Computational Intelligence Seminar (8 November, 2005, Waseda University, Tokyo, Japan)

遺伝的アルゴリズムへの統計力学的アプローチ

Statistical Physics Approaches to Genetic Algorithms

鈴木譲

Joe Suzuki [1]

大阪大学大学院理学研究科

〒 560-0043 大阪府豊中市待兼山町 1-1

Abstract

We address genetic algorithms (GAs) in a different way. The GA selects individuals with high fitness values. This corresponds to decrease the temperature of the Boltzmann distribution at each generation. Then, only the individuals with the highest fitness values survive. However, this does not happen because the length L of each individual is assumed to be large. Otherwise, exhaustive search can be executed. We consider several strategies to estimate the Boltzmann distribution without consuming exponential computation of L. How to apply field/Bethe/Kikuchi approximations and factorizations of the distribution are discussed. The same problem can be solved by differentiating the Kullback-Leibler information under several constraints like as done in solving probablistic inference based on generalized belief propagation.

1 まえがき

遺伝的アルゴリズム (GA) の 1 アプローチとして、Estimation of Distribution Algorithms (EDA) という方法がある。GA の集団をベイジアンネットワーク (BN) の学習に必要な訓練例とみなす。そして、現世代の集団 (M 個の個体) で適合度の高い一部の $N(\leq M)$ 個の個体たちから、それらを発生させた分布を推定し、その分布に基づいて新たに M 個の個体をランダムに発生させる。$N \leq M$ であるので、毎回選択圧力がかかり、適合度の高い個体ばかりが集団を占めるようになる。

そもそも、EDA の estimation とは、ボルツマン分布の推定を意味する。GA で、集団内の個体の個数 M を十分に大きくし、交叉と突然変異を行わないものとしよう。このとき、個体 i の適合度が $f(i)$ であれば、世代 t での温度の逆数を β_t として、$e^{\beta_t \log f(i)} = (f(i))^{\beta_t}$ に比例した選択になり、世代 t とともに β_t を十分に大きくすると、適合度最大の個体が生き残る。ただ、個体の長さを L とすると、個体が 2 進で表現される場合、分母 $\sum_{i \in \{0,1\}^L} e^{\beta_t \log f(i)}$ の計算に L の指数時間の計算が必要となる。そこで、そのまま実現するのはやめて、集団の大きさ M を有限とし、交叉と突然変異で集団の確率的に揺らがせ、集団中の個体の適合度の比からボルツマン分布を推定している。それこそが GA の仕掛けであって、エルゴードな有限マルコフ連鎖であれば、交叉や突然変異にこだわることはないというのである。

この EDA に関して、Heinz Mühlenbein は、"The Estimation of Distributions and the Minimum Relative Entropy Principle" (to appear in Evolutionary Computation 2005) を含む一連の論文で、重要な視点を提示した。まず、GA の場合、個体の長さ L が大きいと仮定する。L が小さければ、GA など用いずと

[1] E-mail: suzuki@math.sci.osaka-u.ac.jp

も全探索すればよい。しかし、それを頂点数 L の BN の推定問題 (L の指数時間かかる) に帰着させていたのでは、計算量的な問題に直面する。そこで、BN の推定に帰着するのではなく、ボルツマン分布を事前に因数分解して、各因子に含まれる変数の個数を制限して、分布の推定および乱数の発生で計算量を削減すべきであるとした。

第 2 に、ボルツマン分布を推定する際に、統計物理における平均場近似、Bethe 近似などの方法が概念が適用できるとし、ベイジアンネットワーク (BN) の推論で用いられているような変分方程式を解く方法が、GA の問題に適用できる可能性があることを示した。逆に、変分方程式で解かれていた問題が GA でも解けるのではないかという期待もある。

本稿の構成は以下のようになる。2. では、伝統的な GA を概観し、公理論的な意味での GA を定義する。3. では、伝統的な GA には属さないが、公理論的な GA に属すものとして、“まずボルツマン分布ありきとしての GA”を定義する。4. では、EDA を解くための基本となる BN の構造推定に関して概観する。5. では、BN の構造が平均場近似の形であたえらる場合の EDA の性質をみる。6. では、BN の構造が平均場近似の形であたえらる場合の EDA の性質と、BN の構造が未知の場合に適用される Chow-Liu アルゴリズムおよびその一般化をみる。7. では、Heinz Mühlenbein によって提案された、ボルツマン分布の因数分解を、GA の適合度が加法的に分解可能な形で与えられた場合とグラフィカルモデルが与えられた場合、について説明する。また、最大エントロピー法との関係を述べる。8. では、既存の統計物理的方法を、Mühlenbein のボルツマン分布の因数分解の概念にしたがって、一般的な形で整理しなおす。

なお、2.-6. では Kullback-Leibler 情報量として $D(p||q)$ という形、7.-8. では Kullback-Leibler 情報量として $D(q||p)$ という形のものを用いる。p はボルツマン分布、q はそれを推定なり近似した分布である。統計的推測や情報理論では $D(p||q)$ が、統計物理では $D(q||p)$ が用いられているようである。

2 公理論的な GA

各個体が長さ L の 2 進列で表現されるものとする。$k = 1, 2, \cdots, L$ 番目の遺伝子座が $\alpha_k (\geq 2)$ 通りの値をとるような一般的な個体表現も可能である。しかし、本稿の議論では、そのような一般化をしても同じ結論が得られるので、簡単のため $\alpha_k = 2$ とする。したがって、

$$J := \{\underbrace{0\cdots0}_{L}, \underbrace{0\cdots1}_{L}, \cdots, \underbrace{1\cdots1}_{L}\}$$

が個体全体の集合となる。

また、集団はつねに M 個の個体を含むものとし、適合度関数を $f : J \to \mathbf{R}^+$ として、適合度 $f(i)$ の大きな個体 $i \in J$ を含む集団の生成にむけて世代交代を繰り返すものとする。

2.1 伝統的な GA

以下、$c[i]$ で、集団中の個体 $i \in J$ の頻度をあらわす。また、等しい長さの 2 進列 2 個 x, y に対して、$x \oplus y$ および $x \otimes y$ で、それぞれビットごとに排他的論理和および論理積をとって得られる 2 進列をあらわす。また、$|i|$ で $i \in J$ の 1 の数、$\bar{i}$ で $i \in J$ の 0,1 を反転した 2 進列をあらわす。

(ルーレット) 選択: 集団から

$$\frac{c[i]f(i)}{\Sigma_{j\in J}c[j]f(j)}\ , \tag{1}$$

の確率で個体 $i \in J$ を選択する。

交叉: 2 個の個体 $k, j \in J$ を、ある確率で 2 個の個体 $k \otimes i \oplus \bar{i} \otimes j, j \otimes i \oplus \bar{i} \otimes k \in J$ に置き換えて $(i \in J)$、そのどちらかの個体を確率 $1/2$ で選択する。

突然変異: 各個体 $k \in J$ をある確率 μ_i で $(, i \in J)k \oplus i \in J$ に置き換える。通常は、遺伝子座の各 0,1 を反転する確率を $\mu > 0$ として、

$$\mu_i = \mu^{|i|}(1-\mu)^{L-|i|}$$

とすることが多い。

これらの遺伝的操作を用いて、現世代の集団から次世代の集団に世代交代を行う。

アルゴリズム 1 (世代交代)

