Bregman divergence: fusing regression and classification with parametric and nonparametric estimation (preliminary report)

Zhang, C.M. and Jiang, Y.
Bregman Divergence: Fusing Regression and Classification with Parametric and Nonparametric Estimation

Chunming Zhang
Department of Statistics
University of Wisconsin-Madison
cmzhang@stat.wisc.edu

Yuan Jiang
Department of Statistics
University of Wisconsin-Madison
jiangy@stat.wisc.edu

Drafted on July 11, 2006

Abstract

In statistical learning, regression and classification concern different types of the output variable and the predictive accuracy is quantified by different error measures. Various types of loss functions have been proposed in classification rules and the machine learning literature. This paper presents a unified account of statistical properties of Bregman divergence, under which nearly all of the commonly used loss functions in regression and classification, associated with their prediction errors, can be studied simultaneously. The similarity of regression and classification formulated under the framework of Bregman divergence enables us to establish some new theoretical results demonstrating the role of parametric and nonparametric estimation played in the performance of classification procedures and related machine learning techniques.

Key words and Phrases: Asymptotic Normality; Bayes Optimal Rule; Consistency; Local Polynomial Regression; Loss Function; Prediction Error; Statistical Learning.
Short title: Bregman Divergence: Fusing Regression and Classification
1 Introduction

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

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

$$T = \{(x_i, y_i), i = 1, \ldots, n\},$$

the goal of a supervised learning algorithm is to use (1.1) to construct a prediction rule for a future output \(Y\) at the observed value of the input or feature variable \(X\). Depending on the nature of \(Y\), the predictive error is quantified by different error measures. For example, the quadratic loss has nice mathematical 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.

Owing to the nature of output variable in classification, the choice of loss functions plays an important role in defining and understanding the bias, variance and prediction error for the classification rule (Tibshirani, 1996). Bias plus variance decomposition for loss functions in classification is an active area of research, including Geman, Bienenstock, and Doursat (1992), Kohavi and Wolpert (1996), Friedman (1997), Wolpert (1997), Heskes (1998), Domingos (2000), James (2003), among many others.

In this paper, we present a unified account of statistical properties of the Bregman divergence (Bregman, 1967), under which nearly all of the commonly used loss functions in regression and classification, associated with their prediction errors, can be studied simultaneously. The similarity of regression and classification formulated under the framework of Bregman divergence enables us to establish some new theoretical results demonstrating the role of parametric and nonparametric estimation played in the performance of classification procedures and related machine learning techniques.

This paper is organized as follows. Section 2 introduces the Bregman divergence and investigates its statistical properties. Section 3 discusses the duality between the the Bregman divergence and its generating function. Section 4 studies the asymptotic behaviors of parametric estimation under the Bregman divergence, whereas Section 5 establishes the asymptotic distribution of nonparametric function estimation under the Bregman divergence. Section 6 applies results developed in Sections 2-5 to classification. Section 7 presents simulation evaluations. Section 8 concludes this paper by a discussion. All technical details are postponed to the Appendix.
2 Statistical Properties of Bregman Divergence

In this section, we discuss a range of statistical properties of the Bregman divergence, which acts as general loss functions in regression and classification. As will be shown in Sections 3 and 6, this class of loss functions unifies nearly all of the commonly used loss functions in the literature, like the general symmetric loss $L(a, b) = L(b, a)$ studied in James (2003).

2.1 Bregman divergence

For a given concave function $q$, Bregman (1967) introduced a device for constructing a bivariate function,

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu),$$

(2.1)
defined for $(\nu, \mu) \in M_0 \times M_1$, where $M_k$ is the set of points at which the $k$th derivative of $q$ exists. A graphical illustration of $Q$ associated with $q$ is displayed in Figure 1. The concavity requirement on $q$ ensures the non-negativity of $Q$. 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 Pietra, and Della Pietra (1997) for certain examples of Bregman divergence in machine learning literature, and Efron (1986, 2004) for the estimation of prediction error under BD.

[ Put Figure 1 about here ]

2.2 Convexity and Bartlett identity for BD $Q$

2.2.1 Convexity

We first consider the convexity of a BD $Q(\nu, \mu)$ in its first argument. Since $q$ is concave, it is straightforward to show that

$$Q(\lambda \nu_1 + (1 - \lambda) \nu_2, \mu) \leq \lambda Q(\nu_1, \mu) + (1 - \lambda) Q(\nu_2, \mu), \quad \text{for any } 0 \leq \lambda \leq 1.$$ 

Thus, for any fixed $\mu \in M_1$, $Q(\nu, \mu)$ is convex in its first argument $\nu \in M_0$. This property, by further assuming $\nu \in M_2$, can also be seen from calculating derivatives of (2.1):

$$\frac{\partial Q(\nu, \mu)}{\partial \nu} = -q'(\nu) + q'(\mu),$$

$$\frac{\partial^2 Q(\nu, \mu)}{\partial \nu^2} = -q''(\nu) \geq 0.$$
We then discuss the convexity of \( Q(\nu, \mu) \) 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 \( \mu \). For a general discussion, let us assume stronger smoothness conditions on \( q \). It follows from (2.1) that

\[
\frac{\partial Q(\nu, \mu)}{\partial \mu} = q'(\mu) + (-1)q'(\mu) + (\nu - \mu)q''(\mu) = (\nu - \mu)q''(\mu), \quad \mu \in \mathcal{M}_2, \tag{2.2}
\]

\[
\frac{\partial^2 Q(\nu, \mu)}{\partial \mu^2} = (\nu - \mu)q''(\mu) - q''(\mu), \quad \mu \in \mathcal{M}_3. \tag{2.3}
\]

