3A303-13

適応型荷重摂動学習の汎化能力に関する統計力学的解析

◎ 三好 亮介† 前田 裕‡ 三好 誠司¶

† 関西大院 理工学研究科 ‡¶関西大 システム理工

†k329499@kansai-u.ac.jp, ¶miyoshi@kansai-u.ac.jp

1 はじめに

学習とは観測データを用いてその背後にあるデータの生成過程を推定することである．教師つき学習においては観測データは教師の入出力であり，これは例題とも呼ばれる．学習はバッチ学習とオンライン学習 [1] に大別できる．バッチ学習においては与えられたいくつかの例題を繰り返し使用する．この場合，生徒が適切な自由度を持っていればすべての例題に正しく答えられるようになるが，それまでに長い時間が必要である．また，多くの例題を蓄えておくメモリが必要である．これに対してオンライン学習では一度使った例題は捨ててしまう．この場合，過去に使った例題に対して生徒が必ず正しく答えられるとは限らないが，多くの例題を蓄えておくためのメモリが不要であり，また時間的に変化する教師にも追随できるなどの利点がある [2].

ところで, ある関数を最大あるいは最小とする可調整パラメータを求めるという問題は, 最適化の問題として広く取り扱われている [3]. このとき, 逐次解法としての一般的なアプローチは勾配法である. 制御をはじめとする多くの分野では勾配法を基本とした手法を用いることが多く, この手法は適切なパラメータ修正則を与えてくれる. しかしながら, 一方で関数の微分として勾配を用いることができない場合には, この手法を用いることができない. この場合, その関数の値は求めることができるという条件で, 摂動を用いて微分値を差分近似するという方法が考えられる. しかし, 可調整パラメータの数が多い場合, この単純な差分近似による勾配の計算手法は, 関数の値を求める回数が増加し, 適用が難しい場合が多い. これに対し, すべての可調整パラメータに同時に摂動を加える同時摂動最適化が, Spall[4], Alespector[5], Cauwenberg[6] および前田 [7] らによって, それぞれ独立に提案されている. 一方, 機械学習の分野において, 学習機械のパラメータである結合荷重の調整に摂動を利用する方法が提案されており, 荷重摂動学習と呼ばれている [8]. 同時摂動最適化を学習に適用する場合を考えるとこれは荷重摂動学習と等価である [8].

オンライン学習を統計力学的に解析する場合, 自己平均性を仮定するためにノルムの小さい入力が毎回独立に生成されると考える [2], [9] が, この入力を摂動として用いるならば荷重摂動学習の解析にオンライン学習の解析手法をそのまま適用することができる [10]. この解析手法により, 荷重摂動学習はパーセプトロン学習と同じ漸近特性を有することが明らかにされている [10].

本稿では, 荷重摂動学習の漸近特性を改良した適応型荷重摂動学習を提案する. またその汎化能力について統計力学的手法を用いて解析する.

2 モデル

本稿では教師と生徒がいずれも単純パーセプトロンであるようなモデルを扱う [2]. 教師と生徒の結合荷重をそれぞれ $\boldsymbol{B}$, $\boldsymbol{J}$ とし, またそれぞれに同じ入力 $\boldsymbol{x}$ が入力されるとする. 教師 $\boldsymbol{B} = (B_1, \cdots, B_N)$, 生徒 $\boldsymbol{J} = (J_1, \cdots, J_N)$, 入力 $\boldsymbol{x} = (x_1, \cdots, x_N)$ は N 次元ベクトルであり, 教師 $\boldsymbol{B}$ の各要素 B_i は平均 0, 分散 1 のガウス分布に従い独立に生成され, 不変であるとする. 生徒 $\boldsymbol{J}$ の初期値 J^0 の各要素 J_i^0 は平均 0, 分散 1 のガウス分布に従い独立に生成されるものとし, $\boldsymbol{B}$ と $\boldsymbol{J}$ の方向余弦は R であるとする. また, 入力 $\boldsymbol{x}$ の各要素 x_i は平均 0, 分散 $1/N$ のガウス分布に従い独立に生成されるものとする. ここで, $v^m = \boldsymbol{B} \cdot \boldsymbol{x}^m$, $u^m l^m = \boldsymbol{J}^m \cdot \boldsymbol{x}^m$ とすると, v と u は平均 0, 分散 1, 共分散 R の二次元ガウス分布に従う確率変数となる [2]. ここで, 教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}$ の誤差を $\epsilon^m \equiv \Theta(-u^m v^m)$ で定義しておく. $\Theta(\cdot)$ はステップ関数である. すなわち,