1. 選択を 2 回繰り返して、個体 $i, j \in J$ を得る。
2. 個体 i, j に交叉を施して、個体 $k \in J$ を得る。
3. 個体 k に突然変異を施して、$k' \in J$ を得る。
4. 1.-3. を M 回繰り返して、M 個の個体を得る (次世代の集団を得る)。

ランダムに初期世代の集団を発生した後、世代交代を有限回繰り返す処理を遺伝的アルゴリズム (GA) とよぶ。

2.2 エルゴードな有限マルコフ連鎖

前節の GA では、ある世代から次の世代への集団間の確率的推移が有限マルコフ連鎖で表現できることがわかる。各個体の頻度を表すベクトル

$$(c[i])_{i\in J},\ \sum_{i\in J} c[i] = M,\ c[i] \geq 0$$

が同じ集団を、同じ状態にある (同値な) 集団とよび、大きさ 2^L のベクトル $(c[i])_{i\in J}$ で代表させる。例えば、$L = 2$ の場合、$J = \{00, 01, 10, 11\}$ であり、さらに $M = 3$ であれば、$(c[00], c[01], c[10], c[11])$ は、

$$(0,0,0,3), (0,0,1,2), (0,0,2,1), (0,0,3,0), (0,1,0,2), (0,1,1,1), (0,1,2,0),$$

$$(0,2,0,1), (0,2,1,0), (0,3,0,0), (1,0,0,2), (1,0,1,1), (1,0,2,0), (1,1,0,1),$$

$$(1,1,1,0), (1,2,0,0), (2,0,0,1), (2,0,1,0), (2,1,0,0), (3,0,0,0).$$

の 20 個の状態をとりうる。また、2 個の集団 $\underbrace{10, 11, 10}_{M=3}$ および $\underbrace{10, 10, 11}_{M=3}$ は、$(c[00], c[01], c[10], c[11])$ が同じ $(0,0,2,1)$ であるので同値である。

状態の集合 S は有限集合で、L, M が決まれば一意にきまる。そして、GA の適合度、交叉、突然変異などの世代交代を設定することは、状態 $s \in S$ から状態 $t \in S$ への推移確率 $Q = (Q(t|s))_{t,s \in S}$ を設定することに他ならない。

そして、推移確率 Q から $qQ = q$ の解として、定常確率 $q = (q(s))_{s \in S}$ が一意に定まる。

ここで、U^* を、適合度 $f(i)$ が最大の値をもつ同じ個体 $i \in J$ のみからなる集団 (の状態) の集合とする。また、$g_p : \mathbf{R}^+ \to \mathbf{R}^+$ を $p \in \mathbf{R}^+$ (選択圧力とよぶ) および $0 < a < b$ に関して、

1. $g_p(a) < g_p(b),\ p > 0$
2. $\dfrac{g_p(b)}{g_p(a)} \to \infty,\ p \to \infty$

となる任意の関数とする。たとえば、$g_p(x) = x^p$ は、上記 2 条件を満足している。そして、(1) を

$$\frac{c[i] g_p(f(i))}{\Sigma_{j \in J} c[j] g_p(f(j))} \ . \tag{2}$$

で置き換える。このとき、任意の交叉について、突然変異確率を μ、選択圧力を p として、

$$\lim_{p \to \infty} \lim_{\mu \to 0} q(s) \qquad \begin{cases} > 0, & s \in U^* \\ = 0, & s \in S - U^* \end{cases} \tag{3}$$

が成立する (De Silva-Suzuki, 2005)。

しかし、(3) はいわゆる GA の示すポジティブな性能の一面にすぎず、いわゆる GA の最適性を保証するものではない。逆に、GA の本来の目的が、高い確率で適合度の高い個体を集団に含ませることであれば、何も交叉や突然変異などの特定の遺伝的操作にたよる必要もないであろう。そのためのエルゴード な有限マルコフ連鎖を用意するのであれば、何でもありということになる。ここで、エルゴード 性とは、推移行列 Q の $k \geq 1$ 乗 Q^k の成分がすべて正になる性質をさす。すなわち、任意の状態から別の任意の状態に、有限回の世代交代で推移しうるという性質である。例えば、突然変異確率 μ が正であれば、この条件は満足される。

3 “まずボルツマン分布ありき”としての GA

3.1 無限の集団サイズで、ボルツマン分布の変化をみる

定数 $\Delta\beta > 0$ と、各個体 $i \in J$ に対して、

$$g(i) := \frac{1}{\Delta\beta} \log f(i) \tag{4}$$

とおくと、$f(i)$ の最大化と $g(i)$ の最大化は、同値である。ここで、集団の大きさ M が十分に大きいとし、交叉、突然変異を行わない状況を考える。各個体が世代 t で確率 $p_t(i)$ で存在しているとすると、(1) は

$$p_{t+1}(i) = \frac{p_t(i) \exp\{\Delta\beta g(i)\}}{\sum_{j \in J} p_t(j) \exp\{\Delta\beta g(j)\}} \tag{5}$$

とかける。したがって、$\beta_0 > 0$ として、$\beta_t := \beta_0 + t\Delta\beta$ とかくと、

$$p_t(i) = \frac{\exp\{\beta_t g(i)\}}{\sum_{j \in J} \exp\{\beta_t g(j)\}} \tag{6}$$

(ボルツマン分布) となるので、$t \to \infty$ で g 最大 (すなわち f 最大) の個体以外では、$p_t(i)$ の値が 0 となる。以下では、$1/\beta_t$ を世代 t における温度とよぶ。

しかしながら、$p_t(i)$ の値を計算するには、$j \in J = \{0,1\}^L$ での $\exp\{\beta_t g(j)\}$ の和を求めることが必要となる。また、GA では大前提として、L が十分大きいことを仮定している。したがって、(5) の分母の和を計算したり、その分子の各 $i \in J$ の比を計算することは、不可能である (L に対して指数的となる)。

3.2 有限の集団サイズで、ボルツマン分布の変化をみる

GA は、集団の大きさ M を有限としている。もし、交叉や突然変異がないと、マルコフ連鎖のエルゴード性が維持できなくなる。選択だけを行えば、状態間の推移が限定した範囲でしか行われないので、ボルツマン分布が推定できなくなる。

世代交代数 t とともに、β_t が大きくなり (温度が下がり)、選択圧力が大きくなる。温度を下げながら、エルゴード性を保ちながら、ボルツマン分布を推定し続け、十分に温度をさげていくと、適合度最大の個体が残る。

具体的には、各世代交代 (t 世代目) では、以下の手順をふむ。

アルゴリズム 2 (EDA (第 t 世代目))

1. M 個の個体から適合度の高い $N_t(\leq M)$ 個を選び、
2. N_t 個の個体からボルツマン分布 $p_t(i)$ を推定する
3. 推定された分布にしたがって、ランダムに個体を M 個生成する。

ここで、$N_t \leq M$ であり、適合度の高い個体のみからボルツマン分布を推定することによって、選択圧力がかかっていることに注意したい。実際に、$\Delta\beta_t$ と N_t がどう対応するかなどは、わかるすべもない。しかし、公理論的な GA の条件を満足している (エルゴードな有限マルコフ連鎖を形成している) のだから、GA である。

ボルツマン分布の推定の仕方により詳細は異なるが、これらの方法を総称して Estimation of Distribution Algorithms (EDA) とよぶ。

