確率文法の推定――ベイズ法によるアプローチ

Estimation in Stochastic Grammar—Bayesian Approach

星　野　　力

要　約　近年，不確実さやルールの複雑さを持つ大量のデータを活用する手段として，例からの学習による統計的アプローチの有効性が示され注目を集めている．特に，学習対象が複雑な構造を持つ場合は，従来の統計学の枠組みにはおさまらない問題が生じ，新しい理論および手法の開発が必要となる．本稿では，学習対象が確率的な文法構造を持つ場合について，ベイズ法の有効性を学習モデルの自由エネルギーの漸近的評価を用いて理論的に示すことを目的としている．解析の結果，ある事前分布の設定によっては，変分自由エネルギーは識別可能なモデルに比べて，顕著に小さくなることが分かった．このことは，確率文法のパラメータ推定における変分ベイズ法の有用性を示唆するものである．

Abstract　Recently, the statistical approach is widely used on many fields which is faced with bulk of complex data. However, when the learning object has hierarchical structure, the conventional statistics cannot apply directly because of the mathematical structure of the models. In this paper, we consider the stochastic grammar which is the standard method for speech recognition and bioinformatics.The stochastic grammar has hierarchical structure and the model is non-identifiable. We evaluate the asymptotic free energy learning with the variational Bayes. The result shows that the variational free energy was much smaller than the identifiable models.

1. はじめに

1.1 背景

近年，WWW の爆発的な普及や POS システムの整備等，これまででは考えられなかった規模の大量データが蓄積されることが日常的に起こっている．そしてこの傾向は，ユビキタスコンピューティングやセンサーネットワークによるデータ流通の遍在化，情報の問題として定式化され始めた遺伝子技術など新しい分野の合流も含め，より加速していく方向にあることは容易に予想される．

これらの蓄積されたデータを活用する際に，データに含まれる不確実性や，ルールを具体的に記述することの困難を解消するために，例からの学習に基づく統計的な手法を用いたアプローチが広く研究されている．

本稿で対象とする確率文法は，与えられた記号列からその背後にある文法を推測する問題に対して広く用いられている統計モデルである．隠れマルコフモデルが正規文法を確率化したものであるのに対して，確率文脈自由文法はチョムスキー階層で一つ上の言語クラスにあたる文脈自由文法を確率化したものに相当する[6,11]．隠れマルコフモデルは，まず音声認識の分野で標準的な手法として確立され，最近ではロボットの状況認識や時系列のクラスタリングなど，さまざまな分野に応用を広げている．一方，自然言語や RNA の解析など，データの系列が入れ子の構造を持つ場合は，隠れマルコフモデルでは表現できないので，確率文脈自由文法が使

われる．確率文脈自由文法はデータマイニング等の工学的なシステムにおいて適用が摸索され始めた段階にある．

さらに，与えられた文字列から背後にある構造を推定する確率文法の学習は，人間における言語獲得（特に文法の獲得）のモデル化にもなっている．実際に幼児が言語を習得する過程を考えてみよう．幼児はただ外部から浴びるように例となる会話や文を与えられるだけで，外国語を習得するときのように単語の種類や文法の規則を明示的に指示されるわけではない．それでも私達は品詞の種類やその組み合わせ規則などの言語体系を自動的に習得し，言語の世界に入っていくことができるわけで，そこには言語の構造を推定する何らかのメカニズムが働いていると考えられる．このメカニズムについては，人は全く白紙の状態で生まれすべては生後の学習によるとする“タブララサ”仮説から，人の脳には生まれつき言語器官が備わっていて生後は与えられる言語に応じて細かいパラメータを調節するという“原理とパラメータ”仮説まで非常に広範囲な提案が行われており，言語学，脳科学，進化心理学などを含めて探求されている[3,14]．この問題に対して確率文法は，統計モデルとして数学的に定式化されているため，“言語を記述する文法の複雑さと例として与えられる文の数に対して，対象となる学習アルゴリズムを適用した結果，どの程度の精度で文法を推定することが可能であるか”について，理論的な基盤を提示することができる．

