The Canadian Journal of Statistics
Vol. 37, No. 1, 2009, Pages 119–139
La revue canadienne de statistique

New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation

Chunming ZHANG*, Yuan JIANG and Zuofeng SHANG

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

Key words and phrases: Asymptotic normality; Bayes optimal rule; consistency; local polynomial regression; loss function; prediction error.

MSC 2000: Primary 62F12, 62G20; secondary 62E20, 60F99.

Abstract: In statistical learning, regression and classification concern different types of the output variables, and the predictive accuracy is quantified by different loss functions. This article explores new aspects of Bregman divergence (BD), a notion which unifies nearly all of the commonly used loss functions in regression and classification. The authors investigate the duality between BD and its generating function. They further establish, under the framework of BD, asymptotic consistency and normality of parametric and nonparametric regression estimators, derive the lower bound of their asymptotic covariance matrices, and demonstrate the role that parametric and nonparametric regression estimation play in the performance of classification procedures and related machine learning techniques. These theoretical results and new numerical evidence show that the choice of loss function affects estimation procedures, whereas has an asymptotically relatively negligible impact on classification performance. Applications of BD to statistical model building and selection with non-Gaussian responses are also illustrated. *The Canadian Journal of Statistics* 37: 119–139; 2009 © 2009 Statistical Society of Canada

Résumé: En apprentissage statistique, la régression et la classification demandent différents types de variables de sortie et la précision prédictive est quantifiée par des fonctions de perte différentes. Cet article explore des nouveaux aspects de la divergence de Bregman (DB), une notion qui unifie presque toutes les fonctions de perte usuelles utilisées en régression et en classification. Les auteurs étudient la dualité entre la divergence de Bregman et sa fonction génératrice. De plus, ils établissent, dans le cadre DB, la cohérence asymptotique et la normalité des estimateurs de régression paramétrique et non paramétrique. Ils ont aussi obtenu une borne inférieure de leur matrice de variance-covariance asymptotique et ils ont démontré le rôle que les estimateurs de régression paramétrique et non paramétrique jouent dans la performance des procédures de classification et les techniques d'apprentissage machine. Ces résultats théoriques et de nouvelles évidences numériques semblent indiquer que le choix de la fonction de perte affecte les procédures d'estimation tandis qu'il a un impact non significatif sur les performance de classification. Cet article présente aussi des applications de la divergence de Bregman à la construction de modèles statistiques et à la sélection avec des variables non gaussiennes. *La revue canadienne de statistique* 37: 119–139; 2009 © 2009 Société statistique du Canada

1. INTRODUCTION

In statistical learning, the primary goals of regression and classification seem to be kept separate. Regression methods concern the "orderable" output variable and aim to estimate the regression function at points of the input variable, whereas the primary interest of classification rules for the "categorical" output variable is to forecast the most likely class label for the output.

As discussed in Friedman (1997), both regression and classification can be viewed from the common perspective of real valued prediction. Namely, given the training sample

* *Author to whom correspondence may be addressed.*
E-mail: cmzhang@stat.wisc.edu

$$\mathcal{T} = \{(\mathbb{X}_i, Y_i), i = 1, \ldots, n\}, \tag{1}$$

the goal of a supervised learning algorithm is to use (1) to construct a prediction rule for a future output Y at the observed value of the input variable $\mathbb{X}$. Depending on the nature of Y, the predictive error is quantified by different error measures. For example, the quadratic loss has nice analytical properties and is usually used in regression. However it is clearly not the most suitable loss function in classification problems where the 0–1 loss (or misclassification loss) is more realistic and commonly used in classification.

This article aims to study new aspects of Bregman divergence (BD), a notion which unifies nearly all of the commonly used loss functions in regression and classification. Particularly, we investigate the duality between BD and its generating function, which is shown to capture all the important statistical properties of BD. We further establish, under the framework of BD, asymptotic consistency and normality of parametric and nonparametric regression estimators, derive the lower bound of their asymptotic covariance matrices, and demonstrate the role that parametric and nonparametric regression estimation play in the performance of classification procedures and related machine learning techniques. The results will provide a more global insight into regression and classification methods.

There has been extensive research on relating divergence minimization to regression and classification. See Altun & Smola (2006), Nguyen, Wainwright & Jordan (2008) and references therein. This article takes a different approach. For instance, our inverse method not only provides necessary and sufficient conditions for a given loss being a BD but also derives an explicit formula for solving the generating function. From another perspective, our study reveals that under mild regularity conditions, the asymptotic distribution of the parametric regression estimator, under BD, relies on the loss only through the second derivative of its generating function; such dependence continues to arise from local regression estimation in the varying-coefficient regression model (Hastie & Tibshirani, 1993) with multivariate predictors, but, curiously, is entirely absent from a univariate nonparametric regression model. In the former two cases (i.e., parametric and varying-coefficient estimation), we manifest that if the generating function satisfies a "*generalized Bartlett identity*," then the asymptotic covariance matrix of the estimator achieves the lower bound; in the third case (i.e., univariate nonparametric regression estimation), we bypass the need to seek an optimal loss function. Besides, the impact of the estimation error of the regression estimator on the misclassification risk can analytically be assessed. These theoretical results, and new numerical evidence, show that the choice of loss function affects estimation procedures, whereas has an asymptotically relatively negligible impact on classification performance. Moreover, we illustrate that BD offers a versatile and useful tool for statistical model building and selection with non-Gaussian responses.

This article is organized as follows. Section 2 briefly reviews BD and quantifies a range of its new statistical properties. Section 3 explores the duality between BD and its generating function. Section 4 studies the asymptotic behaviours of parametric estimation under BD, whereas Section 5 establishes the asymptotic distribution of nonparametric function estimation under BD. Section 6 applies results developed in Sections 2–5 to classification. Section 7 presents simulation evaluations and demonstrates the applicability of BD to model building and selection. Technical details are postponed to Appendix.

2. STATISTICAL PROPERTIES OF BREGMAN DIVERGENCE

2.1. Bregman Divergence

For a given concave q-function, Brègman (1967) introduced a device for constructing a bivariate function,

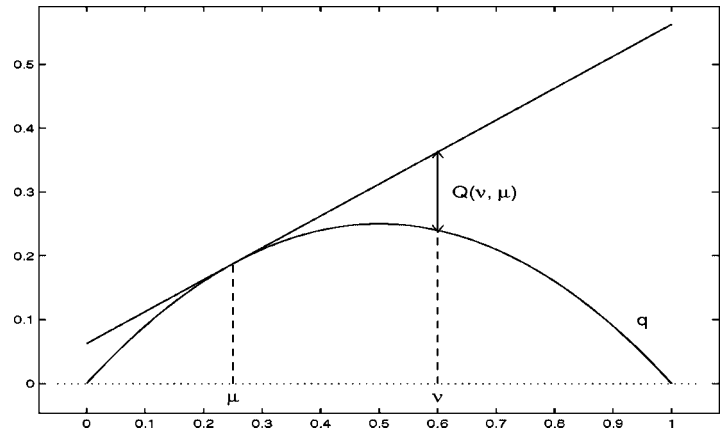

FIGURE 1: Illustration of $Q(\nu, \mu)$ as defined in (2). The concave curve is the q-function; the two dashed lines indicate locations of ν and μ; the solid strict line is $q(\mu) + (\nu - \mu)q'(\mu)$; the length of the vertical line with arrows at each end is $Q(\nu, \mu)$.

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu), \tag{2}$$

defined for $(\nu, \mu) \in \mathcal{M}_0 \times \mathcal{M}_1$, where $\mathcal{M}_k$ is the set of points at which the kth derivative of the q-function exists. Figure 1 displays Q and the corresponding q-function. It is readily seen that the concavity of q ensures the non-negativity of Q. Moreover, for a strictly concave q-function, $Q(\nu, \mu) = 0$ is equivalent to $\nu = \mu$. However, since $Q(\nu, \mu)$ is not generally symmetric in arguments, Q is not a "metric" or "distance" in the strict sense. Hence, we call Q the "Bregman divergence" (BD) and call q the "generating function" of Q. Refer to Lafferty, Della Piestra & Della Piestra (1997) for certain examples of BD in the machine learning literature, Lafferty (1999) and Azoury & Warmuth (2001) for the use of BD in the construction and analysis of online learning algorithms, Kivinen & Warmuth (1999) for a generalized boosting update using BD, Grünwald & Dawid (2004) for decision-based divergence function, Efron (1986, 2004) for the estimation of prediction error under BD, and Altun & Smola (2006) for the duality of divergence minimization and statistical inference methods.

2.2. Convexity and Bartlett Identity for BD Q

Part I: Convexity. We first examine the convexity of a BD $Q(\cdot, \cdot)$ in its first argument. Since the q-function is concave, it is straightforward to show that for any fixed $\mu \in \mathcal{M}_1$, $Q(\nu, \mu)$ is convex in its first argument $\nu \in \mathcal{M}_0$.

We then discuss the convexity of a BD $Q(\cdot, \cdot)$ in its second argument. Note that some functions $Q(\nu, \mu)$, like those associated with $q(\mu) = -\mu^4$ and the misclassification loss (in Section 3.3), are non-convex in its second argument μ. For a general discussion, let us assume stronger smoothness conditions on the q-function. It follows from (2) that

$$\frac{\partial Q(\nu, \mu)}{\partial \mu} = (\nu - \mu)q''(\mu), \quad \mu \in \mathcal{M}_2, \tag{3}$$

$$\frac{\partial^2 Q(\nu, \mu)}{\partial \mu^2} = (\nu - \mu)q'''(\mu) - q''(\mu), \quad \mu \in \mathcal{M}_3. \tag{4}$$

Thus, for fixed $\nu \in \mathcal{M}_0$, $Q(\nu, \mu)$ is typically a V-shaped function of μ, but is not necessarily convex in μ, even if stronger assumptions are imposed on q. Furthermore, it can be deduced from (3) and (4) that all the kth order partial derivatives of $Q(\nu, \mu)$ with respect to $\mu \in \mathcal{M}_{k+1}$ are linear in ν.

Part II: Bartlett identity. In what follows, we denote by $m(\mathtt{x}) = E(Y|\mathtt{X} = \mathtt{x})$ the conditional regression function, and by $\mathrm{var}(Y|\mathtt{X} = \mathtt{x})$ the conditional variance function. From (3) and (4), we observe that

$$E\left\{\frac{\partial Q(Y, m(\mathtt{x}))}{\partial m(\mathtt{x})}\Big|\mathtt{X} = \mathtt{x}\right\} = 0, \quad \text{and} \quad E\left\{\frac{\partial^2 Q(Y, m(\mathtt{x}))}{\partial m(\mathtt{x})^2}\Big|\mathtt{X} = \mathtt{x}\right\} = -q''(m(\mathtt{x})).$$

Accordingly, under a BD Q, the Bartlett identity

$$E\left\{\frac{\partial^2 Q(Y, m(\mathtt{x}))}{\partial m(\mathtt{x})^2}\Big|\mathtt{X} = \mathtt{x}\right\} = E\left[\left\{\frac{\partial Q(Y, m(\mathtt{x}))}{\partial m(\mathtt{x})}\right\}^2\Big|\mathtt{X} = \mathtt{x}\right]$$

holds if and only if

$$q''(m(\mathtt{x})) = -1/\mathrm{var}(Y|\mathtt{X} = \mathtt{x}). \tag{5}$$

This result provides a likelihood view point of BD.

