メトロポリス的モンテカルロ法
(Gibbs Sampler)
の
分類の問題への応用

Version 9.3

1996年 4月 10日

統計数理研究所

伊庭幸人

目 次

1 本論文の主題 **3**

2 問題の説明 **5**
2.1 対比較の問題と組み分けモデル 5
2.2 ベイズ的解釈 7
2.3 統計物理の問題との関連 7
2.4 ２グループに分類する場合の考察と３グループに分類する場合との比較 8

3 メトロポリス法の適用と生成されたマルコフ過程のダイナミクス **11**
3.1 簡単な例 11
3.2 基底状態の縮退とマルコフ過程のダイナミクス 12
3.3 準安定状態とマルコフ過程のダイナミクス 13
3.3.1 温度零の場合 13
3.3.2 有限温度の場合 16

4 統計量の計算 **20**
4.1 MAP 解の計算 (Simulated Annealing) 20
4.2 周辺分布の計算 (1) 22
4.3 周辺分布の計算 (2) 22
4.4 メトロポリス法の緩和 25

5 まとめ **29**

6 謝辞 **30**

7 注 **31**

8 付録　ベイズ統計と統計物理 **37**
8.1 大規模統計モデルと統計物理アナロジー 37
8.2 メトロポリス的モンテカルロ法 (Gibbs Sampler) 38
8.3 メトロポリス的モンテカルロ法のベイズ統計への応用 . . . 40
8.3.1 Simulation without Annealing 40
8.3.2 ABIC(TYPE II Likelihood) の計算 42
8.3.3 simulated annealing 43

9 付録の注 **44**

1　本論文の主題

本論文はメトロポリス的モンテカルロ法のベイズ統計への応用について，模擬データの数値実験を通じて考察したものである．

基本的な立場は，計算法として統計物理の方法を使うことであるが，その過程で起こってくるさまざまな現象を理解するために，統計物理の諸概念が背景として重要であることを示してゆくつもりである．以下では，統計物理との類似にもとづいて用いた用語を必要に応じて「」で囲って明示することにする．

統計モデルに対するメトロポリス的モンテカルロ法の応用は，現在のところ画像処理・空間パターンの解析の分野が先行している [Geman and Geman(1984), Geman and Geman(1986), Geman and McClure(1987), Ogata and Tanemura(1984,1985,1989), Ogata(1990), Penttinen(1984), Poggio et al.(1985), Marroquin et al.(1987)]．これらの問題は，ミクロな要素 (画素・ラベル) を極めて多く含み，粗視化ということがはっきりした意味を持っているという点で，特に統計物理アナロジーに馴染み易いと考えられる．しかし，メトロポリス法の適用をはじめとする統計物理アナロジーは統計の問題一般に役立つ可能性があるので，画像・空間のみに集中せず，全く違ったタイプの問題への適用も研究すべきであろう．

特に興味深い問題のグループとして，分類のための離散ラベルを持つような問題がある．これらの中には画像問題に含まれるもの (たとえばテクスチャーのラベル付け) もあるが，空間構造を含まないものも多い．これらの問題は空間構造を含まないかわりに，いろいろな独自の構造を含んでいるので，統計学的見地からも，物理系との対応という見地からも興味深い．離散パラメータは数が多くても実質的な自由度は少ないため，なかなか overfit にはならず，特に informative prior を用いない場合でも数が多くなりうる．

いわゆる finite mixture の問題はその 1 例である．これは，各標本が g個の分布 $\{P_k, k = 1..g\}$ のいずれから抽出されたか不明であるという条件のもとで，分布のパラメータと各標本がどの分布に由来するかを同時に推定するという問題である．この種の問題は，各標本にそれがどこから抽出されたかを示す離散パラメータ $\{S_i\}(S_i \in \{1, 2..g\})$ が付属していると考えることによって，多数の離散パラメータを含んだ問題として解釈しなおすことができる．いろいろな問題，たとえば，外れ値 (outlier) の検出問題 [Kitagawa, Akaike (1982)]，クラスター分析 (たとえば [Gordon (1981)] p.49-53 参照) などが finite mixture の問題の形に表わせる．finite mixture の問題は統計物理の問題として見たとき，パラメータ $\{S_i\}$ の間に直接の「相互作用」がないため，画像問題に比べると簡単である．しか

し，問題を少し変えて，データそのものが何種類かの分布から引き出されるとする代わりに，データを作り出す要素が何種類かある場合を考えると，パラメータ $\{S_i\}$ の間に直接の相互作用を持つモデルが生ずる．

このような種類の問題は沢山あるが，これらにメトロポリス的モンテカルロ法を応用することは，統計の立場からはまだあまり行われていない (注 1)．ここでは，その典型的なものとして，“対比較デ– タに基づく分類” あるいは” ラベル付けによる対比較デ– タの解析” の問題をとりあげる．この課題では，リーグ戦の対戦の勝敗表がデータとして与えられているときに，各個体がいくつかのグループのどれかから抽出されているとして，各個体の所属と各グループ間の強弱が問われる．この問題は比較的簡単な数式で記述できるが，本質的な問題点の多くを含んでいる．

メトロポリス的モンテカルロ法について，この例を通じてとくに強調したいことは，

- 「有限温度」では特有の興味深い現象が見られること．
- MAP 解だけでなく，期待値や周辺分布が計算できることが重要だということ．

  このことは Marroquin その他も論じている (付録参照) が，ここでは，Marroquin の画像処理の例とはかなり違う状況で定量的な考察を行う．
- 期待値の計算に応用した場合，ある程度の組合せ的複雑さがあっても有効なこと (ここでいう “組合せ的複雑さ” の内容については，本文中で 2 グループに分類する場合と 3 グループに分類する場合の比較などを通じてあきらかにする予定である．)．

などである．

簡単な一般論を付録につけたので，メトロポリス的モンテカルロ法およびそのベイズ統計への応用に関してなじみのない読者はそちらを先に読まれたい．この付録は著者による 2 編の総合報告 (未発表) をもとにしたものである．

2　問題の説明

本節では，本論文で扱う問題を，ベイズ統計の枠組みのなかで提示し，統計物理の問題との類似について説明する．また，モデルの性質について 2 グループに分類する場合と 3 グループに分類する場合の相違に重点を置いて考察する．

2.1　対比較の問題と組み分けモデル

デ− タとして対比較デ− タを考える．対比較デ− タとは，いくつかのものを何回も比較した結果をデ− タとしたもので，もっとも身近な例としては相撲などのリ− グ戦の対戦成績 (表 1) などがある．複数の人による食品の味比べなどを考えてもよい．

確率構造を決める出発点として，2 項分布を考える．K個の対象 (相撲なら力士) に番号 i を付け，i が jに勝った数を $\{n_{ij}\}$ とする (以下では $\forall i n_{ii}=0$ と置く)．すると，i が jに勝つ確率を $\{p_{ij}\}$ として，$\{n_{ij}\}$ で指定される勝敗のパタ− ンが実現される確率は，

$$L(n_{ij} \mid p_{ij}) = \frac{\Pi_{ij} n_{ij}}{\Pi_{i>j}(n_{ij}+n_{ji})} \Pi_{ij} p_{ij}^{n_{ij}} \tag{1}$$

$$p_{ji} = 1 - p_{ij} \tag{2}$$

となる．

これをそのまま使うと，$p_{ij}(i \neq j)$ の最尤推定値は

$$p_{ij} = \frac{n_{ij}}{(n_{ij}+n_{ji})} \tag{3}$$

という常識的な結果になる．

各 (i,j) の対戦数が 1 のオ− ダ− しかないときはパラメ− タの数が多すぎて，このままでは不安定になるので，パラメ− タを取り直したいろいろなモデルが考えられている．その中でもっとも有名なのは，いわゆる Bradley-Terry モデルで，3 すくみや苦手 / カモの関係がないという条件を付ける．

ここでは，別のパラメ− タの減らし方として，対象がいくつかのグル− プにわかれている場合を考えてみる．それぞれのグル− プ内では勝負は互角であり，グル− プ外の相手に対する勝率は自分と相手がそれぞれどのグル− プに属しているかで決まると考える．グル− プ数が 2 なら “強弱” にわけることになるが，グル− プ数が 3 以上では 3 すくみもありうる (図 1 a-b)．

このとき，連続なパラメ− タの数はグル− プ数が 2 のとき 1 個，3 のとき 3 個であるが，そのほかに分類の仕方に対応する K 個の離散パラメ− タ $\{S_i\}$ が現われる．これらは各対象がどのグル− プに属するか決めるもので，グル− プ数が 2 のとき 2 つの値，3 のときは 3 つの値のいずれかをとるものとする．

この場合の尤度関数は，

$$L(\{n_{ij}\} \mid \{S_i\}, \{p\}) = const. \cdot \Pi_{ij} f_p(S_i, S_j)^{n_{ij}} \tag{4}$$

とかける．ただし，$const.$は $\{p\}$ や $\{S_i\}$ に依存しない定数であり，f_pは以下の式で示される．

- 2 グループに分類する場合

$$S_i \in \{\pm 1\} \tag{5}$$

$$f_p(1,-1) = p \tag{6}$$

$$f_p(-1,1) = 1-p \tag{7}$$

$$f_p(1,1) = f_p(-1,-1) = 0.5 \tag{8}$$

- 3 グループに分類する場合

$$S_i \in \{1,2,3\} \tag{9}$$

$$\begin{aligned} f_p(1,2) &= 1 - f_p(2,1) = p1 && (10)\\ f_p(2,3) &= 1 - f_p(3,2) = p2 && (11)\\ f_p(3,1) &= 1 - f_p(1,3) = p3 && (12)\\ f_p(1,1) &= f_p(2,2) = f_p(3,3) = 0.5 && (13) \end{aligned}$$

今後 $\{p\}$ と書いたときは，2 グループに分類する場合は p を，3 グループに分類する場合は $\{p1, p2, p3\}$ を意味することにする．

以上では $\{p\}$ をデータから推定する場合を考えたが，$\{p\}$ が与えられていて，組み分けのみを推定する場合も考えられる．このような場合もそれなりに興味深く，決してトリビアルではない．以下では $\{p\}$ が与えられた場合を “p 既知”，$\{p\}$ をデータから推定する場合を “p 推定” とそれぞれ呼んで区別することにする．

2.2 ベイズ的解釈

このモデルをベイズ的に解釈する．そのためには，前節で定めた尤度関数の他に，各パラメータについて事前分布を与えなくてはならない．本稿では特にことわる場合を除いて，以下のような “ignorant prior ” を用いる．

- 離散パラメータ $\{S_i\}$ については，“可能な状態を等確率でとる” という事前分布を仮定する．可能な状態とは 2 グループの場合 $\{\pm 1\}$，3 グループの場合 $\{1, 2, 3\}$ である．
- 連続パラメータ $\{p\}$ については，事前分布として $(0, 1)$ の一様分布を仮定する．

各グループに属する要素の数が大きく違っている場合は，$\{S_i\}$ に関する上記の事前分布は良くない．これに対しては，たとえば，事前分布を

$$\pi(\{S_i\}) = \frac{\exp(h \sum S_i)}{\cosh^K(h)} \tag{14}$$

のように与えて，ハイパーパラメータ h を ABIC 法 (Type II 最尤法) によってデータから推定することも考えられるが，本論文ではこの方向には深入りしないことにする．

2.3 統計物理の問題との関連

(4) 式から 「エネルギー」を，

$$E_{pos} = -\sum_{ij} n_{ij} \log f_p(S_i, S_j) \tag{15}$$

と定義する．これを用いると事後分布は

$$P(\{S_i\}, \{p\}) = \frac{\exp(-E_{pos}(\{S_i\}, \{p\}))}{\sum_{\{S\},\{p\}} \exp(-E_{pos}(\{S_i\}, \{p\}))} \tag{16}$$

とギブス分布の形にかける．但し，ここで $\sum_{\{S\},\{p\}}$ は $\{S\}, \{p\}$ のあらゆる可能な組合せについての和である．グループ数が 2 のとき S_i は ± 1 の値をとる 「イジングスピン」である．またグループ数が 3 のときは S_i は (広義の) 「ポッツスピン」になる．