1.2　何が問題か

確率モデルの性能を評価する重要な指標として，汎化誤差と自由エネルギーと呼ばれる量がある．詳しくは 3 章で定義するが，直感的に汎化誤差は，与えられたデータ数のもとで，確率モデルが学習した結果，どの程度まで真の分布に近づき得るかを測る量であり，その値が小さいということは，うまく真の分布をまねできているということになる．また自由エネルギーは，与えられたデータのもとでの学習モデルの尤もらしさを表わす量と関係しており，学習モデルがデータに対して尤もらしい程小さな値をとる．文法の例で言えばデータが文法 A から生成されているか，もしくは文法 B から生成されているかを比較する文法の同定問題に使われる．

確率文法におけるそれらの値の理論解析で主要な障害となっているのが，識別不能性と呼ばれる性質である．パラメータ θ を持つ学習モデル $p(x \mid \theta)$ において，パラメータから確率分布への写像が一対一であるとき，学習モデルを識別可能，そうでない場合は，識別不能と定義する．確率文法やニューラルネットワークなど階層的な構造を持つ多くの学習モデルは，識別不能である．識別不能なモデルでは，フィッシャー情報量行列が縮退し，学習モデルのパラメータ空間には多数の特異点が存在する．そのため，フィッシャー情報量行列が正定値であることを仮定して導出されてきた多くの統計学的な方法が適用できず，学習モデルの汎化性能や理論的に保証された構造の同定など，基本的な性質の解明がいまだ十分には行なわれていない．

ベイズ法においては，代数解析的手法を用いて，識別不能なモデルも含めて，自由エネルギーのデータ数が無限大に増えていく場合の漸近的な振る舞いが明らかにされた[18,19]．その結果，自由エネルギーと汎化誤差は，同じパラメータ数の識別可能なモデルより，著しく小さくなることが示され，識別不能なモデルの優位性が明らかになった．しかしながら，ベイズ法を行なう際には，事後分布による平均操作における多重積分を実行する必要があるが，この操作が計算量的に必ずしも容易ではないことが知られている．

そこで，計算量を削減するための様々な近似法が提案されている．現在広く用いられている

手法には，二つのアプローチがある．一つめはサンプリング法であり，代表的な方法として事後分布に収束するマルコフ連鎖を用いる MCMC 法（マルコフ連鎖モンテカルロ法）がある[5]．MCMC 法の利点として，手法の汎用性と，繰り返し計算により極限における事後分布への収束が保証されているが，計算量が多いことや，収束判定が難しいことなどの問題点がある．二つめは決定論的方法で，代表的なものとしてラプラス法や変分ベイズ法がある．変分ベイズ法は，パラメータ上の確率分布で操作しやすいものを用いて，真の事後分布までのカルバック情報量を最小化することにより事後分布の近似を行う方法であり，少ない計算量と実世界の問題での有用性が報告されている[1,2,12]．しかしながら，その近似精度や汎化誤差などの理論的な性質は不明な点が多く，その数理的な構造を解明し，応用へと広げていくことが望まれている．

変分ベイズ法の近似精度に関する研究としては，混合指数型分布における変分自由エネルギーの漸近形が明らかにされている[17]．本稿では，識別不能な確率文法の変分ベイズ法における，自由エネルギーのデータ数が無限に大きくなっていく場合の漸近的な評価とその応用について議論を行なう．

2. 確率文法と識別性

2.1 確率文法の定式化

この章では，本稿で扱う確率文脈自由文法の定式化をおこなう．まず，一般性を失なうことなく，文法がチョムスキー標準形で書かれていることを仮定する．終端記号が M 個ある場合，長さ L の観測系列は，$X_i=\{x_{i1},\cdots,x_{iL}\}\in\{1,\cdots,M\}^L$ となる．与えられる系列の数を n とすると，観測データは，$X^n=\{X_1,\cdots,X_n\}$と表わせる．学習モデルは K 個の非終端記号を持つとする．そのとき，学習モデルは以下のように書き下せる．

$$
\begin{aligned}
p(x|\theta) &= \sum_{t\in T} p_t(x|\theta) \\
\theta &= \{a,b\},\ a=\{a^i_{jk}\}\ (1\le i,j,k\le K),\ b=\{b_{im}\}\ (1\le i\le K,\ 1\le m\le M)
\end{aligned}
\tag{1}
$$