Thus, for fixed \( \nu \), \( Q(\nu, \mu) \) is typically a \( V \)-shaped function of \( \mu \), but not necessarily convex in \( \mu \), even if stronger smoothness assumptions are imposed on \( q \). The non-convexity poses challenges to estimation procedures under BD. Furthermore, it can be deduced from (2.2)–(2.3) that all the \( k \)-order partial derivatives of \( Q(\nu, \mu) \) with respect to \( \mu \in \mathcal{M}_{k+1} \) are linear in \( \nu \).

### 2.2.2 Bartlett identity

In what follows, we denote by

\[ m(x) = E(Y|X = x) \]

the conditional regression function, and by \( \text{var}(Y|X = x) \) the conditional variance function. From (2.2)–(2.3), we observe that

\[
E\left\{ \frac{\partial Q(Y, m(x))}{\partial m(x)} \right\}_{X = x} = 0, \quad E\left\{ \frac{\partial^2 Q(Y, m(x))}{\partial m(x)^2} \right\}_{X = x} = -q''(m(x)).
\]

The Bartlett identity

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

holds if and only if

\[
q''(m(x)) = -\frac{1}{\text{var}(Y|X = x)}. \tag{2.4}
\]

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 \( E\{Q(Y, \mu(X))\} \), associated with a BD \( Q \), is minimized by the optimal “Bayes” rule \( m(X) \).
Theorem 1 Assume that $Q$ is a BD as defined in (2.1). Assume that $m(x)$ is measurable and $m(x) \in M_1$. Then among all measurable functions $\mu$ such that $\mu(x) \in M_1$,

$$\arg \min_{\mu \in M_1} E\{Q(Y, \mu(X))\} = m(X).$$

Throughout the rest of the paper, we assume that $(x_i, y_i)$ in the training sample (1.1) are independent pairs from a common distribution of $(x, y)$, and that $\hat{m}(x)$ is the estimate of $m(x)$, based on the training sample. Suppose that a test point $(x^o, y^o)$ follows the distribution of $(x, y)$ and is independent of the training sample. The conditional prediction error (PE) of the rule $\hat{m}(x)$ is defined as

$$r(x) = E\{Q(Y^o, \hat{m}(x))|x = x\},$$

and the expected prediction error by $E\{Q(Y^o, \hat{m}(x))|x = x\}$. Theorem 2 below indicates that when projected on the Bayes rule $m(x)$, the expected PE under BD satisfies the Pythagorean equality.

Theorem 2 Assume that $Q$ is a BD as defined in (2.1). Assume that $m(x) \in M_1$. Define

$$r_B(x) = E\{Q(Y^o, m(x))|x = x\} = q(m(x)) - E\{q(Y^o)|x = x\}. \tag{2.5}$$

Then the conditional PE has the decomposition,

$$r(x) = r_B(x) + Q(m(x), \hat{m}(x)), \tag{2.6}$$

and the expected PE fulfills the Pythagorean equality,

$$E\{Q(Y^o, \hat{m}(x))|x = x\} = E\{Q(Y^o, m(x))|x = x\} + E\{Q(m(x), \hat{m}(x))\}. \tag{2.7}$$

It follows that the Bayes rule $m(x)$ minimizes the conditional and expected PEs.

On the right side of (2.7), the first term is $r_B(x)$, an irreducible error due to the nature of the response, whereas the second term, $E\{Q(m(x), \hat{m}(x))\}$, corresponds to the functional estimation error. Hence, the function estimate $\hat{m}$ affects the prediction error only through the estimation error. However, in practice, the actual $m(x)$ is usually unknown. Recently, Efron (2004) derived an insightful result of covariance penalty on the in-sample prediction of a future output $Y_i^o$ at $x_i$, where $Y_i^o$ is an independent copy of $Y_i$. Adapting Efron's result to the current formulation, it follows that

$$E\{Q(Y_i^o, \hat{m}(x_i))|x_i\} = E\{Q(Y_i, \hat{m}(x_i))|x_i\} + 2\text{cov}(\hat{\lambda}_i, Y_i|x_i), \quad i = 1, \ldots, n, \tag{2.8}$$

where $\hat{\lambda}_i = q'(\hat{m}(x_i))/2$ and $x = (x_1, \ldots, x_n)$ denotes the entire training sequence of $x_i$. This result indicates that estimation of the average prediction error can be obtained from estimating the covariance penalty associated with $q$. 

