論　文

不定性を用いた分布間情報量の拡張形式に関する検討

内田　真人[†]　　塩谷　浩之[††]

A Study on an Extended Formula of Divergence Measures Using Invariance

Masato UCHIDA[†] and Hiroyuki SHIOYA[††]

あらまし　一般の凸関数を用いて定義される f-divergence は，分布間情報量の一般化表現型として用いられる．本論文では，f-divergence の不定性を用いることで，正値有限測度への適用を背景とした新しい分布間情報量のクラスを導く．これは，これまでよく用いられてきた分布間情報量とは特異なクラスの情報量であるが，統計的なデータ処理を容易にする情報量や，アンサンブル学習を陽に定式化して解析するのに適した情報量を含むことを明らかにすることで，適用対象に応じて分布間情報量を拡張することの有効性の一端を示す．

キーワード　f-divergence の不定性，α-divergence，正値有限測度，アンサンブル学習

1. まえがき

古くから知られる分布間情報量であるカルバック情報量 $D_K(\cdot\|\cdot)$ や α-divergence $D_\alpha(\cdot\|\cdot)$ を統一的に表現するクラスとして f-divergence $D_f(\cdot\|\cdot)$ がある [2]．これは，カルバック情報量のもつ統計的十分性を満たす正統的な一般化クラスとして知られている．

f-divergence の形式を定める凸関数 f に関して，見通しを良くする性質がある．それは，分布間情報量としての同値性の条件，すなわち，凸関数としては異なる f_1，f_2 について，ある条件が満たされた場合，$D_{f_1}(\cdot\|\cdot) = D_{f_2}(\cdot\|\cdot)$ が成り立つことである．[5] では，この性質を f-divergence の不定性と呼んでおり，それは，ある任意に与えられた凸関数を用いた f-divergence の関係不等式の簡易な証明に貢献している．

本研究では，まず，α-divergence を任意の正値有限測度に拡張された情報量にした表現形式（[4] の p.48）が，f-divergence の不定性で表されることに立ち戻る．そして，その視点と，差異が比較される確率密度の一対一写像による変換を用い，新しい分布間情報量を導入する．以上は，分布間情報量のある種の拡張ではあるが，広い表現形式を目的とするものではなく，通常のクラスの f-divergence では得られない性質をもつ分布間情報量を導き出すことを目的としている．この結果として，導入された分布間情報量の中には，サンプルの代入操作による簡便な学習を可能にするものが含まれていることや，適用範囲を分散が同一の 1 次元ガウス密度関数族に限定することで，[7] において定義された分布間情報量 $\mathcal{E}D_\alpha(\cdot\|\cdot)$ を自然に導出できることを示す．一方，サンプルの代入操作による学習を可能にする分布間情報量の特別な場合を用いることで，[8] と同様の枠組みに基づくアンサンブル学習を定式化し，その学習特性に関して議論する．

2. 準　　備

定義，基本性質について述べる．集合 $\boldsymbol{\mathcal{Z}}$ $(\subset \mathbb{R}^d)$ 上の確率密度関数全体を

$$\mathcal{P}(\boldsymbol{\mathcal{Z}}) \stackrel{\text{def}}{=} \left\{ p \;\middle|\; p : \boldsymbol{\mathcal{Z}} \to \mathbb{R}, p(\boldsymbol{z}) > 0 \ (\forall \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}), \int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{z}) d\boldsymbol{z} = 1 \right\}$$

とする．このとき，$\forall p, q \in \mathcal{P}(\boldsymbol{\mathcal{Z}})$ 間の f-divergence は，次のように定義される [2]．

$$D_f(p\|q) \stackrel{\text{def}}{=} \int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{z}) f\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right) d\boldsymbol{z}$$

ただし，$f(u)$ は $(0, \infty)$ 上で定義される凸関数で，

[†] 日本電信電話株式会社　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, NTT Corporation, Musashino-shi, 180–8585 Japan
[††] 室蘭工業大学，室蘭市
Muroran Institute of Technology, Muroran-shi, 050–8585 Japan

$f(1)=0$ を満たし，$u=1$ で狭義凸である．

また，ある二つの凸関数 $\tilde{f}$ と f について，$\tilde{f}(u)=f(u)+c(1-u)$ $(\forall c\in\mathbb{R})$ を満たすならば，$D_f(p\|q)=D_{\tilde{f}}(p\|q)$ が成り立つ．これを f-divergence の不定性と呼ぶ．特に，$c\in[\partial f^-(1),\partial f^+(1)]$ とおくと $\tilde{f}(u)\geq 0$（等号は $u=1$ のときのみ）であることから，$D_{\tilde{f}}(p\|q)\geq 0$ となることが分かる（等号は $p=q$ のときのみ）．ただし $\partial f^-(u)$，$\partial f^+(u)$ は左右の劣微係数とする．

このことに注目すると，$c\in[\partial f^-(1),\partial f^+(1)]$ としたときに，正値有限測度全体

$$\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})\stackrel{\text{def}}{=}\left\{p\;\middle|\;p:\boldsymbol{\mathcal{Z}}\to\mathbb{R},p(\boldsymbol{z})>0\ (\forall\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}),\ \int_{\boldsymbol{\mathcal{Z}}}p(\boldsymbol{z})d\boldsymbol{z}<\infty\right\}$$