ただし，T は長さ L の系列を生成する木全体の集合で，t は実際に生成された木を表し，木 t が確定することと，隠れ変数である非終端記号の生成列が確定することは同値である．さらに，各パラメータa^i_{jk}は，非終端記号 i が非終端記号の組（j, k）を生成する確率を表わし，b_{im} は，非終端記号 i が終端記号 m を出力する確率を表す．また，$\{a, b\}$にはそれぞれ，

$$
a^i_{ii}=1-\sum_{(j,k)\neq(i,i)} a^i_{jk},\ b_{iM}=1-\sum_{m=1}^{M-1} b_{im}
$$

の拘束条件があるものとする．

2.2 識別不能性の例

ここで，学習モデルの状態数が真の分布より多く，冗長な場合どのようなことが起きているか例を挙げて示す．これは統計学ではモデル選択と呼ばれる確率文法の構造を推定する問題を扱うときに必然的に起こる状況である[16]．最も簡単な場合である出力が 2 値の Left-to-Right 型の隠れマルコフモデルを考える．真の分布 $P_0(x)$ は一つの隠れ状態を持ち，出力確率のパラメータは，$b^*\in R$ で与えられるとする．真の分布から発生したデータについて，2 つの隠れ

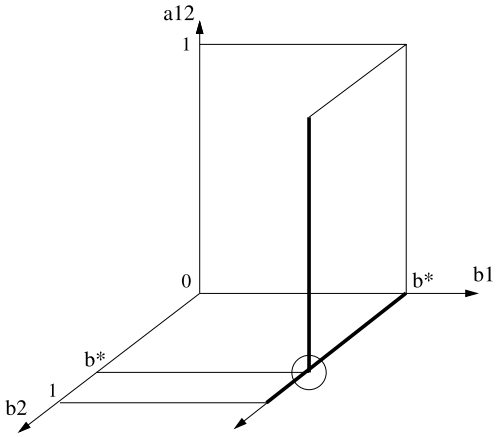

図 1　隠れマルコフモデルの識別不能性

状態を持つ学習モデルで学習を行なう．その時，学習モデルのパラメータ空間上で真の分布と一致する集合は図 1 のような 2 つの交差する集合（代数多様体になる）から構成される．太線は真の分布と一致するパラメータの集合であり，円の中心が特異点になっている．

$$\{b_1 = b^*,\ 0 \le b_2 \le 1, a_{12} = 0\} \cup \{b_1 = b_2 = b^*, 0 \le a_{12} \le 1\}$$

この集合は，(b_1, b_2, a_{12}) = (b^*, b^*, 0) に特異点を持つ．特異点上では，フィッシャー情報量行列が縮退しているため，正規分布での近似を行うことができず，このことが従来のモデル選択法である AIC, BIC, MDL 等[16]の適用を困難にしている．

3.　ベイズ法と変分ベイズ法

3.1　ベイズ法

この節では, ベイズ法の一般論について述べる. 真の分布 $P_0(x)$ から独立に発生した n 個のサンプル $X^n = \{X_1, \cdots, X_n\}$が与えられたものとする．学習モデルを θ をモデルパラメータ，$\varphi(\theta)$ をパラメータの事前分布とする．

そのとき，パラメータの事後分布は

$$p(\theta|X^n) = \frac{1}{Z(X^n)} \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)$$

で与えられる．ただし，$Z(X^n) = \int \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)d\theta$ は規格化定数であるが，これは学習モデルと事前分布の尤もらしさを与える量で周辺尤度もしくは証拠などと呼ばれ，通常異なる学習モデルを比較する場合にはこの値が大きい方を優位であるとみなす．