この場合，温度にあたるものはないが，以下では技巧的に 「温度」T を導入して変形した分布

$$P_T(\{S_i\}, \{p\}) = \frac{\exp(-\frac{E_{pos}(\{S_i\},\{p\})}{T})}{\sum_{\{S\},\{p\}} \exp(-\frac{E_{pos}(\{S_i\},\{p\})}{T})} \tag{17}$$

も考えることにする．$T = 1$ が事後分布に相当する．

デ– タ n_{ij}を 「quench された乱れ」と見なすと，p 既知の場合，(15) 式は，ランダムな結合を持つスピンモデルに対応することがわかる．E_{pos} が S_iの対の関数の和の形をしていることは，「スピングラス模型」(Mezard et al. (1986), Binder and Young (1986), Takayama (1989)) と同じであるが，この関数 $E_{pos}(S_i, S_j)$ が S_iと S_jに関して対称でないという違いがある (注 2)．次節で示すように，2 グループに分類する問題の場合は，この非対称性を S_iについて 1 次の項に押し付けてしまうことができて，モデルはランダム磁場内の (ランダム結合の) 強磁性イジング模型と同じになる．しかし，3 グループに分類する場合はそうはいかない．

p 推定の場合には J_pも確率変数である．この場合のギブス分布は，統計物理ではあまり見られない形のものになる．

2.4 2 グループに分類する場合の考察と 3 グループに分類する場合との比較

2 グループに分類する場合は，“勝ち越している要素と負け越している要素に分類する” という簡単な方法で ‘一致推定量’ が得られる．データが仮定した通りのモデルから作られたとすると，データが十分多い場合，この方法が正しい分類を与えることは明らかである．

しかし，この簡便法は最良の方法ではない．2 つのグループの要素の数が大きく異なっていて，しかも強さの差が大きい場合を考えよう．いま，強いグループに属している要素が少ないとする．すると強いグループに属している要素は弱いグループに属している要素と多く対戦するので，勝ち越すことがほぼ確実である．ところが，数の少ない側のグループ，いまの場合，弱い方のグループに属している要素を考えると，それらは自分と同じグループの要素と多く対戦するので，統計的ゆらぎによって負け越さないものが生じる可能性がある．これが誤分類の原因になる．この誤りを減らすためには，どのくらい勝ち越せば強いといえるのかをデータに基いて決める必要が出てくる．これをうまくやるところに，いまの場合の最尤推定あるいはベイズ推定の意味がある．

2 グループに分類する場合について，(15) 式を S_iを用いて具体的に書き直すと，

$$E_{pos} = C + \sum_i H_i S_i + \sum_{(ij)} J_{ij} S_i S_j \tag{18}$$

$$J_{ij} = \frac{k_{ij}}{4} \log 4p(1-p) \tag{19}$$

$$H_i = -\frac{n_i^+ - n_i^-}{4} \log \frac{p}{1-p} \tag{20}$$

$$C = -\frac{1}{4} \log \frac{p(1-p)}{4} \sum_{(ij)} k_{ij} \tag{21}$$

となる．ここで，$k_{ij} = n_{ij} + n_{ji}$は i と jの対戦数であり，$n_i^+ = \sum_j n_{ij}$, $n_i^- = \sum_j n_{ji}$ はそれぞれ要素 i の総勝ち数，総負け数である．また，$\sum_{(ij)} = 1/2 \sum_{i,j}$は要素の対に関する和である．

上式で，相互作用項$\sum_{ij} J_{ij} S_i S_j$が無視できる場合が簡便法で十分な場合に対応する．この場合，各分類の事後確率は$\sum_i H_i S_i$の部分だけで決まる．対戦数 k_{ij}が固定されている場合 $(J_{ij} \equiv J)$ を考えると，相互作用項は $J(\sum_i S_i)^2$となるが，両グループの要素数が同数に近い場合，$\sum_i S_i$ はほぼ零であり，相互作用項は消える．また，p が 0.5 に近くなると J_{ij}と H_i はともに零に近くなるが，J_{ij}のほうが速く零になるのでやはり相互作用項は重要でなくなる．これらはさきに述べた直観と一致している．さらに，どの組分けも同程度起こり易いという事前分布のもとでは，データが非常に少ない場合 (答えが分からない場合) は，$\sum_i S_i$は零に近くなるので，やはり相互作用項は無視できる．結局，簡便法ともっときちんとした扱いの差がでるのは，要素数が非常に違っていて，p が大きく，しかも，データの量が多くも少なくもないときであるといえる．

2 グループに分類する場合の簡単さは

$$n_{ij} \log f_p(S_i, S_j) + n_{ji} \log f_p(S_j, S_i) = c_{ij} + h_i(S_i) + h_j(S_j) + g_{ij}(S_i, S_j) \tag{22}$$

のように分解したとき，g_{ij}が下記の $(1),(2),(3)$ の条件を満たすようにとれる事実に表われている．

1. 対称性　$g_{ij}(S_i, S_j) = g_{ij}(S_j, S_i)$

2. 「強磁性」$g_{ij}(S_i, S_i) \geq g_{ij}(S_i, S_j)$
   これは，$4p(1-p) \leq 1$ より，$J_{ij} \geq 0$ となることからわかる．

3. 各要素の間の対戦数 $k_{ij} = n_{ij} + n_{ji}$が (i, j) に依存せず一定なら，関数 g_{ij}も (i, j) に依存しない．i に依存するのは $\{S_i\}$ について 1 次の項 $h_i(S_i)$ だけである．

これに対して，3 グループに分類する場合には一般にはこれらの条件のいずれも満たされないことが示せる (注 3)．2 グループに分類する場合の簡単さは，2 グループに分類するモデルが Bradley-Terry モデルにさらに制限を加えたものになっていることからも予想できることである．これに対

して，3 グループに分類するモデルは，‘3 すくみ’ で象徴されるような苦手 / カモの関係を含んでいるため，本質的にもっと複雑になるわけである．

3　メトロポリス法の適用と生成されたマルコフ過程のダイナミクス

(15) 式を 「エネルギー」として，メトロポリス的モンテカルロ法を適用することにより，分布 (17) での期待値を計算することができる．以下では，狭義のメトロポリス法のアルゴリズムを用いたが，その具体的な詳細は，注 4 に譲ることにする．付録の一般論も必要に応じて参照されたい．

本節では，いま論じている問題にメトロポリス的モンテカルロ法を適用した際に生じるさまざまな現象について述べる．これらの現象の多くは統計物理で知られているものであるが，このような具体的な情報処理の問題でそれがどのように現われるかを調べることは意義深いと思われる．

具体的には，まず，対称性からくる 「縮退」の効果について述べ，次に局所的極値とマルコフ過程のダイナミクスの関係について論じる．

本節及び次節では実際に 2 つないし 3 つのグループに分かれているとして作成した模擬データを主として用いる．模擬データの作り方の詳細は注 5 を参照されたい (対戦数が端数になっているのは，確率的に対戦させているからである．詳細は注 5．)．

3.1　簡単な例

詳しい研究に入る前に簡単な例を示しておく．

メトロポリス的モンテカルロ法を実行すると，系の状態の列が生成され，この列 (マルコフ過程の見本) に関する 「時間平均」が事後分布での期待値になる．表 1 のデータを “2 グループに分類する” モデルで解析した場合についての例を図 2 a に示す (p 推定の場合，$T = 1$)．縦方向にメトロポリス法の MCS 数 (「時間」，上から下に流れる) がとってあり，そのときどきの $\{S_i\}$ の状態 (snapshot) が白丸と黒丸で示されているここでは，$0 < p < 1$ のかわりに $0.5 \leq p < 1$ と制限しているので，常に白丸が ‘強者’ である．p の値は左に書かれている．輪島，北の海がいつも白，若の花と貴の花は白と黒両方に変わり，他の 2 人はいつも黒のままといった様子が読みとれる．

図 2 b には 3 グループに分類する場合 (p 既知) の結果が示されている．図の見方は図 2 a と同じ，データは模擬データである．図 2 b の場合，多くの種類の状態が出現しており，目で見ても状況がよく分からないが，このような列を用いてさまざまな量の期待値を計算することが可能なわけである．

3.2　基底状態の縮退とマルコフ過程のダイナミクス

2 つのグループに分類する場合に，$0 < p < 1$ とすると，2 つの組の要素をそっくり入れ換え，かつ，p を $1-p$ に置き換えた状態も等しい事後確率を持つことになる (「縮退」). この条件のもとでメトロポリス法を適用すると，「温度」Tが大きいときには，$0 < p < 0.5$ と $0.5 < p < 1$ の両方の状態が出現しうるが，Tが小さくなると (問題の条件が良いときは) どちらかの半分への閉じ込めが自然に起こって，残りの半分の状態は計算時間内では出現しなくなる. 有限系に対してはメトロポリス法の収束は理論的には保証されているので，長い時間待てばいつかは反対側の状態が出現する筈であるが，それに要する時間が極めて長くなるのである. これは，磁性体で自発磁化を生じるのと似た機構である.

2 グループに分類する場合には $0 < p < 1$ の代わりに $0.5 \leq p < 1$ とすれば問題の縮退は消えて，結果の解釈が容易になる (実は前出の図 2 a でも $0.5 \leq p < 1$ としている.).

3 グループに分類する場合には，3 種の置き換えの数，すなわち 6 通りの「縮退」がある. 「有限温度」で起きることは 2 グループに分類する場合と本質的に同じで，メトロポリス法で生成した状態は，Tが非常に大きいときには状態空間の中をまったくでたらめに動いているが，(推定の条件の良い場合)Tが小さくなると 6 つの “正解” のどれか近くに滞在する時間が長くなり，さらにTが小さくなると 6 個のうち 1 個の近傍に閉じ込められるようになる.

この場合の縮退は 2 グループに分類する場合のように簡単に除くわけにはいかない. このような場合，置き換えについて不変な事象, たとえば，“ある要素がある要素に勝つ”，“ある要素とある要素が同じグループである” についての事後確率だけが意味を持つ. そうでない事象，たとえば “ある要素があるグループに属する” に関する事後確率は，6 つの置き換えについて平均すればつねに $1/3$ となり意味がない. $T = 0$ では，分布は 6 個 (偶然的な縮退を考えれば 6 の整数倍個) の MAP 解の上のδ分布の混合分布になるが，このときも 6 つ全部について平均するならやはり $1/3$ になる.

3 グループに分類する場合について正解とそれの 5 通りの置き換え (図 3 参照) を考え，メトロポリス法で生成した状態とそれぞれとの重なりをプロットしたのが，(図 4 a – f) である. 縦軸は重なりの大きさ，横軸はマルコフ過程の「時間」(メトロポリス法のステップ,MCS) である. (d → e → f) の順に生成された状態との重なりが最も大きいものが入れ替わっているのがわかる.