5
2.4 Additive decomposition of bias and variance under $Q$

For the estimator $\hat{m}(x), \ E(\hat{m}(x))$ is called the aggregated estimator. Following the notion introduced in Tibshirani (1996) for aggregated estimator, we define the aggregated effect (AE) of $\hat{m}(x)$ under the BD $Q$ by

$$\text{AE}(\hat{m}(x)) = E\{Q(Y^o, \hat{m}(x))|X^o = x\} - E\{Q(Y^o, E\{\hat{m}(x)\})|X^o = x\}.$$

We can show by the application of (2.7) an alternative expression for AE,

$$\text{AE}(\hat{m}(x)) = E\{Q(m(x), \hat{m}(x))\} - Q(m(x), E\{\hat{m}(x)\}).$$

This applied to (2.7) yields the bias/variance decomposition,

$$E\{Q(Y^o, \hat{m}(x))|X^o = x\} = E\{Q(Y^o, m(x))|X^o = x\} + \text{bias}^2(\hat{m}(x)) + \text{var}(\hat{m}(x)) + \{\text{AE}(\hat{m}(x)) - \text{var}(\hat{m}(x))\}, \quad (2.9)$$

in which

$$\text{bias}^2(\hat{m}(x)) = Q(m(x), E\{\hat{m}(x)\}) = E\{Q(Y^o, E\{\hat{m}(x)\})|X^o = x\} - E\{Q(Y^o, m(x))|X^o = x\},$$

$$\text{var}(\hat{m}(x)) = E\{Q(\hat{m}(x), E\{\hat{m}(x)\})\}.$$

The results above indicate that $\text{bias}^2(\hat{m}(x)) = SE(\hat{m}(x))$ and $\text{AE}(\hat{m}(x)) = \text{VE}(\hat{m}(x))$, where $SE(\hat{m}(x))$ and $\text{VE}(\hat{m}(x))$ are as defined in James (2003).

The bias-variance decomposition under BD extends the conventional bias-variance decomposition under a quadratic loss. To see this, note that if $Q(\nu, \mu)$ is convex in $\mu$, then $\text{AE}(\hat{m}(x)) \geq 0$; for $Q(\nu, \mu)$ non-convex in $\mu$, $\text{AE}(\hat{m}(x))$ could be negative. If $Q$ is a quadratic loss, $\text{AE}(\hat{m}(x)) = \text{var}(\hat{m}(x))$ and thus the last term in (2.9) vanishes.

3 Duality between $Q$ and $q$

Given any concave $q$-function, $Q(\nu, \mu)$ is well defined by (2.1). In this section, we aim to address the inverse question: Given a loss function $Q(Y, \mu)$, how to recover the generating function $q$? Investigating this inverse problem has important implications. First, as can be seen from (2.5), the ideal Bayes rule serves as a benchmark with the lowest possible risk, which only depends on $q$. Second, evaluation of the estimation error in (2.7) requires not only $Q(Y, \mu)$, but also the full knowledge of $Q(\nu, \mu)$. However, as we will show, the form $Q(Y, \mu)$ alone could not fully determine $Q(\nu, \mu)$. There is a need to recover the $q$-function before obtaining $Q(\nu, \mu)$. Third, in order to use (2.8) for estimating the prediction errors, the forms of $q$ and hence $q'$ are needed. Fourth, as we will show in Section 3.4, once we know the $q$-function for binary classification, the complexity in constructing loss functions for multi-class classification will be much reduced. Nonetheless, solving the $q$-function is non-trivial and some approach via differential equation will be unnecessarily complicated.
**Definition 1** We call functions $q_1$ and $q_2$ equivalent if there exist constants $a$ and $b$ such that the equality

$$q_1(\mu) = q_2(\mu) + a\mu + b$$

holds for all $\mu \in M_1$.

Throughout the rest of the paper, we will not distinguish between equivalent functions $q_1$ and $q_2$, since they will generate an identical $Q$ function.

We first consider the ideal case in which the bivariate function $Q(\nu, \mu)$ is fully known for all possible values $\nu$ and $\mu$. Our Lemma 1 below offers an analytical representation (3.2) of the $q$-function.

**Lemma 1** Consider a bivariate function $Q(\cdot, \cdot)$, defined on $N_0 \times N_1$, which fulfills $Q(\nu, \nu) \equiv 0$ for all $\nu \in N_1$. Let $N_2$ be the set of points $\mu$ at which $\frac{\partial Q(\nu, \mu)}{\partial \mu}$ exists. Assume that $N_2$ is an interval except at most countably many points. Then Conditions A and B below are equivalent:

A. There exists a function $q$ which is concave and not identically zero such that (2.1) holds;

B. For all $\nu \in N_1$, $\mu \in N_2$ and $\nu \neq \mu$,

$$\frac{\partial Q(\nu, \mu)}{\partial \mu} \leq 0, \quad \text{and is free of } \nu.$$  

(3.1)

Moreover, if Condition B holds, then the generating function $q$ is given by

$$q(\mu) = \int_\alpha^\mu \frac{\mu - s}{\nu - s} \frac{\partial Q(\nu, s)}{\partial s} ds, \quad \mu \in N_2.$$  

(3.2)

It should be emphasized that in (3.1) the second half of condition, “and is free of $\nu$”, could not be dropped. A counter example is given a non BD function $Q(\nu, \mu) = (\nu - \mu)^4$, which satisfies the first half of (3.1) but violates the second half.

