Boosting Algorithmの情報幾何

村田 昇

(早稲田大学理工学術院)

概 要

KL-divergence のある種の一般化である Bregman divergence (U-divergence) を用いることによって boosting アルゴリズムを包括的に捉え，その幾何学的意味を考えるための方法論を紹介する．

1 はじめに

近年，インターネットの急速な広がりにより大量のデータの収集あるいは共有が可能となり，また計算機の性能向上によってこうした大規模データを高速に処理することも現実的となり，実世界の膨大なデータを効率の良く正確に処理する要請が強まっている．こうした情勢を予見してか，90 年代以降判別やパタン認識の分野では様々な新しいアイデアが提案されてきた．その中でも boosting アルゴリズムは特筆すべきもののひとつとして挙げられるであろう．そもそもの出発点は Kearns and Valiant (1988) による

> ランダムな判別 (random guess) よりちょっとだけ賢い程度の弱学習器を，いくらでも正確な強学習器に増強 (boost) することはできるだろうか？

という問い掛けであったが，それに対して最初に肯定的な回答を示したのが Schapire (1990) である．実は 80 年代に盛んであったニューラルネットワークなどの並列分散処理の研究の中でも，集団学習 (ensemble learning, modular learning) という考え方はあり，様々な形で最適化の逐次化，分割などが提案され，その性能が解析されていた．しかしながら，その方法が主に学習器の構造に依存したヒューリスティクスにもとづいていたため，一般的な学習の枠組に広がることはなかったと思われる．一方，boosting は Freund and Schapire (1997); Schapire et al. (1998) をはじめとしてしっかりした理論的背景をもって，なおかつ一般的な枠組で研究されたため，その地位を堅固なものとして成功を収めることができたといえるのではないだろうか．

本稿では，boosting アルゴリズムを統計的に定式化し，そのアルゴリズムの背後にある幾何学的構造を明らかにすることを目的とする．これは Lebanon and Lafferty (2001) によって指摘された考え方を基礎とするが，より一般的な枠組で扱うために，KL-divergence をその特殊な形として含む Bregman divergence を導入する．判別問題の学習を，Bregman divergence を擬距離関数として用いた条件付測度の空間における推定問題として捉え，アルゴリズムの意味を情報幾何的な観点 (Amari, 1985; Amari and Nagaoka, 2000) から考察することを試みる．

2 U-Boost アルゴリズム

本稿で考えるのは，特徴量 $\boldsymbol{x} \in \mathcal{X}$ に対して対応するラベル $y \in \mathcal{Y}$ を予測する判別問題である．判別器はできるだけ一般的なものを考えるため，$\boldsymbol{x}$ に対して全てのラベルの集合 $\mathcal{Y}$ の部分集合 $\mathcal{C}$

を返す集合値関数 h

$$h : \boldsymbol{x} \in \mathcal{X} \mapsto \mathcal{C} \subset \mathcal{Y} \tag{1}$$

を考える．また，判別器 h に対して決定関数

$$f(\boldsymbol{x}, y) = \begin{cases} 1, & \text{if } y \in h(\boldsymbol{x}), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

を定義しておく．複数の判別器を用いた多数決判別器は，決定関数の線形結合を用いて

$$H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) \tag{3}$$

により定義される．このとき，単調増加な凸関数 U を用いて U-Boost は以下のように定義される．

U-Boost algorithm

入力: n 個の例題集合 $\{(\boldsymbol{x}_i, y_i); \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n\}$.

初期化: 分布 $D_1(i, y) = 1/n(|\mathcal{Y}| - 1)$ $(i = 1, \ldots, n)$ および結合決定関数 $F_0(\boldsymbol{x}, y) = 0$.