置き換えについて不変な量をプロットした場合は，図 4 のような入れ

替わりは見られない．図 5 に示したのは，各 「時刻」で定まる “対象 i が対象 jに勝つ確率”p_{ij}がを使って定義した “情報量損失”

$$inf = \sum_{ij} p_{ij}^{true} \log \frac{p_{ij}^{true}}{p_{ij}} \tag{23}$$

のプロットである．図 5 は，図 4 と対応するものであるが，図 4 で入れ替わりが生じても，こちらの値は階段状の変化を示さない．

いままで p 推定の場合を考えたが，p 既知の場合，2 グループのときは自明の場合 $(p = 0.5)$ 以外は，p 推定の場合に論じたような問題の構造に由来する縮退がないのに対し，3 グループのときはそのような縮退がありうる．たとえば，$p1^{true} = p2^{true} = p3^{true} = 0.8$ の場合には完全な 3 重の縮退がある．もっと複雑なのは，$p1 = 0.79, p2 = 0.80, p3 = 0.81$ といった場合で，完全な縮退はないが，近似的に縮退した 3 つの状態があることになる．

3.3 準安定状態とマルコフ過程のダイナミクス

以下ではまず 「温度零」の場合を調べ，次に本題の有限温度の場合を考察する．

3.3.1 温度零の場合

メトロポリス法で 「温度」Tを零とすると，S_i及び $\{p\}$ を動かすときに，E_{pos}が小さくなるか等しい場合のみ許すことになるので，一種の局所的な逐次改良法になる．以下では，この型の逐次改良法をアルゴリズム M と呼ぶ．別の方法として，$\{S_i\}$ を与えれば，$E_{pos}(\{S_i\}, \{p\})$ を最小にする $\{p\} = \{p_0\}$ が解析的に求められることを利用して，あらかじめ $\{p\}$ を消去しておく方法も考えられる．$\{p\} = \{p_0\}$ を代入した値 $E_0(\{S_i\})$ を小さくするように $\{S_i\}$ を逐次改良するわけである．これをアルゴリズム C と呼ぶことにする (注 6)．以下では，大部分の場合について，アルゴリズム M とアルゴリズム C の結果を両方示すが，これらの間に大きな差は見いだせなかった．

状態空間の構造が複雑な問題では，局所的な逐次改良法は局所的な極値につかまる可能性が高い．これに対して，状態空間の構造が簡単な場合には考えられる組合せの数 (系の状態の数) が多くても，逐次改良法は高い確率で大域的な解に収束すると思われる．以下では，おなじデ− タに対して $\{S_i\}$ の初期値 (及び途中で使う乱数) を変えて何回も解を求めることにより，いま考えている例がどちらに属するかを調べる．

なお，本論文で (あるアルゴリズムでの)“局所的極値” という場合，実際には “局所的極大” か “局所的極小” のどちらか一方のみを指すが，いちいち書き分けるのは面倒なので，単に “局所的極値” とのみ書くことにする．このようにしても混乱はないものと思う．

3.3.1.1　2 グループに分類する場合　2 グループに分類する場合について，本当に 2 グループに分かれているとして作成した模擬データを用いて作った結果を図 6 ，図 7 に示す．模擬データ 25 組について各 100 回ずつ求めた解を示す (アルゴリズム C．アルゴリズム M による結果も定性的には大差はない．)．それぞれ上から，正解との重なり，p の推定値，得られた$-E_{pos}$を示す．正解との重なりとは，正解と一致する S_i の割合であるが，ここでは，縮退の存在を考慮して “好意的に解釈” したものを用いている (詳しくは注 7 参照)．同じ模擬データに関する結果がひとつの縦の点線上に示されている．

それぞれの縦線上にある丸の数が 100 回の試行で検出された局所的極値の数 (各々の測定量で区別できるもの) を表わしている．M と C どちらのアルゴルズムの場合も，解は必ずしも一意ではないが，ことなる解の数はさほど多くない．とくに，$-E_{pos}$の値が大きく異なるような局所的な極値は稀である．これは，グループの大きさが同じ $(30{+}30)$ で $\{S_i\}$ 間の「相互作用」の効果が少ない場合 (図 6) だけでなく，両グループの大きさがひどく違う $(10{+}50)$ 場合 (図 7) にも当てはまる．

データの源が，実際は 2 グループでなく 3 グループからなる場合はどうだろうか．図 8 はそれぞれの要素数が 5 個,20 個,5 個の 3 グループからなっているデータ (図 9 a) に対して先に述べたのと同じ実験を行ったものである．用いたアルゴリズムは C である．図 8 の一番上の左は 3 グループのうち強い 2 つをまとめた場合 (図 9 c) との重なりを，右は 3 グループのうち弱い 2 つをまとめた場合 (図 9 b) との重なりをあらわしている (いずれも，‘好意的’ な解釈 (注 7) ではなく’ なま’ の値である．)．2 段目，3 段目は図 6 -図 7 と同じく p の推定値と得られた$-E_{pos}$ が示されている．

図 8 を見ると，模擬データによって，1 個の解しかないものもある一方で相当多くの異なる解を持つものもある．一般に，本当は 3 グループに分かれているデータを 2 グループに分類する方法としては，3 すくみでない場合，“ 3 グループのうち弱い方の 2 つを 1 つにまとめる”(図 9 b) “ 3 グループのうち強い方の 2 つを 1 つにまとめる ”(図 9 c)，“ 中途半端なところで分類する ”(図 9 d) などの可能性があるが，図 8 の例の中には，これら 3 通りのそれぞれに近い解をすべて局所的極値として持つものもある．しかし，解を多く持つ場合も，$-E_{pos}$や p のばらつきが小さいという

点では前と同じであって，後述の 3 グループに分類する場合に比べると，はるかに単純といって差し支えない．

アルゴリズム M の場合も上述のアルゴリズム C の場合と似ているが，こちらの場合は，初期値として用いる p の値によって結果がことなる場合があることに注意しなければならない (局所的極値のうち特定のものしか見られなかったりする)．

2 グループに分類するモデルが局所的極値という点からみて比較的簡単な振舞いをする理由としては，まず，(2.4) 節で説明したような意味でのモデルの簡単さがあげられる．また，事前分布として ‘どのような分け方も同程度確からしい’ という分布をとっているので，推定の条件が悪いとき 2 つのグループの大きさが同じくらいになりがちである．この場合，2 グループに分類するモデルの特徴として $\{S_i\}$ 間の 「相互作用」の効果が少なくなるので，局所的極値は出にくい．

以上は p 推定の場合を扱ったが，p 既知の場合に起きる現象についても簡単に触れておく．とくに p として，‘間違った’p を与えた場合に興味がある．実際に 2 グループに分かれているとして作成した模擬データに対して，p を模擬データの作成に使用した p よりうんと大きくすると，(全部$*$) もしくは(全部.) の解が出現する．これは，あまり実力の差がないのに無理に大きな差のある 2 群に分けようとした結果，全部同じ群になってしまったということを表わしており，自然な結果である．

この場合，与える p の値が微妙な場合は，全部同じグループに属するというの解と正解に近い解がともに局所的極値として現われる．この例を図１０ $(\mathrm{a}\text{-}\mathrm{d})$ に示す．この種の振舞いは，$\{Si\}$ の事前分布として他のタイプの事前分布 (たとえば (14)) をとった場合には p 推定の場合にも出現する可能性がある．

3.3.1.2　3 つに分類する場合　2 グループに分類する場合と同様の実験を 3 グループに分類する場合にも行った．実際に 3 グループに分かれているとして作成した模擬データ 25 組について 100 回ずつ求めた解 (p 推定) の$-E_{pos}$をプロットしたものを図１１，図１２に示した．図１１-図１２の見方は図 6-図 8 と同様である．アルゴリズムは C．

対戦数が多く非常に条件のよい図１１の場合でも，局所的極値のなかには$-E_{pos}$の値の大きく異なったものがしばしば見られる．対戦数が少ない図１２の場合には，この傾向はますます甚しい．図１２と同じデータについて $p1$ をプロットしたものを図１３に示すが，$p1$ の値も極値ごとに非常にばらついている．図１２に対応するアルゴリズム M による結果 (但し解を求めた回数は２０回) を図１４に示すが，多くの局所的極値がある

という意味では本質的な差は認められない．

以上はｐ推定の場合の結果を示したが，対応するｐ既知の場合 (正しいｐを与えた場合) にもやはり局所的極値が数多く見られる．結局，3 グループに分類する問題は，ｐ推定でもｐ既知でも，2 グループに分類する問題より遥かに組合せ的な複雑さが大きいといえる．

3.3.2　有限温度の場合

次に，"有限温度」の場合を考える．ここでは，狭義のメトロポリス法が用いられる．

3.3.2.1　2 つに分類する場合　模擬データが本当に 2 グループに分かれているとして作られたものである場合には，とくに面白いことは起きない．これに対して，3 グループに分かれているとして作ったデータをメトロポリス法で処理した場合には，2 つの分け方の間を間欠的に移動する場合がある．このような列の実例を図１５に示す．これは図 8 に用いた模擬データの 1 番目について温度 $T = 1$ で生成された列である．図 9 c(ないし図 9 d) に近いパターンの中に図 9 b に近いパターンが間欠的に出現している．なお，このような現象は対象が 3 グループに分かれている場合にいつも起きるわけではなく，模擬データを作成する条件によっては，はっきりした交替現象が見られない場合もある．

ここで，複数のパターン間を間欠的に移動するという表現をしたが，ここでパターンといっているのはひとつの状態 (グループ分けの仕方と p の値) というより，いま考えているダイナミクスのもとで，"近い" 状態の集まりである．いくつかの集まりについて，1 つの集まりのメンバがしばらく出現したかと思うと，別の集まりのメンバが出現する，ということを間欠的に繰り返すわけである．以下では，このような状況を表わすのに，"準安定状態" という用語を使って，"準安定状態の間を間欠的に移動する" と表現しよう．また，メトロポリス法をランダムな初期値からはじめたとき，ある状態のまわりにしばらく滞在したあと，別のもっと重みの大きい状態に移り，その周辺に長時間滞在する場合がある．このような場合にも準安定状態という言葉を使って "準安定状態にしばらくトラップされた (閉じ込められた)" ということにしよう．このように定義された準安定状態という言葉は，厳密な内容をもつわけではないが，ある種の状況を特徴付けるのは非常に便利な概念である．

最も普通な準安定状態は，$T = 0$ での局所的極値に対応するものである．前に説明した縮退した状態間の入れ替わりは，この特殊な例である．この場合，Tが十分大きいときは，極値は準安定状態と認識されないが，

Tが小さくなると状態間のエネルギー障壁が越えにくくなって，準安定状態として見えてくる．しかし，準安定状態という概念はもっと広い．たとえば，図１６のように，互いによく連結された状態の集まりが２つあって，その間の通路が一本しかない場合，局所的極値がなくても準安定状態が出現しうる (注 8)．局所的極値に対応しない準安定状態についてはあとでまた触れる．

準安定状態が分布に寄与する割合 (メトロポリス法での出現時間の割合) については，誤解しやすい点があるので注意しておく．ある準安定状態の平均的な$-E_{pos}$を$-\bar{E}$とするとき，その状態の寄与する割合は$\exp(-\bar{E}/T)$ではなく，これに準安定状態に含まれる状態数 (統計的多重度) を掛けたものと考えるべきである．準安定状態に含まれると見なされる状態数は大きな数になりうるので，この差は大きい．

3.3.2.2　3 つに分類する場合　組合せ的複雑さという点から興味深いのは 3 グループに分類する場合なので，3 グループに分類する場合のマルコフ鎖のダイナミクスはとくに詳しく調べる価値がある．メトロポリス法で温度Tを変えていくと，

- $T = \infty$: 完全に乱雑 (すべての状態が等確率で出現)．
- $T = 1$: 事後分布が生成される．
- $T \to 0$: MAP 解 (または局所的極値)．

と変化するわけであるが，$T = 0$ で見られたような組合せ的複雑さが$T \neq 0$でどういう形で表れるかに興味がある．

まず，p 既知の場合について例をあげる．以下の例は，

- 実際に 3 グループに分かれている．
- 各グループの大きさは10，総数 K =30．
- $p1^{true} = p2^{true} = p3^{true} = 0.8$(完全 3 すくみ)．
- 各対の対戦数 0.6．

という条件で作った模擬データの解析例から選んだ．

測定する量は，(3.2) 節でも用いた “情報量損失” と

$$inf = \sum_{ij} p_{ij}^{true} \log \frac{p_{ij}^{true}}{p_{ij}} \tag{24}$$

とエネルギー

$$-E_{pos} = \sum_{ij} n_{ij} \log f_p(S_i, S_j) \tag{25}$$

である．前者は正解を知っていないと計算できない量で，正解との距離をあらわし，正解に近いほど小さくなる．後者は正解を知らないでも計算できる量で事後確率が大きいほど大きい．

図１７は，これらの量をマルコフ鎖の「時間」(MCS) に対してプロットしたものである ($\{p\}$) は正解 (0.8,0.8,0.8) を与えている．)．同じ模擬データについての結果がさまざまな温度Tに対して示されている．この例については次のようなことがわかる．

