自然勾配法は多層パーセプトロンの隠れ層の相関に影響されない

井上 真郷[†‡]　朴 慧暎[†]　岡田 真人[†]

†理化学研究所 脳科学総合研究センター 脳数理研究チーム

〒351-0198　埼玉県和光市広沢 2-1

‡京都大学大学院医学研究科 耳鼻咽喉科・頭頸部外科学

〒606-8507　京都府京都市左京区聖護院河原町 54

E-mail: †{minoue, hypark, okada}@brain.riken.go.jp

あらまし　多層パーセプトロンには隠れ層ユニットの入れ換え対称性が存在する．勾配法では入れ換え対称性に起因する鞍点が存在し，学習の過程でのプラトーの原因の一つになっている．教師の隠れ層ユニットの持つ重みベクトルが似ている場合，この鞍点の影響が強くなることが予想される．しかしながら，これまで重みベクトルの相関がどのように鞍点とプラトーに影響を与えるかを議論した例はほとんど存在しない．本論文では，この相関の影響をソフトコミティーマシンにオンライン学習を適用した場合について統計力学的な手法で議論した．最急降下法では，重みベクトルの相関が増えるにしたがって，プラトーが長くなることがわかった．次に自然勾配法を用いる場合を議論した．強い相関がある場合，鞍点まわりでのフィッシャー情報行列の特異性が強くなる．そのため，自然勾配法が上手く働かない可能性も考えられるが，学習の刻みが小さい極限で，相関によらずにプラトーが消失することがわかった．また，鞍点まわりのダイナミクスを求めることにより，解析的にもこれらの知見を裏付ける結果が得られたので報告する．

キーワード　自然勾配法, パーセプトロン, ソフトコミティーマシン, 特異点, 鞍点, プラトー

Natural Gradient Descent is not Affected by the Correlation of Hidden Layer Units in Multilayer Perceptrons

Masato INOUE[†‡], Hyeyoung PARK[†] and Masato OKADA[†]

†Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute,
2-1, Hirosawa, Wako, Saitama 351-0198, Japan

‡Dept. of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto University
54, Kawara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan

E-mail: †{minoue, hypark, okada}@brain.riken.go.jp

Abstract　The permutation symmetry of the hidden units in multilayer perceptrons causes the saddle structure and plateaus of the learning dynamics in gradient learning methods. The correlation of the weight vectors in the teacher network is supposed to affect this saddle structure resulting in the prolonged learning time, but this mechanism is still unclear. In this paper, we discuss it with regard to the soft committee machines and the on-line learning using statistical mechanics. Conventional steepest gradient descent needs longer time depending on the correlation of the weight vectors. On the other hand, natural gradient descent has no plateaus in the limit of the small learning rate even though the weight vectors have the strong correlation, which worsen the singularity of the Fisher information matrix. Analytical results supports these dynamics around the saddle point.

Keyword　Natural gradient descent, perceptron, soft committee machine, singularity, saddle, plateau

1. はじめに

多層パーセプトロンは隠れユニットに入れ換え対称性があるため，ユニット同士が同じ重みベクトルを持つ場合にそれらを区別できなくなる．このことは，誤差関数を用いて勾配法で学習を行う場合に鞍点構造を生み出し，学習のプラトーの主要な原因となる[1,2]．教師ネットワークの隠れユニット同士に相関がある場合は，学習が進めば生徒ネットワークにも同様の相関が生じるため，必然的にこの鞍点周囲へと近づくことになり，それに伴ってプラトーも増大することが予想される．

一方，自然勾配法はパラメータ空間をリーマン空間とみた場合に，最適な方向へパラメータを動かせるため[3]，鞍点周囲であっても，プラトーに陥ることなく素直に学習が進む可能性が考えられる．また別の見方では，鞍点は自然勾配法での特異点でもあるため，この付近で学習が不安定になる可能性も考えられる．自然勾配法は一般に Fisher 情報行列の逆行列を求めることが難しいといわれるが，パーセプトロンに関しては Yang と Amari [4]によって定式化されているため，本質的な困難は無い．また，逆行列の計算を回避する手法も研究されている[5].

様々な学習手法は一般にバッチとオンラインの二通りの学習のさせ方が存在する．今回我々は，系の鞍点周囲でのダイナミクスを解析するに当たって, 後者を対象とした. これは，後者では学習サンプルを一度しか使用しないため，系の状態と学習サンプルが独立であるため，解析が容易になるメリットがある．

オンライン学習で，更に Saad と Solla [2]が提案した統計物理学的手法を用いると，入力次元が無限大の時の系のダイナミクスを正確に追うことが可能となる．Rattray と Saad [6]は自然勾配法に関しても同手法を用いてソフトコミティーマシンのダイナミクスを解析しているが，自然勾配法が鞍点構造をどう変換するか，又，教師の相関が鞍点構造をどのように悪化させるかといった，多層パーセプトロンの本質的な問題に関しては未だ殆ど議論されていないのが現状である．