In realistic statistical applications, loss functions are directly given by $Q(Y, \mu)$ for a random variable $Y$, rather than by $Q(\nu, \mu)$. However, the form of $Q(Y, \mu)$ alone, which is $Q(\nu, \mu)$ evaluated at the point $\nu = Y$, does not necessarily fully determine the form of $Q(\nu, \mu)$ for all $\nu$. Namely, naively substituting $Y$ in $Q(Y, \mu)$ with $\nu$ may fail to fully retrieve $Q(\nu, \mu)$, and hence we may not use Lemma 1 to recover the $q$-function; similarly this may incorrectly evaluate the term of estimation error in (2.7). For example, consider the loss function between a binary random variable $Y$ and $\mu$ given by

$$Y(1 - \mu)^2 + (1 - Y)\mu^2.$$  

(3.3)

This loss function results from setting $\nu = Y$ in the quadratic loss, $Q(\nu, \mu) = (\nu - \mu)^2$. Nonetheless, simply replacing $Y$ in (3.3) by $\nu$ leads to $\nu(1 - \mu)^2 + (1 - \nu)\mu^2$, which is
clearly not the actual loss \((\nu - \mu)^2\). Similar problems will arise from other loss functions in classification.

Theorem 3 below reveals that \(q(\mu)\) can be obtained from the reduced form \(Q(Y, \mu)\),

\[
Q(Y, \mu) = -q(Y) + q(\mu) + (Y - \mu)q'(\mu), \quad \mu \in \mathcal{M}_1,
\]

(3.4)

without the full knowledge of \(Q(\nu, \mu)\) as required in Lemma 1. Corollaries 1 and 2 further indicate that if \(E\{Q(Y, \mu)|X = x\}\) depends on the conditional distribution of \(Y\) only through its conditional regression \(m(x)\), then the generating function \(q\) can be constructed more easily and under weaker smoothness assumptions on \(Q\).

**Theorem 3** For a random variable \(Y\), consider a loss function \(Q(Y, \mu)\), defined for \(\mu \in \mathcal{N}_1\), which fulfills \(Q(Y, Y) \equiv 0\) for all \(Y \in \mathcal{N}_1\). Let \(\mathcal{N}_2\) be the set of points \(\mu\) at which \(\frac{\partial Q(Y, \mu)}{\partial \mu}\) exists. Assume that \(\mathcal{N}_2\) is an interval except at most countably many points. Then Conditions A' and B' below are equivalent:

**A'**. There exists a function \(q\) which is concave and not identically zero such that (3.4) holds;

**B'**. For all \(\mu \in \mathcal{N}_2\) and \(Y \neq \mu\),

\[
\frac{\partial Q(Y, \mu)}{Y - \mu} \leq 0, \quad \text{and is free of } Y.
\]

(3.5)

Moreover, if Condition B' holds, then the generating function \(q\) is given by

\[
q(\mu) = \int_a^\mu \mu - s \frac{\partial Q(Y, s)}{Y - s} ds, \quad \mu \in \mathcal{N}_2.
\]

(3.6)

**Corollary 1** Consider a function \(Q(Y, \mu)\), defined for \(\mu \in \mathcal{N}_1\). Suppose that \(E\{Q(Y, \mu)|X = x\}\) depends on the conditional distribution of \(Y\) only through its conditional expectation \(m(x)\). Denote by \(E(\mu; m(x))\) the expression \(E\{Q(Y, \mu)|X = x\}\), with all additive terms independent of \(\mu\) removed. If Condition A' holds, then the generating function \(q\) is given by

\[
q(\mu) = E(\mu; \mu), \quad \mu \in \mathcal{N}_1.
\]

(3.7)

**Corollary 2** Assume that \(Y|X = x \sim \text{Bernoulli}(m(x))\). Consider a function \(Q(Y, \mu)\), defined for \(\mu \in \mathcal{N}_1\), which fulfills \(Q(Y, Y) \equiv 0\) for all \(Y \in \mathcal{N}_1\). If A' holds, then the generating function \(q\) is given by

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

(3.8)

In particular, \(q(\mu)\) given by the form (3.8) satisfies

\[
q(Y) \equiv 0,
\]

\[
q'(\mu) = Q(1, \mu) - Q(0, \mu), \quad \mu \in \mathcal{N}_1.
\]

(3.9)
In summary, for a given loss function \( Q(Y, \mu) \), we first check the inequality (3.5). If it is valid, then apply (3.6), (3.7), or (3.8) to obtain the \( q \)-function. Applications of (3.6), (3.7), and (3.8) are illustrated below to three situations, ranging respectively from general to specific assumptions on the distribution of \( Y \).

3.1 Quasi-likelihood function: application of (3.6)

A quasi-likelihood function \( Q^* \) was introduced in the statistical literature to relax the distributional assumption on a random variable \( Y \) via the specification,

\[
\frac{\partial Q^*(\mu; Y)}{\partial \mu} = \frac{Y - \mu}{V(\mu)},
\]

in which it is assumed that \( \text{var}(Y|X = x) = \sigma^2 V\{E(Y|X = x)\} \) for a nuisance parameter \( \sigma^2 > 0 \) and a given positive function \( V \). It is easy to see from (3.10) that \( Q^*(\mu; Y) \leq Q^*(Y; Y) \). Setting \( Q(Y, \mu) = Q^*(Y; Y) - Q^*(\mu; Y) \), then \( Q(Y, Y) = 0 \) and

\[
\frac{\partial Q(Y, \mu)}{\partial \mu} = \frac{Y - \mu}{V(\mu)},
\]

thus (3.5) is satisfied. An application of (3.6) yields the generating \( q \)-function,

\[ q(\mu) = \int_{\mu}^{\infty} \frac{s - \mu}{V(s)} ds. \]