- 前と同様，高温では準安定状態間の交替ははっきりしない (図１７ a)．
- ある程度温度を下げると，はっきりした準安定状態間の交替が見られるようになる (図１７ c)．温度を下げるにつれて，交替の平均間隔はのびてゆく (図１７ d)．

  いまの場合，交替が見られるのはinfのグラフのみである．これは２つの準安定状態の$-E_{pos}$にほとんど差がないためである．infは正解を知らないと計算できない量であるが，対応する $\{S_i\}$ のパターンの列を見ていれば，交替現象が起きていることはある程度わかる (図１８)．
- さらに温度を下げると，ついには完全に一方にトラップされる (図１７ e-f)．この場合，初期値によってどちらの状態にも行くし，閉じ込めが起こる直前まで 2 状態の両方がともに相当の割合で出現していることからも，メトロポリス法が観測時間内で緩和していないことは確実である．

いまの場合，p 既知で $p1^{true} = p2^{true} = p3^{true}$だから，問題の構造に由来する完全な 3 重の縮退があるが，測定している量inf,$-E_{pos}$は置き換えで不変な量なのでこれらの縮退した状態間の行き来は見えていないことに注意されたい．交替して出現する状態は本質的に異なったものである．この例は筆者が見たうちでもっとも見事な間欠的交替現象を示した例であって，やや極端な場合であるが，温度の効果をよく表わしている．

$-E_{pos}$が小さい準安定状態の場合，高温では簡単にエネルギー障壁を越えて抜け出せるが，低温では長時間閉じ込められるという例がしばしば見られる．この例を図１９ (a-b) に示した．

- 温度 $T = 1$ ではいくつか初期条件を変えて実験したが，どの場合もすみやかに $-E_{pos}$ の大きい状態に落ち着いた (図１９ a)．
- 温度を下げると，初期条件によっては $-E_{pos}$ の小さい状態にかなりの時間とどまるようになる (図１９ b)．これが甚だしくなれば，利用できる計算時間内には抜け出すことができなくなる．

温度のこのような効果は simulated annealing 法の基本となるものである．

実際にメトロポリス法の計算を行った時，特定の準安定状態に閉じ込められているのか，正しく緩和しているのか判定することは容易でない．とくに問題なのは，少数の初期条件からしか達することのできない準安定状態がある場合である．図１９ a の場合も，本当にすべての初期条件について速やかに緩和するという保証はないわけである．隠れた準安定状態があって，その事後分布への寄与が小さいが無視はできない場合には，その状態は甚だしく間欠的に出現することになる．緩和の問題については，あとでまた触れることにする．

p 推定の場合の例を図２０に示した．図２０でも一種の交替現象が見られる．この場合，$\{S_i\}$ のパターンの列を見ると，時間的に激しく変化する状態とあまり動かない状態とが交互に出現する (図２１ a-b)．この現象についてはまだ十分に研究を行っていないが，激しく変化する状態の方は $T = 0$ での局所的極値に関係づけるより，準安定状態としての ‘乱雑な状態’ と考えた方がよいのではないかと思われる (注 9)．

4　統計量の計算

この節では，統計量を実際に計算してみて，メトロポリス的モンテカルロ法の有用性を評価する．まず，simulated annealing 法による MAP 解の計算について調べ，次に周辺分布の計算への応用 (simulation without annealing) を検討する．

4.1　MAP 解の計算 (Simulated Annealing)

3 グループに分類する場合には，逐次改良法が多くの局所的極値にトラップされることを示した．この場合に simulated annealing 法によってどの程度状況が改善されるかを調べよう．

前に有限温度での振舞いを調べたので，Tを連続的に下げて anneal(『焼きなまし』) したときの系の挙動はその結果から予想できるが，その例をひとつあげておく．図２２ a では温度を $T = 4.0$ から $T = 0$ まで 1000MCS かけて直線的に下げている (p 推定)．温度を下げていくと，ある温度 (この場合 $T = 1$ あたり) で，突然『凍結』しはじめ，特定の組分けに収束しているのがわかる．同じ条件で，anneal を開始する温度を $T = 2.0$ にして，anneal の速さを半分にしたのが図２２ b である．『凍結』の前後の部分にかける時間が増えた結果，図２２ a に比べて時間をかけておだやかに凍結が起こっているのがわかる．

問題は simulated annealing 法のアルゴリズムとしての有効性であるが，多くの場合に，200MCS から 2000MCS の計算時間をかけた annealing で，得られる結果及びその再現性は $T = 0$ の逐次改良法に比べて著しく改善されることがわかる．しかし，2000MCS の annealing を行うのと同じ計算時間で逐次改良法が 100 回以上も実行可能であることを考えると，これは公平な比較とはいえない．そこで，以下では，各模擬データについて初期値を変えて 100 回ずつ逐次改良法を行い，その結果得られた$-E_{pos}$の組と 2000MCS かけた simulated annealing 法 (1 回だけ) によって得られた解を$-E_{pos}$を比較することにした (図２３− 図２５)．いずれも p 推定の場合である．温度の下げ方は図２２のように直線的な下げ方をした．出発する温度は適当に決めたが，図２３，図２４では何通りか変えて良いと思われるものを選んである．

図２３-図２４は図１１-図１２と同じ模擬データ 25 個についての結果である．どちらも $\{p^{true}\} = (0.8, 0.8, 0.8)$(完全３すくみ) で正解のグループの大きさが (20+20+20) の場合であるが，図２３の対戦数が 0.4 であるのに対し，図２４は対戦数が 0.2 と少なく，推定の条件が悪い．図２５は $\{p^{true}\}$ の値は同じだが，対戦数が 0.15 とさらに少なく，正解のグループ

の大きさが (10+10+40) の場合である．annealing で求められた解の正解との重なり (25 個の平均値) はそれぞれ，96 % (図２３)，77 % (図２４)，50 % (図２５) であった (注 7 の意味で “好意的な解釈” をした場合)．図２３-図２５では，ある解が逐次改良法で得られたことはわかっても，それぞれが何回得られたかは不明なので，それをヒストグラムで示したのが図２６− 図２８である．図２３− 図２５とあわせて見られたい．

これらから分かることは以下の通りである．

- 一番条件のよい図２３/図２６の場合，ほとんどの模擬データについて，1 回から 10 回程度の逐次改良法の試行で真の MAP 解と思われるもの (100 回の逐次改良法で得られた最良解) が得られてしまう．この場合でも，1 回のみの逐次改良法では非常に悪い局所的極値につかまる可能性は無視できないが，数十回の試行が意味を持つのは２５個の模擬データのうち１つか２つである．それにもかかわらず，annealing の結果は完璧ではなく３個ほどの悪い解を得ている (このうち１つは$-E_{pos}$の差が極めて小さいので図２３ではわかり難い)．このような失敗はスケジュールを最適化することで防げるかもしれないが，いずれにせよ，この例は annealing の効用を論じるには簡単過ぎるように思われる．

  ヒストグラム (図２６) を見ると，ほとんどの場合に$-E_{pos}$ の大きい側に鋭いピークがある．このピークは MAP 解のみの場合もあるが，それに近いいくつかの局所的極値からなる場合もある．それ以外に少数の悪い ($-E_{pos}$の小さい) 局所的極値がある．

- 図２４/図２７の場合は微妙であって，逐次改良法 20 回との比較なら annealing のほうが良く見えるが，100 回となると差は僅かである．(図２９に 20 回繰り返しとの比較を示した．)．図２７ではほとんどの場合にピークの位置は MAP 解の近くにはなく，逐次改良法が$-E_{pos}$の小さい局所的極値につかまる可能性が大きいことを示している．

- 図２５/図２８の場合も図２４ / 図２７と似ている．いくぶん annealing に分があるようにも見えるが，差は僅かである．図２３ / 図２６の場合とは逆に，このような例では，条件が悪すぎて，いくら努力してもまともな推定結果を出すことが難しいので，$-E_{pos}$の大きい値を追求するのは意味がないように思われる．しかし，準安定状態と$-E_{pos}$の値と推定結果の良さの関係は非常に微妙なので，このような考えは正しくないかもしれない (注 10)．

以上の結果によれば，この問題では，simulated annealing 法が単なる逐次改良法の繰り返しよりはっきり優れているとはいえないことになる．しかし，アニールする場合の出発温度，温度の変え方やメトロポリス法のアルゴリズムの詳細などさまざまな要素を変えることで結果が改良される可能性は残されている (注 11)．

4.2 周辺分布の計算 (1)

simulated annealing 法は興味深い方法であるが，ベイズ統計への応用を考えた場合は，$T \to 0$ としてしまわずに，$T = 1$ で生成した事後分布自体をもっと活用する方法がないかと考えるのが自然である．

事後分布の利用法として，まず考えられるのは “ある要素があるグループに属する確率を計算することによって分類の確かさを知る” ということである．ただ，いまの場合，“ある要素があるグループに属する” という事象はグループの名前の付け方の置換に対して不変でないから，このままでは意味をもたない．しかし，前に述べたように，2 グループに分類する場合には，p を $0.5 \leq p < 1$ に制限して縮退を除くことができる．これは，“ある要素が強い方のグループに属する” という事象の事後確率 Q_iを計算していると見られる．Q_iは $0.5 \leq p < 1$ の制限を付けた場合の S_iの期待値 $\langle S_i \rangle_{pos}^{p \geq 0.5}$を使って，

$$Q_i = \frac{1 + \langle S_i \rangle_{pos}^{p \geq 0.5}}{2}, \tag{26}$$

と書ける．

メトロポリス法で求めた $\langle S_i \rangle_{pos}^{p \geq 0.5}$の例を図３０に示す．これは，図１５に対応するものである．３つのグループをまとめて 2 グループにする仕方で所属が変わる部分については，$| Q_i |$ が 1 より小さくなっていることが分かる．このような例でメトロポリス法が利用可能な計算時間内に収束する保証は一般にはないが，いまの場合は非常にすみやかに収束した．

4.3 周辺分布の計算 (2)

３グループに分類する問題の場合，前節と同じような量を定義するのは難しいが，その他のグループの名前の付け方の置換に対して不変な量について同様の問題を考えることができる．ここではこのような量の例として、“ i が j に勝つ確率”p_{ij}を推定することを考えてみる．

ベイズの立場からは，どのような効用関数を設定するかで，用いるベ

き推定量が違ってくるが，ここでは，前にも用いた情報量損失

$$inf = \sum_{ij} p_{ij}^{true} \log \frac{p_{ij}^{true}}{p_{ij}} \tag{27}$$

を基準にすることにしよう (注 12)．

(27) のもとでの p_{ij}の最良の推定値は p_{ij}の事後分布による期待値 $\langle p_{ij} \rangle_{pos}$ である．このことは，inf を基準にした場合，MAP 解 p_{ij}^{MAP} よりも $\langle p_{ij} \rangle_{pos}$ の方が良い結果を与えることを意味している．MAP 解の場合，各状態での p_{ij}^{MAP} は $p1, p2, p3$ または $1-p1, 1-p2, 1-p3$ のどれかの値をとるが，p_{ij}の事後分布による期待値 $\langle p_{ij} \rangle_{pos}$はさまざまな値をとりうる．効用関数としての情報量損失が “自信がない” 場合には $p_{ij} \sim 0.5$ と答えるほうが有利にできていることを考えると，期待値の有利さは無視できないものになることが予想される．

以下の数値実験はこのことを確かめるとともに，$\langle p_{ij} \rangle_{pos}$を求める方法としてのメトロポリス法の有効性をテストすることを目的とする．以下では，簡単のために p 既知の場合を考える．

