Robust boosting algorithm against mislabeling in multi-class problems

Takashi Takenouchi  
Nara Institute of Science and Technology  
ttakashi@is.naist.jp

Shinto Eguchi  
Institute of Statistical Mathematics, Japan and  
Department of Statistical Science, Graduate University of Advanced Studies  
Minami-azabu, Tokyo 106-8569, Japan  
eguchi@ism.ac.jp

Noboru Murata  
Faculty of Science and Engineering  
Waseda University  
3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan  
noboru.murata@eb.waseda.ac.jp

Takafumi Kanamori  
Department of Computer Science and Mathematical Informatics,  
Graduate School of Information Science  
Nagoya University  
Furocho, Chikusaku, Nagoya 464-8603, Japan  
kanamori@is.nagoya-u.ac.jp

Abstract

We discuss robustness against mislabeling in multi-class labels for classification problems and propose two algorithms of boosting, called the normalized Eta-Boost.M and Eta-Boost.M based on the Eta-divergence. Those two boosting algorithms are closely related with models of mislabeling in which the label is erroneously exchanged into others. For the two boosting algorithms, theoretical aspects supporting the robustness for mislabel is explored. We apply the proposed two boosting methods for synthetic and real datasets, to investigate performance of methods focusing on the robustness, and confirm the validity of the proposed methods.

1 Introduction

Classification and pattern recognition in statistical learning theory is now heading in a new direction. A number of learning algorithms of boosting have been proposed since AdaBoost was first introduced in [1], and their efficiency has been elucidated for classification problems from various points of view [2, 3]. The idea is based on combining different learning machines to build a high-performance machine.

In the present paper, we introduce a new probabilistic formalization for multi-class classification algorithms. In principle, probabilistic formalization covers various situations connecting feature vectors with labels in accordance
with the degree of random variability. There are different strategies for dealing with different types of classification problems. For example, precision goals depend on data contexts in problems for classifying hand-written characters, or printed ones in character recognition. We focus on the classification based on empirical examples with existing noise in labeling feature vectors. Regarding the binary classification problem, a contamination model of a conditional probability of the label for the feature vector is introduced in [4]. Therefore, the practical conditional probability is considered to be a mixture of the assumed probability distribution and the incorrect probability due to mislabeling.

In this paper, we propose Eta-Boost.M for multi-class classification as an extension of the binary $\eta$-Boost algorithm [4]. Eta-Boost.M consists of a single-parameter family of boosting algorithms with the parameter $\eta$, which subsumes the logistic discriminant analysis with $\eta = 0$. This is based on Eta-divergence between densities, which need not have a total mass of 1. The Kullback-Leibler divergence yields a pair of boosting algorithms, AdaBoost and LogitBoost [5]. According to this analogy, Eta-divergence has a pair of boosting algorithms, Eta-Boost.M and normalized Eta-Boost.M associated with two models of mislabeling. The key idea is to focus on a uniform distribution on the labels. We consider that the classification task becomes more difficult when the conditional probability distribution, given a feature vector, becomes closer to the uniform distribution. From this point of view we propose a mislabeling model for the conditional probability such that the probability of a mislabel is relatively high when the supposed conditional probability distribution, given the feature vector, is faced to a difficult situation as described above. This is a probabilistic extension of the mislabeling model discussed in [4]. We discuss several versions of Eta-loss functions which associated with the mislabeling model.

2 Probabilistic model for mislabeling

2.1 Formulation

Let us take a probabilistic formulation for a classification problem. Let $X$ be a feature vector in a space $\mathcal{X}$ and $Y$ be a label in a finite set $\mathcal{Y} = \{1, \cdots, G\}$. Suppose that an underlying probability distribution of $(X, Y)$ is

$$r(x, y) = q(x)p(y|x),$$

and we introduce a formal discussion in an idealized situation where the distribution $r(x, y)$ is completely known. Thus we have complete knowledge of the conditional probability of $Y = y$ given $X = x$, say, $p(y|x)$. Then the Bayes rule provides the optimal rule of classification for a given feature vector $x$:

$$F^\text{bayes}_p(x) = \arg\max_{y \in \mathcal{Y}} p(y|x).$$

The sense of optimality of $F^\text{bayes}_p$ is justified by minimizing the error rate of a classification rule $F(x)$ in the class of all classification rules and the error rate
is defined by

\[
\text{Err}(F, p, q) = \text{Prob}_{p, q}(F(X) \neq Y)
\]

\[
= 1 - \sum_{y \in \mathcal{Y}} \int_{X_{X, Y}} p(y|x)q(x)dx,
\]

where \(X_{X, Y} = \{x | x \in X, F(x) = y\}\). As the optimality measure, the error rate can be extended to various criteria including the credit score and the area under the ROC curve. See [6] for a discussion of the relevant statistics. Let \(H(x, y)\) be a discriminant function associates with a classification rule \(F_H(x)\) by

\[
F_H(x) = \arg\max_{y \in \mathcal{Y}} H(x, y).
\]

(3)

In a class of discriminant functions, \(H\) is said to be Bayes optimal if \(F_H = F_{\text{bayes}}\).

Let \(f(x) \in \mathcal{Y}\) be a weak classification rule that needs a relatively low computational cost to be constructed. Note that \(f(x)\) is a multiple value function and it returns a subset of \(\mathcal{Y}\). For a given classification rule \(f(x)\), we define a weak hypothesis by

\[
h_f(x, y) = I(y \in f(x)).
\]

(4)

The weak hypothesis \(h_f(x, y)\) is a kind of a discriminant function corresponding to the classification rule \(f(x)\) and let \(\mathcal{H}\) be the class of all weak hypotheses. The aim of a Boosting algorithm is to construct a discriminant function by linearly combining \(T\) weak hypotheses as

\[
H_T(x, y) = \sum_{t=1}^{T} \alpha_t h_t(x, y), h_t \in \mathcal{H}.
\]

2.2 Mislabling model

In many cases of classification problems, the class membership probability \(p(y|x)\) is modeled by the conventional logistic model \(p_{0,H}(y|x)\) constructed by the discriminant function \(H(x, y)\) as

\[
p_{0,H}(y|x) = \frac{\exp(H(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(H(x, y'))},
\]

(5)

and the discriminant function \(H\) is estimated based on a given training dataset \(\{(x_i, y_i), i = 1, \cdots, n\}\). In practice, the class labels in the dataset are frequently determined by human judgments with varying degrees of uncertainty. Failure prediction and medical diagnostics are typical examples of such uncertainty. Thus, we often face a situation in which part of labels of the given dataset are wrongly observed and then have to consider a model in which some contamination may be added to the data generation or an observation mechanism in the labeling process. In a context of the binary classification problem, Copas [7]...
considered a contamination model with a constant probability and discussed statistical properties. The model was extended to consider a probability of contamination depending on the input $x$ with an algorithm of Boosting method [4] and it was confirmed that an algorithm considering mislabeling worked well compared with other methods such as original AdaBoost, MadaBoost [8] and AdaBoost\textsuperscript{reg} [9].

In this paper, we extend the concept of the contamination model to the multiclass classification problem. We consider a probabilistic approach to mislabel problems as follows. Let us define

$$\mathcal{D} = \left\{ \delta(x) \mid x \in \mathcal{X}, 0 \leq \delta(x) < \frac{1}{G} \right\}$$

and consider the following model:

$$p_\delta(y|x) = (1 - (G - 1)\delta(x))p(y|x) + \sum_{k \neq y} \delta(x)p(k|x).$$

(6)

This is a model of the posterior distribution of the class label $y$ given the feature vector $x$, such that a class labeling with the original posterior distribution $p(y|x)$ is changed by mislabeling from $y$ to $G - 1$ other labels with equal probability $\delta(x)$. We call $p_\delta(y|x)$ a mislabel model with mislabel probability $\delta(x)$. It will be shown that the Bayes rule based on $p_\delta(y|x)$ is invariant with that based on the original posterior $p(y|x)$, while a lower bound for error rate increases for any $\delta(x)$.

In classification problems, one of the most difficult situations is suggested by a case in which the posterior distribution does not depend on any feature vector $x$, that is $p(y|x) = p_U(y) = 1/G$, where $p_U$ is a uniform distribution on $\mathcal{Y}$. If we choose a probability of mislabel setting by $\delta(x) = 1/G$, then we observe that $p_\delta(y|x) = p_U(y)$ and the model $p_\delta(y|x)$ has no prediction ability. We will propose a more sensible form of $\delta(x)$ in a subsequent discussion.

