

適応型荷重摂動学習の統計力学

三好 亮介[†] 前田 裕^{††} 三好 誠司^{††}

[†] 関西大学大学院 理工学研究科 〒564-8680 大阪府吹田市山手町 3-3-35

^{††} 関西大学 システム理工学部 〒564-8680 大阪府吹田市山手町 3-3-35

E-mail: †k329499@kansai-u.ac.jp, ††{maedayut,miyoshi}@kansai-u.ac.jp

あらまし 学習機械の可変パラメータに摂動を加える学習則として荷重摂動学習が提案されている。荷重摂動学習の汎化能力は統計力学的手法で解析されており、その結果パーセプトロン学習と同一の漸近特性を有することが明らかになっている。本稿ではよく知られている学習則であるパーセプトロン学習とアダトロン学習の違いを考察し、それを荷重摂動学習に適用することにより適応型荷重摂動学習を提案する。さらにその汎化能力を統計力学的手法で解析する。その結果、提案する学習則がアダトロン学習と同一の優れた漸近特性を有することが明らかになる。

キーワード オンライン学習, 荷重摂動, 同時摂動, 汎化誤差, 統計力学

Statistical Mechanics of Adaptive Weight Perturbation Learning

Ryousuke MIYOSHI[†], Yutaka MAEDA^{††}, and Seiji MIYOSHI^{††}

[†] Graduate School of Science and Engineering, Kansai University

3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680 Japan

^{††} Faculty of Engineering Science, Kansai University

3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680 Japan

E-mail: †k329499@kansai-u.ac.jp, ††{maedayut,miyoshi}@kansai-u.ac.jp

Abstract The weight perturbation learning was proposed as a learning rule which adds perturbation to the variable parameters of learning machines. Generalization performance of the weight perturbation learning was analyzed by statistical mechanical methods. The weight perturbation learning has the same asymptotic generalization property as the Perceptron learning. In this paper we consider difference between the Perceptron learning and the AdaTron learning which are well-known learning rules. Applying the consideration to the weight perturbation learning, we propose the adaptive weight perturbation learning. The generalization performance of the proposed rule is analyzed by statistical mechanical methods. Consequently, it is shown that the proposed learning rule has an outstanding asymptotic property corresponding to the AdaTron learning.

Key words on-line learning, weight perturbation, simultaneous perturbation, generalization error, statistical mechanics

1. ま え が き

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従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。

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、よく知られている学習則であるパーセプトロン学習とアダトロン学習の違いを考察し、それを荷重摂動学習に適用することにより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 B , J とし、またそれぞれに同じ入力 x が入力されるとする。教師 $B = (B_1, \dots, B_N)$, 生徒 $J^m = (J_1^m, \dots, J_N^m)$, 入力 $x^m = (x_1^m, \dots, x_N^m)$ は N 次元ベクトルであり、教師 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。ここで m は時刻ステップである。生徒 J^m 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、 B と J^m の方向余弦は R^m であるとする。また、入力 x^m の各要素 x_i^m 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、 $v^m = B \cdot x^m$, $u^m l^m = J^m \cdot x^m$ とすると、 v^m と u^m 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 B と生徒 J^m 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、

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

である。

生徒 J^m は入力 x^m とそれに対する教師 B の出力を使い、更新を行う。すなわち

$$J^{m+1} = J^m + f^m x^m \quad (2)$$

である。 f^m 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^m はそれぞれ

$$f^m = \eta \operatorname{sgn}(v^m) \quad (3)$$

$$f^m = \eta \Theta(-u^m v^m) \operatorname{sgn}(v^m) \quad (4)$$

$$f^m = \eta |u^m| \Theta(-u^m v^m) \operatorname{sgn}(v^m) \quad (5)$$

である [2]。ここで η は学習係数であり、本稿では正の定数とする。

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^m は入力 x に関する誤差 ϵ^m の平均であり、本稿で考えているモデルの場合

$$\epsilon_g^m = \frac{1}{\pi} \cos^{-1} R^m \quad (6)$$

である [2]。以下では時刻ステップ m を適宜省略する。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]。