そこで，今回我々は，ソフトコミティーマシンでのオンライン学習で，わざと鞍点周囲を通過するよう教師ネットワークを有相関にし，最急降下法と自然勾配法でのダイナミクスを調べたので報告する．

2. モデル

ここでは，教師ネットワークに中間層のユニット数 M，生徒ネットワークにユニット数 K のソフトコミティーマシンを考え，生徒の出力には正規分布ノイズ $n \sim \mathcal{N}(0,\sigma^2)$ が加わるものとする．

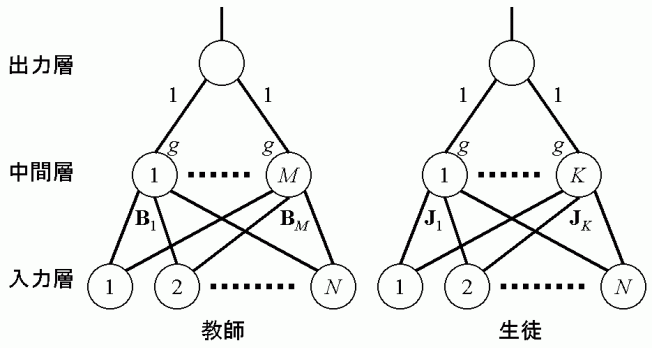

図 1　ソフトコミティーマシン

$$\zeta \equiv f_{\mathbf{B}}(\boldsymbol{\xi}), \qquad f_{\mathbf{B}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{M} g\left(\mathbf{B}_k^T \boldsymbol{\xi}\right) \tag{1}$$

$$\zeta' \equiv f_{\mathbf{J}}(\boldsymbol{\xi}) + n, \quad f_{\mathbf{J}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{K} g\left(\mathbf{J}_k^T \boldsymbol{\xi}\right) \tag{2}$$

$\boldsymbol{\xi}$ は N 次元入力ベクトル，$\mathbf{B}_i$ は教師ネットワークの i 番目の中間層ユニットの入力に対する N 次元重みベクトルで，$\mathbf{J}_i$ も同様である．T は転置を表し，g は中間層の出力関数である．ここでは，自然勾配法を適用するために生徒の出力にノイズを付加する．生徒ネットワークの入力 $\boldsymbol{\xi}$ と出力 ζ' の同時確率分布は

$$p_{\mathbf{J}}(\boldsymbol{\xi},\zeta') \equiv p(\boldsymbol{\xi})\, p_{\mathbf{J}}(\zeta' \mid \boldsymbol{\xi}) \tag{3}$$

$$p_{\mathbf{J}}(\zeta' \mid \boldsymbol{\xi}) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\zeta' - f_{\mathbf{J}}(\boldsymbol{\xi})\right)^2}{2\sigma^2} \right) \tag{4}$$

となる．$N \times K$ 次元パラメータベクトル

$$\mathbf{J} \equiv \left[\mathbf{J}_1^T \quad \mathbf{J}_2^T \quad \cdots \quad \mathbf{J}_K^T \right]^T \tag{5}$$

を調節して，$p_{\mathbf{J}}(\boldsymbol{\xi},\zeta')$ で学習対象の教師ネットワークの入力 $\boldsymbol{\xi}$ と出力 ζ の同時確率分布

$$p(\boldsymbol{\xi},\zeta) \equiv p(\boldsymbol{\xi})\, \delta\left(\zeta - f_{\mathbf{B}}(\boldsymbol{\xi})\right) \tag{6}$$

を近似することを考える．δ はデルタ関数である．一回の学習データ $\{\boldsymbol{\xi}, \zeta\}$ に対する誤差関数を式(4)の条件付確率分布の対数損失を用いて以下のように定義する．

$$\begin{aligned} \varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta) &\equiv -\ln p_{\mathbf{J}}(\zeta \mid \boldsymbol{\xi}) + c_0 \\ &= \frac{1}{2\sigma^2}\left(\zeta - f_{\mathbf{J}}(\boldsymbol{\xi})\right)^2 \end{aligned} \tag{7}$$

ここで，$c_0 \equiv -\ln\sqrt{2\pi\sigma^2}$ は定数である．汎化誤差は学習データの確率分布で平均した誤差である．

$$\varepsilon_g(\mathbf{J}) \equiv \left\langle \varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta) \right\rangle_{\{\boldsymbol{\xi},\zeta\}} \tag{8}$$

式(7)，(8)は $\zeta \equiv f_{\mathbf{B}}(\boldsymbol{\xi})$ であることから，

$$\varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta) = \varepsilon(\mathbf{J},\boldsymbol{\xi}) \equiv \frac{1}{2\sigma^2}\left(f_{\mathbf{B}}(\boldsymbol{\xi}) - f_{\mathbf{J}}(\boldsymbol{\xi})\right)^2 \tag{9}$$

$$\varepsilon_g(\mathbf{J}) = \left\langle \varepsilon(\mathbf{J},\boldsymbol{\xi}) \right\rangle_{\{\boldsymbol{\xi}\}} \tag{10}$$