繰り返し: $t = 1, \ldots, T$

step 1: 分布 D_t のもとでの決定関数 f (判別器 h) の誤差を

$$\epsilon_t(f) = \sum_{i=1}^{n} \sum_{y \neq y_i} \frac{f(\boldsymbol{x}_i, y) - f(\boldsymbol{x}_i, y_i) + 1}{2} D_t(i, y)$$

で定義する．決定関数の集合 $\mathcal{F}$ から，この誤差を最小化するもの (実際には最小値を達成しない近似解でよい) を 1 つ選ぶ．

$$f_t(\boldsymbol{x}, y) = \arg\min_{f \in \mathcal{F}} \epsilon_t(f).$$

step 2: 信頼度 α_t を以下の式で計算する．

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} U\Big(F_{t-1}(\boldsymbol{x}_i, y) - F_{t-1}(\boldsymbol{x}_i, y_i) + \alpha \left(f_t(\boldsymbol{x}_i, y) - f_t(\boldsymbol{x}_i, y_i)\right)\Big).$$

step 3: 結合決定関数 F_t と分布 D_t を更新する．

$$F_t(\boldsymbol{x}, y) = F_{t-1}(\boldsymbol{x}, y) + \alpha_t f_t(\boldsymbol{x}, y),$$

$$D_{t+1}(i, y) \propto U'\left(F_t(\boldsymbol{x}_i, y) - F_t(\boldsymbol{x}_i, y_i)\right), \quad \text{ただし } \sum_{i=1}^{n} \sum_{y \neq y_i} D_{t+1}(i, y) = 1.$$

出力: 多数決判別器を構成する．

$$H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} F_T(\boldsymbol{x}, y) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y).$$

関数 U が指数関数の場合に step 2 と 3 はそれぞれ

$$\textbf{step 2: } \alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t(f_t)}{\epsilon_t(f_t)}, \qquad \textbf{step 3: } D_{t+1}(i,y) \propto \exp\{F_t(\boldsymbol{x}_i,y) - F_t(\boldsymbol{x}_i,y_i)\}$$

となり，上記のアルゴリズムは AdaBoost に帰着される．このため U-Boost は AdaBoost の一般化であると考えることができる．

このアルゴリズムは最適化の観点から見ると，関数 U で定義される結合決定関数 F に関する損失関数

$$L_U(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} U(F(\boldsymbol{x}_i,y) - F(\boldsymbol{x}_i,y_i))$$

を逐次的に最小化していると考えることができる．また，どんな関数 U に対しても，選択された決定関数と次時刻の誤差との間には

$$\epsilon_{t+1}(f_t) = \frac{1}{2} \quad (\forall t = 1,2,\ldots,T-1)$$

という関係が成り立つ．これは時刻 t において選ばれた決定関数 f_t が，更新された分布 D_{t+1} のもとでは正答率が 0.5，すなわちランダムな判別と性能的には同等になることを意味している．言い換えると，1 つ前に選ばれた判別器が最も不得意となるように分布の更新が進むことを意味している．以下では，こうしたアルゴリズムの特徴付けが統計的・情報幾何的な観点からどのような意味を持っているのかを考えていく．

3 有限測度の空間と Bregman Divergence

判別問題を統計的な枠組で取り扱うために，条件付確率の空間

$$\mathcal{P} = \left\{ m(y|\boldsymbol{x}) \;\middle|\; \sum_{y\in\mathcal{Y}} m(y|\boldsymbol{x}) = 1 \text{ (a.e. } \boldsymbol{x}) \right\} \tag{4}$$

と，それを含むより広い正値の条件付測度の空間

$$\mathcal{M} = \left\{ m(y|\boldsymbol{x}) \;\middle|\; \sum_{y\in\mathcal{Y}} m(y|\boldsymbol{x}) < \infty \text{ (a.e. } \boldsymbol{x}) \right\} \tag{5}$$

を定義しておく．統計の問題では $\mathcal{P}$ の中にモデルを設定して推測を行うのが一般的であるが，後述するように boosting ではモデルを $\mathcal{M}$ の中に広げてアルゴリズムを構成していると考えることができ，これがその特徴のひとつとなっている．

次に空間 $\mathcal{M}$ の 2 点の距離 (擬距離) を測るために，$\mathcal{M}$ 上の Bregman divergence を定義する．

