アンサンブル学習の解析と拡張

内田　真人[†]　　塩谷　浩之[††]
伊達　　惇[††]

Analysis and Extension of Ensemble Learning

Masato UCHIDA[†], Hiroyuki SHIOYA[††],
and Tsutomu DA-TE[††],

[†] NTT サービスインテグレーション基盤研究所，武蔵野市
NTT Service Integration Laboratories, Musashino-shi, 180–8585 Japan
[††] 北海道大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkaido University, Sapporo-shi, 060–8628 Japan

あらまし　汎化能力向上の目的から，個々に学習された複数の学習機械の出力を適当に重み付けしたものを予測値として用いるというアンサンブル学習が提案されている．本論文ではこの学習法の特性を，負の対数ゆう度関数を損失関数とする枠組みを用いて解析する．また，この枠組みを拡張し，アンサンブル学習を意味づける．

キーワード　アンサンブル学習，平均予測誤差，α-ダイバージェンス

1. まえがき

個々に学習された複数の学習機械が与えられたとき，汎化能力の高い予測機械をどのように得るかという問題は重要である．アンサンブル学習は，与えられた機械の出力を適当に重み付けして予測値とすることでこの問題への対応を試みる手法であり，次のように定式化されている [4].

M 個の異なる学習機械を考え，各々，入力 $\boldsymbol{x} = (x_1, \cdots, x_m)^T$ $(\in \mathbb{R}^m)$ に対して $f(\boldsymbol{x}; \boldsymbol{\theta}_i)$ $(\in \mathbb{R})$ を出力するものとする $(i = 1, \cdots, M)$．ただし，$\boldsymbol{\theta}_i = (\theta_i^{(1)}, \cdots, \theta_i^{(k_i)})^T$ $(\in \mathbb{R}^{k_i})$ は学習機械 $f(\boldsymbol{x}; \boldsymbol{\theta}_i)$ の修正可能なパラメータであり，$f(\boldsymbol{x}; \boldsymbol{\theta}_i)$ は $\boldsymbol{\theta}_i$ に関して高階微分可能とする．また，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 $p_*(\boldsymbol{x}, y) (= q(\boldsymbol{x}) p_*(y|\boldsymbol{x}))$ に従って互いに独立に $\sum_{i=1}^M n_i$ 個観測されたとし，これを $D_{n_i} = \{(\boldsymbol{x}_1^{(i)}, y_1^{(i)}), \cdots, (\boldsymbol{x}_{n_i}^{(i)}, y_{n_i}^{(i)})\}$ と書く．このとき，2 乗誤差関数を損失関数とするアンサンブル学習とは

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} \sum_{(\boldsymbol{x}, y) \in D_{n_i}} (y - f(\boldsymbol{x}; \boldsymbol{\theta}_i))^2 \qquad (1)$$

により与えられる

$$\bar{f}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}, \boldsymbol{\beta}) = \sum_{i=1}^M \beta_i f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i) \qquad (2)$$

を予測機械として用いることをいう．ただし

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

である．

本論文では上記の定式化を特別な場合として含む枠組みを用い，アンサンブル学習をパラメータ推定の立場から漸近論的に解析する．また，この枠組みに基づいてアンサンブル学習のアルゴリズム構造を検討する．

2. 準　　備

2.1 アンサンブル学習の別表現

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

$$p_G(y|g(\boldsymbol{x})) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y - g(\boldsymbol{x}))^2\right\}$$

を定めると（σ は正定数），式 (1) は

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} \left\{ - \sum_{(\boldsymbol{x}, y) \in D_{n_i}} \log p_G\left(y|f(\boldsymbol{x}; \boldsymbol{\theta}_i)\right) \right\} \qquad (3)$$

と書き換えられ，式 (2) は

$$p_G(y|\bar{f}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}, \boldsymbol{\beta})) = \frac{\prod_{i=1}^M p_G(y|f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i))^{\beta_i}}{\int_{\mathbb{R}} \prod_{i=1}^M p_G(y|f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i))^{\beta_i} dy} \qquad (4)$$

と恒等的に等しい [1].