となる．

本論文では学習サンプル $\{\boldsymbol{\xi}, \zeta\}$ が与えられる毎に $\mathbf{J}$ を更新し，同じ学習サンプルは二度と用いないオンライン学習を議論す

る．通常の最急降下法では，$\mathbf{J}$ の更新則は一回の学習データに対して，適当な学習割合 η を設定して，

$$\Delta\mathbf{J} = -\frac{\eta}{N}\frac{\partial\varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta)}{\partial\mathbf{J}} \tag{11}$$

とする．これが自然勾配法では，パラメータ $\mathbf{J}$ の Fisher 情報行列

$$\mathbf{G} \equiv \left\langle \frac{\partial \ln p_{\mathbf{J}}(\boldsymbol{\xi},\zeta')}{\partial\mathbf{J}} \frac{\partial \ln p_{\mathbf{J}}(\boldsymbol{\xi},\zeta')}{\partial\mathbf{J}^T} \right\rangle_{\{\boldsymbol{\xi},\zeta'\}} \tag{12}$$

の逆行列を用いて

$$\Delta\mathbf{J} = -\frac{\eta}{N}\mathbf{G}^{-1}\frac{\partial\varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta)}{\partial\mathbf{J}} \tag{13}$$

とする[3]. 式(3)から $\mathbf{G}$ を計算すると，小行列形式で以下のように表せる．

$$\begin{aligned} &\mathbf{G} = \frac{1}{\sigma^2}\mathbf{A}, \quad \mathbf{A} \equiv \begin{bmatrix} \mathbf{A}_{1,1} & \cdots & \mathbf{A}_{1,K} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{K,1} & \cdots & \mathbf{A}_{K,K} \end{bmatrix} \\ &\mathbf{A}_{ij} = \left\langle g'\left(\mathbf{J}_i^T\boldsymbol{\xi}\right) g'\left(\mathbf{J}_j^T\boldsymbol{\xi}\right)\boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_{\{\boldsymbol{\xi}\}} \end{aligned} \tag{14}$$

但し，g' は g の導関数である．

3. 理論

3.1. オーダーパラメータと汎化誤差

ここでは，入力次元 $N \to \infty$ での巨視的なネットワークのふるまいを解析する．$N \to \infty$ のため，N 次元ベクトル $\boldsymbol{\xi}$，$\mathbf{B}_i$，$\mathbf{J}_i$ は実際には用いず，代わりにオーダーパラメータであるこれらのベクトル同士の相関を用いて系を記述することができる[2,6].

入力を平均 0, 分散 1 で互いに無相関な正規分布からなる N 次元ベクトル

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{15}$$

とする．更に，

$$x_i \equiv \mathbf{J}_i^T\boldsymbol{\xi}, \quad y_i \equiv \mathbf{B}_i^T\boldsymbol{\xi} \tag{16}$$

と定義すると，正規分布する確率変数同士の和は正規分布なので，$x_i \sim \mathcal{N}\left(0,\mathbf{J}_i^T\mathbf{J}_i\right)$，$y_i \sim \mathcal{N}\left(0,\mathbf{B}_i^T\mathbf{B}_i\right)$ となる．また，それぞれの相関は，

$$\begin{aligned} \left\langle x_i x_j \right\rangle_{\{\boldsymbol{\xi}\}} &= \mathbf{J}_i^T \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_{\{\boldsymbol{\xi}\}} \mathbf{J}_j = \mathbf{J}_i^T\mathbf{J}_j \\ \left\langle x_i y_j \right\rangle_{\{\boldsymbol{\xi}\}} &= \mathbf{J}_i^T \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_{\{\boldsymbol{\xi}\}} \mathbf{B}_j = \mathbf{J}_i^T\mathbf{B}_j \\ \left\langle y_i y_j \right\rangle_{\{\boldsymbol{\xi}\}} &= \mathbf{B}_i^T \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_{\{\boldsymbol{\xi}\}} \mathbf{B}_j = \mathbf{B}_i^T\mathbf{B}_j \end{aligned} \tag{17}$$

となる．従って，

$$\boldsymbol{z} \equiv \left[x_1,\ldots,x_K,y_1,\ldots,y_M\right]^T \tag{18}$$

とすると，これは多変量正規分布 $\mathcal{N}(\mathbf{0},\mathbf{C})$ に従う．

$$p(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^{K+M}|\mathbf{C}|}}\exp\left\{-\frac{1}{2}\boldsymbol{z}^T\mathbf{C}^{-1}\boldsymbol{z}\right\} \tag{19}$$

