Geometry of U-Boost Algorithms

Noboru Murata[1], Takashi Takenouchi[2], Takafumi Kanamori[3], Shinto Eguchi[2,4]
[1]School of Science and Engineering, Waseda University
[2]Department of Statistical Science, Graduate University of Advanced Studies
[3]Department of Mathematical and Computing Sciences, Tokyo Institute of Technology
[4]Institute of Statistical Mathematics, Japan

(revised May 30, 2003)

Abstract

In this paper, we discuss a geometric understanding of the boosting algorithm in the framework of information geometry. We propose U-Boost learning algorithms, which is a generalization of AdaBoost, based on the Bregman divergence defined by a generic convex function U. This approach allows us to discuss statistical properties such as consistency and robustness based on a probabilistic assumption for a training data.

1 Introduction

In the last decade, several novel developments for classification and pattern recognition have been done mainly along statistical learning theory (see for example, MacLachlan, 1992; Bishop, 1995; Vapnik, 1995; Hastie et al., 2001). Several important approaches have been proposed and implemented into feasible computational algorithms. One promising direction is "boosting" which is a method of combining many learning machines trained by simple learning algorithms. Intuitively speaking, the key idea of boosting algorithm is to assort important and unimportant examples according whether machines are good at or weak in learning those examples.

In this paper, we focus on AdaBoost (Freund and Schapire, 1997) type algorithms. We propose a class of boosting algorithms, U-Boost, which is naturally derived from the Bregman divergence. This proposal gives an extension of the geometry discussed by Lebanon and Lafferty (2001) based on the Bregman divergence from the viewpoint of information geometry.

2 AdaBoost Algorithm

Let us consider a classification problem where for a given feature vector $\boldsymbol{x}$ in some space $\mathcal{X}$, the corresponding label y in a discrete set $\mathcal{Y}$ is predicted. For given n examples $\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$, the AdaBoost algorithm (AdaBoost.M1) consists of following steps.

step 1: Initialize $D_1(i) = 1/n, i = 1, \ldots, n$.

step 2: For $t = 1, \ldots, T$

This manuscript is a summary of the previous report:

Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S., "Information geometry of U-Boost and Bregman divergence." Research Memorandum No.860 (Institute of Statistical Mathematics).

which is available at `http://www.murata.elec.waseda.ac.jp/~mura/paper/uboost.pdf` .

- Define the error rate under the distribution D_t as

$$\epsilon_t(h) = \mathrm{Prob}_{D_t}\{h(\boldsymbol{x}) \neq y\} = \sum_{i=1}^{n} I(h(\boldsymbol{x}_i) \neq y_i) D_t(i),$$

  where I is the indicator function defined by

$$I(A) = \begin{cases} 1, & A \text{ is true} \\ 0, & \text{otherwise} \end{cases}.$$

- Select a machine h_t based on the error rate $\epsilon_t(h_t)$.
- Set ϵ_t and α_t as

$$\epsilon_t = \epsilon_t(h_t), \quad \alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right).$$

- Update the distribution D_t by

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t}, & \text{if } h_t(\boldsymbol{x}_i) = y_i \\ e^{\alpha_t}, & \text{otherwise.} \end{cases}$$

  where Z_t is a normalization constant to ensure $\sum_{i=1}^{n} D_{t+1}(i) = 1$.

step 3: Output the final decision as the majority vote

$$H(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t I(h_t(\boldsymbol{x}) = y)$$

AdaBoost can be regarded as a procedure of optimizing an exponential loss with an additive model (Friedman et al., 2000)

$$L(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y \in \mathcal{Y}} \exp(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i)), \quad \text{where } F(\boldsymbol{x}, y) = \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y).$$

By adopting different loss functions, several variations of AdaBoost are proposed, such as MadaBoost (Domingo and Watanabe, 2000), where the loss function

$$L(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y \in \mathcal{Y}} \phi(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i)), \quad \text{where } \phi(z) = \begin{cases} z + \frac{1}{2} & z \geq 0, \\ \frac{1}{2}\exp(2z) & \text{otherwise,} \end{cases}$$

is used instead of the exponential loss.

For constructing algorithms, the notion of the loss function is useful, because the various algorithms are derived based on the gradient descent and line search methods. Also the loss function controls the confidence of the decision, which is characterized by the margin (Schapire et al., 1998). However, the statistical properties such as consistency and efficiency are not apparent, because the relationship between loss functions and the distributions realized by combined machines is unclear so far.