$$\Theta(\mathrm{x}) = \begin{cases} +1 & \text{if } \mathrm{x} \geq 0 \\ 0 & \text{if } \mathrm{x} < 0 \end{cases} \tag{1}$$

である.

生徒 $\boldsymbol{J}$ は入力 $\boldsymbol{x}$ とそれに対する教師 $\boldsymbol{B}$ の出力を使い,

更新を行う．すなわち

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + f^m \boldsymbol{x}^m \tag{2}$$

である．f は更新量を表す関数であり学習則により決定される．よく知られている学習則であるヘブ学習, パーセプトロン学習, アダトロン学習の f はそれぞれ

$$f^m = \eta \mathrm{sgn}(v^m) \tag{3}$$
$$f^m = \eta \Theta(-u^m v^m)\mathrm{sgn}(v^m) \tag{4}$$
$$f^m = \eta |u^m| \Theta(-u^m v^m)\mathrm{sgn}(v^m) \tag{5}$$

である [2]．ここで η は学習係数である．

3　オンライン学習の統計力学的解析

統計的学習理論の目的のひとつは汎化誤差を理論的に求めることである．汎化誤差 ϵ_g は入力 $\boldsymbol{x}$ に関する誤差 ϵ の平均であり, 本稿で考えているモデルの場合

$$\epsilon_g = \frac{1}{\pi}\cos^{-1} R \tag{6}$$

である [2]．式 (6) が示すように汎化誤差 ϵ_g は巨視的変数 R の関数である．R のダイナミクスを記述する連立微分方程式は熱力学的極限における自己平均性に基づき決定論的な形で以下のように導くことができる [2]．

$$\frac{dl}{dt} = \langle fu \rangle + \frac{\langle f^2 \rangle}{2l}, \qquad \frac{dr}{dt} = \langle fv \rangle \tag{7}$$

ここで

$$r = Rl \tag{8}$$

である．式 (7) には 3 つのサンプル平均 $\langle fu \rangle$, $\langle f^2 \rangle$, $\langle fv \rangle$ が含まれている．ヘブ学習, パーセプトロン学習, アダトロン学習の場合, これらは解析的に以下のように求めることができる [2]．

ヘブ学習

$$\langle fu \rangle = \frac{2\eta R}{\sqrt{2\pi}}, \quad \langle fv \rangle = \eta\sqrt{\frac{2}{\pi}}, \quad \langle f^2 \rangle = \eta^2 \tag{9}$$

パーセプトロン学習

$$\langle fu \rangle = -\langle fv \rangle = \eta \frac{R-1}{\sqrt{2\pi}}, \quad \langle f^2 \rangle = \frac{\eta^2}{\pi}\cos^{-1} R \tag{10}$$

アダトロン学習

$$\langle fu \rangle = \frac{\eta}{\pi}(R\sqrt{1-R^2} - \cos^{-1}\mathrm{R}) \tag{11}$$
$$\langle fv \rangle = \frac{\eta}{\pi}(1-R^2)^{3/2} + R\langle fu \rangle \tag{12}$$
$$\langle f^2 \rangle = -\eta \langle fu \rangle \tag{13}$$