ここでは，単に p_{ij}^{MAP}と $\langle p_{ij} \rangle_{pos}$ を比較するだけでなく，温度を $0 \leq T \leq 1.5$ でいろいろ変えて，それぞれの Tでメトロポリス法で求めた p_{ij} の期待値 $\langle p_{ij} \rangle_T^{metro}$ を比較することにした．これらの期待値は，もし計算が時間内に十分緩和していれば，温度 Tでの p_{ij}の期待値 $\langle p_{ij} \rangle_T$となる．T が小さいところでは，時間内に緩和することは期待できないが，高い温度から出発して目的の温度までの anneal(図３１) を十分ゆっくりと行うなら，計算時間内に可能な範囲で MAP 解の “まわり” での平均をとったものが得られる．後者は $T \to 0$ で MAP 解 (のひとつ) に収束する (注 13)．

２通りの条件の模擬データそれぞれ 50 組ずつに関する結果が図３２-図３３の左側である．横軸が温度 T，縦軸がメトロポリス法で求めた期待値 $\langle p_{ij} \rangle_T^{metro}$の正解と比較した場合の情報量損失

$$Ainf = \sum_{ij} p_{ij}^{true} \log \frac{p_{ij}^{true}}{\langle p_{ij} \rangle_T^{metro}} \tag{28}$$

の 50 組の平均値である．$Ainf$の値が小さいほど $\langle p_{ij} \rangle_T^{metro}$は正解に近いことになるが，結果はほぼ予想どおりで，$Ainf$は温度 $T = 1$ の近くで最小値をとる．

$T = 1$ でのメトロポリス法の緩和については次の節で検討するが，$T = 1$ での $Ainf$の値はおおむね正しいと思ってよいようである．$T = 0$ に関しては，anneal の時間を変えた結果に加えて，2000 回の逐次改良法の結果も比較のため記入してある．$-E_{pos}$の値に差が出る回数も差の大き

さも，anneal の時間や逐次改良法の試行回数を増やすにつれて減って行き，対応する inf も差がなくなっていくので，ここで示した結果を MAP 解に対する inf と見なしてよいだろう．それ以外の温度については，低温ではアニールに使う時間，平均をとる時間への依存性が残っている場合もあるが，もともと参考のために測ったわけであるから，詳しく追求することはしない．

図３２-図３３の左側と同じ内容を平均値でなく箱型図で示したものが図３４-図３５である，最大値，第３四分位数，中央値，第１四分位数，最小値がかかれている．第３四分位数では $T = 1$ の近くのくぼみがはっきり見えるのに対し，第１四分位数ではもっと T の小さいところでくぼむことが分かる．また T が大きくなると，(inf の意味で) 良く復元できるデータとそうでないデータの差が少なくなって，散らばりが小さくなる様子も分かる．

これに関連した傾向は，図３６でもはっきり見える．図３６は，個々の模擬データについて，$T = 0.6$ と $T = 1.0$ の $Ainf$ を両軸にプロットしたものである (図３２と同じ結果を用いた)．結果的に非常によい推定結果を与える場合については T を大きくすることは損になるだけであるが，よくない場合についての損失が減るので，平均値としては $T = 1$ の周辺で最もよくなるという結果になっている．これは期待された結果である．

$Ainf$ の値が $T = 1$ で良くなるのは，もちろん p_{ij} の期待値を考えているからで，メトロポリス法で出現する個々の状態について inf を考えれば，これは $T = 0$ で最も良い値をとると考えられる (もちろん模擬データに関する平均と事後分布に関する平均 (「時間」平均) を両方考えての話である)．これを定量的に確かめるために，

$$Xinf = \sum_{ij} \langle p_{ij}^{true} \log \frac{p_{ij}^{true}}{p_{ij}} \rangle_T^{metro} \tag{29}$$

を定義する．$Ainf$ が $\{p_{ij}\}$ の期待値と正解のずれであるのに対し，$Xinf$ はメトロポリス法で作られた列からランダムに抜きだしたパターンひとつひとつの $\{p_{ij}\}$ と正解 $\{p_{ij}^{true}\}$ のずれの期待値である．

図３２-図３３の右側に $Xinf(50$ 個の模擬データについての平均値) のプロットを示したが，予想通りこれはほぼ T の単調増加関数になっている．両者とも $T = 0.4$ のところで下がるのが気になるが，これは模擬データが 50 組しかないための誤差で説明できるようである．

以上で，効用関数として情報量損失 $inf(27)$ を用いた場合，MAP 解 $(T \to 0)$ を用いるより，事後分布による期待値 $(T = 1)$ を使った方が平均的にはよいことが確かめられた．また，事後分布による期待値 $(T = 1)$ を計算するために，メトロポリス法が有効なこともわかった．

p 推定の場合についても同じ問題を考えることができるが，最大の相違点は，MAP 解として $p1, p2, p3$ のどれかがほぼ正確に 0 または 1 である解が得られる可能性があることである．この場合，分類 $\{S_i\}$ が少しでも間違っていれば，情報量損失 $Ainf$ は無限大になってしまう．これは，MAP 解が効用関数 inf に不適合であることを端的にしめす例であって，$T = 1$ での期待値では非常に起こりにくい．また，予備的な実験によれば，これらの極端な場合を除いて考えても，inf で評価する限り MAP 解より期待値がすぐれているようである．

4.4 メトロポリス法の緩和

本節では，前節の計算が温度 $T = 1$ で収束しているかどうかをチェックする．

まず考えられるチェックの方法は，初期値を変えて何回も計算して，結果が変わらないことを確認することである．この場合，本当の意味でその温度で緩和している場合には，anneal なしでも結果は初期条件によらないことが期待される．十分 anneal したとき結果が初期値によらないというだけでは不十分である．前節の問題に関して，高温から出発して目的の温度まで anneal した結果と $\{S_{ij}\}$ がランダムな初期条件から出発して anneal せずにいきなり与えられた温度で計算した結果を比較したのが図３７-図３８である (50 個のデータに関する平均値)．いきなり与えられた温度で計算した結果については，計算時間をいろいろ変えて比較してある．$T = 1$ ではよく一致していることがわかる．これに対して Tの小さい側では，結果は一致せず，計算時間内では全く緩和していないことが分かる．

５０個の平均値でなく，個々の模擬データに関する比較でも，$T = 1$ では良い一致が見られる．ただし，詳しくみると僅かな系統的ずれが見られるものもある．たとえば，図３８のうち 4000MCS 捨てて 8000MCS の平均をとった場合 (X) と 50000MCS 捨てて 150000MCS の平均をとった場合 (--) の比較を図３９に示すが，前者の $Xinf$ がやや大きいように見える．これは，マルコフ鎖の最初の平均に算入せずに捨てる部分が 4000MCS ではまだ足りないためではないかと思われる．捨てる部分の長さを増やすとはっきりしたずれはなくなるが，統計的なゆらぎが大きいため，どの程度捨てれば問題ないかを厳密に確かめるのは難しい．図３９でひとつだけ大きくずれている点は図１７と同じ模擬データに対するもので，ずれが大きいのは結果の分散がもともと大きいためであって，捨てる部分が足りないためのバイアスではないと思われる．

2 回の結果が一致しただけでは不十分であるし，結果の分散を評価す

ることもできない．そこで，初期値と乱数を変えて１０回ずつ計算を行い結果を比較することにした．すべての模擬データについてこれを行うのは大変なので，2 通りの条件のそれぞれの場合について 50 組の模擬データの中から１０組ずつ選んだものを対象とすることにした．図４０，図４１は $Ainf$ の収束を示している．期待される収束先の目安として，独立に行った $150000\mathrm{MCS}$ の計算による期待値 (ランダムな初期条件からはじめて $50000\mathrm{MCS}$ とばしたあとに $150000\mathrm{MCS}$ かけて期待値を計算したもの) を横棒で記入してある．

対戦数 0.4 の場合 (図４０)，順調に収束している様子がわかる．対戦数 0.6 の場合 (図４１) は収束しにくいものと収束しやすいものの差が顕著である．これは，要素間の 「結合」が多いと，$-E_{pos}$ の大きい状態と小さい状態の $-E_{pos}$ の差が大きくなるので，正解を得る可能性が増える一方で，運わるく正解とかけ離れた 'にせ' の解ができた場合には寿命の長い準安定状態になりやすいというふうに理解できる．しかし，散らばりの大きいものも計算時間 (MCS) を増やすと順調に散らばりが減少している．なお，図４１で一番結果の散らばりが大きいデータは，実は $3.3.2.2$節の図１７に対応するものである．2 つの顕著な準安定状態をもつことが緩和を遅くしていると理解できる．

比較のために $T = 0.2$ の場合に同じ模擬データに対して，同じ実験を行ったのが，図４２，図４３である．いずれの場合も緩和していないものがあるのが明白である．これらの場合の多くは，1 回の長い計算を行ってトレンドの有無を調べたとすると，誤って収束していると判定する可能性があり，初期値を変えて調べることの重要性が分かる．

以上の結果は，この例では $T = 1$ でのメトロポリス法が可能な計算時間内に緩和しており，前節の結果が正しいことを示唆しているが，厳密にいえば次のような問題点が残されている．本論文で行ったチェックで一番問題なのは，(1) の点であるが，一般的に重要なのは，$(2),(3)$ であり，緩和の判定を行う上で困難な問題を提起している．

1. いまの場合，$Ainf$ ，$Xinf$,$-E_{pos}$ などの量で収束の判定を行った．これはもともとこれらの値の測定に興味があり，その延長として収束の問題に取り組んだためであるが，この方法では，これらの量が非常に近い準安定状態があった場合に判定を誤ってしまう場合がある．とくに $T = 0$ で，正解に近い解の場合には inf の値はほとんど正解との重なり (整数) で決ってしまうため，この種の一致の可能性は無視できない．$T = 1$ では揺らぎによって $Xinf$,$Ainf$ は半端な値をとるので，真の値が完全に一致する心配はかえって少なくなるが，逆に統計的ゆらぎのために見分けられない場合も起こりうる．この問

題を避けるためには，十分長い計算で得られた $\{p_{ij}\}$ の期待値を仮の解として，別の初期値からはじめた計算で得られた $\{p_{ij}\}$ がそこに収束していくか否かを測定するのが望ましい．いずれにせよ，実際の場合には正解が分からないので，こちらの方法をとらざるおえない．

2. 事後分布への寄与が比較的小さい準安定状態は初期値を変えてチェックしても見落としやすいと思われ，緩和時間の上限を推定する上で，厄介な存在である．寄与の大きさが小さいが完全には無視できない程度で，緩和時間が長い準安定状態があると，その状態はマルコフ鎖の中に長い平均時間間隔で出現することになる．この例を図４４に示した ($T = 1$，図３３に用いた模擬データのひとつに対する結果)．図４４の場合，inf の大きい準安定状態がまれに出現するため，$10000\mathrm{MCS}$ 程度の平均値をとっても $Xinf$ の分布はガウス的にはならず相当裾の長い，非対称な分布になる (図４５)．この例の場合，準安定状態の緩和時間は数百 MCS から長くても $2000\mathrm{MCS}$ 程度と思われるが，原理的にはもっと長い場合もありうるわけである．

3. 万一，非常に重要な準安定状態を見逃すことがあれば，重大な誤りが起こる．これに関しては，無秩序相が準安定状態になる場合 (3.3.2.2 節及び注 9 参照) に特に注意が必要と思われる．これは，身近な物理的現象でいえば，水を摂氏零度以下にしても過冷却になって氷にならないといった場合にあたる．この場合，ランダムな初期条件 (　氷」) から出発すると，どの初期条件に関しても，ほぼ同じ結果 (　氷」) を得るが，MAP 解 (　氷」) から出発するとその近くに留まり，全く違う結果を得ることになる．この可能性を完全に除外するのは，全部の状態を調べていない以上不可能であるが，ある程度のチェックは可能である．たとえば，多数回の逐次改良法で求めた MAP 解と思われるものを初期条件とした計算を行って，ランダムな初期条件の場合と比較するということが考えられる．この方法は $T = 0$ では無秩序相が準安定状態になっていないが，$T = 1$ ではなっている場合には有効だと思われる．いま考えている 2 つの例 (p 既知) の場合に実際にこのチェックを行ったが，$Ainf$,$Xinf$ 等で見ているかぎり問題はなかった．また，正解が分かっていることを利用して，正解を初期条件として出発させるテストも行ったが，やはり特に問題はなさそうであった．しかし，一般的には，この種の問題に十分注意を払う必要があるように思われる．

$T = 1$ でのメトロポリス法はどんなデータに対しても可能な時間内に

緩和するというわけではない (注 14) . 実際， p 推定の場合や p がすべて等しくない p 既知の場合には，同じような模擬データに対してもここで扱った例よりずっと収束しにくくなることが予想される . しかし，これは組合せ的な複雑さがある程度以上ある問題を扱った以上やむ得ないことであり，むしろ，うまく計算できる場合があることの方に注目すべきであろう . この場合，緩和しているか否かの判定を誤ったのでは，メトロポリス法の利点を生かすことができず，かえって間違った結果を得ることになってしまうので，緩和の判定の問題はもっと研究されるべきである .