また，未来のサンプル X_{n+1} に対する予測は，学習モデルを事後分布で平均化した予測分布

$$p(X_{n+1}|X^n) = \int p(X_{n+1}|\theta)p(\theta|X^n)d\theta \tag{2}$$

で行なう．

自由エネルギーは周辺尤度の対数の符号反転,

$$F(X^n) = -\log Z(X^n)$$

で定義される．自由エネルギーは，確率的複雑さとも呼ばれ，定義よりこの値が小さいほど周辺尤度は大きくなる．よって自由エネルギーの挙動を調べることにより，モデル選択問題の数理的な基礎を確立することができる．また，n 個のサンプルが与えられたもとでの，真の分布と学習モデルのカルバック情報量であり，モデルにおけるベイズ法の性能を表現する汎化誤差は，

$$G(X^n) = \int p_0(x) \log \frac{p_0(x)}{p(x|X^n)} dx$$

で定義される．ベイズ法の汎化誤差は，自由エネルギーの増分に等しいことが知られている．

ベイズ法においては，特異モデルも含めて，自由エネルギーの漸近的な評価が与えれている[18]．自由エネルギーのサンプルの出方に対する平均を

$$F(n) = \langle -\log \int \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)d\theta \rangle_{p_0(X^n)} - nS$$

で定義する．ただし，記法 $\langle\rangle_{p(x)}$ は $p(x)$での平均を表わすこととし，S は真の分布のエントロピーで，学習モデルに依存しない量

$$S = -\int p_0(x) \log p_0(x) dx$$

で定義される．

そのとき $F(n)$は以下の漸近形を持つことが示されている[18]．

$$F(n) = \lambda \log n + o(\log n) \quad (n \to \infty)$$

ここで，λは正の有理数である．汎化誤差については上で述べた自由エネルギーとの関係から

$$G(n) = \frac{\lambda}{n} + o(\frac{1}{n})$$

であることが分かる．また，以降λをベイズ法の学習係数と呼ぶ．識別可能なモデルにおいては，学習モデルのパラメータ数が d であるとき，$\lambda = \frac{d}{2}$である．識別不能なモデルに関しては，一般にベイズ法で学習した場合にはλの値は$\frac{d}{2}$よりも小さくなり，最尤法を用いた場合にはλの値が$\frac{d}{2}$より大きくなることが示される．これは，識別不能なモデルでは，最尤法が過学習を起こす一方で，ベイズ法は過学習を防ぎ，逆に識別不能なモデルのエントロピーの大きさを利用した精度のよい推定が可能であることを表わしている．

3.2 変分ベイズ法

次に変分ベイズ法[1]について述べる．前章で述べたとおり，識別不能モデルにおいてはベイズ法が有効であることが分かっているが，予測分布（式（2））の計算における事後分布での平均操作が計算量的に困難であるため，様々な近似法が提案されている．変分ベイズ法もその一つであるが，自由エネルギーや汎化誤差はベイズ法と同等ではなく，その相違を考察することが本稿の目的となる．

観測されるサンプル X^n=$\{X_1,\cdots,X_n\}$とそれに対応する隠れ変数 Y^n=$\{Y_1,\cdots,Y_n\}$の組（X^n, Y^n）を完全サンプルセットと呼ぶ．ベイズ法の自由エネルギーは，

$$F(X^n) = -\log \int \sum_{Y^n} \prod_{i=1}^{n} p(X_i, Y_i|\theta)\varphi(\theta)d\theta$$
$$= -\log \int \sum_{Y^n} p(X^n, Y^n|\theta)\varphi(\theta)d\theta$$

と表わされる．ここで，$\sum_{Y^n}$は隠れ変数のすべての組み合わせで取る．変分ベイズ法は，自由エネルギーを任意の試験分布 $q(Y^n, \theta)$ を用いて近似する手法である．イェンセンの不等式を使うと，自由エネルギーは，

$$F(X^n) = -\log \int \sum_{Y^n} q(Y^n, \theta) \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n, \theta)} d\theta$$
$$\leq -\int \sum_{Y^n} q(Y^n, \theta) \log \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n, \theta)} d\theta \equiv F_v(X^n)$$

