社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

オンライン学習における適応自然勾配法の評価

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・頭頚部外科学
〒 606–8507 京都府京都市左京区聖護院河原町 54
††† 科学技術振興事業団 戦略的創造研究推進事業 (さきがけ研究 21)「協調と制御」研究領域
〒 351–0198 埼玉県和光市広沢 2–1

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

あらまし　適応自然勾配法は自然勾配法における Fisher 情報行列の逆行列を近似的に求めるもので，1) 入力の分布を予め知っている必要が無い，2) 入力次元を N とすると，$O(N^3)$ だった計算コストを $O(N^2)$ にする，という利点がある．本研究ではソフトコミティーマシンでのオンライン学習の枠組みで，統計力学的手法を用いて，適応自然勾配法の性能を評価した．その結果，ネットワークパラメータの更新割合と Fisher 情報行列の逆行列の推定値の更新割合との比を小さくすれば，自然勾配法とほぼ同等の性能が出ることが分かった．

キーワード　適応自然勾配法，自然勾配法，多層パーセプトロン，ソフトコミティーマシン，プラトー，統計力学

Evaluation of Adaptive Natural Gradient Descent in On-Line Learning

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

† Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351–0198, Japan
†† Department of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine,
Kyoto University, Kyoto 606–8507, Japan
††† "Intelligent Cooperation and Control," PRESTO, JST, Saitama 351–0198

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

Abstract　Adaptive natural gradient descent (ANGD) approximates the inverse of Fisher information matrix directly. Compared to natural gradient descent (NGD), it has following advantages ; 1) it works without any knowledge of the input distribution, and 2) it reduces the caliculation cost from $O(N^3)$ to $O(N^2)$. In this report, we evaluate ANGD using statistical mechanics techniques in respect of on-line learning of soft committee machines. We also show that it has almost the same performance to NGD if the ratio of the update speed of network parameters to the one of the estimator of the inverse of Fisher information matrix is set small enough.

Key words　adaptive natural gradient descent, natural gradient descent, multilayer perceptron, soft committee machine, plateau, statistical mechanics

1. はじめに

多層パーセプトロンを，誤差関数を定義して勾配法で学習させる場合，隠れユニットの入れ換え対称性に起因する鞍点構造が生ずる [1], [2]．特に最急降下法では，学習の過程で鞍点近くを通過するため，この時に学習曲線のプラトーが生じ，大きな問題となっている．互いに同構造の教師ネットワークと生徒ネットワークを用意し，教師のパラメータを教師が出す出力をもとに生徒が学ぶというケースで議論すると，教師の隠れユニット間で，入力に対する重みベクトルに強い相関がある場合，このプラトーは更に増加することが分かっている [6]．一方，自然勾配法 [3]〜[5] では，学習割合を十分小さく設定すれば，教師の隠れユニット同士の相関に関係なく，プラトーは殆ど生じないという望ましい性質がある [6]．

しかし，自然勾配法は反面，1) 特異点付近での学習が不安定になる，2) 入力の分布を予め知っている必要がある，3)Fisher 情報行列の逆行列の計算が時間がかかる，等の欠点が知られている．この内，1) は依然未知の部分が大きいが，2,3) については，適応自然勾配法が代替方法として提案されている [7], [8]．そこで今回我々は，自然勾配法に見られた望ましい性質が，適応自然勾配法の近似によってどの程度変化しているかを評価した．

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．オンライン学習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．更に統計力学的手法を用いると，本来確率的にしか表されない学習のダイナミクスが，入力次元 $N \to \infty$ で漸近的に決定論的になるため，確率的シミュレーションを繰返し行わなくとも，一般的な性質を引き出し易くなる [1]．

今回我々は，多層パーセプトロンを簡略化したソフトコミティーマシンのオンライン学習の枠組みで，自然勾配法と適応自然勾配法の違いを，統計力学的手法を用いて評価したので報告する．

2. モデル

教師ネットワークは中間層のユニット数 M 個，生徒ネットワークは K 個のソフトコミティーマシンを考える (図 1)．

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

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

但し，$\bullet^T$ は転置を表す．ここで，$\boldsymbol{B}_i \in \mathbb{R}^N$ 及び $\boldsymbol{J}_i \in \mathbb{R}^N$ は，i 番目のユニットの，入力 $\boldsymbol{\xi} \in \mathbb{R}^N$ に対する重みベクトルである．教師ネットワークの確率分布は，入力分布を $p(\boldsymbol{\xi})$ とすると，

$$p(\boldsymbol{\xi}, \zeta) \equiv p(\boldsymbol{\xi}) \delta(\zeta - f_{\boldsymbol{B}}(\boldsymbol{\xi})) \tag{3}$$

となる．但し，$\delta(\bullet)$ はデルタ関数である．生徒ネットワークには，正規分布ノイズ $n \sim \mathcal{N}(0, \sigma^2)$ が加えられることとすると，確率分布は，

$$p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta') \equiv p(\boldsymbol{\xi}) p_{\boldsymbol{J}}(\zeta' | \boldsymbol{\xi}) \tag{4}$$

$$p_{\boldsymbol{J}}(\zeta' | \boldsymbol{\xi}) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\{\zeta' - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2}{2\sigma^2} \right) \tag{5}$$

となる．学習によってパラメータベクトル $\boldsymbol{J} \equiv [\boldsymbol{J}_1^T, \boldsymbol{J}_2^T, ..., \boldsymbol{J}_K^T]^T$

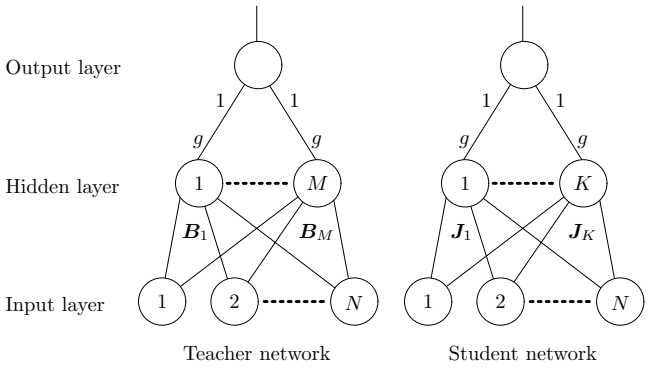

図 1　ソフトコミティーマシン

$\in \mathbb{R}^{KN}$ を少しずつ更新することで，生徒の確率分布を教師の確率分布に近づけることを考える．

与えられた個々の正解サンプル $\{\boldsymbol{\xi}, \zeta\}$ に対する誤差関数を，対数損失を用いて

$$\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) \equiv -\ln p_{\boldsymbol{J}}(\zeta | \boldsymbol{\xi}) + c_0 = \frac{1}{2\sigma^2} \{\epsilon_d(\boldsymbol{\xi})\}^2 \tag{6}$$

$$\epsilon_d(\boldsymbol{\xi}) \equiv f_{\boldsymbol{J}}(\boldsymbol{\xi}) - f_{\boldsymbol{B}}(\boldsymbol{\xi}) \tag{7}$$