定義 1 (Bregman divergence)**.** U を R 上の狭義凸関数とする．導関数 $u = U'$ は単調関数であり，逆関数 $\xi = (u)^{-1}$ が存在することに注意する．$\mathcal{M}$ の 2 つの点 $p(y|\boldsymbol{x})$，$q(y|\boldsymbol{x})$ に対し，$\mu(\boldsymbol{x})$ のもとでの Bregman cross-entropy を

$$H_U(p,q;\mu) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} \{U(\xi(q(y|\boldsymbol{x}))) - p(y|\boldsymbol{x})\xi(q(y|\boldsymbol{x}))\}\,\mu(\boldsymbol{x})d\boldsymbol{x} \tag{6}$$

で定義し，p から q への Bregman divergence を 2 つの cross-entropy の差として

$$D_U(p,q;\mu) = H_U(p,q;\mu) - H_U(p,p;\mu) \tag{7}$$

で定義する．

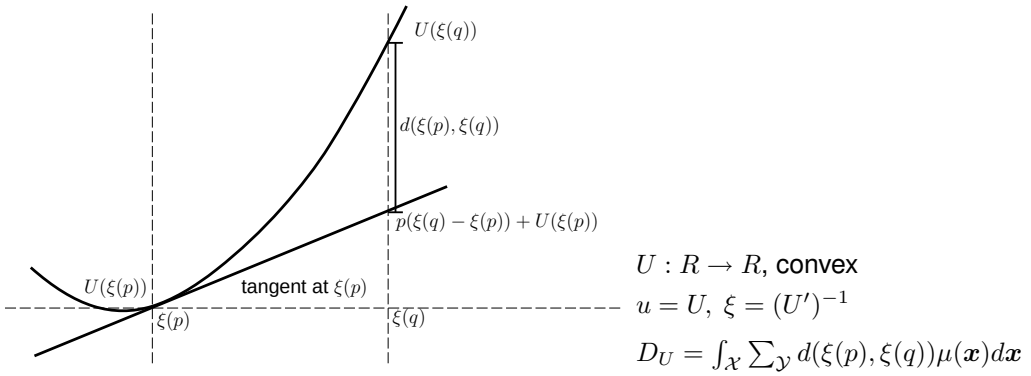

図 1: Bregman divergence.

図 1 は，2 つの測度 p と q の差が関数 U により測られる様子を直観的に示したものである．ここで用いた Bregman divergence の定義は逆関数 ξ による変換を含むため通常の定義より繁雑となっている．これは cross-entropy の第 2 項を関数 p で重み付けた和とすることによって，p が経験分布となっても計算できる形にするためである．

さて Bregman divergence の正値性から任意の p,q に対して

$$H_U(p,q;\mu) \geq H_U(p,p;\mu)$$

が成り立つので，p を固定したときの q に関する Bregman divergence の最小化は cross-entropy の最小化と等価である．

$$\arg\min_q D_U(p,q;\mu) = \arg\min_q H_U(p,q;\mu). \tag{8}$$

これは KL-divergence の最小化と最尤推定の関係と同じであることに注意する．

最後に，与えられた例題 $\{(\boldsymbol{x}_i, y_i); i = 1,\ldots,n\}$ に対して経験分布を次のように定義する．

$$\textbf{条件付:}\ \tilde{p}(y|\boldsymbol{x}) = \begin{cases} I(y = y_i), & \text{if } \boldsymbol{x} = \boldsymbol{x}_i, \\ \dfrac{1}{|\mathcal{Y}|}, & \text{otherwise,} \end{cases} \qquad \textbf{周辺:}\ .\tilde{\mu}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i).$$

ここでは簡単のため，1 つの特徴量に対して 1 つのラベルのみが例題として与えられる consistent data assumption (Lebanon and Lafferty, 2001) を課しているが，一般化は容易である．これらの経験分布を用いると，与えられた例題に対してある条件付測度の集合 $\mathcal{Q}$ の中で Bregman divergence の意味で最適な測度は

$$\tilde{q} = \arg\min_{q\in\mathcal{Q}} H_U(\tilde{p},q;\tilde{\mu}) = \arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{y\in\mathcal{Y}} U(\xi(q(y|\boldsymbol{x}_i))) - \xi(q(y_i|\boldsymbol{x}_i))\right] \tag{9}$$