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

ここで

$$r = Rl \quad (8)$$

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

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

パーセプトロン学習

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

アダトロン学習

$$\langle fu \rangle = \frac{\eta}{\pi} (R\sqrt{1-R^2} - \cos^{-1} R) \quad (11)$$

$$\langle fv \rangle = \frac{\eta}{\pi} (1-R^2)^{3/2} + R\langle fu \rangle \quad (12)$$

$$\langle f^2 \rangle = -\eta\langle fu \rangle \quad (13)$$

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]。

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

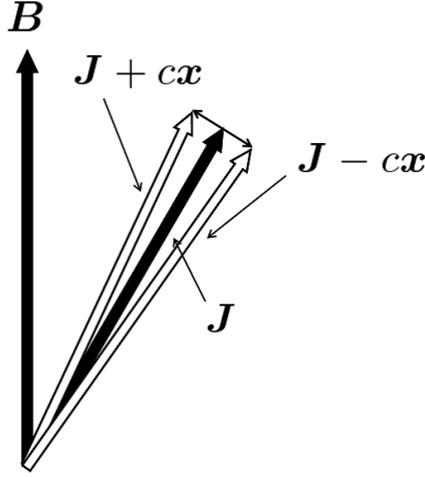


図1 荷重摂動学習における更新

Fig. 1 Update of weight perturbation learning.

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。荷重摂動学習の更新量は

$$f^m = -\frac{\eta}{2c} g^m \quad (15)$$

$$g^m = \Theta(-v^m (\mathbf{J}^m + c\mathbf{x}^m) \cdot \mathbf{x}^m) - \Theta(-v^m (\mathbf{J}^m - c\mathbf{x}^m) \cdot \mathbf{x}^m) \quad (16)$$

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

$$\langle f^u \rangle = -\frac{\eta}{c} \int_{-\frac{c}{l}}^{\frac{c}{l}} Du u H \left(\frac{Ru}{\sqrt{1-R^2}} \right) \quad (17)$$

$$\langle f^v \rangle = \frac{\eta}{c} \int_0^\infty Dv v \left(H \left(\frac{-\frac{c}{l} + Rv}{\sqrt{1-R^2}} \right) - H \left(\frac{\frac{c}{l} + Rv}{\sqrt{1-R^2}} \right) \right) \quad (18)$$

$$\langle f^2 \rangle = \frac{\eta^2}{2c^2} \int_{-\frac{c}{l}}^{\frac{c}{l}} Du u H \left(-\frac{Ru}{\sqrt{1-R^2}} \right) \quad (19)$$

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, \times , \square , \circ , \triangle 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

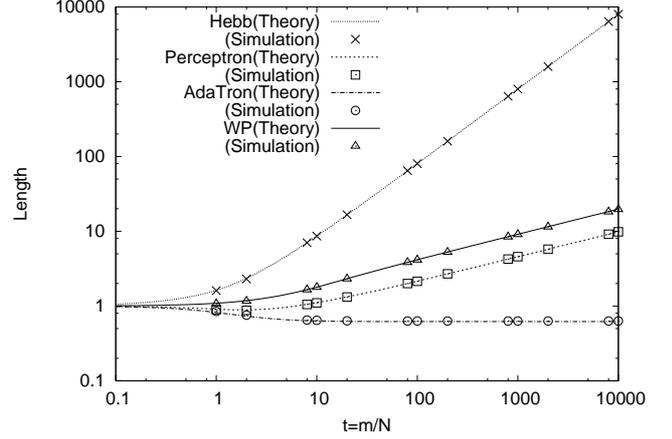


図2 生徒の長さ l のダイナミクス

Fig. 2 Dynamical behaviors of student's length l .