で定義する．但し，$c_0 \equiv -\ln \sqrt{2\pi\sigma^2}$ はパラメータが完全に一致している ($\boldsymbol{J} = \boldsymbol{B}$) 時の誤差を 0 にする定数である．また，汎化誤差を

$$\epsilon_g(\boldsymbol{J}) \equiv \langle \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) \rangle_{\{\boldsymbol{\xi}, \zeta\}} \tag{8}$$

で定義する．但し，$\langle \bullet \rangle_{\{\circ\}}$ は，確率変数 $\circ$ での平均である．

勾配法のパラメータ更新則は一般に，

$$\Delta \boldsymbol{J} = -\frac{\eta}{N} \boldsymbol{M} \nabla_{\boldsymbol{J}} \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) \tag{9}$$

で表され，$\boldsymbol{M} = \boldsymbol{I}$(単位行列) とするのが通常の最急降下法，$\boldsymbol{M} = \boldsymbol{G}^{-1}$($\boldsymbol{G}$ は $\boldsymbol{J}$ の Fisher 情報行列) が自然勾配法 [3] で，$\boldsymbol{G}^{-1}$ の代わりにその推定値 $\widehat{\boldsymbol{G}}^{-1}$ を用いるのが適応自然勾配法 [7], [8] である．Fisher 情報行列は，

$$\begin{aligned} \boldsymbol{G} &\equiv \left\langle [\nabla_{\boldsymbol{J}} \ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')] [\nabla_{\boldsymbol{J}} \ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')]^T \right\rangle_{\{\boldsymbol{\xi}, \zeta'\}} \\ &= \frac{1}{\sigma^2} \left\langle [\nabla f] [\nabla f]^T \right\rangle_{\{\boldsymbol{\xi}\}} \end{aligned} \tag{10}$$

で定義される．但し，∇f は，$\nabla_{\boldsymbol{J}} f_{\boldsymbol{J}}(\boldsymbol{\xi})$ の意である．$\boldsymbol{G}$ をオンラインで推定するには，与えられる個々の $\boldsymbol{\xi}$ に対して，更新割合 ($0 < \frac{\rho}{N} < 1$) を設定して，

$$\Delta \widehat{\boldsymbol{G}} \equiv \frac{\rho}{N} \left[-\widehat{\boldsymbol{G}} + \frac{1}{\sigma^2} [\nabla f] [\nabla f]^T \right] \tag{11}$$

で，更新していく方法が考えられる．この時，$\widehat{\boldsymbol{G}}^{-1}$ の差分は Sherman-Morrison 公式を用いて正確に，

$$\Delta \widehat{\boldsymbol{G}}^{-1} = \frac{\frac{\rho}{N}}{1 - \frac{\rho}{N}} \left[\widehat{\boldsymbol{G}}^{-1} - \frac{\frac{1}{\sigma^2} \widehat{\boldsymbol{G}}^{-1} [\nabla f] [\nabla f]^T \widehat{\boldsymbol{G}}^{-1}}{1 + \frac{\rho}{N} \left\{ \frac{1}{\sigma^2} [\nabla f]^T \widehat{\boldsymbol{G}}^{-1} [\nabla f] - 1 \right\}} \right] \tag{12}$$

となる．これが適応自然勾配法で，$\boldsymbol{G}$ でなく $\boldsymbol{G}^{-1}$ を直接推定できるため，逆行列を計算する手間が省ける (逆行列の計算は $O((NK)^3)$，適応自然勾配法の計算は $O((NK)^2)$ のコストである)．また，入力分布 $p(\boldsymbol{\xi})$ を必要としない．$\widehat{\boldsymbol{G}}^{-1}$ の初期値は正定値対称行列であればよいが，$\sigma^2 \boldsymbol{I}$ とすると，汎化誤差を除いて系が σ^2 に依存しなくなる．

3. 理　　論

3.1 オーダーパラメータ

入力を $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ とすると，系の N 次元ベクトル全て $(\boldsymbol{\xi},\ \boldsymbol{B}_i,\ \boldsymbol{J}_i)$ を原点周りで回転させても本質的に同等であるので，これらの相関のみで系を記述できる [1], [4]．即ち，

$$\begin{array}{ll} \boldsymbol{x} \equiv [x_i]_{i=1,\ldots,K}, & x_i \equiv \boldsymbol{J}_i^T \boldsymbol{\xi} \\ \boldsymbol{y} \equiv [y_i]_{i=1,\ldots,M}, & y_i \equiv \boldsymbol{B}_i^T \boldsymbol{\xi} \\ \boldsymbol{Q} \equiv [Q_{ij}]_{i,j=1,\ldots,K}, & Q_{ij} \equiv \boldsymbol{J}_i^T \boldsymbol{J}_j \\ \boldsymbol{R} \equiv [R_{ij}]_{i=1,\ldots,K,j=1,\ldots,M}, & R_{ij} \equiv \boldsymbol{J}_i^T \boldsymbol{B}_j \\ \boldsymbol{T} \equiv [T_{ij}]_{i,j=1,\ldots,M}, & T_{ij} \equiv \boldsymbol{B}_i^T \boldsymbol{B}_j \end{array} \tag{13}$$

