Entropy **2015**, *17*, 5673-5694; doi:10.3390/e17085673

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Binary Classification with a Pseudo Exponential Model and Its Application for Multi-Task Learning †

Takashi Takenouchi [1,*], Osamu Komori [2] and Shinto Eguchi [2]

[1] Future University Hakodate, 116-2 Kamedanakano, Hakodate Hokkaido 041-8655, Japan
[2] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; E-Mails: komori@ism.ac.jp (O.K.); eguchi@ism.ac.jp (S.E.)

† This paper is an extended version of our paper published in Proceedings of the MaxEnt 2014 Conference on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Amboise, France, 21–26 September 2014.

* Author to whom correspondence should be addressed; E-Mail: ttakashi@fun.ac.jp; Tel.: +81-138-34-6415.

Academic Editors: Frédéric Barbaresco and Ali Mohammad-Djafari

Received: 11 May 2015 / Accepted: 3 August 2015 / Published: 6 August 2015

Abstract: In this paper, we investigate the basic properties of binary classification with a pseudo model based on the Itakura–Saito distance and reveal that the Itakura–Saito distance is a unique appropriate measure for estimation with the pseudo model in the framework of general Bregman divergence. Furthermore, we propose a novel multi-task learning algorithm based on the pseudo model in the framework of the ensemble learning method. We focus on a specific setting of the multi-task learning for binary classification problems. The set of features is assumed to be common among all tasks, which are our targets of performance improvement. We consider a situation where the shared structures among the dataset are represented by divergence between underlying distributions associated with multiple tasks. We discuss statistical properties of the proposed method and investigate the validity of the proposed method with numerical experiments.

Keywords: multi-task learning; Itakura–Saito distance; pseudo model; un-normalized model

1. Introduction

In the framework of multi-task learning problems, we assume that there are multiple related tasks (datasets) sharing a common structure and can utilize the shared structure to improve the generalization performance of classifiers for multiple tasks [1,2]. This framework has been successfully employed in various kind of applications, such as medical diagnosis. Most methods utilize the similarity among tasks to improve the performance of classifiers by representing the shared structure as a regularization term [3,4]. We tackle this problem using a boosting method, which makes it possible to adaptively learn complicated problems with low computational cost. The boosting methods are notable implementations of the ensemble learning and try to construct a better classifier by combining weak classifiers. AdaBoost is the most popular boosting method, and many variations, including TrAdaBoost for the multi-task learning [5], have been developed. In face recognition [6], as well as web search ranking [7], the computational efficiency of boosting is paid attention to in the framework of multi-task learning.

In this paper, we firstly reveal that AdaBoost can be derived by a sequential minimization of the Itakura–Saito (IS) distance between an empirical distribution and a pseudo measure model associated with a classifier. The IS distance is a special case of the Bregman divergence [8] between two positive measures and is frequently used for non-negative matrix factorization (NMF) in the region of signal processing [9,10]. Secondly, we propose a novel boosting algorithm for the multi-task learning based on the IS distance. We utilize the IS distance as a discrepancy measure between pseudo models associated with tasks and incorporate the IS distance as a regularizer into AdaBoost. The proposed method can capture the shared structure, *i.e.*, the relationship between underlying distributions by considering the IS distance between pseudo models constructed by classifiers. We discuss the statistical properties of the proposed method and investigate the validity of the regularization by the IS distance with small experiments using synthetic datasets and a real dataset.

This paper is organized as follows. In Section 2, basic settings are described, and a divergence measure is introduced. In Section 3, we briefly introduce the IS distance, which is a special case of the Bregman divergence, and investigate the relationship between a well-known ensemble algorithm, AdaBoost and estimation with a pseudo model using the Itakura–Saito distance. In Section 4, we propose a method for multi-task learning, which is derived from a minimization of the weighted sum of divergence, and the performance of the proposed methods is examined in Section 5 using a synthetic dataset and a real dataset (a short version of this article has been presented as a conference paper [11]; some theoretical results and numerical experiments are added to the current version).

2. Settings

In this study, we focus on binary classification problems. Let $\boldsymbol{x}$ be an input and $y \in \mathcal{Y} = \{\pm 1\}$ be a class label. Let us assume that J datasets $\mathcal{D}_j = \{\boldsymbol{x}_i^{(j)}, y_i^{(j)}\}_{i=1}^{n_j}$ $(j = 1, \ldots, J)$ are given, and let $p_j(y|\boldsymbol{x})r_j(\boldsymbol{x})$ and $\tilde{p}_j(y|\boldsymbol{x})\tilde{r}_j(\boldsymbol{x})$ be an underlying distribution and an empirical distribution associated with the dataset $\mathcal{D}_j$, respectively. Here, we assume that each conditional distribution of y given $\boldsymbol{x}$ is written as:

$$p_k(y|\boldsymbol{x}) = p_0(y|\boldsymbol{x}) + \delta_k(\boldsymbol{x})y \tag{1}$$

where $p_0(y|\boldsymbol{x})$ is a common conditional distribution for all datasets and $\delta_k(x)$ is a term that is specific to the dataset $\mathcal{D}_k$. Note that $\sum_{y\in\mathcal{Y}} \delta_k(\boldsymbol{x})y = 0$ holds, because $p_k(y|\boldsymbol{x})$ is a probability distribution. While a discriminant function F_k is usually constructed using only the dataset $\mathcal{D}_k$, the multi-task learning aims to improve the performance of the discriminant function for each dataset $\mathcal{D}_k$ with the help of datasets $\mathcal{D}_j$ ($j \neq k$). For this purpose, we consider a risk minimization problem defined with a pseudo model and the Itakura–Saito (IS) distance, which is a discrepancy measure frequently used in a region of signal processing.

Let $\mathcal{M} = \left\{ m(y) \;\middle|\; 0 \leq \sum_{y\in\mathcal{Y}} m(y) < \infty \right\}$ be a space of all positive finite measures over $\mathcal{Y}$. The Itakura–Saito distance between $p, q \in \mathcal{M}$ is defined as:

$$\mathrm{IS}(p, q; r) = \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{ \log \frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})} - 1 + \frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})} \right\} d\boldsymbol{x} \tag{2}$$

where $r(\boldsymbol{x})$ is a marginal distribution of $\boldsymbol{x}$ shared by p, $q \in \mathcal{M}$. Note that the IS distance is a kind of statistical version of the Bregman divergence [12], which makes it possible to directly plug-in the empirical distribution. We observe that $\mathrm{IS}(p, q; r) \geq 0$ and $\mathrm{IS}(p, q; r) = 0$ if and only if $p = q$. Banerjee *et al.* [13] showed that there exists a unique Bregman divergence corresponding to every regular exponential family, and the Itakura–Saito distance is associated with the exponential distribution.

3. Itakura–Saito Distance and Pseudo Model

3.1. Parameter Estimation with the Pseudo Model

Let $q_F(y|\boldsymbol{x})$ be an (un-normalized) pseudo model associated with a function $F(\boldsymbol{x})$,

$$q_F(y|\boldsymbol{x}) = \exp(F(\boldsymbol{x})y). \tag{3}$$

Note that $q_F(y|\boldsymbol{x})$ is not a probability function, *i.e.*, $\sum_{y\in\mathcal{Y}} q_F(y|\boldsymbol{x}) \neq 1$ in general. If $q_F(y|\boldsymbol{x})$ is normalized, the model reduces to the classical logistic model as:

$$\bar{q}_F(y|\boldsymbol{x}) = \frac{\exp(F(\boldsymbol{x})y)}{\exp(F(\boldsymbol{x})) + \exp(-F(\boldsymbol{x}))}. \tag{4}$$

When the function F is parameterized by $\boldsymbol{\theta}$, the maximum likelihood estimation (MLE) $\mathrm{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \bar{q}_F(y_i|\boldsymbol{x}_i)$ or equivalently minimization of the (extended) Kullback–Leibler (KL) divergence is a powerful tool for the estimation of $\boldsymbol{\theta}$, and the MLE has properties such as asymptotic consistency and efficiency under some regularity conditions. Here, we consider parameter estimation with the pseudo model Equation (3) rather than the normalized model Equation (4).

Proposition 1. *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, we observe:*

$$\operatorname*{argmin}_F \mathrm{IS}(p, q_F; r) = F_0, \tag{5}$$

$$\operatorname*{argmin}_F \mathrm{IS}(q_F, p; r) = F_0. \tag{6}$$

Proof. See Appendix A □

On the other hand, when we consider an estimation based on the extended KL divergence, *i.e.*, $\mathrm{argmin}_F \mathrm{KL}(p, q_F; r)$ where:

$$\mathrm{KL}(p, q; r) = \int r(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \{p(y|\boldsymbol{x}) \log \frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})} - p(y|\boldsymbol{x}) + q(y|\boldsymbol{x})\} d\boldsymbol{x}, \tag{7}$$

we observe the following.