本論文では上記の対応関係に注目し，負の対数ゆう度関数を損失関数とするアンサンブル学習の定式化と解析を行う．これは，2 乗誤差関数を損失関数とするアンサンブル学習を特別な場合として含む．次節では，そのための補題を用意する．

2.2 補題の準備

集合 $\boldsymbol{\mathcal{Z}}$ 上の確率分布全体を $\mathcal{P}(\boldsymbol{\mathcal{Z}})$ とおく．すなわち

$$\mathcal{P}(\boldsymbol{\mathcal{Z}}) \stackrel{\text{def}}{=} \Big\{ p | p : \boldsymbol{\mathcal{Z}} \to \mathbb{R}, p(\boldsymbol{z}) \geqq 0 \ (\forall \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}),$$

$$\int_{\mathcal{Z}} p(\boldsymbol{z})d\boldsymbol{z} = 1\Big\}$$

とおく．ここで，$p_i(\boldsymbol{z})$ $(\in \mathcal{P}_i(\mathcal{Z}) \subset \mathcal{P}(\mathcal{Z}))$ を用いて $(i = 1, \cdots, M)$，新たな確率分布 $\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})$ $(\in \mathcal{P}(\mathcal{Z}))$ を

$$\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i}}{\int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z}}$$

と定義する．ただし

$$\int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z} < \infty$$

$$\sum_{i=1}^{M} \beta_i = 1, \quad \beta_i > 0$$

$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)^T \in \mathbb{R}^M$$

とする．このとき，以下の補題が成立する．ただし，$D(\cdot\|\cdot)$ は Kullback ダイバージェンスを表す．

［補題 1］ 上の定義のもとで

$$D(p\|\bar{p}_{\boldsymbol{\beta}}) = \sum_{i=1}^{M} \beta_i D(p\|p_i) + \log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z}$$

が成り立つ． □

補題 1 から，仮に，添字 i によらず $D(p\|p_i)$ の値が同じであれば，$D(p\|\bar{p}_{\boldsymbol{\beta}})$ の $\boldsymbol{\beta}$ に関する最小化は

$$-\min_{\boldsymbol{\beta}} \log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z}$$

を求めることに相当することがわかる．この量は Chernoff 情報量と呼ばれる [1].

［補題 2］ 上の定義のもとで

$$-\log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z} = \sum_{i=1}^{M} \beta_i D(\bar{p}_{\boldsymbol{\beta}}\|p_i) \geqq 0$$

が成り立つ． □

また，次が知られている [1].

［補題 3］ 上の定義のもとで

$$\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}) = \arg \min_{p(\boldsymbol{z}) \in \mathcal{P}(\mathcal{Z})} \sum_{i=1}^{M} \beta_i D(p\|p_i)$$

が成り立つ．別表現として

$$\sum_{i=1}^{M} \beta_i D(\bar{p}_{\boldsymbol{\beta}}\|p_i) \leqq \sum_{i=1}^{M} \beta_i D(\tilde{p}\|p_i)$$

が成り立つ．ここで $\tilde{p}(\boldsymbol{z})$ は $\mathcal{P}(\mathcal{Z})$ の任意の要素を表す． □

3. 解　　析

3.1 問題設定

2.2 の仮定を満たす確率分布 $p_i(\boldsymbol{z}) \stackrel{\text{def}}{=} p(\boldsymbol{z};\boldsymbol{\theta}_i)$ $(\in \mathcal{P}_i(\mathcal{Z}) \subset \mathcal{P}(\mathcal{Z}))$ について $(i = 1, \cdots, M)$

$$\begin{aligned}\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}) &\stackrel{\text{def}}{=} \bar{p}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta}) \\ &\stackrel{\text{def}}{=} \frac{\prod_{i=1}^{M} p(\boldsymbol{z};\boldsymbol{\theta}_i)^{\beta_i}}{\int_{\mathcal{Z}} \prod_{i=1}^{M} p(\boldsymbol{z};\boldsymbol{\theta}_i)^{\beta_i} d\boldsymbol{z}} \end{aligned} \tag{5}$$

とおく．ただし

$$\boldsymbol{\theta}_i = (\theta_i^{(1)}, \cdots, \theta_i^{(k_i)})^T \in \mathbb{R}^{k_i}$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_M^T)^T \in \mathbb{R}^{\sum_{i=1}^{M} k_i}$$