について f-divergence を拡張できる．すなわち，$\forall\tilde{p},\tilde{q}\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ について $D_{\tilde{f}}(\tilde{p}\|\tilde{q})\geq 0$ となり，その等号が $\tilde{p}\equiv\tilde{q}$ のときかつそのときに限り成立する．ただし，この場合は不定性が必ずしも成立しない．

同様に，$\mathbb{R}_+\stackrel{\text{def}}{=}\{x|x>0,x\in\mathbb{R}\}$ 上の一対一写像 g を用いれば，$D_{\tilde{f}}(g\circ p\|g\circ q)\geq 0$ となり，その等号が $p\equiv q$ のときかつそのときに限り成立する．ただし，$p,q\in\mathcal{P}(\boldsymbol{\mathcal{Z}})$ かつ $g\circ p,g\circ q\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ とした．

本論文では，以上の考察に基づき，いくつかの新しい分布間情報量を導出し，その性質について考察する．

3. α-divergence の拡張

2. では，f-divergence の不定性を利用することで，通常の f-divergence のクラスには含まれない分布間情報量が導出できることを示した．そこで本章では，まず，**2.** の視点に基づき α-divergence を直接的に拡張し，α-divergence との関係が明確な新たな分布間情報量 (γ,β)-divergence を導出する (**3. 1**)．次に，(γ,β)-divergence の $\beta=1+\gamma$ の場合である $(\gamma,1+\gamma)$-divergence について，[7] で定義されている分布間情報量 $\mathcal{E}D_\gamma(p\|q)$ との関係や，学習問題における位置付けに関して考察する (**3. 2**)．最後に，f-divergence の不定性を利用して α-divergence を拡張することで見通し良く得られるいくつかの関係式を示す (**3. 3**)．

3. 1 (γ,β)-divergence

α-divergence は

$$f_\alpha(u)=\frac{1}{1-\alpha}\left\{\frac{1}{\alpha}(1-u^\alpha)-(1-u)\right\}\qquad(1)$$

を用いて

$$\begin{aligned}&D_\alpha(p\|q)\\&\stackrel{\text{def}}{=}\int_{\boldsymbol{\mathcal{Z}}}p(\boldsymbol{z})f_\alpha\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right)d\boldsymbol{z}\\&=\frac{1}{\alpha(1-\alpha)}\int_{\boldsymbol{\mathcal{Z}}}\Big\{(1-\alpha)p(\boldsymbol{z})+\alpha q(\boldsymbol{z})\\&\quad-p(\boldsymbol{z})^{1-a}q(\boldsymbol{z})^\alpha\Big\}d\boldsymbol{z}\end{aligned}$$

と定義される [4]．ここで，$\alpha\in(0,1)$ とする．このとき，$\alpha\to 0$ とすることでカルバック情報量 $D_K(\cdot\|\cdot)$ が得られる [4]．ただし

$$D_K(p\|q)\stackrel{\text{def}}{=}\int_{\boldsymbol{\mathcal{Z}}}\left\{q(\boldsymbol{z})-p(\boldsymbol{z})+p(\boldsymbol{z})\log\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right\}d\boldsymbol{z}$$

である $(\forall p,q\in\mathcal{P}(\boldsymbol{\mathcal{Z}}))$．なお，既述のように，$f$-divergence の不定性に基づき，$f_\alpha(u)$ とは異なる凸関数を用いて α-divergence を定義することが可能であるが，ここでは後の議論の簡単のため，あらかじめ $f'_\alpha(1)=0$ となるように $f_\alpha(u)$ を設定した．

次に

$$g_\beta(u)\stackrel{\text{def}}{=}u^\beta,\quad\beta\neq 0$$

を用いて，α-divergence を

$$\begin{aligned}&D_\alpha(g_\beta\circ p\|g_\beta\circ q)\\&=\frac{1}{\alpha(1-\alpha)}\int_{\boldsymbol{\mathcal{Z}}}\Big\{(1-\alpha)p(\boldsymbol{z})^\beta+\alpha q(\boldsymbol{z})^\beta\\&\quad-p(\boldsymbol{z})^{(1-\alpha)\beta}q(\boldsymbol{z})^{\alpha\beta}\Big\}d\boldsymbol{z}\end{aligned}$$

と拡張する．このとき，$g_\beta(u)$ が $\mathbb{R}_+$ 上の一対一写像であることから

$$\begin{aligned}&D_\alpha(g_\beta\circ p\|g_\beta\circ q)\geq 0,\quad\forall p,q\in\mathcal{P}(\boldsymbol{\mathcal{Z}})\\&D_\alpha(g_\beta\circ p\|g_\beta\circ q)=0\quad\text{iff}\quad p\equiv q\end{aligned}$$

が成り立つ．

更に，後の議論の簡単のため $\gamma=\alpha\beta$ とおき，$D_\alpha(g_\beta\circ p\|g_\beta\circ q)$ から α を消去すると

$$D_{\frac{\gamma}{\beta}}(g_\beta\circ p\|g_\beta\circ q)$$

$$=\frac{\beta^2}{\beta-\gamma}\int_{\boldsymbol{Z}}\left\{\frac{1}{\gamma}(p(\boldsymbol{z})^{\beta}-p(\boldsymbol{z})^{\beta-\gamma}q(\boldsymbol{z})^{\gamma})-\frac{1}{\beta}(p(\boldsymbol{z})^{\beta}-q(\boldsymbol{z})^{\beta})\right\}d\boldsymbol{z}$$

