3A303-12

入力が相関を有するオンライン学習に関する統計力学的解析

○中尾 健人 † 鳴川 雄太 ‡ 三好 誠司 ¶
† 関西大院 理工学研究科 ‡ 株式会社 ダイヘン ¶ 関西大 システム理工

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．教師あり学習において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に大別できる [1]．バッチ学習は与えられた一定数の例題を繰り返し使用する．バッチ学習の利点は生徒が適度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ということである．しかし，それに至るまでに長時間が必要であり，例題を蓄えるためのメモリが必要になる．一方，オンライン学習は一つの例題が与えられたら，生徒はそれに応じて自分自身を変化させ，その例題は破棄して二度と使用しない．オンライン学習の場合は過去に出た例題に正しく答えられるとは限らない．しかし，例題を蓄えるためのメモリが不要であり，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]．

最近，時間的あるいは空間的な観点で興味深いいくつかのモデルがオンライン学習の枠組みで解析されているが，それらはいずれも入力が独立に生成される場合を扱っている [2]．しかしながら，実際の応用の場面においては入力が相関を有する場合が多くみられる．線形な学習機械が相関入力を用いる場合についてはすでに解析されている [3] が，パターン認識の応用を考えると，学習機械が非線形な場合について解析しておくことは意義がある．そこで本稿では，学習機械が非線形パーセプトロンであり，入力が相関を有する場合のオンライン学習について統計力学的手法を用いて理論的に解析する．

本稿では学習則としてヘブ学習，パーセプトロン学習，アダトロン学習を扱う．解析を行った結果，ヘブ学習とアダトロン学習では入力の相関が大きいほど学習が遅くなるのに対し，パーセプトロン学習では入力の相関の影響を受けないことが明らかになる．

2 モデル

本稿では教師，生徒の二個の非線形パーセプトロンを考え，それぞれの結合荷重を $\boldsymbol{B}, \boldsymbol{J}^m$ とする．ここで m は時刻ステップである．なお，本論文では簡単のため，教師の結合荷重，生徒の結合荷重のことをそれぞれ単に教師，生徒と呼ぶことにする．教師 $\boldsymbol{B} = (B_1, \ldots, B_N)^T$，生徒 $\boldsymbol{J}^m = (J_1^m, \ldots, J_N^m)^T$，入力 $\boldsymbol{x}_k^m = (x_{k1}^m, \ldots, x_{kN}^m)^T$，$k = 1, \ldots, K$ は N 次元ベクトルであり，教師の各要素 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．生徒の初期値の各要素 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．教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}^m$ の関係を図 1 に示す．方向余弦 $\cos\theta_R$ を R と書くことにする．

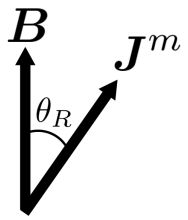

図 1: 教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}^m$

入力 $\boldsymbol{x}_k^m$ の要素 x_{ki}^m は式 (1)-(4) のように生成される．

$$\boldsymbol{\xi}^m = (\xi_1^m, \ldots, \xi_N^m)^T \tag{1}$$

$$\xi_i^m \sim \mathcal{N}(0, 1) \tag{2}$$

$$\boldsymbol{x}_k^m = (x_{k1}^m, \ldots, x_{kN}^m)^T, (k = 1, \ldots, K) \tag{3}$$

$$P\left(x_{ki}^m = \pm\frac{\xi_i^m}{\sqrt{N}}\right) = \frac{1 \pm a}{2} \tag{4}$$

式 (2) は $\boldsymbol{\xi}^m$ の各要素 ξ_i^m が平均 0，分散 1 のガウス分布に従って独立に生成されることを意味する．式 (3)，(4) は $\boldsymbol{\xi}^m$ と方向余弦 a である入力 $\boldsymbol{x}_k^m$ を K 個生成することを表している．このように，親ベクトル $\boldsymbol{\xi}^m$ を中心にしてその周りに方向余弦 a である入力 x_k^m を K 個生成し，この K 個をひとつのセットとして学習に用いる．$\boldsymbol{\xi}^m$ と $\boldsymbol{x}_k^m$ の関係を図 2 に示す．