である．$f_{\boldsymbol{J}}(\boldsymbol{\xi}), f_{\boldsymbol{B}}(\boldsymbol{\xi})$ は $\boldsymbol{\xi}$ の代わりに $\boldsymbol{x}, \boldsymbol{y}$ を用いてそれぞれ計算できる．また，$\boldsymbol{Q}$ は生徒の重みベクトル同士の相関，$\boldsymbol{T}$ は教師の重みベクトル同士の相関，$\boldsymbol{R}$ は生徒と教師の重みベクトル間の相関である．これにより，入力次元 N が大きい時でも，系のパラメータ数を N に依存しない形に圧縮できる．また，

$$\boldsymbol{z} \equiv [\boldsymbol{x}^T, \boldsymbol{y}^T]^T, \quad \boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{T} \end{bmatrix} \tag{14}$$

とすると，$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$ となる．

汎化誤差は，オーダーパラメータで表すことができる．

$$\epsilon_g = \frac{1}{2\sigma^2} \int p(\boldsymbol{z}) d\boldsymbol{z}\ \{\epsilon_d(\boldsymbol{z})\}^2 \tag{15}$$

$$\epsilon_d(\boldsymbol{z}) = f_{\boldsymbol{J}}(\boldsymbol{x}) - f_{\boldsymbol{B}}(\boldsymbol{y}) = \sum_{k=1}^{K+M} a_k g(z_k) \tag{16}$$

$$a_i \equiv \begin{cases} 1 & (i = 1, ..., K) \\ -1 & (i = K+1, ..., K+M) \end{cases} \tag{17}$$

また，ここから先 $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$，$g$ の一次導関数を $g'(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}$ とすると，汎化誤差は解析的に求まる (付録 1 参照).

$$\epsilon_g = \frac{1}{\pi\sigma^2} \sum_{i,j=1}^{K+M} a_i a_j \arcsin \frac{C_{ij}}{\sqrt{\{C_{ii}+1\}\{C_{jj}+1\}}} \tag{18}$$

3.2 最急降下法

オーダーパラメータの更新式も，オーダーパラメータで表すことができる [1].

$$\nabla_{\boldsymbol{J}i} \epsilon_{\boldsymbol{J}} = \frac{1}{\sigma^2} \epsilon_d(\boldsymbol{z}) \nabla_{\boldsymbol{J}i} f_{\boldsymbol{J}} \tag{19}$$

$$\nabla_{\boldsymbol{J}i} f_{\boldsymbol{J}} = g'(x_i) \boldsymbol{\xi} \tag{20}$$

となるので，最急降下法の場合の $\boldsymbol{J}$ の更新則は

$$\Delta \boldsymbol{J}_i = -\frac{\eta}{\sigma^2 N} \epsilon_d(\boldsymbol{z}) g'(x_i) \boldsymbol{\xi} \tag{21}$$

となり，オーダーパラメータの更新式は

$$\begin{aligned} \Delta R_{ij} &= [\Delta \boldsymbol{J}_i^T] \boldsymbol{B}_j = -\frac{\eta}{\sigma^2 N} \epsilon_d(\boldsymbol{z}) g'(x_i) y_j \\ \Delta S_{ij} &\equiv [\Delta \boldsymbol{J}_i^T] \boldsymbol{J}_j = -\frac{\eta}{\sigma^2 N} \epsilon_d(\boldsymbol{z}) g'(x_i) x_j \\ \Delta Q_{ij} &= [\Delta \boldsymbol{J}_i^T] \boldsymbol{J}_j + [\Delta \boldsymbol{J}_j^T] \boldsymbol{J}_i + [\Delta \boldsymbol{J}_i^T][\Delta \boldsymbol{J}_j] \\ &= \Delta S_{ij} + \Delta S_{ji} + \frac{\eta^2 \{\boldsymbol{\xi}^T \boldsymbol{\xi}\}}{\sigma^4 N^2} \{\epsilon_d(\boldsymbol{z})\}^2 g'(x_i) g'(x_j) \end{aligned} \tag{22}$$

となる．(ΔS_{ij} は便宜上導入した．以下も S については同様．)

次に，時間 α を導入し，一回の更新で $1/N$ の時間が経過するものとし，$N \to \infty$ の極限で，系のパラメータがどのように変化するかを考える．極限をとることで，系のダイナミクスが一意に決まるので，解析が容易になる．具体的には，例えば ΔR_{ij} は確率変数であるが，現在時刻 α より微小時間 τ 後の値は，他のパラメータが固定されているものとして，

$$\begin{aligned} R_{ij}(\alpha+\tau) &= R_{ij}(\alpha) + \lim_{N\to\infty} \sum_{k=0}^{\tau N-1} \Delta R_{ij} \\ &= R_{ij}(\alpha) + \tau \left\langle -\frac{\eta}{\sigma^2} \epsilon_d(\boldsymbol{z}) g'(x_i) y_j \right\rangle_{\{\boldsymbol{z}\}} \end{aligned} \tag{23}$$

と，確率変数でなくなる．従って，現在の微分値は，

$$\begin{aligned} \frac{\partial R_{ij}(\alpha)}{\partial \alpha} &= \lim_{\tau\to 0} \frac{R_{ij}(\alpha+\tau) - R_{ij}(\alpha)}{\tau} \\ &= \left\langle -\frac{\eta}{\sigma^2} \epsilon_d(\boldsymbol{z}) g'(x_i) y_j \right\rangle_{\{\boldsymbol{z}\}} \end{aligned} \tag{24}$$

となる．他も同様で，簡潔のため行列形式で表すと，

$$\begin{aligned} \frac{\partial \boldsymbol{R}}{\partial \alpha} &= -\frac{\eta}{\sigma^2} \left\langle \boldsymbol{\delta} \boldsymbol{y}^T \right\rangle_{\{\boldsymbol{z}\}} \\ \frac{\partial \boldsymbol{S}}{\partial \alpha} &\equiv -\frac{\eta}{\sigma^2} \left\langle \boldsymbol{\delta} \boldsymbol{x}^T \right\rangle_{\{\boldsymbol{z}\}} \\ \frac{\partial \boldsymbol{Q}}{\partial \alpha} &= \frac{\partial \boldsymbol{S}}{\partial \alpha} + \frac{\partial \boldsymbol{S}^T}{\partial \alpha} + \frac{\eta^2}{\sigma^4} \left\langle \boldsymbol{\delta} \boldsymbol{\delta}^T \right\rangle_{\{\boldsymbol{z}\}} \end{aligned} \tag{25}$$

となる．但し，

$$\boldsymbol{\delta} \equiv [\delta_i]_{i=1,\ldots,K}, \quad \delta_i \equiv \epsilon_d(\boldsymbol{z}) g'(x_i) \tag{26}$$

とした．また，ΔQ_{ij} に残っていた $\boldsymbol{\xi}$ は，$N \to \infty$ で $\boldsymbol{\xi}^T \boldsymbol{\xi} \to N$ となる．平均の計算については付録 1 参照.

