数理モデル化と問題解決　41－2
（２００２．　９．　２０）

拡張混合表現を用いた学習と遺伝子型分布

塩谷浩之 †　内田真人 ‡
† 室蘭工業大学　‡ NTT サービスインテグレーション基盤研究所
† 〒 050-8585 室蘭市 / Email:shioya@csse.muroran-it.ac.jp
‡ 〒 180-8585 武蔵野市 / Email:uchida.masato@lab.ntt.co.jp

アブストラクト: 線型混合は，混合ガウスモデルなどよく用いられる一般的な混合システムの形式である．一方，学習機械の混合によるアンサンブルを基盤とする確率モデルにおいては，指数型分布族の性質を持つ線形混合とは異なる拡張混合形式が用いられ，新たな工学的応用が期待される．そこで本研究では，拡張表現モデリングの応用として，遺伝的アルゴリズムにおける遺伝子型分布に関する条件付確率的推移モデルに適用する．そして，遺伝的アルゴリズムの混合モデル的解析について述べる．

キーワード：遺伝子型分布，無限母集団，拡張混合表現，分布間情報量，アンサンブル学習

Genotype Distributions and Learning Procedure using an Extended Mixture Formula

Hiroyuki Shioya †　Masato Uchida‡
† Muroran Institute of Technology　‡ NTT Service Integration Laboratories
† Muroran-shi, 050-8585 / Email:shioya@csse.muroran-it.ac.jp
‡ Musashino-shi, 180-8585 / Email: uchida.masato@lab.ntt.co.jp

Abstract: A linear mixture is a standard formula used well in Gaussian mixture distribution model. However an extended mixture formula, which has an exponential property, is used in the stochastic model based on an ensemble of the mixture learning machines, and the mixture formula is expected to be applied to a new engineering application. In this paper, we apply the modeling of the mixture formula to the conditional probability model with respect to the transition of genotype distributions in the genetic algorithms, and we show some results concerning to the modeling of the mixture system of the genetic algorithms.

Keywords: genotype distribution, infinite population, an extended mixture, information divergence, ensemble learning

1　まえがき

遺伝的アルゴリズム (GA) において，世代ごとに施される遺伝子母集団への操作 (GA オペレーション) は，遺伝子母集団の遺伝子型の存在割合を示している遺伝子型分布の変換と見なさる．前世代の遺伝子集団の遺伝子型分布を条件とし，GA により，次世代にある遺伝子型分布が得られる条件付確率モデルが，説明的によく用いられる [2]．あまり多くない遺伝子母集団で，適度な計算コストで準最適解を求められる実用上の GA の良さとは別に，遺伝子型分布に着目されており，その中で，無限個の遺伝子母集団を仮定した研究が行われている．[3] [4]．無限個の遺伝子母集団 GA とは確率的に次世代が定められる有限集団 GA の決定論的版であり，GA の学習論的様相が十分に感じられる．

本研究では，無限遺伝子母集団 GA と統計的学習との関連を固め，混合モデルを用いた統計的学習問題を，遺伝子型分布に関する変換写像モデルに適応することで，GA に関する混合システムモデルを構成する．通常は，混合 Gauss モデルで用いられる線形混合の形式を混合モデルとしているが，本論では，指数型混合による拡張混合表現を用いた混合モデル [8] を用い，マルコフ的な有限遺伝子集団 G A の確率モデルに導入することで得られた結果について述べる．

2　遺伝子型分布と確率モデル

2.1　遺伝子型分布

サイズ n の遺伝子個体群 A_n の各個体は長さ l のビット列で表現されている．すると A_n は Ω 上の分布 $\boldsymbol{x} = (x_1, x_2, \ldots, x_{2^l})^T$ で特徴付けられる ($x_k = n_k/n$ $(k = 1, \ldots, 2^l)$, n_k は遺伝子型 k の個体の数，^{T} は転置)．このとき A_n は l ビットの遺伝子個体の型表現全体 Λ，すなわち

$$\Lambda = \left\{ \boldsymbol{x} \;\middle|\; \sum_{k=1}^{2^l} x_k = 1,\ x_k \geq 0 \ \ (\forall k) \right\} \tag{1}$$