2.3. Bayes Rule and Pythagorean Equality

The assessment and estimation of prediction error play an important role in developing reliable prediction rules in regression and classification. Theorem 1 states that the risk $E\{Q(Y, \mu(\mathtt{X}))\}$, associated with a BD Q, is minimized (with respect to μ) by the optimal "Bayes" rule $m(\mathtt{X})$. The proof follows from the definition (2) and can be found in Banerjee, Guo & Wang (2005).

Theorem 1. *Suppose that Q is a BD as defined in (2) for $(\nu, \mu) \in \mathcal{M}_0 \times \mathcal{M}_1$ and $Y \in \mathcal{M}_0$ is a random variable. Assume that $m(X)$ is measurable and $m(x) \in \mathcal{M}_1$. Then among all measurable functions μ such that $\mu(x) \in \mathcal{M}_1$, $\arg\min_{\mu \in \mathcal{M}_1} E\{Q(Y, \mu(X))\} = m(X)$.*

Throughout the rest of the article, we assume that $(\mathtt{X}_i, Y_i)$, $i = 1, \ldots, n$, in the training sample (1) are independent pairs from a common distribution of $(\mathtt{X}, Y)$, and that $\widehat{m}(\mathtt{x})$ is an estimate of $m(\mathtt{x})$, based on the training sample. Suppose that a test point $(\mathtt{X}^o, Y^o)$ follows the distribution of $(\mathtt{X}, Y)$ and is independent of the training sample. The conditional prediction error (cPE) of the rule $\widehat{m}(\mathtt{x})$ is defined as

$$r(\mathtt{x}) = E\{Q(Y^o, \widehat{m}(\mathtt{x}))|\mathcal{T}, \mathtt{X}^o = \mathtt{x}\},$$

and the expected prediction error (ePE) is defined by $E\{Q(Y^o, \widehat{m}(\mathtt{x}))|\mathtt{X}^o = \mathtt{x}\}$. Theorem 2 supplies an additive decomposition of cPE and indicates that ePE under BD, when projected on the Bayes rule $m(\mathtt{x})$, satisfies the Pythagorean equality.

Theorem 2. *Suppose that Q is a BD as defined in (2) for $(\nu, \mu) \in \mathcal{M}_0 \times \mathcal{M}_1$ and $Y \in \mathcal{M}_0$ is a random variable. Assume that $m(X)$ is measurable and $m(x) \in \mathcal{M}_1$. Define*

$$r_B(x) = E\{Q(Y^o, m(x))|X^o = x\} = q(m(x)) - E\{q(Y^o)|X^o = x\}. \tag{6}$$

Then cPE *has the decomposition,*

$$r(x) = r_B(x) + Q(m(x), \widehat{m}(x)), \tag{7}$$

and ePE *fulfills the Pythagorean equality,*

$$E\{Q(Y^o, \widehat{m}(x))|X^o = x\} = E\{Q(Y^o, m(x))|X^o = x\} + E\{Q(m(x), \widehat{m}(x))\}. \tag{8}$$

Thus both cPE *and* ePE *are minimized, with respect to $\widehat{m}(x)$, at the Bayes rule $m(x)$.*

On the right side of (8), the first term is $r_B(\mathrm{x})$, an irreducible error due to the nature of the response, whereas the second term, $E\{Q(m(\mathrm{x}), \widehat{m}(\mathrm{x}))\}$, corresponds to the function estimation error. Hence, the function estimate $\widehat{m}$ affects the prediction error only through the function estimation error.

3. DUALITY BETWEEN Q AND q

Given any concave q-function, the bivariate function $Q(\nu, \mu)$ is well defined by (2). In this section, we aim to address the inverse question: Given a loss function $Q(Y, \mu)$, how to recover the generating q-function? Investigating this inverse problem has the following important implications:

I. As can be seen from (6), the ideal Bayes rule under a BD Q serves as a benchmark, achieving the lowest possible risk which only depends on the q-function.
II. Evaluation of the function estimation error in (8) requires not only $Q(Y, \mu)$, but also the full knowledge of $Q(\nu, \mu)$. However, in general, $Q(Y, \mu)$ alone may not entirely determine $Q(\nu, \mu)$. For example, consider $Q(Y, \mu) = Y(1-\mu)^2 + (1-Y)\mu^2$ where Y is a binary random variable. There is a need to recover the q-function before obtaining $Q(\nu, \mu)$.
III. As we will show in Section 3.4, once we know the q-function for binary classification, the counterpart and loss function for multi-class classification can easily be accommodated.
IV. The asymptotic distributions of parametric and nonparametric regression estimators under BD depend on Q only through q''. See Theorems 5 and 8.

Nonetheless, explicitly solving the generating q-function from a given Q-loss is non-trivial and some approach via differential equation will be unnecessarily complicated. Theorem 3 reveals that the q-function can explicitly be obtained from $Q(Y, \mu)$,

$$Q(Y, \mu) = -q(Y) + q(\mu) + (Y - \mu)q'(\mu). \tag{9}$$

Theorem 3. *Let $a = \mu_0 < \mu_1 < \cdots < \mu_K < \mu_{K+1} = b$ and set $\mathcal{I} = \cup_{j=0}^{K}(\mu_j, \mu_{j+1})$, where $K \geq 0$. For a random variable Y taking values only in $[a, b]$, suppose that $Q(Y, \mu)$ is a loss function defined for $\mu \in \mathcal{I} \cup \{a, b\}$ which fulfills that $Q(Y, Y) = 0$ for all Y. Assume that $Q(Y, \mu)$ is continuous for $\mu \in \mathcal{I} \cup \{a, b\}$ and $(\partial/\partial\mu)Q(Y, \mu)$ is continuous for $\mu \in \mathcal{I}$. Then Conditions* **A** *and* **B** *are equivalent:*

A. *There exists a concave function q such that (9) holds for all $\mu \in \mathcal{I} \cup \{a, b\}$;*
B. *For all $\mu \in \mathcal{I}$ and $Y \neq \mu$,*

$$\frac{1}{Y - \mu}\frac{\partial Q(Y, \mu)}{\partial \mu} \leq 0, \quad \text{and is free of } Y; \tag{10}$$

for $j = 1, \ldots, K$, $Q(Y, \mu_j+) = \lim_{\mu \downarrow \mu_j} Q(Y, \mu)$ and $Q(Y, \mu_j-) = \lim_{\mu \uparrow \mu_j} Q(Y, \mu)$ finitely exist, and for $Y \neq \mu_j$,

$$\frac{Q(Y, \mu_j+) - Q(Y, \mu_j-)}{Y - \mu_j} \leq 0, \quad \text{and is free of } Y.$$

Moreover, if Condition **B** *holds, then the generating q-function is piecewisely given by*

$$q(\mu) = \int_{\mu_j}^{\mu} \frac{\mu - s}{Y - s}\frac{\partial Q(Y, s)}{\partial s}\,\mathrm{d}s + (\mu - \mu_j)C_j + D_j, \tag{11}$$

for $\mu \in [\mu_j, \mu_{j+1})$, $j = 0, \ldots, K$, *where* $C_0 = D_0 = 0$ *and for* $j = 1, \ldots, K$, $C_j = \sum_{k=1}^{j}(p_k + \delta_k)$, *and* $D_j = \sum_{l=1}^{j}\{\int_{\mu_{l-1}}^{\mu_l} \frac{\mu_l - s}{Y - s} \frac{\partial Q(Y,s)}{\partial s}\, \mathrm{d}s + (\mu_l - \mu_{l-1})C_{l-1}\}$, *in which* $p_j = \int_{\mu_{j-1}}^{\mu_j} \frac{1}{Y-s} \frac{\partial Q(Y,s)}{\partial s}\, \mathrm{d}s$ *and* $\delta_j = \{Q(Y, \mu_j+) - Q(Y, \mu_j-)\}/(Y - \mu_j)$.

Remark 1. *Throughout the article, we will not distinguish between equivalent functions* q_1 *and* q_2 *in* $\mathcal{M}$ *(i.e., there exist constants a and b such that* $q_1(\mu) = q_2(\mu) + a\mu + b$ *for all* $\mu \in \mathcal{M}$*), since equivalent q-functions will generate an identical Q-function. In the f-divergence setting, the non-uniqueness of the corresponding generating function was also stated in Nguyen, Wainwright & Jordan (2008).*

On the other hand, if $E\{Q(Y, \mu)|\mathrm{X} = \mathrm{x}\}$ depends on the conditional distribution of Y only through its conditional regression function $m(\mathrm{x})$, then Corollaries 1–2 will produce the generating q-function more easily and under weaker smoothness assumptions on Q than Theorem 3.

Corollary 1. *Suppose that* $Q(Y, \mu)$ *is a loss function, well defined for* $\mu \in \mathcal{N}_1$ *and a random variable Y. Assume that* $E\{Q(Y, \mu)|X = x\}$ *depends only on* μ *and* $m(x)$*. Denote by* $\mathcal{E}(\mu; m(x))$ *the expression* $E\{Q(Y, \mu)|X = x\}$*, with all additive terms independent of* μ *removed. If (9) holds for all* $\mu \in \mathcal{N}_1$*, then the generating q-function is given by*

$$q(\mu) = \mathcal{E}(\mu; \mu), \quad \mu \in \mathcal{N}_1. \tag{12}$$

Corollary 2. *Assume that* $Y|X = x \sim$ Bernoulli$\{m(x)\}$*. Suppose that* $Q(Y, \mu)$ *is a loss function, defined for* $\mu \in \mathcal{N}_1 \subseteq [0, 1]$*, which fulfills* $Q(Y, Y) \equiv 0$ *for all* $Y \in \{0, 1\}$ *and* $\{0, 1\} \subseteq \mathcal{N}_1$*. If (9) holds for all* $\mu \in \mathcal{N}_1$*, then the generating q-function is given by*

$$q(\mu) = \mu Q(1, \mu) + (1 - \mu)Q(0, \mu), \quad \mu \in \mathcal{N}_1. \tag{13}$$

In particular, the q-function given by the form (13) satisfies

$$q(Y) \equiv 0, \quad q'(\mu) = Q(1, \mu) - Q(0, \mu), \quad \mu \in \mathcal{N}_1. \tag{14}$$

In summary, to recover the q-function from a given loss $Q(Y, \mu)$, we first check Condition **B** for the validity of (9). Then apply (11), or (12), or (13) to acquire the q-function. Applications of (11), (12), and (13) are illustrated below to three situations, ranging respectively from general to specific assumptions on the probability distribution of Y.