なおヘブ学習の場合, 式 (9) を式 (7) に代入した具体的な連立微分方程式は以下のように解析的に解くことができる [2], [11]．

$$r = \eta\sqrt{\frac{2}{\pi}}t, \quad l^2 = \frac{2\eta^2}{\pi}t^2 + \eta^2 t + 1 \tag{14}$$

4　荷重摂動学習

学習機械のパラメータである結合荷重の調整に摂動を利用する方法は荷重摂動学習と呼ばれている [8]．前節で述べたようにオンライン学習を統計力学的に解析する場合, ノルムの小さい入力が毎回独立に生成されると考えるが, この入力を摂動として用いるならば荷重摂動学習にオンライン学習の解析手法をそのまま適用することができる [10]．荷重摂動学習における更新を図 1 に示す．

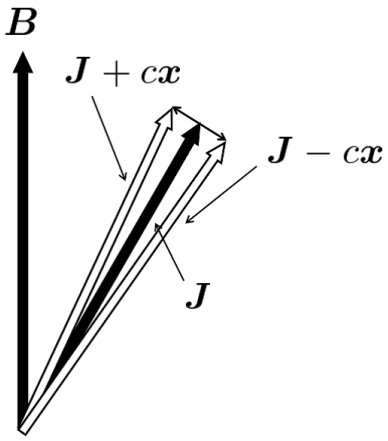

図 1　荷重摂動学習における更新

荷重摂動学習の更新量は

$$f^m = -\frac{\eta}{2c} g^m \tag{15}$$
$$g^m = \Theta(-v^m(\boldsymbol{J}^m + c\boldsymbol{x}^m)\cdot \boldsymbol{x}^m) - \Theta(-v^m(\boldsymbol{J}^m - c\boldsymbol{x}^m)\cdot \boldsymbol{x}^m) \tag{16}$$

である [10]．ここで c は正の定数である．またサンプル平均は

$$\langle fu \rangle = -\frac{\eta}{c}\int_{-\frac{c}{l}}^{\frac{c}{l}} Duu H\left(\frac{Ru}{\sqrt{1-R^2}}\right) \tag{17}$$
$$\langle fv \rangle = \frac{\eta}{c}\int_0^\infty Dv v\left(H\left(\frac{-\frac{c}{l}+Rv}{\sqrt{1-R^2}}\right) - H\left(\frac{\frac{c}{l}+Rv}{\sqrt{1-R^2}}\right)\right) \tag{18}$$
$$\langle f^2 \rangle = \frac{\eta^2}{2c^2}\int_{-\frac{c}{l}}^{\frac{c}{l}} Du H\left(-\frac{Ru}{\sqrt{1-R^2}}\right) \tag{19}$$

となる [10]．ここで $H(u) \equiv \int_u^\infty Dx$, $Dx \equiv \frac{dx}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ である．

式 (6)-(14), (17)-(19) を用いて理論的に計算される生徒の長さ l, 汎化誤差 ϵ_g のダイナミクスを計算機実験の結果と重ねて図 2, 図 3 に示す．これらの図において曲線は理論計算の結果を表し, ○, ×, △, □のシンボルは計算機実験の結果を表す．また WP は荷重摂動のことを指す．理論計算はヘブ学習については解析解である式 (6), (8), (14) をプロットした．また, パーセプトロン学習, アダトロン学習に関しては連立微分方程式である式 (7), (10)-(13) は解析的に解けないためルンゲ・クッタ法を用いて数値的に解いた．荷重摂動学習はサンプル平均 式 (17)-(19) の積分が解析的に実行できないのでシンプソン則とルンゲ・クッタ法を併用して数値的に解いた [10]．なおシンプソン則の積分範囲は $-5 \sim +5$, 積分刻みは 0.01, ルンゲ・クッタ法の時間刻みは 0.01 とした．一方, 計算機実験は次元 $N = 10^3$ で実行し, ϵ_g は各時点で 10^5 個のランダム入力の中で, 教師と生徒の出力が異なる入力の個数をカウントすることにより算出している．なお, $\eta = 1$, $c = 1$ とした． 図 2, 図 3 より理論と計算機実験はよく一