Proposition 2. *Let F_0 be a function $F_0 (\neq 0)$ and $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, we observe:*

$$F_{\mathrm{KL},1} = \operatorname*{argmin}_F \mathrm{KL}(p, q_F; r) \neq F_0, \tag{8}$$

$$F_{\mathrm{KL},2} = \operatorname*{argmin}_F \mathrm{KL}(q_F, p; r) \neq F_0. \tag{9}$$

Proof. See Appendix B. □

Remark 1. *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, minimizer Equation (8) or (9) of the extended KL divergence attains the Bayes rule, i.e.,*

$$\mathrm{sgn}(F_{\mathrm{KL},1}(\boldsymbol{x})) = \mathrm{sgn}(F_{\mathrm{KL},2}(\boldsymbol{x})) = \mathrm{sgn}\left(\frac{1}{2} \log \frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}\right). \tag{10}$$

The proposition and the remark show that the extended KL divergence is not completely appropriate for estimation with the pseudo model.

3.2. Characterization of the Itakura–Saito Distance

In this section, we investigate the characterization of the Itakura–Saito distance for estimation with the pseudo model, in the framework of the Bregman U-divergence. Firstly, we briefly introduce the statistical version of Bregman U-divergence [12]. The statistical version of Bregman U-divergence is a discrepancy measure between positive measures in $\mathcal{M}$ defined by a generating function U and enables us to directly plug-in the empirical distribution for estimation. [12] proposed a general boosting-type algorithm for classification using the Bregman U-divergence and discussed properties of the method from the viewpoint of information geometry [14]. By changing the generating function U, the Bregman U-divergence can have a useful property as robustness against noise. For example, the β-divergence is a special case of the Bregman U-divergence and is frequently used for robust estimation in the context of unsupervised learning, such as clustering or component analysis [15,16]. Another example of the Bregman U-divergence is the η-divergence, which is employed to robustify the classification algorithm and is closely related to probability models of mislabeling [17,18].

Let U be a monotonically-increasing convex function and ξ be an inverse function of U', the derivative of U. From the convexity of the function U, the function ξ is a monotonically-increasing function. The statistical version of Bregman U-divergence between two measures $p, q \in \mathcal{M}$ is defined as follows.

$$D_U(p, q; r) = \int r(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \{U(\xi(q(y|\boldsymbol{x}))) - U(\xi(p(y|\boldsymbol{x}))) - p(y|\boldsymbol{x}) (\xi(q(y|\boldsymbol{x})) - \xi(p(y|\boldsymbol{x})))\} d\boldsymbol{x}. \tag{11}$$

Note that the function ξ should be defined at least on $z > 0$.

Remark 2. *The KL divergence and the Itakura–Saito distance are special cases of the Bregman U-divergence Equation (11) with generating functions $U(z) = \exp(z)$ and $U(z) = -\log(c - z) + c_1$ $(z < c)$, where c and c_1 are constants, respectively.*

Here, we introduce the concept of reflection-symmetric for characterization of the IS distance.

Definition 3. *A function $f(z)$ is reflection-symmetric if:*

$$f(z) = f\left(z^{-1}\right) \tag{12}$$

holds for all $z \neq 0$.

If the function f is reflection-symmetric, we observe that:

$$\lim_{z\to 0} f(z) = \lim_{z\to\infty} f(z). \tag{13}$$

Because of this property, the reflection-symmetric function often has a singular point at $z = 0$, and to investigate the behavior of the function, we can employ the Laurent series as:

$$f(z) = c + \sum_{k=1}^{\infty} \left(a_k z^k + b_k z^{-k}\right). \tag{14}$$

Note that if the function f is holomorphic over R, $b_k = 0$ for all k, and the Laurent series is equivalent to the Taylor series.

Remark 3. *If the function f is reflection-symmetric and holomorphic over R, $a_k = b_k = 0$ holds for all k, and then, f is a constant function.*

For the Bregman U-divergence Equation (11), we observe the following Lemma.

Lemma 4. *Let F_0 be an arbitrary function, $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution and $q_F(\boldsymbol{x})$ be the pseudo model Equation (3). If the Bregman U-divergence associated with the function U attains:*

$$F_0 = \operatorname*{argmin}_F D_U(p, q_F; r), \tag{15}$$

a function $\xi'(z)z^2$ derived from U is reflection-symmetric. In addition, if the Bregman U-divergence associated with the function U attains:

$$F_0 = \operatorname*{argmin}_F D_U(q_F, p; r), \tag{16}$$

a function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ derived from U is reflection-symmetric.

Proof. See Appendix C. □

Remark 4. *Proposition 1 implies that the function ξ associated with the IS distance satisfies Lemma 4.*

Remark 5. *Propositions imply that the function U, i.e., Bregman U-divergence, attains Equation (15) or (16) is not unique and there exists divergences satisfying Equation (15) or (16), other than the Itakura–Saito distance. For example, a function:*

$$\xi(z) = -2z^{-\frac{2}{3}} - z^{-\frac{4}{3}} \tag{17}$$

satisfies $\xi'(z)z^2 = \frac{4}{3}(z^{1/3} + z^{-1/3})$, and then, $\xi'(z)z^2$ is reflection-symmetric. The associated generating function U is written as:

$$U(z) = \int^z \xi^{-1}(z')dz' = -4\frac{-2 + \sqrt{1-z}}{\sqrt{-1 + \sqrt{1-z}}} + C_1 \tag{18}$$

where C_1 is a constant.

In the following theorem, we reveal the characterization of the Itakura–Saito distance for estimation with the pseudo model Equation (3) and the Bregman U-divergence.

Theorem 5. *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution and $q_F(\boldsymbol{x})$ be the pseudo model Equation (3). If conditions:*

$$F_0 = \operatorname*{argmin}_F D_U(p, q_F; r), \tag{19}$$

$$F_0 = \operatorname*{argmin}_F D_U(q_F, p; r) \tag{20}$$

simultaneously hold, then $U(z) = -\log(-z)$, i.e., $D_U(p, q; r)$ is the Itakura–Saito distance $\mathrm{IS}(p, q; r)$.

Proof. See Appendix D. □

Remark 6. *If we assume that a function $\xi'(z)z^2$ derived from U is reflection-symmetric and holomorphic over R, $\xi'(z)z^2$ is a constant function from Remark 3. Then, we obtain $\xi(z) = c + \frac{b_1}{z}$ where c, b_1 are constants, implying that the associated divergence is equivalent to the Itakura–Saito distance.*

3.3. Relationship with AdaBoost

The IS distance between the underlying conditional distribution $p(y|\boldsymbol{x})$ and the pseudo model $q_F(y|\boldsymbol{x})$ is written as:

$$\begin{aligned}\mathrm{IS}(p, q_F; r) &= C + \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{ F(\boldsymbol{x})y + \frac{p(y|\boldsymbol{x})}{q_F(y|\boldsymbol{x})} \right\} d\boldsymbol{x} \\ &= C + \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} p(y|\boldsymbol{x}) e^{-F(\boldsymbol{x})y} d\boldsymbol{x},\end{aligned} \tag{21}$$

where C is a constant, and Equation (21) is equivalent to an expected loss of AdaBoost, except for the constant term. Then, sequential minimization of an empirical version of Equation (21) is equivalent to the algorithm of AdaBoost, which is the most popular boosting method for the binary classification. Furthermore, [12,19] discussed that a gradient-based boosting algorithm can be derived from the minimization of the KL divergence or the Bregman U-divergence between the underlying

distribution and a pseudo model. An important difference between these frameworks and our framework Equation (21) is the employed pseudo model. The pseudo model employed by the previous frameworks assumes a condition called "consistent data assumption" and is defined with the empirical distribution, implying that the pseudo model varies depending on the dataset. On the other hand, the pseudo model Equation (3) employed in Equation (21) is fixed against the dataset as usual statistical models.

The IS distance between two pseudo models $q_F(y|\boldsymbol{x})$ and $q_{F'}(y|\boldsymbol{x})$ is written as,

$$\begin{aligned}\mathrm{IS}(q_F, q_{F'}; r) &= \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \{F'(\boldsymbol{x})y - F(\boldsymbol{x})y - 1 + \exp(F(\boldsymbol{x})y - F'(\boldsymbol{x})y)\}\, d\boldsymbol{x} \\ &= 2 + \int r(\boldsymbol{x}) \{\exp(F(\boldsymbol{x}) - F'(\boldsymbol{x})) + \exp(F'(\boldsymbol{x}) - F(\boldsymbol{x}))\}\, d\boldsymbol{x}. \end{aligned} \tag{22}$$

Note that $\mathrm{IS}(q_{F'}, q_F; r) = \mathrm{IS}(q_F, q_{F'}; r)$ holds for arbitrary q_F and $q_{F'}$, while the IS distance itself is not necessarily symmetric. Furthermore, note that the symmetric property does not hold for normalized models $\bar{q}_F$ and $\bar{q}_{F'}$.

4. Application for Multi-Task Learning

There are two main types of frameworks for multi-task learning [20,21].

Case 1 : There is a target dataset $\mathcal{D}_k$, and our interest is to construct a discriminant function F_k utilizing remaining datasets $\mathcal{D}_j$ $(j \neq k)$ or *a priori* constructed discriminant functions F_j $(j \neq k)$.

Case 2 : Our interest is to simultaneously construct better discriminant functions $F_1, \ldots, F_J$ using all J datasets $\mathcal{D}_1, \ldots, \mathcal{D}_J$ by utilizing shared information among datasets.