5　まとめ

本論文では，離散ラベルをもつ分類の問題にメトロポリス的モンテカルロ法を応用することを提案し，具体的な例として対比較に基づく分類の問題をとりあげて研究した．

はじめに，この問題は 2 グループに分類する場合と 3 グループに分類する場合で大きく違った性質を持つことを示し，以下ではこれらの 2 つを対照しつつ論じた．

‘マルコフ過程のダイナミクス’ の節では，縮退や局所的極値がある場合に有限温度でメトロポリス法のしめす振舞いを調べた．そこで見いだされた振舞いは統計物理のモデルで見られるものと似ているが，それらがこのような具体的な統計の問題で見られることは興味深い．

‘統計量の計算’ の節では，この問題におけるメトロポリス法のアルゴリズムとしての有効性を評価した．MAP 解を求める方法としての simulated annealing 法と $T = 1$ の周辺分布の計算法としてのメトロポリス法 (simulation without annealing) を両方とりあげた．まず，simulated annealing 法と逐次改良法の多数回繰り返しを比較したが，simulated annealing 法が特に優れているという結果は得られなかった．つぎに，メトロポリス法で期待値や周辺分布を計算することの利点について考察し，模擬データを用いた数値実験でこれを確認した．また，この実験では逐次改良法で相当多くの局所的な極値が見いだされる場合でも，$T = 1$ では組合せ的な複雑さが弱められて利用可能な計算時間内に計算ができる場合があることも示された．これらの結果は周辺分布の計算法としてのメトロポリス法の有効性を示しているといえる．最後の部分では，緩和しているかどうかの判定の問題を扱った．

6　謝辞

熱心に議論をして下さり，また資料を教えて頂いた ATR 視聴覚研究所の川人光男氏，岡本俊明氏，本郷節之氏，乾俊郎氏に感謝します．統計数理研究所の田辺国士教授と尾形良彦助教授には有益なご助言を頂きました．また研究の初期段階における東京大学物性研究所の鈴木淳史氏との議論は非常な励ましになりました．統計数理研究所の田村義保助教授，桂康一氏，丸山直昌助教授には計算機の使用法についていつも教えていただいています．また柳本武美教授，北川源四郎助教授をはじめとする統計数理研究所の教官の方々は統計学についての初歩的な質問に親切に答えて下さいました．この場をかりて感謝の意を表したく思います．松山香織さんには図の作成などを手伝って頂きました．

7　注

7.0.0.1　**注 1 関連研究**　神経回路網の分野では，Hinton らの (狭義の) ボルツマンマシンの ‘読みだし”(あるいは “動作”) 相が本論文でいう離散パラメータを含む分類の問題に近いといえる．ただし，統計では多くの場合に ‘問題の構造に合わせてモデルを作る’ という立場であるのに対し，神経回路網としてのボルツマンマシンは ‘万能’ モデルをめざしているという点で見方の違いがあるまた，モデル自体も本論文で論じるモデルは，(少なくとも 3 グループに分類する問題の場合には) ボルツマンマシンとはかなり違っている．

本節ではボルツマンマシンという言葉をここでは Hinton らのオリジナルのモデル (イジング型神経回路網) の意味で用いたが，広義にはメトロポリス的モンテカルロ法の工学や統計学への応用一般をさす場合もあるようである．この意味では，本論文の例ももちろんボルツマンマシンの一種ということになる．

その他，関連する話題として，A.F.J.Smith は大規模な finite mixture の問題を Gibbs Sampler で扱っていると聞いたが詳細は不明である (シンポジウムでの私信)．また，群内分散を最小化する方法によるクラスター分析に simulated annealing 法を適用した研究があるが，統計モデルという意識があるかどうかは不明である．

7.0.0.2　**注 2**　**非対称性**　これは，神経回路網でいう “シナプス結合の非対称性” とは全く違うことを注意しておく．非対称シナプスの場合は，そもそも E_{pos} にあたるものが書けない　「非平衡系」である．これに対して，ここで扱う系は，少々変わっているが　「熱平衡系」である．

7.0.0.3　**注 3**　**3 つに分類する場合の難しさの証明**　p 既知の場合に性質 (1) と (2) が成り立たない場合があることをきちんと示そう．ある $(i,j)=(k,l)$ について，$n_{k,l}=1$ かつ $n_{l,k}=0$ となったとする．$(i,j)\neq(k,l)$ については，すべて $n_{ij}=n_{ji}=0$ とすると，(22) 式より，

$$-E_{pos}=\log f_p(S_k,S_l)=c_{kl}+h_k(S_k)+h_l(S_l)+g_{kl}(S_k,S_l) \qquad (30)$$

となる．g_{ij} の対称性 $g_{ij}(S_i,S_j)=g_{ij}(S_j,S_i)$ を用いると $f^{1,2}=f_p(1,2)$ 等の略記法を用いて，

$$f^{1,2}+f^{2,3}+f^{3,1}-f^{1,1}-f^{2,2}-f^{3,3}=f^{1,3}+f^{3,2}+f^{2,1}-f^{1,1}-f^{2,2}-f^{3,3} \qquad (31)$$

が成り立たなくてはならない．しかし，たとえば，$p1 = p2 = p3 = 0.8$ の場合を考えると，

$$f^{1,2} = f^{2,3} = f^{3,1} = \log(0.8) \tag{32}$$

$$f^{1,3} = f^{3,2} = f^{2,1} = \log(0.2) \tag{33}$$

$$f^{1,1} = f^{2,2} = f^{3,3} = \log(0.5) \tag{34}$$

であるから，この場合に性質 $(1),(2)$ が成り立たないことはすぐ分かる．ポイントは (31) 式の両辺が g_{ij}のみで書けることにある．

性質 (3) が成り立たない場合があることも，ある $(i,j) = (k,l)$ について $n_{k,l} = 1$ かつ $n_{l,k} = 0$，別の $(i,j) = (\hat{k},\hat{l})$ について $n_{\hat{k},\hat{l}} = 0$ かつ $n_{\hat{l},\hat{k}} = 1$ であり，この 2 組以外の (i,j) については $n_{ij} = n_{ji} = 0$ の場合を考えて，(31) 式のに相当するものを考察することにより示すことができる．

7.0.0.4　注 4　アルゴリズムの詳細　アルゴリズムとしては，(狭義の) メトロポリス法を使用した．P 推定の場合，次の 2 つの操作を繰り返すことになる．

- **Procedure Sflip**　(系の大きさ K に比例した時間を要する)
  1. ランダムに i を選ぶ．
  2. S_i^{new}をランダムに決める．
  3. S_iを S_i^{new}に変えたときの確率の比 $r = P_{new}/P_{old}$を計算する．
  4. $(0,1)$ の一様乱数 rnd を生成し，$rnd < r$なら $S_i \to S_i^{new}$とする．

- **Procedure Pflip**　(系の大きさの 2 乗 K^2に比例した時間を要する)
  1. $(-pskip, pskip)$ の一様乱数 $\triangle p$ を用いて，$p_{new} = p + \triangle p$ とする．
  2. $p \to p_{new}$としたときの確率の比 Rを計算する．
  3. $(0,1)$ の一様乱数 rnd を用いて，$rnd < R$なら $p \to p_{new}$とする．

Procedure Pflip は 2 グループに分類する場合について書いたが，3 グループに分類する場合もほとんど同じで，$p1, p2, p3$ を順番にそれぞれ動

かしてやればよい．“跳躍限界” を与える定数 $pskip$ ははじめにきめて固定しておくが，この選び方が悪いと収束の速さがひどく遅くなるので注意が必要である．普通，“受理” される率が 50 % ぐらいになるように選ぶのが良いといわれている (大きすぎても良くない．)．p は $0 < p < 1$ を満たさなくてはならないが，これは各試行ではみだしたものを無条件で捨てることで実現できる．可能な領域の端に近いからといって，“跳躍限界” を変えることは一般に誤りであることに注意しよう．不用意にこれを行うと，詳細釣合を満たさなくなる。

これら 2 つの procedure を用いて，次のような操作を実行すればよい (K は対象の総数)．

```
for mcs=1 to MCSMAX do
begin
  for n=1 to K do
    begin
      Sflip;
    end;
  Pflip;
end;
```

P 既知の場合は，Sflip のみを用いて，

```
for mcs=1 to MCSMAX do
begin
  for n=1 to K do
    begin
      Sflip;
    end;
end;
```

となる．

外側のループを 1 回まわることを 1 MCS(Monte Carlo Step) と定義するのが自然である．いまの問題では，2 次元の画像モデルなどと異なり、各要素 S_i ががすべての (もしくは総数の一定の割合の) 要素と相互作用しているので，1 MCS に要する計算量は系の大きさの 2 乗に比例する．

実際の計算に要する時間は，要素数 30，p 既知で 3 つのグループに分ける場合，統計数理研究所の HITAC M682H で約 50 秒/10000MCS であった．これは，いろいろな量を計算する手間も含めての時間で，コーディングも計算効率を考慮しない原始的なものを用いた場合である．

7.0.0.5　注 5　模擬データの作成法　模擬デ– タを作る手順は，以下のようになる (2 種の場合で説明する) . 模擬デ– タを作るためには，(1) 要素の全数 K ，(2) 強と弱が何個ずつか，(3) 強が弱に勝つ確率 p^{true} ，(4) 各要素対の対戦数を決めなくてはならない . (1)-(3) ははじめにあたえてえておく . (4) については，次の 2 通りのやり方を試みた .

- A　(対戦数固定) 対戦数をはじめから一定に決めておく .
- B　(対戦数乱数) 任意の対の間で一定の確率で対戦が行われるとする . “対戦確率 (平均対戦数)” をはじめに与えておき，各リ– グ戦ごとにそれぞれの対が対戦をするかどうか乱数で決める .

本稿に示した結果は，図 4 -図 5 を除いてすべて方法 B で作った模擬データを用いた . 一対の対戦数が端数になっているのはこのためである . (1)-(4) が決まったら乱数を引いて各要素に “試合” をさせ，ひと通りリ– グ戦が終ったら，その結果を 1 組の模擬デ– タとする . これを繰り返し，必要な個数，たとえば 50 組の模擬デ– タを得るわけである。

7.0.0.6　注 6　逐次改良法のアルゴリズムの詳細　アルゴリズム M と C のそれぞれに，細部の違いでいくつかの種類がある .

アルゴリズム M では，T を極めて小さな値にとって，メトロポリス法のプログラムをそのまま用いた . この場合，試行の回数 (MCS) は十分大きいと思われる値に固定した . このほか，改良版として，E_{pos} が同じ場合は遷移しないことにして，終了条件のチェックをつけ加えたものも用いた (アルゴリズム M' と呼ぶことにする .) .

アルゴリズム C にも，E_0 の値が真に小さくなる場合のみに遷移がおこるもの (3 つのグループに分類する場合) と同じ E_0 でも遷移するもの (2 つのグループに分類する場合) の違いがある . 念のため，3 つのグループに分類する場合の一部を同じ E_0 でも遷移するアルゴリズムでやり直してみたが，準安定状態の数が大幅に減ることはなかった . アルゴリズム C でも，変化させようとする S_i はランダムに選んだ .