3 Bregman Divergence

Let us consider the space of all the positive finite measures over $\mathcal{Y}$ conditioned by $\boldsymbol{x} \in \mathcal{X}$

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

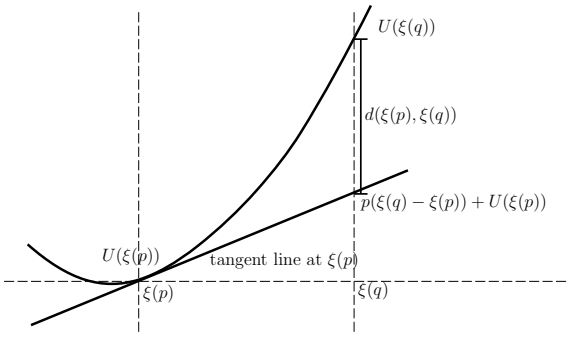
The Bregman divergence is a pseudo-distance for measuring the discrepancy between two functions. We define the Bregman divergence between two conditional measures as follows.

Definition 1 (Bregman divergence). Let U be a strictly convex function on R, then its derivative $u = U'$ is a monotone function, which has the inverse function $\xi = (u)^{-1}$. For $p(y|\boldsymbol{x})$ and $q(y|\boldsymbol{x})$ in $\mathcal{M}$, the Bregman divergence from p to q is defined by

$$D_U(p,q) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} \Big[\big\{U(\xi(q)) - U(\xi(p))\big\} - p\big\{\xi(q) - \xi(p)\big\}\Big] d\mu, \tag{2}$$

where $d\mu = \mu(d\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}$.

As easily seen, the Bregman divergence is not symmetric with respect to p and q in general, therefore it is not a distance.



$U : R \to R$, convex
$\xi : u^{-1} = (U')^{-1}$

Note: $u(\xi(p)) = p$

Figure 1: Bregman divergence.

The advantage of the form (2) is allowing us to plug in the empirical distribution directly. To see this, let us decompose the Bregman divergence into

$$D_U(p,q) = L_U(p,q) - L_U(p,p), \tag{3}$$

where

$$L_U(p,q) = \int \sum \{U(\xi(q)) - p\xi(q)\}\, d\mu. \tag{4}$$

Note that L_U can be regarded as a loss function, and since the Bregman divergence is non-negative, that is $D_U(p,q) \geq 0$, the loss is bounded below by

$$L_U(p,q) \geq L_U(p,p).$$

Now we consider a problem in which q is optimized with respect to $D_U(p,q)$ for fixed p. Picking out the terms which depend on q, the problem is simplified as

$$\operatorname*{argmin}_q D_U(p,q) = \operatorname*{argmin}_q L_U(p,q). \tag{5}$$

In $L_U(p,q)$, the distribution p appears only for taking the expectation of $\xi(q)$, therefore the empirical distribution $\tilde{p}$ is used without any difficulty as

$$L_U(\tilde{p},q) = \frac{1}{n}\sum_{i=1}^{n} \Big\{ \sum_{y\in\mathcal{Y}} U(\xi(q(y|\boldsymbol{x}_i))) - \xi(q(y_i|\boldsymbol{x}_i)) \Big\}, \tag{6}$$

which we refer as the empirical U-loss, and the optimal distribution for given examples is defined by

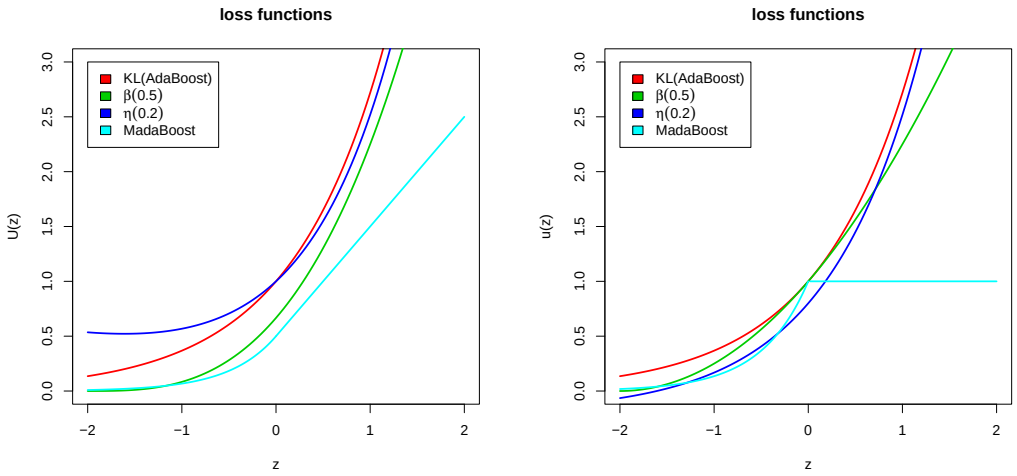
$$\tilde{q} = \operatorname*{argmin}_q L_U(\tilde{p},q).$$