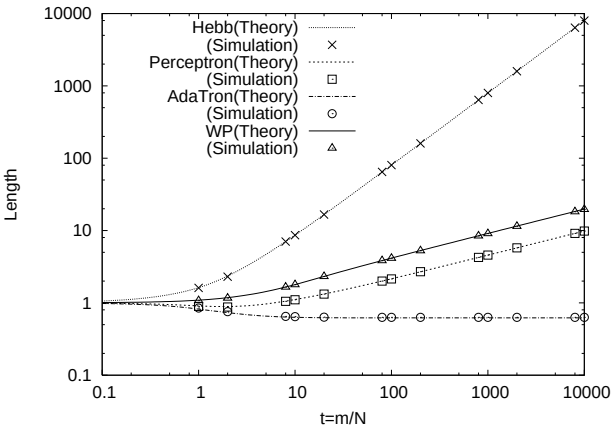

図 2　生徒の長さ l のダイナミクス

致しており, 本節で述べた理論解析の結果が正しいことがわかる．ここで計算機実験は有限の N で実行しているので自己平均性は厳密には破れており, 理論と若干のずれが見られる．

図 3 より, パーセプトロン学習, 荷重摂動学習の漸近特性が $\epsilon_g \sim O(t^{-\frac{1}{3}})$ であるのに対し, アダトロン学習の漸近特性は $\epsilon_g \sim O(t^{-1})$ であり非常に優れていることがわかる．アダトロン学習はパーセプトロン学習の適応型である．そこで次節においてはパーセプトロン学習とアダトロン学習の関係を考察することにより適応型の荷重摂動学習を提案し, その汎化能力を統計力学的な手法を用いて解析する．図 2 よりアダトロン学習だけが生徒の長さ l が