となる．ただし，γ の定義より $\frac{\gamma}{\beta}\in(0,1)$ である．本論文では，$\frac{1}{\beta^2}D_{\frac{\gamma}{\beta}}(g_\beta\circ p\|g_\beta\circ q)$ を，(γ,β)-divergence と呼び $D_{\gamma,\beta}(p\|q)$ と書く．定義から明らかに，$\beta=1$ の場合が γ-divergence に相当することになる．

3.2 $(\gamma,1+\gamma)$-divergence

$(\gamma,1+\gamma)$-divergence は

$$D_{\gamma,1+\gamma}(p\|q)=\int_{\boldsymbol{Z}}\left\{\frac{1}{\gamma}p(\boldsymbol{z})(p(\boldsymbol{z})^{\gamma}-q(\boldsymbol{z})^{\gamma})-\frac{1}{1+\gamma}(p(\boldsymbol{z})^{1+\gamma}-q(\boldsymbol{z})^{1+\gamma})\right\}d\boldsymbol{z}$$

と与えられる．ただし，$\gamma>0$ とする．これは，[3] において，本論文とは異なる形式に基づき定義されており，独立成分分析へ応用されている．

ここで，分散が同一の 1 次元ガウス密度関数族を $\mathcal{G}$ とおく．このとき，$\forall p,q\in\mathcal{G}$ について

$$\int_{\mathbb{R}}p(x)^{1+\gamma}dx=\int_{\mathbb{R}}q(x)^{1+\gamma}dx,\quad\gamma>0$$

となる．したがって，$\forall p,q\in\mathcal{G}$ について

$$D_{\gamma,1+\gamma}(p\|q)=\frac{1}{\gamma}\int_{\mathbb{R}}\{p(x)(p(x)^{\gamma}-q(x)^{\gamma})\}dx$$

が成り立つ．これは，[7] において定義された情報量 $\mathcal{E}D_\gamma(p\|q)$ に定数倍を除いて一致する．しかし，$\mathcal{E}D_\gamma(\cdot\|\cdot)$ は $\mathcal{G}$ に限定的な分布間情報量として定義されており，$\mathbb{R}$ 上の確率密度関数全体に関しては必ずしも分布間情報量としての性質を満たさない．そのため，この意味において $(\gamma,1+\gamma)$-divergence は $\mathcal{E}D_\gamma(\cdot\|\cdot)$ の一般化とみなすことができる．

また，統計モデル $\mathcal{Q}(\boldsymbol{Z})$ $(\subset\mathcal{P}(\boldsymbol{Z}))$ として，q の汎関数

$$C(q)\stackrel{\text{def}}{=}\int_{\boldsymbol{Z}}q(\boldsymbol{z})^{1+\gamma}d\boldsymbol{z},\quad q\in\mathcal{Q}(\boldsymbol{Z})$$

の値が解析的に求められるものを用いた場合，すなわち，$\forall q\in\mathcal{Q}(\boldsymbol{Z})$ について，$C(q)$ の計算値が $\mathcal{Q}(\boldsymbol{Z})$ のもつモデルパラメータに対する依存の有無にかかわらず明示的に求められる場合，[7] と同様の議論によりサンプルの代入操作に基づく簡便な学習アルゴリズムを構成できる．このことについて以下で説明する(注1)．

まず

$$\begin{aligned}\hat{q}(\boldsymbol{z})&\stackrel{\text{def}}{=}\arg\min_{q\in\mathcal{Q}(\boldsymbol{Z})}D_{\gamma,1+\gamma}(p_*\|q)\\&=\arg\min_{q\in\mathcal{Q}(\boldsymbol{Z})}\int_{\boldsymbol{Z}}\left\{-\frac{1}{\gamma}p_*(\boldsymbol{z})q(\boldsymbol{z})^{\gamma}+\frac{1}{1+\gamma}q(\boldsymbol{z})^{1+\gamma}\right\}d\boldsymbol{z}\\&=\arg\min_{q\in\mathcal{Q}(\boldsymbol{Z})}\left\{-\frac{1}{\gamma}\mathbb{E}_{p_*}[q(\boldsymbol{Z})]+\frac{1}{1+\gamma}C(q)\right\}\end{aligned}$$

に注意する．ただし，p_* は推定の対象となる確率密度関数，$\hat{q}$ は推定された確率密度関数，$\mathbb{E}_p$ は p に関する期待値を表す．推定の対象となる確率密度関数 p_* が未知の場合，$\mathbb{E}_{p_*}[q(\boldsymbol{Z})]$ は計算不可能であるが，以下では，p_* からの i.i.d (independently and identically distributed) データ系列により定まる経験確率密度関数 p_e が得られているという標準的な問題設定を考える．このとき，p_* に関する期待値の代わりに p_e に関する期待値を代用すると，上式は

$$\hat{p}_i(\boldsymbol{z})=\arg\min_{q\in\mathcal{Q}(\boldsymbol{Z})}\left\{-\frac{1}{\gamma}\mathbb{E}_{p_e}[q(\boldsymbol{Z})]+\frac{1}{1+\gamma}C(q)\right\}\tag{2}$$

と置き換えられ，サンプルの代入操作による簡便な学習アルゴリズムを構成できる(注2)．なお，実用上，この学習アルゴリズムを実行する際は，モデルパラメータに関するこう配法などを用いることになるが，この場合は必ずしも式 (2) における最小解が得られるとは限らない．