で与えられる．

4 ピタゴラスの定理と直交葉層化

次に，空間 $\mathcal{M}$ の中の点の擬距離を Bregman divergence を用いて決めた場合に，重要となる性質をまとめておく．まず，3 つの点に関する最も基本的な定理を述べる．

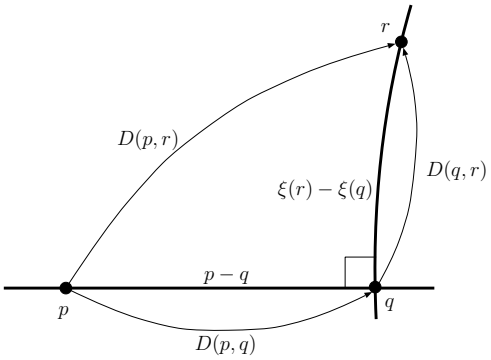

図 2: Bregman divergence におけるピタゴラスの定理.

定理 1 (ピタゴラスの定理). $\mathcal{M}$ の中の 3 点 p, q, r を考える．もし $p-q$ と $\xi(r)-\xi(q)$ が μ のもとで直交する

$$\langle p-q, \xi(r)-\xi(q)\rangle_\mu = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} \left(p(y|\boldsymbol{x}) - q(y|\boldsymbol{x})\right)\left(\xi(r(y|\boldsymbol{x})) - \xi(q(y|\boldsymbol{x}))\right)\mu(\boldsymbol{x})d\boldsymbol{x} = 0$$

のであれば，以下の関係が成り立つ.

$$D_U(p, r; \mu) = D_U(p, q; \mu) + D_U(q, r; \mu) \tag{10}$$

ここで直交性の定義のために用いられる $p-q$ と $\xi(r)-\xi(q)$ という表現に注意しておく．この表現から自然に導かれる性質の良い 2 つの部分空間が次のように定義される.

定義 2 (U-平坦部分空間). 1 点 $q_0 \in \mathcal{M}$ と，Λ を有限の添字の集合とする決定関数の集合 $\mathcal{F} = \{f_\lambda(\boldsymbol{x}, y); \lambda \in \Lambda\}$ を用いて

$$\mathcal{Q}_U(q_0, \mathcal{F}) = \left\{ q \in \mathcal{M} \;\middle|\; q(y|\boldsymbol{x}) = u\Big(\xi(q_0(y|\boldsymbol{x})) + \sum_{\lambda\in\Lambda} \alpha_\lambda f_\lambda(\boldsymbol{x}, y)\Big),\ \alpha_\lambda \in R \right\} \tag{11}$$

で定義される条件付測度の集合を U-平坦部分空間と呼ぶ.

定義 3 (m-平坦部分空間). 1 点 $p_0 \in \mathcal{M}$ を通り $\mathcal{Q}_U$ に垂直な $\mathcal{M}$ の部分空間

$$\begin{aligned}\mathcal{T}(p_0, \mu, \mathcal{F}) &= \left\{ p \in \mathcal{M} \;\middle|\; \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} (p(y|\boldsymbol{x}) - p_0(y|\boldsymbol{x})) f_\lambda(\boldsymbol{x}, y)\mu(\boldsymbol{x})d\boldsymbol{x} = 0,\ \forall\lambda \in \Lambda \right\} \\ &= \left\{ p \in \mathcal{M} \;\middle|\; \langle p - p_0, f_\lambda\rangle_\mu = 0,\ \forall\lambda \in \Lambda \right\} \end{aligned} \tag{12}$$

を m-平坦部分空間と呼ぶ.