であり，$p(\boldsymbol{z};\boldsymbol{\theta}_i)$ は $\boldsymbol{\theta}_i$ に関して高階微分可能とする．また，ある確率分布 $p_*(\boldsymbol{z})$ に従って互いに独立に $\sum_{i=1}^{M} n_i$ 個のサンプルが観測されたとし，これを $D_{n_i} = \{\boldsymbol{z}_1^{(i)}, \cdots, \boldsymbol{z}_{n_i}^{(i)}\}$ と書く．このとき，**2.1** に対応させて，負の対数ゆう度関数を損失関数とするアンサンブル学習を

$$\hat{\boldsymbol{\theta}}_i = \arg \min_{\boldsymbol{\theta}_i} \left\{ -\sum_{z \in D_{n_i}} \log p(\boldsymbol{z};\boldsymbol{\theta}_i) \right\} \tag{6}$$

によって与えられる $\bar{p}(\boldsymbol{z};\hat{\boldsymbol{\theta}},\boldsymbol{\beta})$ を求めることであると定義し，その汎化能力を平均予測誤差

$$E_{\hat{\theta}}[D(p_*(\boldsymbol{z})\|\bar{p}(\boldsymbol{z};\hat{\boldsymbol{\theta}},\boldsymbol{\beta}))] \tag{7}$$

で評価する．ただし，E は添字の従う確率分布に関する期待値を表す．また

$$\mathcal{P}_i(\mathcal{Z}) = \{p(\boldsymbol{z};\boldsymbol{\theta}_i) | \boldsymbol{\theta}_i \in \mathbb{R}^{k_i}\}$$

$$\mathcal{P}_1(\mathcal{Z}) \subset \cdots \subset \mathcal{P}_M(\mathcal{Z})$$

$$\exists \boldsymbol{\theta}_i^* \quad s.t. \quad p_*(\boldsymbol{z}) = p(\boldsymbol{z};\boldsymbol{\theta}_i^*) \in \mathcal{P}_i(\mathcal{Z})$$

とする．更に

$$\check{\boldsymbol{\Theta}}_i = \left\{ \boldsymbol{\theta} \middle| \text{ 正射影 } : \boldsymbol{\theta} \mapsto \boldsymbol{\theta}_i, \boldsymbol{\theta}_i \in \mathbb{R}^{k_i}, \boldsymbol{\theta} \in \mathbb{R}^{k_M} \right\}$$

とし，$\boldsymbol{\theta}_i \in \mathbb{R}^{k_i}$ と $\check{\boldsymbol{\theta}}_i \in \check{\boldsymbol{\Theta}}_i$ を同一視する．ただし，$\check{\boldsymbol{\theta}}_i$ のうち $\boldsymbol{\theta}_i$ と対応しない部分は $\boldsymbol{\theta}_j^*$ と一致するものと仮定する $(\forall j > i)$．以下では，記法の簡単のため $\check{\boldsymbol{\theta}}_i$ と $\boldsymbol{\theta}_i$ を区別せずに用いる．例えば，$\boldsymbol{\theta}_i + \boldsymbol{\theta}_j$ は $\check{\boldsymbol{\theta}}_i + \check{\boldsymbol{\theta}}_j$ を意味する $(i < j)$．また，この仮定のもとで $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}^*$ となる $(i = 1, \cdots, M)$.

3.2 平均予測誤差の上界と下界

式 (7) を直接解析することは困難であるため，その上界と下界を評価する．

［補題 4］ 式 (6) で表される $\hat{\boldsymbol{\theta}}_i$ について，n_i が十分大きいとき $(i=1,\cdots,M)$，適当な正則条件のもとで

$$E_{\hat{\theta}}[D(p(\boldsymbol{z};\tilde{\boldsymbol{\theta}})\|p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i))] = \frac{k_i}{2n_i} - \frac{k_i}{\sum_{l=1}^M n_l} + \frac{1}{2}\frac{\sum_{j=1}^M n_j k_j}{(\sum_{l=1}^M n_l)^2}$$

が成り立つ．ただし

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^M n_i \hat{\boldsymbol{\theta}}_i}{\sum_{i=1}^M n_i}$$

である． □

補題 1，2，3，4より，次が成り立つ．

［定理 1］ n_i が十分大きいとき $(i=1,\cdots,M)$，適当な正則条件のもとで

$$\frac{\sum_{i=1}^M \beta_i k_i}{\sum_{j=1}^M n_j} - \frac{1}{2}\frac{\sum_{i=1}^M n_i k_i}{(\sum_{j=1}^M n_j)^2} \leqq E_{\hat{\theta}}[D(p_*(\boldsymbol{z})\|\bar{p}(\boldsymbol{z};\hat{\boldsymbol{\theta}},\boldsymbol{\beta}))] \leqq \sum_{i=1}^M \frac{\beta_i k_i}{2n_i}$$