3.1. Quasi-Likelihood Function: Application of (11)

A quasi-likelihood function Q^* was introduced in Wedderburn (1974) to relax the distributional assumption on a random variable Y via the specification, $\partial Q^*(\mu; Y)/\partial\mu = (Y - \mu)/V(\mu)$, in which it is assumed that $\mathrm{var}(Y|\mathrm{X} = \mathrm{x}) = \sigma^2 V\{E(Y|\mathrm{X} = \mathrm{x})\}$ for a nuisance parameter $\sigma^2 > 0$ and a given continuous function $V > 0$. Since Condition **B** is satisfied, (11) yields

$$q(\mu) = \int_{-\infty}^{\mu} \frac{s - \mu}{V(s)}\, \mathrm{d}s. \tag{15}$$

3.2. Exponential Family of Probability Functions: Application of (12)

A special case of the quasi-likelihood function is the exponential family distribution, where the conditional probability function of Y given $\mathrm{X} = \mathrm{x}$ is $f_{Y|\mathrm{X}}(y|\mathrm{x}) = \exp[\{y\theta(\mathrm{x}) - b(\theta(\mathrm{x}))\}/a(\psi) +$

$c(y, \psi)$], for some known functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$, where $\theta(\mathrm{x})$ is called a canonical parameter and ψ is called a dispersion parameter, respectively. For the commonly used Kullback–Leibler divergence (or the deviance loss) defined by $Q(Y, \mu) = 2[Y(\widetilde{\theta} - \theta) - \{b(\widetilde{\theta}) - b(\theta)\}]$, where $b'(\widetilde{\theta}) = Y$ and $b'(\theta) = \mu$, assume that b' is a bijection. Since Condition **B** is fulfilled, (12) gives

$$q(\mu) = 2\{b(\theta) - \mu\theta\}, \quad \text{where } b'(\theta) = \mu. \tag{16}$$

This approach is more convenient than (11) for obtaining the q-function.

3.3. Binary Response: Application of (13)

For a binary variable $Y|\mathrm{X} = \mathrm{x} \sim \text{Bernoulli}\{m(\mathrm{x})\}$, its distribution belongs to the exponential family. We apply (13) to obtain $q(\mu)$ and $Q(\nu, \mu)$ associated with four types of loss functions $Q(Y, \mu)$, where $0 < \mu < 1$.

Example 1. The exponential loss used in AdaBoost (Hastie, Tibshirani & Friedman, 2001) is $Q(Y, \mu) = \mathrm{e}^{-(2Y-1)F(\mu)}$, where $F(\mu) = 2^{-1}\ln\{\mu/(1-\mu)\}$. Since Condition **B** is fulfilled, (13) gives that $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$. Hence (2) gives the Q-function, $Q(\nu, \mu) = [\{\mu(1-\nu)\}^{1/2} - \{\nu(1-\mu)\}^{1/2}]^2/\{\mu(1-\mu)\}^{1/2}$.

Example 2. The misclassification loss is

$$Q(Y, \mu) = |Y - c|\,\mathrm{I}\{Y \neq \mathrm{I}[\mu > c]\}, \tag{17}$$

where $\mathrm{I}[\cdot]$ is an indicator function, and the constant $c \in (0, 1)$ denotes the misclassification cost with $c = 1/2$ for equal cost. Since Condition **B** is fulfilled, (13) gives that

$$q(\mu) = \min\{\mu(1-c), (1-\mu)c\}. \tag{18}$$

Hence from (2), we get

$$Q(\nu, \mu) = |\nu - c|\,\mathrm{I}\{\mathrm{I}[\nu > c] \neq \mathrm{I}[\mu > c]\}. \tag{19}$$

Example 3. The polynomial loss is $Q(Y, \mu) = |Y - \mu|^k = Y(1-\mu)^k + (1-Y)\mu^k$, where $k > 0$ is a constant. For $Y \neq \mu$, we observe that $(\partial/\partial\mu)Q(Y, \mu) = -k(Y - \mu)|Y - \mu|^{k-2}$. Thus Condition **B** is fulfilled if and only if $k = 2$ which corresponds to the quadratic loss. As a result, the absolute loss with $k = 1$ does not belong to BD. For the quadratic loss, (13) gives that $q(\mu) = \mu(1-\mu)$. Hence from (2), we obtain the Q-function, $Q(\nu, \mu) = (\nu - \mu)^2$.

Example 4. The twice negative binomial log-likelihood is $Q(Y, \mu) = -2\{Y\ln(\mu) + (1 - Y)\ln(1-\mu)\}$, where $0\ln(0) = 0$ by definition. Since Condition **B** is fulfilled, (13) gives that $q(\mu) = -2\{\mu\ln(\mu) + (1-\mu)\ln(1-\mu)\}$. Hence (2) gives the Q-function, $Q(\nu, \mu) = 2[\nu\ln(\nu/\mu) + (1-\nu)\ln\{(1-\nu)/(1-\mu)\}]$.

3.4. Multi-Class Response

For a multi-class response $Y \in \mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$, where $K \geq 2$, we now demonstrate that the notion of Bregman divergence can be generalized from binary responses. For vectors $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$, both of which are discrete probability measures, the definition (2) can be extended from scalar arguments to vectors $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ as follows,

$$Q(\boldsymbol{\nu}, \boldsymbol{\mu}) = -q(\boldsymbol{\nu}) + q(\boldsymbol{\mu}) + (\boldsymbol{\nu} - \boldsymbol{\mu})^{\mathrm{T}}\nabla q(\boldsymbol{\mu}),$$

where $\nabla q(\boldsymbol{\mu}) = ((\partial/\partial\mu_1)q(\boldsymbol{\mu}), \ldots, (\partial/\partial\mu_K)q(\boldsymbol{\mu}))^{\mathrm{T}}$ is the gradient vector of q. Define $\mathbf{y} = (\mathrm{I}(Y = \mathcal{G}_1), \ldots, \mathrm{I}(Y = \mathcal{G}_K))^{\mathrm{T}}$. In addition, the choice of $q(\boldsymbol{\mu})$ can be extended from the counterpart developed in Section 3.3 for binary responses. The scheme is illustrated below with four types of multi-class loss functions.

Example 1. Define $q(\boldsymbol{\mu}) = \sum_{j=1}^{K} 2\{\mu_j(1-\mu_j)\}^{1/2}$. Then

$$Q(\boldsymbol{\nu}, \boldsymbol{\mu}) = \sum_{j=1}^{K} [\{\mu_j(1-\nu_j)\}^{1/2} - \{\nu_j(1-\mu_j)\}^{1/2}]^2 / \{\mu_j(1-\mu_j)\}^{1/2}$$

This gives the exponential loss $Q(\mathbf{y}, \boldsymbol{\mu}) = \sum_{j=1}^{K} \mathrm{e}^{-\{2\mathrm{I}[Y=\mathcal{G}_j]-1\}F(\mu_j)}$, where $F(\mu) = 2^{-1}\ln\{\mu/(1-\mu)\}$. Hastie, Tibshirani & Friedman (2001, p. 310) mentioned "*We know of no natural generalization of the exponential criterion for K classes.*" The derivation we provide above indeed generalizes the exponential loss from 2-classes to K-classes.

Example 2. Define $q(\boldsymbol{\mu}) = 1 - \max_{1\le j\le K}\mu_j$ and $\mathsf{k}^*(\boldsymbol{\mu}) = \arg\max_{1\le j\le K}\mu_j$. Then $Q(\boldsymbol{\nu}, \boldsymbol{\mu}) = \nu_{\mathrm{k}^*(\boldsymbol{\nu})} - \nu_{\mathrm{k}^*(\boldsymbol{\mu})}$, which gives the misclassification loss, $Q(\mathbf{y}, \boldsymbol{\mu}) = \mathrm{I}[Y \ne \mathcal{G}_{\mathrm{k}^*(\boldsymbol{\mu})}]$.

Example 3. Define $q(\boldsymbol{\mu}) = \sum_{j=1}^{K}\mu_j(1-\mu_j)$. Then $Q(\boldsymbol{\nu}, \boldsymbol{\mu}) = \sum_{j=1}^{K}(\nu_j - \mu_j)^2$. This gives the quadratic loss $Q(\mathbf{y}, \boldsymbol{\mu}) = \sum_{j=1}^{K}(y_j - \mu_j)^2$.

Example 4. Define $q(\boldsymbol{\mu}) = -\sum_{j=1}^{K}\mu_j\ln(\mu_j)$. Then $Q(\boldsymbol{\nu}, \boldsymbol{\mu}) = \sum_{j=1}^{K}\nu_j\ln(\nu_j/\mu_j)$. This gives the relative entropy, and the corresponding multinomial deviance (or negative log-likelihood), $Q(\mathbf{y}, \boldsymbol{\mu}) = -\sum_{j=1}^{K}\mathrm{I}(Y = \mathcal{G}_j)\ln(\mu_j)$.

4. PARAMETRIC ESTIMATION UNDER BREGMAN DIVERGENCE

In this section, we aim to study the parametric estimation of $m(\mathrm{x})$ under BD. This is achieved by estimating $F(m(\mathrm{x}))$, for a known link function $F(\cdot)$. Define $\eta(\mathrm{x}) = F(m(\mathrm{x}))$. We assume that $\eta(\mathrm{x}) = \alpha_0 + \mathrm{x}^{\mathrm{T}}\boldsymbol{\alpha}$ for some unknown parameters α_0 and $\boldsymbol{\alpha}$. Based on the independent training data $\{(\mathrm{x}_i, Y_i)_{i=1}^n\}$ from the population (x, Y), the minimum BD parametric estimator $(\widehat{\alpha}_0, \widehat{\boldsymbol{\alpha}})$ of $(\alpha_0, \boldsymbol{\alpha})$ is defined to be the minimizer of the criterion function,

$$\ell_n(\alpha_0, \boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, F^{-1}(\alpha_0 + \mathrm{x}_i^{\mathrm{T}}\boldsymbol{\alpha})), \tag{20}$$

where the loss function Q is a Bregman divergence. Define $\mathrm{x} = (X_1, \ldots, X_d)^{\mathrm{T}}$ and $\widetilde{\mathrm{x}} = (1, \mathrm{x}^{\mathrm{T}})^{\mathrm{T}}$. Set $\widetilde{\boldsymbol{\alpha}} = (\alpha_0, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$, which is in a parameter space $\Theta \subset \mathbb{R}^{d+1}$. The asymptotic consistency and normality of the minimum BD parametric estimator are presented in Theorems 4 and 5, respectively.