3.3 自然勾配法

自然勾配法の場合も同様に計算できる [5]．まず $\boldsymbol{G} = [\boldsymbol{G}_{ij}]_{i,j=1,\ldots,K}$ をオーダーパラメータを用いて表す．これは，

$$\boldsymbol{\lambda} \equiv [\lambda_{ij}]_{i,j=1,\ldots,K}$$

$$\boldsymbol{\Lambda} \equiv [\boldsymbol{\Lambda}_{ij}]_{i,j=1,\ldots,K}, \quad \widetilde{\boldsymbol{\Lambda}} \equiv [\widetilde{\boldsymbol{\Lambda}}_{ij}]_{i,j=1,\ldots,K} \tag{27}$$

$$\boldsymbol{\Lambda}_{ij} = \left[\widetilde{\boldsymbol{\Lambda}}_{ij} - \lambda_{ij} \boldsymbol{I}\right] \boldsymbol{Q}^{-1}, \quad \widetilde{\boldsymbol{\Lambda}}_{ij} = \boldsymbol{\Lambda}_{ij} \boldsymbol{Q} + \lambda_{ij} \boldsymbol{I} \tag{28}$$

を用いて，

$$\boldsymbol{G}_{ij} = \sigma^2 \left[\lambda_{ij} \boldsymbol{I} + \boldsymbol{U} \boldsymbol{\Lambda}_{ij} \boldsymbol{U}^T\right] \tag{29}$$

$$= \sigma^2 \left[\lambda_{ij} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T\right] + \boldsymbol{U} \widetilde{\boldsymbol{\Lambda}}_{ij} \boldsymbol{V}^T\right] \tag{30}$$

の二形式で表せる．また $\boldsymbol{G}^{-1} = [\boldsymbol{G}_{ij}^{-1}]_{i,j=1,\ldots,K}$ も，

$$\boldsymbol{\theta} \equiv [\theta_{ij}]_{i,j=1,\ldots,K}$$

$$\boldsymbol{\Theta} \equiv [\boldsymbol{\Theta}_{ij}]_{i,j=1,\ldots,K}, \quad \widetilde{\boldsymbol{\Theta}} \equiv [\widetilde{\boldsymbol{\Theta}}_{ij}]_{i,j=1,\ldots,K} \tag{31}$$

$$\boldsymbol{\Theta}_{ij} = \left[\widetilde{\boldsymbol{\Theta}}_{ij} - \theta_{ij} \boldsymbol{I}\right] \boldsymbol{Q}^{-1}, \quad \widetilde{\boldsymbol{\Theta}}_{ij} = \boldsymbol{\Theta}_{ij} \boldsymbol{Q} + \theta_{ij} \boldsymbol{I} \tag{32}$$

を用いて，

$$\boldsymbol{G}_{ij}^{-1} = \sigma^2 \left[\theta_{ij} \boldsymbol{I} + \boldsymbol{U} \boldsymbol{\Theta}_{ij} \boldsymbol{U}^T\right] \tag{33}$$

$$= \sigma^2 \left[\theta_{ij} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T\right] + \boldsymbol{U} \widetilde{\boldsymbol{\Theta}}_{ij} \boldsymbol{V}^T\right] \tag{34}$$

の二形式で表せることが分かっている [4]．但し，$\boldsymbol{U} \equiv$

$[\boldsymbol{J}_1 \cdots \boldsymbol{J}_K]$ は $N \times K$ 行列，$\boldsymbol{V} \equiv \boldsymbol{U}[\boldsymbol{U}^T\boldsymbol{U}]^{-1} = \boldsymbol{U}\boldsymbol{Q}^{-1}$ は $\boldsymbol{U}$ の一般逆行列で $\boldsymbol{V}^T\boldsymbol{U} = \boldsymbol{I}$ を満たす．また，λ_{ij}，θ_{ij} はスカラー，$\boldsymbol{\Lambda}_{ij}$，$\widetilde{\boldsymbol{\Lambda}}_{ij}$，$\boldsymbol{\Theta}_{ij}$，$\widetilde{\boldsymbol{\Theta}}_{ij}$ は $K \times K$ 行列で，オーダーパラメータより求めることができる (付録 2 参照)．また，$\boldsymbol{\Lambda}_{ij}$ と $\widetilde{\boldsymbol{\Lambda}}_{ij}$ は互いに表裏の関係になって，片方が求まれば $\lambda_{ij}, \boldsymbol{Q}$ を用いて残りも求まる．$\boldsymbol{\Theta}_{ij}$ と $\widetilde{\boldsymbol{\Theta}}_{ij}$ についても同様である．二通りの表現を用いるのは，それぞれに計算式が簡潔になる利点があるためである．

$\boldsymbol{J}$ の更新則は，

$$\Delta \boldsymbol{J}_i = -\frac{\eta}{\sigma^2 N} \sum_{k=1}^{K} \boldsymbol{G}_{ik}^{-1}[\delta_k \boldsymbol{\xi}] \tag{35}$$

なので，$\boldsymbol{G}^{-1}$ を含む他のパラメータが固定されているとして，$N \to \infty$ の極限を計算すると

$$\begin{aligned}
\frac{\partial \boldsymbol{R}}{\partial \alpha} &= -\eta \left[\boldsymbol{\theta} \left\langle \boldsymbol{\delta y}^T \right\rangle_{\{\boldsymbol{z}\}} + \boldsymbol{\Xi R} \right] \\
\frac{\partial \boldsymbol{S}}{\partial \alpha} &\equiv -\eta \left[\boldsymbol{\theta} \left\langle \boldsymbol{\delta x}^T \right\rangle_{\{\boldsymbol{z}\}} + \boldsymbol{\Xi Q} \right] \\
\frac{\partial \boldsymbol{Q}}{\partial \alpha} &= \frac{\partial \boldsymbol{S}}{\partial \alpha} + \frac{\partial \boldsymbol{S}^T}{\partial \alpha} + \eta^2 \boldsymbol{\theta} \left\langle \boldsymbol{\delta \delta}^T \right\rangle_{\{\boldsymbol{z}\}} \boldsymbol{\theta}^T
\end{aligned} \tag{36}$$