で特徴付けられ，n が既知のとき，A_n は対応する分布 $\boldsymbol{x}$ から一意に定まる．

個体群 A_n からその次世代の個体群を生成する GA のオペレーションは，A_n の遺伝子型分布 $\boldsymbol{x}$ に依存する分布にしたがって Ω から n 個のサンプルを抽出と考える．この

「$\boldsymbol{x}$ に依存する分布」を，写像 $\boldsymbol{G}:\Lambda\to\Lambda$ を用いて $\boldsymbol{G}(\boldsymbol{x})$ と書く．現世代の個体群の分布が $\boldsymbol{x}$ のとき，次世代の個体群の分布が $\boldsymbol{y}$ となる条件付き確率 $P^m_{\boldsymbol{G}}(\boldsymbol{y}|\boldsymbol{x})$ は次の多項分布で与えられる [2]．

$$P^m_{\boldsymbol{G}}(\boldsymbol{y}|\boldsymbol{x}) \stackrel{\text{def}}{=} n!\prod_{k=1}^{K}\frac{(\boldsymbol{G}(\boldsymbol{x})_k)^{ny_k}}{(ny_k)!} \tag{2}$$

ただし，$K=2^l$，$\boldsymbol{G}(\boldsymbol{x})$ は Vose-map [3] による無限個の遺伝子母集団の決定論的写像とする．$\boldsymbol{G}$ は，適応度分布 $\boldsymbol{F}$ および遺伝的オペレータ $\boldsymbol{M}$ の合成写像 $\boldsymbol{M}\cdot\boldsymbol{F}$ で表され，その付随するパラメータを $\boldsymbol{\theta}$ とする．前世代の遺伝子型分布 $\boldsymbol{x}$ を用い，次世代の遺伝子型分布を $\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x})$ とする．Λ_n における $\boldsymbol{y}$ の確率は，$P_{\boldsymbol{G}}$ の期待値ベクトル $\boldsymbol{G}$ を用い以下のように表現される．

$$\Pr\{\boldsymbol{y}|D_k(\boldsymbol{y}\|\boldsymbol{G}(\boldsymbol{x}))>\epsilon\}\le h(n)\exp\{-n(\epsilon-\log K)\} \tag{3}$$

ただし $D_k(\cdot\|\cdot)$ はカルバックの情報量，$h(n)$ は n に関する正値関数．

無限個の遺伝子母集団を仮定することで，遺伝的オペレータによる遺伝子型分布の更新は決定論的更新，すなわち次世代の遺伝子型分布が決定論的に定まる．有限の個体群サイズ n をもつ GA は，Λ_n を状態集合，式 (2) を状態 $\boldsymbol{x}$ ($\in\Lambda_n$) から状態 $\boldsymbol{y}$ ($\in\Lambda_n$) への推移確率で表現される．

2.2 確率モデルと情報量最小化

ある問題がコーディングされた未知の遺伝的アルゴリズム GA* からの遺伝子型分布に関するデータ $\boldsymbol{D}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_{|\boldsymbol{D}|},\boldsymbol{y}_{|\boldsymbol{D}|})\}$ が各々独立に観測されたとする．多項分布モデルの負の対数尤度は，

$$\begin{aligned}&-\log P^m_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x}) = -\log\frac{n!}{\prod_{k=1}^{K}(ny_k)!}\\ &+\quad nD_k(\boldsymbol{y}\|\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x}))-nH(\boldsymbol{y})\end{aligned} \tag{4}$$