We first remark from (6) that

$$p_\delta(y|x) = \left(1 - G\delta(x)\right)p(y|x) + \delta(x),$$

(7)

which implies that the Bayes rule based on the original posterior $p(y|x)$ is the same as that based on the current posterior $p_\delta(y|x)$ for any $\delta(x) \in \mathcal{D}$.

**Remark 1.** For any $\delta \in \mathcal{D}$, we observe $F_{p_\delta}^{\text{bayes}}(x) = F_{p_\delta}^{\text{bayes}}(x)$.

*Proof.** For any $y, y' \in \mathcal{Y}$ and $x \in \mathcal{X}$, we observe

$$p_\delta(y|x) - p_\delta(y'|x) = (1 - G\delta(x))(p(y|x) - p(y'|x)).$$

(8)

Because $\delta(x)$ is in $\mathcal{D}$, $1 - G\delta(x)$ is positive and then order relations of posteriors of $y$ are invariant. \qed
Secondly, we investigate lower bounds of error rates for both cases of $p(y|x)$ and $p_{\delta}(y|x)$. Let us consider a lower bound $\text{errbd}(p, q)$ of the error rate under $p(y|x)q(x)$,

$$\text{errbd}(p, q) = \text{Err}(F_{p}^{\text{bayes}}, p, q).$$

Then we obtain the following theorem.

**Theorem 1.** For any $\delta \in \mathcal{D}$, we observe

$$\text{errbd}(p_{\delta}, q) \geq \text{errbd}(p, q).$$

**Proof.** From the Remark 1, a region $\mathcal{X}_{F_{p}^{\text{bayes}}, y}$ coincides with $\mathcal{X}_{F_{p_{\delta}}^{\text{bayes}}, y}$ and then, we observe

$$\text{errbd}(p_{\delta}, q) - \text{errbd}(p, q) = \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}_{F_{p_{\delta}}^{\text{bayes}}, y}} \{p(y|x) - p_{\delta}(y|x)\} q(x)dx.$$

Because of the equation (7), we obtain

$$\text{errbd}(p_{\delta}, q) - \text{errbd}(p, q) = \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}_{F_{p_{\delta}}^{\text{bayes}}, y}} G\{p(y|x) - p_{U}(y)\} \delta(x)q(x)dx.$$  
(10)

Here we observe for any class label $y \in \mathcal{Y}$ that

$$x \in \mathcal{X}_{F_{p_{\delta}}^{\text{bayes}}, y} \Rightarrow p(y|x) = \max_{k \in \mathcal{Y}} p(k|x) \geq p_{U}(y).$$

Hence any integrand in (10) is non-negative, which completes the proof. $\square$

Finally, we consider which form of the mislabel probability $\delta(x)$ is appropriate. In practical situation, it is just conceivable that a failure of labeling in the observation process is apt to occur in a region where the classification based on the original posterior $p(y|x)$ is difficult. Then it is necessary to model $\delta(x)$ in accordance with the difficulty of the classification task. For this purpose, it is natural to consider the following situation: if the feature vector $x$ is in a region of $\mathcal{X}$ in which a maximum of $p(y|x)$ on $y \in \mathcal{Y}$ is almost 1, then $\delta(x)$ should smaller, while if $x$ is in a region of $\mathcal{X}$ in which the maximum is almost $1/G$, then $\delta(x)$ should be larger. The easiest case is suggested when $p(y|x)$ becomes $I(y = F_{p}^{\text{bayes}}(x))$, while the most difficult case is suggested when $p(y|x)$ becomes $p_{U}(y)$. To quantify a degree of the classification easiness, we use the Kullback-Leibler divergence between $p_{U}(y)$ and $p(y|x)$,

$$d_{Y}(p_{U}, p(\cdot|x)) = \sum_{y \in \mathcal{Y}} p_{U}(y) \log \frac{p_{U}(y)}{p(y|x)}$$

$$= \log \frac{1}{G} - \frac{1}{G} \sum_{y \in \mathcal{Y}} \log p(y|x).$$

(11)

(12)
which measures the discrepancy between distributions and satisfies that
\[ p(\cdot|x) \rightarrow p_U \Rightarrow d_Y(p_U, p(\cdot|x)) \rightarrow 0, \]
and
\[ p(\cdot|x) \rightarrow 1(y = F_{\text{Bayes}}(x)) \Rightarrow d_Y(p_U, p(\cdot|x)) \rightarrow \infty. \]
In the following section 4, we will proposed a learning algorithm associated with a mislabeling probability
\[ \delta(x) = v(d_Y(p_U, p(\cdot|x))) \]
where \( v \) is a monotonically decreasing fixed function such that \( v(0) \leq 1/G \) and \( \lim_{t \to \infty} v(t) = 0. \)
While the discussion here is consistently based on the model (6), the model is rather restrictive. For example, we can extend to the model to consider a situation in which a mislabeling probability depends on the class label \( y \) as follows:
\[
\begin{align*}
    p(\cdot|y) &= Z(x) \left\{ \left( 1 - \sum_{k \neq y} \delta_k(x) \right) p(y|x) + \sum_{k \neq y} \delta_k(x) p(k|x) \right\} \\
    &= Z(x) \left\{ \left( 1 - \sum_{k \in Y} \delta_k(x) \right) p(y|x) + \sum_{k \in Y} \delta_k(x) p(k|x) \right\},
\end{align*}
\]
where \( \Delta = (\delta_1, \cdots, \delta_G) \), \( \delta_k(x) \in \mathcal{D} \) is a probability of mislabeling in which the label is erroneously observed as \( k \) and
\[
Z(x) = \left( 1 - \sum_{k \in \mathcal{Y}} \delta_k(x) \right) + G \sum_{k \in \mathcal{Y}} \delta_k(x) p(k|x).
\]
This model would be much more complex and flexible than the model (6), however in this paper, we focus on the model (6).

3 Bregman \( U \)-loss function and constant volume bias condition

Throughout the present paper, we assume that the probability density function of the feature vector \( x \) is fixed to \( q(x) \) and \( \bar{q}(x) \) is its empirical version. Let
\[
\mathcal{P} = \left\{ m(y|x) \left| x \in \mathcal{X}, m(y|x) > 0, \sum_{y \in \mathcal{Y}} m(y|x) = 1 \right. \right\}
\]
be the space of all the conditional probability distributions of $Y$ given $X = x$ and let

$$
\mathcal{M} = \left\{ m(y|x) \Big| x \in \mathcal{X}, m(y|x) > 0, \sum_{y \in \mathcal{Y}} m(y|x) < \infty \right\}
$$

be the extended space of $\mathcal{P}$ with positive finite mass. Thus $m(y|x)$ in $\mathcal{M}$ does not always satisfy $\sum_{y \in \mathcal{Y}} m(y|x) = 1$.

In this formulation, we take a convex and monotonically increasing function $U$ on the real line. For $m, \mu \in \mathcal{M}$, the Bregman $U$-divergence $D_U$ defined over $\mathcal{M} \times \mathcal{M}$ is given by

$$
D_U(m, \mu) = \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} \Phi(m, \mu) dx.
$$

The function $\Phi(a, b)$ is defined as

$$
\Phi(a, b) = U \left( a \right) - U \left( b \right) - (a - b) U \left( 0 \right),
$$

where $u(z) = U'(z)$ and $u^{-1}$ is the inverse function of $u$. From the convexity of $U$, $\Phi(a, b) \geq 0$ for any $a > 0, b > 0$. We observe that $D_U(m, \mu) \geq 0$ by definition. Also, if we set $U(z) = \exp(z)$, $U$-divergence is reduced to the extended Kullback-Leibler divergence.