となる．但し，

$$\boldsymbol{\Xi} \equiv [\Xi_{ij}]_{i,j=1,\ldots,K},\ \Xi_{ij} \equiv \sum_{k,l=1}^{K} \langle \delta_k x_l \rangle_{\{\boldsymbol{z}\}} [\boldsymbol{\Theta}_{ik}^T]_{lj} \tag{37}$$

とした．

3.4 適応自然勾配法

ここでは，次の二つの近似を使う．1)$\boldsymbol{J}$ と $\widehat{\boldsymbol{G}}^{-1}$ の更新に用いる $\boldsymbol{\xi}$ は同一であるが，そうすると二回目以降の更新で $\boldsymbol{J}$ と $\widehat{\boldsymbol{G}}^{-1}$ に共通の偏りが生じて，微分を求める際にパラメータを固定できなくなるため，別々の独立な $\boldsymbol{\xi}$ を用いるものとして計算する．本来は $\boldsymbol{J}$ 及び $\widehat{\boldsymbol{G}}^{-1}$ のダイナミクスを同時に議論するべきであるが，この仮定により，それぞれの時間微分が別個の平均操作で求まるようになる．尚，η, ρ が共に小さい時はこの近似が成立する．

2) 二点目は，式 (12) の平均操作は分母に確率変数があり難しいため，代わりに $\widehat{\boldsymbol{G}}$ のダイナミクスを求めてその逆行列で代用する点である．$\widehat{\boldsymbol{G}}$ の現在時刻 α より微小時間 τ 後の値は，

$$\begin{aligned}
\widehat{\boldsymbol{G}}(\alpha+\tau) = &\left\{1 - \frac{\rho}{N}\right\}^{\tau N} \widehat{\boldsymbol{G}}(\alpha) \\
&+ \sum_{k=0}^{\tau N - 1} \frac{\rho}{N} \left\{1 - \frac{\rho}{N}\right\}^{k} \frac{1}{\sigma^2} [\nabla f][\nabla f]^T
\end{aligned} \tag{38}$$

で，$\boldsymbol{J}$ が固定されているものとして $N \to \infty$ により，

$$\widehat{\boldsymbol{G}}(\alpha+\tau) = e^{-\rho\tau} \widehat{\boldsymbol{G}}(\alpha) + \{1 - e^{-\rho\tau}\} \boldsymbol{G} \tag{39}$$

$$\frac{\partial \widehat{\boldsymbol{G}}}{\partial \alpha} = \lim_{\tau \to 0} \frac{1 - e^{-\rho\tau}}{\tau} \left[-\widehat{\boldsymbol{G}} + \boldsymbol{G} \right] = \rho \left[-\widehat{\boldsymbol{G}} + \boldsymbol{G} \right] \tag{40}$$

となる．$\widehat{\boldsymbol{G}}^{-1}$ を $[\widehat{\boldsymbol{G}}]^{-1}$ で代用するので，

$$\begin{aligned}
\frac{\partial \widehat{\boldsymbol{G}}^{-1}}{\partial \alpha} &= -\widehat{\boldsymbol{G}}^{-1} \frac{\partial \widehat{\boldsymbol{G}}}{\partial \alpha} \widehat{\boldsymbol{G}}^{-1} \\
&= \rho \left[\widehat{\boldsymbol{G}}^{-1} - \widehat{\boldsymbol{G}}^{-1} \boldsymbol{G} \widehat{\boldsymbol{G}}^{-1} \right]
\end{aligned} \tag{41}$$

とできる．ρ が小さい時はこの近似が成立する．

次に，$\widehat{\boldsymbol{G}}^{-1} = [\widehat{\boldsymbol{G}}_{ij}^{-1}]_{i,j=1,\ldots,K}$ をオーダーパラメータで表すことを考える．今，$\widehat{\boldsymbol{G}}^{-1}$ が先の $\boldsymbol{\theta}, \boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}$ に相当する $\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ を用いて常に表せると仮定する．即ち，

$$\begin{aligned}
&\boldsymbol{\omega} \equiv [\omega_{ij}]_{i,j=1,\ldots,K} \\
&\boldsymbol{\Omega} \equiv [\boldsymbol{\Omega}_{ij}]_{i,j=1,\ldots,K}, \quad \widetilde{\boldsymbol{\Omega}} \equiv [\widetilde{\boldsymbol{\Omega}}_{ij}]_{i,j=1,\ldots,K}
\end{aligned} \tag{42}$$

$$\boldsymbol{\Omega}_{ij} = \left[\widetilde{\boldsymbol{\Omega}}_{ij} - \omega_{ij} \boldsymbol{I} \right] \boldsymbol{Q}^{-1}, \quad \widetilde{\boldsymbol{\Omega}}_{ij} = \boldsymbol{\Omega}_{ij} \boldsymbol{Q} + \omega_{ij} \boldsymbol{I} \tag{43}$$

を用いて，

$$\widehat{\boldsymbol{G}}_{ij}^{-1} = \sigma^2 \left[\omega_{ij} \boldsymbol{I} + \boldsymbol{U} \boldsymbol{\Omega}_{ij} \boldsymbol{U}^T \right] \tag{44}$$

$$= \sigma^2 \left[\omega_{ij} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T \right] + \boldsymbol{U} \widetilde{\boldsymbol{\Omega}}_{ij} \boldsymbol{V}^T \right] \tag{45}$$

とできると仮定する．すると，$\frac{\partial \widehat{\boldsymbol{G}}^{-1}}{\partial \alpha}$ も $\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ を用いて表せる．

$$\begin{aligned}
\frac{\partial \widehat{\boldsymbol{G}}_{ij}^{-1}}{\partial \alpha} = \rho\sigma^2 \Big[&[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T \right] \\
&+ \boldsymbol{U} \left[\widetilde{\boldsymbol{\Omega}} - \widetilde{\boldsymbol{\Omega}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\Omega}} \right]_{[ij]} \boldsymbol{V}^T \Big]
\end{aligned} \tag{46}$$