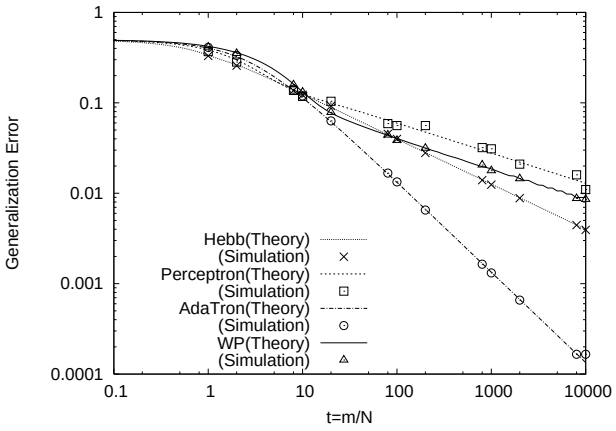

図 3　汎化誤差 ϵ_g のダイナミクス

0.5 に収束しつつあり, 短くなっていることがわかる．

5　適応型荷重摂動学習の提案

5.1　パーセプトロン学習とアダトロン学習

$\eta = 1$ の場合, パーセプトロン学習, アダトロン学習の更新式はそれぞれ以下のようになる．

$$f = \quad \Theta(-uv)\mathrm{sgn}(v) \tag{20}$$

$$f = |u|\Theta(-uv)\mathrm{sgn}(v) \tag{21}$$

またパーセプトロン学習とアダトロン学習の更新の様子をそれぞれ図 4, 図 5 に示す．

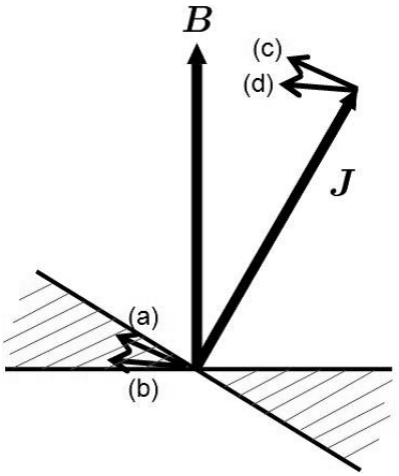

図 4　パーセプトロン学習における更新

式 (20), (21) よりパーセプトロン学習もアダトロン学習も更新が行われるのは教師と生徒の出力が異なる場合、すなわち図 4, 図 5 の斜線部に入力が生成された場合である．パーセプトロン学習の場合, 入力 $\boldsymbol{x}$ が図 4 の (a), (b) のとき更新ベクトルはそれぞれ (c), (d) である．いずれ

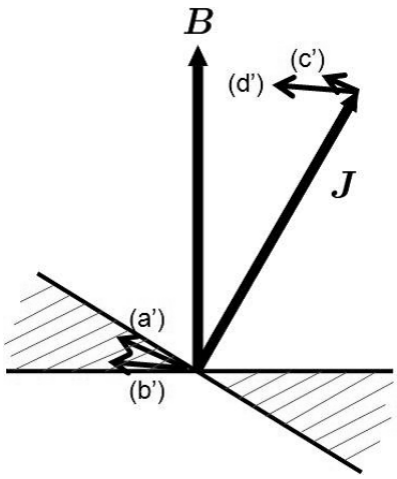

図 5　アダトロン学習における更新

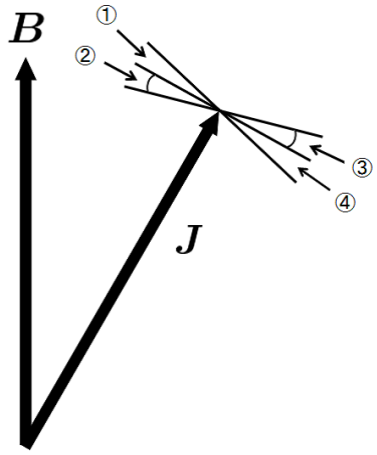

図 6　適応型荷重摂動学習の提案

の場合も更新ベクトルのノルムは変わらない．それに対してアダトロン学習の場合, 入力 $\boldsymbol{x}$ が図 5 の (a'), (b') のとき更新ベクトルはそれぞれ (c'), (d') である．このとき更新により生徒 $\boldsymbol{J}$ の長さがどうなるかについて考察する．パーセプトロン学習の場合, 生徒が長くなる方向にも短くなる方向にも更新ベクトルが一定の大きさで更新される．結果的には図 2 のように長さ l は発散する．しかしながらアダトロン学習の場合は, 生徒ベクトルの長さが長くなる方向には更新ベクトルが小さく, 短くなる方向には更新ベクトルが大きいため図 2 のように長さ l は 0.5 に漸近する [9]．パーセプトロン学習とアダトロン学習のこの違いが汎化誤差の漸近特性の違いの理由である．

5.2　適応型荷重摂動学習の提案

荷重摂動学習の更新量は式 (15),(16) であり, その更新の様子は図 1 である．荷重摂動学習の更新においては, 入力 $\boldsymbol{x}$ が生徒とほぼ直交する角度で生成されないと更新が行われないため, まず入力が生徒と直交する場合を考える．実際には次元 N が有限であれば, 生徒 $\boldsymbol{J}$ と入力 $\boldsymbol{x}$ が直交からややずれても更新は行われるため, 実際に更新が行われる入力 $\boldsymbol{x}$ の範囲は図 6 の ① から ④ である．

ここで, アダトロン学習のように生徒の長さが短くなる方向に大きく更新することを考える．そのためには入力 $\boldsymbol{x}$ が ② と ③ に生成された場合に $|u|$ に比例する更新を行えばよい．すなわち, 更新量 f を

$$f^m = -\frac{\eta}{2c}\,|u^m|\,g^m\,\Theta(u^m g^m) \tag{22}$$

$$g^m = \Theta(-v^m(\boldsymbol{J}^m + c\boldsymbol{x}^m)\cdot\boldsymbol{x}^m) - \Theta(-v^m(\boldsymbol{J}^m - c\boldsymbol{x}^m)\cdot\boldsymbol{x}^m) \tag{23}$$