Theorem 4. *Let $\widetilde{\boldsymbol{\alpha}}^{(0)}$ denote the true value of $\widetilde{\boldsymbol{\alpha}}$. Suppose that Q is a BD. Assume Condition* **C** *in Appendix. Define $\widehat{\widetilde{\boldsymbol{\alpha}}} = (\widehat{\alpha}_0, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}})^{\mathrm{T}}$. Then as $n \to \infty$, the minimum BD parametric estimator $\widehat{\widetilde{\boldsymbol{\alpha}}}$ is asymptotically consistent to $\widetilde{\boldsymbol{\alpha}}^{(0)}$.*

Theorem 5. *Let $\widetilde{\boldsymbol{\alpha}}^{(0)}$ denote the true value of $\widetilde{\boldsymbol{\alpha}}$. Suppose that Q is a BD. Assume Condition* **D** *in Appendix. Define $\widehat{\widetilde{\boldsymbol{\alpha}}} = (\widehat{\alpha}_0, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}})^{\mathrm{T}}$. Then as $n \to \infty$, the minimum BD parametric estimator $\widehat{\widetilde{\boldsymbol{\alpha}}}$ is asymptotically normal, $\sqrt{n}\{\widehat{\widetilde{\boldsymbol{\alpha}}} - \widetilde{\boldsymbol{\alpha}}^{(0)}\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathsf{H}_0^{-1}\Omega_0\mathsf{H}_0^{-1})$, where $\Omega_0 = E[\mathrm{var}(Y|X)\{q''(m(X)\}^2\{F'(m(X))\}^{-2}(\widetilde{\widetilde{XX}}^{\mathrm{T}})]$ and $\mathsf{H}_0 = -E[q''(m(X))\{F'(m(X))\}^{-2}(\widetilde{\widetilde{XX}}^{\mathrm{T}})]$.*

Theorem 5 reveals that the asymptotic distribution of the parametric estimator under BD depends on the choice of the loss function Q only through the second derivative of its generating q-function. As will be seen from nonparametric estimators in Section 5, this dependence continues to arise (in Theorem 8) from the varying-coefficient regression model with multivariate predictors, but is entirely relieved (in Theorem 7) from the univariate nonparametric regression model.

Is there an optimal choice of the q-function such that the asymptotic covariance matrix in Theorem 5 achieves its lower bound? (For two symmetric matrices A and B, we say $A \geq B$ if $A - B$ is non-negative definite.) Theorem 6 manifests that the optimal q-function satisfies the "*generalized Bartlett identity*,"

$$q''(m(\mathtt{x})) = -c/\mathrm{var}(Y|\mathtt{X} = \mathtt{x}), \quad \text{for a constant } c > 0, \tag{21}$$

which includes the conventional Bartlett identity (5) as a special case.

Theorem 6. *If the q-function satisfies (21), then the asymptotic covariance matrix of $\widehat{\widetilde{\boldsymbol{\alpha}}}$ in Theorem 5 achieves the lower bound* $(E[1/\mathrm{var}(Y|X)\{F'(m(X))\}^{-2}\widetilde{X}\widetilde{X}^{\mathrm{T}}])^{-1}$.

Using the criterion (21), we are able to show that the q-functions given in (15) and (16) for quasi-likelihood function and exponential family of probability functions respectively satisfy the "*generalized Bartlett identity*." In the particular case of binary responses, the binomial deviance loss satisfies the "*generalized Bartlett identity*," but the misclassification loss, quadratic loss and exponential loss do not.

5. NONPARAMETRIC ESTIMATION UNDER BREGMAN DIVERGENCE

In this section, we study the nonparametric estimation of $m(\mathtt{x})$ under BD, via local estimation techniques. Recall that the conventional local polynomial estimation (Fan & Gijbels, 1996) is conducted under the quadratic loss, and the local-likelihood estimation (Tibshirani & Hastie, 1987) is conducted under the deviance loss for exponential family responses. In contrast, the loss function in the current article is the broader class of Bregman divergence.

5.1. Univariate Nonparametric Regression

To facilitate presentations, we first assume that $\mathtt{X}$ is univariate. Suppose that the function $\eta(\cdot) = F(m(\cdot))$ has a $(p+1)$th continuous derivative at a fitting point x. Let $\beta_j(x) = \eta^{(j)}(x)/j!$, $j = 0, 1, \ldots, p$. For X_i close to x, the Taylor expansion implies that

$$\eta(X_i) \doteq \beta_0(x) + (X_i - x)\beta_1(x) + \cdots + (X_i - x)^p\beta_p(x) \equiv \mathbf{x}_i(x)^{\mathrm{T}}\boldsymbol{\beta}(x),$$

in which $\mathbf{x}_i(x) = (1, (X_i - x), \ldots, (X_i - x)^p)^{\mathrm{T}}$ and $\boldsymbol{\beta}(x) = (\beta_0(x), \ldots, \beta_p(x))^{\mathrm{T}}$. Based on the independent training data $\{(X_i, Y_i)_{i=1}^n\}$, the vector of local parameters $\boldsymbol{\beta}(x)$ can be estimated by the local minimum BD nonparametric estimator $\widehat{\boldsymbol{\beta}}(x) = (\widehat{\beta}_0(x), \ldots, \widehat{\beta}_p(x))^{\mathrm{T}}$ which minimizes

$$\ell_n(\boldsymbol{\beta}; x) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, F^{-1}(\mathbf{x}_i(x)^{\mathrm{T}}\boldsymbol{\beta}))K_h(X_i - x), \tag{22}$$

with respect to $\boldsymbol{\beta}$, in which $K_h(\cdot) = K(\cdot/h)/h$ is re-scaled from a kernel function K and $h > 0$ is termed a bandwidth parameter. Then, the local BD estimates of $\eta(x)$ and $m(x)$ are given by $\widehat{\eta}(x) = \widehat{\beta}_0(x)$ and $\widehat{m}(x) = F^{-1}(\widehat{\eta}(x))$, respectively. We now define $S = (\mu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\mu_k = \int t^k K(t)\,\mathrm{d}t$ and $S^* = (\nu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\nu_k = \int t^k K^2(t)\,\mathrm{d}t$. The asymptotic distribution of the local BD estimator is delivered in Theorem 7.

Theorem 7. *Define $\eta(x) = F(m(x))$ and $H = \mathrm{diag}\{1, h, \ldots, h^p\}$. Assume that the degree p is odd. Suppose that Q is a BD. Suppose that Condition* **E** *in Appendix holds. If $n \to \infty$, $nh \to \infty$ and $nh^{2p+3} = O(1)$, then for the local minimum BD estimator $\widehat{\boldsymbol{\beta}}(x)$ at an interior point x of the design density, we have that*

$$\sqrt{nh}\left[H\{\widehat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x)\} - S^{-1}\boldsymbol{c}_p \frac{\eta^{(p+1)}(x)}{(p+1)!}h^{p+1}\right] \xrightarrow{\mathcal{L}} N(\mathbf{0}, S^{-1}S^*S^{-1}v(x)/f_X(x)),$$

where $\boldsymbol{c}_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}$ and $v(x) = \mathrm{var}(Y|X = x)\{F'(m(x))\}^2$.

Theorem 7 has the following useful consequences: The asymptotic distributions of the local BD estimator $\widehat{\boldsymbol{\beta}}(x)$, and hence $\widehat{\eta}(x)$ and $\widehat{m}(x)$, do not depend on either the choice of the loss function Q, or the distributional assumption of Y, but rely on the choice of the link function F. The adaptation of the local BD estimator, with one-dimensional predictors, to the choice of the loss function is an interesting result that has not been seen in the literature. Thus, Theorem 7 indeed gains new insight into nonparametric function estimation.

From a function estimation perspective, Theorem 7 enables us to derive the asymptotically optimal bandwidth by minimizing certain criterion. If the criterion is $\mathrm{AMISE}_{\widehat{\eta}}(h)$, the asymptotic mean integrated squared error of $\widehat{\eta}(\cdot)$, then the minimizer of $\mathrm{AMISE}_{\widehat{\eta}}(h)$ is

$$h_{\mathrm{AMISE}}(\widehat{\eta}) = C_p(K)\left[\frac{\int \mathrm{var}(Y|X = x)\{F'(m(x))\}^2 w(x)/f_X(x)\,\mathrm{d}x}{\int \{\eta^{(p+1)}(x)\}^2 w(x)\,\mathrm{d}x}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \quad (23)$$

where $w(\cdot) \geq 0$ is a weight function, and $C_p(K)$ is a constant depending only on the degree and kernel of the local regression. In practice, the data-driven optimal bandwidth can be selected via cross-validation.

From a classification point of view, since a classifier only depends on the sign of $F^{-1}(\widehat{\eta}(x))$, the classification performance using the margin-based loss functions, such as the quadratic loss, log-likelihood loss and the exponential loss, are expected to be similar. In the context of boosting, Bühlmann & Yu (2003) found comparable performances between L_2Boost and LogitBoost. We will revisit this issue in more detail in Section 6.

5.2. Varying-Coefficient Regression Model

This section extends the techniques of Section 5.1 to a useful class of multi-predictor models. Consider multivariate predictor variables, consisting of a scalar U and a vector $\mathrm{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$. For the response variable Y, define by $m(u, \mathrm{x}) = E(Y|U = u, \mathrm{X} = \mathrm{x})$ the conditional mean regression function, where $\mathrm{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$. The varying-coefficient model assumes that

$$F(m(U, \mathrm{X})) = \eta(U, \mathrm{X}) = \sum_{j=1}^{d} a_j(U)X_j = \mathrm{X}^{\mathrm{T}}\boldsymbol{A}(U), \quad (24)$$

for a vector $\boldsymbol{A}(u) = (a_1(u), \ldots, a_d(u))^{\mathrm{T}}$ of unknown smooth coefficient functions.

We first describe the local minimum BD estimation of $\boldsymbol{A}(u)$, based on the independent observations $\{(U_i, \mathrm{X}_i, Y_i)_{i=1}^n\}$. Assume that $a_j(\cdot)$'s are $(p+1)$-times continuously differentiable at a fitting point u. Put $\boldsymbol{A}^{(\ell)}(u) = (a_1^{(\ell)}(u), \ldots, a_d^{(\ell)}(u))^{\mathrm{T}}$. Denote by $\boldsymbol{\beta}(u) = (\boldsymbol{A}(u)^{\mathrm{T}}, \ldots, \boldsymbol{A}^{(p)}(u)^{\mathrm{T}}/p!)^{\mathrm{T}}$ the $d(p+1)$ by 1 vector of coefficient functions along with their derivatives, $\mathbf{u}_i(u) = (1, (U_i - u), \ldots, (U_i - u)^p)^{\mathrm{T}}$, and $\mathbf{I}_d$ a $d \times d$ identity matrix. For observed covariates U_i close to the point u,

$$\boldsymbol{A}(U_i) \doteq \boldsymbol{A}(u) + (U_i - u)\boldsymbol{A}^{(1)}(u) + \cdots + (U_i - u)^p\boldsymbol{A}^{(p)}(u)/p! = \{\mathbf{u}_i(u) \otimes \mathbf{I}_d\}^{\mathrm{T}}\boldsymbol{\beta}(u),$$

in which the symbol $\otimes$ denotes the Kronecker product, and thus, $\eta(U_i, \mathrm{X}_i) \doteq \{\mathbf{u}_i(u) \otimes \mathrm{X}_i\}^{\mathrm{T}} \boldsymbol{\beta}(u)$. The local minimum BD estimator $\widehat{\boldsymbol{\beta}}(u)$ minimizes the criterion function,

$$\ell_n(\boldsymbol{\beta}; u) = \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, F^{-1}(\{\mathbf{u}_i(u) \otimes \mathrm{X}_i\}^{\mathrm{T}} \boldsymbol{\beta})) K_h(U_i - u).$$

The first d entries of $\widehat{\boldsymbol{\beta}}(u)$ supply the local minimum BD estimates $\widehat{\boldsymbol{A}}(u)$ of $\boldsymbol{A}(u)$, and the local minimum BD estimates of $\eta(u, \mathrm{x})$ and $m(u, \mathrm{x})$ are given by $\widehat{\eta}(u, \mathrm{x}) = \mathrm{x}^{\mathrm{T}} \widehat{\boldsymbol{A}}(u)$ and $\widehat{m}(u, \mathrm{x}) = F^{-1}(\widehat{\eta}(u, \mathrm{x}))$, respectively. Theorem 8 establishes the limiting distribution of $\widehat{\boldsymbol{\beta}}(u)$.