3.2 Exponential family of probability functions: application of (3.7)

A special case of the quasi-likelihood function is the exponential family distribution, where the conditional probability function of \( Y \) given \( X = x \) is

\[ f_{Y|X}(y|x) = \exp\{y\theta(x) - b(\theta(x))\}/a(\psi) + c(y, \psi), \]

for some known functions \( a(\cdot), b(\cdot), \) and \( c(\cdot, \cdot) \), where \( \theta(x) \) is called a canonical parameter and \( \psi \) is called a dispersion parameter, respectively. Let's consider the commonly used Kullback-Leibler divergence (or the deviance loss) defined by

\[
L(Y, \mu) = 2[Y\tilde{\theta} - \theta - \{b(\tilde{\theta}) - b(\theta)\}]
\]

\[
= 2[-Y\theta + b(\theta) + Y\tilde{\theta} - b(\tilde{\theta})],
\]

where \( b'(\tilde{\theta}) = Y \) and \( b'(\theta) = \mu \). Then \( L(Y, Y) = 0 \). Define \( Q(Y, \mu) = L(Y, \mu) \). Assume that \( b' \) is a bijection. It follows that

\[
\frac{\partial Q(Y, \mu)}{\partial \mu} = 2\left\{ -Y \frac{d\theta}{d\mu} + \frac{db(\theta)}{d\mu} \right\} = 2\left\{ -Y \frac{d\theta}{d\mu} + \frac{db(\theta)}{d\theta} \frac{d\theta}{d\mu} \right\}
\]

\[
= -2(Y - b'(\theta)) \frac{d\theta}{d\mu} = \frac{2(Y - \mu)}{b''(\theta)}.
\]
Thus (3.5) is fulfilled. Since $E\{Y \tilde{\theta} - b(\tilde{\theta})|X = x\}$ is independent of $\mu$, $E(\mu; m(x)) = 2\{-m(x)\theta + b(\theta)\}$. We obtain from (3.7) the generating $q$-function,

$$q(\mu) = 2\{b(\theta) - \mu \theta\}, \quad \text{where} \quad b'(\theta) = \mu. \quad (3.11)$$

Compared with (3.6), this approach for obtaining $q$ seems to be more convenient.

### 3.3 Binary response: application of (3.8)

For a binary variable $Y|X = x \sim \text{Bernoulli}(m(x))$, its distribution belongs to the exponential family. Because of the importance of binary responses in classification, we will examine four types of loss functions $Q(Y, \mu)$, where $0 < \mu < 1$, determine which of them is a Bregman divergence and illustrate how to obtain $q(\mu)$, $q'(\mu)$ and $Q(\nu, \mu)$.

#### 3.3.1 Misclassification loss

The misclassification loss is $Q(Y, \mu) = |Y - c|1\{Y \neq I[\mu > c]\}$, where $1[.]$ is an indicator function, and the constant $0 < c < 1$ denotes the misclassification cost with the most common choice $c = 1/2$ for equal cost. This loss function can be rewritten as

$$Q(Y, \mu) = Y(1-c)I[\mu < c] + (1-Y)cI[\mu > c]. \quad (3.12)$$

Note that $Q(Y, Y) = 0$ and for $\mu \in \mathcal{N}_1 = \{\mu : \mu \neq c\}$,

$$\partial Q(Y, \mu)/\partial \mu = 0,$$

thus (3.5) is fulfilled. Define $\text{sgn}(\mu)$ to be 1 if $\mu > 0$ and $-1$ if $\mu < 0$. We obtain from (3.8) and (3.9) that

$$q(\mu) = \mu(1-c)I[\mu < c] + (1-\mu)cI[\mu > c]$$

$$= 2^{-1}(c - \mu)\text{sgn}(\mu - c) + 2^{-1}\{c + (1 - 2c)\mu\}$$

$$= \min\{\mu(1-c), (1-\mu)c\}, \quad (3.13)$$

$$q'(\mu) = -2^{-1}\text{sgn}(\mu - c) + 2^{-1}(1-2c), \quad \text{for} \quad \mu \in \mathcal{N}_1. \quad (3.14)$$

Hence from (2.1), it is easy to verify the bivariate $Q$ function,

$$Q(\nu, \mu) = -2^{-1}(\nu - \mu)\text{sgn}(\nu - c) - 2^{-1}\{c + (1 - 2c)\nu\}$$

$$+ 2^{-1}(c - \mu)\text{sgn}(\mu - c) + 2^{-1}\{c + (1 - 2c)\mu\}$$

$$- 2^{-1}(\nu - \mu)\text{sgn}(\mu - c) - 2^{-1}(\nu - \mu)(1 - 2c)$$

$$= 2^{-1}(\nu - c)\{\text{sgn}(\nu - c) - \text{sgn}(\mu - c)\}$$

$$= 2^{-1}|\nu - c|\text{sgn}(\nu - c)\{\text{sgn}(\nu - c) - \text{sgn}(\mu - c)\}$$

$$= 2^{-1}|\nu - c|\{1 - \text{sgn}(\nu - c)\text{sgn}(\mu - c)\}$$

$$= |\nu - c|I\{\text{sgn}(\nu - c) \neq \text{sgn}(\mu - c)\}$$

$$= |\nu - c|I\{I[\nu > c] \neq I[\mu > c]\}. \quad (3.15)$$
3.3.2 Polynomial loss

The polynomial loss is $Q(Y, \mu) = |Y - \mu|^k = Y(1 - \mu)^k + (1 - Y)\mu^k$ where $k > 0$ is a constant. In this case, $Q(Y, Y) = 0$. For $Y \neq \mu$, we observe that

$$\frac{\partial Q(Y, \mu)}{\partial \mu} = -k|Y - \mu|^{k-1}\text{sgn}(Y - \mu) = -k(Y - \mu)|Y - \mu|^{k-2},$$

which gives

$$\frac{\partial Q(Y, \mu)}{Y - \mu} = -k|Y - \mu|^{k-2}.$$

Thus (3.5) 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 $\mathcal{BD}$.