が成り立つ． □

［系 1］ $n=n_i$，$k=k_i$ のとき $(i=1,\cdots,M)$，n が十分大きければ，適当な正則条件のもとで

$$\frac{k}{2Mn} \leqq E_{\hat{\theta}}[D(p_*(\boldsymbol{z})\|\bar{p}(\boldsymbol{z};\hat{\boldsymbol{\theta}},\boldsymbol{\beta}))] \leqq \frac{k}{2n}$$

が成り立つ． □

3.3 考　　察

系 1 の下界は Mn 個のサンプルを用いたときの平均予測誤差，上界は n 個のサンプルを用いたときの平均予測誤差を表している [2]．したがって，Mn 個のサンプルが一括で与えられている場合には，アンサンブル学習は有効ではなく，n 個のサンプルを用いて学習した予測機械のみが M 個与えられている場合にはアンサンブル学習は有効であるということがわかる．この様子は [4] において実験的に示されていた．また，一般にアンサンブル学習の有効性の根拠は，補題 2 で与えられる量が正値となることにあることが容易にわかる．一方，新たなサンプル $D_{n'}=\{\boldsymbol{z}'_1,\cdots,\boldsymbol{z}'_{n'}\}$ が与えられた場合には

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{-\sum_{\boldsymbol{z}\in D_{n'}} \log\bar{p}(\boldsymbol{z};\hat{\boldsymbol{\theta}},\boldsymbol{\beta})\right\} \tag{8}$$

によって $\boldsymbol{\beta}$ を設定することができるが，本論文では，これに関する解析結果は与えられていない．

4. α-ダイバージェンスへの拡張

前節までに扱ってきたアンサンブル学習のアルゴリズムの本質は

$$\hat{p}_i(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg\min_{p(\boldsymbol{z})\in\mathcal{P}_i(\boldsymbol{\mathcal{Z}})} D(p_*\|p) \tag{9}$$

$$\bar{\hat{p}}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg\min_{p(\boldsymbol{z})\in\mathcal{P}(\boldsymbol{\mathcal{Z}})} \sum_{i=1}^M \beta_i D(p\|\hat{p}_i) \tag{10}$$

$$\hat{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\beta}} D(p_*\|\bar{\hat{p}}_{\boldsymbol{\beta}}) \tag{11}$$

と表される 3 段階の最小化操作に帰着される．なぜならば，式 (1)，(3)，(6) は式 (9) に，式 (2)，(4)，(5) は式 (10) に，更に，式 (8) は式 (11) に対応するからである．

ところで，最近では [5] に見られるように，学習問題における α-ダイバージェンスの重要性が論じられている．そこで本節では式 (9)，(10)，(11) の最小化操作が，α-ダイバージェンスを用いた場合に自然に拡張できることを明らかにする．ただし，α-ダイバージェンスは次のように定式化されている $(0<\alpha<1)$ [3]．

$$D^{(\alpha)}(p\|q) = \frac{1}{\alpha(1-\alpha)}\left\{1-\int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{z})^\alpha q(\boldsymbol{z})^{1-\alpha} d\boldsymbol{z}\right\}$$

まず，補題 3 は次のように拡張される．

［定理 2］ 上の定義のもとで

$$\bar{p}^{(\alpha)}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\}^{\frac{1}{1-\alpha}}}{\int_{\boldsymbol{\mathcal{Z}}}\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\}^{\frac{1}{1-\alpha}} d\boldsymbol{z}} = \arg\min_{p(\boldsymbol{z})\in\mathcal{P}(\boldsymbol{\mathcal{Z}})} \sum_{i=1}^M \beta_i D^{(\alpha)}(p\|p_i)$$

が成り立つ． □

［系 2］ 上の定義のもとで

$$\lim_{\alpha\to 1}\bar{p}^{(\alpha)}_{\boldsymbol{\beta}}(\boldsymbol{z}) = \bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})$$

$$\lim_{\alpha\to 0}\bar{p}^{(\alpha)}_{\boldsymbol{\beta}}(\boldsymbol{z}) = \sum_{i=1}^M \beta_i p_i(\boldsymbol{z})$$