(注1)：この説明は，$(\gamma,1+\gamma)$-divergence が $\mathcal{E}D_\gamma(\cdot\|\cdot)$ を特別な場合として含むということから，[7] の議論と重複する部分があるが，$\mathcal{E}D_\gamma(\cdot\|\cdot)$ が分散が同一の 1 次元ガウス密度関数族に対してのみ適用可能なことに対し，$(\gamma,1+\gamma)$-divergence は任意の確率密度関数に対して適用可能であるという差異がある．本論文では，この特徴を用いることで，$(\gamma,1+\gamma)$-divergence の $\gamma=1$ の場合である $(1,2)$-divergence を用いたアンサンブル学習の定式化を行う（**4.** を参照）．$(1,2)$-divergence に基づくアンサンブル学習におけるアンサンブル学習機械は混合確率密度関数で表されるため（**4.** の系を参照），$\mathcal{E}D_\gamma(\cdot\|\cdot)$ を用いた定式化はできない．

(注2)：この性質は，一般の divergence については必ずしも成り立たない．例えば，$D_\alpha(p_*\|q)$ の場合，$\int_{\boldsymbol{Z}}p_*(\boldsymbol{z})^{1-\alpha}q(\boldsymbol{z})^{\alpha}d\boldsymbol{z}$ という項が含まれているので，この項を計算するために，平滑化手法等を用いて，与えられたサンプルから p_* に相当する確率密度関数 $\hat{p}_*$ を生成する必要がある．一方，$D_{\gamma,1+\gamma}(p_*\|q)$ の場合には，$\int_{\boldsymbol{Z}}p_*(\boldsymbol{z})q(\boldsymbol{z})^{\gamma}d\boldsymbol{z}$ という項が含まれているが，これは p_* に関する期待値の形式をとっているため，単純に p_* を p_e に置き換えることができる．

上記の例として，$\mathcal{Q}(\boldsymbol{\mathcal{Z}})$ を分散可変の 1 次元ガウス密度関数族とした場合が挙げられる．この場合，$C(q)$ は分散パラメータに依存する関数として明示的に求められるため，簡便な学習アルゴリズムを構成できる．なお，分散が定数の 1 次元ガウス密度関数族を用いた場合は $C(q)$ の値が定数となるため，$D_{\gamma,1+\gamma}(p\|q)$ の $\forall q \in \mathcal{Q}(\boldsymbol{\mathcal{Z}})$ に対する最小化と $\mathcal{E}D_{\gamma}(p\|q)$ の $\forall q \in \mathcal{Q}(\boldsymbol{\mathcal{Z}})$ に対する最小化は同じ意味をもつ．

3.3 (γ,β)-divergence に関する関係式

本節では，f-divergence の不定性を利用するという新しい視点に基づき α-divergence を拡張することで見通し良く得られるいくつかの命題を示す．

次の命題は，$D_{\alpha}(\cdot\|\cdot)$ と $D_{1-\alpha}(\cdot\|\cdot)$ の関係 [4] を利用する事で導かれる(注3)（証明については付録を参照）．

［命題 1］ $\forall p,q \in \mathcal{P}(\boldsymbol{\mathcal{Z}})$ について

$$D_{\gamma,\beta}(p\|q) = D_{\beta-\gamma,\beta}(q\|p)$$

が成り立つ． □

次の命題は，$(\gamma,1)$-divergence（すなわち，γ-divergence）と $(\gamma,1+\gamma)$-divergence のいずれもがカルバック情報量を特別な場合として含むことを示している．証明は容易なので省略する．

［命題 2］ $\forall p,q \in \mathcal{P}(\boldsymbol{\mathcal{Z}})$ について

$$\lim_{\gamma\to 0} D_{\gamma,1}(p\|q) = \lim_{\gamma\to 0} D_{\gamma,1+\gamma}(p\|q) = D_K(p\|q)$$

が成り立つ． □

ところで，(γ,β)-divergence は f-divergence の不定性を用いずに，ある特別な不等式を用いることでも定義することができる．このことにより，(γ,β)-divergence の非負性や等号条件を成立させる背景を別の視点から与えることが可能となる．また，この不等式を用いることで，(γ,β)-divergence に関する定量的な評価を可能とする様々な関係式や，α-divergence の差により表現される分布間情報量を導くことができる．しかし，本論文では，f-divergence の不定性に基づいた議論を主題としているため，(γ,β)-divergence の別定義に関する議論は付録に記載する．

4. $(1,2)$-divergence に基づくアンサンブル学習への応用

複数の学習機械を統合し，最終的な予測結果を得る手法はアンサンブル学習と総称される．学習機械の統合方法としては，単純平均や重み付き平均により統合するという方法 [9] が知られているが，最近では，学習の手続きに関する様々な設定方法が提案されている(注4)．一方，単純平均や重み付き平均によるアンサンブル学習については，確率モデルを適当に設定することで，カルバック情報量に関する 3 段階の最小化操作に書き下せることや，その枠組みが α-divergence を用いた場合にも拡張可能であることが示されている [8]．そこで本章では，[8] と同様の考え方に基づき $(1,2)$-divergence

$$\begin{aligned} &D_{1,2}(p\|q) \\ &= \int_{\boldsymbol{\mathcal{Z}}} \left\{ (p(\boldsymbol{z})^2 - p(\boldsymbol{z})q(\boldsymbol{z})) \right. \\ &\qquad \left. -\frac{1}{2}(p(\boldsymbol{z})^2 - q(\boldsymbol{z})^2) \right\} d\boldsymbol{z} \\ &= \frac{1}{2}\int_{\boldsymbol{\mathcal{Z}}} (p(\boldsymbol{z}) - q(\boldsymbol{z}))^2 d\boldsymbol{z}, \qquad p,q \in \mathcal{P}(\boldsymbol{\mathcal{Z}}) \end{aligned}$$