3.3 進化とボルツマン分布との関係

10 年以上前から、GA の論文では、John Holland の “Adaptation in Nature and Artificial Systems” というテキストを参考文献にあげて、進化の過程を模倣して適合度の高い個体を生き残らせる情報処理である、という説明をかかげているものが多い。最近の De Silva-Suzuki(2005) のように、GA のポジティブな性質を数学的に証明する論文も出てきたが、従来は、Holland のスキーマ定理のみが、GA の性能を保証する唯一の根拠となっていた。個体の L ビット中のビットパターンをスキーマとよぶ。あるスキーマをもつ個体の集合 $H(\subseteq J)$ について、世代 t での頻度が $(c_t[j])_{j\in J}$ であったとする。このとき、適合度のスキーマでの平均値

$$m(H,t) := \frac{\sum_{i\in H} c[i]f(i)}{\sum_{j\in H} c[j]} \tag{7}$$

が、適合度の個体全体での平均値 $m(J,t)$ と比較して大きいときに、そのスキーマが適合度を高める要因になり、GA は適合度の高いスキーマを発見する情報処理である (ビルディングブロック仮説) とされている。

スキーマ定理とは、具体的に

$$Em(H,t+1) \geq m(H,t)(1-\chi(H))(1-\mu)^{d(H)} \tag{8}$$

なる不等式をさす。ここで、E を世代 t が $(c[j])_{j\in J}$ であるもとでの平均、$\chi(H)$ を H のビットパターンが交叉で壊れる確率、$d(H)$ を H のビットパターンの長さとした。

世代交代ごとに選択圧力が大きくなるのだから、交叉や選択を無視すれば、$m(H,t+1) \geq m(H,t)$ が期待できる (左辺で平均 E をとれば当然正しい) ので、(8) は定性的に自明な結論といわざるを得ない。また、同じ文献で Holland は、集団サイズ M が十分に大きいとしたときに、$P(H,t) = \sum_{i\in H} p_t(i)$ が

$$\frac{dP(H,t)}{dt} = P(H,t)[m(H,t) - m(J,t)] \tag{9}$$

を満たすことが、よい探索アルゴリズムであるために必要であり、GA はこの性質を満足していると主張した。(9) は、離散値 $t = 0, 1, \cdots$ で微分する、確率過程の実現値 $m(H,t)$ と確率 $P(H,t)$ が混在するなど、数学的に正しくない表現である。しかし、交叉および選択をしない、集団サイズが十分に大きいという仮定の下で、(6) で $\beta_0 = 0, \Delta\beta = 1, \beta_t = t$ とおき、(7) で $p_t(i) = c[i]/M$ とおくと、(9) が成立する。実際、

$$\frac{dp_t(i)}{dt} = p_i(t)[g(i) - \sum_{j\in J} g(j)p_j(t)]$$

$$\frac{dP(H,t)}{dt} = \sum_{k\in J} p_k(t) \frac{\sum_{i\in J} p_t(i)\{g(i) - \sum_{j\in J} g(j)p_j(t)\}}{\sum_{j\in J} p_j(t)} = P(H,t)(m(H,t) - m(J,t))$$

が成立する。すなわち、ボルツマン分布の仮定があって初めて、(9) が成立するのである。

4 BN の学習との関係

一般的には、各世代でベイジアンネットワーク (BN) の構造とパラメータを推定することになる。

N 個の確率変数 $X^{(1)}, X^{(2)}, \cdots, X^{(L)}$ の同時確率分布が各 $(x^{(1)}, x^{(2)}, \cdots, x^{(N)}) \in \{0,1\}^L$ に対して、

$$P(X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)}, \cdots, X^{(L)} = x^{(L)})$$

で与えられるものとする。この L 個の確率変数間の条件付独立性を BN で記述する際には、L 個の確率変数間の順序 $<$ をきめる。一般性を失うことなく、

$$X^{(1)} < X^{(2)} < \cdots < X^{(L)}$$

を仮定する。$X^{(i)}$ の生起は、一般には $\{X^{(1)} = x^{(1)}, \cdots, X^{(i-1)} = x^{(i-1)}\}$ すべてではなく、$\{X^{(k)}\}_{k\in\pi^{(i)}}$, $\pi^{(i)} \subseteq \{1, 2, \cdots, i-1\}$ に依存する ($\pi^{(1)} = \{\}$)。あるいは、$\{X^{(k)}\}_{k\in\pi^{(i)}}$ が与えられたときに、$\{X^{(i)}\}$ と $\{X^{(k)}\}_{k\in\{1,2,\cdots,i-1\}\setminus\pi^{(i)}}$ が条件付独立であるといってよい。したがって、同時確率分布は

$$\prod_{i=1}^{N} P(X^{(i)} = x^{(i)} | X^{(1)} = x^{(1)}, \cdots, X^{(i-1)} = x^{(i-1)}) \quad = \quad \prod_{i=1}^{N} P(X^{(i)} = x^{(i)} | (X^{(k)} = x^{(k)})_{k\in\pi^{(i)}})$$

とかける。この $\pi = (\pi^{(1)}, \pi^{(2)}, \cdots, \pi^{(L)})$ を BN の構造という。また、各 $i = 1, 2, \cdots, L$ と $(X^{(k)} = x^{(k)})_{k\in\pi^{(i)}}$ について、

$$\sum_{x^{(i)}\in\mathcal{X}^{(i)}} P(X^{(i)} = x^{(i)} | (X^{(k)} = x^{(k)})_{k\in\pi^{(i)}}) = 1$$

を満たすような

$$\theta^{(i)} \quad = \quad (P(X^{(i)} = x^{(i)} | (X^{(k)} = x^{(k)})_{k \in \pi^{(i)}}))_{x^{(i)} \in \mathcal{X}^{(i)}, x^{(k)} \in \mathcal{X}^{(k)}, k \in \pi^{(i)}}$$

がきまる。$\theta = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(L)})$ を BN のパラメータとよぶ。確率変数間の順序が指定されれば、BN は構造 π とパラメータ θ で記述される。

推定問題は、以下のように定式化できる。未知の BN$= (\pi, \theta)$ から発生した M 個の訓練例 $x \in \{0,1\}^{LM}$

$$\begin{aligned} X^{(1)} = x_1^{(1)}, X^{(2)} = x_1^{(2)}, \cdots, X^{(L)} = x_1^{(L)} \\ X^{(1)} = x_2^{(1)}, X^{(2)} = x_2^{(2)}, \cdots, X^{(L)} = x_2^{(L)} \\ \cdots \\ X^{(1)} = x_M^{(1)}, X^{(2)} = x_M^{(2)}, \cdots, X^{(L)} = x_M^{(L)} \end{aligned}$$

から、ベイズの意味で最適な構造を推定する。詳細は省略するが、構造やパラメータの事前確率が与えられると、訓練例 x のもとで事後確率最大の構造 $\hat{\pi}$ が求まる。

ここで、注意しなければならないことは、π の候補が L とともに指数的に増えてくることである。実際、各 $j = 1, 2, \cdots, L$ で比較すべき $\pi^{(j)}$ の候補の数は、2^{j-1} 個だけあり、合計 $\sum_{j=1}^{L} 2^{j-1} = 2^L - 1$ 個の事後確率を計算しなければならない。