が成り立つ． □

また，補題 1，2は次のように拡張される．

［定理 3］ 上の定義のもとで

$$C(\alpha)D^{(\alpha)}(p\|\bar{p}^{(\alpha)}_{\boldsymbol{\beta}}) = \sum_{i=1}^M \beta_i D^{(\alpha)}(p\|p_i) - \sum_{i=1}^M \beta_i D^{(\alpha)}(\bar{p}^{(\alpha)}_{\boldsymbol{\beta}}\|p_i)$$

が成り立つ．ただし

$$C(\alpha) \stackrel{\text{def}}{=} \sum_{i=1}^{M} \beta_i \int_{\mathcal{Z}} \{\bar{p}_\beta^{(\alpha)}(\boldsymbol{z})\}^\alpha \{p_i(\boldsymbol{z})\}^{1-\alpha} d\boldsymbol{z}$$

である．また

$$\lim_{\alpha\to 1} C(\alpha) = \lim_{\alpha\to 0} C(\alpha) = 1$$

が成り立つ． □

以上から，α-ダイバージェンスを用いたアンサンブル学習のアルゴリズムは次のように定義するのが自然である．

$$\hat{p}_i^{(\alpha)}(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg\min_{p(z)\in\mathcal{P}_i(\mathcal{Z})} D^{(\alpha)}(p_*\|p)$$

$$\bar{\hat{p}}_\beta^{(\alpha)}(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg\min_{p(z)\in\mathcal{P}(\mathcal{Z})} \sum_{i=1}^{M} \beta_i D^{(\alpha)}(p\|\hat{p}_i^{(\alpha)})$$

$$\hat{\boldsymbol{\beta}}^{(\alpha)} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\beta}} D^{(\alpha)}(p_*\|\bar{\hat{p}}_\beta^{(\alpha)})$$

5. むすび

本論文では，負の対数ゆう度関数を損失関数とするアンサンブル学習を定式化し，その平均予測誤差の上界と下界を求めた．この結果，すべてのサンプルが一括で与えられている場合には，アンサンブル学習は有効ではなく，学習済みの学習機械のみが与えられている場合にはアンサンブル学習は有効であるということがわかった．また，一般にアンサンブル学習の有効性の根拠は，補題 2 で与えられる量が正値となることにあるがわかった．これらの結果は 2 乗誤差関数を損失関数とする場合にも成立する．また，アンサンブル学習のアルゴリズムの本質が，式 (9)，(10)，(11) で表される，Kullback ダイバージェンスに関する 3 段階の最小化操作にあることから，これを α-ダイバージェンスを用いる場合に拡張した．今後の課題としては，$\boldsymbol{\beta}$ の推定についての考察や，これまでに筆者らが行った研究 [6] との関連性についての考察などが挙げられる．

文　献

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

[2] 坂本慶行，石黒真木夫，北川源四郎，情報量統計学，情報科学講座 A・5・4，共立出版，1983.

[3] 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.

[4] 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, pp.2512–2521, Sept. 1997.

[5] 松山泰男，“α-EM アルゴリズムとその基本的性質，” 信学論（D-I），vol.J82-D-I, no.12, pp.1347–1358, Dec. 1999.

[6] 内田真人，塩谷浩之，伊達　惇，“非ベイズ的付加項を用いた多層パーセプトロンの学習，” 信学論（D-II），vol.J83-D-II, no.6, pp.1572–1576, June 2000.

付　録

1. 補題 1 の証明

定義より

$$\begin{aligned} &D(p\|\bar{p}) \\ &= \int_{\mathcal{Z}} p(\boldsymbol{z}) \log \prod_{i=1}^{M} \left(\frac{p(\boldsymbol{z})}{p_i(\boldsymbol{z})}\right)^{\beta_i} d\boldsymbol{z} \\ &\quad + \int_{\mathcal{Z}} p(\boldsymbol{z}) \left\{ \log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z}')^{\beta_i} d\boldsymbol{z}' \right\} d\boldsymbol{z} \\ &= \sum_{i=1}^{M} \beta_i D(p\|p_i) + \log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z} \end{aligned}$$