で上から押さえることができる．これを$p(X^n, Y^n, \theta) = p(Y^n, \theta|X^n)p(X^n)$を用いて以下のように書き替える．

$$F_v(X^n) = F(X^n) + \int \sum_{Y^n} q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{p(Y^n, \theta|X^n)} d\theta$$

この式から$F_v(X^n)$の最小化は，試験分布 $q(Y^n, \theta)$ から真の事後分布 $p(Y^n, \theta|X^n)$ へのカルバック情報量の最小化に等しくなっていること，および，等号が成立するのは，試験分布が真の事後分布と等しい場合に限ることが分かる．また，ベイズ法と変分ベイズ法の確率的複雑さの差が変分ベイズ法による事後分布の近似精度となっている．

変分ベイズ法においては，実現しやすい試験分布 $q(X^n, \theta)$ を構成するために，隠れ変数とパラメータが独立な形 $q(X^n)r(\theta)$ に制限したものを考察する．その時，制限されたもとでの自由エネルギー（以後，変分自由エネルギーと呼ぶ）$\overline{F}(X^n)$は，

$$F(X^n) \leq -\int \sum_{Y^n} q(Y^n)r(\theta) \log \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n)r(\theta)} d\theta$$
$$= -\int \sum_{Y^n} q(Y^n)r(\theta) \log \frac{p(X^n, Y^n|\theta)}{q(Y^n)} d\theta + \int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta \equiv \overline{F}(X^n)$$

と表わされる．

$\overline{F}(X^n)$は任意の分布$q(Y^n)$, $r(\theta)$に関する汎関数であり，その$\overline{F}(X^n)$を最小にする$q(Y^n)$および$r(\theta)$が満たすべき関係式を導出できることが知られている．実際，$\int r(\theta)d\theta = 1$ の制約のもと$\overline{F}(X^n)$を$r(\theta)$で変分し 0 と置く．その時，最適なパラメータの分布が満たす条件は

$$r(\theta) = C_r \exp\langle \log p(X^n, Y^n|\theta)\varphi(\theta)\rangle_{q(Y^n)} \tag{3}$$

となる．ただし，C_r は規格化定数である．同様に最適な隠れ変数の分布も C_q を規格化定数として，

$$q(Y^n) = C_q \exp\langle \log p(X^n, Y^n|\theta)\rangle_{r(\theta)} \tag{4}$$

となる．変分ベイズ法は，$\overline{F}(X^n)$の最小化を，式（3）と式（4）を繰り返し解くことにより行なわれる．学習モデル$p(X^n, Y^n)$を隠れ変数を持つ指数分布から選択し，モデルに対応する共役事前分布を用いた場合は，EM アルゴリズムと類似した非常に効率的なアリゴリズムが導出される．各繰り返しについて$\overline{F}(X^n)$が単調に減少すること，局所的な最小値に収束することが保証されている．確率文法の場合には，EM アリゴリズムを効率的に実行するための手段として，隠れマルコフモデルに対しては，Forward-Backward アルゴリズム，確率文脈自由文法については，Inside-Outside アルゴリズムが知られており[11,13]，それを変分ベイズ法に拡張したものも提案されている[12]．

4. 変分ベイズ法における自由エネルギーの漸近形

試験関数$q(Y^n)$および$r(\theta)$として$\overline{F}(X^n)$を最小にするものが得られたときの$\overline{F}(X^n)$の値をサンプルの出方について平均した値を$\overline{F}(n)$と書く．最近，筆者は共同研究者と協力して，サンプル数 n が大きくなるときの$\overline{F}(n)$の漸近形を明らかにした[7,8]．以下に確率文脈自由文法の場合の結果を述べる．隠れマルコフモデルに対しても同様の議論が成り立つ．変分自由エネルギーのサンプルの出方による平均を

$$\begin{aligned}\overline{F}(n) &= \langle \overline{F}(X^n) \rangle_{p_0(X^n)} - nS \\ &= \langle \min_{q,r} \{ -\langle \log \frac{p(X^n, Y^n|\theta)}{q(Y^n)} \rangle_{q(Y^n)r(\theta)} + \log p_0(X^n) + \int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta \} \rangle_{p_0(X^n)}\end{aligned} \tag{5}$$