$\mathbf{C}$ は分散共分散行列で以下のように表せる．

$$\mathbf{C} \equiv \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{R}^T & \mathbf{T} \end{bmatrix} \tag{20}$$

但し，

$$\begin{aligned} &\mathbf{Q} \equiv \mathbf{J}'^T\mathbf{J}' = \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,K} \\ \vdots & \ddots & \vdots \\ Q_{K,1} & \cdots & Q_{K,K} \end{bmatrix} \\ &\mathbf{R} \equiv \mathbf{J}'^T\mathbf{B}' = \begin{bmatrix} R_{1,1} & \cdots & R_{1,M} \\ \vdots & \ddots & \vdots \\ R_{K,1} & \cdots & R_{K,M} \end{bmatrix} \\ &\mathbf{T} \equiv \mathbf{B}'^T\mathbf{B}' = \begin{bmatrix} T_{1,1} & \cdots & T_{1,M} \\ \vdots & \ddots & \vdots \\ T_{M,1} & \cdots & T_{M,M} \end{bmatrix} \\ &\mathbf{J}' \equiv \begin{bmatrix} \mathbf{J}_1 & \cdots & \mathbf{J}_K \end{bmatrix},\ \mathbf{B}' \equiv \begin{bmatrix} \mathbf{B}_1 & \cdots & \mathbf{B}_M \end{bmatrix} \end{aligned} \tag{21}$$

である．$\mathbf{Q}$，$\mathbf{R}$ は系のオーダーパラメータである．

汎化誤差は，

$$\varepsilon_g(\mathbf{J}) = \int dz\, p(z)\frac{1}{2\sigma^2}\left(\sum_{k=1}^{M} g(y_k) - \sum_{k=1}^{K} g(x_k)\right)^2 \tag{22}$$

となる．ここ以降では

$$g(x) \equiv \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{23}$$

とすると，汎化誤差は

$$\begin{aligned} \varepsilon_g(\mathbf{J}) = \frac{1}{\pi\sigma^2}\Bigg\{ &-2\sum_{i=1}^{K}\sum_{j=1}^{M}\arcsin\frac{R_{ij}}{\sqrt{Q_{ii}+1}\sqrt{T_{jj}+1}} \\ &+\sum_{i,j=1}^{K}\arcsin\frac{Q_{ij}}{\sqrt{Q_{ii}+1}\sqrt{Q_{jj}+1}} + \sum_{i,j=1}^{M}\arcsin\frac{T_{ij}}{\sqrt{T_{ii}+1}\sqrt{T_{jj}+1}}\Bigg\}, \end{aligned} \tag{24}$$

となり，オーダーパラメータのみで表せる[2].

3.2. オーダーパラメータダイナミクス

次にオーダーパラメータのダイナミクスを求める．

$$\frac{\partial\varepsilon(\mathbf{J},\boldsymbol{\xi},\zeta)}{\partial\mathbf{J}_i} = \frac{1}{\sigma^2}g'(x_i)\left[\sum_{k=1}^{K} g(x_k) - \sum_{k=1}^{M} g(y_k)\right]\boldsymbol{\xi} \tag{25}$$

であるので，式(11)は

$$\Delta\mathbf{J}_i = -\frac{\eta}{\sigma^2 N}\delta_i\boldsymbol{\xi} \tag{26}$$

$$\delta_i \equiv g'(x_i)\left[\sum_{k=1}^{K} g(x_k) - \sum_{k=1}^{M} g(y_k)\right] \tag{27}$$

となる．同様にオーダーパラメータの更新式も次式で与えられる．

$$\begin{aligned} \Delta R_{ij} &= \left(\mathbf{J}_i + \Delta\mathbf{J}_i\right)^T\mathbf{B}_j - \mathbf{J}_i^T\mathbf{B}_j \\ &= \left(\Delta\mathbf{J}_i^T\right)\mathbf{B}_j \\ &= -\frac{\eta}{\sigma^2 N}\delta_i y_j \end{aligned} \tag{28}$$

$$\begin{aligned} \Delta Q_{ij} &= \left(\mathbf{J}_i + \Delta\mathbf{J}_i\right)^T\left(\mathbf{J}_j + \Delta\mathbf{J}_j\right) - \mathbf{J}_i^T\mathbf{J}_j \\ &= \left(\Delta\mathbf{J}_i^T\right)\mathbf{J}_j + \mathbf{J}_i^T\left(\Delta\mathbf{J}_j\right) + \left(\Delta\mathbf{J}_i^T\right)\left(\Delta\mathbf{J}_j\right) \\ &= -\frac{\eta}{\sigma^2 N}\left(\delta_i x_j + \delta_j x_i\right) + \frac{\eta^2}{\sigma^4 N^2}\delta_i\delta_j\boldsymbol{\xi}^T\boldsymbol{\xi} \end{aligned} \tag{29}$$

