

$$y = 2 + 3x_1^{-1}x_2^3 + 4x_3x_4^{1/2}x_5^{-1/3}. \quad (38)$$

ただし，説明変数には法則に現れる $x_1, \cdots, x_5$ のほかに，この法則とは無関係な変数 $x_6, \cdots, x_9$ を加える☆．各説明変数の値は領域 $(0, 1)$ の中でランダムに生成し，基準変数 $y$ の値は上記法則に従って計算した後，平均 0，標準偏差 0.1 の正規ノイズを加えて求める．この手順を繰り返し，サンプル数 $N = 200$ のデータを生成した．また，各説明変数 $(x_1, \cdots, x_9)$ には以下の変数変換を施した☆☆．

$$\ln \widetilde{x}_k = \frac{1}{\mathrm{std}(\ln x_k)} \ln x_k - \frac{\mathrm{mean}(\ln x_k)}{\mathrm{std}(\ln x_k)}. \quad (39)$$

ただし，mean() は平均，std() は標準偏差を訓練サンプル集合に対して計算する関数である．

実験では，ニューラルネット重みやペナルティ変数の初期値を次のように設定した．入力–中間ユニット間の重み $w_{jk}$ については，平均 0，標準偏差 1 の正規分布に基づいて独立に生成した．出力ユニットのバイアス $w_0$ は 0 に設定した．ペナルティ変数 $\boldsymbol{\lambda}$ の初期値は 0 に設定した．つまり，行列 $\boldsymbol{\Lambda}$ の初期値は単位行列となる．一方，重みやペナルティ変数学習の終了条件については，勾配ベクトルが十分に小さいときとした．すなわち，重み $\boldsymbol{\Theta}$ の学習では，

$$\max_m \left\{ \left| \frac{\partial \mathcal{F}(\boldsymbol{\Theta});\boldsymbol{\lambda}}{\partial \theta_m} \right| \right\} < 10^{-6} \quad (40)$$

ペナルティ変数 $\boldsymbol{\lambda}$ の学習では，

☆ 冗長な変数を加える理由は，一般には，学習前に必要な変数数が未知だからである．

☆☆ 変数変換導入の動機は，学習が困難となるスケールの現実データに対処することであるが，本実験は変数変換不変性の検証を目的としているので，オリジナルも変数変換後も学習可能なスケールに制限している．

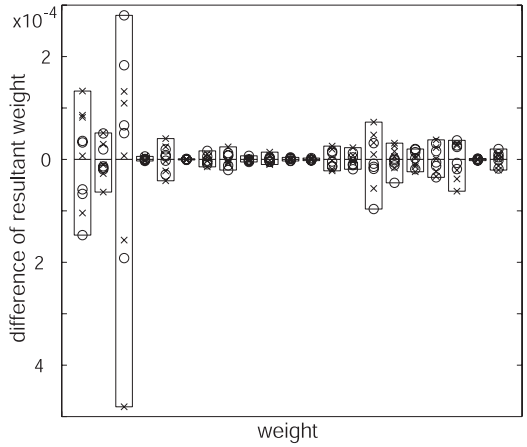


**図 1** ニューラルネット重みの試行ごとの差
Fig. 1 Difference of resultant weights.

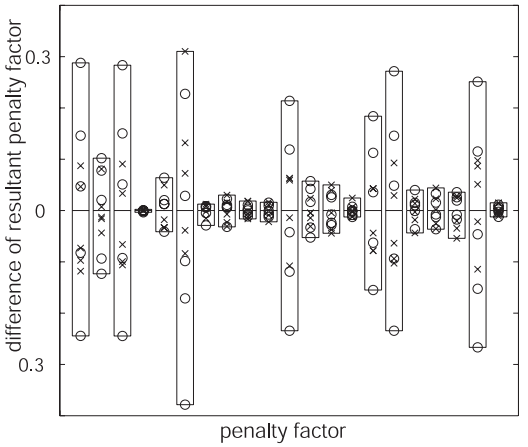


**図 2** ペナルティ変数の試行ごとの差
Fig. 2 Difference of resultant penalty factor.

$$\max_m \left\{ \left| \frac{\partial \mathcal{CV}(\boldsymbol{\lambda})}{\partial \lambda_m} \right| \right\} < 10^{-6} \tag{41}$$

のときとした．中間ユニット数は真の個数 $J = 2$，$\varsigma(w_1)$ と $\varsigma(w_2)$ はともに真の符号（正）とし，交差検証誤差の計算は 1 つ抜き法（$S = N$）を用いた．

### 4.3 実験結果

同一の訓練サンプル集合に対して，ニューラルネット重みの初期値のみを変えて，変数変換前後でそれぞれ 5 回（合計 10 回）の試行を行った．**図 1** に，試行ごとの結果の各重み（$w_0, w_1, w_2, w_{11}, \cdots, w_{19}, w_{21}, \cdots, w_{29}$）を，全試行の平均からの差で示す．ただし，変数変換後の重みは，変数変換前で対応する重みへ，式 (36) を用いて変換して示した．また，変数変換前の試行結果は “○”，変換後のものは “×” で示した．図より，全試行結果の差は大変小さく，特に，変数変換にともなう差は，試行ごとの差の範囲内であることが分かる．