4.1. Case 1

In this section, we focus on the above first framework. Let us assume that discriminant functions $F_j(\boldsymbol{x})$ $(j \neq k)$ are given or are constructed by an arbitrary binary classification method. Then, let us consider a risk function:

$$\begin{aligned} L_k(F_k) &= \mathrm{IS}(p_k, q_{F_k}; r_k) + \sum_{j\neq k} \lambda_{k,j}\, \mathrm{IS}(q_{F_k}, q_{F_j}; r_k) \\ &= \int r_k(\boldsymbol{x}) \left\{ \sum_{y\in\mathcal{Y}} p_k(y|\boldsymbol{x}) e^{-F_k(\boldsymbol{x})y} + \sum_{j\neq k} \lambda_{k,j} \left\{ e^{F_k(\boldsymbol{x}) - F_j(\boldsymbol{x})} + e^{F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})} \right\} \right\} d\boldsymbol{x}, \end{aligned} \tag{23}$$

where $\lambda_{k,j} \geq 0$ $(j \neq k)$ are regularization constants. Note that the risk function depends on functions F_j $(j \neq k)$, and the second term becomes small when the target discriminant function F_k is similar to functions $F_j (j \neq k)$ in the sense of the IS distance; and the second term corresponds to a regularizer incorporating the shared information among datasets into the target function F_k. Furthermore, note that the marginal distribution r_k is shared in the second term for the ease of implementation and the simplicity of theoretical analysis.

An empirical version of Equation (23) is written as:

$$\bar{L}_k(F_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(e^{-F_k(\boldsymbol{x}_i^{(k)}) y_i^{(k)}} + \sum_{j\neq k} \lambda_{k,j} \left(e^{F_k(\boldsymbol{x}_i^{(k)}) - F_j(\boldsymbol{x}_i^{(k)})} + e^{F_j(\boldsymbol{x}_i^{(k)}) - F_k(\boldsymbol{x}_i^{(k)})} \right) \right). \tag{24}$$

An algorithm is derived by sequential minimization of Equation (24) by updating F_k to $F_k + \alpha f$, *i.e.*, $(\alpha, f) = \mathrm{argmin}_{\alpha,f} \bar{L}_k(F_k + \alpha f)$, where f is a weak classifier and α is a coefficient [22].

(1) Initialize the function to F_k^0, and define weights for the i-th example with a function F as:

$$w_1(i; F) = \frac{e^{-F(\boldsymbol{x}_i^{(k)})y_i^{(k)}}}{Z_1(F)},$$
$$w_2(i; F) = \frac{\sum_{j \neq k} \lambda_{k,j} e^{f(\boldsymbol{x}_i^{(k)})(F(\boldsymbol{x}_i^{(k)}) - F_j(\boldsymbol{x}_i^{(k)}))}}{Z_2(F)}$$

where:

$$Z_1(F) = \sum_{i=1}^{n_k} e^{-F(\boldsymbol{x}_i^{(k)})y_i^{(k)}},$$
$$Z_2(F) = \sum_{i=1}^{n_k} \sum_{j \neq k} \lambda_{k,j} \left(e^{F(\boldsymbol{x}_i^{(k)}) - F_j(\boldsymbol{x}_i^{(k)})} + e^{-F(\boldsymbol{x}_i^{(k)}) + F_j(\boldsymbol{x}_i^{(k)})} \right).$$

(2) For $t = 1, \ldots, T$

(a) Select a weak classifier $f_k^t \in \{\pm 1\}$, which minimizes the following quantity:

$$\varepsilon(f) = \frac{Z_1(F_k^{t-1})}{Z_1(F_k^{t-1}) + Z_2(F_k^{t-1})} \varepsilon_1(f) + \frac{Z_2(F_k^{t-1})}{Z_1(F_k^{t-1}) + Z_2(F_k^{t-1})} \varepsilon_2(f). \tag{25}$$

where $\varepsilon_1(f) = \sum_{i=1}^{n_k} w_1(i; F_k^{t-1}) \, \mathrm{I}(f(\boldsymbol{x}_i^{(k)}) \neq y_i^{(k)})$ and $\varepsilon_2(f) = \sum_{i=1}^{n} w_2(i; F_k^{t-1})$.

(b) Calculate a coefficient of f_k^t by $\alpha_k^t = \frac{1}{2} \log \frac{1 - \varepsilon(f_k^t)}{\varepsilon(f_k^t)}$.

(c) Update the discriminant function as $F_k^t = F_k^{t-1} + \alpha_k^t f_k^t$.

(3) Output $F_k^T(\boldsymbol{x}) = F_k^0(\boldsymbol{x}) + \sum_{t=1}^{T} \alpha_k^t f_k^t(\boldsymbol{x})$.

In Step 1, F_k^0 is typically initialized as $F_k^0(\boldsymbol{x}) = 0$. The quantity Equation (25) is a mixture of two terms: $\varepsilon_1(f)$ is a weighted error rate of the classifier f, and $\varepsilon_2(f)$ is the sum of weights $w_2(f)$, which represents the degree of discrepancy between f and $F - F_j$. $\varepsilon_2(f)$ becomes large when F is updated by f as departed from F_j. Note that if we set $\lambda_{k,j} = 0$ for all j, the risk function Equation (24) coincides with that of AdaBoost, and the above derived algorithm reduces to the usual AdaBoost.

Because the empirical risk function Equation (24) is convex with respect to F or F', we can consider another version of the risk function as:

$$\bar{L}_k(F_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(e^{-F_k(\boldsymbol{x}_i^{(k)})y_i^{(k)}} + \lambda_k \left(e^{F_k(\boldsymbol{x}_i^{(k)}) - \bar{F}_k(\boldsymbol{x}_i^{(k)})} + e^{-F_k(\boldsymbol{x}_i^{(k)}) + \bar{F}_k(\boldsymbol{x}_i^{(k)})} \right) \right) \tag{26}$$

where $\bar{F}_k(\boldsymbol{x}) = \sum_{j \neq k} \frac{\lambda_{k,j}}{\lambda_k} F_j(\boldsymbol{x})$. The risk function is upper bounded by the risk function Equation (24), implying that the effect of regularization by the shared information is weakened. The derived algorithm is almost the same as the one derived from Equation (24).

4.2. Case 2

In this section, we consider simultaneous construction of discriminant functions $F_1, \ldots, F_J$ by minimizing the following risk function:

$$L(F_1, \ldots, F_J) = \sum_{j=1}^{J} \pi_j L_j(F_j) \tag{27}$$

where $\pi_j (j = 1, \ldots, J)$ is a positive constant satisfying $\sum_{j=1}^{J} \pi_j = 1$ and L_k is defined in Equation (23).

Though we can directly minimize the empirical version of Equation (27), a derived algorithm is complicated and is computationally heavy. Then, we derive a simplified algorithm utilizing the algorithm shown in Case 1 in which a target dataset is fixed.

(1) Initialize functions $F_1, \ldots, F_J$.
(2) For $t = 1, \ldots, T$:
   (a) Randomly choose a target index $k \in \{1, \ldots, J\}$.
   (b) Update the function F_k using the algorithm in Case 1 by S steps, with fixed functions F_j $(j \neq k)$.
(3) Output learned functions $F_1, \ldots, F_J$.

Note that the empirical risk function cannot be monotonically decreased because the minimization of $L_k(F_k)$ is a trade-off of the first term and the second regularization term, and a decrease of $L_k(F_k)$ does not necessarily mean a decrease of the regularization term.

4.3. Statistical Properties of the Proposed Methods

In this section, we discuss the statistical properties of the proposed methods. Firstly, we focus on Case 1, and the minimizer F_k^* of the risk function Equation (23) satisfies the following:

$$\left.\frac{\delta L_k(F_k)}{\delta F_k(\boldsymbol{x})}\right|_{F_k = F_k^*} \propto -p_k(+1|\boldsymbol{x}) e^{-F_k^*(\boldsymbol{x})} + p_k(-1|\boldsymbol{x}) e^{F_k^*(\boldsymbol{x})} + \sum_{j \neq k} \lambda_{k,j} \left\{ e^{F_k^*(\boldsymbol{x}) - F_j(\boldsymbol{x})} - e^{F_j(\boldsymbol{x}) - F_k^*(\boldsymbol{x})} \right\} = 0, \tag{28}$$

which implies:

$$F_k^*(\boldsymbol{x}) = \frac{1}{2} \log \frac{p_k(+1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{k,j} \exp(F_j(\boldsymbol{x}))}{p_k(-1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{k,j} \exp(-F_j(\boldsymbol{x}))}, \tag{29}$$

or equivalently:

$$p_k(y|\boldsymbol{x}) = p_{0,k}(y|\boldsymbol{x}) \left(1 + \sum_{j \neq k} \lambda_{k,j} \exp(-F_j(\boldsymbol{x}) y) \right) - p_{0,k}(-y|\boldsymbol{x}) \sum_{j \neq k} \lambda_{k,j} \exp(F_j(\boldsymbol{x}) y), \tag{30}$$

where $p_{0,k}(y|\boldsymbol{x}) = \frac{\exp(F_k^*(\boldsymbol{x}) y)}{\exp(F_k^*(\boldsymbol{x})) + \exp(-F_k^*(\boldsymbol{x}))}$. This can be interpreted as a probabilistic model of asymmetric mislabeling [17,18]. In Equation (29), the confidence of classification is discounted by the results of remaining discriminant functions when the classifier $\mathrm{sgn}(F_k^*(\boldsymbol{x}))$ makes a different decision from these of $\mathrm{sgn}(F_j(\boldsymbol{x}))$ $(j \neq k)$.