ここで時間概念を導入し，一回の更新で $\Delta\alpha = 1/N$ 時間が経過するものとすると，$N \to \infty$ で R_{ij}, Q_{ij} の時間微分が求まる．

$$\begin{aligned}\frac{\partial R_{ij}}{\partial \alpha} &= \lim_{\Delta\alpha\to 0}\frac{\Delta R_{ij}}{\Delta\alpha} = \lim_{N\to\infty} N\Delta R_{ij} = -\frac{\eta}{\sigma^2}\left\langle \delta_i y_j \right\rangle_{\{z\}} \\ &= -\frac{\eta}{\sigma^2}\psi_{ij}\end{aligned} \tag{30}$$

これは厳密には，時間 α での確率変数 R_{ij} が $N \to \infty$ で確率収束することから証明される．同様に，

$$\frac{\partial Q_{ij}}{\partial \alpha} = -\frac{\eta}{\sigma^2}\left(\varphi_{ij} + \varphi_{ji}\right) + \frac{\eta^2}{\sigma^4}\upsilon_{ij} \tag{31}$$

となる．但し，$\boldsymbol{\xi}^T\boldsymbol{\xi} \to N$ となることを用いた．また，

$$\psi_{ij} \equiv \left\langle \delta_i y_j \right\rangle_{\{z\}},\quad \varphi_{ij} \equiv \left\langle \delta_i x_j \right\rangle_{\{z\}},\quad \upsilon_{ij} \equiv \left\langle \delta_i \delta_j \right\rangle_{\{z\}} \tag{32}$$

と定義した．

自然勾配法についても同様にダイナミクスを求めることが可能である．まず，Fisher 情報行列の逆行列であるが，

$$\mathbf{A}^{-1} \equiv \frac{1}{\sigma^2}\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{A}_{1,1}^{-1} & \cdots & \mathbf{A}_{1,K}^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{K,1}^{-1} & \cdots & \mathbf{A}_{K,K}^{-1} \end{bmatrix} \tag{33}$$

と定義すると，$\mathbf{A}_{ij}^{-1}$ は，

$$\mathbf{A}_{ij}^{-1} = \theta_{ij}\mathbf{I} + \mathbf{J}'\boldsymbol{\Theta}_{ij}\mathbf{J}'^T \tag{34}$$

という形で表せる[4]．θ_{ij}，$\boldsymbol{\Theta}_{ij}$ はそれぞれスカラー，$K\times K$ 行列で，オーダーパラメータ $\mathbf{Q}$ を用いて表せる．

自然勾配法での $\mathbf{J}$ の更新式は式(13)より

$$\Delta\mathbf{J}_i = -\frac{\eta}{N}\sum_{k=1}^{K}\delta_k \mathbf{A}_{ik}^{-1}\boldsymbol{\xi} \tag{35}$$

となり，各々の時間微分は最急降下法と同様に

$$\frac{\partial R_{ij}}{\partial \alpha} = -\eta\sum_{k=1}^{K}\left(\theta_{ik}\psi_{kj} + \varphi_{k\bullet}\boldsymbol{\Theta}_{ki}R_{\bullet j}\right) \tag{36}$$

$$\begin{aligned}\frac{\partial Q_{ij}}{\partial \alpha} &= -\eta\sum_{k=1}^{K}\left(\theta_{ik}\varphi_{kj} + \theta_{jk}\varphi_{ki} + \varphi_{k\bullet}\boldsymbol{\Theta}_{ki}Q_{\bullet j} + \varphi_{k\bullet}\boldsymbol{\Theta}_{kj}Q_{\bullet i}\right) \\ &\quad + \eta^2\sum_{k,l=1}^{K}\theta_{ik}\theta_{jl}\upsilon_{kl},\end{aligned} \tag{37}$$

となる．ここで，$\varphi_{s\bullet}$ は，$\{\varphi_{ij}\}_{i,j=1,\ldots,K}$ を行列に見立てた時の第 s 行を意味する．$R_{\bullet j}$ は行列 $\mathbf{R}$ の第 j 列で，他も同様である．また，$\boldsymbol{\Theta}_{ij}^T = \boldsymbol{\Theta}_{ji}$ を用いた．

4. 数値実験

ここでは解析を簡単にするため，中間層のユニット数を以下のようにする．

$$K = M = 2 \tag{38}$$

また，パラメータの初期条件が

$$\begin{aligned} T_{1,1} &= T_{2,2},\quad Q_{1,1} = Q_{2,2} \\ R_{1,1} &= R_{2,2},\quad R_{1,2} = R_{2,1}\end{aligned} \tag{39}$$

となるよう設定した．系の対称性から，学習が進んでもこの条件は崩れない．具体的な数値は，

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\quad \mathbf{R} = \begin{bmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{bmatrix},\quad \mathbf{T} = \begin{bmatrix} 1 & T_{1,2} \\ T_{1,2} & 1 \end{bmatrix} \tag{40}$$