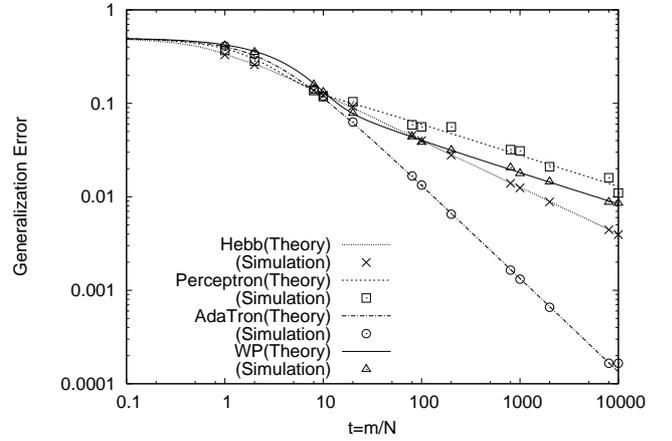


図3 汎化誤差 ϵ_g のダイナミクス

Fig. 3 Dynamical behaviors of generalization error ϵ_g .

(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を用いて数値的に解いた [10]. なおシンプソン則の積分範囲は $-5 \sim +5$, 積分刻みは 0.001, ルンゲ・クッタ法の時間刻みは 0.001 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 で実行しているため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, 荷重摂動学習の漸近特性は $\epsilon_g \sim O(t^{-\frac{1}{3}})$ であり, これはパーセプトロン学習のそれと等しいことがわかる。

5. 適応型荷重摂動学習

5.1 パーセプトロン学習とアダトロン学習

図3よりアダトロン学習の漸近特性は $\epsilon_g \sim O(t^{-1})$ であり非

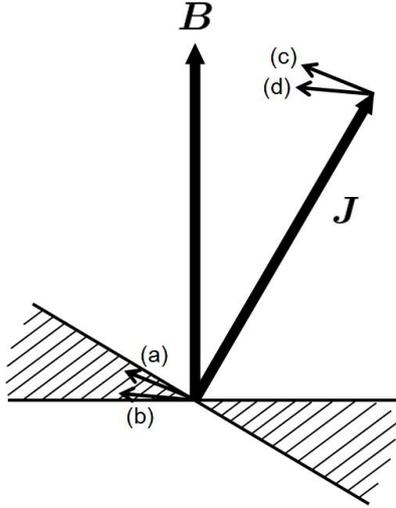


図4 パーセプトロン学習における更新
Fig. 4 Update of Perceptron learning.

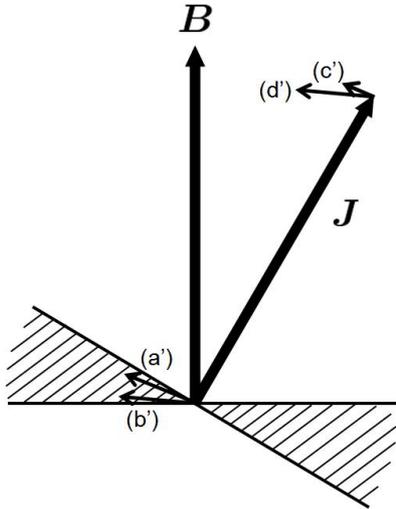


図5 アダトロン学習における更新
Fig. 5 Update of AdaTron learning.

常に優れていることがわかる。また図2よりアダトロン学習だけが生徒の長さ l が単調減少することがわかる。アダトロン学習はパーセプトロン学習の適応型であると言える。そこで本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。 $\eta = 1$ 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

$$f^m = \Theta(-u^m v^m) \text{sgn}(v^m) \quad (20)$$

$$f^m = |u^m| \Theta(-u^m v^m) \text{sgn}(v^m) \quad (21)$$

である。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。

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産された場合である。パーセプトロン学習の場合、入力 x^m 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。いずれの場合も更新ベクトルのノ