Remark 7. *$F_k^*(\boldsymbol{x}) \geq 0$ does not mean $p_k(+1|\boldsymbol{x}) \geq \frac{1}{2}$ unless $F_j(\boldsymbol{x}) = \frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})}$ holds.*

Proposition 6. *Let us assume that $F_j(\boldsymbol{x})$ satisfies:*

$$\frac{\exp(F_j(\boldsymbol{x})y)}{\exp(F_j(\boldsymbol{x})) + \exp(-F_j(\boldsymbol{x}))} = p_0(y|\boldsymbol{x}) + \epsilon_j(\boldsymbol{x})y,\ ||\epsilon_j(\boldsymbol{x})|| \ll 1. \tag{31}$$

Then, Equation (29) can be approximated as:

$$F_k^*(\boldsymbol{x}) \simeq \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})} + \frac{1}{2P^2}\frac{P\delta_k(\boldsymbol{x}) + \sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})}{P + \lambda_k} \tag{32}$$

where $P = \sqrt{p_0(+1|\boldsymbol{x})p_0(-1|\boldsymbol{x})}$ and $\lambda_k = \sum_{j\neq k}\lambda_{k,j}$.

Proof. We obtain Equation (32) by considering the Taylor expansion of Equation (29). □

We observe that a discrepancy derived by δ_k is moderated by the mixture of ϵ_j when perturbations ϵ_j are independently and identically distributed.

Proposition 7. *Let $\eta_j(\boldsymbol{x}) = F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})$ be a difference between two functions. Then, F_k^* can be approximated as:*

$$F_k^*(\boldsymbol{x}) \simeq \frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})} + \frac{1}{P}\sum_{j\neq k}\lambda_{k,j}\eta_j(\boldsymbol{x}). \tag{33}$$

Proof. See Appendix E. □

Proposition 8. *Let $\bar{F}_k^*$ be a minimizer of the risk function Equation (23) with $\lambda_{k,j} = 0 (j \neq k)$. Then, we observe:*

$$\left(\bar{F}_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2 \geq \left(F_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2, \tag{34}$$

i.e., the proposed method improves the performance in the sense of the squared error, when:

$$|\delta_k(\boldsymbol{x})| \geq \frac{|\sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})|}{\lambda_k} \tag{35}$$

holds.

Proof. See Appendix F. □

Secondly, we consider a property of the algorithm for Case 2.

Proposition 9. *Let $r(\boldsymbol{x}) = r_j(\boldsymbol{x})$ $(j = 1, \ldots, J)$ be a common marginal distribution shared by all tasks. Then, the minimizer of the risk function is written as:*

$$F_k(\boldsymbol{x}) = \frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x}) + \sum_{j\neq k}\lambda_{jk}e^{F_j(\boldsymbol{x})}}{p_k(-1|\boldsymbol{x}) + \sum_{j\neq k}\lambda_{jk}e^{-F_j(\boldsymbol{x})}}, \tag{36}$$

where $\lambda_{jk} = \lambda_{k,j} + \frac{\pi_j}{\pi_k}\lambda_{k,j}$.

Proof. See Appendix G. □

The only difference from Equation (28) is that regularization is strengthened by $\frac{\pi_j}{\pi_k}\lambda_{j,k}$, and then, the same propositions in Section 4.1 hold for Equation (36).

4.4. Comparison of Regularization Terms

The proposed method incorporates the regularization term defined by the IS distance into AdaBoost. In this section, we discuss a property of the regularization term.

Proposition 10. *Let $\epsilon(\boldsymbol{x})$ be a perturbation function satisfying $|\epsilon(\boldsymbol{x})| \ll 1$. Then, we observe:*

$$\mathrm{KL}(\bar{q}_F, \bar{q}_{F+\epsilon}; r) \simeq \int 2r(\boldsymbol{x})\epsilon(\boldsymbol{x})^2 \bar{q}_F(+1|\boldsymbol{x})\bar{q}_F(-1|\boldsymbol{x})d\boldsymbol{x}, \tag{37}$$

$$\mathrm{KL}(q_F, q_{F+\epsilon}; r) \simeq \int \frac{r(\boldsymbol{x})}{2}\epsilon(\boldsymbol{x})^2 \frac{1}{\sqrt{\bar{q}_F(+1|\boldsymbol{x})\bar{q}_F(-1|\boldsymbol{x})}}d\boldsymbol{x}, \tag{38}$$

$$\mathrm{IS}(\bar{q}_F, \bar{q}_{F+\epsilon}; r) \simeq \int 2r(\boldsymbol{x})\epsilon(\boldsymbol{x})^2 \sum_{y\in\mathcal{Y}} \bar{q}_F(y|\boldsymbol{x})^2 d\boldsymbol{x}, \tag{39}$$

$$\mathrm{IS}(q_F, q_{F+\epsilon}; r) \simeq \int r(\boldsymbol{x})\epsilon(\boldsymbol{x})^2 d\boldsymbol{x}. \tag{40}$$

Proof. We obtain these approximations by considering the Taylor expansion up to second order. □

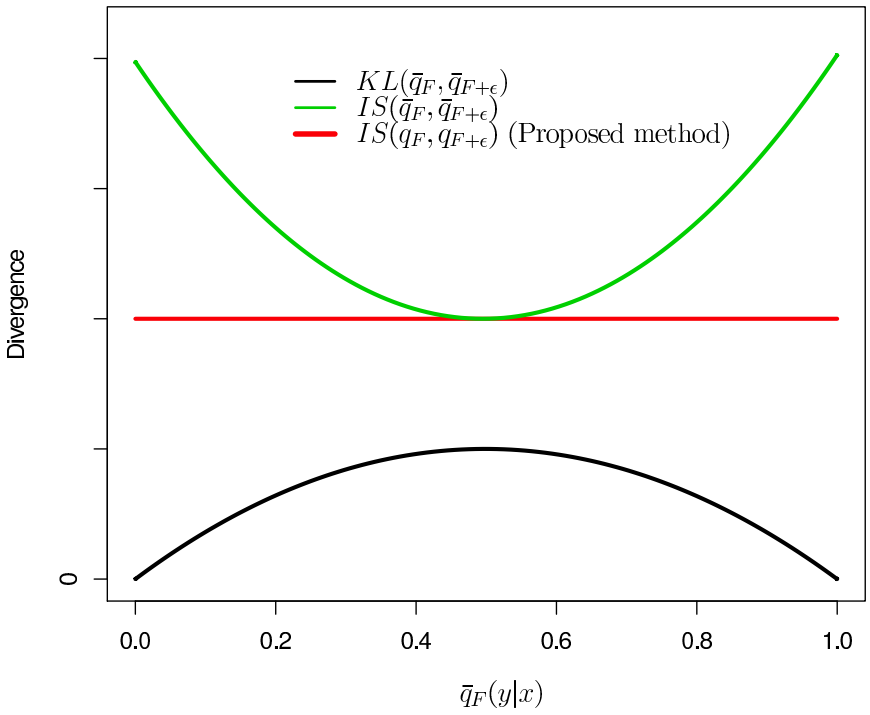

Figure 1. Values of divergences (regularization terms) against $\bar{q}_F$.

Figure 1 shows values of divergences against a value of $\bar{q}_F(\boldsymbol{x})$. Those relations implies that the KL divergence Equation (37) emphasizes a region of input $\boldsymbol{x}$ whose conditional distribution $\bar{q}_F(\boldsymbol{x})$ is nearly equal to $\frac{1}{2}$, *i.e.*, the classification boundary, while the IS distance Equation (39) focuses on a region of

$\boldsymbol{x}$ whose conditional distribution is nearly equal to zero or one. The IS distance between pseudo model Equation (40), *i.e.*, the proposed method, considers the intermediate of Equations (37) and (39). This implies that the regularization Equation (40) with the IS distance puts more focus on a region far from the classification boundary compared to Equation (37), while Equation (39) tends to relatively ignore the region near the classification boundary. Furthermore, note that the employment of Equation (40) makes it possible to derive the simple algorithm shown in Section 4.1.

5. Experiments

In this section, we investigate the performance of the proposed multi-task algorithm with synthetic datasets and a real dataset.

5.1. Synthetic Dataset

Firstly, we investigate the performance of the proposed method using two synthetic datasets within the situation described in Case 2. We compared the proposed method with AdaBoost trained with an individual dataset and AdaBoost trained with all datasets simultaneously. We employed the boosting stump (the boosting stump is a decision tree with only one node) as the weak classifier and fixed as $\pi_j = 1/J$. A boosting-type method has a hyper-parameter T, the step number of boosting, and the proposed method additionally has the hyper-parameter $\lambda_{k,j}$. In the experiment, we determined these parameters T and $\lambda_{k,j}$ by the validation technique. Especially, we investigated two kinds of scenarios for the determination of $\lambda_{k,j}$.