が成立する．

2. 補題 2 の証明

定義より

$$\begin{aligned} &-\log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z} \\ &= -\int_{\mathcal{Z}} \bar{p}_\beta(\boldsymbol{z}) \left\{ \log \int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(\boldsymbol{z}')^{\beta_i} d\boldsymbol{z}' \right\} d\boldsymbol{z} \\ &= \int_{\mathcal{Z}} \bar{p}_\beta(\boldsymbol{z}) \log \prod_{i=1}^{M} \left(\frac{\bar{p}_\beta(\boldsymbol{z})}{p_i(\boldsymbol{z})}\right)^{\beta_i} d\boldsymbol{z} \\ &= \sum_{i=1}^{M} \beta_i D(\bar{p}_\beta\|p_i) \end{aligned}$$

が成立する．不等号は，$\boldsymbol{\beta}$ の定義と Kullback ダイバージェンスの非負性より明らか．

3. 補題 4 の証明

$-\log p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i)$ を $\tilde{\boldsymbol{\theta}}$ の周りで Taylor 展開し，高次の項を無視することで

$$\begin{aligned} &-\log p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i) \\ &= -\log p(\boldsymbol{z};\tilde{\boldsymbol{\theta}}) - (\hat{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}})^T \nabla \log p(\boldsymbol{z};\tilde{\boldsymbol{\theta}}) \\ &\quad - \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}})^T \nabla\nabla^T \log p(\boldsymbol{z};\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}) \end{aligned}$$

が得られる．ただし

$$\nabla = \left(\frac{\partial}{\partial\theta^{(1)}}, \cdots, \frac{\partial}{\partial\theta^{(k_M)}}\right)^T$$

であり，$\nabla\nabla^T \log p(\boldsymbol{z};\boldsymbol{\theta})$ は $\log p(\boldsymbol{z};\boldsymbol{\theta})$ の Hessian 行列 $(\partial^2 \log p(\boldsymbol{z};\boldsymbol{\theta})/\partial\theta^{(i)}\partial\theta^{(j)})$ を表す．よって

$$
\begin{aligned}
&D(p(\boldsymbol{z};\tilde{\boldsymbol{\theta}})\|p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i)) \\
&= \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}})^T G(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}) \\
&= \frac{1}{2}\left\{1 - \frac{n_i}{\sum_{l=1}^M n_l}\right\}^2 (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)^T G(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*) \\
&\quad + \frac{1}{2}\sum_{j=1,j\neq i}^M \left\{\frac{n_j}{\sum_{l=1}^M n_l}\right\}^2 \\
&\quad \times (\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*)^T G(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*) \\
&\quad + \sum_{i\neq j} N_{ij}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)^T G(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*) \quad \text{(A·1)}
\end{aligned}
$$

が成り立つ．ただし，$G(\boldsymbol{\theta})$ は Fisher 情報行列 $E_{\boldsymbol{z}}[\nabla \log p(\boldsymbol{z};\boldsymbol{\theta})\nabla \log p(\boldsymbol{z};\boldsymbol{\theta})^T]$ であり，N_{ij} は $\hat{\boldsymbol{\theta}}_i$, $\hat{\boldsymbol{\theta}}_j$ に依存しない定数である．また，各 n_i $(i = 1,\cdots,M)$ が十分大きいとき，適当な正則条件のもとで

$$
\hat{\boldsymbol{\theta}}_i \sim N\left(\boldsymbol{\theta}^*, \frac{1}{n_i} G(\boldsymbol{\theta}^*)^{-1}\right)
$$

であり [2]，かつ，各 $\hat{\boldsymbol{\theta}}_i$ $(i = 1,\cdots,M)$ は独立であることから

$$
\tilde{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}^*, \frac{1}{\sum_{i=1}^M n_i} G(\boldsymbol{\theta}^*)^{-1}\right)
$$

となる．そこで，式 (A·1) において高次の項を無視して $G(\tilde{\boldsymbol{\theta}}) = G(\boldsymbol{\theta}^*)$ と置き換える．以上から

$$
\begin{aligned}
&E_{\hat{\boldsymbol{\theta}}}[D(p(\boldsymbol{z};\tilde{\boldsymbol{\theta}})\|p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i))] \\
&= E_{\hat{\boldsymbol{\theta}}_1}\cdots E_{\hat{\boldsymbol{\theta}}_M}[D(p(\boldsymbol{z};\tilde{\boldsymbol{\theta}})\|p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i))] \\
&= \frac{1}{2}\left\{1 - \frac{n_i}{\sum_{l=1}^M n_l}\right\}^2 \frac{k_i}{n_i} \\
&\quad + \frac{1}{2}\sum_{j=1,j\neq i}^M \left\{\frac{n_j}{\sum_{l=1}^M n_l}\right\}^2 \frac{k_j}{n_j} \\
&= \frac{k_i}{2n_i} - \frac{k_i}{\sum_{l=1}^M n_l} + \frac{1}{2}\frac{\sum_{j=1}^M n_j k_j}{(\sum_{l=1}^M n_l)^2}
\end{aligned}
$$