全試行の平均値を結果重みとして採用したときの発見例を以下に示す．

$$\begin{aligned} y = & 1.9270 \\ & + 3.0287 x_1^{-0.9954} x_2^{+2.9839} x_3^{+0.0000} x_4^{+0.0028} \\ & \times x_5^{-0.0048} x_6^{-0.0009} x_7^{-0.0009} x_8^{+0.0000} x_9^{+0.0015} \\ & + 4.0356 x_1^{-0.0006} x_2^{-0.0028} x_3^{+0.9777} x_4^{+0.4935} \\ & \times x_5^{-0.3251} x_6^{-0.0013} x_7^{-0.0019} x_8^{-0.0000} x_9^{-0.0066} \end{aligned}$$

ただし，図 1 の縦軸のスケールに合わせて，各重みの値は小数点第 4 位までで四捨五入した．この発見例とともに，全試行結果は真の法則にかなり近いことが分かる．

**図 2** には，図 1 に対応するペナルティ変数学習の結果を示す．図より，ニューラルネット重みのときよりも差は大きくなるものの，変数変換にともなう差は，試行ごとの差の範囲内であることが分かる．したがって，重みとペナルティ変数双方の学習において，データが不要な説明変数を含み，かつ，基準変数値がある程度のノイズを含む場合でも，変数変換に対して学習が不変となることをこの例で数値的にも確認できた．なお，ペナルティ変数学習の結果のばらつきに対して，重み学習の結果は比較的頑健であると考えられる．また，図 2 からは，変数変換前の結果の分散が比較的大きいことが分かる．よって，適切な変数変換を施せば，安定した学習を行えることが示唆されるものの，その詳しい性質などについては，今後の研究課題の 1 つとなる．

## 5. おわりに

本稿では，正則化項付きニューラルネット学習において，データ変数の線形変換に対する不変性について論じた．すなわち，ニューラルネット重みだけでなく，交差検証誤差を目的関数としてペナルティ変数を学習する枠組みにおいて，学習の不変性を保証する自乗値ペナルティ項の構成法を提案した．実験では，指数が整数に限定されない実数指数の多項式型法則の発見問題において，データが不要な説明変数を含み，かつ，基準変数値がある程度のノイズを含む場合でも，変数変換前後で学習結果が不変となることを検証した．今後は，さらに幅広い問題への適用実験を行い，提案法の有効性を検証する予定である．

## 参考文献

1) Bishop, C.M.: *Neural networks for pattern recognition*, Clarendon Press (1995).
2) Gill, P.E., Murray W. and Wright, M.H.:

*Practical optimization*, Academic Press (1981).
3) Luenberger, D.G.: *Linear and nonlinear programming*, Addison-Wesley (1984).
4) Nakano, R. and Saito, K.: Discovery of a set of nominally conditioned polynomials, *Proc. 2nd Int. Conference on Discovery Science, LNAI 1721*, pp.287–298 (1999).
5) 中野良平，斉藤和巳：多変量データからの多項式型法則の発見，発見科学とデータマイニング，森下真一，宮野　悟（編），bit 別冊号，pp.217–227 (2000).
6) Ripley, B.D.: *Pattern recognition and neural networks*, Cambridge University Press (1996).
7) 斉藤和巳，中野良平：コネクショニストアプローチによる数法則の発見，情報処理学会論文誌，Vol.37, No.9, pp.1708–1716 (1996).
8) Saito, K. and Nakano, R.: Partial BFGS update and efficient step-length calculation for three-layer neural networks, *Neural Computation*, Vol.9, No.1, pp.239–257 (1997).
9) Saito, K. and Nakano, R.: Second-order learning algorithm with squared penalty term, *Neural Computation*, Vol.12, No.3, pp.709–729 (2000).
10) 斉藤和巳，中野良平：交差検証誤差最小化による重要重みの発見，電子情報通信学会論文誌，Vol.J84-D-II, No.1, pp.178–187 (2001).
11) Seber, G.A.F. and Wild, C.J.: *Nonlinear regression*, John Wiley & Sons (1989).
12) Stone, M.: Cross-validatory choice and assessment of statistical predictions (with discussion), *Journal of the Royal Statistical Society B*, Vol.64, pp.111–147 (1974).