1. We set that $\lambda_{k,j} = \lambda$ for all j, k and determined λ.
2. We set that $\lambda_{k,j} = \frac{\lambda}{\mathrm{IS}\left(q_{\hat{F}_k}, q_{\hat{F}_j}; r_k\right)}$ where $\hat{F}_j$ is a discriminant function constructed by AdaBoost with the dataset $\mathcal{D}_j$ and determined λ.

Scenario 2 can incorporate more detailed information about the relationship between tasks, and the proposed method can ignore the information of tasks having less shared information. In summary, we compared the following four methods:

A : The proposed method with $\lambda_{k,j}$ determined by Scenario 1.
B : The proposed method with $\lambda_{k,j}$ determined by Scenario 2.
C : AdaBoost trained with an individual dataset.
D : AdaBoost trained with all datasets simultaneously.

We utilized 80% of the training dataset for training of classifiers and the remaining 20% for the validation. We repeated the above procedure 20 times and observed the averaged performance of the methods.

5.1.1. Dataset 1

We set the number J of tasks to three and assume that a marginal distribution of $\boldsymbol{x}$ is a uniform distribution on $[-1, 1]^2$, and a discriminant function F_j $(j = 1, 2, 3)$ associated with each dataset is

generated by $F_j(\boldsymbol{x}) = (1 + c_{j,2})(x_1 - c_{j,1}) - x_2$, where $c_{j,1} \sim \mathcal{N}(0, 0.2^2)$ and $c_{j,2} \sim \mathcal{N}(0, 0.1^2)$. In addition, we randomly added a contamination noise on label y. Under these settings, we generated a training dataset, including 400 examples, and a test dataset, including 600 examples. Generated datasets are shown in Figure 2. We observe that each discriminant function and noise structure are different from the other two.

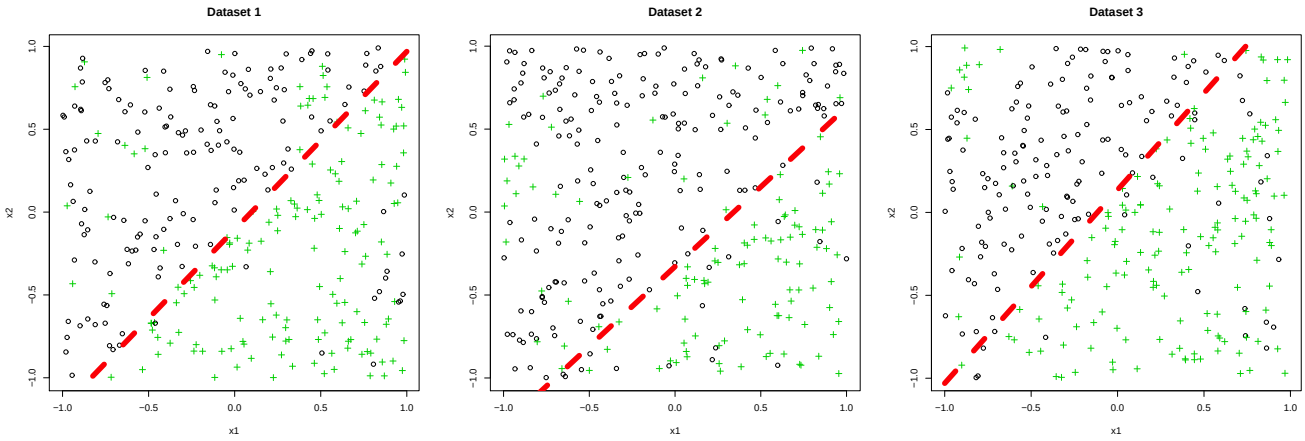

Figure 2. The three generated datasets and decision boundaries.

Figure 3 shows boxplots of the test errors of each method for datasets $\mathcal{D}_j$ ($j = 1, 2, 3$). We observe that the proposed method consistently outperforms individually trained AdaBoost, and AdaBoost trained with all datasets simultaneously. The figure shows that the proposed method can incorporate shared information among datasets into classifiers.

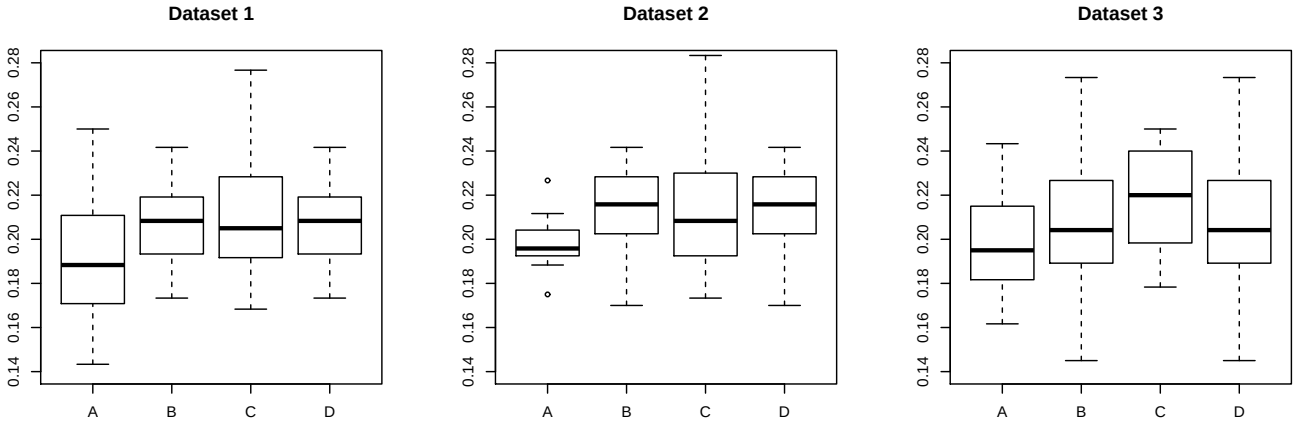

Figure 3. Boxplots of the test error of each method: A—proposed method with λ in Scenario 1; B—proposed method with λ in Scenario 2; C—AdaBoost trained with the individual dataset; D—AdaBoost trained with all datasets simultaneously; for three datasets, over the 20 simulation trials.

5.1.2. Dataset 2

We set the number J of tasks to 6 and assume that a marginal distribution of $\boldsymbol{x}$ is a uniform distribution on $[-1, 1]^2$. Discriminant functions associated with each dataset are generated by:

$$F_j(\boldsymbol{x}) = \begin{cases} (1 + c_{j,2})(x_1 - c_{j,1}) - x_2, & j = 1, 2, 3, \\ -(1 + c_{j,2})(x_1 - c_{j,1}) + x_2, & j = 4, 5, 6, \end{cases}$$

where $c_{j,1} \sim \mathcal{N}(0, 0.1^2)$ and $c_{j,2} \sim \mathcal{N}(0, 0.1^2)$. In addition, we randomly added a contamination noise on label y. Under these settings, we generated training dataset, including 400 examples, and the test dataset, including 600 examples. Generated datasets are shown in Figure 4. We observe that Datasets 1, 2 and 3 share a structure, and Datasets 4, 5 and 6 share another structure.