“伝統的な GA”、“まずボルツマン分布ありきとしての GA”のいずれであっても、個体の長さ L が十分に大きいことが仮定される。もし、L が小さいのであれば、全探索で $f(i)$ が最大の $i \in J$ を見出せばよいからである。したがって、一般的な BN の学習に帰着させることは、計算量的困難さに直面することになる。

5 GA における平均場近似

そこで、以下のような戦略を考える。個体 $i = (x^{(1)}, x^{(2)}, \cdots, x^{(L)}) \in \{0,1\}^L$ について、

$$q_t^{(k)}(x^{(k)}) = \sum_{(x^{(j)})_{j \neq k}} p_t(x^{(1)}, x^{(2)}, \cdots, x^{(L)})$$

とおき、$q_t^{(k)}(j), j = 0, 1$ を

$$\hat{q}_t^{(k)}(1) := \frac{c_1 + a_1}{M + a_1 + a_0} \tag{10}$$

で推定する。ただし、c_1 を世代 t で k 番目の遺伝子座が $x^{(k)} = 1$ の個体の数、$a_0, a_1 > 0$ は、$q_t^{(k)}(x^{(k)})$ の事前分布からきまる定数とした (M が十分に大きければ、$a = 0 = a_1 = 1/2$ などとして差し支えない)。

各世代交代で、アルゴリズム 2 の 3 段階を経ることは同じだが、2. で $k = 1, 2, \cdots, L$ ごとに (10) を推定し、それを用いて、3. で $k = 1, 2, \cdots, L$ ごとに $x^{(k)}$ を発生させる。この手順は、Univariate Marginal Distribution Algorithm (UMDA) とよばれる。$p_t(i)$ を近似 $q_t(i) = \prod_{k=1}^{L} q_t^{(k)}(x^{(k)})$ の推定値 $\hat{q}_t(i) = \prod_{k=1}^{L} \hat{q}_t^{(k)}(x^{(k)})$ で置き換えることになる。$\hat{q}_t(i)$ が真の値 $q_t(i)$ であったとしても、Kullback-Leibler 情報量

$$D(p_t || q_t) = \sum_{i \in J} q_t(i) \log \frac{q_t(i)}{p_t(i)}$$

が大きくなり、精度が心配になる。つまり、ビット間の相互作用の影響を無視しているために、適合度の高いスキーマがあっても、それを壊している可能性がある。しかし、$O(L)$ の計算時間で世代交代が完了す

る。また、$W(t)=\sum_{i\in J} q_t(i)g(i)$, $q_t^{(k)}=q_t^{(k)}(x^{(k)}), x^{(k)}=0,1$ とおくと、Wright の方程式

$$q_{t+1}^{(k)}=q_t^{(k)}+q_t^{(k)}(1-q_t^{(k)})\frac{\partial W(t)/\partial q_t^{(k)}}{W(t)} \tag{11}$$

が成立し、ダイナミクスの解析が容易に行える。

6 GA における Bethe 近似

$V=\{1,2,\cdots,L\}$ に対して、巡回経路を含まないように $E=\{\{j,k\}|j,k\in V, j\neq k\}$ を設定する。ここで、巡回経路とは、E の部分集合 E' で、E' の要素の中に V の要素がちょうど 2 回ずつ現れるものである。このとき $p_t(i)$ を

$$q_t(i)=\frac{\prod_{\{k,l\}\in E} q_t^{(k,l)}(x^{(k)},x^{(l)})}{\prod_{k\in V}[q_t^{(k)}(x^{(k)})]^{N(k)-1}} \tag{12}$$

で近似する。ここで、$N(k)$ は $k\in e$ となる $e\in E$ の個数である。すなわち、E の与え方によって、近似の仕方が異なる。この値をさらに推定値 $\hat{q}_t(i)$ で置き換えることになる。具体的に、個体 $i=(x^{(1)},x^{(2)},\cdots,x^{(L)})\in\{0,1\}^L$ について、

$$q_t^{(k,l)}(x^{(k)},x^{(l)})=\sum_{(x^{(j)})_{j\neq k,l}} p_t(x^{(1)},x^{(2)},\cdots,x^{(L)})$$

とおき、$q_t^{k,l}(j,h), j,h=0,1$ を

$$\hat{q}_t^{(k,l)}(j,h):=\frac{c_{j,h}+a_{j,h}}{M+a_{0,0}+a_{0,1}+a_{1,0}+a_{1,1}} \tag{13}$$

で推定する。ただし、$c_{j,h}$ を世代 t で k,l 番目の遺伝子座が $x^{(k)}=j, x^{(l)}=h$ の個体の数、$a_{j,h}>0$ は、$q_t^{(k,l)}(j,h)$ の事前分布からきまる定数である (M が十分に大きければ、$a_{j,h}=1/4$ などとして差し支えない)。

最初から E を決めずに、与えられた p_t から、Kullback-Leibler 情報量 $D(p_t||q_t)$ を最小にするように E を決定する方法もある。まず、$k,l\in V, k\neq l$ の相互情報量を

$$I(k,l):=\sum_{x^{(k)},x^{(l)}} q_t^{(k,l)}(x^{(k)},x^{(l)})\log\frac{q_t^{(k,l)}(x^{(k)},x^{(l)})}{q_t^{(k)}(x^{(k)})q_t^{(l)}(x^{(l)})} \tag{14}$$

で定義する。

アルゴリズム 3 (Chow-Liu アルゴリズム)　1. 各 $k,l\in V, k\neq l$ について、$I(k,l)$ を計算する。

2. $E\cup\{\{k,l\}\}$ が巡回経路を持たない $e=\{k,l\}\notin E, k\neq l$ の中で、$I(k,l)$ が最大かつ非負となるものを e^* とし、$E\cup\{e^*\}$ を E とおく。

3. そのような $\{k,l\}\notin E, k\neq l$ がなくなるまで 2. を繰り返す。

ここで、

$$D(p_t||q_t)=\sum_{k\in V} H(k)-\sum_{\{k,l\}\in E} I(k,l)$$

とできる点に注意したい。ただし、

$$H(k):=\sum_{x^{(k)}} -q_t^{(k)}(x^{(k)})\log q_t^{(k)}(x^{(k)}) \tag{15}$$

である。アルゴリズム 3 のように、greedy に探索しても、Kruscal の極大木アルゴリズムに基づいているので、p_t と (12) の q_t の Kullback-Libler 情報量 $D(p_t||q_t)$ が最小になる。アルゴリズム 3 から、$V = \{1,2,3\}$, $I(1,2) > I(2,3) > I(3,1) > 0$ であれば、$E = \{\{1,2\},\{2,3\}\}$ となる。また、$V = \{1,2,3,4\}$, $I(1,2) > I(1,3) > I(2,3) > I(1,4) > I(2,4) > I(3,4)$ であれば、$E = \{\{1,2\},\{1,3\},\{1,4\}\}$ となる。