を用いたアンサンブル学習を定式化する．

この $(1,2)$-divergence は，分布間情報量同士について成り立つ関係不等式（例えば，$D_{1,2} \leq D_v$，$D_v \stackrel{\text{def}}{=} D_{f(u)=|u-1|}$）において扱われているが，学習における有効な扱いは，それほどなされていない．なお，自明なことであるが，$(1,2)$-divergence は f-divergence のクラスには含まれない．

まず，新たな確率密度関数 $\bar{p}^{1,2}_{\boldsymbol{\beta}}$ $(\in \mathcal{P}(\boldsymbol{\mathcal{Z}}))$ を，p_i $(\in \mathcal{P}_i(\boldsymbol{\mathcal{Z}}) \subset \mathcal{P}(\boldsymbol{\mathcal{Z}}))$ を用いて

$$\bar{p}^{1,2}_{\boldsymbol{\beta}}(\boldsymbol{z}) = \sum_{i=1}^{M} \beta_i p_i(\boldsymbol{z})$$

と定義する $(i = 1,\cdots,M)$．ただし

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

(注3)：命題 1 は $(\gamma,1+\gamma)$-divergence を含む一般の場合において成り立つ関係式となっている．既述のように (γ,β)-divergence の $\beta = 1+\gamma$ の場合である $(\gamma,1+\gamma)$-divergence は，文献 [3] において本論文とは全く異なる形式から定義されているが，この定義は α-divergence の性質を利用したものとはなっていないため，$D_{\alpha}(\cdot\|\cdot)$ と $D_{1-\alpha}(\cdot\|\cdot)$ の関係を利用して命題 1 を直接的に導くことができない．

(注4)：データのリサンプリング方法，統合する学習機械の学習方法，学習機械の統合方法等に改良が施された様々なアンサンブル学習が提案されている．こうした中で提案されている Boosting [10] や Bagging [11] に比べ，個々に学習された複数の学習機械の予測結果を単純平均・重み付き平均により統合するというアンサンブル学習 [8], [9] は単純ではあるが，様々に提案されているアンサンブル学習の共通的な特徴（複数の学習機械を統合し，最終的な予測結果を得るという特徴）を反映したものと考えられる．ただし，本論文で扱うアンサンブル学習の枠組みと Boosting や Bagging とのかかわりについては今後の課題とする．

$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)^T$$

である．すなわち，$\bar{p}^{1,2}_{\boldsymbol{\beta}}$ は混合確率密度関数である．このとき，$(1,2)$-divergence と混合確率密度関数 $\bar{p}^{1,2}_{\boldsymbol{\beta}}$ との間には以下の関係が成り立つ（証明については付録を参照）．

［補題 1］ 次が成り立つ．

$$D_{1,2}(p\|\bar{p}^{1,2}_{\boldsymbol{\beta}}) = \sum_{i=1}^{M} \beta_i D_{1,2}(p\|p_i) - \sum_{i=1}^{M} \beta_i D_{1,2}(\bar{p}^{1,2}_{\boldsymbol{\beta}}\|p_i)$$

□

また，補題 1 の右辺の第 2 項が p に依存しないことから次が得られる．

［系 ］ 次が成り立つ．

$$\bar{p}^{1,2}_{\boldsymbol{\beta}}(\boldsymbol{z}) = \arg \min_{p(\boldsymbol{z})\in\mathcal{P}(\boldsymbol{\mathcal{Z}})} \sum_{i=1}^{M} \beta_i D_{1,2}(p\|p_i)$$

□

ここで，$(1,2)$-divergence を用いたアンサンブル学習を [8] と同様にして

$$\hat{p}_i(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg \min_{p(\boldsymbol{z})\in\mathcal{P}_i(\boldsymbol{\mathcal{Z}})} D_{1,2}(p_*\|p) \tag{3}$$

$$\bar{\hat{p}}^{1,2}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg \min_{p(\boldsymbol{z})\in\mathcal{P}(\boldsymbol{\mathcal{Z}})} \sum_{i=1}^{M} \beta_i D_{1,2}(p\|\hat{p}_i) \tag{4}$$

$$\hat{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\beta}} D_{1,2}(p_*\|\bar{\hat{p}}^{1,2}_{\boldsymbol{\beta}}) \tag{5}$$

と定義する(注5)．

ただし，p_* は推定の対象となる真の確率密度関数，各 $\hat{p}_i$ $(i = 1, \cdots, M)$ は事前に推定された確率密度関数（個々の学習機械），$\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}$ は最終的に推定された確率密度関数（アンサンブル学習機械）である．以上より，$(1,2)$-divergence に基づくアンサンブル学習においては，混合確率密度関数を用いて学習対象をモデル化するということが分かる．なお，式 (4) は上記の系に相当するため定義可能 (well-defined) である．

また，補題 1 を用いることで次が得られる（証明については付録を参照）．

［定理 ］ 次が成り立つ $(i = 1, \cdots, M)$．

$$D_{1,2}(p_*\|\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}) = D_{1,2}(p_*\|p_i) - D_{1,2}(\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}\|p_i) \tag{6}$$