U-平坦部分空間は別の書き方をすれば

$$\xi(q) - \xi(q_0) = \sum_{\lambda\in\Lambda} \alpha_\lambda f_\lambda(\boldsymbol{x}, y)$$

となり，ξ で変換された条件付測度がアフィン空間となっていることを表している．一方 m-平坦部分空間は $\mathcal{F}$ を法線方向とする $\mathcal{M}$ の中のアフィン部分空間である．またこれら 2 つの部分空間は q^* をはさんで，図 3 に示すような関係にあることに注意する.

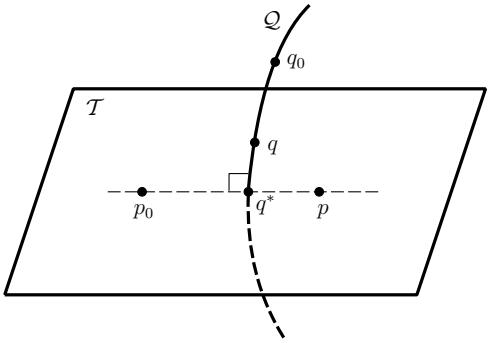

図 3: $\mathcal{Q}_U$ と $\mathcal{T}$ の幾何学的な関係.

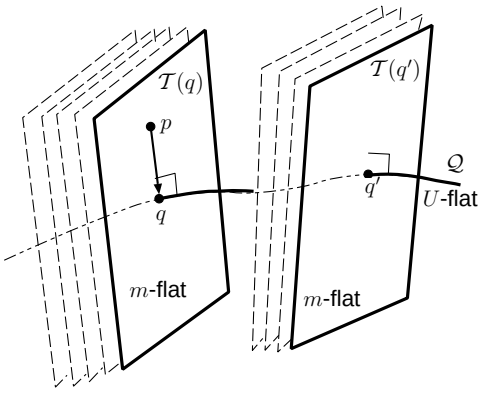

図 4: Bregman divergence から導かれる直交葉層化.

図 3 の関係を繰り返し用いると，図 4 のように $\mathcal{M}$ を $\mathcal{Q}_U$ に直交する m-平坦部分集合で分割することができる．

$$\bigcup_{q\in\mathcal{Q}_U} \mathcal{T}(q) = \mathcal{M},$$
$$\mathcal{T}(q) \cap \mathcal{T}(q') = \phi, \text{ if } q \neq q'.$$

個々の部分空間 $\mathcal{T}(q)$ は葉と呼ばれ，分割 $\{\mathcal{T}(q); q \in \mathcal{Q}_U\}$ は直交葉層化と呼ばれる．このとき，葉 $\mathcal{T}(q)$ に含まれる任意の点から Bregman divergence の意味で最も近い $\mathcal{Q}_U$ 上の点は，$\mathcal{T}(q)$ と $\mathcal{Q}_U$ の交点となる．この関係を用いると，次の 2 つの最適化問題が等価な解を与えることを示すことができる．

定理 2. p に関する最適化問題

$$q_0 \text{ を固定して } p \in \mathcal{T}(p_0) \text{ に関して } D_U(p, q_0; \mu) \text{ を最小化} \tag{13}$$

と，q に関する最適化問題

$$p_0 \text{ を固定して } q \in \mathcal{Q}_U(q_0) \text{ に関して } D_U(p_0, q; \mu) \text{ を最小化} \tag{14}$$

は，$\mathcal{Q}_U$ と $\mathcal{T}$ の交点

$$q^* = \arg\min_{p\in\mathcal{T}} D_U(p, q_0; \mu) = \arg\min_{q\in\mathcal{Q}_U} D_U(p_0, q; \mu), \tag{15}$$

を同じ解として与える．

この関係は図 5 のようにまとめられる．この関係，および divergence と cross-entropy の最小化の同値性から，モデル $\mathcal{Q}_U$ の中で経験分布 $\tilde{p}$ に最も近い点を求める，すなわち $H_U(\tilde{p}, q; \tilde{\mu})$ を最小とする q を求める最適化問題は，経験分布を含む性質の良い m-平坦な部分空間と U-平坦なモデルの空間の交点を求める問題に帰着される．