となる．ただし $H(\cdot)$ はエントロピ関数．データ全体に対する負の対数尤度の最小化は，観測データ数 $|\boldsymbol{D}|$ を N，データ $\boldsymbol{D}$ に関する Λ_n 上の分布を $P^{\Lambda_n}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}$，$\boldsymbol{x}$ の入力分布を Q として，

$$\begin{aligned}&\arg\min_{\boldsymbol{\theta}} D_k(P^{\Lambda_n}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}(\boldsymbol{D})\|Q_{\boldsymbol{x}}P^m_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x}))\\ \simeq\ &\arg\min_{\boldsymbol{\theta}}\sum_{d=1}^{|\boldsymbol{D}|}D_k(\boldsymbol{y}_d\|\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x}_d))\end{aligned} \tag{5}$$

となる．ただし $\lim_{N\to\infty}P^{\Lambda_n}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}(\boldsymbol{D})=P^m_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}$ とする．多項分布モデルの学習とは，出力遺伝子型分布と決定論的写像による遺伝子型分布とのカルバック情報量を最小化することに相当する．

全く別のアプローチでは，GA による遺伝子型分布の変換を一般の入出力機械とみなすことで，データから学習させることも可能である．その際，確率モデルを

$$P^g_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{C}\exp\{-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x})\|^2\} \tag{6}$$

とすると (C は正定数)，最尤法により，

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{D}}=\arg\min_{\boldsymbol{\theta}\in\Theta}\frac{1}{2|D|}\sum_{j=1}^{|\boldsymbol{D}|}\|\boldsymbol{y}_j-\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\|^2 \tag{7}$$

を求めることで学習が行われる．式 (7) において，最小化される項 $\|\boldsymbol{y}_j-\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\|^2$ は，Quadratic divergence ($D_Q(\cdot\|\cdot)$) と呼ばれる．よって，適当な平滑化手法[1]によって得られた $\boldsymbol{R}$ 上のデータ分布を，$P^{\boldsymbol{R}}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}(\boldsymbol{D})$ とすると，

$$\begin{aligned}&\arg\min_{\boldsymbol{\theta}} D_k(P^{\boldsymbol{R}}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}(\boldsymbol{D})\|Q_{\boldsymbol{x}}P^g_{\boldsymbol{G}_{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x}))\\ \simeq\ &\arg\min_{\boldsymbol{\theta}}\sum_{d=1}^{|\boldsymbol{D}|}D_Q(\boldsymbol{y}_d\|\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{x}_d))\end{aligned} \tag{8}$$

となる．ただし観測データ数 $N=|\boldsymbol{D}|$ とし，

$$\boldsymbol{\theta}^*=\lim_{N\to\infty}\arg\min_{\boldsymbol{\theta}}D_k(P^{\boldsymbol{R}}_{\boldsymbol{G}_{\boldsymbol{\theta}^*}}(\boldsymbol{D})\|Q_{\boldsymbol{x}}P^g_{\boldsymbol{G}_{\boldsymbol{\theta}}}) \tag{9}$$

を満たすとする．以上から，前提とする確率モデルが多項分布モデルであるか，ガウスモデルであるかの選択は，入出力学習機械の学習の立場からは，最小化する情報量の選択と等価であることが分かる．

定理 1 決定論的写像を $\boldsymbol{G}^*$ とする未知の *GA* からの独立なデータ D から，式 *(2)* を確率モデルとし，学習アルゴリズムに従い，学習結果 $\hat{\boldsymbol{G}}$ が得られたとする．その結果，$\epsilon>0$ に対し，$D_k(P^m_{\boldsymbol{G}^*}\|P^m_{\hat{\boldsymbol{G}}})<\epsilon$ を達成したならば，決定論的写像 $\boldsymbol{G}^*$ および $\hat{\boldsymbol{G}}$ について，

$$D_k(\boldsymbol{G}^*\|\hat{\boldsymbol{G}})<\frac{\epsilon}{n},\quad D_Q(\boldsymbol{G}^*\|\hat{\boldsymbol{G}})\le 2\sqrt{\frac{\epsilon}{n}}. \tag{10}$$