Theorem 8. *Define $H = \mathrm{diag}\{1, h, \ldots, h^p\}$ and $\mathbf{H} = H \otimes \mathbf{I}_d$. Assume that the degree p is odd. Assume that Q is a BD. Suppose that Condition $\mathbf{E}'$ in Appendix holds. If $n \to \infty$, $nh \to \infty$ and $nh^{2p+3} = O(1)$, then for the local minimum BD estimator $\widehat{\boldsymbol{\beta}}(u)$ at an interior point u of the design density, we have that*

$$\sqrt{nh} \left[\mathbf{H}\{\widehat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u)\} - \left\{ S^{-1} \boldsymbol{c}_p \otimes \frac{\boldsymbol{A}^{(p+1)}(u)}{(p+1)!} \right\} h^{p+1} \right]$$
$$\xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \left[S^{-1} S^* S^{-1} \otimes \{\Gamma(u)^{-1} \Delta(u) \Gamma(u)^{-1}\} \right] / f_U(u) \right),$$

where $\Delta(u) = E[\mathrm{var}(Y|U = u, X)\{q''(m(u, X))\}^2 \{F'(m(u, X))\}^{-2} XX^T | U = u]$ *and* $\Gamma(u) = -E[q''(m(u, X))\{F'(m(u, X))\}^{-2} XX^T | U = u]$.

Apparently, the asymptotic distribution of the local BD estimator $\widehat{\boldsymbol{\beta}}(u)$ relies on the q-function. Does this dependence contradict the previous Theorem 7 which shows the lack of dependence on q? Here we add an explanation. It is easily seen that when the dimension of the predictor X is $d = 1$ and $X_1 \equiv 1$, varying-coefficient models reduce to the particular case of univariate nonparametric regression models. As a result, the q'' term in $\Gamma(u)^{-1} \Delta(u) \Gamma(u)^{-1}$ is cancelled. Therefore, no contradiction arises from Theorems 7 and 8. Furthermore, similar to Theorem 6, if $q''(m(u, \mathrm{x})) = -c/\mathrm{var}(Y|U = u, \mathrm{X} = \mathrm{x})$ for a constant $c > 0$, then $\Gamma(u)^{-1} \Delta(u) \Gamma(u)^{-1}$ in the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}(u)$ in Theorem 8 achieves the lower bound

$$\left(E\left[1/\mathrm{var}(Y|U = u, \mathrm{X})\{F'(m(u, \mathrm{X}))\}^{-2} \mathrm{X}\mathrm{X}^{\mathrm{T}} | U = u \right] \right)^{-1}.$$

6. APPLICATIONS TO CLASSIFICATION

In this section, we focus on the binary response $Y \in \{0, 1\}$, for which $m(\mathrm{x}) = P(Y = 1|\mathrm{X} = \mathrm{x})$. Classification aims to produce a classification rule, $\widehat{Y}(\mathrm{x}) \in \{0, 1\}$, for the class label Y at every input point x of X. The optimal rule is to minimize the misclassification risk $E\{L(Y, \widehat{Y}(\mathrm{X}))\}$ with the loss function $L(Y_1, Y_2) = |Y_1 - c|\, \mathrm{I}[Y_1 \neq Y_2]$ for $Y_1 \in \{0, 1\}$, $Y_2 \in \{0, 1\}$ and $c \in (0, 1)$.

6.1. Function Estimation Error and Classification Error

For $c = 1/2$ representing equal misclassification costs, it is well known that the optimal classifier is the Bayes rule $Y_B(\mathrm{x}) = \mathrm{I}[m(\mathrm{x}) > 1/2]$. Since the true class probabilities $m(\mathrm{x})$ are usually unknown, probability estimates $\widehat{m}(\mathrm{x})$ via function estimation procedures can be used to form a classification rule, that is, $\widehat{Y}(\mathrm{x}) = \mathrm{I}[\widehat{m}(\mathrm{x}) > 1/2]$. In this case, Friedman (1997) studied the way in which function estimation error of $\widehat{m}(\mathrm{x})$ affects the misclassification rate, and illustrated with the naive Bayes estimator and nearest-neighbour estimator.

We now investigate the extent to which function estimation error of the minimum BD estimators $\widehat{m}(\mathrm{x})$ affects the classification error, in the nonstandard situation with unequal misclassification costs. It is easy to show that the Bayes classifier $Y_B(\mathrm{x}) = \mathrm{I}[m(\mathrm{x}) > c]$ minimizes the misclassification risk $E\{L(Y, \widehat{Y}(\mathrm{X}))\}$ with respect to $\widehat{Y}$. We now study the classification performance of $\widehat{Y}(\mathrm{X}) = \mathrm{I}[\widehat{m}(\mathrm{X}) > c]$, which substitutes the true function in the Bayes rule by the function estimate. The corresponding misclassification loss is $L(Y, \widehat{Y}(\mathrm{X}))$, which agrees with $Q(Y, \widehat{m}(\mathrm{X}))$ in (17) associated with q in (18). Combining (6) and (14), we deduce $r_B(\mathrm{x}) = \min[m(\mathrm{x})(1-c), \{1-m(\mathrm{x})\}c]$. Applying (19) to (7) yields $Q(m(\mathrm{x}), \widehat{m}(\mathrm{x})) = |m(\mathrm{x}) - c|\mathrm{I}\{Y_B(\mathrm{x}) \neq \mathrm{I}[\widehat{m}(\mathrm{x}) > c]\}$. Thus the function estimation error of $\widehat{m}(\mathrm{x})$ in (8) is $E\{Q(m(\mathrm{x}), \widehat{m}(\mathrm{x}))\} = |m(\mathrm{x}) - c|P\{Y_B(\mathrm{x}) \neq \mathrm{I}[\widehat{m}(\mathrm{x}) > c]\}$, in which

$$P\{Y_B(\mathrm{x}) \neq \mathrm{I}[\widehat{m}(\mathrm{x}) > c]\} = \mathrm{I}[m(\mathrm{x}) \leq c]P\{\widehat{m}(\mathrm{x}) > c\} + \mathrm{I}[m(\mathrm{x}) > c]P\{\widehat{m}(\mathrm{x}) \leq c\}. \quad (25)$$

According to Theorems 5, 7, and 8, the minimum BD estimator $\widehat{m}(\mathrm{x})$ has an asymptotically normal distribution. Thus, by using the idea of normal approximation similar to that of Friedman (1997), an asymptotic approximation of the probability in (25) is provided by

$$\Phi\left(\mathrm{sign}\{c - m(\mathrm{x})\}\mathrm{sign}[E\{\widehat{m}(\mathrm{x})\} - c]\frac{|E\{\widehat{m}(\mathrm{x})\} - c|}{\sqrt{\mathrm{var}\{\widehat{m}(\mathrm{x})\}}}\right),$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution. Thus, when $m(\mathrm{x})$ and the aggregated predictor $E\{\widehat{m}(\mathrm{x})\}$ are on the same side of the classification boundary $\{\mathrm{x} : m(\mathrm{x}) = c\}$, the misclassification risk will decrease as $E\{\widehat{m}(\mathrm{x})\}$ departs farther away from c irrespective of the function estimation bias $m(\mathrm{x}) - E\{\widehat{m}(\mathrm{x})\}$; when $m(\mathrm{x})$ and $E\{\widehat{m}(\mathrm{x})\}$ are on opposite sides of the classification boundary, the misclassification risk will increase with the distance between $E\{\widehat{m}(\mathrm{x})\}$ and c.

6.2. Relation Between Margin-Based Loss Function and BD

Intuitively, the misclassification loss (17) should be used as the training loss, since it is the loss function used to evaluate the performances of classifiers. However, this function is neither convex nor continuous in μ, and causes problems for computation. Therefore many margin-based loss functions are used as training loss functions in many classification procedures (Shen et al., 2003).

The margin-based loss function is expressed in the form,

$$V(Y^*F(\mu)),$$

where $Y^* = 2Y - 1 \in \{-1, +1\}$ and Y^*F is called the "margin" with F playing the role similar to that in Section 5. Margin-based loss functions have been motivated as being upper bounds of the misclassification loss and have been widely used in the machine learning literature. One important application of margin-based loss functions, such as the exponential loss function (Freund & Schapire, 1997; Friedman, Hastie & Tibshirani, 2000), is to show the convergence rate of boosting procedures (Schapire, 2002) and in turn indirectly bound the misclassification error.

We are interested in studying the role of margin-based loss functions in classification. We will show that most of the commonly used margin-based loss functions are BD. Thus the results on the prediction error conveyed by Theorem 2 makes the comparison much easier. For a given loss function $L(Y, \mu)$, we first illustrate how to represent it by a margin-based loss function $V(Y^*F(\mu))$.

Lemma 1. *Suppose that $L(Y, \mu)$ is a loss function for a binary variable Y. Assume that Conditions* **F** *and* **G** *below hold:*

F. $L(0, \mu) = L(1, 1 - \mu)$;

G. $F(\mu)$ *is monotone increasing, satisfying* $F(1 - \mu) = -F(\mu)$, *and* $F^{-1}(s)$ *is right-continuous* (*or equivalently, continuous*), *where* $F^{-1}(s) = \inf\{u : F(u) \geq s\}$.

Define $V(s) = L(1, F^{-1}(s))$. *Then* $L(Y, \mu) = V(Y^*F(\mu))$.

Lemma 1 actually supplies a simple method to represent L in the form of $V(Y^*F)$. Condition **F** holds in the equal-cost misclassification loss and all other loss functions in Section 3.3. The choice of F satisfying Condition **G** is flexible and certainly not unique. However, since L is not necessarily a BD, its alternative form $V(Y^*F)$ in Lemma 1 is not necessarily either. Recall that we have shown in Section 3.3 the misclassification loss is BD. Theorem 9 asserts that under very mild conditions on V and F, the (centralized) margin-based loss function is indeed a BD.