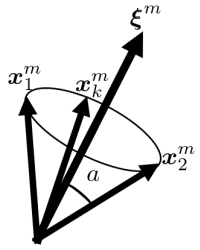

図 2: 相関のある入力

本稿では次元 N が $N \to \infty$ の熱力学的極限を考える．このとき，

$$||\boldsymbol{B}|| = \sqrt{N}, \quad ||\boldsymbol{J}^0|| = \sqrt{N}, \quad ||\boldsymbol{x}_k^m|| = 1 \tag{5}$$

となる．生徒 $\boldsymbol{J}$ のノルムは一般的に時間経過とともに変化するが，初期値 $||\boldsymbol{J}^0||$ に対する比を l とし，これを生徒の長さと呼ぶことにする．すなわち，$||\boldsymbol{J}^m|| = l^m\sqrt{N}$ である．今回，非線形パーセプトロンを考えているので，教師の出力は $\mathrm{sgn}(v)$，生徒の出力は $\mathrm{sgn}(ul)$ である．ここで，

$$v_k^m = \boldsymbol{B} \cdot \boldsymbol{x}_k^m, \qquad u_k^m l^m = \boldsymbol{J}^m \cdot \boldsymbol{x}_k^m \tag{6}$$

であるとする．このとき v_k と u_k 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．

各時刻ステップにおける生徒 $\boldsymbol{J}$ の更新式は

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + \sum_{k=1}^{K} f_k^m \boldsymbol{x}_k^m \tag{7}$$

とする．ここで f_k^m は更新量を表す関数で学習則によって決まる．非線形パーセプトロンの代表的な学習則にはヘブ学習，パーセプトロン学習，アダトロン学習の 3 つがあり，それぞれの更新量は

ヘブ学習

$$f_k^m = \eta \mathrm{sgn}(v_k^m) \tag{8}$$

パーセプトロン学習

$$f_k^m = \eta \Theta(-u_k^m v_k^m)\mathrm{sgn}(v_k^m) \tag{9}$$

アダトロン学習

$$f_k^m = \eta |u_k^m| \Theta(-u_k^m v_k^m)\mathrm{sgn}(v_k^m) \tag{10}$$

である．ここで $\Theta(\cdot)$ はステップ関数

$$\Theta(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

であり，η は学習係数である．

3 理論

3.1 汎化誤差

統計的学習理論の目的の一つは汎化誤差 ϵ_g を理論的に求めることである．汎化誤差 ϵ_g とは未知の入力 $\boldsymbol{x}$ に関する誤差 ϵ の平均である．いま，教師の出力と生徒の出力が一致する場合の誤差 ϵ を 0，一致しない場合の誤差 ϵ を 1 と決めると，汎化誤差 ϵ_g は教師の出力と生徒の出力の異なる確率に等しい．汎化誤差 ϵ_g は式 (12)-(14) のように計算される．

$$\epsilon_g = \langle \epsilon \rangle \tag{12}$$

$$= \int d\boldsymbol{x} P(\boldsymbol{x}) \epsilon \tag{13}$$

$$= \int du dv P(u, v) \epsilon \tag{14}$$

ここで式 (12) の $\langle \cdot \rangle$ は多くの入力 $\boldsymbol{x}$ に関する誤差 ϵ の平均を表している．これを $\boldsymbol{x}$ の確率密度 $P(\boldsymbol{x})$ を用いて表したものが式 (13) であるが，これは N 重積分である．いま $N \to \infty$ を考えているので，このままではこの積分の実行は困難である．しかし，$\boldsymbol{x}$ で決まる v と u が平均 0，分散 1，共分散 R の二次元ガウス分布に従う確率変数であるので，この積分の積分変数を v と u に置き換えて式 (14) のように書くことができる．このガウス積分を解析的に実行することにより，汎化誤差 ϵ_g は

$$\epsilon_g = \frac{1}{\pi} \cos^{-1} R \tag{15}$$

と求まる [2, 4].