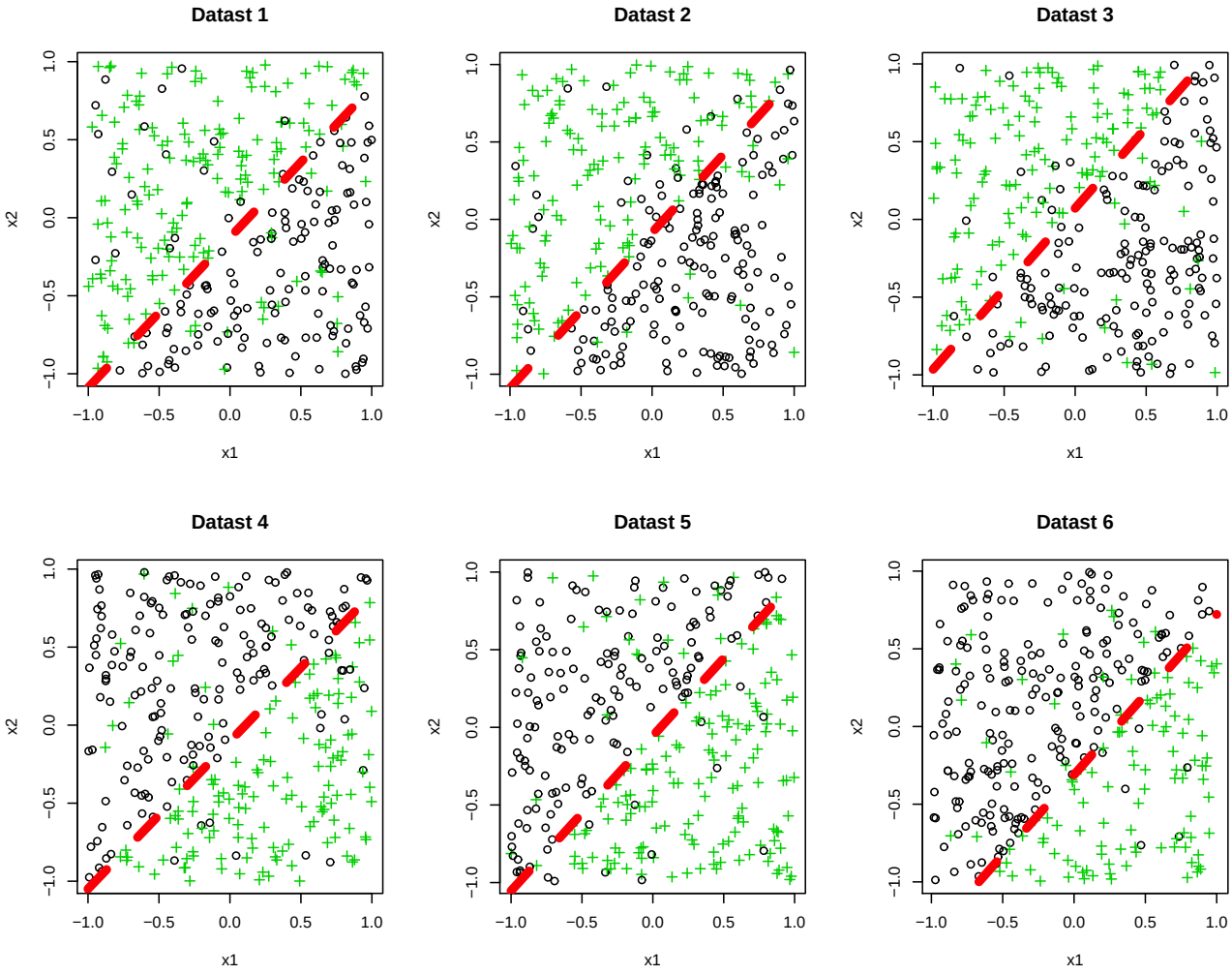

Figure 4. The six generated datasets and decision boundaries.

Figure 5 shows boxplots of the test errors of each method for datasets $\mathcal{D}_j$ ($j = 1, \ldots, 6$). We omitted the result of AdaBoost trained with all datasets simultaneously (D) from the figure, because its performance is significantly worse than those of the other methods: the median of classification errors is around 0.5. This is because the structures of Datasets 1, 2, 3 and Datasets 4, 5, 6 are opposite, and the labeling of concatenated dataset seems to be random. We observe that the proposed method with Scenario 2 (B) improves performance against individually-trained AdaBoost (C) and the proposed method in Scenario 1 (A). This is because the structure shared among Datasets 1, 2 and 3 does not have

information about Datasets 4, 5 and 6 (and *vice versa*), and Method (B) can ignore the influence of the irrelevant information by adjusting $\lambda_{k,j}$ responding to $\mathrm{IS}(q_{\hat{F}_j}, q_{\hat{F}_k}; r_k)$. Note that the performance of Method (A) is not so degraded, because the regularization parameter $\lambda_{k,j}$ was determined, so as to be zero, implying AdaBoost trained with the individual dataset.

Figure 6 shows examples of classification boundaries estimated by Methods A, B, C and D, for Dataset 6.

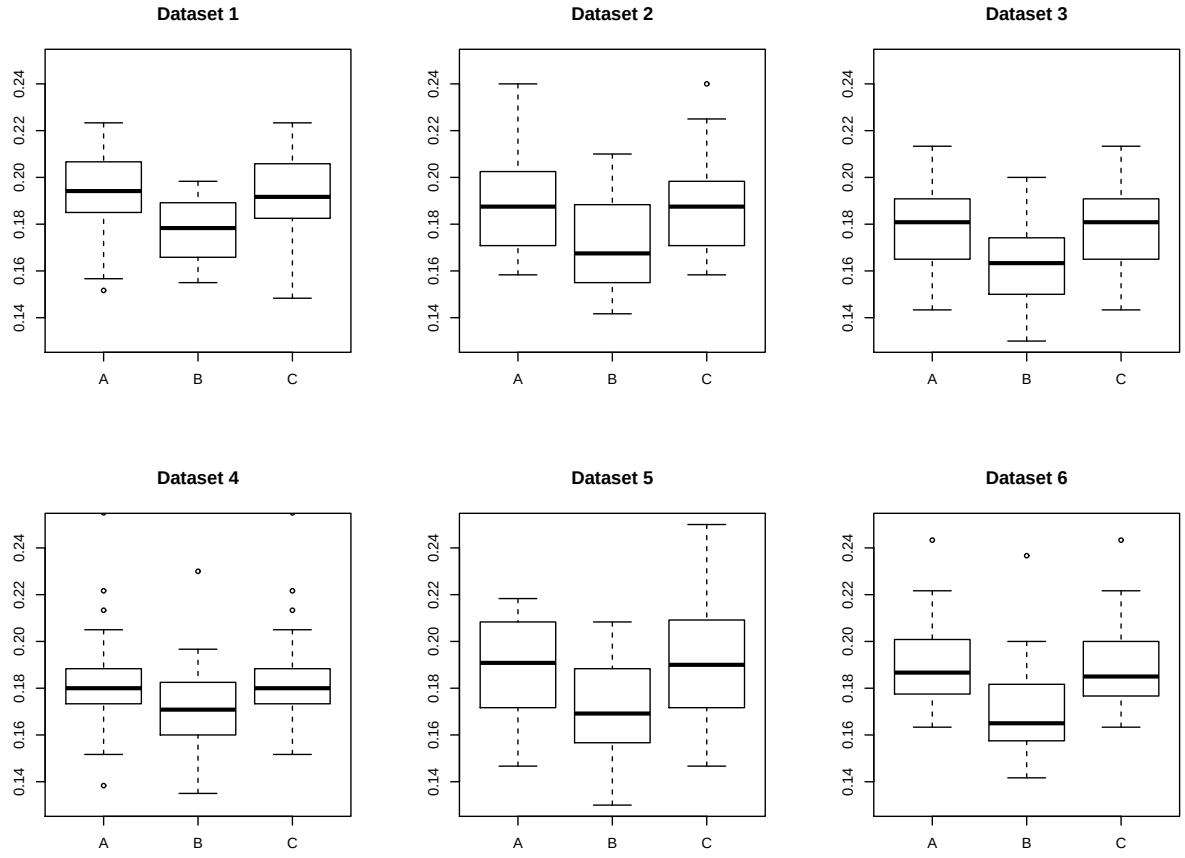

Figure 5. Boxplots of the test error of each method: A, Proposed method with λ in Scenario 1; B, proposed method with λ in Scenario 2; C, AdaBoost trained with the individual dataset ; for 6 datasets, over the 20 simulation trials.

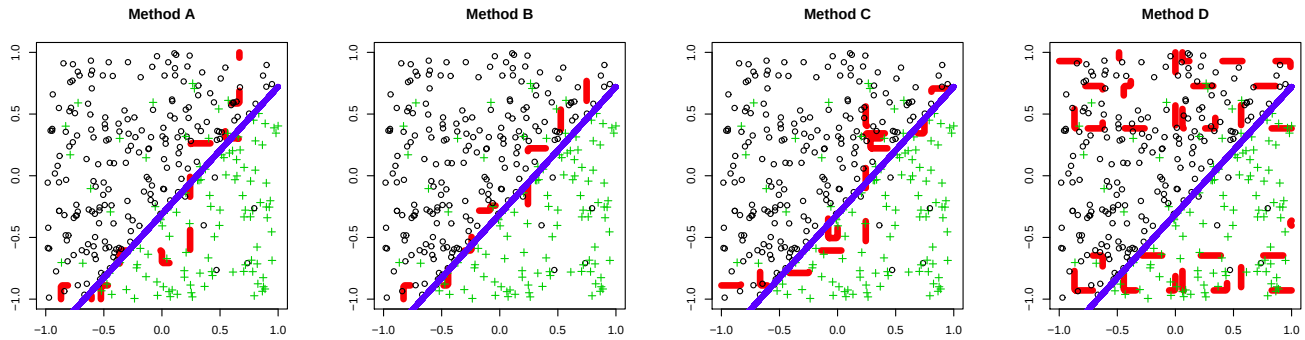

Figure 6. Classification boundaries by Methods A, B, C and D for Dataset 6. The blue line is the true classification boundary, and the red line represents the estimated classification boundary.

5.2. Real Dataset: School Dataset

In this section, we compared the proposed method (Scenario 2) to the a binary decision tree-based ensemble method, called extremely randomized trees (ExtraTrees) [23], applying to a real dataset, "school data", reported from the Inner London Education Authority [24]. The dataset consists of examination records of 15,362 students from 139 secondary schools, *i.e.*, we had 139 tasks. The dimension of input x is 27, in which original variables that are categorical were transformed into dummy variables. The original target variable y_0 represents score values in the range $[1, 70]$, and we transformed the target variable y_0 to a binary variable as:

$$y = \mathrm{sgn}(y_0 - 20).$$

We set the threshold to 20 to balance the ratio of classes ($-1 : +1 = 7930 : 7432$). We randomly divided the dataset of each tasks into 80% of the training dataset and remaining 20% test dataset. In addition, we used 20% of the divided training dataset as a validation dataset to determine the hyper-parameter λ and step number T. We repeated the above procedure 20 times and observed the average performance of the methods. Figure 7 shows the medians of error rates over 20 trials, by the proposed method and the ExtraTrees for 139 tasks. The horizontal axis indicates an index of a task, which is ranked in increasing order of the median error rate of the ExtraTrees. We observe that the proposed method is comparable to the ExtraTrees and especially has an advantage for tasks, in which the error rates of the ExtraTrees are large.