This is equivalent with the well-known relationship between the maximum likelihood estimation and the minimization of the Kullback-Leibler divergence. Related discussions can be found in Eguchi and Kano (2001), in which the divergences are derived based on the pseudo-likelihood.

The followings are examples of the convex function U.

Example 1 (U-functions).

Kullback-Leibler:	$U(z) = \exp(z)$
β-type:	$U(z) = \frac{1}{\beta+1}(\beta z + 1)^{\frac{\beta+1}{\beta}}$
η-type:	$U(z) = \exp(z) - \eta z$
MadaBoost:	$U(z) = \begin{cases} z + \frac{1}{2} & z \geq 0, \\ \frac{1}{2}\exp(2z) & z < 0, \end{cases}$



(a) (b)

Figure 2: Examples of U-functions. (a) Shapes of U-functions. (b) Derivatives of U-functions.

4 Pythagorean Relation and Orthogonal Foliation

Let us define the inner product of functions of $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ by

$$\langle f, g \rangle = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} f(\boldsymbol{x}, y) g(\boldsymbol{x}, y) \mu(d\boldsymbol{x})$$

and define that f and g are orthogonal if $\langle f, g \rangle = 0$. Then the Pythagorean relation for the Bregman divergence is stated as follows.

Lemma 1 (Pythagorean relation). *Let p, q and r be in $\mathcal{M}$. If $p - q$ and $\xi(r) - \xi(q)$ are orthogonal, $\langle p - q, \xi(r) - \xi(q) \rangle = 0$, the relation*

$$D_U(p, r) = D_U(p, q) + D_U(q, r) \tag{7}$$

holds.

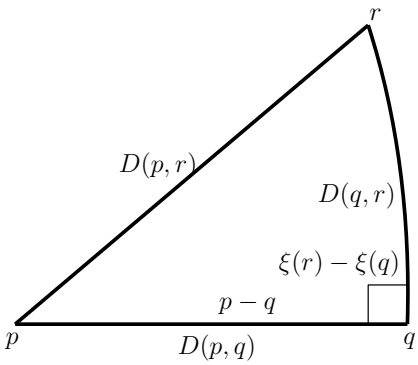


Figure 3: Pythagorean relation for Bregman divergence.

Now we consider subspaces suitable for the nature of the Bregman divergence. Let us consider a set of conditional measures with a fixed $q_0 \in \mathcal{M}$ and a set of functions $\boldsymbol{f} = \{f_t(\boldsymbol{x}, y); t = 1, \ldots, T\}$ and a function b of $\boldsymbol{x}$ and $\boldsymbol{\alpha}$, written in the form of

$$\begin{aligned}\mathcal{Q}_U &= \mathcal{Q}_U(q_0, \boldsymbol{f}, b) \\ &= \left\{ q \in \mathcal{M} \;\middle|\; q_{\boldsymbol{\alpha}} = u\big(\xi(q_0) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) - b(\boldsymbol{x}, \boldsymbol{\alpha})\big) \right\}, \end{aligned} \tag{8}$$

where $\boldsymbol{\alpha} = \{\alpha_t \in R; t = 1, \ldots, T\}$. In other words, $\mathcal{Q}_U$ consists of functions such that

$$\xi(q) - \xi(q_0) = \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) - b(\boldsymbol{x}, \boldsymbol{\alpha}),$$

which means $\mathcal{Q}_U$ is a subspace spanned by $\boldsymbol{f}$ and b and includes q_0. $\mathcal{Q}_U$ is called a U-flat subspace, which we refer the U-model in the following.