と定義し，以下の条件（A 1），（A 2），（A 3），（A 4）を仮定する．

（A 1）真の分布 $p_0(x)$は K_0 個の非終端記号，M 個の終端記号を持ち，θ^*を定数の集合として，

$$\begin{aligned}p_0(x|\theta^*) &= \sum_{t \in T} p_{0t}(x|\theta^*) \\ \theta^* &= \{a^*, b^*\} \\ a^* &= \{a^{*i}_{jk}\} \ (1 \le i, j, k \le K_0, \ (j,k) \ne (i,i)) \\ b^* &= \{b^*_{im}\} \ (1 \le i \le K, \ 1 \le m \le M-1) \\ a^{*i}_{ii} &= 1 - \sum_{(j,k) \ne (i,i)} a^i_{jk} \\ b^*_{iM} &= 1 - \sum_{m=1}^{M-1} b_{im}\end{aligned}$$

と書ける．ただし，確率文脈自由文法においては隠れマルコフモデルと同様に，特定不能性の問題があるが[4,10]，K_0 はこのパラメトリゼーションのもとでの最小の非終端記号の数であるとする．

（A 2）すべての真のパラメータ$\{a^{*i}_{jk}, b^*_{im}\}$は正値であり，このパラメトリゼーションのもと最小のパラメータ数になっているとする．

$$\theta^* = \{\{a^{*i}_{jk} > 0\}, \{b^*_{im} > 0\}\}$$

（A 3）学習モデルは式（1）で与えられ，真の分布を含む．つまり，学習モデルの非終端記

号の数 K は $K_0 \leq K$ を満たす．

（A 4）a=$\{a^i_{jk}\}$および，b=$\{b_{im}\}$の事前分布は，ディリクレ分布で与えられ，そのハイパーパラメータは$\phi_0 > 0$，$\xi_0 > 0$ を満たすとする．

$$\varphi(a) = \prod_{i=1}^{K} \frac{\Gamma(K^2\phi_0)}{\Gamma(\phi_0)^{K^2}} \prod_{j=1,k=1}^{K} (a^i_{jk})^{\phi_0 - 1}$$

$$\varphi(b) = \prod_{i=1}^{K} \frac{\Gamma(M\xi_0)}{\Gamma(\xi_0)^{M}} \prod_{m=1}^{M} b_{im}^{\xi_0 - 1}$$

定理 1．条件（*A 1*）から（*A 4*）の仮定のもと，変分自由エネルギーの平均は，サンプル数 *n* が無限大の極限で，

$$\overline{F}(n) = \overline{\lambda} \log n + O(1) \quad (n \to \infty)$$

を満たす．ただし，

$$\overline{\lambda} = \begin{cases} \frac{K_0(K_0^2-1)+K_0(M-1)}{2} + K_0(K^2 - K_0^2)\phi_0 \\ \quad (\phi_0 \leq \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}) \\ \frac{K(K^2-1)+K(M-1)}{2} \\ \quad (\phi_0 > \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}) \end{cases} \tag{6}$$

で与えられる．

証明の方針について簡単に述べる．学習モデルが真の分布を含んでいるので，式(5)の始めの 2 つの項は合わせて $O(1)$以下であることから，事前分布から事後分布までのカルバック情報量である最後の項を考えればよいことが分かり，それを具体的に期待十分統計量で書き下すと，

$$\begin{aligned} & K(r(\theta)||\varphi(\theta)) \\ &= \sum_{i=1}^{K} \Big\{ \log\Gamma(n_i + K^2\phi_0) - n_i \Psi(n_i + K^2\phi_0) - \sum_{j,k=1}^{K} \{\log\Gamma(n^i_{jk} + \phi_0) - n^i_{jk}\Psi(n^i_{jk} + \phi_0)\} \Big\} \qquad (7) \\ &\quad + \sum_{i=1}^{K} \Big\{ \log\Gamma(\hat{n}_i + M\xi_0) - \hat{n}_i \Psi(\hat{n}_i + M\xi_0) - \sum_{m=1}^{M} \{\log\Gamma(n_{im} + \xi_0) - n_{im}\Psi(n_{im} + \xi_0)\} \Big\} \\ &\quad + const \qquad (8) \end{aligned}$$