5 判別のためのモデル

最後に，条件付確率を用いて判別規則を決める問題が，確率とは限らない条件付測度の空間での探索を用いて解かれる仕組みについて説明する．これは次に述べる判別規則の不変性を用いる．

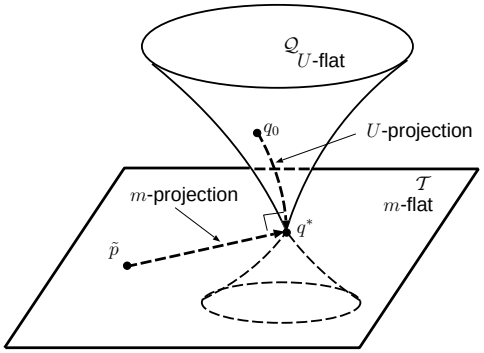

図 5: U-Boost アルゴリズムにおける 2 つの最適化問題の幾何学的解釈.

シフト不変性: $b(\boldsymbol{x})$ を任意の $\boldsymbol{x}$ の関数とする．$u(\xi(q)-b)$ に基づく判別規則は q に基づく判別規則と同等である.

$$\arg\max_{y\in\mathcal{Y}} \xi(q(y|\boldsymbol{x})) = \arg\max_{y\in\mathcal{Y}} \{\xi(q(y|\boldsymbol{x})) - b(\boldsymbol{x})\}$$

スケール不変性: $c(\boldsymbol{x})$ を任意の $\boldsymbol{x}$ の正値関数とする．$c(\boldsymbol{x})q(y|\boldsymbol{x})$ に基づく判別規則は $q(y|\boldsymbol{x})$ に基づく判別規則と同等である.

$$\arg\max_{y\in\mathcal{Y}} \xi\big(c(\boldsymbol{x})q(y|\boldsymbol{x})\big) = \arg\max_{y\in\mathcal{Y}} \xi\big(q(y|\boldsymbol{x})\big)$$

この 2 つの不変性により，直接確率を表していない条件付測度を用いても一致性のある判別が可能となる．アルゴリズムを構成する上で重要なのは主に前者で，直観的にはアルゴリズムにとって都合良く $b(\boldsymbol{x})$ を選べば良いことになる．実際 $\tilde{f}$ を

$$\tilde{f}_t(\boldsymbol{x}) = \sum_{y\in\mathcal{Y}} \tilde{p}(y|\boldsymbol{x}) f_t(\boldsymbol{x},y) = \begin{cases} f_t(\boldsymbol{x}_i,y_i), & \text{if } \boldsymbol{x}=\boldsymbol{x}_i, \\ \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_t(\boldsymbol{x},y), & \text{otherwise.} \end{cases}$$

で定義し，$\mathcal{F} = \{f_t(\boldsymbol{x},y) - \tilde{f}_t(\boldsymbol{x}); t=1,\ldots,T\}$ から構成されるモデル

$$\mathcal{Q}_U^{\mathrm{emp}}(q_0,\mathcal{F}) = \left\{ q\in\mathcal{M} \;\middle|\; \xi(q(y|\boldsymbol{x})) = \xi(q_0(y|\boldsymbol{x})) + \sum_{t=1}^{T} \alpha_t \big(f_t(\boldsymbol{x},y) - \tilde{f}_t(\boldsymbol{x})\big) \right\}. \tag{16}$$

が重要となる．これは b を

$$b(\boldsymbol{x},\boldsymbol{\alpha}) = \sum_{t=1}^{T} \alpha_t \tilde{f}_t(\boldsymbol{x}),$$

としているモデルである．このモデルは経験分布 $\tilde{p}$ に依存してしまうが，これに直交する m-平坦部分空間

$$\mathcal{T}(q) = \left\{ p\in\mathcal{M} \;\middle|\; \langle p-q, f_t-\tilde{f}_t\rangle_{\tilde{\mu}} = 0,\ \forall t \right\},\ q\in\mathcal{Q}_U^{\mathrm{emp}} \tag{17}$$