が成り立つ．□

情報量最小化による統計的推定において，用意した統計モデルに情報源となる真の分布が含まれる場合は，どの分布間距離 (divergence) を用いてもデータ数 $N\to\infty$ で真の分布のパラメータを達成することができるが，データにノイズが混入するときは，異なる divergence を用いた場合，異なる学習結果が得られる．よって次の系が得られる．

系 1 遺伝子型分布に関するデータの情報源である GA が，少なくとも *SGA* であり，データ観測の際にノイズを含まない場合には，データが十分に多く観測される場合，モデル *(2)* の学習結果と，*(6)* における学習結果の差異は十分に小さい．

[1]例えば $\boldsymbol{R}$ 上の確率密度関数モデル $\{P_\theta(x):\theta\in\Theta,\int_{\boldsymbol{R}}P_\theta(x)dx=1,\ P_\theta(x)\ge 0,\ \forall\theta\}$ において，観測データ全体の経験分布 P_e とモデル P_θ との分布間情報量 $D(\cdot\|\cdot)$ による比較を行う場合，観測データから，ある種の補間操作による平滑化手法により，$\boldsymbol{R}$ 上の密度関数 $\bar{P}_e$ を生成することで，$D(\bar{P}_e\|P_\theta)$ が計算される．

3 混合学習機械と遺伝子型分布

3.1 混合システムの学習

一般の入出力学習機械を，遺伝子型分布の変換モデルに対応させたモデル (6) を元に，[6] においては，複数の入出力システムの混合 (例えば，NGnet 等) を導入することで，ある最適化問題に対して適用された遺伝的アルゴリズムを情報源とする遺伝子型分布データから，エキスパートによる分解的表現を実現している．

ところで，複数の学習機械を組み合わせ，汎化性能を向上させるシステムの学習は，アンサンブル学習と呼ばれる [7]．その枠組みとは，M 個の異なる入出力システム f_i について，各学習機械はパラメータ $\boldsymbol{\theta}_i$ を持ち，各システムへの入力 $\boldsymbol{x}$ に対する出力 $y_i = f_{\boldsymbol{\theta}_i}(\boldsymbol{x})$ を各学習機械の出力とする．入力 $\boldsymbol{x}$ と望ましい出力 y からなるサンプルの組 $(\boldsymbol{x}, y)$ が，ある確率分布 $p_*(\boldsymbol{x}, y)(= q(\boldsymbol{x})p_*(y|\boldsymbol{x}))$ にしたがって互いに独立に $\sum_{i=1}^{M} n_i$ 個観測されたとし，これを $D_{n_i} = \{(\boldsymbol{x}_1^{(i)}, y_1^{(i)}), \cdots, (\boldsymbol{x}_{n_i}^{(i)}, y_{n_i}^{(i)})\}$ とする．自乗誤差関数を損失関数とするアンサンブル学習とは

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} \sum_{(\boldsymbol{x},y)\in D_{n_i}} (y - f_{\boldsymbol{\theta}_i}(\boldsymbol{x}))^2 \tag{11}$$

により与えられる

$$f^l_{\hat{\boldsymbol{\theta}},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{i=1}^{M} \beta_i f_{\hat{\boldsymbol{\theta}}_i}(\boldsymbol{x}) \tag{12}$$

を予測機械として用いることをいう．ただし

$$\sum_{i=1}^{M} \beta_i = 1, \quad \beta_i > 0$$
$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)^T \in \boldsymbol{R}^M$$
$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_M^T)^T \in \boldsymbol{R}^{\sum_{i=1}^{M} k_i}.$$

ある関数 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ $(\in \boldsymbol{R})$ を用いて，$\boldsymbol{x}$ に対する y の条件付き確率分布

$$P^g_{f_{\boldsymbol{\theta}}}(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2\right\} \tag{13}$$