Next let us consider an mixture-flat subspace in $\mathcal{M}$ which passes a point $q = q_{\boldsymbol{\alpha}} \in \mathcal{Q}_U$,

$$\begin{aligned}\mathcal{F}_U(q) &= \mathcal{F}_U(q, \boldsymbol{f}, b) \\ &= \left\{ p \in \mathcal{M} \;\middle|\; \langle p - q, f_t - b'_t(\boldsymbol{\alpha}) \rangle = 0, \ \forall t \right\}, \end{aligned} \tag{9}$$

where

$$b'_t(\boldsymbol{x}, \boldsymbol{\alpha}) = \frac{\partial b(\boldsymbol{x}, \boldsymbol{\alpha})}{\partial \alpha_t}.$$

By these definitions, $\mathcal{Q}_U$ and $\mathcal{F}_U(q)$ are orthogonal at q, that is,

$$\left\langle p - q, \frac{\partial}{\partial \alpha_t} \xi(q) \right\rangle = \langle p - q, f_t - b'_t(\boldsymbol{\alpha}) \rangle = 0, \ \forall p \in \mathcal{F}_U(q), \ \forall t.$$

A set $\{\mathcal{F}_U(q); q \in \mathcal{Q}_U\}$ is called a foliation of $\mathcal{M}$, which covers the whole space $\mathcal{M}$ as

$$\begin{aligned}&\bigcup_{q \in \mathcal{Q}} \mathcal{F}_U(q) = \mathcal{M}, \\ &\mathcal{F}_U(q) \cap \mathcal{F}_U(q') = \phi, \text{ if } q \neq q'. \end{aligned}$$

To put it in other words, $\mathcal{M}$ is decomposed into an orthogonal foliation by giving $\mathcal{Q}_U$.

The function b must be determined by the constraint on the U-model such as from the statistical or computational convenience. From a statistical point of view, following two specific cases are plausible:

- **unnormalized:** $b = 0$,
- **normalized:** b is chosen so that $\sum_{y\in\mathcal{Y}} q(y|\boldsymbol{x}) = 1$.

5 U-Boost Algorithm

Using the Bregman divergence as a loss function, a generic form of the U-Boost algorithm is naturally introduced as follows.

U-Boost algorithm

step 1: Initialize $q_0(y|\boldsymbol{x})$. (In usual case, set $\xi(q_0) = 0$ for simplicity.)

step 2: For $t = 1, \ldots, T$

- Select a machine h_t so that

$$\langle \tilde{p} - q_{t-1}, f_t - b_t'(\alpha = 0)\rangle \neq 0,$$

  where

$$f_t(\boldsymbol{x}, y) = \begin{cases} \frac{1}{2}, & y \in h_t(\boldsymbol{x}), \\ -\frac{1}{2}, & \text{otherwise}, \end{cases}$$

  and $b_t(\boldsymbol{x}, \alpha)$ is an auxiliary function to satisfy an imposed constraint.
- Construct $\mathcal{Q}_t$,

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

- Find q_t and corresponding α_t which minimizes $D_U(\tilde{p}, q)$,

$$\begin{aligned} q_t &= \operatorname*{argmin}_{q\in\mathcal{Q}_t} D_U(\tilde{p}, q) \\ &= \operatorname*{argmin}_{q\in\mathcal{Q}_t} \sum_{i=1}^{n} \left\{ \sum_{y\in\mathcal{Y}} U(\xi(q(y|\boldsymbol{x}_i))) - \xi(q(y_i|\boldsymbol{x}_i)) \right\}. \end{aligned}$$

step 3: Output the final decision as the majority vote,

$$H(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y).$$

The optimization procedure in step 2 is geometrically interpreted as shown in Fig. 4. For a U-model $\mathcal{Q}_t$, we can consider an orthogonal foliation $\mathcal{F}_t(q)$ as

$$\mathcal{F}_t(q) = \left\{ p \in \mathcal{M} \;\middle|\; \langle p - q, f_t - b_t'(\alpha)\rangle = 0 \right\}, \quad \forall q = q_\alpha \in \mathcal{Q}_t. \tag{10}$$

Then we can find a leaf $\mathcal{F}_t(q_*)$ which passes the empirical distribution $\tilde{p}$, and the optimal model at step t is determined by $q_t = q_*$.

6 Properties of U-Boost

In this section, we discuss some properties of the U-Boost algorithms from the statistical point of view.