We consider a pseudo conditional probability $\mu(y|x)$ in $\mathcal{M}$, which is connected with a discriminant function $H(x, y)$ as

$$
\mu(y|x) = u \left( H(x, y) - b_H(x) \right),
$$

where $b_H(x)$ is a bias function depending on the $H(x, y)$ but not on the label $y$. In [10], we proposed bias functions satisfying the normalized condition and the moment matching condition, and investigated statistical properties. The function $u(z)$, the derivative of $U(z)$, is the positive monotone increasing function from the convexity and monotonicity of $U(x)$ and the classification rule associated with $H(x, y)$ is not changed by $u(z)$ or any choice of $b_H(x)$, that means

$$
\arg\max_{y \in \mathcal{Y}} H(x, y) = \arg\max_{y \in \mathcal{Y}} u(H(x, y)) = \arg\max_{y \in \mathcal{Y}} u(H(x, y) - b_H(x)).
$$

By substituting $m(y|x) = p(y|x)$ and $\mu(y|x) = u(H(x, y) - b_H(x))$ into $D_U(m, \mu)$ respectively and omitting terms that do not affect an optimization of $H(x, y)$, we obtain the $U$-loss function for the discriminant function $H(x, y)$ as

$$
L_U(H) = \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} \left\{ U \left( H(x, y) - b_H(x) \right) - p(y|x)(H(x, y) - b_H(x)) \right\} dx.
$$

See [10] for detailed discussions. In fact, this $U$-loss function is equipped with a reasonable property in which the loss has a minimum when $H(x, y)$ generates
the Bayes rule. Thus the function \( U \) generates the \( U \)-divergence, and the \( U \)-loss function \( L_U \). Then a variety of the \( U \)-loss functions corresponds to that of the function \( U \). An empirical version of the loss (17) and a general algorithm derived from it, are shown in Appendix A.

In the next subsection, we propose a new bias function, a constant volume condition, which produces a statistical model associated with the function \( U \).

### 3.1 Condition of the bias function and associative model

Let us discuss the class of optimal discriminant functions which are obtained by minimizing the given \( U \)-loss function (17) associated with three types of conditions for the bias function \( b_H(x) \). For the first condition, we obtain the following theorem.

**Theorem 2 (Constant volume condition).** Let \( H^* \) be a minimizer of \( U \)-loss function \( L_U(H) \) under the condition:

\[
\sum_{y \in Y} U(H(x, y) - b_H(x)) = C,
\]

where \( C \) is a constant. Then \( H^* \) satisfies

\[
p(y|x) = \frac{u(H^*(x, y) - b_{H^*}(x))}{\sum_{y' \in Y} u(H^*(x, y') - b_{H^*}(x))},
\]

and hence \( H^* \) is Bayes optimal.

**Proof.** We seek a minimizer of (17) under constraint (18) and find that it is sufficient to minimize (17) conditional on \( x \) (cf. [11]). By setting a variational derivative of (17) conditional on \( x \) to 0, we observe

\[
0 = \frac{\partial}{\partial H(x, y)} \sum_{y' \in Y} \{U(H(x, y') - b_H(x)) - p(y'|x)(H(x, y') - b_H(x))\} \bigg|_{H=H^*}
\]

\[
= -p(y|x) + u(H^*(x, y) - b_{H^*}(x))
\]

\[
+ \left( 1 - \sum_{y' \in Y} u(H^*(x, y') - b_{H^*}(x)) \right) \frac{\partial b_{H^*}(x)}{\partial H(x, y)} \bigg|_{H=H^*}.
\]

Also, we observe the following relation from a variational derivative of the condition (18).

\[
\forall y \in Y, \frac{\partial b_{H^*}(x)}{\partial H(x, y)} = \frac{u(H(x, y) - b_{H^*}(x))}{\sum_{y' \in Y} u(H(x, y') - b_{H^*}(x))}.
\]

By substituting (21) into (20), we conclude (19). The denominator of the right-hand side of (19) does not depend on \( y \) and the function \( u \) is the monotonically increasing one. Thus, \( H^* \) is Bayes optimal.

\( \square \)
We describe condition (18) as the constant volume condition. Note that we shall denote the object under condition (18) by one with a superscript \( v \) as a \( U \)-loss function \( L_U^v \) or a bias function \( b_H^v \). The constant value \( C \) is conventionally fixed to \( G_U(0) \), where we impose \( b_H^v(x) = 0 \) as \( H(x, 1) = \cdots = H(x, G) = 0 \).

Under condition (18), the \( U \)-loss function is reduced to

\[
L_U^v(H) = C - \int_X q(x) \sum_{y \in \mathcal{Y}} p(y|x)(H(x, y) - b_H^v(x)) \, dx.
\]

We will omit the first term \( C \) since the optimization of the discriminant function \( H \) does not depend on the term. An algorithm is derived from an empirical version of (22) as in the case of Appendix A and we note an error rate property of the algorithm under the constant volume condition in Appendix B.

For the second and third conditions, detailed properties have been previously investigated in [10], and we observe the following corollaries.

**Corollary 1** (Normalized condition). Let \( H^* \) be a minimizer of \( U \)-loss function \( L_U(H) \) under the following conditions:

\[
\sum_{y \in \mathcal{Y}} u(H(x, y) - b_H(x)) = 1.
\]

Then \( H^* \) satisfies (19) and is Bayes optimal.

**Proof.** From equation (20) and condition (23), we immediately obtain the relation

\[
p(y|x) = \frac{u(H^*(x, y) - b_H(x))}{\sum_{y' \in \mathcal{Y}} u(H^*(x, y') - b_H(x))}.
\]

Furthermore, this relationship shows that \( H^* \) is Bayes optimal. \hfill \Box

**Corollary 2** (Moment matching condition). Let \( H^* \) be a minimizer of \( U \)-loss function \( L_U(H) \) under the conditions:

\[
\sum_{y \in \mathcal{Y}} p(y|x)(H(x, y) - b_H(x)) = 0.
\]

Then, \( H^* \) satisfies (19) and is Bayes optimal.

**Proof.** By differentiating condition (25), we observe

\[
0 = \left. \frac{\partial}{\partial H(x, y)} \sum_{y' \in \mathcal{Y}} p(y'|x)(H(x, y') - b_H(x)) \right|_{H=H^*} = \left. p(y|x) - \frac{\partial b_H(x)}{\partial H(x, y)} \right|_{H=H^*}.
\]

By substituting (26) into (20), we conclude that (19), which shows the Bayes optimality of \( H^* \). \hfill \Box
If we set $U(z) = \exp(z)$, the right-hand side of (19) with those three types of bias functions, reduces to be the logistic model, which is used in the context of conventional statistical classification.

We describe condition (23) as a normalized condition and condition (25) as a moment matching condition. Details of the properties of those two bias conditions in the context of the Boosting algorithm were discussed in [10]. The normalized condition constrains the pseudo conditional model (16) in $M$ to be the probabilistic model in $P$, and derives the conventional logistic discriminant analysis with the function $U(z) = \exp(z)$. It is pointed out that the original AdaBoost is derived from the minimization of the extended Kullback-Leibler divergence between the empirical distribution and the extended exponential model in $M$ under the moment matching condition (See [5]). This condition is also used in statistical inference. In the following context, we denote an object under the normalized condition (23) with a superscript $n$, and an object under the moment matching condition (25) with a superscript $m$, as $L_nU$, $b_nH$, $L_mU$ and $b_mH$, respectively. The $U$-loss function $L_mU$ is written as

$$L_mU(H) = \int_X q(x) \sum_{y \in \mathcal{Y}} U(H(x, y) - b_m(x)) dx.$$  

(27)

3.2 Consistency for the binary case

For the binary classification problem, we can assume that the label $y$ is in $\{1, -1\}$ for convenience of calculation, and then the discriminant function $H(x, y)$ is represented by $H(x, y) = F(x)y$ where a label is predicted by a sign of the function $F(x)$, $\text{sgn}(F(x))$. Then the $U$-loss function (17) is rewritten as

$$L_U(F) = \int_X q(x) \sum_{y \in \{1,-1\}} U(F(x)y - b(x; F)) - p(y|x)(F(x)y - b(x; F)) dx.$$ (27)

For the constant volume condition and the normalized condition, the function $b(x; F)$ is represented as $b(F(x))$, since $b(x; F) = b(x; -F)$ holds and $b$ depends on $x$ through $F(x)$. Let us consider a function

$$V(z) = U(-z - b(z)) + U(z - b(z)) + z + b(z).$$

Then the above loss function (27) is rewritten as $V$-loss function as follows:

$$L_V(F) = \int_X q(x) \{p(1|x)V(-F(x)) + p(-1|x)V(F(x))\} dx.$$ (28)

For a fixed $x$, we focus on a point-wise property of the $V$-loss and denote $p(1|x)$ by $\xi$ and $F(x)$ by $\gamma$. Let us define

$$C_\xi(\gamma) = \xi V(-\gamma) + (1 - \xi)V(\gamma),$$ (29)

for $0 \leq \xi \leq 1$. The large sample behavior of the algorithm based on the above $V$-loss function was investigated.
Definition 1. The \( V \)-loss is classification calibrated if, for any \( \xi \neq \frac{1}{2} \),
\[
\inf_{\gamma \in (2\xi - 1) \leq 0} C_{\xi}(\gamma) > \inf_{\gamma \in R} C_{\xi}(\gamma).
\]

The classification calibration implies that a minimizer of \( C_{\xi}(\gamma) \) with respect to \( \gamma \in R \) has the same sign as \( \xi - \frac{1}{2} \). If the loss function \( V \) is classification calibrated, asymptotic convergence to the optimal value of loss implies that the decision function reaches the Bayes optimal decision rule. The following theorem is shown in [12].