とした．

ここで，各変数の大雑把な大きさを N を基準に整理しよう．入力ベクトル $\boldsymbol{\xi}$ の長さは $O(\sqrt{N})$ である．一方，x_i 及び y_i は非線形関数 g の引数であるので，$O(1)$にしなければならない．すると，教師及び生徒の重みベクトルの長さ$\|\mathbf{B}_i\|$，$\|\mathbf{J}_i\|$は $O(1)$ にすべきである．故に T_{ii}，Q_{ii} も $O(1)$となる．また，入力次元 N が高次元の場合，ランダムに決定された初期 $\mathbf{J}_i$ と $\mathbf{B}_i$ の相関 R_{ii} は $O(1/\sqrt{N})$ である．以上から，上記初期値を決定した．

教師ベクトル同士の相関($T_{1,2}$)を変化させて，最急降下法と自然勾配法での汎化誤差の時間変化を比較した．他のパラメータは以下のように設定した．

$$\eta = 10^{-2},\quad \sigma^2 = 10^{-1} \tag{41}$$

尚，Fisher 情報行列の逆行列はこのケースでは

$$\begin{aligned}&\theta_{1,1} = \theta_{2,2} = c\sqrt{a} \\ &\boldsymbol{\Theta}_{1,1} = d\sqrt{a}\begin{pmatrix} 2a(a-b) - bQ_{1,1} & bQ_{1,2} \\ bQ_{1,2} & -bQ_{1,1} \end{pmatrix} \\ &\boldsymbol{\Theta}_{2,2} = d\sqrt{a}\begin{pmatrix} -bQ_{1,1} & bQ_{1,2} \\ bQ_{1,2} & 2a(a-b) - bQ_{1,1} \end{pmatrix} \\ &\theta_{1,2} = \theta_{2,1} = -c\sqrt{b} \\ &\boldsymbol{\Theta}_{1,2} = \boldsymbol{\Theta}_{2,1} = -d\sqrt{b}\begin{pmatrix} a(1+Q_{1,1}) - b^2 & aQ_{1,2} \\ aQ_{1,2} & a(1+Q_{1,1}) - b^2 \end{pmatrix} \\ &a \equiv (Q_{1,1}+1)^2 - Q_{1,2}^2,\quad b \equiv 2Q_{1,1} + 1, \\ &c \equiv \frac{\pi}{2}\frac{\sqrt{ab}}{a-b},\quad d \equiv \frac{c}{a^2 - b^2}\end{aligned} \tag{42}$$

となる．

汎化誤差の時間変化を図 2 に表す. 最急降下法(図 2a)では，汎化誤差は，教師ベクトル同士の相関が強くなるにつれてプラトーが増大するのに対し，自然勾配法(図 2b)では初期相関 $R_{1,1}$ に比して η を十分小さく取った場合, 汎化誤差は教師ベクトル同士の相関に関係なく，プラトーが殆どみられず，ほぼ指数的に減少した．

次にオーダーパラメータのダイナミクスを図 3 に表す．学習によって変化するパラメータは全部で 4 つ($Q_{1,1}$, $Q_{1,2}$, $R_{1,1}$, $R_{1,2}$) あるため，全てを表示できないので，ここでは R を採用した．即ち，R は生徒ベクトルと教師ベクトルの相関であるので，最初はほとんど無相関(=0)であり，学習が進むにつれ完全に一致(=1)するようになる．図には，二つの生徒の位置の時間変化をプロットし, 座標はそれぞれ($R_{1,1}$, $R_{1,2}$), ($R_{2,1}$, $R_{2,2}$)で，自分の教師との相関及び，自分の教師ではない教師との相関である (対称条件のため, 二つのプロットは鏡像になる)．正解地点は◇印：(1, $T_{1,2}$)，($T_{1,2}$, 1)となる．