また、真の同時分布 q_t ではなく、推定値 $\bar{q}_t$

$$\bar{q}_t^{(k,l)}(j,h) := \frac{c_{j,h}}{M} \tag{16}$$

を q_t として、アルゴリズム 3 を適用すれば、平均対数尤度が最大となる。

さらに、対数尤度すなわちサンプルの適合性だけで木を生成することで満足しない場合、

$$\sum_{i \in J} -c[i] \log \bar{q}_t(i) + \frac{k(E)}{2} d_n \tag{17}$$

を最小にする情報量基準的な方法を適用してもよい。ここで、$\{d_n\}$ は正の実数列で、たとえば MDL であれば $d_n = \log n$ となる。また、$k(E)$ は $\bar{q}_t$ を記述するための独立なパラメータの個数とした。第 1 項はサンプルのグラフへの適合性、第 2 項はグラフの複雑さをあらわす形になっている。各 $x^{(k)}$ が 0,1 の 2 値ではなく、$1,2,\cdots,\alpha_k$ のいずれかの値をとるものとする。このとき、アルゴリズム 3 の $I(k,l)$ を

$$\bar{I}(k,l) := I(k,l) - \frac{(\alpha_k - 1)(\alpha_l - 1)}{2} d_n \tag{18}$$

とすると、(17) が最小となることがわかっている (Suzuki, 1993)。しかも、(18) は負の値をとりうるので、$G = (V,E)$ は一般には木ではなく森になる。また、$\alpha_k \neq 2$ の場合、$d_n = 0$ の場合と $d_n \neq 0$ の場合で、辺の選択順序が異なってくる。

7　ボルツマン分布の因数分解

世代交代のたびに個体の長さ L の指数的な計算量を用いて構造推定を行うのは、不合理であることがわかる。また、構造に関しては事前にわかっていることが多い。そこで、適合度関数やビットの生成規則に関する事前知識を利用して、ボルツマン分布を因数分解できれば、集団内の個体の頻度からパラメータを推定できる (Factorized Distribution Algorithm, FDA)。そして、その推定されたパラメータを用いて、次世代の集団をランダムに生成できる。

その際に、

$$p_t(x^{(1)},\cdots,x^{(L)}) = \prod_{j=1}^{L} p_t(x^{(j)}|x^{(1)},\cdots,x^{(j-1)})$$

の形では、首尾よく因数分解したことにはならない。因数分解した各因子に含まれる変数の数が少なければ (L によらない定数で抑えられれば)、それらの同時確率の推定と乱数発生における計算量が少なくてすむ。しかし、各因子はスキーマに対応し、因子に含まれる変数の個数はそのスキーマの定義するビット長に相当する。長いビットのスキーマが含まれるということは、その問題が GA が扱う問題として難しいことを意味し、さらなる近似手法が必要となる。

7.1 加法的に分解可能な関数が与えられる場合

関数 $g : \{0,1\}^L \to \mathbf{R}^+$ を $I := \{1, 2, \cdots, L\}$ の部分集合たち $s_1, s_2, \cdots, s_m$ を用いて、

$$g(\mathbf{x}) = \sum_k g_k(\mathbf{x}_{s_k}) \tag{19}$$

と書く ((19) の形で書くこと自体は可能だが、$\{g_k\}, \{s_k\}$ を見出すことが難しい場合がある)。以降、記法が複雑になるので、$(x^{(j)})_{j \in s_k}$ などを $\mathbf{x}_{s_k}$ などと書くものとする。

$s_1, s_2, \cdots, s_m$ がある条件を満たすときに、前述の EDA の概念が効果的に適用できることを示す。I の部分集合の長さ m の列 $\{s_k\}_{k=1}^m$ から、I の部分集合の長さ m の列たち $\{d_k\}_{k=1}^m$(histories)、$\{b_k\}_{k=1}^m$(residuals), $\{s_k\}_{k=1}^m$(separators) を以下のように構成する。

$$d_0 = \{\},\ d_k := \cup_{j=1}^k s_j,\ b_k := s_k \backslash d_{k-1},\ c_k := s_k \cap d_{k-1}$$

そして、以下の 3 条件を満足するときに、$\{s_k\}_{k=1}^m$ は running intersection property (RIP) を満たすという。

1. $b_k \neq \{\}, k = 1, 2, \cdots, m$
2. $d_m = \{x_1, x_2, \cdots, x_m\}$
3. 各 $k = 2, 3, \cdots, m$ について、$c_k \subseteq s_j$ なる $j < k$ が存在

もし、RIP が満足されない場合、$s_1, s_2, \cdots, s_m$ の順序を変えて試みる。また、$s_1, s_2, \cdots, s_m$ の $s_{k_1}, \cdots, s_{k_l}$ を $s_{k_1} \cup \cdots \cup s_{k_l}$ におきかえる (merge) 方法もある。しかし、その場合、因数分解された因子に含まれる変数の数は多くなる。ただ、$s_1 = 1, m = 1$ とおけば RIP は自明に成立する。したがって、merge によって RIP を満足する因数分解は必ず得られる。問題は、因子に含まれる各変数の個数である。

このとき、ボルツマン分布 $p_t(\mathbf{x}) := e^{\beta_t g(\mathbf{x}))} / \sum_{\mathbf{x}'} e^{\beta_t g(\mathbf{x}')}$ は、以下のように因数分解できる。

$$p_t(\mathbf{x}) = \prod_{k=1}^m p_t(\mathbf{x}_{b_k} | \mathbf{x}_{c_k}) = \prod_{k=1}^m \frac{p_t(\mathbf{x}_{b_k}, \mathbf{x}_{c_k})}{p_t(\mathbf{x}_{c_k})} \tag{20}$$

例えば、$J_{23}, J_{31}, J_{12} \in \mathbf{R}^+$ として、

$$g(x^{(1)}, x^{(2)}, x^{(3)}) = J_{23} x^{(2)} x^{(3)} + J_{31} x^{(3)} x^{(1)} + J_{12} x^{(1)} x^{(2)}$$

であれば、$I = \{1, 2, 3\}, s_1 = \{2, 3\}, s_2 = \{3, 1\}, s_3 = \{1, 2\}$ となり、

$$d_0 = \{\}, d_1 = \{2, 3\}, d_2 = \{1, 2, 3\}, d_3 = \{1, 2, 3\}$$

$$b_1 = \{2, 3\}, b_2 = \{1\}, b_3 = \{\}$$

$$c_1 = \{\}, c_2 = \{3\} \subseteq s_1, c_3 = \{1, 2\} \not\subseteq s_1, s_2$$

となり、RIP の条件 1 と条件 3 が満足されない。s_1, s_2, s_3 の要素を入れ替えても同様である。

$$g(x^{(1)}, x^{(2)}, x^{(3)}) = J_{23} x^{(2)} x^{(3)} + J_{31} x^{(3)} x^{(1)} + J_4 x^{(4)}$$

であれば、$I = \{1, 2, 3, 4\}, s_1 = \{2, 3\}, s_2 = \{3, 1\}, s_3 = \{4\}$ となり、

$$d_0 = \{\}, d_1 = \{2, 3\}, d_2 = \{1, 2, 3\}, d_2 = \{1, 2, 3, 4\}$$

$$b_1 = \{2,3\}, b_2 = \{1\}, b_3 = \{4\}$$

$$c_1 = \{\}, c_2 = \{3\} \subseteq s_1, c_3 = \{\} \subseteq s_1$$

となり、RIP の条件すべてが満足される。したがって、

$$p_t(x^{(1)}, x^{(2)}, x^{(3)}) = p_t(x^{(2)}, x^{(3)}) p_t(x^{(1)}|x^{(3)}) p_t(x^{(4)}) = \frac{p_t(x^{(2)}, x^{(3)}) p_t(x^{(3)}, x^{(1)}) p_t(x^{(4)})}{p_t(x^{(3)})}$$