Theorem 3. The following conditions are equivalent.

1. \( V \)-loss is classification calibrated.

2. For every sequence of measurable functions \( F_i : X \to R \), \( i = 1, 2, \ldots \), and every probability distribution \( q(x) p(y|x) \) on \( X \times \{1, -1\} \),
\[
\lim_{i \to \infty} \int_X q(x) \sum_{y \in Y} p(y|x)V(-yF_i(x))dx = \int_X q(x) \sum_{y \in Y} p(y|x)V(-yF_{\text{bayes}}(x))dx
\]
implies
\[
\lim_{i \to \infty} \int_X q(x) \sum_{y \in Y} p(y|x)I(Y \neq \text{sgn}(F_i(x)))dx
= \int_X q(x) \sum_{y \in Y} p(y|x)I(Y \neq \text{sgn}(F_{\text{bayes}}(x)))dx.
\]

Moreover, the following lemma has been claimed.

Lemma 1. If \( V(z) \) is a convex, differentiable at \( z = 0 \) and \( V'(0) > 0 \) holds, the \( V \)-loss is classification calibrated.

From the Lemma 1, we obtain the following two lemmas for the \( U \)-loss under the constant volume condition and the normalized condition. Note that in the following, we will denote \( b(z) \) by \( b \) and a derivative with respect to \( \gamma \) by \( \gamma \) if the context is clear.

Lemma 2. Under the constant volume condition, the \( U \)-loss is classification calibrated.

Proof. From the condition (18), we obtain
\[
U(\gamma - b) + U(-\gamma - b) = C,
\]
and the function \( V \) is written as \( V'(\gamma) = C + \gamma + b' \) where \( b' \) is a bias function satisfying (31). By differentiating the condition (31) with respect to \( \gamma \), we obtain
\[
0 = u(\gamma - b')(1 - b') + u(-\gamma - b'(-1 - b'),
\]

and equivalently
\[ b^v = \frac{-u(-\gamma - b^v) + u(\gamma - b^v)}{u(-\gamma - b^v) + u(\gamma - b^v)}. \]

Then we observe that the first and second derivatives of \( V^v \) with respect to \( \gamma \) are

\[
\begin{align*}
V^v(\gamma) &= \frac{2u(\gamma - b^v)}{u(-\gamma - b^v) + u(\gamma - b^v)}, \\
V^{vv}(\gamma) &= \frac{4u(-\gamma - b^v)u'(\gamma - b^v) + u(\gamma - b^v)^2u'(-\gamma - b^v)}{(u(-\gamma - b^v) + u(\gamma - b^v))^2}.
\end{align*}
\]

Now we assumed that the function \( U \) is the convex function with a positive derivative. Consequently, we observe \( V^{v\prime}(\gamma) > 0 \) which implies a convexity of \( V^v(\gamma) \) and \( V^{\prime\prime}(0) = 1 > 0 \). Then the \( U \)-loss under the constant volume condition is classification calibrated. \( \square \)

**Lemma 3.** Under the normalized condition, the \( U \)-loss is classification calibrated.

**Proof.** The function \( V \) is written as \( V^n(\gamma) = U(-\gamma - b^n) + U(\gamma - b^n) + \gamma + b^n \) where \( b^n \) is a bias function satisfying the normalized condition (23). Using the condition, we obtain

\[ u(\gamma - b^n) + u(-\gamma - b^n) = 1. \]

The first and second derivative of \( V^n \) are calculated as

\[
\begin{align*}
V^{n\prime}(\gamma) &= u(-\gamma - b^n)(-1 - b^n\prime) + u(\gamma - b^n)(1 - b^n\prime) + 1 + b^n\prime, \\
V^{n\prime\prime}(\gamma) &= u'(-\gamma - b^n)(-1 - b^n\prime)^2 + u'(\gamma - b^n)(1 - b^n\prime)^2.
\end{align*}
\]

From those equations, we observe

\[ V^{n\prime}(0) > 0, V^{n\prime\prime}(\gamma) > 0, \]

which concludes that \( U \)-loss is classification calibrated under the normalized condition. \( \square \)

For the moment matching condition, the form of loss function is differ from the one of \( V \)-loss (28), because the bias function is defined using the posterior distribution \( p(y|x) \). Then \( U \)-loss under the moment matching condition is definitely not classification calibrated. However, we can remark the following property, which seems to be essential for results in [12].

**Remark 2.** Under the moment matching condition, the \( U \)-loss satisfies (30).

For the multi-class problem, the concept of the classification calibrated was proposed [13]. It will be a future work that \( U \)-losses for multi-class problems are also classification calibrated.
For the special form of the function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z$, robust methods in the binary labels setting have been proposed under both the normalized and the moment matching conditions. Model (6) with the constant mislabeling probability, which is associated with the normalized condition, was statistically analyzed in [7]. The $\eta$-Boost for binary classification problems in [4], which is a modified version of AdaBoost for the purposes of robustness, is introduced from the moment matching condition and is associated with a probabilistic model of $x$-dependent mislabeling. The associative model suggests that the larger the proportion of mislabeling, the nearer the feature vector is to the decision boundary. This model is useful and practical in many applications of pattern recognition. A major objective of this paper is to extend this $\eta$-Boost for binary classification to that for multi-class classification problems, and we recognize failure of the direct extension of binary-class $\eta$-Boost as based on the moment matching condition. Therefore, we present a new idea for an associated model of mislabeling in multi-class situations based on the constant volume condition (18). In the following section, we primarily focus on the constant volume condition (18), which leads to a mislabeling model (13) suitable for the multi-class problem. Before that, however, we consider the normalized condition (23) that leads to the constant mislabeling probability model by way of introduction.

4 Eta-divergence and Eta-loss

Let us focus on the specific form of the function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z (0 \leq \eta < 1)$ in the class of the Bregman $U$-divergences. By applying the argument in Section 3 to $U_\eta$, we observe that Eta-divergence is

$$D_{U_\eta}(m, \mu) = \int_X q(x) \sum_{y \in Y} \left\{ m(x, y) - m(x, y) + (m(x, y) - \eta) \log \frac{m(x, y) - \eta}{m(x, y) - \eta} \right\} dx,$$

and so the corresponding loss functions are

$$L_\eta(H) = \int_X q(x) \sum_{y \in Y} \left\{ (1 - \eta) \exp \left( H(x, y) - b_H(x) \right) + \eta(H(x, y) - b_H(x)) \right\} dx.$$

Fixing $\eta = 0$, we obtain the extended Kullback-Leibler divergence $D_{U_0}$ and observe the relationship $D_{U_0}(p, q) = D_{U_\eta}(p + \eta, q + \eta)$.

As noted above, we first introduce a method associated with the normalized condition, which leads to a constant probability of mislabeling; second, we discuss a method associated with the constant volume condition, which induces the model of mislabeling, (13).
4.1 The normalized Eta-Boost.M: The normalized condition and the associative model

The normalized condition given in (23) constrains the pseudo conditional model in (16) to be the probabilistic model. From the normalized condition with the function

$$u(z) = U(z) = (1 - \eta) \exp(z) + \eta,$$

we obtain the following relation:

$$1 = \sum_{y \in Y} u_y(H(x, y) - b_H^n(x))$$

$$= (1 - \eta) \exp(-b_H^n(x)) \sum_{y \in Y} \exp(H(x, y)) + \eta G.$$  

The bias function can be written as

$$b_H^n(x) = \log \frac{1 - \eta}{1 - G\eta} + \log \sum_{y \in Y} \exp(H(x, y)), \quad (34)$$

and the loss function is written as

$$L^n_H(H) = (1 - \eta G) \left( 1 + \log \frac{1 - \eta}{1 - G\eta} \right) - \int x q(x) \sum_{y \in Y} (p(y|x) - \eta) \log p_{0,H}(y|x) dx, \quad (35)$$

where $p_{0,H}(y|x)$ is the logistic model constructed by the discriminant function $H(x, y)$, defined by (5). If $\eta$ is set to 0, the above loss function is a negative log likelihood of the logistic model, and the optimization of (35) leads to the logistic discriminant analysis. From the loss function (35), we can derive normalized Eta-Boost.M as shown in Appendix A.