は非常に単純な構造を持ち，特に $\xi(q_0)=0$ としたとき，$H_U(\tilde{p},q;\tilde{\mu})$ を $F=\xi(q)$ に関して書き直した損失関数は

$$L_U^{\mathrm{emp}}(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} U\big(F(\boldsymbol{x}_i,y) - F(\boldsymbol{x}_i,y_i)\big) \tag{18}$$

となる．ここから AdaBoost をはじめとする様々なアルゴリズムを含む一般形が導かれる．

なお統計的に重要なもうひとつのモデルは，条件付確率分布を考える次のモデルである．

$$\mathcal{Q}_U^{\mathrm{norm}}(q_0,\mathcal{F})=\left\{q\in\mathcal{P}\;\middle|\;\xi(q(y|\boldsymbol{x}))=\xi(q_0(y|\boldsymbol{x}))+\sum_{t=1}^{T}\alpha_t f_t(\boldsymbol{x},y)-\phi(\boldsymbol{x},\boldsymbol{\alpha})\right\} \tag{19}$$

ここで ϕ は $\sum_{y\in\mathcal{Y}} q(y|\boldsymbol{x})=1$ とするための正規化因子であり，前節の意味ではこのモデルは厳密に平坦ではない．この場合 F の損失関数を書き下すと

$$L_U^{\mathrm{norm}}(F)=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{y\in\mathcal{Y}}U\big(F(\boldsymbol{x}_i,y)-\phi(\boldsymbol{x}_i,\boldsymbol{\alpha})\big)-\big\{F(\boldsymbol{x}_i,y_i)-\phi(\boldsymbol{x}_i,\boldsymbol{\alpha})\big\}\right] \tag{20}$$

となり，一般には ϕ を含んだ非線形最適化が必要となる．LogitBoost(Friedman et al., 2000) はこの特殊な場合となっている．

6 U-Boost の幾何学的解釈

以上の考察を踏まえ，U-Boost アルゴリズムを幾何的に書き直してみよう．

U-Boost algorithm

入力: n 個の例題集合 $\{(\boldsymbol{x}_i,y_i);\boldsymbol{x}_i\in\mathcal{X},y\in\mathcal{Y},i=1,\ldots,n\}$

初期化: $q_0(y|\boldsymbol{x})$ (通常，簡単のため $\xi(q_0)=0$ とする)

Do for $t=1,\ldots,T$

step 1: $f_t-b'_t$ ができるだけ $q_{t-1}-\tilde{p}$ と同じ方向を向くように判別器 h_t を選ぶ．

$$\text{maximize}\;\;\langle q_{t-1}-\tilde{p},f_t-b'_t\rangle_{\tilde{\mu}}$$

step 2: 1 次元のモデル

$$\mathcal{Q}_t=\left\{q\;\middle|\;\xi(q(y|\boldsymbol{x}))=\xi(q_{t-1}(y|\boldsymbol{x}))+\alpha f_t(\boldsymbol{x},y)-b_t(\boldsymbol{x},\alpha),\;\alpha\in R\right\}$$

と，その直交葉層化 $\{\mathcal{T}(q);q\in\mathcal{Q}_t\}$ を構成し，経験分布 $\tilde{p}$ を含む葉と $\mathcal{Q}_t$ の交点から α_t を求める．このとき α_t は以下を満している．

$$\alpha_t=\arg\min_{q\in\mathcal{Q}_t}\sum_{i=1}^{n}\left[\sum_{y\in\mathcal{Y}}U\big(\xi(q(y|\boldsymbol{x}_i))\big)-\xi(q(y_i|\boldsymbol{x}_i))\right].$$

step 3: q_t を更新する．

$$q_t(y|\boldsymbol{x})=u\Big(\xi(q_{t-1}(y|\boldsymbol{x}))+\alpha_t f_t(\boldsymbol{x},y)-b_t(\boldsymbol{x},\alpha_t)\Big)$$