の形に因数分解される。

7.2 グラフィカルモデルが与えられる場合

変数間の依存関係が、無向のグラフィカルモデルで与えられる場合を考えよう。

確率変数間の条件付独立性を、グラフ $G = (V, E)$ の依存分離性で表現してみよう。$X, Y, U \subseteq V$ ($X \cup Y \cup Z = V, X \cap Y = Y \cap Z = Z \cap X = \{\}$) をグラフの頂点の集合と見て、$X, Y$ が Z で依存分離的であること (X の頂点から Y の頂点への経路すべてが Z の頂点を 1 個以上含むこと) を、$< X|Z|Y >_G$ と書く。また、$X, Y, Z \subseteq V$ を確率変数の集合と見て、X, Y が Z で条件付独立であること ($x \in X$ と $p(y,z) > 0$ なる $y \in Y, z \in Z$ に対して $p(x|y,z) = p(x|z)$) を、$I(X, U, Z)$ と書く。辺集合 E が $\{\{u,v\}|u,v \in V, u \neq v\}$(完全グラフ) であれば、任意の $x \in X, y \in Y$ について、Z を通らない x, y を結ぶ経路が必ず存在し (たとえば、x, y を直接結ぶ経路)、条件付独立性のチェックが不要であるので、

$$< X|Z|Y >_G \Longrightarrow I(X, Z, Y) \tag{21}$$

(I-map) は自明に成立する。辺集合 E を小さくしていき、どのような辺を削除しても (21) が成立しなくなる (極小 I-map) とき、$G = (V, E)$ を確率変数の集合 V の (無向の) グラフィカルモデル (Pearl 1988) とよぶ。無向のグラフィカルモデルをマルコフネットワーク、有向のグラフィカルモデルをベイジアンネットワークとよぶ。いずれにせよ、確率変数の条件付独立性を表示したものになる。

本節では、$V = \{x^{(1)}, \cdots, x^{(L)}\}$ として、V の無向のグラフィカルモデル $G = (V, E)$ が与えられたとする。ここでは、前節と同様の因数分解を行うために、G から以下の条件を満たす木 T(接合木, junction tree) を生成する。

1. T の頂点 (クラスタ) は、G の頂点集合 V の部分集合に対応し、G の各クリークに対して、それを含む T のクラスタが存在する。

2. T の 2 クラスタ $C_1, C_2 \subseteq V$ を結ぶ経路上の各クラスタは、$C_1 \cap C_2$ の各元を含む。

上記で、クリークとは、V の部分集合 C で、C の任意 2 頂点を結ぶ辺が E に含まれ、しかも C にさらに 1 頂点を加えた $C' \subseteq V$ に関しては、この性質がいえなくなるものをいう。(無向) グラフ G に対して、接合木は一般には複数存在する。接合木を求めるには、たとえば、以下の手順をふむ。

アルゴリズム 4 (接合木アルゴリズム)

1. G に適当な辺を加え、長さ 4 以上の弧を含まない巡回経路を除去する (三角化)。

2. G のクリーク $C_1, \cdots, C_m$ を T のクラスタとする。

3. 頂点集合を $\mathcal{V} := \{C_1, \cdots, C_m\}$、$C_1, C_2 \in \mathcal{V}$ を結ぶ辺 $C_1 \cap C_2$ の効用 $c(C_1, C_2)$ を $C_1 \cap C_2$ に含まれる要素の個数として、Kruscal の極大木アルゴリズムを適用して、接合木 T を生成する。

極大木アルゴリズムでは、有限の頂点集合 $\mathcal{V}$ と頂点間に割り当てられた効用 $c : \mathcal{V} \times \mathcal{V} \to \mathbf{R}^+$ を定義する。そして、結合してもループにならない未連結の $u, v \in \mathcal{V}$ の中で、$c(u, v)$ が正で最大となる 2 頂点を結合して辺とする。そのような条件を満たす $u, v \in \mathcal{V}$ がなくなるまでこの処理を繰り返して、辺集合 $\mathcal{E}$ を得る。そのように greedy に探索しても、木の辺に割り当てられた効用の合計が最大となる (Kruscal, 1956)。

接合木では、辺も頂点も G の頂点集合 V の部分集合となる。接合木の頂点集合 $\mathcal{V}$、辺集合 $\mathcal{E}$ に対して、

$$p_t(V) = \frac{\prod_{v \in \mathcal{V}} p_t(v)}{\prod_{e \in \mathcal{E}} p_t(e)}$$

とかける。

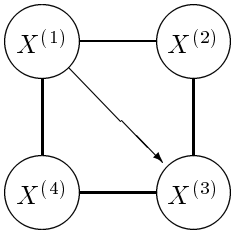

Figure 1: $x^{(1)}$ と $x^{(3)}$ を結んで三角化

たとえば、確率変数 $V = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\}$ のグラフィカルモデルとして、$G = (V, E)$ が図 1 のように与えられているものとする。$x^{(1)}, x^{(3)}$ を辺として結んで三角化すると、クリークは $C_1 = \{x^{(1)}, x^{(2)}, x^{(3)}\}, C_2 = \{x^{(1)}, x^{(3)}, x^{(4)}\}$ の 2 個となる。両者を結ぶ辺は、$C_1 \cap C_2 = \{x^{(1)}, x^{(3)}\}$ となる ($\mathcal{V} = \{C_1, C_2\}, \mathcal{E} = \{C_1 \cap C_2\}$)。したがって、因数分解は

$$p_t(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) = \frac{p_t(x^{(1)}, x^{(2)}, x^{(3)}) p_t(x^{(1)}, x^{(3)}, x^{(4)})}{p_t(x^{(1)}, x^{(3)})}$$

となる。

本稿の議論では、因数分解された各因子に含まれる変数の個数をなるべく少なくすることが目的であった。そして、接合木を求めるアルゴリズムには上記の他にも種々のものがある。ただ、最適化は非常に難しく、頂点の個数 (各 $C \in \mathcal{V}$ の要素の個数) の合計 (因数分解で得られた分母の因子に含まれる変数の合計) を最小にするクリークを求める問題は、NP 困難であることが知られている。

7.3 最大エントロピー原理との関連

$s_1, s_2, \cdots, s_m \in I$ に対して、非負の値 (同時分布とよぶ) $\tilde{q}_k(\mathbf{x}_{s_k}) = \tilde{q}_k(\mathbf{x}_{b_k}, \mathbf{x}_{c_k})$ が 2 条件

$$\sum_{\mathbf{x}_{b_k}, \mathbf{x}_{c_k}} \tilde{q}_k(\mathbf{x}_{b_k}, \mathbf{x}_{c_k}) = 1 \tag{22}$$

$$\sum_{\mathbf{x}_{b_k}} \tilde{q}_k(\mathbf{x}_{b_k}, \mathbf{x}_{c_k}) = \tilde{q}_k(\mathbf{x}_{c_k}) \tag{23}$$

を満足するとき、$\tilde{q}_1, \cdots, \tilde{q}_m$ は、一貫性がある (consistent) という。このとき、RIP の最初の条件 $b_k \neq \{\}, k = 1, \cdots, m$ のもとで、$q(\mathbf{x}) := \prod_{k=1}^m \tilde{q}_k(\mathbf{x}_{b_k} | \mathbf{x}_{c_k})$ は、$\sum q(\mathbf{x}) = 1$ を満たす。そして、

$$q_k(\mathbf{x}_{b_k} | \mathbf{x}_{c_k}) = \tilde{q}_k(\mathbf{x}_{b_k} | \mathbf{x}_{c_k}) \tag{24}$$