From Lemma 1, we observe the following relationship between the true conditional distribution $p(y|x)$ of data and the optimal discriminant function $H^*(x, y) = \arg\min_H L^n_H(H)$:

$$p(y|x) = \frac{u_y(H^*(x, y) - b_H^n(x))}{(1 - \eta) \sum_{y' \in Y} \exp(H^*(x, y')) \exp(H^*(x, y)) + \eta}$$

$$= \frac{(1 - \eta) \exp(-b_H^n(x)) \sum_{y \in Y} \exp(H^*(x, y)) + \eta}{(1 - \eta) \sum_{y' \in Y} \exp(H^*(x, y')) \exp(H^*(x, y)) + \eta}$$

$$= \frac{(1 - \eta(G - 1)) p_{0,H^*}(y|x) + \eta \sum_{y' \neq y} p_{0,H^*}(y'|x)}{(1 - \eta(G - 1)) p_{0,H^*}(y|x) + \eta \sum_{y' \neq y} p_{0,H^*}(y'|x)}. \quad (36)$$

This model represents the mislabeling model with constant probability $\eta$ and is a multi-class version of [7]. In the following subsection, we will introduce a model of mislabeling that depends on the feature $x$, derived from the constant volume condition (18) with the form of function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z$.

4.2 The constant volume condition and the associative model

Under the constant volume condition with the function $U_\eta(z)$, we obtain the following theorem about the explicit form of the bias function $b_H^n(x)$. 


Theorem 4. When we apply the constant volume condition (18) to the function $U_\eta(z) = (1 - \eta)\exp(z) + \eta z$ with the typical choice of constant $C = GU_\eta(0)$, we obtain the explicit form of the bias function $b^*_H$ by solving the following differential equation (21):

\[
b^*_H(x) = \begin{cases} 
\log \frac{1}{G} + \log \sum_{y' \in \mathcal{Y}} \exp(H(x, y')), & \eta = 0, \\
\frac{1}{G} \sum_{y' \in \mathcal{Y}} H(x, y') - \frac{1}{\eta} + K(Q(H)), & 0 < \eta < 1,
\end{cases}
\]

(37)

where $K(z)$ is the inverse function of $k(z) = z\exp(z)$ and

\[
Q(H) = k\left(\frac{1 - \eta}{\eta}\right) \frac{1}{G} \sum_{y' \in \mathcal{Y}} \exp \left( H(x, y') - \frac{1}{G} \sum_{y' \in \mathcal{Y}} H(x, y') \right)
\]

= \left(\frac{1 - \eta}{\eta}\right) \exp(d_\mathcal{Y}(p_U, p_{0,H}(\cdot|x))).
\]

(38)

Here, $d_\mathcal{Y}(\cdot, \cdot)$ is the Kullback-Leibler divergence on $\mathcal{Y} \times \mathcal{Y}$, defined by (11). The $p_U(y|x)$ is the uniform distribution on $\mathcal{Y}$ given $x$ and $p_{0,H}$ is the logistic model (5) constructed by $H$.

Proof. In the following, we will denote $H(x, y)$ by $H_y$ and $b^*_H(x)$ by $b$ when there is no risk of causing confusion.

First, we discuss the case where $\eta = 0$. From the constant volume condition (18), we immediately obtain the relation

\[
\sum_{y \in \mathcal{Y}} \exp(H_y - b) = C = G \exp(0).
\]

(39)

The bias function with $\eta = 0$ is thus written as

\[
b = \log \frac{1}{G} + \log \sum_{y \in \mathcal{Y}} \exp(H_y).
\]

Second, we proceed to the case where $0 < \eta < 1$. By setting

\[
r = \sum_{y \in \mathcal{Y}} \exp(H_y - b),
\]

(40)

we obtain

\[
\frac{\partial r}{\partial H_y} = \exp(H_y - b) - \frac{\partial b}{\partial H_y} \frac{\sum_{y' \in \mathcal{Y}} \exp(H_{y'} - b)}{\sum_{y' \in \mathcal{Y}} \exp(H_{y'} - b) + \eta}
\]

\[
= r \left( p_{0,H}(y|x) - \frac{\partial b}{\partial H_y} \right),
\]

\[
\frac{\partial b}{\partial H_y} = \frac{(1 - \eta) \exp(H_y - b) + \eta}{\sum_{y' \in \mathcal{Y}} \{ \exp(H_{y'} - b) + \eta \}}
\]

\[
= \frac{(1 - \eta) r p_{0,H}(y|x) + \eta}{(1 - \eta) r + \eta G}.
\]
From the first line to the second line, we use (21) with the form \( u = u_y \). From the above two equations, we observe that

\[
\frac{\partial r}{\partial H_y} = \frac{r y G}{(1 - \eta) r + \eta G} \left\{ p_0, H(y|x) - \frac{1}{G} \right\}. \tag{41}
\]

It is possible to solve (41) by the separation of variables:

\[
\int \left( \frac{1 - \eta}{\eta G} + \frac{1}{r} \right) \, dr = \int \left( p_0, H(y|x) - \frac{1}{G} \right) \, dH_y, \tag{42}
\]

\[
\frac{(1 - \eta) r}{\eta G} + \log r = \log \sum_{y' \in \mathcal{Y}} \exp(H_{y'}) - \frac{H_{y}}{G} + C, \tag{43}
\]

where \( C \) is a function of \( H_{y'} \), \( y' \neq y \). Since we can solve as same for all \( y \in \mathcal{Y} \), the following equation is obtained:

\[
\frac{(1 - \eta) r}{\eta G} + \log r = \log \sum_{y' \in \mathcal{Y}} \exp(H_{y'}) - \frac{1}{G} \sum_{y' \in \mathcal{Y}} H_{y'} + C', \tag{44}
\]

where \( C' \) is a constant term that does not depend on \( H_1, \ldots, H_y \). If \( C' \) is subject to the initial condition that \( C = GU_0(0) \), which implies \( b' = 0 \) when \( H_1 = \cdots = H_G = 0 \), we obtain \( C' = \frac{-b}{\eta} \). We can rewrite (44) as

\[
\frac{1 - \eta}{\eta G} r \exp \left( \frac{1 - \eta}{\eta G} r \right) = k \left( \frac{1 - \eta}{\eta G} r \right) \tag{45}
\]

where \( k \) is a constant term. By applying (47) to (44), we obtain

\[
b = \log \frac{1 - \eta}{\eta G} + \log \sum_{y \in \mathcal{Y}} \exp(H_y) - \log K(Q(H)). \tag{46}
\]

Additionally, we observe that

\[
\log z = \log k(K(z)) = \log (K(z) \exp(K(z))) = K(z) + \log K(z). \tag{47}
\]

By applying (47) to (44), we obtain

\[
b = \log \frac{1 - \eta}{\eta G} + \log \sum_{y' \in \mathcal{Y}} \exp(H_{y'}) + K(Q(H)) - \log Q(H)
\]

\[
= \frac{1}{G} \sum_{y' \in \mathcal{Y}} H_{y'} - \frac{1 - \eta}{\eta} + K(Q(H)). \tag{48}
\]

Thus, we conclude the bias function (37). \( \square \)
Note that since $Q(H)$ takes a positive value for any discriminant function $H(x, y)$, $K(z)$ is the positive and monotonically increasing function in our context. The loss function with the bias function (37) has the form of (22) and we observe that the loss function with $\eta = 0$, $L_0^0(H)$, also coincides with the negative log likelihood of the logistic model (5), just as in the case of the loss function (35) with $\eta = 0$ except for a constant term. The algorithm Eta-Boost.M is derived from the sequential minimization of the loss function $L_0^0(H)$.