となる．ただし$n_i, n^i_{jk}, \hat{n}_i$は期待十分統計量である．これをプサイ関数 $\Psi(x)$ と対数ガンマ関数 $\log\Gamma(x)$ の漸近形

$$\Psi(x) = \log x + O(1)$$

$$\log\Gamma(x) = (x - \frac{1}{2})\log x - x + O(1) \quad (x \to \infty)$$

を用いて変形すると，最終的には以下の最適化問題に帰着される．

$$K(r(\theta)||\varphi(\theta)) = \sum_{i=1}^{K}\Big\{(K^2\phi_0 - \frac{1}{2})\log(n_i + K^2\phi_0) - \sum_{j,k=1}^{K}\{(\phi_0 - \frac{1}{2})\log(n_{jk}^{i} + \phi_0)\}\Big\} + \sum_{i=1}^{K}\Big\{(M\xi_0 - \frac{1}{2})\log(\hat{n_i} + M\xi_0) - \sum_{m=1}^{M}\{(\xi_0 - \frac{1}{2})\log(n_{im} + \xi_0)\}\Big\} + O(1) \tag{9}$$

詳細については，文献[7, 8]を参照のこと．

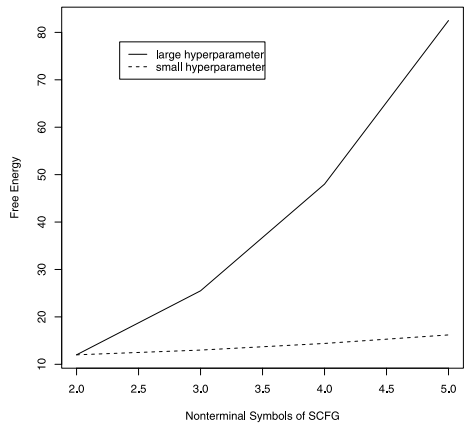

図 2　変分自由エネルギーの模式図

5.　考察

解析の結果から変分自由エネルギーは，非終端記号の生成確率の事前分布ϕ_0によって場合分けされることが分かる．ϕ_0が大きい場合は，学習係数λは学習モデルのパラメータ数に一致し，モデル選択基準として広く使われている BIC[15]に等しい．一方，ϕ_0が小さい場合は，最小値は，冗長な非終端記号を消す条件を満たし，モデルパラメータ数に比べてずっと小さな値をとる．図 2 に変分自由エネルギーの模式図を示す．縦軸は *log n* で正規化した変分自由エネルギーである．真の分布の非終端記号の数が 2，終端記号の数が 10 の場合．上の実線は事前分布のハイパーパラメータϕ_0が大きい場合で BIC と一致する．下の点線は$\phi_0 = 0.1$ の場合である．この結果は，変分ベイズ法が識別不能モデルで最尤推定を行ったときに問題となる過学習の現象をうまく回避する優れた手法であることを示す．

次に結果をモデル選択に応用することを考える．具体的には変分自由エネルギーの漸近形である式（6）が真のモデルの状態数 K_0 を含んでいることを利用する[20]．証明の方針から分かるように変分ベイズ法では，事後分布と事前分布のカルバック情報量 $K(r(\theta)||\phi(\theta))$ が最終的な漸近形を決定するが，この量はディリクレ分布同士のカルバック情報量になるので簡単に計算することができる．確率文脈自由文法の場合に書き下してみる．真の非終端記号数 K_0 より大きい K 個の非終端記号をもち，事前分布のハイパーパラメータが $\phi_0 = \frac{1}{2}$ である学習モデルにおいて，最適化の結果，変分自由エネルギーの最小点が見つかったとすると，真のモデルの非終端記号を表す指標として以下の量が計算できる[9]．

$$K_0 \leftarrow \frac{2K(r(\theta)||\phi(\theta))}{K^2 + M - 2}$$

このようなうまい指標が構成できるので，変分ベイズ法による確率文法の構造推定は理論的に

保証される．さらに，この手法の利点としては，真の状態数を直接推定するので，各モデルを情報量基準に基づいて比較する従来のモデル選択において生じる，組み合わせ的な複雑さを回避することができる．