## 付　　録

### A.1　重みのペナルティ変数に対する微分

$\widehat{\boldsymbol{\Theta}}$ を $\boldsymbol{\lambda}$ の関数として，式 (10) の右辺を $\boldsymbol{\lambda}$ について微分すれば以下を得る．

$$\left(\frac{\partial^2\mathcal{F}(\widehat{\boldsymbol{\Theta}};\boldsymbol{\lambda})}{\partial\widehat{\boldsymbol{\Theta}}\partial\widehat{\boldsymbol{\Theta}}'}\right)\frac{\partial\widehat{\boldsymbol{\Theta}}}{\partial\boldsymbol{\lambda}'}+\frac{\partial\boldsymbol{\Lambda}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}'}\widehat{\boldsymbol{\Theta}}=\mathbf{0}. \quad (42)$$

よって，$\widehat{\boldsymbol{\Theta}}$ の $\boldsymbol{\lambda}$ に対する微分は，式 (42) を変形することにより得られる．

$$\frac{\partial\widehat{\boldsymbol{\Theta}}}{\partial\boldsymbol{\lambda}'}=-\left(\frac{\partial^2\mathcal{F}(\widehat{\boldsymbol{\Theta}};\boldsymbol{\lambda})}{\partial\widehat{\boldsymbol{\Theta}}\partial\widehat{\boldsymbol{\Theta}}'}\right)^{-1}\frac{\partial\boldsymbol{\Lambda}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}'}\widehat{\boldsymbol{\Theta}}. \quad (43)$$

ただし，$(\partial\boldsymbol{\Lambda}(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}')\widehat{\boldsymbol{\Theta}}$ の各 $(i,j)$ 要素は以下である．

$$\left(\frac{\partial\boldsymbol{\Lambda}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}'}\widehat{\boldsymbol{\Theta}}\right)_{i,j}=\sum_{k=1}^{M}\frac{\partial\Lambda_{ik}}{\partial\lambda_j}\theta_k. \quad (44)$$

### A.2　ペナルティ変数数学習での不変性

式 (18) で定義した目的関数 $\widetilde{\mathcal{CV}}$ を $\boldsymbol{\lambda}$ で微分すれば以下を得る．

$$\frac{\partial\widetilde{\mathcal{CV}}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}=\sum_{s=1}^{S}\left(\frac{\partial\widehat{\widetilde{\boldsymbol{\Theta}}}_s}{\partial\boldsymbol{\lambda}'}\right)'\sum_{\nu\in G_s}\frac{\partial\widetilde{\mathcal{E}}^{\nu}(\widehat{\widetilde{\boldsymbol{\Theta}}}_s)}{\partial\widehat{\widetilde{\boldsymbol{\Theta}}}_s}. \quad (45)$$

一方，式 (25) を考慮すれば，$\widehat{\widetilde{\boldsymbol{\Theta}}}_s$ の $\boldsymbol{\lambda}$ に対する微分は以下となる．

$$\frac{\partial\widehat{\widetilde{\boldsymbol{\Theta}}}_s}{\partial\boldsymbol{\lambda}'}=\frac{\partial\widehat{\widetilde{\boldsymbol{\Theta}}}_s}{\partial\widehat{\boldsymbol{\Theta}}_s'}\frac{\partial\widehat{\boldsymbol{\Theta}}_s}{\partial\boldsymbol{\lambda}'}=\boldsymbol{T}^{-1}\frac{\partial\widehat{\boldsymbol{\Theta}}_s}{\partial\boldsymbol{\lambda}'}. \quad (46)$$

また，変数変換前後の重みに対して以下が成り立つ．

$$\frac{\partial\widetilde{\mathcal{E}}^{\nu}(\widetilde{\boldsymbol{\Theta}}_s)}{\partial\widetilde{\boldsymbol{\Theta}}_s}=c^2\boldsymbol{T}'\frac{\partial\mathcal{E}^{\nu}(\boldsymbol{\Theta}_s)}{\partial\boldsymbol{\Theta}_s}. \quad (47)$$

よって，式 (46) と (47) のそれぞれ関係を，式 (45) に代入すれば以下が導ける．

$$\frac{\partial\widetilde{\mathcal{CV}}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}=c^2\frac{\partial\mathcal{CV}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}. \quad (48)$$

(平成 13 年 1 月 31 日受付)
(平成 15 年 9 月 5 日採録)

**斉藤　和巳**（正会員）

昭和 38 年生．昭和 60 年慶應義塾大学理工学部数理科学科卒業．工学博士．同年 NTT 入社．平成 3 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習の研究に従事．現在，NTT コミュニケーション科学基礎研究所主任研究員（特別研究員）．情報処理学会論文賞受賞（平成 9 年）．人工知能学会論文賞受賞（平成 11 年）．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE 各会員．

**中野　良平**（正会員）

昭和 22 年生．昭和 46 年東京大学工学部計数工学科卒業．工学博士．同年，日本電信電話公社（現 NTT）入社．以来，統計解析，分散処理，データベース，人工知能，神経回路網の研究に従事．平成 11 年より，名古屋工業大学知能情報システム学科教授．電気通信普及財団賞（テレコムシステム技術賞）受賞（平成 9 年）．情報処理学会論文賞受賞（平成 9 年）．人工知能学会論文賞受賞（平成 11 年）．電子情報通信学会論文賞受賞（平成 12 年）．電子情報通信学会，人工知能学会，日本神経回路学会各会員．