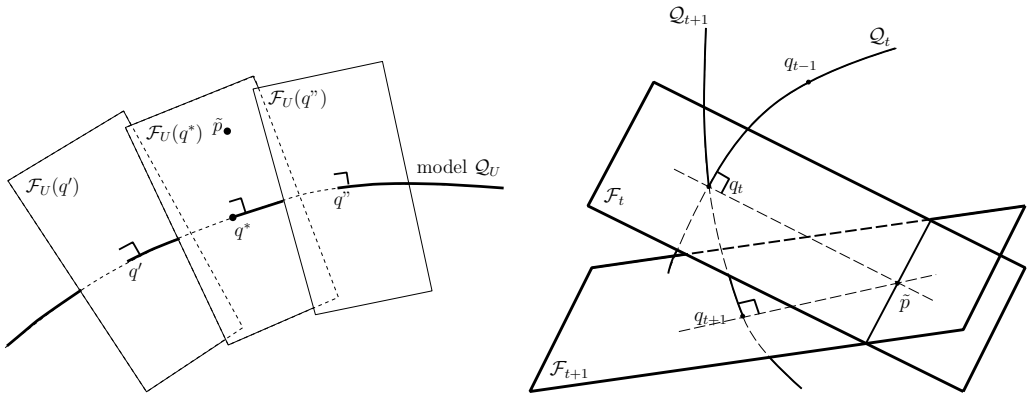


Figure 4: A geometrical interpretation of the U-Boost algorithm: relationship between the optimal point and the empirical distribution (left), and an intuitive image of successive two updates (right).

6.1 Error Rate

One of the important characteristics of the AdaBoost algorithm is the evolution of its weighted error rates, that is, the machine h_t at step t shows the worst performance under the distribution at the next step $t+1$, that is equivalent to random guess. By defining the weighted error as

$$\epsilon_t(h) = \sum_{i=1}^{n} I(h(\boldsymbol{x}_i) \neq y_i) D_t(i),$$

this fact is stated as

$$\epsilon_{t+1}(h_t) = \frac{1}{2}.$$

Similar disposition can be observed in the U-Boost algorithm as follows. Let us define the weight

$$D_t(i, y) = \frac{q_{t-1}(y|\boldsymbol{x}_i)}{Z_t}, \tag{11}$$

where Z_t is a normalization constant defined by

$$Z_t = \sum_{i=1}^{n} \sum_{y \neq y_i} q_{t-1}(y|\boldsymbol{x}_i), \tag{12}$$

and then define the weighted error by

$$\epsilon_t(h) = \sum_{i=1}^{n} \sum_{y \neq y_i} \left(\frac{f(\boldsymbol{x}_i, y) - f(\boldsymbol{x}_i, y_i) + 1}{2} \right) D_t(i, y). \tag{13}$$

Then we can prove

$$\epsilon_{t+1}(h_t) = \frac{1}{2}. \tag{14}$$

This mean that the U-Boost algorithm updates the distribution into the least favorable for the previous machine at each step.

6.2 Consistency and Bayes Optimality

Using the basic property of the Bregman divergence, we can show the consistency of the U-loss.

Lemma 2. *Let $p(y|\boldsymbol{x})$ be the true conditional distribution and $F(\boldsymbol{x}, y)$ be the minimizer of the U-loss $L_U(p, q)$, where $q = u(F)$. The classification rule given by F becomes Bayes optimal, that is,*

$$\hat{y}(\boldsymbol{x}) = \operatorname*{argmin}_{y \in \mathcal{Y}} F(\boldsymbol{x}, y) = \operatorname*{argmin}_{y \in \mathcal{Y}} p(y|\boldsymbol{x}). \tag{15}$$

In the U-Boost algorithm, $F(\boldsymbol{x}, y)$ is chosen from a class of functions which are linear combination of $f_t(\boldsymbol{x}, y); t = 1, \ldots, T$. In the case that the true distribution is not in the considered U-model, the closest point in the model is chosen in the sense of U-loss, however, if the number of boosting is sufficiently large and the functions $f_t; t = 1, \ldots, T$ are diverse, U-model can well approximate the true distribution. See for example Barron (1993); Murata (1996), for the discussion about the richness of the linear combination of simple functions.