但し，$[\bullet]_{[ij]}$ は，i, j 成分ではなく i, j ブロック ($K \times K$ 行列) を指すものとする．もし，$\boldsymbol{J}$ のダイナミクスが固定 ($\frac{\partial \boldsymbol{J}}{\partial \alpha} = \boldsymbol{0}$) で，$\widehat{\boldsymbol{G}}^{-1}$ のみを更新していくなら，$\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ のダイナミクスは

$$\frac{\partial \omega_{ij}}{\partial \alpha} = \rho [\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij} \tag{47}$$

$$\frac{\partial \widetilde{\boldsymbol{\Omega}}_{ij}}{\partial \alpha} = \rho \left[\widetilde{\boldsymbol{\Omega}} - \widetilde{\boldsymbol{\Omega}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\Omega}} \right]_{[ij]} \tag{48}$$

となるが，実際には学習則に従って $\boldsymbol{U}$，$\boldsymbol{V}$ も更新されていくので，$\frac{\partial \widetilde{\boldsymbol{\Omega}}}{\partial \alpha}$ はそれを考慮しなければならない．式 (46) の右辺と式 (45) の右辺を微分した式が等しい

$$\begin{aligned}
&\rho\sigma^2 \left[[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T \right] + \boldsymbol{U} \left[\widetilde{\boldsymbol{\Omega}} - \widetilde{\boldsymbol{\Omega}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\Omega}} \right]_{[ij]} \boldsymbol{V}^T \right] \\
&= \sigma^2 \left[\frac{\partial \omega_{ij}}{\partial \alpha} \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{V}^T \right] - \omega_{ij} \left[\frac{\partial \boldsymbol{U}}{\partial \alpha} \boldsymbol{V}^T + \boldsymbol{U} \frac{\partial \boldsymbol{V}^T}{\partial \alpha} \right] \right. \\
&\left. + \frac{\partial \boldsymbol{U}}{\partial \alpha} \widetilde{\boldsymbol{\Omega}}_{ij} \boldsymbol{V}^T + \boldsymbol{U} \widetilde{\boldsymbol{\Omega}}_{ij} \frac{\partial \boldsymbol{V}^T}{\partial \alpha} + \boldsymbol{U} \frac{\partial \widetilde{\boldsymbol{\Omega}}_{ij}}{\partial \alpha} \boldsymbol{V}^T \right]
\end{aligned} \tag{49}$$

ことから，両辺に左右から $\boldsymbol{V}^T$，$\boldsymbol{U}$ を掛けることで

$$\begin{aligned}
\frac{\partial \widetilde{\boldsymbol{\Omega}}_{ij}}{\partial \alpha} = &\rho \left[\widetilde{\boldsymbol{\Omega}} - \widetilde{\boldsymbol{\Omega}} \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\Omega}} \right]_{[ij]} \\
&+ \boldsymbol{\Omega}_{ij} \left[\frac{\partial \boldsymbol{Q}}{\partial \alpha} - \frac{\partial \boldsymbol{S}}{\partial \alpha} \right] - \boldsymbol{Q}^{-1} \frac{\partial \boldsymbol{S}^T}{\partial \alpha} \boldsymbol{\Omega}_{ij} \boldsymbol{Q}
\end{aligned} \tag{50}$$

を得る．式 (47) と (50) を元の式 (49) に代入して，右辺$-$左辺 を計算すると，

$$[\boldsymbol{I} - \boldsymbol{H}] \frac{\partial \boldsymbol{U}}{\partial \alpha} \boldsymbol{\Omega}_{ij} \boldsymbol{Q} \boldsymbol{V}^T + \boldsymbol{U} \boldsymbol{\Omega}_{ij} \boldsymbol{Q} \frac{\partial \boldsymbol{V}^T}{\partial \alpha} [\boldsymbol{I} - \boldsymbol{H}] \tag{51}$$

となり，これは何れの項も $\boldsymbol{0}$ になるので，確かに成立していることが分かる．但し，$\boldsymbol{H} \equiv \boldsymbol{U}\boldsymbol{V}^T$ は，N 次元ベクトルを $\boldsymbol{U}$ の列ベクトルで張る空間に垂直に落とす直交射影行列，$[\boldsymbol{I} - \boldsymbol{H}]$ は，その垂直成分のみを抽出する行列である．

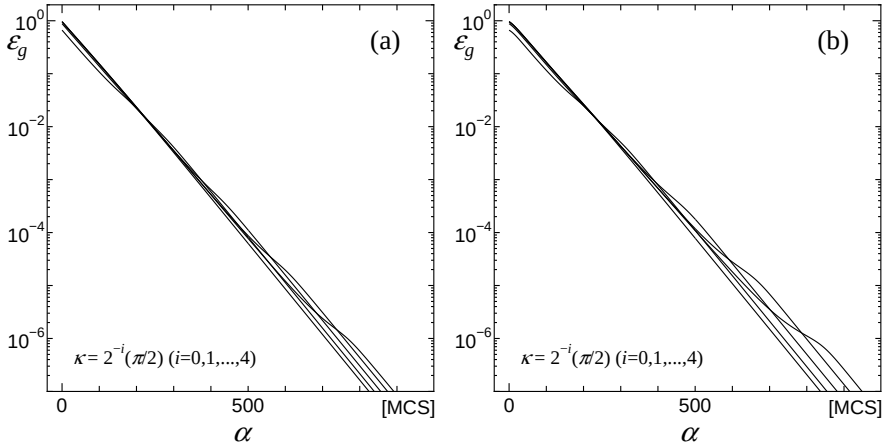

図 2　理論計算による汎化誤差の時間変化．(a) 自然勾配法,
(b) 適応自然勾配法 (近似 1(別個の $\boldsymbol{\xi}$), 2($\widehat{\boldsymbol{G}}^{-1}$ を $[\widehat{\boldsymbol{G}}]^{-1}$ で代用) を使用).