6. おわりに

確率文法の変分ベイズ法における変分自由エネルギーの漸近評価をおこない，変分ベイズ法の有効性を理論的に示した．さらに，解析の結果から導かれる確率文法の構造を推定する新しい方法について議論した．なお本稿で示した結果は，東京工業大学の渡辺澄夫教授，渡辺一帆氏，山崎啓介助手との共同研究に基づいている．

参考文献
［1］ H. Attias,“ Inferring parameters and structure of latent variable models by variational Bayes ”, in *Proc.15th Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 21–20.
［2］ M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD thesis, University College London, 2003.
［3］ T. W. Deacon,“ The Symbolic Species ”, Norton, 1998.
［4］ E. Gassiat and S. Boucheron.“ Optimal error exponents in hidden Markov models order estimation. ”, *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 964–980, 2003.
［5］ W. R. Gilks, S. Richardson, D. J. Speigelhalter（eds）, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.
［6］ J. E. Hopcroft.“ Introduction to Automata Theory, Languages and Computation. ”, Addison–Wesley, 1979.
［7］ T. Hosino, K. Watanabe, and S. Watanabe.“ Stochastic Complexitiy of Variational Bayesian Hidden Markov Models ”, *International Joint Conferenece on Neural Networks*, 2005.
［8］ T. Hosino, K. Watanabe, and S. Watanabe.“ Free Energy of Stochastic Context Free Grammar on Variational Bayes ”, *The 13th International Conference on Neural Information Processing*, 2006.
［9］ T. Hosino, K. Watanabe, K. Yamazaki, and S. Watanabe.“ Model Selection for Stochastic Grammar Based on the Variational Bayes ”, *in preparation*.
[10] H. Ito, S. Amari, and K. Kobayashi.“ Identifiability of hidden Markov information sources and their minimum degrees of freedom ”, *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 324–333, 1992.
[11] 北研二，“ 確率的言語モデル ”，東京大学出版会，1999.
[12] K. Kurihara and T. Sato,“ An Application of the Variational Bayesian Approach to Probabilistic Context–Free Grammars. ”, *International Joint Conference on Natural Language Processing*, 2004.
[13] K. Lari and S. Young,“ The estimation of stochastic context–free grammars using the inside–outside algorithm, *Computer Speech and Language*, vol. 4, pp. 33–56, 1990.
[14] S. Pinker,“ The Language Instinct : How the Mind Creates Language ”, Harper Perennial Modern Classics, 2000.
[15] G. Schwarz,“ Estimating the dimension of a model ”, *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
[16] 下平英寿，伊藤秀一，久保川達也，竹内啓“ 統計科学のフロンティア 3 モデル選択 ”，岩波書店，2004.
[17] K. Watanabe and S. Watanabe,“ Lower bounds of stochastic complexities in variational bayes learning of gaussian mixture models, ” in *Proc. IEEE conference on Cybernetics and Intelligent Systems*, 2004, pp. 99–104.
[18] S. Watanabe,“ Algebraic analysis for non–identifiable learning machines, ” *Neural Computation*, vol. 13, no. 4, pp. 899–933, 2001.
[19] 渡辺澄夫，“ 代数幾何と学習理論，” 森北出版株式会社，2006.

［20］ K. Yamazaki, Kenji Nagata, Sumio Watanabe,“ A New Method of Model Selection Based on Learning Coefficient ” 2005 *International Symposium on Nonlinear Theory and its Applications*, 2005.

執筆者紹介 星 野 力（Tikara Hosino）
2001年日本ユニシス(株)入社．データマイニングソフトの開発に従事．2005年東京工業大学博士後期過程（社会人コース）入学．学習理論について研究している．