7.0.0.7　注 7 正解との重なりの “好意的な解釈”　縮退の存在を考慮して，正解との重なりは “好意的” に解釈したものを用いた．この意味は 2 つのグループに分類する場合には，正解との重なりが 0.5 より小さい場合は 1 から引いたものを正解との重なりと定義しなおすということである．また，3 つのグループに分類する場合には，図 3 の $(\mathrm{a\text{-}f})$ に相当する 6 つの “重なり” のうち最も大きいものを正解との重なりと定義する．本論文で “正解との重なり” という場合，図 4 と図 8 を除いて，“好意的” に解釈したものであるこれに対して，p の方は常に “なま” の値を示した．したがって，図 6 や図 7 でも，p の真の値 p^{true} に近い値と $1-p^{true}$ に近い値の両方が出ている．

7.0.0.8　注 8 自由エネルギー　ここでいいたいことは，統計物理ないし物理化学の用語を使えば，“「活性化エネルギー」だけでなく「活性化エントロピー」も考えなくてはならない”，“準安定状態は有限温度では「エネルギー」の局所的極値ではなく，「自由エネルギー」の局所的極値である” などと表現される．なお，本文では「エントロピー」という用語は誤解を招きやすいので意図的に避け，“多重度”，“場合の数” などで言いかえている．

7.0.0.9　注 9　準安定状態としての無秩序相　統計物理のモデル，たとえば状態数が 5 以上のポッツ模型では，無秩序相 (パラ相) が相転移の温度以下でも準安定状態として存在する $[\mathrm{Wu}(1982)]$．これは液体を冷やしてもなかなか結晶化せずに過冷却となる現象に似ている．図 2 0-図 2 1 の現象はこの種のものであるある可能性がある．

7.0.0.10　注 10 局所的極値の$-E_{pos}$の値と正解との重なりの関係　逐次改良法 (アルゴリズム C) で見出された局所的極値について$-E_{pos}$の値と正解との重なり (好意的に解釈されたもの (注 7)) を両軸にプロットしたグラフを図 4 6-図 4 8 に示す．$-E_{pos}$も正解とのも模擬データの組ごとに違う値のまわりに分布するが，図 4 6-図 4 8 ではそれぞれの模擬データについての平均値を引き去ることで重心を一致させて重ね書きしてある．図 2 3/図 2 6 に対応する場合を図 4 6 に，図 2 4/図 2 7 を図 4 7 に，図 2 5/図 2 8 を図 4 8 に示した．

図 4 6 では両者には強い相関があり，僅かに$-E_{pos}$を改善することで，ほぼ確実に正解との重なりが改善される．図 4 7 でもかなりはっきりした相関があるが，分散が大きくなる．問題の図 4 8 では分散が大きく相関がほとんど見えないが，平均でみれば$-E_{pos}$を大きくする努力に意味がある

という可能性は否定できない．これを確かめるには，膨大なデータを集めなくてはならない．

ここでは解の良さを評価するために，正解との重なり (好意的に解釈されたもの) を用いた．$\inf$ を用いなかったのは，p 推定の場合，$\mathrm{p1,p2,p3}$ のいずれかが 0 か 1 になった場合に無限大になってしまうからであるが，どのような距離で見るかによって結果が変わってくるかもしれない．

7.0.0.11　注 11 画像処理における simulated annealing 法の有効性
画像処理でも $\mathrm{simulated\ annealing}$ 法の有効性について疑問が提起されている [$\mathrm{Porteous\ et.al\ (1989)}$]．しかし，彼らの結果は，これは扱われた事前分布の特性 (線過程を含まない単純なイジング模型) によるところが大きいように思われる．

7.0.0.12　注 12 inf**の定義について**　いま用いている模擬データでは一対の対戦数は 1 か 0 なので，上の定義に使う (i,j) として対戦がある対だけをとるという立場もありうると思うが，ここではそうせずにすべての (i,j) に関する和を考えることにする．

7.0.0.13　注 13 $T=0$ **での** $Ainf$**の定義**　メトロポリス法のアルゴリズムでは，全く同じ E_{pos}の状態が隣接している場合，原理的には $T=0$ でも遷移が起こる．ここでは，便宜上 $T=10^{-6}$としているので，ある程度以上 E_{pos}が近い状態が隣接していれば，実際に遷移が起こることになる．従って，MAP 解が複数あってしかも隣接している場合には，メトロポリス法で求めた期待値 $\langle p_{ij}\rangle_{T=0}^{metro}$は p_{ij}^{MAP}に必ずしも一致しない．しかし，いまの場合はその差は無視できる程度である．

7.0.0.14　注 14 メトロポリス法が実用上緩和しないデータの例　このモデルに対しては，たとえば，$T=0.2$ にすることは，データの勝ち数，負け数をそれぞれ 5 倍にすることと等価である．逆にいえば，勝ち数，負け数をそれぞれ 5 倍にしたものをデータとして与えられたなら，$T=1$ でももとのデータの $T=0.2$ の場合と同じ収束性しか示さないことになる．このことからすぐに，任意のデータについて緩和するとはいえないことがわかる．なお，本論文のようにして作った模擬データの場合，勝ち数，負け数をそれぞれ 5 倍にして作ったデータと実際に対戦数を 5 倍にしてつくったデータの違いは，前者では，いわは雑音にあたる部分も 5 倍になっていることである．

8　付録　ベイズ統計と統計物理

8.1　大規模統計モデルと統計物理アナロジー

熱平衡統計力学とベイズ統計は，どちらも確率分布を扱うので，その意味では類似していて当然である．全く違う問題を扱っているという印象を与えることがあるとすれば，その最大の原因は，統計物理は沢山の変数を扱うための技術や概念が発達しているのに対し，いままでの統計学ではそのようなことが問題にならなかった (あるいは難しいから回避されてきた) ということである．

ところが，計算機の発達にともなって統計モデルは大規模化し，上記のギャップは埋まりつつある．とくに，

- informative prior を使用した場合
- パラメータが離散的な場合

等では，パラメータが多数あっても overfit になるとは限らないので，相当大規模なモデルが出現しうる．たとえば，画像処理ではパラメータが数万個あっても不思議ではないし，本論文で扱ったようなモデルでも，パラメータの数は容易に数十から数百になりうる．そして，実際にこれらのモデルから生ずる事後分布は統計物理で馴染みのあるモデルによって生成されたギブス分布に似ている場合が少なくない．imformative prior の場合は事前分布についても同様である．

事後分布の場合は “データ” が含まれているので，類比の対照になる統計物理のモデルも対応する要素 (たとえば非一様な外場や結合) を含まなくてはならず，そういった系についての統計物理での知見は多くない．

しかし，計算法に関していえば，統計物理で使われているさまざまな方法が非一様な系にも適用できる場合もあり，現在この方向での試みがさかんになって来ている．その代表的なものが，本論文の主題であるメトロポリス的モンテカルロ法 (Gibbs Sampler) である．また，ここでは解説しないが，平均場近似や繰り込み群，転送積分法など，他の統計物理の手法の応用も考えられはじめている (もっとも転送積分法は非線形フィルターといわれているものとほとんど同じである．)．

また，“データ” が完全にランダムな場合 (雑音ばかりの場合) に相当する場合については “ランダム系の統計物理” として物理でもかなり研究されている．たとえば，画像の大規模ベイズモデルから生ずる事後分布は，データを quenched randomness と見たとき，『ランダム磁場型」のスピンモデルに似ているし，本論文で扱った問題の事後分布は 『ランダム結

合型」ないし「スピングラス型」である．これらの知識が統計に役立つこともあるかもしれない．

歴史的にみると，統計物理と大規模モデルを用いる統計学の類似性を認識し，応用することは，いろいろなグループによって独立にはじめられたと考えられる．Hinton と Sejnowski のボルツマンマシン [Hinton and Sejnowski(1986)]，Smolensky の Harmony Theory[Smolensky(1986)], Geman らのメトロポリス的モンテカルロ法による画像処理の研究 [Geman and Geman(1984), Geman and Geman(1986), Geman and Geman(1987) Geman and McClure(1987)] は最もよく知られた例である．また，画像の texture の生成をメトロポリス的モンテカルロ法で行った研究としては，[Cross and Jain(1983)] がよく知られている．このうち，Geman らの研究に関しては，Grenander の示唆によるところが大きいらしい．マルコフ場と統計物理の関連についての初期の研究については，[Geman and Geman(1984)] の文献を参照されたい．統計数理研究所の尾形と種村は点配置の問題と液体の統計力学の類似から出発して独自の路線をすすんでいる [Ogata,Tanemura (1981,1984,1985,1989)] マルコフ場を用いた画像処理の研究 [Ogata(1990)] はその路線上にあるといえる．統計物理学者による最近の研究 (理論) としては，[Bialek and Zee(1987), Zee(1988)] がある．

筆者は 1986 年から 1987 年にかけて統計数理研究所のグループの ABIC に関する仕事を学んだ際に Geman らと似たようなことを考え，それ以来この方向の研究をしている．正確にいうと，その時すでに [Geman and Geman(1986)] を見ていたが，simulated annealing 法の単なる応用例だとしか思っておらず，読み直してはじめて意味が分かった．Hinton, Sejnowski, Smolensky などいわゆる PDP グループの主張に接したのはもっとあとである．統計物理とベイズ統計の対応については表 2 にまとめた．

8.2 メトロポリス的モンテカルロ法 (Gibbs Sampler)

この方法は，普通，モンテカルロ法とかメトロポリス法とか呼ばれているが，前者では意味が広すぎるし，後者は逆に特定のダイナミクス (狭義のメトロポリス法) を指すことがあるので，ここではメトロポリス的なモンテカルロ法と呼ぶことにする．統計学者は **Gibbs Sampler** とも呼ぶようである (この用語は物理学者には通じないと思う)．

この方法の骨子は定常分布が欲しい分布になるようなマルコフ鎖を構成して，乱数を引いてそれをシュミレートすることにあり，統計物理での経験では，ギブス分布の性質を知るための方法として極めて有効であることが知られている [Metoropolis et al. (1953), Binder(1986)]．

定常分布が状態空間 $\{\alpha\}$ の上の欲しい分布 $\{P_\alpha^\infty\}$ になるマルコフ鎖を具体的に作るためには，系全体の状態の総数が有限な系の場合，マルコフ鎖の遷移確率 $\{W(\alpha \to \beta)\}$ を詳細釣合の式

$$P_\alpha^\infty W(\alpha \to \beta) = P_\beta^\infty W(\beta \to \alpha) \tag{35}$$

をみたすようにとればよい．このとき，$W(\alpha \to \beta) \neq 0$ なる (α, β) を結んだグラフが連結であるなどの条件がみたされるようにすれば，定常分布が唯一に定まって，$\{P_\alpha^\infty\}$ に一致することが示せる．

統計物理の場合，生成したい分布 $\{P_\alpha^\infty\}$ は Gibbs 分布

$$\{P_\alpha^\infty\} = \{\frac{\exp(-E_\alpha/T)}{Z}\} \tag{36}$$

の形をしているが，$W(\alpha \to \beta)$ は P_α^∞ と P_β^∞ の比で決まるので，遷移確率の表式は計算の難しい規格化定数 (分配関数)Z を含む必要がない．

(35) 式だけからは $\{W(\alpha \to \beta)\}$ は一意に定まらない．可能な W の中から物理的に自然なものを選ぶことになるが，普通，状態を定義する変数のうち同時に変化するのは 1 つもしくは少数のみとする (局所的なダイナミクス)．最も多く使われる局所的なダイナミクスの決め方としては (狭義の) メトロポリス法と熱浴法がある．本文の例では前者が用いられている．

理論的にはマルコフ鎖が構成できれば，マルコフ鎖の緩和時間より長い間隔で抜き出したサンプルが，定常分布 $\{P_\alpha^\infty\}$ からのランダムサンプルと見なせることを利用して，任意の量の期待値，周辺分布を計算することができる．物理量 A の期待値は各時刻 t_i での値 $A(t_i)$ を用いて，

$$\langle A \rangle = lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} A(t_i) \tag{37}$$