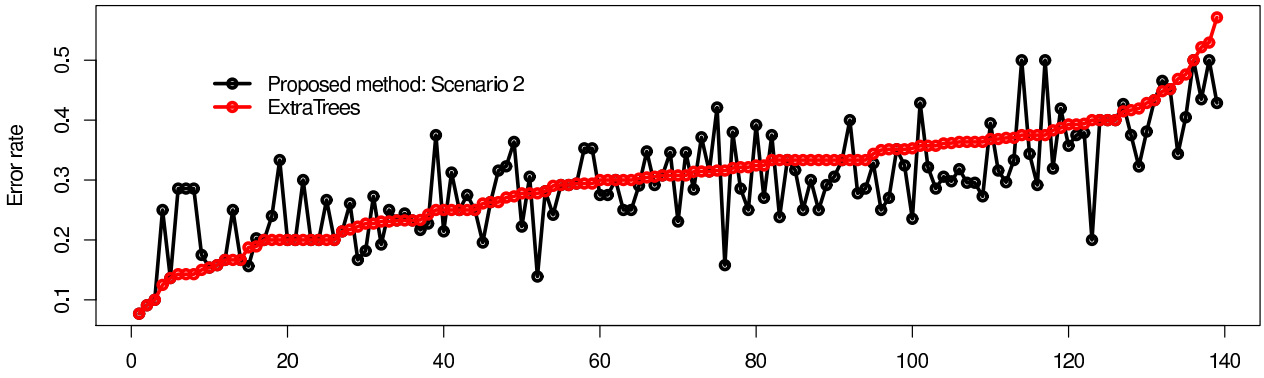

Figure 7. Medians of error rates by the proposed method and extremely randomized trees (ExtraTrees) for 139 tasks. The horizontal axis represents an index of a task, and the vertical axis indicates the median of error rates over 20 trials. Tasks are ranked in increasing order of the median error rate of the ExtraTrees.

6. Conclusions

In this paper, we investigate the properties of binary classification with the pseudo model and reveal that minimization of the Itakura–Saito distance between the empirical distribution and the pseudo model is equivalent to AdaBoost and provides suitable properties for the binary classification. In addition, we pointed out that the Itakura–Saito distance is a unique divergence, having a suitable property for estimation with the pseudo model in the framework of the Bregman divergence. Based on the framework,

we proposed a novel binary classification method for the multi-task learning, which incorporates shared information among tasks into the targeted task. The risk function of the proposed method is defined by the mixture of IS distance. The IS distance between pseudo models can be interpreted as the regularization term, incorporating shared information among tasks into the binary classifier for the target task. We investigated statistical properties of the risk function and derived computationally-feasible boosting-based algorithms. Furthermore, we considered a mechanism for the adjustment of the degree of information sharing and numerically investigated the validity of the proposed methods.

Acknowledgments

This study was partially supported by a Grant-in-Aid for Young Scientists (B), 25730018, from MEXT, Japan. Shinto Eguchi and Osamu Komori were supported by the Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology (CREST).

Author Contributions

Takashi Takenouchi made major contributions to employing the Itakura–Saito divergence, and Shinto Eguchi gave a proof for the characterization associated with the divergence. Takashi Takenouchi and Osamu Komori contributed to the statistical discussion for the multi-task learning.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

A. Proof of Proposition 1

By a variational calculation, a minimizer of Equation (5) satisfies:

$$\frac{\delta \,\mathrm{IS}(p, q_F; r)}{\delta F(\boldsymbol{x})} \propto \frac{e^{F_0(\boldsymbol{x})-F(\boldsymbol{x})} - e^{-F_0(\boldsymbol{x})+F(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}} = 0, \tag{41}$$

and $F = F_0$ satisfies the above equation for an arbitrary F_0, which concludes Equation (5). Furthermore,

$$\frac{\delta \,\mathrm{IS}(q_F, p; r)}{\delta F(\boldsymbol{x})} \propto \left(e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}\right)\left(e^{F(\boldsymbol{x})-F_0(\boldsymbol{x})} - e^{-F(\boldsymbol{x})+F_0(\boldsymbol{x})}\right) = 0, \tag{42}$$

and $F = F_0$ satisfies the above equation for an arbitrary F_0, concluding Equation (6).

B. Proof of Proposition 2

By a straightforward variational calculation, we observe that a minimizer $F_{\mathrm{KL},1}$ of Equation (8) satisfies:

$$\begin{aligned}\frac{\delta \,\mathrm{KL}(p, q_F; r)}{\delta F(\boldsymbol{x})} &\propto -p(+1|\boldsymbol{x}) + p(-1|\boldsymbol{x}) + \exp(F_{\mathrm{KL},1}(\boldsymbol{x})) - \exp(-F_{\mathrm{KL},1}(\boldsymbol{x})) \\ &= \frac{-e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}} + e^{F_{\mathrm{KL},1}(\boldsymbol{x})} - e^{-F_{\mathrm{KL},1}(\boldsymbol{x})} = 0,\end{aligned} \tag{43}$$

and $F_{\mathrm{KL},1} = F_0$ means $F_0(\boldsymbol{x}) = 0$ ($\forall \boldsymbol{x}$), which concludes Equation (8). Furthermore, for Equation (9), $F_{\mathrm{KL},2}$ satisfies:

$$\begin{aligned}&\frac{\delta \mathrm{KL}(q_F, p; r)}{\delta F(\boldsymbol{x})}\\ \propto&(F_{\mathrm{KL},2}(\boldsymbol{x}) - F_0(\boldsymbol{x}))(e^{F_{\mathrm{KL},2}(\boldsymbol{x})} + e^{-F_{\mathrm{KL},2}(\boldsymbol{x})}) + (e^{F_{\mathrm{KL},2}(\boldsymbol{x})} - e^{-F_{\mathrm{KL},2}(\boldsymbol{x})}) \log(e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})})\\ =&0,\end{aligned}$$

and $F_{\mathrm{KL},2} = F_0$ means $F_0(\boldsymbol{x}) = 0$ ($\forall \boldsymbol{x}$), concluding Equation (9).

C. Proof of Lemma 4

If Equation (15) holds, F_0 satisfies:

$$\begin{aligned}\left.\frac{\delta D_U(p, q_F; r)}{\delta F(\boldsymbol{x})}\right|_{F=F_0} &= \left(1 - \frac{1}{\sum_{y\in\mathcal{Y}} q_{F_0}(y|\boldsymbol{x})}\right) \sum_{y\in\mathcal{Y}} y\xi'(q_{F_0}(y|\boldsymbol{x}))q_{F_0}(y|\boldsymbol{x})^2\\ &\propto \xi'(e^{F_0(\boldsymbol{x})})e^{2F_0(\boldsymbol{x})} - \xi'(e^{-2F_0(\boldsymbol{x})})e^{-2F_0(\boldsymbol{x})}\\ &= 0.\end{aligned}$$

By setting $z = e^{F_0(\boldsymbol{x})}$, we have $z^2\xi'(z) = z^{-2}\xi'(z^{-1})$, and the function $\xi'(z)z^2$ is reflection-symmetric.

If Equation (16) holds, F_0 satisfies:

$$\begin{aligned}&\left.\frac{\delta D_U(q_F, p; r)}{\delta F(\boldsymbol{x})}\right|_{F=F_0}\\ =&\sum_{y\in\mathcal{Y}} yq_{F_0}(y|\boldsymbol{x}) \left\{\xi(q_{F_0}(y|\boldsymbol{x})) - \xi(\bar{q}_{F_0}(y|\boldsymbol{x}))\right\}\\ =&e^{F_0(\boldsymbol{x})}\left\{\xi(e^{F_0(\boldsymbol{x})}) - \xi\left(\frac{e^{F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}}\right)\right\} - e^{-F_0(\boldsymbol{x})}\left\{\xi(e^{-F_0(\boldsymbol{x})}) - \xi\left(\frac{e^{-F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}}\right)\right\}\\ =&0,\end{aligned}$$

implying that the function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ is reflection-symmetric.

D. Proof of Theorem 5

For the proof of the theorem, we firstly prepare the following lemmas.

Lemma 11. *Let $f(z)$ be a reflection-symmetric and holomorphic function on $z \neq 0$. Then, $a_k = b_k$ holds for all $k \geq 1$.*

Proof. The function f can be expressed as Equation (14), and let us assume that there exists an integer k_0, such that $a_{k_0} \neq b_{k_0}$. From the reflection-symmetric property, we have:

$$(a_{k_0} - b_{k_0})(z^{k_0} - z^{-k_0}) = 0 \tag{44}$$

for all $z > 0$, which contradicts $a_{k_0} \neq b_{k_0}$. □

Lemma 12. *Let $\xi(z)$ be a holomorphic function on $z \neq 0$. If two functions:*

$$\xi'(z)z^2, \text{ and } z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\} \tag{45}$$

are both reflection-symmetric, then $\xi(z) = \frac{c_1}{z} + c_0$.