□

上記の定理は，アンサンブル学習機械 $\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}$ の予測精度 $D_{1,2}(p_*\|\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}})$ が，個々の学習機械 $\hat{p}_i$ の予測精度 $D_{1,2}(p_*\|p_i)$ と比較して $(i = 1, \cdots, M)$，どの程度改善されるかを示している．改善の量は右辺の第 2 項 $D_{1,2}(\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}\|p_i)$ で与えられ，かつ，この量はアンサンブル学習を構成するために用意された個々の学習機械と，重み $\hat{\boldsymbol{\beta}}$ により定まるため，p_* を直接的に利用することなく評価可能となっている．更に，一般性を失うことなく

$$D_{1,2}(p_*\|p_M) = \arg \min_{i=1,\cdots,M} D_{1,2}(p_*\|p_i)$$

とすれば，アンサンブル学習機械の予測精度 $D_{1,2}(p_*\|\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}})$ は，アンサンブル学習を構成するために用意された個々の学習機械の中で最良の学習機械の予測精度 $D_{1,2}(p_*\|p_M)$ よりも $D_{1,2}(\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}\|p_M)$ だけ改善されることが分かる．以上のように，上記の定理は，式 (3)，(4)，(5) により定義されるアンサンブル学習の学習結果に関する定量的な評価となっている．

また，$\bar{\hat{p}}^{1,2}_{\hat{\boldsymbol{\beta}}}$ を混合型 1 次元ガウス密度関数とした場合，すなわち，各 $\mathcal{P}_i(\boldsymbol{\mathcal{Z}})$ $(i = 1, \cdots, M)$ を 1 次元ガウス密度関数族（分散が同一とは限らない）とした場合

$$\int_{\mathbb{R}} (\bar{p}^{1,2}_{\boldsymbol{\beta}}(x))^2 dx$$

の値は解析的に求められる．したがって，$(1,2)$-divergence を用いたアンサンブル学習はサンプルの代入操作に基づく簡便な学習アルゴリズムにより実現することができる．このことに関する詳細については **3. 2** を参照されたい．また，式 (5) においても，p_* を p_e で置き換えることで $\hat{\boldsymbol{\beta}}$ を導出することができる．

ところで

$$D_{dK}(p\|q) \stackrel{\text{def}}{=} D_K(q\|p)$$

と定義される D_{dK} を用いたアンサンブル学習（ [8] の **4.** における $\alpha \to 0$ の場合）で利用される確率密度関

(注5)：入出力機械の学習問題は条件付き確率密度関数の学習に帰着できる場合があることが知られている．しかし，入力の確率密度関数が機械のパラメータに依存しない場合，条件付き確率密度関数の学習は入力と出力に関する同時確率密度関数の学習と等価になる．そのため，本論文では一般の確率密度関数を用いて $(1,2)$-divergence に基づくアンサンブル学習を定式化している．なお, [8] においても，こうした理由から同時確率密度関数に基づくアンサンブル学習の定式化を行っている．ただし, [8] の式 (4) に記述されているように，本論文や [8] の枠組みは条件付き確率密度関数の学習にも適用可能である．

数は $\bar{p}_{\boldsymbol{\beta}}^{1,2}$ に一致している．しかし，D_{dK} を用いたアンサンブル学習の場合には，サンプルの代入操作に基づく学習アルゴリズムが構成できないことから実用の観点における有効性はないといえる．

5. むすび

本論文では，f-divergence の不定性に着目して α-divergence を拡張することで，新たな分布間情報量 (γ,β)-divergence を導き，いくつかの考察を加えた．その結果，適用範囲を分散が同一の 1 次元ガウス密度関数族に限定した場合，$(\alpha, 1+\alpha)$-divergence は，[7] において導入された情報量 $\mathcal{E}D_\alpha(\cdot\|\cdot)$ に一致することが分かった．また，$(\alpha, 1+\alpha)$-divergence は $\mathcal{E}D_\alpha(\cdot\|\cdot)$ と同様にして，サンプルの代入操作に基づく簡便な学習アルゴリズムを構成することが可能であることが分かった．このことは，[7] で扱われている学習問題において利用可能となる確率モデルの設定の幅が広がったことを意味する．そのため，今後，様々な確率モデルに対する本研究での拡張の有効性，及び学習機構を有するシステムへの実装などが期待される．

また，補題 2 により与えられる $h_{\gamma,\beta}(u)$ に関する不等式を用いることで (γ,β)-divergence を別の視点から導出した．更に，この不等式を用いることで，(γ,β)-divergence について成り立ついくつかの関係式や，α-divergence の差として表現される分布間情報量を導いた．

一方で，$(1,2)$-divergence を用いたアンサンブル学習を定式化し，その学習特性に関する定量的評価と計算手続きに関する定性的評価を行った．

文　献

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

[2] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” Studia Sci. Math. Hunger., vol.2, pp.299–318, 1967.

[3] M. Minami and S. Eguchi, “Robust blind source separation by beta divergence,” Neural Comput., vol.14, no.8, pp.1859–1886, 2002.

[4] 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.

[5] 塩谷浩之，長岡浩司，伊達　惇，“f-divergence に関する新しい不等式と最大値および学習問題への応用，” 信学論 (A), vol.J77-A, no.4, pp.720–726, April 1994.

[6] 塩谷浩之，佐藤佳久，伊達　惇，“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.

[7] 塩谷浩之，伊達　惇，“α-ダイバージェンスを利用した一般化された 2 乗誤差最小学習，” 信学論 (D-II), vol.J84-D-II, no.12, pp.2690–2695, Dec. 2001.

[8] 内田真人，塩谷浩之，伊達　惇，“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, July 2001.

[9] 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.