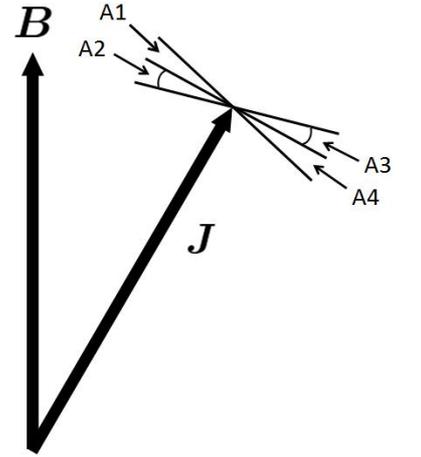


図6 適応型荷重摂動学習
Fig. 6 Adaptive weight perturbation learning.

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 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 J^m 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、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 x^m 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 J^m と入力 x^m 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 x^m の範囲は図6のA1からA4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 x^m がA2とA3に生成された場合に $|u^m|$ に比例する更新を行えばよい。すなわち、更新量 f^m を

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

$$g^m = \Theta(-v^m (J^m + c x^m) \cdot x^m) - \Theta(-v^m (J^m - c x^m) \cdot x^m) \quad (23)$$

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

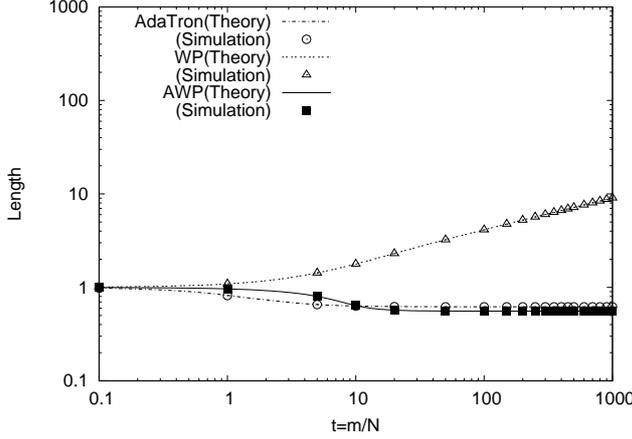


図 7 生徒の長さ l のダイナミクス

Fig. 7 Dynamical behaviors of student's length l .

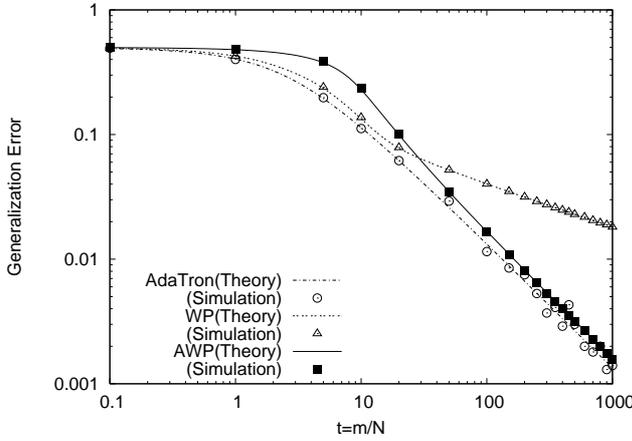


図 8 汎化誤差 ϵ_g のダイナミクス

Fig. 8 Dynamical behaviors of generalization error ϵ_g .

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, \circ , \times , \triangle , \square 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

$$\langle fu \rangle = \int dudv P(u, v) f(v, u, l) u \quad (26)$$

$$\langle fv \rangle = \int dudv P(u, v) f(v, u, l) v \quad (27)$$

$$\langle f^2 \rangle = \int dudv P(u, v) f(v, u, l)^2 \quad (28)$$

の積分が解析的に実行できないのでシンプソン則を用いて数値的に解いた [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, ルング・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。また, $\eta = 1, c = 1$ とした。