Proof. We can express the function $\xi(z)$ by a Laurent series as:

$$\xi(z) = c + \sum_{k=1}^{\infty}\left(a_k z^k + b_k z^{-k}\right). \tag{46}$$

Then, we have:

$$\begin{aligned}\xi'(z)z^2 &= \sum_{k=1}^{\infty} k\left(a_k z^{k+1} - b_k z^{-k+1}\right)\\ &= -b_1 - 2b_2 z^{-1} + \sum_{k=1}^{\infty}\left(k a_k z^{k+1} - (k+2)b_{k+2}z^{-k-1}\right).\end{aligned} \tag{47}$$

Because of the assumption of reflection-symmetry for $z^2\xi'(z)$ and Lemma 11, we have $b_2 = 0$ and $ka_k = -(k+2)b_{k+2}$ for all $k \geq 1$. Thus, we obtain:

$$\begin{aligned}\xi(z) &= \int -\frac{b_1}{z^2} + \sum_{k=1}^{\infty} a_k\left(kz^{k-1} + kz^{-k-3}\right)dz\\ &= c + b_1 z^{-1} + \sum_{k=1}^{\infty} a_k\left(z^k - \frac{k}{k+2}z^{-k-2}\right).\end{aligned} \tag{48}$$

Then, we have:

$$\begin{aligned}&z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}\\ =&b_1(1-(z+z^{-1})) + \sum_{k=1}^{\infty} a_k\left\{z^{k+1}(1-(z+z^{-1})^{-k}) - \frac{k}{k+2}z^{-k-1}(1-(z+z^{-1})^{k+2})\right\}.\end{aligned} \tag{49}$$

From Equation (48) and the assumption of the reflection-symmetry of the function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$, we observe that for all z,

$$\begin{aligned}z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\} - z^{-1}\left\{\xi(z^{-1}) - \xi\left(\frac{z^{-1}}{z+z^{-1}}\right)\right\} &= \sum_{k=1}^{\infty} a_k h_k(z)\\ &= 0\end{aligned} \tag{50}$$

where:

$$h_k(z) = \left(z^{k+1} - z^{-k-1}\right)\left\{1 - (z+z^{-1})^{-k} + \frac{k}{k+2}\left\{1 - (z+z^{-1})^{k+2}\right\}\right\}. \tag{51}$$

Since $\{h_k(z)\}_{k=1}^{\infty}$ is functionally independent, we conclude that $a_k = 0$ for all $k \geq 1$ or, equivalently, $\xi(z) = c + \frac{b_1}{z}$. □

We now give a proof for Theorem 5 using Lemma 12.

Proof. If condition Equations (19) and (20) hold, functions $\xi'(z)z^2$ and $z\left\{\xi(z)-\xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ are both reflection-symmetric from Lemma 4. From Lemma 12, the reflection-symmetric property of these two functions implies $\xi(z)=\frac{b_1}{z}+c$. Since the function ξ should be defined on $z>0$, the generating function U derived from ξ is written as:

$$U(z)=\int \xi^{-1}(z)dz=b_1\log(c-z)+c_1\ (z<c) \tag{52}$$

where c_1 is a constant and $b_1<0$ holds because of the convexity of function U. Then, we have $U(\xi(z)))=b_1\log(-b_1)-b_1\log z+c_1(z>0)$, and the associated divergence is equivalent to the IS distance, *i.e.*,

$$\begin{aligned} D_U(p,q;r)&=\int r(\boldsymbol{x})\sum_{y\in\mathcal{Y}}\left\{-b_1\log\frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})}-p(y|\boldsymbol{x})\left\{\frac{b_1}{q(y|\boldsymbol{x})}-\frac{b_1}{p(y|\boldsymbol{x})}\right\}\right\}d\boldsymbol{x}\\ &=-b_1\int r(\boldsymbol{x})\sum_{y\in\mathcal{Y}}\left\{\log\frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})}+\frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})}-1\right\}d\boldsymbol{x}\\ &=-b_1\,\mathrm{IS}(p,q;r), \end{aligned} \tag{53}$$

up to the constant $-b_1$. □

E. Proof of Proposition 7

From Equation (28), we observe:

$$\begin{aligned} &F_k^*(\boldsymbol{x})\\ &=\log\frac{\sqrt{p_k(+1|\boldsymbol{x})+\frac{1}{4p_k(-1|\boldsymbol{x})}\left(\sum_{j\neq k}\lambda_{k,j}\left\{e^{-\eta_j(\boldsymbol{x})}-e^{\eta_j(\boldsymbol{x})}\right\}\right)^2}-\frac{1}{2\sqrt{p_k(-1|\boldsymbol{x})}}\sum_{j\neq k}\lambda_{k,j}\left\{e^{-\eta_j(\boldsymbol{x})}-e^{\eta_j(\boldsymbol{x})}\right\}}{\sqrt{p_k(-1|\boldsymbol{x})}}\\ &\simeq\frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})}+\frac{1}{P}\sum_{j\neq k}\lambda_{k,j}\eta_j(\boldsymbol{x}). \end{aligned}$$

F. Proof of Proposition 8

We observe that:

$$\begin{aligned} &\left(\bar{F}_k^*(\boldsymbol{x})-\frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2-\left(F_k^*(\boldsymbol{x})-\frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2\\ &=\frac{1}{4P^4(P+\lambda_k)^2}\left(\lambda_k\delta_k(\boldsymbol{x})-\sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})\right)\left((\lambda_k+2P)\delta_k(\boldsymbol{x})+\sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})\right), \end{aligned}$$

which implies the proposition.

G. Proof of Proposition 9

The minimizer of the risk function Equation (27) satisfies:

$$\begin{aligned}\frac{\delta L(F_1,\ldots,F_J)}{\delta F_k} \propto & e^{F_k(\boldsymbol{x})}\left\{\pi_k p_k(-1|\boldsymbol{x})+\sum_{j\neq k}(\pi_k\lambda_{k,j}+\pi_j\lambda_{j,k})e^{-F_j(\boldsymbol{x})}\right\}\\ & - e^{-F_k(\boldsymbol{x})}\left\{\pi_k p_k(+1|\boldsymbol{x})+\sum_{j\neq k}(\pi_k\lambda_{k,j}+\pi_j\lambda_{j,k})e^{F_j(\boldsymbol{x})}\right\}\\ =&0,\end{aligned}$$

implying Equation (36).

References

1. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75.
2. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
3. Argyriou, A.; Pontil, M.; Ying, Y.; Micchelli, C.A. A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems 19*; MIT Press: Cambridge, MA, USA, 2007.
4. Evgeniou, A.; Pontil, M. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*; MIT Press: Cambridge, MA, USA, 2007.
5. Dai, W.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 193–200.
6. Wang, X.; Zhang, C.; Zhang, Z. Boosted multi-task learning for face verification with applications to web image and video search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp.142–149.
7. Chapelle, O.; Shivaswamy, P.; Vadrevu, S.; Weiinberger, K.; Zhang, Y.; Tseng, B. Multi-task learning for boosting with application to web search ranking. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 1189–1198
8. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
9. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
10. Lefevre, A.; Bach, F.; Févotte, C. Itakura–Saito nonnegative matrix factorization with group sparsity. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech, 22–27 May 2011; pp. 21–24.
11. Takenouchi, T.; Komori, O.; Eguchi, S. A novel boosting algorithm for multi-task learning based on the Itakura–Saito divergence. In Proceedings of the Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21–26 September 2014; pp. 230–237.

12. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
13. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
14. Amari, S.; Nagaoka, H. *Methods of Information Geometry of Translations of Mathematical Monographs*; Oxford University Press: Providence, RI, USA, 2000; Volume 191.
15. Mihoko, M.; Eguchi, S. Robust blind source separation by beta divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
16. Cichocki, A.; Cruces, S.; Amari, S.I. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
17. Takenouchi, T.; Eguchi, S. Robustifying AdaBoost by adding the naive error rate. *Neural Comput.* **2004**, *16*, 767–787.
18. Takenouchi, T.; Eguchi, S.; Murata, T.; Kanamori, T. Robust boosting algorithm against mislabeling in multi-class problems. *Neural Comput.* **2008**, *20*, 1596–1630.
19. Lafferty, G.L.J. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems 14*; MIT Press: Cambridge, MA, USA, 2002.
20. Evgeniou, T.; Pontil, M. Regularized multi-task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 109–117.
21. Xue, Y.; Liao, X.; Carin, L.; Krishnapuram, B. Multi-task learning for classification with Dirichlet process priors. *J. Mach. Learn. Res.* **2007**, *8*, 35–63.
22. Mason, L.; Baxter, J.; Bartlett, P.; Frean, M. Boosting algorithms as gradient decent in function space. In *Advances in Neural Information Processing Systems 11*; MIT Press: Cambridge, MA, USA, 1999.
23. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42.
24. Goldstein, H. Multilevel modelling of survey data. *J. R. Stat. Soc. Ser. D* **1991**, *40*, 235–244.