最急降下法(図 3a)では，□から出発し，△で折り返した後，

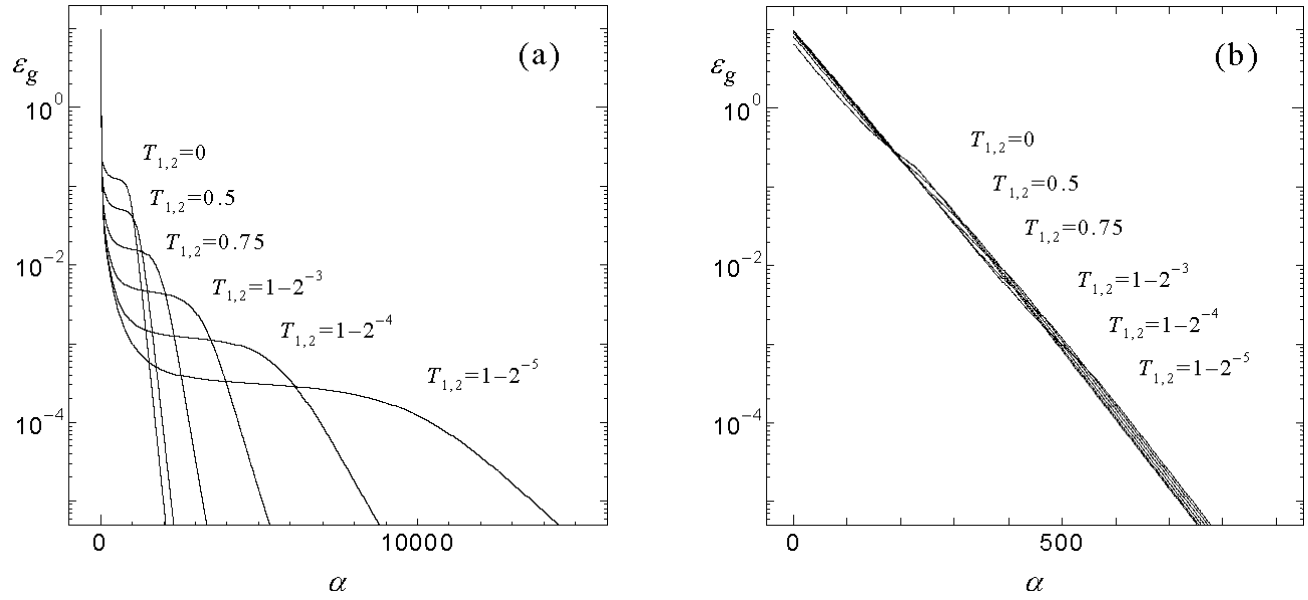

図 2　(a)最急降下法，(b)自然勾配法での汎化誤差の時間変化．(b)では全ての曲線がほぼ重なっている．

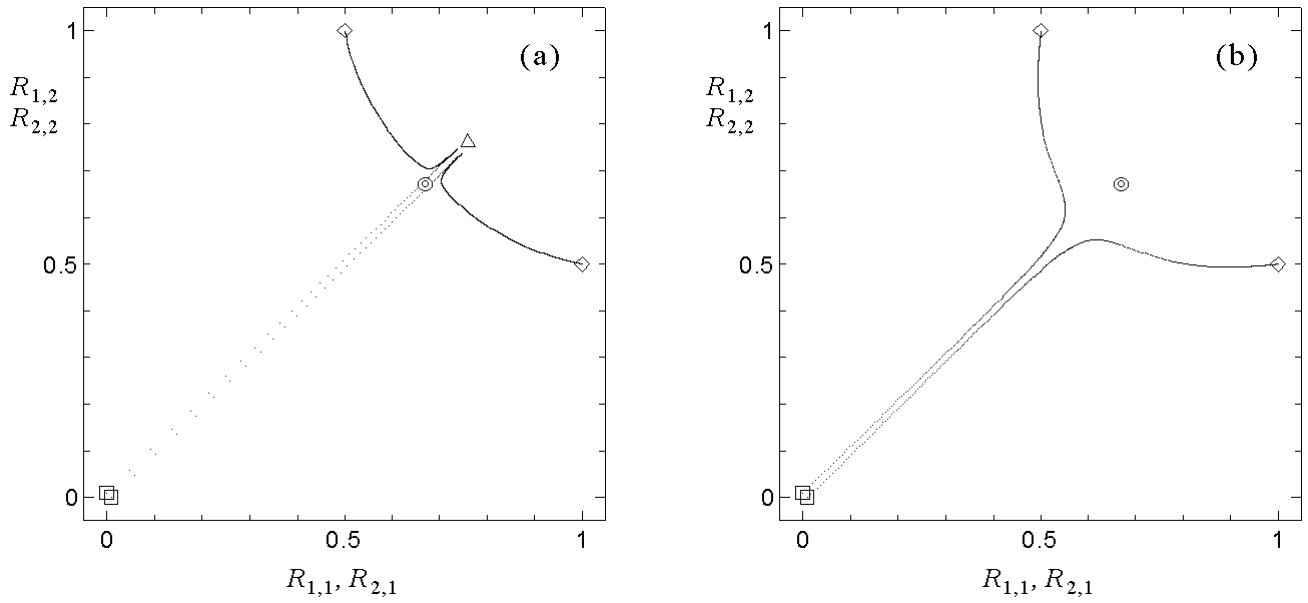

図 3　(a)最急降下法，(b)自然勾配法でのオーダーパラメータ R の時間変化．教師同士の相関 $T_{1,2}=0.5$．□：スタート地点，△：折り返した地点，◎：鞍点，◇：正解地点．

◎に近づき，最後に◇に到達した．同じ所を往復しているようにみえるのは，残りのパラメータ（$Q_{1,1}$，$Q_{1,2}$）を表示していないためで，実際には同じ所を通っているわけではない．