3.2 ダイナミクスを記述する連立微分方程式

式 (15) が示すように汎化誤差 ϵ_g は巨視的変数 R の関数であるため，我々は R を知りたい．そのダイナミクスを記述する連立微分方程式を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, 3, 4].

$$\frac{dl^2}{dt} = K \left\langle f_k^2 \right\rangle + K(K-1) \left\langle f_k f_{k'} \right\rangle a^2 + 2Kl \left\langle f_k u_k \right\rangle \tag{16}$$

$$\frac{dr}{dt} = K \left\langle f_k v_k \right\rangle \tag{17}$$

ここで

$$r \equiv Rl \tag{18}$$

である．式 (16)，(17) には 4 つのサンプル平均 $\langle\cdot\rangle$ が含まれる．これらについては具体的な学習則ごとに求める必要がある．

3.3 ヘブ学習

ヘブ学習の場合，サンプル平均は

$$\langle f_k v_k\rangle = \frac{2\eta R}{\sqrt{2\pi}},\ \langle f_k u_k\rangle = \eta\sqrt{\frac{2}{\pi}},\ \left\langle f_k^2\right\rangle = \eta^2 \quad (19)$$

$$\langle f_k f_{k'}\rangle = \eta^2\left(1-\frac{2}{\pi}\cos^{-1}a^2\right) \quad (20)$$

である [2]．式 (19)，(20) を式 (16)，(17) に代入して得られる具体的な微分方程式は解析的に解くことができる．微分方程式の初期条件として $R=0, l=1$ を用いると，

$$l = \eta K\sqrt{\frac{2}{\pi}}t\Bigg(1+\frac{\pi}{2\eta^2K^2}t^{-2} + \frac{\pi}{2K}\left(1+a^2(K-1)\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)\right)t^{-1}\Bigg)^{\frac{1}{2}} \quad (21)$$

$$R = \Bigg(1+\frac{\pi}{2\eta^2K^2}t^{-2} + \frac{\pi}{2K}\left(1+a^2(K-1)\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)\right)t^{-1}\Bigg)^{-\frac{1}{2}} \quad (22)$$

と求まる．式 (15)，(22) より汎化誤差 ϵ_g は

$$\epsilon_g = \frac{1}{\pi}\tan^{-1}\Bigg(\frac{\pi}{2K}\Bigg(1+a^2(K-1)\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)\Bigg)t^{-1}+\frac{\pi}{2\eta^2K^2}t^{-2}\Bigg)^{\frac{1}{2}} \quad (23)$$

となる．さらに t が大きいとき，

$$l \simeq \eta K\sqrt{\frac{2}{\pi}}t \quad (24)$$

$$R \simeq 1-\frac{\pi}{4K}\left(1+a^2(K-1)\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)\right)t^{-1} \quad (25)$$

$$\epsilon_g \simeq \sqrt{\frac{1}{2\pi K}\left(1+a^2(K-1)\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)\right)}t^{-\frac{1}{2}} \quad (26)$$

となる．さらに $K\to\infty$，$a=0$ を考えるとそれぞれ

$$\epsilon_g \simeq \sqrt{\frac{a^2}{2\pi}\left(1-\frac{2}{\pi}\cos^{-1}a^2\right)}t^{-\frac{1}{2}} \quad (27)$$

$$\epsilon_g \simeq \sqrt{\frac{1}{2\pi K}}t^{-\frac{1}{2}} \quad (28)$$

となる．式 (27) より入力に相関がある場合，同時に用いる入力の数 K を増やしても汎化誤差 ϵ_g には下限があることがわかる．式 (28) より入力に相関がない場合には汎化誤差 ϵ_g は $K^{\frac{1}{2}}$ に反比例し減少していくことがわかる．