For the quadratic loss, we obtain from (3.8) and (3.9) that

$$q(\mu) = \mu(1 - \mu)^2 + (1 - \mu)\mu^2 = \mu(1 - \mu), \quad q'(\mu) = 1 - 2\mu.$$

Hence from (2.1), we obtain the $Q$ function,

$$Q(\nu, \mu) = -\nu(1 - \nu) + \mu(1 - \mu) + (\nu - \mu)(1 - 2\mu)$$

$$= -\nu + \nu^2 + \mu - \mu^2 - 2\nu\mu + \nu + 2\mu^2 - \mu = \nu^2 - 2\nu\mu + \mu^2$$

$$= (\nu - \mu)^2.$$

3.3.3 Binomial deviance

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. In this case, $Q(Y, Y) = 0$ and

$$\frac{\partial Q(Y, \mu)}{\partial \mu} = -\frac{2(Y - \mu)}{\mu(1 - \mu)},$$

thus (3.5) is fulfilled. We obtain from (3.8) and (3.9) that

$$q(\mu) = -2\{\mu \ln(\mu) + (1 - \mu) \ln(1 - \mu)\}, \quad q'(\mu) = 2\ln\left(\frac{1 - \mu}{\mu}\right).$$

Hence from (2.1), we obtain the $Q$ function,

$$Q(\nu, \mu) = 2[\nu \ln(\nu) + (1 - \nu) \ln(1 - \nu) - \mu \ln(\mu) - (1 - \mu) \ln(1 - \mu)$$

$$+ (\nu - \mu) \ln(1 - \mu) - (\nu - \mu) \ln(\mu)]$$

$$= 2[\nu \ln(\nu) + (1 - \nu) \ln(1 - \nu) - \nu \ln(\mu) - (1 - \nu) \ln(1 - \mu)]$$

$$= 2[\nu \ln(\nu/\mu) + (1 - \nu) \ln\{(1 - \nu)/(1 - \mu)\}].$$
3.3.4 Exponential loss

The exponential loss used in AdaBoost is $Q(Y, \mu) = e^{-(2Y-1)F(\mu)}$, where $F(\mu) = \frac{1}{2} \ln \left( \frac{\mu}{1-\mu} \right)$. In this case, it is easy to see that $Q(Y, \mu) = Y \left( \frac{1-\mu}{\mu} \right)^{1/2} + (1 - Y) \left( \frac{\mu}{1-\mu} \right)^{1/2}$, $Q(Y, Y) = 0$, and

$$\frac{\partial Q(Y, \mu)}{\partial \mu} = -\frac{Y - \mu}{2\mu(1-\mu)^{3/2}},$$

thus (3.5) is fulfilled. We obtain from (3.8) and (3.9) that

$$q(\mu) = \mu \left( \frac{1-\mu}{\mu} \right)^{1/2} + (1 - \mu) \left( \frac{\mu}{1-\mu} \right)^{1/2} = 2\mu(1-\mu)^{1/2},$$

$$q'(\mu) = \frac{1 - 2\mu}{\mu(1-\mu)^{1/2}}.$$

Hence from (2.1), we obtain the $Q$ function,

$$Q(\nu, \mu) = -2\nu(1-\nu)^{1/2} + 2\mu(1-\mu)^{1/2} + (\nu - \mu) \frac{1 - 2\mu}{\mu(1-\mu)^{1/2}}$$

$$= \left[ -2\nu(1-\nu)\mu(1-\mu)^{1/2} + 2\mu(1-\mu) + (\nu - \mu)(1-2\mu) \right] / \{ \mu(1-\mu)^{1/2} \}$$

$$= \left[ -2\nu(1-\nu)\mu(1-\mu)^{1/2} + \mu - 2\mu\nu + \nu \right] / \{ \mu(1-\mu)^{1/2} \}$$

$$= \left[ -2\nu(1-\nu)\mu(1-\mu)^{1/2} + \mu(1-\nu) + \nu(1-\mu) \right] / \{ \mu(1-\mu)^{1/2} \}$$

$$= \left[ (\mu(1-\nu))^{1/2} - \{ \nu(1-\mu) \}^{1/2} \right] / \{ \mu(1-\mu)^{1/2} \}.$$

3.4 Multi-class response

For a multi-class response $Y \in G = \{G_1, \ldots, G_K\}$, where $K \geq 2$, we will demonstrate that the Bregman divergence can be generalized from that for binary responses. For vectors $\nu = (\nu_1, \ldots, \nu_K)^T$ and $\mu = (\mu_1, \ldots, \mu_K)^T$ which are discrete probability measures, the definition (2.1) can be extended from scalar arguments to vectors $\nu$ and $\mu$ as follows,