を定めると（σ は正定数），式 (11) は

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i}\left\{-\sum_{(\boldsymbol{x},y)\in D_{n_i}} \log P^g_{f_{\boldsymbol{\theta}}}(y|\boldsymbol{x})\right\} \tag{14}$$

と書き換えられ，式 (12) は以下に恒等的に等しい [1]．

$$P^e_{f^l_{\hat{\boldsymbol{\theta}}_i},\boldsymbol{\beta}}(y|\boldsymbol{x}) = \frac{\prod_{i=1}^{M} p^g_{f_{\hat{\boldsymbol{\theta}}_i}}(y|\boldsymbol{x})^{\beta_i}}{\int_{\mathbb{R}} \prod_{i=1}^{M} p^g_{f_{\hat{\boldsymbol{\theta}}_i}}(y|\boldsymbol{x})^{\beta_i} dy} \tag{15}$$

式 (6) を確率モデルとする場合，アンサンブル学習の定式化に沿えば，[6] において，複数の GA システム出力の線形混合が用いられている．よって，式 (6) のシステム混合モデルは，式 (15) とするのが妥当であることがわかる [8]．

[6] においては，基底となるエキスパートは，遺伝子座の交換によって用意されるものであるので，データに対して複数の学習試行によって用意されるアンサンブルとは異なるが，複数の学習機械の準備について，上記の違いを許容する．

3.2 拡張混合システム表現

多項分布を用いた確率モデルでの学習は，カルバック情報量の最小化であったので．アンサンブル学習の意味で用意されるべきシステム群 $\boldsymbol{G}_{\hat{\boldsymbol{\theta}}_i}(\boldsymbol{x})$ は，

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{D}_i} D_k(\boldsymbol{y}\|\boldsymbol{G}_{\boldsymbol{\theta}_i}(\boldsymbol{x})) \tag{16}$$

によって得られる．アンサンブル学習の場合，データを $\boldsymbol{D} = \cup_i \boldsymbol{D}_i = \cup_i \{(x_j, y_j^*)\}_j^i$ と表現する．各システムを個別化するために $\boldsymbol{G}_{\hat{\boldsymbol{\theta}}_i}(\boldsymbol{x})$ を $\boldsymbol{G}_i(\boldsymbol{x})$ とする．[6] における最適化問題の分解表現の意味では，データからの学習で各システムが用意されるのではなく，学習済みのシステムを用い，遺伝子座を入れ替えることで，システム群が用意されることとなる．

多項分布モデルの場合，各 GA システムの出力の線形混合を用いても，これまで述べたアンサンブル的枠組みにはならないが，式 (15) の形式の混合によってアンサンブル的枠組みが得られる．具体的には，以下の定理で表現される．

定理 2 確率モデル *(2)* において，以下の拡張混合表現

$$\boldsymbol{G}^e_{\boldsymbol{\beta}}(\boldsymbol{x})_k = \frac{\prod_{j=1}^{M} (\boldsymbol{G}_j(\boldsymbol{x})_k)^{\beta_j}}{\sum_{s=1}^{K} \prod_{j=1}^{M} (\boldsymbol{G}_j(\boldsymbol{x})_s)^{\beta_j}} \tag{17}$$

は，式 *(15)* と同等な形式の以下に等しい．

$$P^m_{\boldsymbol{G}^e_{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\prod_j^M P_{\boldsymbol{G}_j}(\boldsymbol{y}|\boldsymbol{x})^{\beta_j}}{\sum_{\Lambda_n} \prod_j^M P_{\boldsymbol{G}_j}(\boldsymbol{y}|\boldsymbol{x})^{\beta_j}} \tag{18}$$