We discuss a relation between $L_0^0(H)$, the Eta-loss function (33) with the constant volume condition and the optimal discriminant function $H^*(x, y)$. From the theorem 2, we observe

$$p(y|x) = \frac{(1 - \eta) \exp(H^*(x, y) - b^*_H(x)) + \eta}{\sum_{y' \in \Omega} (1 - \eta) \exp(H^*(x, y') - b^*_H(x)) + \eta G}$$

$$= (1 - (G - 1)\delta(H^*)) p_{0,H^*}(y|x) + \delta(H^*) \sum_{y' \neq y} p_{0,H^*}(y'|x), \quad (49)$$

where $\delta(H)$ can be interpreted as a probability of mislabeling,

$$\delta(H) = \frac{\eta}{\sum_{y' \in \Omega} (1 - \eta) \exp(H(x, y') - b^*_H(x)) + \eta G} \frac{1}{G \{1 + K\{k \left(\frac{1 - \eta}{\eta}\right) \exp(d_Y(p_U, p_{0,H^*}(\cdot|x)))\}\}}.$$ 

(50)

The $\delta(H^*)$ represents greater association depending on $x$ in the following sense. From the property of the KL-divergence, we obtain

$$\delta(H^*) \leq \frac{1}{G \{1 + K\{k \left(\frac{1 - \eta}{\eta}\right)\}\}} = \frac{\eta}{G}$$

(51)

and the equality holds if and only if $p_{0,H^*} = p_U$. The optimal discriminant function $H^*(x, y)$ has no information about a class label of the feature $x$. The probability of mislabeling, $\delta(H)$, rapidly decreases as the feature vector $x$ diverges from the common decision boundary because the function $K(\cdot)$ with a non-negative argument is a monotonically increasing function. Figure 1 shows $\delta(H)$ against $\eta$ and KL-divergence, $d_Y(p_U, p_{H,0}(\cdot|x))$ at $G = 10$. Eta-boost.M provides a mislabeling mechanism such that a smaller proportion is given in accordance with the difficulty of classification of the input $x$. In other words, Eta-Boost.M reduces risk of a hasty decision for a region where it is difficult to discriminate.

4.3 Comparison of Eta-Boost.M with the binary $\eta$-Boost

For the binary classification problem with a label $y \in \{1, -1\}$, the binary $\eta$-Boost has been proposed by considering the model of mislabeling associated
with the input $x$ in [4]. In this subsection, we compare Eta-Boost.M with the binary $\eta$-Boost algorithm. The binary $\eta$-Boost algorithm includes the original AdaBoost as a special case and is derived from the sequential minimization of the mixture of the exponential loss function and the naive error loss function. Although a formal extension of binary $\eta$-Boost to a multi-class classification problem is possible, the statistical properties related to the mislabeling are unclear as in the previous subsection. Consequently, we focus on the degree of deviation from the uniform distribution as a criterion for classification difficulty, and technically leap from the binary $\eta$-Boost to Eta-Boost.M by introducing the constant volume condition.

In the binary $\eta$-Boost algorithm, the discriminant function $H$ is defined by $H(x, y) = F(x)y$, where $F(x)$ is a binary classifier, and a probability of mislabeling is modeled as the following form:

$$\delta'(H) = \frac{1}{2 \left\{ 1 + \frac{1}{\eta} \exp(d_Y(p_U, p_{H, 0}(\lvert x \rvert))) \right\}}.$$  \hfill (52)

Then the mislabeling probability (50) coincides with (52) if we substitute $k(z) = z \exp(z)$ for the identity function $z$. If the quantity $d_Y(p_U, p_{H, 0}(\lvert x \rvert))$ is sufficiently small, we can approximate (50) and (52) as

$$\delta(H) \approx \frac{\eta}{G} - \frac{\eta^2(1 - \eta)}{G} d_Y(p_U, p_{H, 0}(\lvert x \rvert)), \hfill (53)$$

Figure 1: The probability of mislabeling against $\eta$ and KL-divergence, at $G = 10$. 

![Figure 1](image-url)
Those equations imply that Eta-Boost.M assumes a higher probability of mislabeling compared with binary \( \eta \)-Boost and tends to make a conservative decision about class labels. Figure 2 shows the probability of mislabeling of each method against \( \eta \) when we fix the KL-divergence \( d_Y \) to 0.5.

\[
\delta'(H) \approx \frac{\eta}{G} + \frac{\eta(1-\eta)}{G} d_Y(p_{U,H,H}, p_{U,H,0}(\cdot|x)).
\]

(54)

Figure 2: The probability of mislabeling of two methods against \( \eta \). The quantity \( d_Y \) is fixed to 0.5.

### 4.4 Weight property of Eta-Boost.M

The general algorithm derived from the \( U \)-loss function \( L_U(H) \) is described in Appendix A, and the error rate property of the algorithm under the constant volume condition is shown in Appendix B. In this section, we discuss the property of the weight associated with the function \( U_\eta \) under the normalized condition and the constant volume condition. From (63) and (36) or (49), the weight function under the normalized condition, \( D^n_{i,y}(i,y) \), and one under the constant volume condition, \( D^v_{i,y}(i,y) \), are written as

\[
D^n_{i,y}(i,y) = \frac{1}{N} \left( (1-\eta(G-1))p_{0,H_{-i}}(y|x) + \eta \sum_{y' \neq y} p_{0,H_{-i}}(y'|x) \right),
\]

(55)
and

\[ D^t(i, y) = \frac{1}{N} \left( (1 - \delta(H_{t-1})(G - 1))p_0H_{t-1}(y|x) + \delta(H_{t-1}) \sum_{y' \neq y} p_0H_{t-1}(y|x) \right). \]  (56)

While the weight function (55) assumes the constant probability of mislabeling throughout the algorithm’s process, (56) depends on the probability of mislabel \( \delta(H_{t-1}) \) when the learning process proceeds because \( d_Y(p_U; p_0H_{t-1}(:|x)) \) tends to take larger values as the algorithm progresses. This implies that the weight function is regularized to avoid excessively premature discrimination.

5 Simulation studies

In this section, we investigate the performance of the proposed methods using artificial datasets in the context of the multi-class classification. Also, we apply methods to UCI repository datasets [14]. For comparison, we employ several methods, namely AdaBoost.M2, AdaBoost.M1W type algorithm [15] which is Grploss method [3] with decision stumps making hard decisions, linear discriminant analysis, quadratic discriminant analysis, and rpart (a decision tree in the R [16]). The weak hypothesis (4) is constructed by the decision stump,

\[ \hat{y}(x) = \begin{cases} S, & x_j \geq b, \\ \mathcal{Y} \setminus S, & \text{otherwise}, \end{cases} \]

where \( S \) is a subset of \( \mathcal{Y} \), \( x_j \) is the \( j \)-th variable of \( x \), and \( b \) is a threshold value.

In our experiments, we considered all patterns as the subset \( S \).

5.1 Synthetic dataset

We examined a case in which a training dataset was generated by model (49) and the optimal discriminant function is

\[ H^*(x, y) = \begin{cases} x_1, & y = 1, \\ -x_1, & y = 2, \\ 20 - \frac{x_1^2 + x_2^2}{3}, & y = 3. \end{cases} \]  (57)

Let \( q(x) \) be the probability density function of the two-dimensional uniform distribution on \((-10, 10) \times (-10, 10)\), and we fix \( \eta = 0.5 \) in the conditional probability of class label (49). We generated 50 different training datasets and test datasets and investigated the averaged performance of Eta-Boost.M. Tuning parameters \( \eta, T \) were determined by 5-fold cross-validation for each training dataset. Test dataset contains 3000 observations and a typical training dataset containing 300 observations and the classification boundary of the Bayes rule are shown in Figure 3.

Contaminated examples were particularly observed near the boundary. The probability of mislabeling (50) associated with (57) is shown in Figure 4 and we
Figure 3: A typical training dataset and a classification boundary of Bayes rule.

Figure 4: The mislabeling probability (50) associated with (57).
can confirm that the probability has the maximum value at the intersection of boundaries of all classes.

Figure 5 shows the averaged test errors of Eta-Boost.M, normalized Eta-Boost.M, AdaBoost.M2 and AdaBoost.M1W method in 50 different datasets against the learning step \( t \). We observed that Eta-Boost.M at the learning step \( T_{\text{Eta}} \), performed well and attained lower error compared to AdaBoost.M2 at the learning steps \( T_{\text{Ada}} \).

![Graph showing test errors comparison](image)

Figure 5: The average of test errors of Eta Boost.M(\( \eta = 0.2 \)), AdaBoost.M2, normalized Eta-Boost.M(\( \eta = 0.05 \)) and AdaBoost.M1W method.

Figure 6 shows the averaged test error and cross-validation error of Eta-Boost.M and test error of normalized Eta-Boost.M against \( \eta \). Eta-Boost.M outperformed AdaBoost.M2, AdaBoost.M1W, Linear discriminant analysis(0.32339), Quadratic discriminant analysis (0.22271) and rpart (0.19827) in terms of test error for appropriate \( \eta \), and we could determine the appropriate \( \eta \) and \( T \) by cross-validation error. The error of normalized Eta-Boost.M tended to diverge as \( \eta \) increased, and we found that appropriate tuning of \( \eta \) was difficult for this dataset. Additionally, we applied multi-class version of SVM ([17]) with a polynomial kernel (0.31471) and an RBF kernel (0.14819). The SVM with the RBF kernel outperformed Eta-Boost.M, however its performance is explained by the non-linearity of the kernel function and depends on a selection of kernel function. Our focus is not on the effect of the kernel function but on the method for combining weak learning machines.