Theorem 9. *Suppose that* $V(Y^*F)$ *is a margin-based loss function for a binary variable* Y. *Assume that Conditions* **H** *and* **I** *below hold:*

H. *There exists* F_B *such that* $(\partial/\partial\mu)V(Y^*F_B(\mu))$ *is continuous in* $\mu \in (0, 1)$, *and*

$$F_B'(\mu) \geq 0; \tag{26}$$

I. *For* $\mu \in (0, 1)$, $V'(F_B(\mu)) \leq 0$ *and*

$$\mu V'(F_B(\mu)) = (1 - \mu)V'(-F_B(\mu)). \tag{27}$$

Then the centralized form of $V(Y^*F_B(\mu))$, *that is,*

$$Q(Y, \mu) = V(Y^*F_B(\mu)) - V(Y^*F_B(Y)), \tag{28}$$

is a BD, for which the generating q-function is given by

$$q(\mu) = \mu V(F_B(\mu)) + (1 - \mu)V(-F_B(\mu)), \quad \mu \in [0, 1]. \tag{29}$$

The applicability of Theorem 9 relies on finding F_B. In the case of convex V, Lemma 2 draws connections between F_B and the Bayes rule, whereas Lemma 3 indicates that Condition **H** is also necessary in certain situations.

Lemma 2. *Assume that* $V(s)$ *is continuous and convex. If* $V'(s)$ *exists at* $s \in \mathcal{F}_2$, *then the existence of* $F_B(m(x))$ *such that* $\pm F_B \in \mathcal{F}_2$ *and*

$$F_B(m(x)) = \arg\min_{F \in \mathcal{F}_2} E\{V(Y^*F)|X = x\} \tag{30}$$

implies (27).

Lemma 3. *Assume* $V''(s) \geq 0$ *for all* $s \in \mathcal{F}_2$, *and* F_B *satisfies Condition* **I**. *Then (26) must hold.*

Below, we illustrate that five margin-based loss functions are BD.

Example 1. $V(s) = (1 - s)^2$ for the quadratic loss. We have that $V'(s) = -2(1 - s)$. From (27), we have $F_B(\mu) = 2\mu - 1$. Thus Conditions **H** and **I** are satisfied and $V(Y^*F_B(\mu)) = Y(2 - 2\mu)^2 + (1 - Y)(2\mu)^2$. By (28)–(29), $Q(Y, \mu) = 4(Y - \mu)^2$, and $q(\mu) = 4\mu(1 - \mu)$.

Example 2. $V(s) = (1 - s)^5$ for the arching loss given in Breiman (1998). We have that $V'(s) = -5(1 - s)^4 \leq 0$. From (27), we have $F_B(\mu) = \{\mu^{1/4} - (1 - \mu)^{1/4}\}/\{\mu^{1/4} + (1 - \mu)^{1/4}\}$. Thus

Conditions **H** and **I** are satisfied and $V(Y^*F_B(\mu)) = Y[\{2(1-\mu)^{1/4}\}/\{\mu^{1/4}+(1-\mu)^{1/4}\}]^5 + (1-Y)[\{2\mu^{1/4}\}/\{\mu^{1/4}+(1-\mu)^{1/4}\}]^5$. By (28)–(29), $Q(Y,\mu) = V(Y^*F_B(\mu))$ and $q(\mu) = 2^5\mu(1-\mu)/\{\mu^{1/4}+(1-\mu)^{1/4}\}^4$.

Example 3. $V(s) = \ln(1+\mathrm{e}^{-s})$ for negative log-likelihood. We have that $V'(s) = -1/(1+\mathrm{e}^s) \le 0$. From (27), we have $F_B(\mu) = \ln\{\mu/(1-\mu)\}$. Thus Conditions **H** and **I** are satisfied and $V(Y^*F_B(\mu)) = -\{Y\ln(\mu) + (1-Y)\ln(1-\mu)\}$. By (28)–(29), $Q(Y,\mu) = V(Y^*F_B(\mu))$ and $q(\mu) = -\{\mu\ln(\mu) + (1-\mu)\ln(1-\mu)\}$.

Example 4. $V(s) = \mathrm{e}^{-s}$ for the exponential loss. We have that $V'(s) = -\mathrm{e}^{-s} \le 0$. From (27), we have $F_B(\mu) = 2^{-1}\ln\{\mu/(1-\mu)\}$. Thus Conditions **H** and **I** are satisfied and $V(Y^*F_B(\mu)) = Y\{(1-\mu)/\mu\}^{1/2} + (1-Y)\{\mu/(1-\mu)\}^{1/2}$. By (28)–(29), $Q(Y,\mu) = V(Y^*F_B(\mu))$ and $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$.

Example 5. $V(s) = \max(1-s,0) = (1-s)\,\mathrm{I}[s \le 1]$ for the hinge loss used in support vector machine. We have that V is continuous, convex and $V'(s) = -\mathrm{I}[s<1]$ for $s \in \mathcal{F}_2 = \{s : s \ne 1\}$. By a graphical approach, we observe that the desired F_B to minimize $E\{V(Y^*F)|\mathrm{X} = \mathrm{x}\}$ is $F_B(\mu) = \mathrm{sign}(2\mu-1)$. Thus $V(Y^*F_B(\mu)) = 2\{Y\mathrm{I}[2\mu<1] + (1-Y)\mathrm{I}[2\mu>1]\}$. Define $Q(Y,\mu) = V(Y^*F_B(\mu))$. Since Condition **I** does not hold, we could not use Theorem 9. Nonetheless, we can directly verify Condition **B** in Theorem 3 and apply Corollary 2 to obtain $q(\mu) = 2\{\mu\mathrm{I}[2\mu<1] + (1-\mu)\mathrm{I}[2\mu>1]\}$.

Theorem 7 implies that for the margin-based loss functions $V(Y^*F_B(\mu))$, the asymptotic distribution of (non-regularized) $\widehat{m}(x)$ will be the same. Thus, the classification performance will be similar. Again, the corresponding prediction error can be assessed by the application of Theorem 2. This is particularly useful for the assessment when the true $m(\mathrm{x})$ is known.

7. SIMULATIONS

7.1. Impact of BD on Parametric Regression and Classification

To evaluate the impact of loss functions on parametric regression and classification, we conduct the simulation study. We generate data with two-classes from the following model,

$$\mathrm{X} = (X_1,\ldots,X_6)^{\mathrm{T}} \sim N(\mathbf{0},\mathbf{I}_6), \quad Y|\mathrm{X} = \mathrm{x} \sim \text{Bernoulli}\{m(\mathrm{x})\},$$

where

$$F(m(\mathrm{X})) = \ln\left\{\frac{m(\mathrm{X})}{1-m(\mathrm{X})}\right\} = \alpha_0 + \sum_{j=1}^{6}\alpha_j X_j, \tag{31}$$

with true values of the parameters $\alpha_j = 3\times(-1)^j$, $j = 0,\ldots,6$.

First, we generate 500 sets of random samples $\{(\mathrm{X}_i, Y_i)\}_{i=1}^n$ of size $n = 500$ from the distribution of (X, Y). The estimates $\widehat{\alpha}_j$, $j = 0,\ldots,6$, are numerically obtained to minimize the criterion function (20). Figure 2 compares the boxplots of $\widehat{\alpha}_j - \alpha_j$, $j = 0,\ldots,6$, using the deviance loss and the exponential loss. It is clearly seen that the regression estimates under the deviance loss are more centered around the true values with smaller variabilities than those under the exponential loss. This lends support to Theorem 6 which reveals that regression estimates under the deviance loss achieves the asymptotically lower bound for the covariance matrix of the estimators. From the regression point of view, the deviance loss does exhibit superiority over the exponential loss in the finite-sample cases.

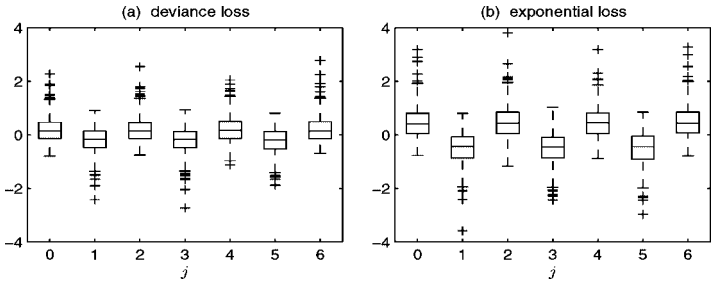

FIGURE 2: Boxplots of $\widehat{\alpha}_j - \alpha_j$, $j = 0, \ldots, 6$ (from left to right in each panel). Panel (a): using the deviance loss; panel (b): using the exponential loss.

TABLE 1: Comparing test misclassification rates using deviance loss and exponential loss.

Loss	Test misclassification rates									
Deviance	0.064	0.079	0.078	0.080	0.067	0.071	0.071	0.061	0.073	0.071
Exponential	0.068	0.075	0.079	0.078	0.066	0.074	0.072	0.060	0.070	0.079

Next, we examine the behaviours of classification procedures constructed under different loss functions in the two-class classification. One single training set of size 500 is used for estimating parameters α_j, $j = 0, \ldots, 6$. Test samples are randomly generated from model (31) of size 1,000. A comparison of the test misclassification rates in 10 sets of test samples is listed in Table 1. The results indicate that the difference from the deviance and exponential loss functions in regression estimates has a negligible impact on the classification performance.

7.2. Penalized BD: Application to Model Building and Selection

Model selection methods, such as Lasso (Tibshirani, 1996), choose the model parameters by minimizing the sum of a quadratic loss plus a penalty on the parameters. To illustrate the application of BD as a versatile and useful tool for statistical model building and selection, we consider the penalized estimator $(\widehat{\alpha}_0, \widehat{\boldsymbol{\alpha}})$ by minimizing the "penalized BD,"

$$\frac{1}{n}\sum_{i=1}^{n} Q(Y_i, F^{-1}(\alpha_0 + \mathrm{X}_i^{\mathrm{T}}\boldsymbol{\alpha})) + \sum_{j=1}^{d} P_{\lambda_n}(|\alpha_j|),$$

where Q is a BD, F is a link and $P_{\lambda_n}(\cdot)$ is a penalty function indexed by a tuning parameter $\lambda_n > 0$. The Lasso uses the L_1-penalty, $P_\lambda(|x|) = \lambda|x|$. Similar to the simulation study in Section 7.1, we generate random samples from the following model,

$$\mathrm{X} \sim N(\mathbf{0}, \mathbf{I}_8), \quad Y|\mathrm{X} = \mathrm{x} \sim \text{Bernoulli}\{m(\mathrm{x})\}, \tag{32}$$

where $F(m(\mathrm{x})) = \text{logit}\{m(\mathrm{x})\} = \alpha_0 + \mathrm{x}^{\mathrm{T}}\boldsymbol{\alpha}$, with $\alpha_0 = 3$ and $\boldsymbol{\alpha} = (0, 0, 1.5, 0, 0, 2, 0, 0)^{\mathrm{T}}$. The penalized estimates are numerically obtained from modifying the LARS algorithm.

First, to examine the effect of penalized and non-penalized regression estimates under BD on model fitting, we generate 100 training sets of size 200. For each training set, the model error (ME) is calculated by $\sum_{l=1}^{L}\{\widehat{m}(\mathrm{x}_l) - m(\mathrm{x}_l)\}^2/L$ at a sequence $\{\mathrm{x}_l\}_{l=1}^{L=5,000}$ randomly generated from (32), and the relative model error (RME) is the ratio of ME using penalized estimators and ME using non-penalized estimators. Table 2 tabulates MRME, the median of RMEs obtained from

TABLE 2: Simulation results from penalized BD estimates.