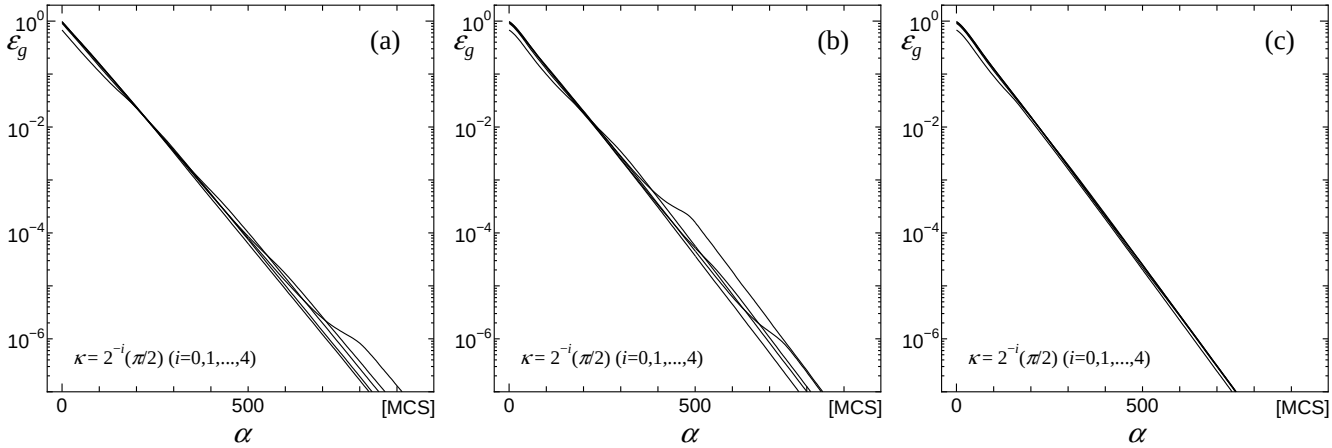

図 3　$N = 500$ での汎化誤差の時間変化．(a) 自然勾配法,
(b) 適応自然勾配法 (近似 1(別個の $\boldsymbol{\xi}$) 使用)，(c) 適応自然勾配法 (近似なし).

$\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ の初期値は，$\boldsymbol{\omega} = \boldsymbol{I}$，$\boldsymbol{\Omega} = \boldsymbol{0}$ とすれば，先の初期値 $\widehat{\boldsymbol{G}}^{-1} = \sigma^2 \boldsymbol{I}$ を満たすことができる．従って，$\widehat{\boldsymbol{G}}^{-1}$ が常に $\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ を用いて表せることが分かる．

$\frac{\partial \boldsymbol{R}}{\partial \alpha}$，$\frac{\partial \boldsymbol{Q}}{\partial \alpha}$ は，前述の自然勾配法のダイナミクスで，$\boldsymbol{\theta}, \boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}$ を $\boldsymbol{\omega}, \boldsymbol{\Omega}, \widetilde{\boldsymbol{\Omega}}$ に置き換えることで同様に計算できる．

4. 結　　果

自然勾配法と適応自然勾配法の理論を，時間幅 $\alpha = 0.1$ で Runge-Kutta 法で数値計算した結果を示す．中間層ユニット数は生徒，教師共に 2 個 ($K = M = 2$) とし，$\boldsymbol{J}$ の更新割合を小さく $\eta = 10^{-2}$，$\widehat{\boldsymbol{G}}^{-1}$ の更新割合を η よりも大きくして $\rho = 10^{-1}$ とした．$\sigma^2 = 1$ は汎化誤差以外には影響しない．生徒，教師の各重みベクトル $\boldsymbol{J}_i, \boldsymbol{B}_i$ については長さを 1 とし，教師間角度 $\kappa \equiv \arccos(\frac{T_{1,2}}{\sqrt{T_{1,1}T_{2,2}}})$ は，幾つか変化させたものを試した．

$$\kappa = 2^{-i}\{\pi/2\} \quad (i = 0, 1, ..., 4) \tag{52}$$

生徒同士の相関及び対応する教師との相関は，$Q_{1,2} = R_{1,1} = R_{2,2} = \frac{1}{\sqrt{500}}$，相手の教師に対する相関は $R_{1,2} = R_{2,1} = \frac{1}{\sqrt{500}} - \frac{2}{\sqrt{500}} \sin\frac{\kappa}{2}$ とした．これらは，$\boldsymbol{J}$ の初期値を $\mathcal{N}(0, \frac{1}{N}\boldsymbol{I})$ でランダムに設定すると，

$$\begin{aligned}
&\langle Q_{ii} \rangle_{\{\boldsymbol{J}\}} = 1 \\
&\sqrt{\langle \{Q_{ij}\}^2 \rangle_{\{\boldsymbol{J}\}}} = \frac{1}{\sqrt{N}} \quad (i \neq j) \\
&\sqrt{\langle \{R_{ij}\}^2 \rangle_{\{\boldsymbol{J}\}}} = \frac{1}{\sqrt{N}} \\
&\sqrt{\langle \{R_{ik} - R_{jk}\}^2 \rangle_{\{\boldsymbol{J}\}}} = \frac{2}{\sqrt{N}} \sin\frac{\kappa}{2} \quad (i \neq j)
\end{aligned} \tag{53}$$

となることから，$N = 500$ を想定した値を用いた．

図 2 に，汎化誤差 ϵ_g の時間変化を示す．(a) の自然勾配法に比べて (b) 適応自然勾配法でもプラトーは殆ど生じず，教師の相関 κ にもあまり影響されていないことが分かる．総じて，$\frac{\eta}{\rho} \ll 1$ の時はプラトーが生じず，そうでない時はプラトーがこの値に応じて増大する傾向があった．これは，$\frac{\eta}{\rho}$ が小さければ，$\boldsymbol{J}$ の変化 (即ち $\boldsymbol{G}^{-1}$ の変化) に $\widehat{\boldsymbol{G}}^{-1}$ の更新が十分ついて行けることによると考えられる．

次に，オリジナルの漸化式 (12) を用いた，$N = 500$ での実シミュレーションを行い，理論を検証した．初期条件は，$\boldsymbol{J} \sim \mathcal{N}(0, \frac{1}{N}\boldsymbol{I})$ とし，他は理論計算と同様とした．結果を図 3 に示す．(a) は自然勾配法，(b)，(c) は適応自然勾配法で，(b) は $\boldsymbol{\xi}$ を $\boldsymbol{J}$ と $\widehat{\boldsymbol{G}}^{-1}$ 用に別個に用意したもの，(c) は本来の近似を用いない手法である．(b)，(c) 共に図 2(b) の理論値と大きくは異なっていないことが分かる．図 3(c) と図 3(b) の相違は，近似 1) の $\boldsymbol{\xi}$ を別個に用意したことによる