Figure 6 only shows the averaged improvement of the proposed methods,
therefore we examine the superiority of the proposed method to the existing method for each datasets. For this purpose, we have investigated the winning rate of Eta-Boost.M in 50 experiments against other methods, shown in Figure 7. Consequently we see that Eta-Boost.M with an appropriate \( \eta \) outperforms other methods with high probability in terms of test error.

![Figure 6: The average of test errors of Eta-Boost.M and normalized Eta-Boost.M and cross-validation errors of Eta-Boost.M are plotted against \( \eta \). Additionally, averaged test errors of AdaBoost.M2, QDF, rpart and AdaBoost.M1W are plotted.](image)

Figures 8 and 9 respectively show boundaries of AdaBoost.M2 and Eta-Boost.M(\( \eta = 0.2 \)). The optimal learning step for each method is estimated by cross-validation. While the boundary of AdaBoost.M2 is influenced by examples that are difficult to classify, Eta-Boost.M tends to construct a comparatively smooth boundary.

### 5.2 UCI repository

We next applied the proposed methods, Eta-Boost.M and normalized Eta-Boost.M to datasets in the UCI repository. Each dataset was originally divided into a training dataset and a test dataset except for the waveform dataset. The waveform dataset was randomly divided into a training dataset and a test dataset. More detailed information on datasets is given in Table 1.
Figure 7: Winning rate of Eta-Boost.M in 50 runs for other methods in the sense of test error.

First, we applied the proposed methods, AdaBoost.M1W method and AdaBoost.M2 for the dataset. In all applied methods, the number $T$ of the learning step should be determined and Eta-Boost.M and normalized Eta-Boost.M additionally had a hyper parameter $\eta$ that should be appropriately tuned. Those parameters $T$, $\eta$ were determined by 5-fold cross-validation using only the training examples as in the case of the experiment with synthetic datasets. Second, we investigated the performance of the proposed methods using the dataset “Thyroid” and “Waveform” with simulating a random mislabeling in the label observation process, where the class label of the dataset was randomly contaminated with a constant probability. We examined two cases of the contamination probability, 0.1 and 0.2. Note that this contamination process was different from that assumed by the statistical model (49) associated with the constant volume condition.

Training and test performances for the original dataset and those two contaminated datasets are shown in Table 2 and the best performance for each dataset is written in boldface. For contaminated datasets, normalized Eta-Boost.M outperforms the other methods, because the mechanism of mislabeling coincides with assumed one. Additionally, Eta-Boost.M worked well even though assumed mechanism of mislabeling is different from the optimal one, (49).
6 Conclusions

We proposed the new class of the loss function associated with the constant volume condition in the Boosting algorithm derived from the minimization of the Bregman $U$-divergence. As a main result, we derived two kinds of robust Boosting algorithms, normalized Eta-Boost.M and Eta-Boost.M, which naturally associate with mechanisms of mislabeling. The former algorithm considers the uniform contamination among multi-class labels, while the latter makes it possible to consider the contamination depending on the input. The method assumes a larger proportion of mislabeling for a region of input space where it is difficult to discriminate. This mechanism reduces the risk of a hasty decision.

We have numerically explored the robustness of two versions of Eta-Boost.M in regard to the appearance of mislabeling with the synthetic dataset. Selection of the tuning parameter $\eta$ in proposed methods was implemented by the $K$-fold cross-validation error rate, which reinforces advantage of proposed methods over AdaBoost.M2 or other existing methods. Also for datasets in the UCI repository, we investigated performance of proposed methods and confirmed that proposed methods attained better performance on robustness for noisy datasets. Theoretical analysis of the large sample behavior of the algorithm for multi-class problem is remaining as a future work.
Appendix A: A Boosting algorithm associated with $U$-loss

Here we derive a Boosting algorithm associated with the $U$-loss. For a given dataset $\{(x_i, y_i) : i = 1, \cdots, N\}$, the empirical version of $U$-loss (17) is written as

$$L_U(H) = \frac{1}{N} \sum_{i=1}^{N} \left\{ -H(x_i, y_i) + b_H(x_i) + \sum_{y \in \mathcal{Y}} U(H(x_i, y) - b_H(x_i)) \right\}.$$ 

We can obtain the algorithm by the sequential minimization of the above loss function. It is assumed that a discriminant function $H_t(x, y)$ is obtained and consider an update from $H_t$ to $H_t + \alpha h$, where $h$ is a new weak hypothesis in
Table 1: Information on datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Total</th>
<th>#Training</th>
<th>#Test</th>
<th>#Attributes</th>
<th>#Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>2310</td>
<td>210</td>
<td>2100</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Satimage</td>
<td>6435</td>
<td>4435</td>
<td>2000</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>Pendigit</td>
<td>10992</td>
<td>7494</td>
<td>3498</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Waveform</td>
<td>5000</td>
<td>3000</td>
<td>2000</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Thyroid</td>
<td>7200</td>
<td>3772</td>
<td>3428</td>
<td>21</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \mathcal{H} \text{ and } \alpha \text{ is a small constant. We then observe that} \]

\[
\bar{L}_U(H_t + ah) - \bar{L}_U(H_t) \\
\approx \frac{\alpha}{N} \sum_{i=1}^{N} \sum_{y \in Y} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i)) \left( h(x_i, y) - \frac{\partial b_{H_t + ah}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right) \right. \\
- \left. \left( h(x_i, y) - \frac{\partial b_{H_t + ah}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right) \right\} \\
= \frac{\alpha}{N} \sum_{i=1}^{N} \sum_{y \in Y} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i)) - \bar{p}(y|x) \right\} \left\{ h(x_i, y) - \frac{\partial b_{H_t + ah}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right\}, \tag{58}
\]

where \( \bar{p}(y|x) \) is an empirical version of the conditional distribution \( p(y|x) \). The quantity (58) is a functional of the hypothesis \( h(x, y) \) and the optimization of \( h \) does not depend on \( \alpha \). We define \( \mathcal{E}_t(h) \) by

\[
\mathcal{E}_t(h) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in Y} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i)) - \bar{p}(y|x) \right\} \left\{ h(x_i, y) - \frac{\partial b_{H_t + ah}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right\}, \tag{59}
\]

and select a new hypothesis \( h_{t+1}(x, y) \) by minimizing \( \mathcal{E}_t(h) \) as

\[ h_{t+1} = \arg\min_{h \in \mathcal{H}} \mathcal{E}_t(h). \]

For the selected hypothesis \( h_{t+1}(x, y) \), we optimize the coefficient \( \alpha_{t+1} \) by

\[ \alpha_{t+1} = \arg\min_{\alpha} \bar{L}_U(H_t + \alpha h_{t+1}). \]

Then we obtain an updated discriminant function \( H_{t+1} = H_t + \alpha_{t+1} h_{t+1} \). A general algorithm of Boosting is summarized as follows.