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)^T \nabla q(\mu),$$

where $\nabla q(\mu) = (\frac{\partial q(\mu)}{\partial \mu_1}, \ldots, \frac{\partial q(\mu)}{\partial \mu_K})^T$ is the gradient vector of $q$. Define $y = (I(Y = G_1), \ldots, I(Y = G_K))^T$. Indeed, the choice of $q(\mu)$ can be made from the counterpart developed in Section 3.3 for binary responses.
3.4.1 Misclassification loss

Define \( q(\mu) = 1 - \max_{1 \leq j \leq K} \mu_j \). We see that
\[
q(\mu) = 1 - \sum_{j=1}^{K} \mu_j \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \mu_k],
\]
\[
\nabla q(\mu) = \left(- \mathbb{I}[1 = \arg \max_{1 \leq k \leq K} \mu_k], \ldots, - \mathbb{I}[K = \arg \max_{1 \leq k \leq K} \mu_k]\right)^T,
\]
\[
Q(\nu, \mu) = \sum_{j=1}^{K} \nu_j \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \nu_k] - \sum_{j=1}^{K} \mu_j \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \mu_k] - \sum_{j=1}^{K} (\nu_j - \mu_j) \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \mu_k]
\]
\[
= \sum_{j=1}^{K} \nu_j \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \nu_k] - \mathbb{I}[j = \arg \max_{1 \leq k \leq K} \mu_k]
= \nu_{\arg \max_{1 \leq k \leq K} \nu_k} - \nu_{\arg \max_{1 \leq k \leq K} \mu_k}.
\]

This gives the misclassification loss,
\[
Q(y, \mu) = \mathbb{I}[y = \hat{y}_{\arg \max_{1 \leq k \leq K} \mathbb{I}[y = \hat{y}_k]}] - \mathbb{I}[y = \hat{y}_{\arg \max_{1 \leq k \leq K} \mu_k}]
= \mathbb{I}[y \neq \hat{y}_{\arg \max_{1 \leq k \leq K} \mu_k}].
\]

3.4.2 Quadratic loss

Define \( q(\mu) = \sum_{j=1}^{K} \mu_j (1 - \mu_j) \). We see that \( \nabla q(\mu) = (1 - 2\mu_1, \ldots, 1 - 2\mu_K)^T \) and thus
\[
Q(\nu, \mu) = -\sum_{j=1}^{K} \nu_j (1 - \nu_j) + \sum_{j=1}^{K} \mu_j (1 - \mu_j) + \sum_{j=1}^{K} (\nu_j - \mu_j) (1 - 2\mu_j)
\]
\[
= \sum_{j=1}^{K} (-\nu_j + \nu_j^2 + \mu_j - \mu_j^2 + \nu_j - 2\nu_j \mu_j - \mu_j + 2\mu_j^2) = \sum_{j=1}^{K} (\nu_j - \mu_j)^2.
\]

This gives the quadratic loss.

3.4.3 Multinomial deviance

Define \( q(\mu) = -\sum_{j=1}^{K} \mu_j \ln(\mu_j) \). We see that \( \nabla q(\mu) = -(\ln(\mu_1) + 1, \ldots, \ln(\mu_K) + 1)^T \) and thus
\[
Q(\nu, \mu) = \sum_{j=1}^{K} \nu_j \ln(\nu_j) - \sum_{j=1}^{K} \mu_j \ln(\mu_j) - \sum_{j=1}^{K} (\nu_j - \mu_j) (\ln(\mu_j) + 1)
\]
\[
= \sum_{j=1}^{K} \{\nu_j \ln(\nu_j) - \mu_j \ln(\mu_j) - \nu_j \ln(\mu_j) - \nu_j + \mu_j \ln(\mu_j) + \mu_j\} = \sum_{j=1}^{K} \nu_j \ln(\nu_j / \mu_j).
\]

This gives the relative entropy, and the corresponding loss function
\[
Q(y, \mu) = -\sum_{j=1}^{K} \mathbb{I}[y = \hat{y}_j] \ln(\mu_j) = -\ln(f(Y))
\]
is the multinomial deviance (or negative log-likelihood).
3.4.4 Exponential loss

Define \( q(\mu) = \sum_{j=1}^{K} 2\{\mu_j(1-\mu_j)\}^{1/2} \). We see that \( \nabla q(\mu) = (\frac{1-2\mu_1}{\mu_1(1-\mu_1)}^{1/2}, \ldots, \frac{1-2\mu_K}{\mu_K(1-\mu_K)}^{1/2})^T \)
and thus

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

This gives the exponential loss function

\[
Q(y, \mu) = \sum_{j=1}^{K} \left( \frac{1-\mu_j}{\mu_j} \right)^{1/2} \{21[Y=g_j]-1\} = \sum_{j=1}^{K} e^{-\{21[Y=g_j]-1\}F(\mu_j)},
\]

where \( F(\mu) = \frac{1}{2} \ln(\frac{\mu}{1-\mu}) \). Hastie, Tibshirani and 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. The key technical component is \( q \) whose construction is not dependent on the number of classes.

Figure 2 illustrates the functions \( q(\mu) \) in the cases of two-classes and three-classes.

4 Parametric Estimation under Bregman Divergence

In this section, we aim to study the parametric estimation of \( m(x) \) under \( \mathcal{B} \mathcal{D} \). This is achieved by estimating \( F(m(x)) \), for a known link function \( F(\cdot) \). Define \( \eta(x) = F(m(x)) \).