3.4 パーセプトロン学習

パーセプトロン学習の場合，サンプル平均は

$$\langle f_k v_k\rangle = -\langle f_k u_k\rangle = \eta\frac{R-1}{\sqrt{2\pi}} \quad (29)$$

$$\left\langle f_k^2\right\rangle = \frac{\eta^2}{\pi}\cos^{-1}R \quad (30)$$

$$\langle f_k f_{k'}\rangle = \eta^2\int dv_k du_k dv_{k'} du_{k'} P(v_k, u_k, v_{k'}, u_{k'}) \times\Theta(-u_k v_k)\mathrm{sgn}(v_k)\Theta(-u_{k'}v_{k'})\mathrm{sgn}(v_{k'}) \quad (31)$$

となる．式 (31) の積分は解析的に実行することができないので数値的に解く必要がある．

3.5 アダトロン学習

アダトロン学習の場合，サンプル平均は

$$\left\langle f_k^2\right\rangle = -\eta\langle f_k u_k\rangle = \eta^2\frac{R\sqrt{1-R}-\cos^{-1}R}{\pi} \quad (32)$$

$$\langle f_k v_k\rangle = \eta\frac{(1-R^2)^{\frac{3}{2}}}{\pi}+R\langle f_k u_k\rangle \quad (33)$$

$$\langle f_k f_{k'}\rangle = \eta^2\int dv_k du_k u_k dv_{k'} du_{k'} u_{k'} P(v_k, u_k, v_{k'}, u_{k'}) \times\Theta(-u_k v_k)\mathrm{sgn}(v_k)\Theta(-u_{k'}v_{k'})\mathrm{sgn}(v_{k'}) \quad (34)$$

となる．式 (34) の積分は解析的に実行することができないので数値的に解く必要がある．

3.6 マルコフ連鎖モンテカルロ法

マルコフ連鎖モンテカルロ法（MCMC）は，乱数を用いて確率分布を再現したり，期待値を計算したりするための一群の手法である．その特徴は離散変数，連続変数を問わず，様々な分布に適用できることである．また，確率変数が高次元ベクトルの場合でも適用できる．MCMC にはメトロポリス法，熱浴法，ギブスサンプラーなどがある [5, 6]．

通常の MCMC では分布が多峰性であったり，局所的である場合に領域の移動に大きな時間がかかってしまい効率が低下する．こうした問題を改善した方法に交換モンテカルロ法がある [5, 6, 7]．本稿でも式 (31)，(34) の数値的積分計算に交換モンテカルロ法を用いる．

4　結果と考察

式 (15)，(16)，(17) と各サンプル平均を用いて理論的に計算される汎化誤差 ϵ_g のダイナミクスと計算機実験の結果を重ねて図 3-8 に示す．各学習則とも曲線が理論でシンボルが計算機実験を示している．なお，学習係数 η は 1 とした．計算機実験においては同時入力数 K が 1 , 10, 100 のときは，次元 N は 10^3 とし，汎化誤差 ϵ_g は各時点でランダム入力を 10^5 個生成し，その中で教師と生徒の出力が異なる入力の数をカウントして算出した．相関入力数 K が 10^3 のときは，次元 N は 10^4 とし，汎化誤差 ϵ_g は各時点でランダム入力を 10^6 個生成し，その中で教師と生徒の出力が異なる入力の数をカウントして算出した．

4.1　ヘブ学習

図 3，4 はヘブ学習の汎化誤差 ϵ_g のダイナミクスである．汎化誤差 ϵ_g の理論は式 (23) である．図 3，4 をみると理論と計算機実験がよく一致しており理論解析の結果は正しいとわかる．ただし理論と計算機実験が若干ずれているのは，計算機実験においては有限の次元で実行しているため自己平均性は厳密には破れているからである．

ヘブ学習は $\mathrm{sgn}(v_k^m)\boldsymbol{x}_k^m$ だけ生徒 $\boldsymbol{J}^m$ を更新する．すなわち，生徒の出力には関係なく生徒 $\boldsymbol{J}^m$ を教師 $\boldsymbol{B}$ に近づける．よって，入力に相関がなければ K 個の例題を同時に使う 1 回の更新は，1 個の例題を使う更新を K 回更新することと等価である．つまり，学習速度は K に比例する．しかし，入力に相関がある場合（$a > 0$）は式 (27) に関して述べたように同時に用いる例題を多くしても汎化誤差には下限があることが図 4 よりわかる．