[10] Y. Freund and R.E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Comput. Syst. Sci., vol.55, no.1, pp.119–139, 1997.

[11] L. Breiman, “Bagging predictors,” Machine Learning, vol.24, pp.123–140, 1996.

付　録

1. (γ,β)-divergence

本章では，f-divergence の不定性を用いずに，ある特別な不等式を用いることで (γ,β)-divergence を導出する．また，この不等式に関する考察を加える．

1.1 不等式を用いた (γ,β)-divergence の導入

次の不等式が成り立つ．証明は容易なので省略する．

［補題 2］ $(\alpha,\beta)\in\mathbb{R}^2\backslash\{(x,y)|x=y \text{ or } x=0 \text{ or } y=0\ ;\ x,y\in\mathbb{R}\}$，$u\in\{x|x\in\mathbb{R}, x\geq 0\}$ について

$$\frac{1}{\beta-\alpha}\left\{\frac{1}{\alpha}(1-u^\alpha)-\frac{1}{\beta}(1-u^\beta)\right\}\geq 0 \qquad \text{(A·1)}$$

が成り立つ．ただし，等号は $u=1$ のとき，かつそのときに限り成立する． □

式 (A·1) の左辺を $h_{\gamma,\beta}(u)$ とおく．このとき，$\forall p,q\in\mathcal{P}(\boldsymbol{\mathcal{Z}})$ と $\frac{\gamma}{\beta}\in(0,1)$ に対して

$$D_{\gamma,\beta}(p\|q)=\int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{z})^\beta h_{\gamma,\beta}\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right)d\boldsymbol{z}$$

が成り立つ．また

$$D_{\gamma,\beta}(p\|q)\geq 0,\quad \forall p,q\in\mathcal{P}(\boldsymbol{\mathcal{Z}})$$
$$D_{\gamma,\beta}(p\|q)=0 \quad \text{iff} \quad p\equiv q$$

が成り立つことは，式 (A·1) によっても理解できる．

式 (A·1) は $\frac{\gamma}{\beta}\in(0,1)$ 以外の (γ,β) についても成立する．したがって，(γ,β)-divergence における (γ,β) の定義域を拡張することが可能であるが，その場合は α-divergence における $\alpha\in(0,1)$ の条件との関連性が不明確になる．

1.2 $h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(u)$ の性質

まず

$$h_{\alpha,\beta}(u)=\frac{1}{\beta-\alpha}\big\{(1-\alpha)f_\alpha(u)-(1-\beta)f_\beta(u)\big\}$$

が成り立つことに注意する（式 (1) 参照）．このことと，補題 1 より次が成り立つ．証明は容易なので省略する．

［命題 3］ $\forall p,q\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ について

$$\begin{aligned}&\int_{\boldsymbol{\mathcal{Z}}}p(\boldsymbol{z})h_{\alpha,\beta}\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right)d\boldsymbol{z}\\&=\frac{1}{\beta-\alpha}\Big\{(1-\alpha)D_\alpha(p\|q)-(1-\beta)D_\beta(p\|q)\Big\}\\&\geq 0\end{aligned}$$

が成り立つ．ただし，$\alpha,\beta\in(0,1)$ かつ $\alpha\neq\beta$ である．また，不等式の等号は $p\equiv q$ のとき，かつそのときに限り成り立つ． □

この性質は，α-divergence の差を用いて新たな分布間情報量の構成が可能であることを意味する点において興味深い．また，この結果を用いることで，$\alpha>\beta$ のとき $\forall p,q\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ について

$$(1-\alpha)D_\alpha(p\|q)\leq(1-\beta)D_\beta(p\|q)$$

を導くことができる [6]．更に，次のような拡張が可能である．証明は容易なので省略する．

［命題 4］ $\forall p,q\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ について

$$\begin{aligned}&\frac{1}{\gamma^2}\int_{\boldsymbol{\mathcal{Z}}}p(\boldsymbol{z})^\gamma h_{\alpha,\beta}\left(\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right)^\gamma\right)d\boldsymbol{z}\qquad\text{(A·2)}\\&=\frac{1}{\beta-\alpha}\big\{(1-\alpha)D_{\gamma\alpha,\gamma}(p\|q)\\&\quad-(1-\beta)D_{\gamma\beta,\gamma}(p\|q)\big\}\geq 0\end{aligned}$$

が成り立つ．ただし，$\alpha,\beta\in(0,1)$ かつ $\alpha\neq\beta$ である．また，不等式の等号は $p\equiv q$ のとき，かつそのときに限り成り立つ． □

また，この結果より $\alpha>\beta$ のとき $\forall p,q\in\tilde{\mathcal{P}}(\boldsymbol{\mathcal{Z}})$ について

$$(1-\alpha)D_{\gamma\alpha,\gamma}(p\|q)\leq(1-\beta)D_{\gamma\beta,\gamma}(p\|q)$$

が成り立つ．

なお，命題 4 において $\gamma=1$ とおくことで，命題 3 が得られる．また，式 (A·2) において，$\beta=1$ とすることで $(\gamma\alpha,\gamma)$-divergence が得られ，$\beta=1$，$\gamma=1$ とすることで α-divergence が得られる．

2. 各種導出

2.1 命題 1

定義より

$$\begin{aligned}D_{\gamma,\beta}(p\|q)&=\frac{1}{\beta^2}D_{\frac{\gamma}{\beta}}(g_\beta\circ p\|g_\beta\circ q)\\&=\frac{1}{\beta^2}D_{1-\frac{\gamma}{\beta}}(g_\beta\circ q\|g_\beta\circ p)\\&=\frac{1}{\beta^2}D_{\frac{\beta-\gamma}{\beta}}(g_\beta\circ q\|g_\beta\circ p)\\&=D_{\beta-\gamma,\beta}(q\|p)\end{aligned}$$