が得られる．

4. 定理 1 の証明

n_i が十分大きいとき，適当な正則条件のもとで

$$
E_{\hat{\boldsymbol{\theta}}_i}[D(p_*(\boldsymbol{z})\|p(\boldsymbol{z};\hat{\boldsymbol{\theta}}_i))] = \frac{k_i}{2n_i}
$$

である $(i = 1,\cdots,M)$ [2]．これと，補題 2 より上界が示される．更に，補題 1，3，4を用いることで下界が示される．

5. 定理 2 の証明

Lagrange 乗数 λ を用いて汎関数 $L(q)$ を

$$
L(q) \stackrel{\text{def}}{=} \sum_{i=1}^M \beta_i D^{(\alpha)}(q\|p_i) + \lambda\int_{\mathcal{Z}} q(\boldsymbol{z})d\boldsymbol{z}
$$

と定義する．ただし，$q(\boldsymbol{z}) \in \mathcal{P}(\boldsymbol{\mathcal{Z}})$ である．このとき

$$
\begin{aligned}
\frac{\partial L(q)}{\partial q(\boldsymbol{z})} &= -\frac{q(\boldsymbol{z})^{\alpha-1}}{1-\alpha}\left\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{(1-\alpha)}\right\} + \lambda \\
&= 0
\end{aligned}
$$

を満たす $q(\boldsymbol{z})$ が求めるものである．ここで

$$
\int_{\mathcal{Z}} q(\boldsymbol{z})d\boldsymbol{z} = 1
$$

となるように λ を選ぶことで定理が得られる．

6. 定理 3 の証明

前半を証明する．まず

$$
\begin{aligned}
&\left[\int_{\mathcal{Z}}\left\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\right\}^{\frac{1}{1-\alpha}} d\boldsymbol{z}\right]^{1-\alpha} \\
&= \int_{\mathcal{Z}} \frac{\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\}^{\frac{1}{1-\alpha}}}{[\int_{\mathcal{Z}}\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z}')^{1-\alpha}\}^{\frac{1}{1-\alpha}} d\boldsymbol{z}']^{\alpha}} \\
&\quad \times\left\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\right\}^{\frac{\alpha-1}{1-\alpha}}\left\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\right\} d\boldsymbol{z} \\
&= \int_{\mathcal{Z}} \{\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(\boldsymbol{z})\}^{\alpha}\left\{\sum_{i=1}^M \beta_i p_i(\boldsymbol{z})^{1-\alpha}\right\} d\boldsymbol{z} \\
&= \sum_{i=1}^M \beta_i \int_{\mathcal{Z}} \{\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(\boldsymbol{z})\}^{\alpha}\{p_i(\boldsymbol{z})\}^{1-\alpha} d\boldsymbol{z} \\
&\stackrel{\text{def}}{=} C(\alpha) \qquad \text{(A·2)}
\end{aligned}
$$

が成り立つ．ここで，定義と式 (A·2) を用いることで

$$
\begin{aligned}
&C(\alpha)D^{(\alpha)}(p\|\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}) \\
&= \frac{1}{\alpha(1-\alpha)}\left\{1 - \int_{\mathcal{Z}} p(\boldsymbol{z})^{\alpha}\right.
\end{aligned}
$$

$$\times \left\{ \sum_{i=1}^{M} \beta_i p_i(\boldsymbol{z})^{1-\alpha} \right\} d\boldsymbol{z} \Bigg\}$$

$$- \frac{1}{\alpha(1-\alpha)} \{1 - C(\alpha)\}$$

$$= \sum_{i=1}^{M} \beta_i D^{(\alpha)}(p \| p_i) - \sum_{i=1}^{M} \beta_i D^{(\alpha)}(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)} \| p_i)$$

が示される．後半は明らか．

（平成 12 年 10 月 17 日受付，13 年 1 月 9 日再受付）