も成立する (因数分解が一意的)。しかし、RIP を仮定しないと、

$$q_k(\mathbf{x}_{b_k}, \mathbf{x}_{c_k}) \neq \tilde{q}_k(\mathbf{x}_{b_k}, \mathbf{x}_{c_k}) \tag{25}$$

となる。たとえば、$s_1 = \{2,3\}, s_2 = \{3,1\}, s_3 = \{1,2\}$ であれば、$b_3 = \{\}$ となるため、

$$q(\mathbf{x}) = \tilde{q}_1(\mathbf{x}_{b_1})\tilde{q}_2(\mathbf{x}_{b_2}|\mathbf{x}_{c_2})$$

となる。しかし、そのような q が、

$$\sum_{x^{(3)}} q(x^{(1)}, x^{(2)}, x^{(3)}) = \tilde{q}_3(x^{(1)}, x^{(2)})$$

を満足しているとは限らない。

$\{b_k\}, \{c_k\}$ の定義から、一貫性がある同時分布 $\tilde{q}_1, \cdots, \tilde{q}_m$ について、任意の $\mathbf{x}$ について

$$\tilde{q}_k(\mathbf{x}_{s_k}) = \sum_{(x^{(j)})_{j \in I \setminus s_k}} q(\mathbf{x}) \tag{26}$$

を満足する q が存在する。最大エントロピー原理とは、(26) を満足する q の中から、特にエントロピー

$$H(q) := -\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x}) \tag{27}$$

を最大にする q を選択すべきであるという原理である。そのような q は一意であり、以下の Iterative Proportional Fitting (IPF) とよばれる手順の収束解に一致することが知られている (Csiszar, 1975)。$q^{(0)}$ を一様分布 (確率をすべて 2^{-L}) として、各 $\mathbf{x} \in \{0,1\}^L$ について、

$$q^{(\tau+1)}(\mathbf{x}) \longleftarrow q^{(\tau)}(\mathbf{x}) \frac{q_k(\mathbf{x}_{s_k})}{\sum_{\mathbf{y}} q^{(\tau)}(\mathbf{x}_{s_k}, \mathbf{y})} \tag{28}$$

を繰り返す手順である。ここで、右辺分母の和は、$\mathbf{x}$ の $I \setminus s_k$ に対応する成分すべての和である。そして、$k = ((\tau - 1) \bmod m) + 1$ としている。

$I = \{1,2,3\}, s_1 = \{2,3\}, s_2 = \{3,1\}, s_3 = \{1,2\}, m = 3$ であれば、

$$q^{(0)}(x^{(1)}, x^{(2)}, x^{(3)}) = 2^{-3}$$

$$q^{(1)}(x^{(1)}, x^{(2)}, x^{(3)}) = \tilde{q}_1(x^{(2)}, x^{(3)})2^{-1}$$

$$q^{(2)}(x^{(1)}, x^{(2)}, x^{(3)}) = \frac{\tilde{q}_1(x^{(2)}, x^{(3)})\tilde{q}_1(x^{(3)}, x^{(1)})}{\sum_{x^{(2)}} \tilde{q}_1(x^{(2)}, x^{(3)})}$$

しかし、これが $\tilde{q}_3(x^{(1)}, x^{(2)})$ に一致する保証はなく、繰り返しにより矛盾を解消し、最後には目的の q を得る。

一方、$I = \{1,2,3\}, s_1 = \{2,3\}, s_2 = \{3,1\}, m = 2$ であれば、

$$q^{(0)}(x^{(1)}, x^{(2)}, x^{(3)}) = 2^{-3}$$

$$q^{(1)}(x^{(1)}, x^{(2)}, x^{(3)}) = \tilde{q}_1(x^{(2)}, x^{(3)})2^{-1}$$

となり、この段階で正しい q が得られ、$q^{(2)} = q^{(3)} = \cdots$ となる。

一般に、RIP が満足されていると、最初の $\tau = 0, 1, \cdots, m$ までの m ステップの繰り返しで、正しい解が得られる。すなわち、(20) は最大エントロピー原理によって得られた解にもなっていることがわかる。

8 BN の推論との関係

統計物理では、同時分布 $\{\tilde{q}_k\}$ を満足する q の中で、エントロピーを最大にするのではなく、与えられた分布 p との Kullback-Leibler 情報量 $D(q||p)$ を最小にする q を用いる方法がよく検討されている。pが一様分布であれば、最大エントロピー法になる。本稿では GA を扱っているので、$g(\mathbf{x})$ 最大化、もしくは $\log f(\mathbf{x})$ 最大化を検討する (統計物理では、マイナスの符号をつけて、エネルギー $-g(\mathbf{x})$ 最小の最適化問題を解くことになる)。以下、平均エネルギー $U(q)$ およびギブスエネルギー $G(q)$ を

$$U(q) := -\sum_{\mathbf{x}} q(\mathbf{x})g(\mathbf{x})$$

$$G(q) := U(q) - H(q)$$

で定義すると、Kullback-Leibler 情報量 $D(q||p)$ は

$$D(q||p) = U(q) - H(q) + \ln Z \tag{29}$$

となる。ただし、$Z = \sum_{\mathbf{x}} e^{g(\mathbf{x})}$ となる。$q(\mathbf{x}) = p(\mathbf{x}) = e^{g(\mathbf{x})}$ のときに限り、$D(q||p) = 0$ となり、$G(q) = -\ln Z$ が得られる。

本節では、前節で扱ってきた平均場近似、因数分解などの手法を、GA を用いずに、ベイジアンネットワーク (BN) の確率推論などで利用されている変分方程式を解く方法で解くことを試みる。すなわち、GA で行われていたような q_k の推定を経ずに、温度を一定のまま、$D(q||p)$ の極値に関する q_k のラグランジュ未定係数法として定式化し、その解 $q_k, k = 1, 2, \cdots, m$ を直接求める。以下では、時刻ごとに温度が変化することはないので、$\beta_t = 1$ とし、$p_t(\mathbf{x}), q_t(\mathbf{x})$ などを、$p(\mathbf{x}), q(\mathbf{x})$ などであらわす。

$\log Z$ は定数であるから、$U(q)$ と $H(q)$ を求めればよい。(19) は、さらに

$$f(\mathbf{x}) = \sum_{s \subseteq I} a_s \mathbf{x}_s \tag{30}$$

の形で書ける。これを、q で平均をとると、

$$U(q) = \sum_{\mathbf{x}} q(\mathbf{x}) f(\mathbf{x}) = \sum_{k=1}^{m} a_s q_k(\mathbf{x}_s) \tag{31}$$

となる。

8.1 平均場近似

各 $q_k, k = 1, 2, \cdots, m$ に含まれる変数を 1 個に制限する ($m = L$)。そして、

$$q(\mathbf{x}) = \prod_{k=1}^{m} q_k(\mathbf{x}) \tag{32}$$

とおくと、

$$H(q) = -\sum_{k=1}^{m} \sum_{x^{(k)}} q(x^{(k)}) \log q(x^{(k)}) \tag{33}$$

となる。そして、変分方程式は、

$$\frac{\partial D(q||p)}{\partial q_k} = \log \frac{q_k}{1 - q_k} + \frac{\partial U}{\partial q_k} = 0$$