が成り立つ．

2.2 補題 1

定義より

$$\begin{aligned}&D_{1,2}(p\|\bar{p}_{\boldsymbol{\beta}}^{1,2})\\&=\int_{\boldsymbol{\mathcal{Z}}}\frac{1}{2}\{p(\boldsymbol{z})^2-2p(\boldsymbol{z})\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z})+(\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z}))^2\}d\boldsymbol{z}\\&=\int_{\boldsymbol{\mathcal{Z}}}\frac{1}{2}\{p(\boldsymbol{z})^2-2p(\boldsymbol{z})\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z})+\sum_{i=1}^M\beta_i p_i(\boldsymbol{z})^2\\&\quad+(\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z}))^2-\sum_{i=1}^M\beta_i p_i(\boldsymbol{z})^2\}d\boldsymbol{z}\\&=\sum_{i=1}^M\beta_i\int_{\boldsymbol{\mathcal{Z}}}\frac{1}{2}\{p(\boldsymbol{z})^2-2p(\boldsymbol{z})p_i(\boldsymbol{z})+p_i(\boldsymbol{z})^2\}d\boldsymbol{z}\\&\quad-\sum_{i=1}^M\beta_i\int_{\boldsymbol{\mathcal{Z}}}\frac{1}{2}\{(\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z}))^2-2\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z})p_i(\boldsymbol{z})\\&\qquad\qquad+p_i(\boldsymbol{z})^2\}d\boldsymbol{z}\\&=\sum_{i=1}^M\beta_i D_{1,2}(p\|p_i)-\sum_{i=1}^M\beta_i D_{1,2}(\bar{p}_{\boldsymbol{\beta}}^{1,2}\|p_i)\end{aligned}$$

が成り立つ．

2.3 定理

一般性を失うことなく $\beta_M=1-\sum_{i=1}^{M-1}\beta_i$ としてよい．このとき

$$\begin{aligned}&\frac{\partial}{\partial\beta_j}\left\{\sum_{i=1}^M\beta_i D_{1,2}(p\|p_i)\right\}\\&=D_{1,2}(p\|p_j)-D_{1,2}(p\|p_M)\end{aligned}$$

及び

$$\frac{\partial}{\partial\beta_j}\left\{\sum_{i=1}^M\beta_i D_{1,2}(\bar{p}_{\boldsymbol{\beta}}^{1,2}\|p_i)\right\}$$

$$
\begin{aligned}
&= \frac{1}{2}\frac{\partial}{\partial\beta_j}\left\{\sum_{i=1}^{M}\beta_i p_i(\boldsymbol{z})^2 - (\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z}))^2\right\}d\boldsymbol{z} \\
&= \frac{1}{2}\{p_j(\boldsymbol{z})^2 - p_M(\boldsymbol{z})^2 \\
&\quad -2\bar{p}_{\boldsymbol{\beta}}^{1,2}(\boldsymbol{z})(p_j(\boldsymbol{z}) - p_M(\boldsymbol{z}))\}d\boldsymbol{z} \\
&= D_{1,2}(\bar{p}_{\boldsymbol{\beta}}^{1,2}\|p_j) - D_{1,2}(\bar{p}_{\boldsymbol{\beta}}^{1,2}\|p_M)
\end{aligned}
$$

が成り立つ ($j = 1, \cdots, M-1$)．このことと，$\hat{\boldsymbol{\beta}}$ の定義より

$$
\begin{aligned}
&\frac{\partial}{\partial\beta_j}D(p\|\bar{p}_{\boldsymbol{\beta}}^{1,2})\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= D_{1,2}(p\|p_j) - D_{1,2}(p\|p_M) \\
&\quad -D_{1,2}(\bar{p}_{\hat{\boldsymbol{\beta}}}^{1,2}\|p_j) + D_{1,2}(\bar{p}_{\hat{\boldsymbol{\beta}}}^{1,2}\|p_M) \\
&= 0
\end{aligned}
$$

が成り立つ ($j = 1, \cdots, M-1$)．したがって，$\forall i, j \in \{1, \cdots, M\}$ について

$$
\begin{aligned}
&D_{1,2}(p\|p_i) - D_{1,2}(\bar{p}_{\hat{\boldsymbol{\beta}}}^{1,2}\|p_i) \\
&= D_{1,2}(p\|p_j) - D_{1,2}(\bar{p}_{\hat{\boldsymbol{\beta}}}^{1,2}\|p_j)
\end{aligned}
$$

が成り立つ．これと補題 1 を用いることで示される．

（平成 14 年 7 月 10 日受付，15 年 5 月 30 日再受付，
12 月 8 日最終原稿受付）

内田　真人　（正員）

平 11 北大・工・情報卒．平 13 同大大学院工学研究科修士課程了．同年日本電信電話（株）入社．統計的学習理論，通信トラヒック分析の研究に従事．

塩谷　浩之　（正員）

平 2 北大・理・数学卒．平 4 同大大学院工学研究科修士課程了．平 7 同大大学院工学研究科博士後期課程了．同年同大工学部助手．現在，室蘭工業大学工学部助教授．数理情報工学の研究に従事．情報処理学会，日本神経回路学会各会員．