ただし β は $\sum_i^M \beta_i = 1$ を満たす．

$D_k(\boldsymbol{y}\|\boldsymbol{G}^e_{\boldsymbol{\beta}}(\boldsymbol{x}))$ について，

$$\begin{aligned} D_k(\boldsymbol{y}\|\boldsymbol{G}^e_{\boldsymbol{\beta}}(\boldsymbol{x})) &= \sum_{j=1}^{M} \beta_j D_k(\boldsymbol{y}\|\boldsymbol{G}_j(\boldsymbol{x})) \\ &+ \log \sum_{s=1}^{K} \prod_{j=1}^{M} \boldsymbol{G}_j(\boldsymbol{x})_s^{\beta_j} \end{aligned} \tag{19}$$

となることから，以下の補題が成り立つ．

補題 1 M 個のアンサンブル的 *GA* システム群 $\boldsymbol{G}_j (j = 1, \cdots, M)$ が用意されている場合，データ $\boldsymbol{D}$ による $\boldsymbol{\beta}$ の

学習,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\boldsymbol{D}} D_k(\boldsymbol{y} \| G^e_{\boldsymbol{\beta}}(\boldsymbol{x})) \tag{20}$$

で与えられる $\hat{\boldsymbol{\beta}}$ について，指数型分布族の性質 *([5])* から，$\forall j$ に対し，

$$D_k(\boldsymbol{y} \| \boldsymbol{G}^e_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x})) = D_k(\boldsymbol{y} \| \boldsymbol{G}_j(\boldsymbol{x})) - D_k(\boldsymbol{G}^e_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}) \| \boldsymbol{G}_j(\boldsymbol{x})) \tag{21}$$

が成り立つ．□

4 むすび

多項分布モデルを確率モデルとする有限遺伝子集団ＧＡの学習は，入出力機械の視点からは，その入出力データに関するカルバック情報量の最小化に帰着される．本論では，アンサンブル的混合システム構成のために，拡張混合表現を新たに導入し，GAの学習モデルである多項分布モデルにおいて拡張混合学習の枠組みを構成した．アンサンブル学習の理論解析との関連についてなどを今後の課題としたい．

参考文献

[1] T.M.Cover and J.A.Thomas : Elements of Information Theory, Wiley-Interscience publication, America, 1991.

[2] T. E. Davis and J. C. Principe : A Markov chain framework for the simple genetic algorithm, Evolutionary Computation, vol.1, no.3, pp.269–288, 1993.

[3] M. D. Vose : Modeling simple genetic algorithms, Evolutionary Computation, vol.3, no.4, pp.453–472, 1996.

[4] M. D. Vose and A. H. Wright : Simple genetic algorithms with linear fitness, Evolutionary Computation, vol.2, no.4, pp.347–368, 1995.

[5] 甘利 俊一, 長岡 浩司, 情報幾何の方法, 岩波講座 応用数学 6 [対象 12], 岩波書店, 1993.

[6] 今井 順一, 塩谷 浩之, 伊達　悼 : 学習機械を利用した遺伝的アルゴリズムのモデリングに関する検討, 信学技報, NC 2001–115, 2002.

[7] 上田 修功, 中野 良平 : アンサンブル学習における汎化誤差解析, 信学論 (D-II), vol.J80-D-II, no.9, pp2512-2521, 1997.

[8] 内田 真人, 塩谷 浩之, 伊達 悼 : アンサンブル学習の解析と拡張, 信学論 (D-II), vol.J84-D-II, no.7, pp.1537-1542, 2001.