とすればよい．

5.3　結果と考察

アダトロン学習, 荷重摂動学習, 適応型荷重摂動学習について理論的に計算される生徒の長さ l, 汎化誤差 ϵ_g のダイナミクスを計算機実験の結果と重ねて図 7, 図 8 に示す．適応型荷重摂動学習については式 (6)-(8), (22), (23) を用いた．これらの図において曲線は理論計算の結果を表し, ○, △, ■のシンボルは計算機実験の結果を表す．また AWP は適応型荷重摂動のことを指す．理論計算について, 適応型荷重摂動学習はサンプル平均

$$\langle fu\rangle = \int dudvP(u,v)f(v,u,l)u \tag{26}$$

$$\langle fv\rangle = \int dudvP(u,v)f(v,u,l)v \tag{27}$$

$$\langle f^2\rangle = \int dudvP(u,v)f(v,u,l)^2 \tag{28}$$

の積分が解析的に実行できないのでシンプソン則とルンゲ・クッタ法を併用して数値的に解いた [10]．なおシンプソン則の積分範囲は $-3 \sim +3$, 積分刻みは u, v がそれぞれ 0.01, 0.0025, ルンゲ・クッタ法の時間刻みは $t = 0 \sim 100$ の区間, $100 \sim 1000$ の区間についてそれぞれ 0.01, 0.1 で行った．一方, 計算機実験は次元 $N = 10^4$ で実行し, ϵ_g は各時点で 10^5 個のランダム入力の中で, 教師と生徒の出力が異なる入力の個数をカウントすることにより算出している．なお, $\eta = 1$, $c = 1$ とした．

図 7, 図 8 を見ると理論と計算機実験はよく一致しており, 本節で述べた理論解析の結果が正しいことがわかる．図 7 より, 適応型荷重摂動学習の生徒の長さは, アダトロン学習と同様に 0.5 に漸近することがわかる．これは 5.1