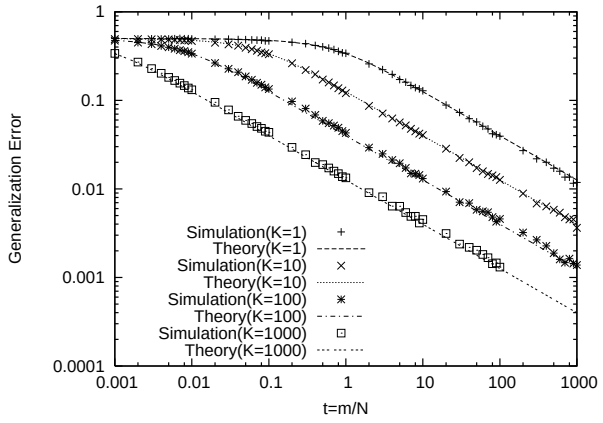

図 3: 汎化誤差 ϵ_g のダイナミクス (ヘブ学習，$a = 0.0$)

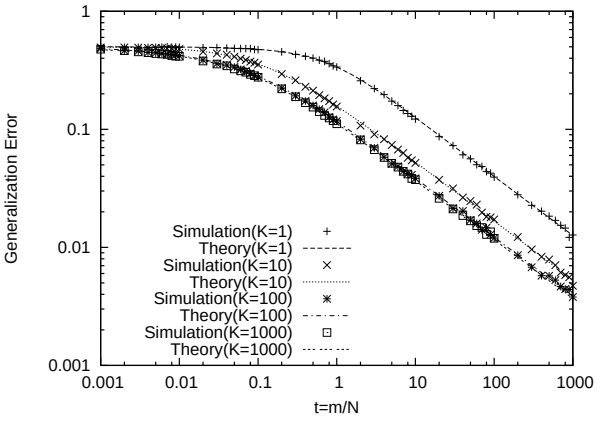

図 4: 汎化誤差 ϵ_g のダイナミクス (ヘブ学習，$a = 0.6$)

4.2　パーセプトロン学習

図 5，6 はパーセプトロン学習の汎化誤差 ϵ_g のダイナミクスである．理論の数値計算の際，式 (16)，(17) に式 (29)-(31) を代入して得られる具体的な連立微分方程式を数値的に解く必要がある．連立微分方程式の数値解法にはルンゲクッタ法を用いた．式 (31) の数値積分に関しては $K = 1, 10, 10^2$ と $K = 10^3$ の $a = 0.0$ の場合はシンプソン則を用いた [8]．$K = 10^3$ の $a = 0.6$ の場合は交換モンテカルロ法を用いた．図 5，6 をみると理論と計算機実験がよく一致しており理論の結果は正しいとわかる．ただし理論と計算機実験が若干ずれているのは，計算機実験においては有限の次元で実行しているため自己平均性は厳密には破れているからである．

パーセプトロン学習は $\Theta(-u_k^m v_k^m)\mathrm{sgn}(v_k^m)\boldsymbol{x}_k^m$ だけ生徒 $\boldsymbol{J}^m$ を更新する．すなわち，教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}$ の出力が異なるときに生徒 $\boldsymbol{J}^m$ を教師 $\boldsymbol{B}$ に近づける．よって，入力に相関が無ければ K 個の例題を同時に使う 1 回の更新は，1 個の例題を使う更新を K 回更新することと等価である．つまり，学習速度は K に比例する．入力に相関がある場合は他二つの学習とは異なっており，同時に用いる例題を多くしても汎化誤差 ϵ_g に下限がないことが図 6 よりわかる．これは汎化誤差 ϵ_g のダイナミクスに関する質的な違いであり，大変興味深い現象である．また，汎化誤差 ϵ_g が初期段階（t = 1〜10）で急速に減少し，その後なだらかに減少していくことがわかる．漸近領域では汎化誤差 ϵ_g は $K^{\frac{1}{3}}$ に反比例することがわかる．