より、

$$q_k = \frac{1}{1+\exp[\partial U/\partial q_k]} \tag{34}$$

とかける。$\{q_k\}$ の初期値を与えてから、(34) を繰り返して収束を待つ。

8.2 Bethe 近似、菊池近似、一般の因数分解

ここでは、$\{s_k\}$ が与えられた場合、どのような手順を踏めばよいかについて、一般論を整理する。

$q(\mathbf{x}) = \sum_{k=1}^{m} \tilde{q}_k(\mathbf{x}_{b_k}|\mathbf{x}_{c_k})$ とおくと、

$$H(q) = -\sum_{k=1}^{m} \sum_{\mathbf{x}_{b_k},\mathbf{x}_{c_k}} q_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) \log \tilde{q}_k(\mathbf{x}_{b_k}|\mathbf{x}_{c_k}) \tag{35}$$

となる。しかし、RIP が満足されないと、一般に $q_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) \neq \tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k})$ であり、

$$H(q) \approx -\sum_{k=1}^{m} \sum_{\mathbf{x}_{b_k},\mathbf{x}_{c_k}} \tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) \log \tilde{q}_k(\mathbf{x}_{b_k}|\mathbf{x}_{c_k}) \tag{36}$$

を用いざるを得ない。最終的に、

$$\begin{aligned}
\mathcal{L} \quad &= U(q) - H(q) \\
&- \sum_{k=1}^{m} \sum_{\mathbf{x}_{c_k}} \lambda_k(\mathbf{x}_{c_k})[\sum_{\mathbf{x}_{b_k}} \tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) - \tilde{q}_k(\mathbf{x}_{c_k})] - \sum_{k=1}^{m} \sum_{\mathbf{x}_{b_k}} \lambda_k(\mathbf{x}_{b_k})[\sum_{\mathbf{x}_{c_k}} \tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) - \tilde{q}_k(\mathbf{x}_{b_k})] \\
&- \sum_{k=1}^{m} \gamma_k[\sum_{\mathbf{x}_{b_k},\mathbf{x}_{c_k}} \tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}) - 1] - \sum_{k=1}^{m} \delta_k[\sum_{\mathbf{x}_{c_k}} \tilde{q}_k(\mathbf{x}_{c_k}) - 1]
\end{aligned} \tag{37}$$

を $\tilde{q}_k(\mathbf{x}_{b_k},\mathbf{x}_{c_k}), \tilde{q}_k(\mathbf{x}_{c_k})$ で偏微分して、解が求まる。

例として、BN の推論の場合に適用してみよう。そして、BN の値が観測された頂点 (確率変数) の値を $\mathbf{y}$、BN の値が観測されていない L 個の頂点 (確率変数) の値を $\mathbf{x}$ として、$\mathbf{y}$ のもとでの $\mathbf{x}$ の条件付確率を $P(\mathbf{x}|\mathbf{y})$ として、

$$g(\mathbf{x}) = -\log P(\mathbf{x}|\mathbf{y})$$

$$U(q) = -\sum_{\mathbf{x}} q(\mathbf{x}) \log P(\mathbf{x}|\mathbf{y})$$

とおく。ここで、$I = \{1,\cdots,L\}$ を観測されていない L 個の頂点とした。各 $\{s_k\} \subseteq I$ は要素をちょうど 2 個含み (Bethe 近似)、頂点を結ぶ辺に相当する。

また、RIP が満足されるのは、無向グラフとしてみたときにループを持たない場合 (singly connected) に限る。その場合、

$$q(\mathbf{x}) = \frac{\prod_{(i,j)} \tilde{q}_k(x^{(i)},x^{(j)})}{\prod_k [\tilde{q}_k(x^{(k)})]^{N(k)-1}} \tag{38}$$

とかける。ただし、分子の積は $x^{(i)}, x^{(j)}$ が辺として結ばれている場合のみかけるものとする。また、$N(k)$ を頂点 k から出ている辺の個数とした。本来、ビリーフプロパゲーション (BP) といって、特にループをもつ場合は $\tilde{q}_k$ の更新を繰り返して収束を待った。しかし、ループを持たない場合には、m ステップの更新で、正しい解が得られる。BP の詳細は多くの解説があるので省略するが、BP が上記のような変分方程式を解く問題に帰着できることは、Weiss-Freeman の Generalized Belief Propagation(2004) によって指摘されている。

Bethe 近似で RIP を満足しなくとも (singly connected でなくとも)、そのグラフィカルモデルから接合木を生成すれば、RIP を満足する $\{s_k\}$ が得られる。逆に、singly connected であれば、それに対応する接合木が存在し、RIP ははじめから満足されている。

9 まとめ

平均場近似であれ、Bethe 近似であれ、一般的な因数分解であれ、

1. GA 的方法は $\beta \to \infty$ の最適解を、BN 的方法は β 有限のボルツマン分布を近似的に求めている。
2. GA 的方法は集団を利用してサンプルを繰り返し発生させるが、BN は変分方程式を解く問題に帰着させる。

の相違はあるが、$\{0,1\}^L \to \mathbf{R}^+$ の最適化問題を、GA 的にも BN 的にも解けることがわかった。したがって、BN 的解くよりも GA 的に解いたほうが効率的であるような問題が存在するかもしれない。残念なことではあるが、いずれの方法でも、変数の個数のなるべく少ない因数分解を探す問題 (各因子の変数の個数の合計を最小にするのは NP 困難) に帰着できることがわかった。

謝辞

SMAMIP のシンポジウムでは統計力学的方法を勉強する機会を、そして今回のチュートリアル講演で発表する機会をくださった東北大学の田中和之氏に感謝します。

参考文献

1. Csiszar, I. (1975) I-divergence geometry of probability distributions and minimization problems. Annals of Probability, 3(1):146–158, 1975
2. Chandi DeSilva and Suzuki, J. (2005) On the Stationary Distribution of GAs with Positive Crossover Probability. GECCO 2005, Washington DC, June 2005.
3. Chow, C. K. Liu, C. N. (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, IT-14(3), 462-467.
4. Heinz Muhlenbein and Thilo Mahnig. (1998) FDA - A scalable evolutionary algorithm for the optimization of additively decomposed functions. in Evolutionary Computation, Volume 7, 1999, 353-376
5. Heinz Muhlenbein and Robin Hons. (2005) The Estimation of Distributions and the Minimum Relative Entropy Principle. to appear in Evolutionary Computation.
6. John Holand. (1975) Adaptation in Nature and Artificial Systems. MIT Press.

7. J. Pearl. (1988) Probabilistic Reasoning in Intelligent Systems. Palo Alto, CA: Morgan Kaufmann.

8. Suzuki, J. (1993) A Construction of Bayesian Networks from Databases Based on the MDL principle. 1993 Uncertainty in Artificial Intelligence conference: 266-273 (1993);
   Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties. IEICE Trans. on Foundamentals, E82-A: 2237-2245 (1999).

9. Suzuki, J. (1995) A Markov Chain Analysis on Simple Genetic Algorithms. IEEE Transactions on Systems, Man and Cybernatics, 25(2) pages 655-659, April 1995.

10. Suzuki, J. (1998) A further result on the Markov chain model of genetic algorithms and its application to a simulated annealing-like strategy. IEEE Transactions on Systems, Man and Cybernatics, 28(1)

11. Yedidia J.S., Freeman W.T. and Weiss Y. (2000) Generalized Belief Propagation. NIPS 2000