1. Setting \( H_0(x, y) = 0 \).
2. for \( t = 1, \cdots, T \)
   (a) Find \( h_t = \arg\min_{h \in \mathcal{H}} \mathcal{E}_{t-1}(h) \).
Table 2: Training and test error rate for datasets in UCI repository and for contaminated datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training Error</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AdaBoost.M2</td>
<td>AdaBoost.M</td>
</tr>
<tr>
<td>Image</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.09290</td>
<td>0.07734</td>
</tr>
<tr>
<td>Pendigit</td>
<td>0.01468</td>
<td>0.01161</td>
</tr>
<tr>
<td>Waveform</td>
<td>0.13033</td>
<td>0.13333</td>
</tr>
<tr>
<td>Waveform(10%)</td>
<td>0.21300</td>
<td>0.19933</td>
</tr>
<tr>
<td>Waveform(20%)</td>
<td>0.29133</td>
<td>0.28967</td>
</tr>
<tr>
<td>Thyroid</td>
<td>0.00080</td>
<td>0.00000</td>
</tr>
<tr>
<td>Thyroid(10%)</td>
<td>0.11612</td>
<td>0.11532</td>
</tr>
<tr>
<td>Thyroid(20%)</td>
<td>0.22004</td>
<td>0.21633</td>
</tr>
<tr>
<td></td>
<td>Normalized Eta-Boost.M</td>
<td>Normalized Eta-Boost.M</td>
</tr>
<tr>
<td>Image</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.04397</td>
<td>0.18625</td>
</tr>
<tr>
<td>Pendigit</td>
<td>0.00961</td>
<td>0.21390</td>
</tr>
<tr>
<td>Waveform</td>
<td>0.11267</td>
<td>0.15367</td>
</tr>
<tr>
<td>Waveform(10%)</td>
<td>0.21967</td>
<td>0.22333</td>
</tr>
<tr>
<td>Waveform(20%)</td>
<td>0.28700</td>
<td>0.30200</td>
</tr>
<tr>
<td>Thyroid</td>
<td>0.00053</td>
<td>0.02174</td>
</tr>
<tr>
<td>Thyroid(10%)</td>
<td>0.11320</td>
<td>0.12540</td>
</tr>
<tr>
<td>Thyroid(20%)</td>
<td>0.21739</td>
<td>0.22243</td>
</tr>
<tr>
<td></td>
<td>AdaBoost.M1W</td>
<td>AdaBoost.M1W</td>
</tr>
<tr>
<td>Image</td>
<td>0.10476</td>
<td>0.12810</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.12950</td>
<td>0.19250</td>
</tr>
<tr>
<td>Pendigit</td>
<td>0.05746</td>
<td>0.24271</td>
</tr>
<tr>
<td>Waveform</td>
<td>0.14650</td>
<td>0.15100</td>
</tr>
<tr>
<td>Waveform(10%)</td>
<td>0.16750</td>
<td>0.16450</td>
</tr>
<tr>
<td>Waveform(20%)</td>
<td>0.15750</td>
<td>0.17800</td>
</tr>
<tr>
<td>Thyroid</td>
<td>0.00642</td>
<td>0.03034</td>
</tr>
<tr>
<td>Thyroid(10%)</td>
<td>0.02100</td>
<td>0.03792</td>
</tr>
<tr>
<td>Thyroid(20%)</td>
<td>0.02392</td>
<td>0.03151</td>
</tr>
</tbody>
</table>

(b) Calculate $\alpha_t = \arg\min_{\alpha} L_U(H_{t-1} + \alpha h_t)$.

(c) Update the discriminant function as $H_t = H_{t-1} + \alpha h_t$.

3. Output the discriminant function $H_T = \sum_{t=1}^{T} \alpha_t h_t(x, y)$.

A label for a new feature $x$ is predicted by $\arg\max_y H_T(x, y)$.

**Appendix B: Error rate property of the algorithm under the constant volume condition**

Here we discuss the criterion (59) for selecting a new weak hypothesis based on the given dataset and the associated empirical distributions $\tilde{q}(x)$ and $\tilde{p}(y|x)$. Under the normalized condition (23) and the moment matching condition (25), it is pointed out that (59) in the general algorithm can be interpreted as a weighted error rate [10]. Also, under the constant volume condition (18), we can rewrite step 2.(a) of the general algorithm as an optimization of a weighted
error rate. By differentiating (18) with respect to \( \alpha \), we obtain

\[
0 = \frac{\partial}{\partial \alpha} \sum_{y \in \mathcal{Y}} U(H(x, y) + \alpha h(x, y) - b_{H+\alpha h}(x)) \bigg|_{\alpha = \beta}
\]

\[
= \sum_{y \in \mathcal{Y}} u(H(x, y) + \beta h(x, y) - b_{H+\beta h}(x)) \left( h(x, y) - \frac{\partial b_{H+\beta h}(x)}{\partial \alpha} \bigg|_{\alpha = \beta} \right),
\]

and then

\[
\frac{\partial b_{H+\alpha h}(x)}{\partial \alpha} \bigg|_{\alpha = \beta} = \frac{\sum_{y \in \mathcal{Y}} u(H(x, y') + \beta h(x, y') - b_{H+\beta h}(x)) h(x, y')}{\sum_{y' \in \mathcal{Y}} u(H(x, y') + \beta h(x, y') - b_{H+\beta h}(x)).}
\]

By substituting (61) with \( \beta = 0 \) into (59), we observe that

\[
\mathcal{E}_t(h) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \tilde{p}(y|x_i) \left( -h(x_i, y) + \frac{\sum_{y' \in \mathcal{Y}} u(H_t(x_i, y') - b_{H_t}(x_i)) h(x_i, y')}{\sum_{y' \in \mathcal{Y}} u(H_t(x_i, y') - b_{H_t}(x_i))} \right)
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\sum_{y' \in \mathcal{Y}} u(H_t(x_i, y') - b_{H_t}(x_i)) h(x_i, y')}{\sum_{y' \in \mathcal{Y}} u(H_t(x_i, y') - b_{H_t}(x_i))} - h(x_i, y_i) \right)
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} (h(x_i, y) - h(x_i, y_i)) \frac{u(H_t(x_i, y) - b_{H_t}(x_i))}{\sum_{y' \in \mathcal{Y}} u(H_t(x_i, y') - b_{H_t}(x_i))}
\]

Let us define the weight for a pair comprising the \( i \)-th example and the index \( y \) at step \( t \) by

\[
D_t(i, y) = \frac{1}{N} \frac{u(H_{t-1}(x_i, y) - b_{H_{t-1}}(x_i))}{\sum_{y' \in \mathcal{Y}} u(H_{t-1}(x_i, y') - b_{H_{t-1}}(x_i))},
\]

and the weighted error rate of the hypothesis \( h \) by

\[
\varepsilon_t(h) = \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \frac{h(x_i, y) - h(x_i, y_i) + 1}{2} D_t(i, y).
\]

We discuss the factor of the right-hand side of the equation (64):

\[
\frac{h(x_i, y) - h(x_i, y_i) + 1}{2} = \begin{cases} 
0 & \text{if } y_i \in f(x_i) \text{ and } y \notin f(x_i), \\
\frac{1}{2} & \text{if } y_i \in f(x_i) \text{ and } y \in f(x_i), \\
\frac{3}{2} & \text{if } y_i \not\in f(x_i) \text{ and } y \not\in f(x_i), \\
1 & \text{if } y_i \not\in f(x_i) \text{ and } y \in f(x_i),
\end{cases}
\]

where \( f(x) \) is a classification rule associated with the weak hypothesis \( h(x, y) \). If the classification machine \( f(x) \) cannot distinguish the label \( y_i \) from another label \( y \), it takes 1/2. On the other hand, it takes 0 when \( x_i \) is correctly classified.
as \( y_i \) and takes 1 when \( x_i \) is incorrectly classified in a situation where \( f(x_i) \) can distinguish the label \( y_i \) from the other label \( y \). We thus obtain the following relationship:

\[
E_t(h) = 2\varepsilon_t(h) - 1.
\]

Consequently, step 2.(a) in the general algorithm is equivalent to a minimization of the weighted error rate (64).

In [10], the weighted error rate property \( \varepsilon_{t+1}(h_t) = 1/2 \) was observed under both the normalized condition and the moment matching condition. We obtain the same property for the weighted error rate (64) as the following lemma.

**Lemma 4.** Under the constant volume condition (18), a relationship

\[
\varepsilon_{t+1}(h_t) = \frac{1}{2},
\]

holds.

**Proof.** The result is immediately shown by the equilibrium condition of the coefficient \( \alpha_t \). Assuming that the discriminant function \( H_{t-1}(x, y) \) is obtained and the algorithm selects the new hypothesis \( h_t(x, y) \), then \( \alpha_t \) satisfies

\[
0 = \frac{\partial}{\partial \alpha} L_U^t(H_{t-1} + \alpha h_t) \bigg|_{\alpha = \alpha_t}
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \left\{ -h_t(x_i, y_i) + \frac{\partial h_{t-1} + \alpha h_t(x_i)}{\partial \alpha} \bigg|_{\alpha = \alpha_t} \right\}
\]

\[
= \sum_{i=1}^{N} \sum_{y \in Y} (h_t(x_i, y_i) - h_t(x_i, y)) D_{t+1}(i, y)
\]

\[
= 2\varepsilon_{t+1}(h_t) - 1.
\]

Thus we conclude (66).

This theorem shows that the selected \( t \)-th hypothesis \( h_t \) can be regarded as the weakest hypothesis or as a random guess under the updated weight \( D_{t+1}(i, y) \) of the next \( (t + 1) \)-th step.

**References**


31