と求められる．

実際には，マルコフ鎖の緩和時間より長い間隔でサンプルを抜き出すのは必ずしも必要でないし，得策でもない．たとえば，図４９ b のようなエネルギー曲面の場合，緩和時間は 2 つの谷底を行き来する時間で決まるが，個々の谷に関する平均はそれよりずっと短い時間で計算できる．このような場合，サンプルを抜き出す間隔を緩和時間よりずっと短くしたほうがあきらかに効率的である．逆にサンプルを短い時間間隔ででとりだす分には特に有害なことはない．また，場合によっては，厳密な意味での緩和時間の何十倍もの時間をかけなくても良い期待値が計算できることもある．たとえば，図４９ a のような (自由) エネルギー曲面の場合，初期状態を星印のところにとった場合，そのまわりの準安定状態から抜け出す

のに長い時間かかるので，厳密な意味でのマルコフ鎖の緩和時間は長くなる．しかし，星印のまわりの状態が期待値にほとんど寄与しないなら，メトロポリス的モンテカルロ法で生成した列のはじめの部分を捨てる (初期緩和の部分を捨てる) ことにより，短時間の平均で精度の良い期待値が計算できる．ただし，スピングラスのような複雑な問題では，初期緩和だと思った部分が実は意味のある部分で，ある程度長時間計算していると再現されるという可能性もあるので十分な注意が必要である．これは，図４９ a と思ったのが図４９ b だった場合に当たる．もちろん，はじめの部分を十分に捨てる習慣自体は悪いことではない．

メトロポリス的なモンテカルロ法は，いわゆる importance sampling を用いたモンテカルロ法の一種とも考えられるが，求める分布が近似的にわかっていなくても使えるという点に特長がある．高次元になってくると，よい重み関数をあらかじめ知ることは非常に難しい．

8.3 メトロポリス的モンテカルロ法のベイズ統計への応用

以下では，メトロポリス的モンテカルロ法のベイズ統計への応用について 3 つの場合に分けて述べる．本文では 2 番目の応用 (ABIC の計算) についてはとりあげていないが，原理的には 3 つのすべてが本文で扱った問題に対して応用可能である．

8.3.1 Simulation without Annealing

ベイズ統計では問題の “答” はパラメータの事後分布である．$\pi(\{x\})$ を事前分布，$L(\{y\} \mid \{x\})$ を尤度関数とするとデータ $\{y\}$ を与えたときの $\{x\}$ の事後分布は

$$P(\{x\}) = \frac{L(\{y\} \mid \{x\})\pi(\{x\})}{\sum_{\{x\}} L(\{y\} \mid \{x\})\pi(\{x\})} \tag{38}$$

で与えられるわけである ($\sum_{\{x\}}$は可能な $\{x\}$ すべてに関する和)．

この式は，分母に高次元の積分 (規格化定数) を含むため，非ガウス的な大規模モデルの場合には，計算することが非常に困難である．しかし，多くの場合に実際に必要とされるのは，事後分布の細部の特徴ではなく，特定の量の事後分布での期待値や大部分のパラメータに関して和をとった分布 (周辺分布) である．この状況は統計物理のそれと類似しており，先に述べたモデル自体の類似性も考え合わせると，統計物理の場合と全く同様の計算法を適用する．すなわち，メトロポリス的モンテカルロ法によって事後確率に比例した割合で各状態が含まれる列を作り出し，期待値や周辺分布の計算を行うわけである．この方法を，本論文で

は **simulation without annealing** と呼ぶことにする．これは，後出の **simulated annealing** を念頭に置いた筆者の造語である．

事前分布を

$$\pi(\{x\}) = \frac{\exp(-E_\pi(\{x\}))}{\sum_{\{x\}} \exp(-E_\pi(\{x\}))} \tag{39}$$

尤度関数を

$$L(\{y\} \mid \{x\}) = \frac{1}{Z_L} \exp(-E_L(\{y\}, \{x\})) \tag{40}$$

とかけば，事後分布は

$$E_{pos}(\{x\}) = E_\pi(\{x\}) + E_L(\{y\}, \{x\}) \tag{41}$$

を用いて，

$$P_{pos}(\{x\}) = \frac{\exp(-E_{pos}(\{x\}))}{\sum_{\{x\}} \exp(-E_{pos}(\{x\}))} \tag{42}$$

とかけ，統計物理 (Gibbs 分布) との類似性があからさまになる (ここで，尤度関数の $\{y\}$ に関する規格化定数 Z_L が $\{x\}$ を含まないことを仮定した．)．これを見ると事後分布の分母の規格化定数は統計物理における分配関数に対応するわけであるが，先に述べた通り，メトロポリス的モンテカルロ法での期待値の計算には，これがあらかじめ分かっている必要はない．いまの場合，「温度」T にあたるものは特に存在しないが，後出の simulated annealing 法と対比させていえば，$T = 1$ にあたるともいえる．

イジング型神経回路網 (狭義のボルツマンマシン) の問題では早くから simulation without anealing が用いられている [Hinton and Sejnowski(1986)]．また，simulation without annealing が画像処理に適用可能なことは，すでに [Geman and Geman(1984)] で示唆されている．ただし，Geman らの初期の仕事はもっぱら後述の simulated annealing 法によるものである．これに対して，Marroquin([Marroquin (1985), Marroquin et al. (1987)])．は anneal しない方法のすぐれた点を強調した．Marroquin の主張は，

- 雑音が多い場合には，“正しい画像とすべての微視的パラメータが厳密に一致する可能性” はほとんどないのに，それを最大にする (MAP 解を求める) 意義は疑わしい．正解との重なりの期待値を最大にする MPM 解 (周辺確率最大解，Maximum Posterior Marginal estimate) を simulation without annealing を用いて計算したほうがよい．
- 実際に MPM 解の方がずっと良い結果を与える例がある．
- simulated annealing より simulation without anealing のほうが速く収束するし，anneal の速度を調節する手間がいらない．

といったところにある．ただし，2番目の点に関しては，主張自体は正しいとしても，テストに用いられたパターンをメトロポリス的モンテカルロ法で生成するときに十分緩和していない疑いが強く，復元に用いられている巨視的なパラメータもデータから推定していないので，例題としては不適当な点がある．また，解の評価基準の性格や良い結果が得られる理由も本論文で議論したものとはかなり違っている．最近では [Geman and McClure(1987), Ogata(1990), York(1989), Smith(1989)] など simulation without anealing の立場をとっている (または併用している) 研究が多くなってきた．

8.3.2 ABIC(TYPE II Likelihood) の計算

メトロポリス的モンテカルロ法によって ABIC 最小化を行いハイパーパラメータを決定する試みが [Geman and McClure(1987)], [Ogata(1990)] らによって試みられている．これらの方法は本質的には，simulation without annealing による期待値の計算に基づくものであって，" 完全データ " の場合にメトロポリス的モンテカルロ法を利用して尤度を最大化するアルゴリズム ([Penttinen(1984),Ogata and Tanemura(1984,1985,1989)] を ABIC に拡張したものとも考えられる．

ABIC(TYPE II Likelihood) とは，パラメータ $\{x\}$ について和をとった尤度

$$-\frac{1}{2}ABIC = \log \sum_{\{x\}} L(\{y\} \mid \{x\})\pi(\{x\}) \tag{43}$$

のことである．これは，前節の記号を使えば，

$$-\frac{1}{2}ABIC = \log \sum_{\{x\}} \exp(-E_{pos}) - \log \sum_{\{x\}} \exp(-E_{\pi}) - \log Z_L \tag{44}$$

と書ける．

E_π が変数 λ を含んでいるとき λ による $ABIC$ の微分は，

$$-\frac{1}{2}\frac{\partial(ABIC)}{\partial\lambda} = \frac{\sum_{\{x\}}(-\frac{dE_\pi}{d\lambda})\exp(-E_{pos})}{\sum_{\{x\}}\exp(E_{pos})} - \frac{\sum_{\{x\}}(-\frac{dE_\pi}{d\lambda})\exp(-E_\pi)}{\sum_{\{x\}}\exp(-E_\pi)} \tag{45}$$

のように $(-\frac{dE_\pi}{d\lambda})$ の事後分布での期待値と事前分布での期待値の差の形になる．ここで，λ はもともと問題に含まれているハイパーパラメータでもよいし，人為的に加えたものでもよい．また，E_L に含まれている変数による微分に関しても同様に期待値による表現が可能である．

これらを利用して，最急降下法で $ABIC$ を最小にするハイパーパラメータを求めるのが [Geman and McClure(1987)] の方法であり，λ につ

いての 1 次元積分を計算して $ABIC$の値を計算するのが [Ogata(1990)] の方法である．厳密な意味ではハイパーパラメータの推定とはいえないが，イジング型神経回路網 (狭義のボルツマンマシン) におけるシナプス荷重の決定に同じようなアルゴリズムを用いた仕事として，[Hinton and Sejnowski(1986)] がある (発表の時期はこれが最も早いのではないかと思われる)．神経回路網の用語では，(45) 式から導かれる最急降下法の式を学習方程式という (付録 注 1)．

8.3.3 simulated annealing

メトロポリス的モンテカルロ法のもうひとつの応用は simulated annealing 法によって，多くの局所的極値をもつ問題の (近似的な)MAP 解を計算することである．

事後分布が，

$$P(\{x_i\}) = \frac{\exp(-E_{pos}(\{x_i\}))}{\sum_{\{x\}} \exp(-E_{pos}(\{x_i\}))} \tag{46}$$

とかけるとしたとき，これに 「温度」Tを導入して変形し，

$$P_T(\{x_i\}) = \frac{\exp(-\frac{E_{pos}(\{x_i\})}{T})}{\sum_{\{x\}} \exp(-\frac{E_{pos}(\{x_i\})}{T})} \tag{47}$$

とする．

ここで，はじめはTを大きくして (「高温で」) メトロポリス的モンテカルロ法による simulation を実行し，ゆっくり $T \to 0$ とする．途中で準安定状態につかまらなければ $T \to 0$ で MAP 解が得られるはずである．

smulated anealing 法の直観的根拠は文字通り，物理における 「焼きなまし」にある．はじめから $T = 0$ でメトロポリス的モンテカルロ法を行えば普通の逐次改良法と同じであるが，ゆっくり冷やす (**anneal**，焼きなまし) 過程を経ている分だけ，準安定状態につかまる率が低く，また，仮につかまっても真の解に近いものが得られることを期待するわけである．

simulation without annealing(期待値の計算) が統計物理・ベイズ統計など確率分布を扱う問題に特有の手法であるのに対し，simulated anealing 法は原理的には最適化問題一般 (たとえば，LSI の設計，巡回セールスマン問題) のための手法で，ベイズ統計はその 1 適用例にすぎないことを注意しておく (一般には E_{pos}にあたるものが先に与えられているとき，(47) 式のようなギブス分布を技巧的に導入することになる．)．

simulated annealing 法は [Kirkpatrick et al. (1983)] によって導入された．Cêrný も同時期に Kirkpatrick らと独立に同様のことを考えた由で

ある．[Tsuda and Kiyono (1964)] は，わずか 2 変数の問題についてではあるが，simulated annealing 法に近い方法で方程式の大域解を求めることを論じている．また，似た発想は Watanabe の研究にもあるようである (クラスター分析への応用，[渡辺 (1978)] 参照)．simulated annealing 法を画像処理の問題に適用して有名になったのは，[Geman and Geman (1984)] であるが，同時期の独立の研究がいくつかある．筆者が読んだものとしては，[Camevali, Coletti and Patarnello (1985)] がある．ただし，モデルは単純で，ベイズ統計との関係も言及されていない．

9　付録の注

9.0.3.1　付録 注 **1**　学習方程式　狭義のボルツマンマシンについては本文の注 1 でも触れたがここで問題にしているのは，“学習” 相の方である．狭義のボルツマンマシンの学習は，統計学の言葉でいえば latent variable を持つ対数線型模型 (**log linear model**) のパラメタ推定の問題を Gibbs Sampler で扱っていることに相当する．