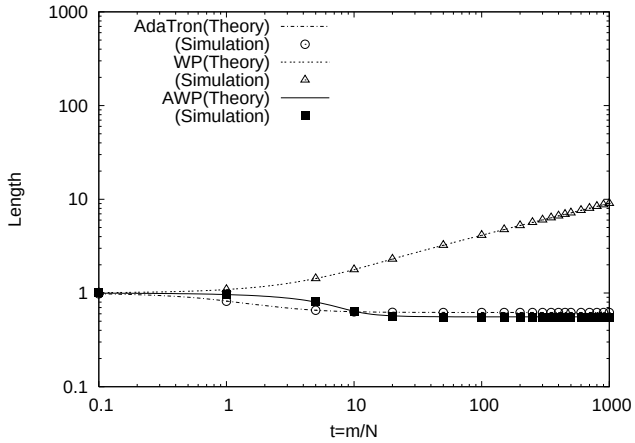

図 7　生徒の長さ l のダイナミクス

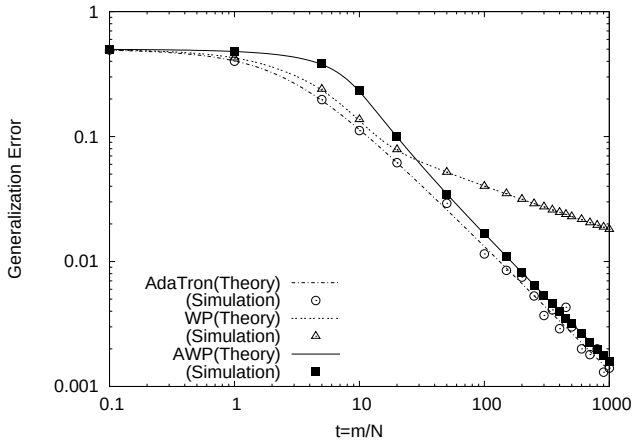

図 8　汎化誤差 ϵ_g のダイナミクス

節で述べた, 生徒ベクトルの長さが短くなる方向だけに更新ベクトルが大きく進んだことによるものである. また, 図 8 より, 荷重摂動学習の漸近特性が $\epsilon_g \sim O(t^{-\frac{1}{3}})$ であるのに対し, 適応型荷重摂動学習の漸近特性は $\epsilon_g \sim O(t^{-1})$ であり, アダトロン学習と一致していることがわかる.

6　むすび

本稿ではまず, 統計力学的な手法を用いてオンライン学習の解析を行う方法 [2] について述べた. その際に汎化誤差を求めるために, 生徒ベクトルの長さ l, 教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}$ の方向余弦 R を導入し, これらのダイナミクスを記述する連立微分方程式を解いた.

次に, 荷重摂動学習の解析 [10] について述べた. パーセプトロン学習, 荷重摂動学習の漸近特性が $\epsilon_g \sim O(t^{-\frac{1}{3}})$ であるのに対し, アダトロン学習の漸近特性は $\epsilon_g \sim O(t^{-1})$ であり非常に優れている.

さらに, パーセプトロン学習とアダトロン学習の関係を考察することにより, 荷重摂動学習の適応版である適応型荷重摂動学習を提案した. 統計力学的手法を用いた解析の結果, その汎化誤差の漸近特性がアダトロン学習と同様に $\epsilon_g \sim O(t^{-1})$ であることが明らかになった.

謝辞

本研究の一部は科学研究費補助金（基盤 (C)21500228）および平成 22 年度関西大学大学院理工学研究科高度化推進研究費によるものです.

参考文献

[1] D. Saad(ed.), On-Line Learning in Neural Networks, Cambridge University Press, 1998.

[2] 三好 誠司, “ オンライン学習の統計力学的解析 ”, システム/制御/情報, Vol.51, No5, pp.216-223, 2007.

[3] 前田 裕, “ 同時摂動型最適化法とその応用 ”, システム/制御/情報, Vol.52, No2, pp.47-53, 2008.

[4] J. C. Spall,“ A stochastic approximation technique for generating maximum likelihood parameter estimates ” Proc. of American Control Conference, pp.1161-1167, 1987.

[5] J. Alespector, R. Meir, B. Yuhas, A. Jayakumar and D. Lippe,“ A parallel gradient descent method for learning in analog VLSI neural networks ”, Advances in neural information processing systems 5 (S.J.Hanson, J.D.Cowan and C.Lee), Morgan Kaufmann Publisher, pp.836-844, 1993.

[6] G. Cauwenberghs,“ A fast stochastic error-descent algorithm for supervised learning and optimization ”, Advances in neural information processing systems 5 (S.J.Hanson, J.D.Cowan and C.Lee), Morgan Kaufmann Publisher, pp.244-251, 1993.

[7] 平野, 前田, 金田, “ 同時摂動を用いたニューラルネットワークの学習則 ”, 平成 4 年電気関係学会関西支部連合大会予稿集, p. G307, 1992.

[8] J. Werfel, X. Xie, H. Seung, “ Learning curves for stochastic gradient descent in linear feedforward networks ”, Neural Computation, 17, pp.2699-2718, 2005.

[9] 西森 秀稔, スピングラス理論と情報統計力学, 岩波書店, 東京, 1999.

[10] S. Miyoshi, H. Hikawa, and Y. Maeda,“ Statistical mechanical analysis of simultaneous perturbation learning ”, IEICE trans. Fundamentals, Vol.E92-A, pp.1-4, No7, July 2009.

[11] E. Domany, J.L. van Hemmen and K. Shulten(eds.): Model of Neural Networks III, Springer, 1996.