Penalty	Loss	Regression MRME	Variable selection		Classification MAMR
			Correct zeros	Incorrect zeros	
L_1	Deviance	0.7636	2.38	0.00	0.1180
	Exponential	0.6849	2.31	0.00	0.1181

100 training sets. Evidently, if the true model has sparse coefficients, the penalized estimators reduce the function estimation error compared with the non-penalized estimators, under both the deviance and exponential loss functions.

Next, to study the utility of penalized estimators in revealing the effects in variable selection under different loss functions, Table 2 lists a column labelled "Correct zeros" as the average number of parameters which are correctly estimated to be zero when the true parameters are zero, and a column labelled "Incorrect zeros" as the average number of parameters which are erroneously estimated to be zero when the true parameters are nonzero. Overall, using either the deviance or exponential loss, penalization helps yield a sparse solution and build a sparse model.

Last, to investigate the performance of classification rules using penalized estimators, we evaluate the average misclassification rate (AMR) for 10 independent test sets of size 10,000. Table 2 reports MAMR, the median of AMRs calculated from 100 training sets. The results indicate that the classification rules constructed from penalized estimators under both exponential and deviance loss perform as well as the Bayes optimal rule (whose MAMR is 11.06%).

7.3. Impact Of BD on Nonparametric Logistic Regression

We generate independent observation pairs $\{(X_i, Y_i)_{i=1}^{400}\}$ according to $X \sim U(0, 1)$ and $Y|X = x \sim \text{Bernoulli}\{m(x)\}$, where $\eta(x) = F(m(x)) = \text{logit}\{m(x)\} = 7[\exp\{-(4x-1)^2\} + \exp\{-(4x-3)^2\}] - 5.5$. To assess the sampling variability of $\widehat{\eta}(x)$ obtained from (22) using local estimation method (with degree $p = 1$ and the Epanechnikov kernel), the AMISE-optimal bandwidth $h = 0.108$ in (23) is used, where $w(\cdot) = f_X(\cdot)$. Figure 3 compares the estimates associated with the deviance loss and exponential loss. Panels (a)–(b) present the estimated curves $\widehat{\eta}(\cdot)$ from three sets of random samples. For illustrative simplicity, panels (c)–(d) give the boxplots of $\widehat{\eta}(x)$ at points $x = 0.1, 0.3, 0.5, 0.7, 0.9$, based on 100 random samples. Clearly, the estimates, as well as their sampling variations, are nearly indistinguishable under the deviance loss and exponential loss. (The results for other choices of h are similar and thus omitted.) This is consistent with the asymptotic theory in Theorem 7.

APPENDIX

We first impose some technical conditions, which may not be the weakest possible.

Condition C:

C1. Θ is compact in $\mathbb{R}^{d+1}$.
C2. X is on a compact support Λ, and X has the design density $f_{\mathrm{X}}(\cdot)$ with $f_{\mathrm{X}}(\cdot) > 0$.
C3. $F(\cdot)$ is a bijection, with $F^{-1}(\cdot)$ continuously differentiable.
C4. q is twice continuously differentiable with $q''(\cdot) < 0$.

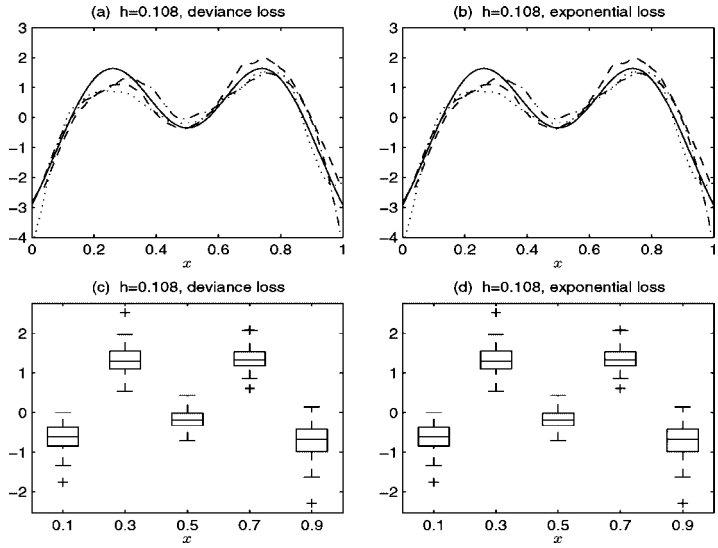

FIGURE 3: Comparison of local BD estimates $\widehat{\eta}(\cdot)$ using the deviance and exponential losses. Panels (a)–(b): solid curve is the true curve $\eta(\cdot)$; the dotted, dashed and dash-dotted curves are $\widehat{\eta}(\cdot)$ from three random samples. Panels (c)–(d): boxplots of $\widehat{\eta}(x)$ from 100 random samples at points $x = 0.1, 0.3, 0.5, 0.7, 0.9$.

C5. There does not exist a vector $\widetilde{\boldsymbol{c}} \neq \mathbf{0}$ such that $\underset{\sim}{\mathrm{X}}^{\mathrm{T}} \widetilde{\boldsymbol{c}} = 0$ almost surely.
C6. $E(|Y|) < \infty$.

Condition D: D1, D2, and D5 are identical to C1, C2, and C5 respectively, and

D3. $F(\cdot)$ is a bijection, with $F^{-1}(\cdot)$ twice continuously differentiable.
D4. $q'''(\cdot)$ is continuous and $q''(\cdot) < 0$.
D6. $E(Y^2) < \infty$.

Condition E:

E1. q is concave, $q''(m(x)) < 0$ and $q^{(4)}(\cdot)$ is continuous in a neighbourhood of $m(x)$.
E2. Let $\mathrm{q}_j(y;\theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$. Assume that $\mathrm{q}_2(y;\theta) > 0$ for $\theta \in \mathbb{R}$ and y in the range of the response variable.
E3. There exists some $\delta > 0$ such that $E(|Y|^{2+\delta}|X = \cdot)$ is bounded in a neighbourhood of x.
E4. The kernel function K is a symmetric probability density function with bounded support.
E5. X has the design density $f_X(\cdot)$ which is continuous in a neighbourhood of x, and $f_X(x) > 0$.
E6. Both $m(\cdot)$ and $\mathrm{var}(Y|X = \cdot)$ are continuous in a neighbourhood of x, and $\mathrm{var}(Y|X = x) > 0$.
E7. $F(\cdot)$ is a bijection. $F'(m(x)) > 0$ and $F^{(3)}(\cdot)$ is continuous in a neighbourhood of $m(x)$. $F^{-1}(\cdot)$ is continuous in a neighbourhood of $\eta(x)$.
E8. $\eta^{(p+1)}(\cdot)$ is continuous in a neighbourhood of x.

Condition E′: E2′ is identical to E2, and

E1′. q is concave, $q''(m(u, \mathrm{x})) < 0$ and $q^{(4)}(\cdot)$ is continuous in a neighbourhood of $m(u, \mathrm{x})$.
E3′. There exists some $\delta > 0$ such that $E(|Y|^{2+\delta}|U = \cdot, \mathrm{X} = \mathrm{x})$ is bounded in a neighbourhood of u, for a.e. x.
E4′. The kernel function K is a symmetric probability density function with bounded support.
E5′. U has the design density $f_U(\cdot)$ which is continuous in a neighbourhood of u, and $f_U(u) > 0$.

E6′. Both $m(\cdot, \mathrm{x})$ and $\mathrm{var}(Y|U = \cdot, \mathrm{X} = \mathrm{x})$ are continuous in a neighbourhood of u, and $\mathrm{var}(Y|U = u, \mathrm{X} = \mathrm{x}) > 0$ for a.e. x.

E7′. $F(\cdot)$ is a bijection. $F'(m(u, \mathrm{x})) > 0$ and $F^{(3)}(\cdot)$ is continuous in a neighbourhood of $m(u, \mathrm{x})$. $F^{-1}(\cdot)$ is continuous in a neighbourhood of $\eta(u, \mathrm{x})$.

E8′. $a_j^{(p+1)}(\cdot)$, $j = 1, \ldots, d$, are continuous in a neighbourhood of u.

E9′. $\Gamma(\cdot)$ is continuous at u and $\Gamma(u) > \mathbf{0}$. $\Delta(\cdot)$ is continuous in a neighbourhood of u and $\Delta(u) > \mathbf{0}$.

E10′. $E(\mathrm{X}\mathrm{X}^T|U = u) > \mathbf{0}$ for a.e. u.

We next introduce some necessary notations and definition.

Notations: Define $\mathrm{q}_j(y; \theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$. Let $\theta = F(\mu)$, which implies $\mu = F^{-1}(\theta)$ and $d\theta/d\mu = F'(\mu)$. Direct calculations via (3)–(4) give that

$$\mathrm{q}_1(y; \theta) = q''(\mu)(y - \mu)/F'(\mu),$$
$$\mathrm{q}_2(y; \theta) = \left[-q''(\mu)F'(\mu) + (y - \mu)\{q'''(\mu)F'(\mu) - q''(\mu)F''(\mu)\}\right]/\{F'(\mu)\}^3.$$

Accordingly, $\mathrm{q}_j(y; \theta)$ is linear in y for fixed θ.

Proof of Theorem 2. The argument is similar to that for Theorem 1 and is thus omitted. ■

Proof of Theorem 3. We first consider the case $K = 0$, that is, $\partial Q(Y, \mu)/\partial\mu$ is continuous for $\mu \in (a, b)$.

If Condition **A** holds, then Condition **B** will directly follow from (3). Conversely, assume that Condition **B** holds. Define a function q by

$$q(\mu) = \int_a^\mu \left\{ \int_a^t \frac{1}{Y - s} \frac{\partial Q(Y, s)}{\partial s} \, \mathrm{d}s \right\} \mathrm{d}t \quad \text{for } \mu \in [a, b]. \tag{33}$$

It follows that q is continuous on $[a, b]$, and

$$q'(\mu) = \int_a^\mu \frac{1}{Y - s} \frac{\partial Q(Y, s)}{\partial s} \, \mathrm{d}s \quad \text{for } \mu \in [a, b].$$

According to Condition **B**, the integrand above is free of Y and less than or equal to zero, so $q'(\mu)$ is monotone decreasing for $\mu \in [a, b]$. Thus the concavity of q is implied. Note that (10) in Condition **B** guarantees the use of Fubini theorem in (33), and thus for $Y \in [a, b]$ and $\mu \in [a, b]$,

$$q(\mu) = \int_{\mu_j}^\mu \frac{\mu - s}{Y - s} \frac{\partial Q(Y, s)}{\partial s} \, \mathrm{d}s \tag{34}$$
$$= \int_a^\mu \frac{\partial Q(Y, s)}{\partial s} \, \mathrm{d}s - (Y - \mu) \int_a^\mu \frac{1}{Y - s} \frac{\partial Q(Y, s)}{\partial s} \, \mathrm{d}s$$
$$= Q(Y, \mu) - Q(Y, a) - (Y - \mu)q'(\mu), \tag{35}$$

which also indicates, by replacing μ by Y in (35) and using $Q(Y, Y) = 0$, the identity $q(Y) = -Q(Y, a)$. This combined with (35) implies the result desired in Condition **A**.