As a special case, we can show the following theorem for the binary classification, where the U-loss is given by

$$L_U(p, q) = \int_{\mathcal{X}} \sum_{y \in \{\pm 1\}} p(y|\boldsymbol{x}) U(-yF(\boldsymbol{x})) \mu(d\boldsymbol{x}).$$

Theorem 1. *The minimizer of the U-loss gives the Bayes optimal, that is,*

$$\{\boldsymbol{x} | F(\boldsymbol{x}) > 0\} = \left\{ \boldsymbol{x} \middle| \log \frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})} > 0 \right\}.$$

Moreover, if

$$\log \frac{u(z)}{u(-z)} = 2z \tag{16}$$

holds, F coincides with the log likelihood ratio

$$F(x) = \frac{1}{2} \log \frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}.$$

U-functions for AdaBoost, LogitBoost, and MadaBoost satisfy the condition (16). The theorem agree with the result in Eguchi and Copas (2001, 2002).

6.3 Most B-robust U-function

Let us consider an estimator of α with the U-function as

$$\alpha_U(q\mu) = \operatorname*{argmin}_{\alpha} \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} q(y|\boldsymbol{x}) U\big(-y(F(\boldsymbol{x}) + \alpha h(\boldsymbol{x}))\big) d\mu(\boldsymbol{x}),$$

where $q\mu$ is the joint distribution of $\boldsymbol{x}$ and y. The robustness of the estimator is measured by the gross error sensitivity (Hampel et al., 1986)

$$\gamma(U, p_0) = \sup_{(\tilde{\boldsymbol{x}}, \tilde{y})} \left\{ \lim_{\epsilon \to +0} \frac{1}{\epsilon} \Big[\alpha_U((1-\epsilon)\, p_0\mu + \epsilon\, \delta(\tilde{\boldsymbol{x}}, \tilde{y})) - \alpha_U(p_0\mu) \Big] \right\}^2, \tag{17}$$

where $\delta(\tilde{\boldsymbol{x}}, \tilde{y})$ is the probability distribution with a point mass at $(\tilde{\boldsymbol{x}}, \tilde{y})$. The gross error sensitivity measures the worst influence which a small amount of contamination can have on the value of the estimator. The estimator which minimizes the gross error sensitivity is called the most B-robust estimator. For a choice of a robust U-function, we show the following theorem.

Theorem 2. *The U-function which derives MadaBoost algorithm minimizes the gross error sensitivity among the U-function with the property of* (16).

7 Conclusion

In this paper, we formulated boosting algorithms as sequential updates of conditional measures, and we introduced a class of boosting algorithms by considering the relation with the Bregman divergence. By this framework, statistical properties such as consistency and robustness are discussed. Still detailed studies on some properties such as the rate of convergence and stopping criteria of boosting, are needed to avoid overfitting problem and so on.

Here we only treated the classification problem, but the formulation can be extended to the case where y is in some continuous space, such as regression and density estimation. This is also remained as a future work.

References

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Information Theory*, 39(3):930–945, May 1993.

C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

C. Domingo and O. Watanabe. MadaBoost: A modification of AdaBoost. In *Proc. of the 13th Conference on Computational Learning Theory, COLT'00*, 2000.

S. Eguchi and J. B. Copas. Recent developments in discriminant analysis from an information geometric point of view. *Journal of the Korean Statistical Society*, 30:247–264, 2001. (The special issue of the 30th anniversary of the Korean Statistical Society).

S. Eguchi and J. B. Copas. A class of logistic type discriminant functions. *Biometrika*, 89: 1–22, 2002.

S. Eguchi and Y. Kano. Robustifing maximum likelihood estimation by psi-divergence. ISM Research Memo 802, The Institute of Statistical Mathematics, 2001.

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.

F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics*. John Wiley and Sons, Inc., 1986.

T. Hastie, R. Tibishirani, and J. Friedman. *The elements of statistical learning*. Springer, New York, 2001.

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.

G. J. MacLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley, New York, 1992.

N. Murata. An integral representation with ridge functions and approximation bounds of three-layered network. *Neural Networks*, 9(6):947–956, 1996.

N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. ISM Research Memorandam 860, Institute of Statistical Mathematics, Tokyo, Japan, 2002.

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 and S. Eguchi. Robustifying AdaBoost by adding the naive error rate. ISM Research Memorandam 859, Institute of Statistical Mathematics, Tokyo, Japan, 2002.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.