図 7, 図 8 より理論と計算機実験はよく一致しており, 本節で

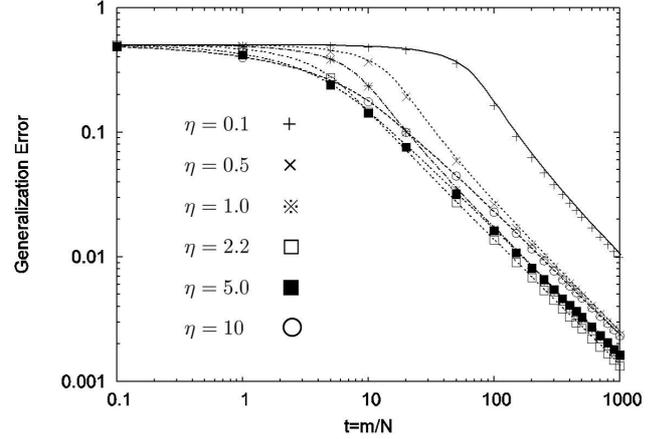


図 9 η を変化させた場合における汎化誤差 ϵ_g のダイナミクス

Fig. 9 Dynamical behaviors of generalization error ϵ_g when η has been changed.

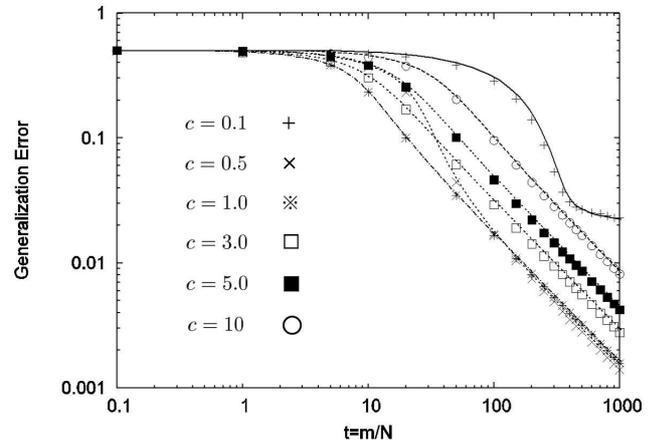


図 10 c を変化させた場合における汎化誤差 ϵ_g のダイナミクス

Fig. 10 Dynamical behaviors of generalization error ϵ_g when c has been changed.

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単調減少することがわかる。これは 5.1 節で述べた, 生徒ベクトルの長さが短くなる方向へ更新ベクトルを大きくした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は $\epsilon_g \sim O(t^{-1})$ であり, アダトロン学習と同一のすぐれた漸近特性を有することがわかる。

次に, η, c を変化させた場合における汎化誤差 ϵ_g のダイナミクスをそれぞれ図 9, 図 10 に示す。こちらの図も同様に曲線は理論計算の結果を表し, $+$, \times , $*$, \square , \triangle , \square のシンボルは計算機実験の結果を表す。図 9 より, η には最適値が存在し, その値は $\eta=2.2$ 付近にあることがわかる。また図 10 より, $c < 1.0$ の場合, 汎化誤差 ϵ_g は特異なふるまいを示すことが分かる。

6. むすび

本稿ではまず, 統計力学的な手法を用いて荷重摂動学習の解析をした結果 [10] について述べた。その結果, 荷重摂動学習の

汎化能力はパーセプトロン学習と同一の漸近特性を有することを確認した。よく知られている学習則であるアダルトロン学習の漸近特性は $\epsilon_g \sim O(t^{-1})$ であり、パーセプトロン学習の $\epsilon_g \sim O(t^{-\frac{1}{3}})$ よりも優れている。そこで、パーセプトロン学習とアダル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 $\epsilon_g \sim O(t^{-1})$ であり、アダルトロン学習と同一の優れた性質を有することが明らかになった。また適応型荷重摂動学習の η, c を変化させた場合に関する汎化誤差の統計力学解析も行った。その結果、学習係数 η は 2.2 付近が最適な学習係数であり、 $c < 1$ において特異なふるまいを示す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。

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. Yuhua, 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, and 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.