もの，図 3(b) と図 2(b) の相違は，近似 2) の$\widehat{\boldsymbol{G}}^{-1}$を$[\widehat{\boldsymbol{G}}]^{-1}$で代用したことによるものと考えられる．

ρ の値が大きい時，適応自然勾配法の理論値は自然勾配法とほぼ同じであったが，近似 1,2 による二つの相違は共に拡大する傾向があり，近似を用いないシミュレーションは，理論値の倍以上の学習スピードを示した．

5. 結　　論

ソフトコミティーマシンのオンライン学習について，統計力学的手法を用いて適応自然勾配法の性能を評価した．その結果，$\frac{\eta}{\rho}$ を十分小さく設定すれば，自然勾配法とほぼ同等の性能を持つことが分かった．また，理論値を求める際に二つの近似を用い，シミュレーションで近似による誤差の程度を評価した．

付　　録

1. ガウス積分

$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}), g(x) \equiv \mathrm{erf}(x/\sqrt{2}), g'(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}$ とすると，以下の積分は解析的に解くことができる [1].

$$\begin{aligned}
\langle g(z_i)g(z_j)\rangle_{\{\boldsymbol{z}\}} &= \frac{2}{\pi}\arcsin\frac{C_{ij}}{\sqrt{\{C_{ii}+1\}\{C_{jj}+1\}}} \\
\langle g'(z_i)g(z_j)z_k\rangle_{\{\boldsymbol{z}\}} &= \frac{2}{\pi}\frac{c_1}{\{C_{ii}+1\}\sqrt{|\boldsymbol{M}_{ij}|}} \qquad \text{(A·1)} \\
\langle g'(z_i)g'(z_j)g(z_k)g(z_l)\rangle_{\{\boldsymbol{z}\}} &= \frac{4}{\pi^2}\frac{1}{\sqrt{|\boldsymbol{M}_{ij}|}}\arcsin\frac{c_4}{\sqrt{c_2 c_3}}
\end{aligned}$$

但し，

$$\boldsymbol{M}_{ij} \equiv \begin{bmatrix} C_{ii}+1 & C_{ij} \\ C_{ij} & C_{jj}+1 \end{bmatrix} \qquad \text{(A·2)}$$

$$c_1 \equiv \begin{vmatrix} C_{ii}+1 & C_{ij} \\ C_{ik} & C_{jk} \end{vmatrix}, c_2 \equiv \begin{vmatrix} \boldsymbol{M}_{ij} & C_{ik} \\ & C_{jk} \\ C_{ik} \quad C_{jk} & C_{kk}+1 \end{vmatrix}$$

$$c_3 \equiv \begin{vmatrix} \boldsymbol{M}_{ij} & C_{il} \\ & C_{jl} \\ C_{il} \quad C_{jl} & C_{ll}+1 \end{vmatrix}, c_4 \equiv \begin{vmatrix} \boldsymbol{M}_{ij} & C_{il} \\ & C_{jl} \\ C_{ik} \quad C_{jk} & C_{kl} \end{vmatrix}$$

これらを用いて，

$$\begin{aligned}
\epsilon_g &= \frac{1}{2\sigma^2}\sum_{i,j=1}^{K+M} a_i a_j \langle g(z_i)g(z_j)\rangle_{\{\boldsymbol{z}\}} \\
\langle \delta_i z_j\rangle_{\{\boldsymbol{z}\}} &= \sum_{k=1}^{K+M} a_k \langle g'(z_i)g(z_k)z_j\rangle_{\{\boldsymbol{z}\}} \qquad \text{(A·3)} \\
\langle \delta_i \delta_j\rangle_{\{\boldsymbol{z}\}} &= \sum_{k,l=1}^{K+M} a_k a_l \langle g'(z_i)g'(z_j)g(z_k)g(z_l)\rangle_{\{\boldsymbol{z}\}}
\end{aligned}$$

が求まる．

2. 自然勾配法

自然勾配法の $\boldsymbol{\lambda}, \boldsymbol{\Lambda}, \widetilde{\boldsymbol{\Lambda}}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}$ は，上記の $\boldsymbol{M}_{ij}$ を用いて以下のように求めることができる [4].

$$\lambda_{ij} = \frac{2}{\pi}\frac{1}{\sqrt{|\boldsymbol{M}_{ij}|}} \qquad \text{(A·4)}$$

$$\boldsymbol{\Lambda}_{ij} = -\lambda_{ij}\begin{bmatrix}\boldsymbol{e}_i & \boldsymbol{e}_j\end{bmatrix}\boldsymbol{M}_{ij}^{-1}\begin{bmatrix}\boldsymbol{e}_i & \boldsymbol{e}_j\end{bmatrix}^T \qquad \text{(A·5)}$$

$$\boldsymbol{\theta} = \boldsymbol{\lambda}^{-1}, \quad \widetilde{\boldsymbol{\Theta}} = \widetilde{\boldsymbol{\Lambda}}^{-1} \qquad \text{(A·6)}$$

但し，$\boldsymbol{e}_i \in \mathbb{R}^K$ は第 i 成分のみ 1 の単位ベクトルである．$\widetilde{\boldsymbol{\Lambda}}$ 及び，$\boldsymbol{\Theta}$ は，式 (28,32) を用いて求まる．

文　　献

[1] D. Saad and A. Solla: Phys. Rev. E **52** (1995) 4225.
[2] K. Fukumizu and S. Amari: Neural Networks **13** (2000) 317.
[3] S. Amari: Neural Comput. **10** (1998) 251.
[4] H. H. Yang and S. Amari: Neural Comput. **10** (1998) 2137.
[5] M. Rattray and D. Saad: Phys. Rev. E **59** (1999) 4523.
[6] M. Inoue, H. Park and M. Okada: J. Phys. Soc. Jpn. **72** (2003) (*in press*).
[7] S. Amari, H. Park and K. Fukumizu: Neural Comput. **12** (2000) 1399.
[8] H. Park, S. Amari and K. Fukumizu: Neural Networks **13** (2000) 755.