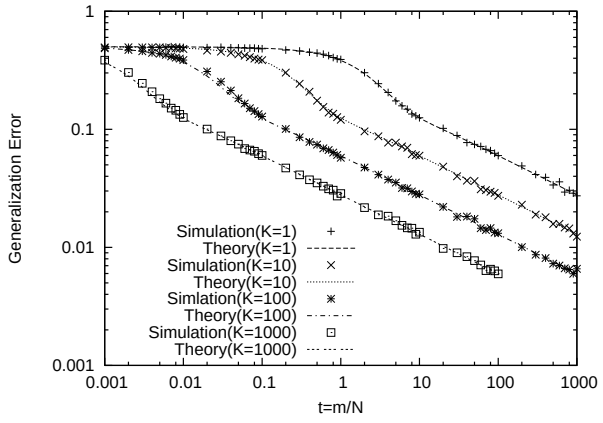

図 5: 汎化誤差 ϵ_g のダイナミクス (パーセプトロン学習, $a = 0.0$)

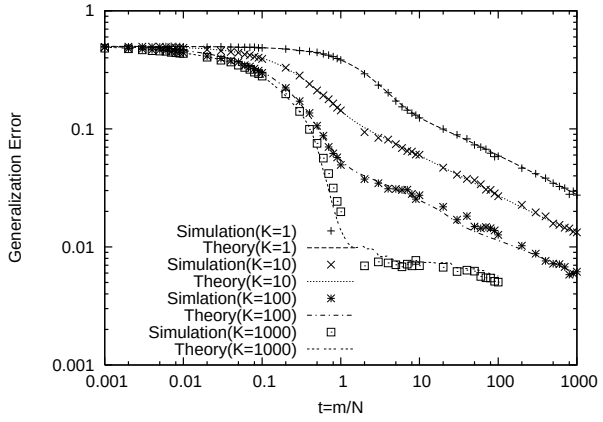

図 6: 汎化誤差 ϵ_g のダイナミクス (パーセプトロン学習, $a = 0.6$)

4.3 アダトロン学習

図 7，8 はアダトロン学習の汎化誤差 ϵ_g のダイナミクスである．理論の数値計算の際，式 (16)，(17) に式 (32)-(34) を代入して得られる具体的な連立微分方程式を数値的に解く必要がある．連立微分方程式の数値解法にはルンゲクッタ法を用いた．式 (34) の数値積分に関してシンプソン則を用いた．[8] 図 7，8 をみると理論と計算機実験がよく一致しており理論の結果は正しいとわかる．ただし理論と計算機実験が若干ずれているのは，計算機実験においては有限の次元で実行しているため自己平均性は厳密には破れているからである．

アダトロン学習は $|u_k^m|\Theta(-u_k^m v_k^m)\mathrm{sgn}(v_k^m)\boldsymbol{x}_k^m$ だけ生徒 $\boldsymbol{J}^m$ を更新する．すなわち，教師 $\boldsymbol{B}$ と生徒 $\boldsymbol{J}$ の出力が異なるときに生徒 $\boldsymbol{J}^m$ を教師 $\boldsymbol{B}$ に近づける．よって，入力に相関がなければ K 個の例題を同時に使う 1 回の更新は，1 個の例題を使う更新を K 回更新することと等価である．つまり，学習速度は K に比例する．しかし，入力に相関がある場合（$a > 0$）は同時に用いる例題を多くしても汎化誤差には下限があることが図 8 よりわかる．