[9] 塩谷 浩之, 佐藤 佳久, 伊達 悼 : ボルツマン機械の学習と擬距離最小化基準, 信学技報, NC99-24 (1999-06)

5 付録

証明 (定理 1) 学習結果 $\hat{\boldsymbol{G}}$ と決定論的写像を $\boldsymbol{G}^*$ とのカルバック情報量について仮定より $D_k(P_{\boldsymbol{G}^*} \| P_{\hat{\boldsymbol{G}}}) < \epsilon$ とする．確率分布とパラメータ空間との 1 対 1 対応から，$D_k(\boldsymbol{G}^* \| \hat{\boldsymbol{G}}) < \tau(\epsilon)$ を満たす単調減少関数 $\tau(\epsilon)$ が存在する．しかし，カルバック情報量から直接展開した場合，タイトな不等式が得られないため，カルバック情報量と関連深い以下の α-divergence

$$D_\alpha(P^m_{\boldsymbol{G}^*} \| P^m_{\hat{\boldsymbol{G}}}) \stackrel{\text{def}}{=} \frac{1}{\alpha(1-\alpha)}\Big(1 - \sum_{\boldsymbol{y}\in\Lambda_n} P^m_{\boldsymbol{G}^*}(\boldsymbol{y})^{1-\alpha} P^m_{\hat{\boldsymbol{G}}}(\boldsymbol{y})^{\alpha}\Big). \tag{22}$$

(ただし $\alpha \in [0,1]$) を用いる．式 (22) の右辺の第 2 項を計算すると，

$$\begin{aligned}
&\sum_{\boldsymbol{y}\in\Lambda_n} P^m_{\boldsymbol{G}^*}(\boldsymbol{y})^{1-\alpha} P^m_{\hat{\boldsymbol{G}}}(\boldsymbol{y})^{\alpha} \\
&= \sum_{\boldsymbol{y}\in\Lambda_n} \Big(\frac{n!}{\prod_{k=1}^K (ny_k)!} \prod_{k=1}^K (\boldsymbol{G}_k^{*ny_k})^{1-\alpha} (\hat{\boldsymbol{G}}_k^{ny_k})^{\alpha} \Big) \\
&= \Big\{ \sum_{s=1}^K \boldsymbol{G}_s^{*1-\alpha} \hat{\boldsymbol{G}}_s^{\alpha} \Big\}^n \Big\{ \sum_{\boldsymbol{y}\in\Lambda_n} \frac{n!}{\prod_{k=1}^K (ny_k)!} \\
&\quad \prod_{k=1}^K \left(\frac{\boldsymbol{G}_k^{*1-\alpha} \hat{\boldsymbol{G}}_k^{\alpha}}{\sum_{s=1}^K \boldsymbol{G}_s^{*1-\alpha} \hat{\boldsymbol{G}}_s^{\alpha}} \right)^{ny_k} \Big\} \\
&= \Big(\sum_{s=1}^K \boldsymbol{G}_s^{*1-\alpha} \hat{\boldsymbol{G}}_s^{\alpha} \Big)^n \\
&= \Big\{ 1 - \alpha(1-\alpha) D_\alpha(\boldsymbol{G}^* \| \hat{\boldsymbol{G}}) \Big\}^n
\end{aligned} \tag{23}$$

定理の仮定と補題 2 の式 (27) から

$$D_\alpha(P^m_{\boldsymbol{G}^*} \| P^m_{\hat{\boldsymbol{G}}}) \le \frac{\epsilon}{1-\alpha}. \tag{24}$$

が得られ，$\lim_{\alpha\to 0} D_\alpha = D_k$ および補題 2 の式 (26) を用い，

$$D_k(\boldsymbol{G}^* \| \hat{\boldsymbol{G}}) < \frac{\epsilon}{n},\ D_Q(\boldsymbol{G}^* \| \hat{\boldsymbol{G}}) \le 2\sqrt{\frac{\epsilon}{n}} \tag{25}$$

が得られる．[Q.E.D.]

補題 2 p, q を任意の確率分布とする．

$$\frac{(D_Q(p\|q))}{4} \le \frac{(D_v(p\|q))}{4} \le D_K(p\|q) \tag{26}$$

$$D_\alpha(p\|q) \le \frac{1}{1-\alpha} D_k(p\|q) \tag{27}$$

ただし $D_v(p\|q) \stackrel{\text{def}}{=} \int |p-q| d\mu$ とする [9]．□