出力: 多数決判別器を構成する．

$$H(\boldsymbol{x})=\arg\max_{y\in\mathcal{Y}}\sum_{t=1}^{T}\alpha_t f_t(\boldsymbol{x},y)$$

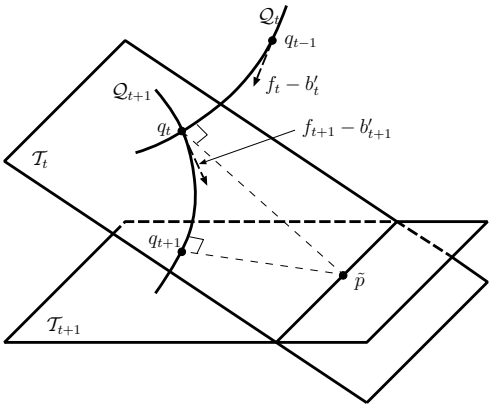

図 6: U-Boost アルゴリズムの幾何学的な解釈.

時刻 $t+1$(t ではなく $t+1$ としていることに注意) の探索においては，step 1 で用いる $q_t - \tilde{p}$ と直交する方向がランダムな判別の空間となる．q_t は $\mathcal{Q}_t$ の中で $\tilde{p}$ に最も近く，q_t における $\mathcal{Q}_t$ の接空間は $q_t - \tilde{p}$ と直交している．このため f_t(正確には $f_t - b'_t$) は時刻 $t+1$ においてはランダムな判別の空間に含まれ,

$$\epsilon_{t+1}(f_t) = \frac{1}{2} \quad (\forall t = 1, 2, \ldots, T-1)$$

が成り立つことになる．以上の幾何学的な解釈を図 6 にまとめておく.

7 おわりに

本稿は Murata et al. (2004) で扱った内容の概略を紹介したものである．本稿では紹介しきれなかったが，ここで述べたような統計的な枠組で扱うことによって，一般的な boosting アルゴリズムの一致性，有効性，頑健性といった性質を統一的に扱うことができる．例えば Bregman divergence の基本的な性質を用いると，どのような凸関数を用いても一致性を示すことができるので，適切な条件のもとで例題の数が十分多ければアルゴリズムは最適解に収束していくことが保証される．では凸関数の選択は何に影響するかと言えば，収束の速さを決める有効性であったり，外れ値やノイズに対する頑健性であったりする．これらを統一的に扱うことによって初めて凸関数の性質を比較することが可能となる.

Murata et al. (2004) の他にも，一致性の条件を詳しく議論した Zhang (2004); Bartlett et al. (2006); Tewari and Bartlett (2007) や，凸関数 U とノイズの関係を詳しく議論した Takenouchi et al. (2008) があるので，興味のある方はこれらも参照して戴きたい．また Bregman divergence については，ベイズ統計とゲーム理論との関連からその特徴付けを行う研究 (Grünwald and Dawid, 2004) などもあり，機械学習や統計においてその重要性が認識されつつあることを付記しておく.

なお boosting に関連する論文については，本稿ではあまり多くの論文を引用することはできなかったが，Murata et al. (2004) の参考文献表などを参照して戴きたい.

参考文献

S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985.

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classication, and risk bounds. *Journal of the American Statistical Association*, 101:138—156, 2006.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

M. Kearns and L. G. Valiant. Learning boolean formulae or finte automata is as hard as foctoring. Technical Report TR-14-88, Harvard University Aiken Computation Laboratry, aug 1988.

G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. Technical Report CMU-CS-01-144, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Aug 2001.

N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of $\mathcal{U}$-boost and bregman divergence. *Neural Computation*, 16:1432–1481, 2004.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

T. Takenouchi, S. Eguchi, N. Murata, and T. Kanamori. Robust boosting algorithm against mislabeling in multiclass problems. *Neural Computation*, 20:1596—1630, 2008.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classication methods. *Journal of Machine Learning Research*, 8:1007—1025, 2007.

T. Zhang. Statistical analysis of some multi-category large margin classication methods. *Journal of Machine Learning Research*, 5:1225—1251, 2004.