自然勾配法(図 3b)では，同様に□から出発した後，◎を避けるようにして◇に到達した．

5. 鞍点でのダイナミクス

次に，上記条件（(23)，(38)，(39)）下で，汎化誤差が鞍点となっている点を解析した．数値実験では，図 3 で◎印で示した所で，最急降下法ではこの地点に近づいてプラトーが起こり，自然勾配法では特異点となっていて，図では避けて通っているように見える点である．

$$Q_{1,1}=Q_{1,2}=\frac{T_{1,1}+T_{1,2}}{T_{1,1}-T_{1,2}+2}$$
$$R_{1,1}=R_{1,2}=\frac{T_{1,1}+T_{1,2}}{\sqrt{2\left(T_{1,1}-T_{1,2}+2\right)}} \tag{43}$$

この点は，二つの生徒ベクトルが互いに等しく，しかも，二つの教師ベクトルが張る平面上で，両方のベクトルからの距離が等しい線上に丁度乗っている状態である．今関心がある

のは，鞍点から離れて正解方向へ向かうことなので，図 4 のような平面上のダイナミクスを議論すれば十分である．そこで，相関を角度を用いて

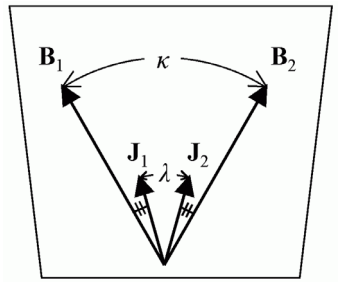

図 4　鞍点は λ = 0 の時で，$\mathbf{J}_1 = \mathbf{J}_2$ は $\mathbf{B}_1$ と $\mathbf{B}_2$ が張る平面上に載っている．

$$\begin{aligned} T_{1,2} &= T_{1,1}\cos\kappa \\ Q_{1,2} &= Q_{1,1}\cos\lambda \end{aligned} \tag{44}$$

と定義して，$\lambda \approx 0$ での鞍点から遠ざかる角速度を調べた．

$\frac{\partial Q_{ij}}{\partial \alpha}$ には η^2 の項が含まれているが，学習割合 η を十分小さくすれば，この項を無視できる．

最急降下法では，

$$\frac{\partial \lambda}{\partial \alpha} = \lambda \frac{4\eta}{\pi\sigma^2} \frac{T_{1,1}\sin^2\kappa}{\left(T_{1,1}(1-\cos\kappa)+2\right)^{\frac{1}{2}}\left(T_{1,1}(3+\cos\kappa)+2\right)^{\frac{3}{2}}} \tag{45}$$

となり，生徒の移動速度は 0 に一次収束しており，更に，教師ベクトル間の相関が高い程速度が小さくなっている．このことは，先のシミュレーションで教師ベクトル間の相関が高い程プラトーが長くなった事と一致する．

自然勾配法では，

$$\frac{\partial \lambda}{\partial \alpha} = \frac{1}{\lambda} 2\eta \tan^2\frac{\kappa}{2} \tag{46}$$

となり，生徒の移動速度は逆数で発散しており，この地点からは急速に弾かれて遠ざかることが分かる．先のシミュレーションでは，そもそもこの地点にはあまり近づかなかったが，このことと合致する．また，例え近づいたとしてもプラトーになることなく，急速に遠ざかることが示唆される．

6. まとめ

ソフトコミティーマシンでのオンライン学習において，最急降下法では教師ネットワークの中間層ユニットの相関に依存してプラトーが長くなるのに対し，自然勾配法では相関にほとんど影響されなかった．また，学習割合を十分小さくとれば，プラトー自体が生じなかった．

プラトーが生じる鞍点付近でのダイナミクスを解析的に求めたところ，最急降下法では鞍点から正解地点へと遠ざかる速度が 0 に一次収束していたのに対し，自然勾配法では無限大に発散していた．このことが，プラトーの有無の原因と思われた．

文　献

[1] K. Fukumizu and S. Amari, "Local minima and plateaus in multilayer neural networks," Neural Networks, vol. **13**, pp. 317-327, 2000.

[2] D. Saad and A. Solla, "On-line learning in soft committee machines," Phys. Rev. E, vol. **52**, pp. 4225-4243, 1995.

[3] S. Amari, "Natural gradient works efficiently in learning," Neural Comput., vol. **10**, pp. 251-276, 1998.

[4] H. H. Yang and S. Amari, "Complexity issues in natural gradient descent method for training multilayer perceptrons," Neural Comput., vol. **10**, pp. 2137-2157, 1998.

[5] H. Park, S. Amari and K. Fukumizu, "Adaptive natural gradient learning algorithms for various stochastic models," Neural Networks, vol. **13**, pp. 755-764, 2000.

[6] M. Rattray and D. Saad, "Analysis of natural gradient descent for multilayer neural networks," Phys. Rev. E, vol. **59**, pp. 4523-4532, 1999.