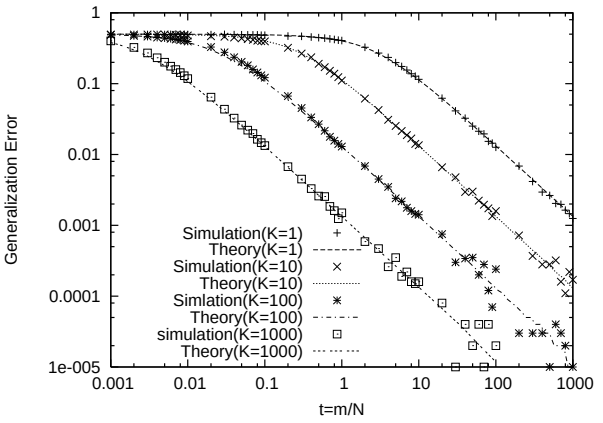

図 7: 汎化誤差 ϵ_g のダイナミクス (アダトロン学習, a=0.0)

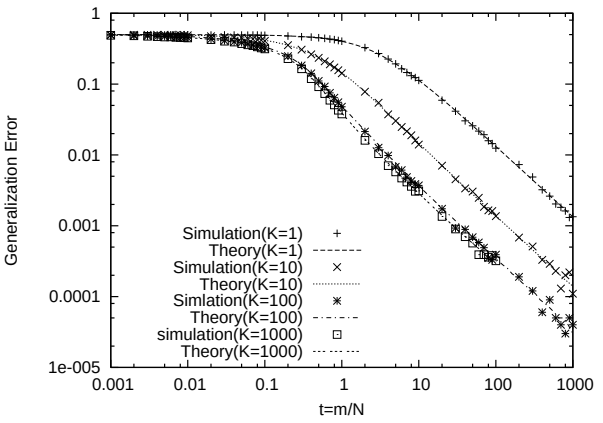

図 8: 汎化誤差 ϵ_g のダイナミクス (アダトロン学習, a=0.6)

5 結論

学習機械が非線形パーセプトロンである場合の入力に相関を有する場合のオンライン学習に統計力学的手法を用いて解析を行った．また，巨視的変数 R と l のダイナミクスを記述する連立微分方程式を決

定論的に導き，微分方程式中に含まれるサンプル平均 $\langle f_k v_k \rangle, \langle f_k u_k \rangle, \langle f_k^2 \rangle, \langle f_k f_{k'} \rangle$ を導出し，これらをもとに汎化誤差を計算した．その理論と計算機実験とよく一致した．ヘブ学習については入力に相関がない場合，反汎化誤差 ϵ_g は $K^{\frac{1}{2}}$ に反比例して減少する．入力に相関がある場合には K を大きくしても下限があることが解析的に明らかになった．パーセプトロン学習については他の二つの学習とは違い，同時に用いる例題を多くしても汎化誤差 ϵ_g に下限がないことが数値的に明らかになった．これは汎化誤差 ϵ_g のダイナミクスに関する質的な違いであり，大変興味深い現象である．また，汎化誤差 ϵ_g が初期段階（t = 1〜10）で急速に減少し，その後なだらかに減少していく．漸近領域では汎化誤差 ϵ_g は $K^{\frac{1}{3}}$ に反比例することが明らかになった．アダトロン学習については入力に相関がない場合，汎化誤差 ϵ_g は K に反比例して減少する．入力に相関がある場合には K を大きくしても下限があることが数値的に明らかになった．

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, No. 5, pp. 216-233, 2007.

[3] C. Seki, S. Sakurai, M. Matsuno, and S. Miyoshi, “A theoretical analysis of on-line learning using correlated examples”, IEICE Trans.Fundamentals, vol. E91-A, no. 9, pp. 2663-2670, Sep. 2008.

[4] 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.

[5] 伊庭 幸人, ベイズ統計と統計物理, 岩波書店, 東京, 2006.

[6] 伊庭 幸人, 種村 正美, 大森 裕浩, 和合 肇, 佐藤 整尚, 高橋 明彦, 計算統計 II, 岩波書店, 東京, 2008.

[7] K. Hukushima and K. Nemoto, “Exchange monte carlo method and application to spin glass simulations”, Journal of Physical Society of Japan, Vol. 65, No. 6, pp. 1604-1608, June 1996.

[8] 戸川 隼人, 数値計算, 岩波書店, 東京, 2002.