For $K \geq 1$, that is, $\partial Q(Y, \mu)/\partial\mu$ has K points of discontinuity in $\mu \in (a, b)$, the proofs are similar to those used in the previous case $K = 0$. In particular, the q-function can be piecewisely defined on each interval as in (34) and made continuous by using the equivalent version of q in Remark 1. We omit the lengthy details. ■

Proof of Corollary 1. It follows from (9) that for $\mu \in \mathcal{N}_1$,

$$E\{Q(Y, \mu)|\mathtt{X} = \mathtt{x}\} = -E\{q(Y)|\mathtt{X} = \mathtt{x}\} + q(\mu) + \{m(\mathtt{x}) - \mu\}q'(\mu).$$

Note $E\{q(Y)|\mathtt{X} = \mathtt{x}\}$ is independent of μ. By the definition, $\mathcal{E}(\mu; m(\mathtt{x})) = q(\mu) + \{m(\mathtt{x}) - \mu\}q'(\mu)$ for $\mu \in \mathcal{N}_1$, which implies $q(\mu) = \mathcal{E}(\mu; \mu)$ for $\mu \in \mathcal{N}_1$. ■

Proof of Corollary 2. For a binary random variable Y, we see that $Q(Y, \mu) = YQ(1, \mu) + (1 - Y)Q(0, \mu)$ and thus $E\{Q(Y, \mu)|\mathtt{X} = \mathtt{x}\} = m(\mathtt{x})Q(1, \mu) + \{1 - m(\mathtt{x})\}Q(0, \mu)$, which depends on the conditional distribution of Y only through $m(\mathtt{x})$. An application of Corollary 1 yields $\mathcal{E}(\mu; m(\mathtt{x})) = m(\mathtt{x})Q(1, \mu) + \{1 - m(\mathtt{x})\}Q(0, \mu)$ and hence $\mathcal{E}(\mu; \mu) = \mu Q(1, \mu) + (1 - \mu)Q(0, \mu)$. The conclusion (13) follows from applying (12). From (13), it follows immediately that $q(Y) = Q(Y, Y)$ and thus $q(Y) \equiv 0$.

To show the second part of (14), we see from (9) that

$$Q(1, \mu) = -q(1) + q(\mu) + (1 - \mu)q'(\mu) \quad \text{for } \mu \in \mathcal{N}_1,$$
$$Q(0, \mu) = -q(0) + q(\mu) + (-\mu)q'(\mu) \quad \text{for } \mu \in \mathcal{N}_1,$$

which implies that $q'(\mu) = Q(1, \mu) - Q(0, \mu) + q(1) - q(0)$ for $\mu \in \mathcal{N}_1$, namely, (14), due to the fact that $q(Y) \equiv 0$. ■

Proofs of Theorem 4 and 5. The lengthy details can be found in Zhang, Jiang & Shang (2007). ■

Proof of Theorem 6. Before showing Theorem 6, we need the following Lemma 4.

Lemma 4. *For appropriately dimensioned random matrices A and B, if $E(BB^{\mathrm{T}})$ is positive definite, then $E(AA^{\mathrm{T}}) \geq E(AB^{\mathrm{T}})\{E(BB^{\mathrm{T}})\}^{-1}E(BA^{\mathrm{T}})$. Moreover, if $B = cA$ for a constant $c \neq 0$, then the inequality becomes an equality.*

Proof. Let $C = A - E(AB^{\mathrm{T}})\{E(BB^{\mathrm{T}})\}^{-1}B$. Then $E(CB^{\mathrm{T}}) = \mathbf{0}$. Thus

$$E(CC^{\mathrm{T}}) = E(AA^{\mathrm{T}}) - E(AB^{\mathrm{T}})\{E(BB^{\mathrm{T}})\}^{-1}E(BA^{\mathrm{T}}),$$

which yields the matrix inequality and equality. This completes the proof of Lemma 4. ■

To show Theorem 6, let $A = \{\mathrm{var}(Y|\mathtt{X})\}^{-1/2}\{F'(m(\mathtt{X}))\}^{-1}\widetilde{\mathtt{X}}$ and $B = -\mathrm{var}(Y|\mathtt{X})q''(m(\mathtt{X}))A$ be two random matrices. Then $\mathsf{H}_0 = E(AB^{\mathrm{T}}) = E(BA^{\mathrm{T}})$ and $\Omega_0 = E(BB^{\mathrm{T}})$. Employing Lemma 4,

$$\mathsf{H}_0^{-1}\Omega_0\mathsf{H}_0^{-1} \geq \{E(AA^{\mathrm{T}})\}^{-1} = \left(E[1/\mathrm{var}(Y|\mathtt{X})\{F'(m(\mathtt{X}))\}^{-2}\widetilde{\mathtt{X}}\widetilde{\mathtt{X}}^{\mathrm{T}}]\right)^{-1},$$

and $\geq$ is $=$ when $q''(m(\mathtt{x})) = -c/\mathrm{var}(Y|\mathtt{X} = \mathtt{x})$ for a constant $c > 0$. ■

Proof of Theorem 7. Note that a varying-coefficient model (24) with $\mathtt{X} = X_1 \equiv 1$ reduces to a univariate nonparametric regression model. Thus Theorem 7 can be deduced from Theorem 8. ■

Proof of Theorem 8. The lengthy details can be found in Zhang, Jiang & Shang (2007). ■

Proof of Lemma 1. For a binary Y, $L(Y, \mu) = YL(1, \mu) + (1 - Y)L(0, \mu)$. Then Condition **F** leads to $L(Y, \mu) = YL(1, \mu) + (1 - Y)L(1, 1 - \mu)$. Condition **G** combined with the continuity

assumption on F^{-1} implies that $1-\mu = F^{-1}(-F(\mu))$. Thus $L(Y,\mu) = YL(1, F^{-1}(F(\mu))) + (1-Y)L(1, F^{-1}(-F(\mu))) = L(1, F^{-1}(Y^*F(\mu)))$. ■

Proof of Theorem 9. We notice that for a generic F,

$$V(Y^*F) = YV(F) + (1-Y)V(-F), \tag{36}$$

$$\partial V(Y^*F)/\partial F = YV'(F) - (1-Y)V'(-F). \tag{37}$$

Combining Condition **I**, we have that for all μ, $\partial V(Y^*F_B(\mu))/\partial\mu = V'(F_B(\mu))F_B'(\mu)\frac{Y-\mu}{1-\mu}$, which along with Condition **H** implies that $V(Y^*F_B(\mu)) \geq V(Y^*F_B(Y))$. It follows that $Q(Y,\mu)$, as defined in (28), satisfies $Q(Y,Y) = 0$, $(\partial/\partial\mu)Q(Y,\mu)$ is continuous, and fulfills Condition **B**. By Theorem 3, (9) holds. An application of Corollary 2, particularly (13), gives the q-function in (29). ■

Proof of Lemma 2. Define $L(F) = E\{V(Y^*F)|\mathrm{X} = \mathrm{x}\}$. From (36), it follows that $L(F) = m(\mathrm{x})V(F) + \{1-m(\mathrm{x})\}V(-F)$ for $F \in \mathcal{F}_2$. Since V is convex, we observed that $L(F)$ is convex in F, and $L'(F) = m(\mathrm{x})V'(F) - \{1-m(\mathrm{x})\}V'(-F)$, for $\pm F \in \mathcal{F}_2$. Thus (30) implies $L'(F_B(m(\mathrm{x}))) = 0$ for any $m(\mathrm{x})$ such that $V'(F_B(m(\mathrm{x})))$ and $V'(-F_B(m(\mathrm{x})))$ exist, namely, (27). ■

Proof of Lemma 3. Taking the derivatives with respect to μ on both sides of (27) leads to

$$V'(F_B(\mu)) + \mu V''(F_B(\mu))F_B'(\mu) = -V'(-F_B(\mu)) - (1-\mu)V''(-F_B(\mu))F_B'(\mu),$$

that is, $\{\mu V''(F_B(\mu)) + (1-\mu)V''(-F_B(\mu))\}F_B'(\mu) = -V'(F_B(\mu)) - V'(-F_B(\mu))$. ■

ACKNOWLEDGEMENTS

The research is supported in part by National Science Foundation grants and Wisconsin Alumni Research Foundation. The authors are grateful to the Editor, Associate Editor, and two anonymous referees for insightful comments and suggestions.

BIBLIOGRAPHY

Y. Altun & A. Smola (2006). Unifying divergence minimization and statistical inference via convex duality. In *Proc. COLT2006*, Springer, Berlin/Heidelberg, pp. 139–153.

K. S. Azoury & M. K. Warmuth (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43, 211–246.

A. Banerjee, X. Guo & H. Wang (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51, 2664–2669.

L. M. Brègman (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7, 620–631.

L. Breiman (1998). Arching classifiers (with discussion). *Annals of Statistics*, 26, 801–824.

P. Bühlmann & B. Yu (2003). Boosting with the L_2 loss: regression and classification. *Journal of American Statistical Association*, 98, 324–339.

B. Efron (1986). How biased is the apparent error rate of a prediction rule? *Journal of American Statistical Association*, 81, 461–470.

B. Efron (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussion). *Journal of American Statistical Association*, 99, 619–642.

J. Fan & I. Gijbels (1996). *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

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

J. Friedman (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Journal of Data Mining and Knowledge Discovery*, 1, 55–77.

J. Friedman, T. Hastie & R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28, 337–407.

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

T. J. Hastie & R. J. Tibshirani, (1993). Varying-coefficient models (with discussion), *Journal of the Royal Statistical Society, Series B*, 55, 757–796.

T. Hastie, R. Tibshirani & J. Friedman (2001). *The Elements of Statistical Learning*, Springer, New York.

J. Kivinen & M. K. Warmuth (1999). Boosting as entropy projection. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Santa Cruz, CA. ACM Press, New York, NY, pp. 134–144.

J. D. Lafferty, S. Della Piestra & V. Della Piestra (1997). Statistical learning algorithms based on Bregman distances. In *Proceedings of the Canadian Workshop on Information Theory*.

J. Lafferty (1999). Additive models, boosting, and inference for generalized divergences. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Santa Cruz, CA. ACM Press, New York, NY, pp. 125–133.

X. Nguyen, M. J. Wainwright & M. I. Jordan (2008). On surrogate loss functions and *f*-divergences. *Annals of Statistics*, in press.

R. E. Schapire (2002). The boosting approach to machine learning: an overview. In *MSRI Workshop on Nonlinear Estimation and Classification*.

X. Shen, G. C. Tseng, X. Zhang & W. H. Wong (2003). On ψ-learning. *Journal of American Statistical Association*, 98, 724–734.

R. Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

R. Tibshirani & T. Hastie (1987). Local likelihood estimation. *Journal of American Statistical Association*, 82, 559–567.

R. W. M. Wedderburn (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439–447.

C. M. Zhang, Y. Jiang & Z. Shang (2007). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Technical Report #1127*, Department of Statistics, University of Wisconsin, Madison, WI.

Received 30 April 2007
Accepted 26 August 2008