We assume that \( \eta(x) = \alpha_0 + x^T \alpha \) for some unknown parameters \( \alpha_0 \) and \( \alpha \). Based on the independent training data \( \{(x_i, y_i)\}_{i=1}^{n} \) from the population \( (x, y) \), the minimum \( \mathcal{B} \mathcal{D} \) estimator \( (\hat{\alpha}_0, \hat{\alpha}) \) of \( (\alpha_0, \alpha) \) is proposed by the minimizer of the criterion function,

\[
\ell_n(\alpha_0, \alpha) = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, F^{-1}(\alpha_0 + x_i^T \alpha)),
\]

where the loss function \( Q \) is a Bregman divergence. Define \( \bar{x} = (x_1, \ldots, x_d)^T \) and \( \bar{X} = (1, \bar{x}^T)^T \). Set \( \bar{\alpha} = (\alpha_0, \bar{\alpha}^T)^T \), which is in a parameter space \( \Theta \subseteq \mathbb{R}^{d+1} \). The asymptotic consistency and normality of \( \bar{\alpha} \) are presented in Theorems 4–5 respectively.

**Theorem 4** Let \( \bar{\alpha}^{(0)} \) denote the true value of \( \bar{\alpha} \). Suppose that \( Q \) is a \( \mathcal{B} \mathcal{D} \). Assume Condition C in the Appendix. Then as \( n \to \infty \), the minimum \( \mathcal{B} \mathcal{D} \) estimator \( \bar{\alpha} \) is asymptotically consistent.

**Theorem 5** Let \( \bar{\alpha}^{(0)} \) denote the true value of \( \bar{\alpha} \). Suppose that \( Q \) is a \( \mathcal{B} \mathcal{D} \). Assume Condition D in the Appendix. Define \( \bar{\alpha} = (\bar{\alpha}_0, \bar{\alpha}^T)^T \). Then as \( n \to \infty \), the minimum \( \mathcal{B} \mathcal{D} \) estimator \( \bar{\alpha} \) is asymptotically normal,

\[
\sqrt{n} \{ \bar{\alpha} - \bar{\alpha}^{(0)} \} \overset{d}{\to} N(0, H_0^{-1} \Omega_0 H_0^{-1}),
\]
where $H_0 = -E[q''(m(X))\{F'(m(X))^2\}^{-2}(XX^T)]$ and $\Omega_0 = E[\text{var}(Y|X)\{q''(m(X))^2\{F'(m(X))^2\}^{-2}(XX^T)]$.

As an illustrative example, consider the exponential family of distributions in Section 3.2 with the $q$-function given in (3.11). For the canonical link $F(\cdot) = (b')^{-1}(\cdot)$, the asymptotic covariance matrix in Theorem 5 reduces to $a(\psi)[E\{XX^T b''(\hat{X}^{(0)}_{\theta})\}]^{-1}$.

Theorem 5 reveals that the asymptotic distribution of the parametric estimator depends on the choice of the loss function $Q$ via the second derivative of its generating $q$-function. This dependence arises from the global nature of the parametric estimation. This stands in stark contrast to the nonparametric counterpart. As will be seen from the next section, the local feature of the nonparametric estimation entirely relieves the dependence on choices of $Q$ from the asymptotic distribution of the nonparametric estimators.

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

$$q''(m(x)) = -\frac{c}{\text{var}(Y|X = x)}, \quad \text{for a constant } c > 0, \quad (4.2)$$

which includes the Bartlett identity (2.4) as a special case.

**Theorem 6** If the $q$-function satisfies (4.2), then the asymptotic covariance matrix of $\hat{\alpha}$ in Theorem 5 achieves the lowest bound

$$\left( E[1/\text{var}(Y|X)\{F'(m(X))^2\}^{-2}(XX^T)] \right)^{-1}.$$  

5 **Nonparametric Estimation under Bregman Divergence**

In this section, we describe the $p$th degree local polynomial estimation of $m(x)$ under BD. The estimation will be discussed in the cases of univariate $x$ and multivariate $x$ respectively.

5.1 **Univariate nonparametric regression**

To facilitate presentations, we first assume that $X$ is univariate. Assume 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

$$\eta(X_i) = \beta_0(x) + (X_i - x)\beta_1(x) + \cdots + (X_i - x)^p\beta_p(x) = x_i(x)^T\beta(x),$$

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

$$
\ell_n(\beta; x) = \frac{1}{n} \sum_{i=1}^{n} Q\{Y_i, F^{-1}(x_i(x)^T \beta)\} K_h(X_i - x),
$$

(5.1)

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

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

$$
\sqrt{nh} \left[ H\{\hat{\beta}(x) - \beta(x)\} - S^{-1} c_p \frac{\eta^{(p+1)}(x)}{(p+1)!} \right] \xrightarrow{p} N(0, S^{-1} S^* S^{-1} v(x) f_X(x)),
$$

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

Theorem 7 has the following important implications. First, the traditional local polynomial estimation is under the quadratic loss for Gaussian responses and the deviance loss for exponential family responses. In contrast, the loss function in the current paper is the broader class of Bregreman divergence. Furthermore, Theorem 7 indicates that the asymptotic distributions of the local BD estimator $\hat{\beta}(x)$, and hence $\hat{\eta}(x)$ and $\hat{m}(x)$, 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 $F$. Thus, Theorem 7 indeed gains further insight into nonparametric function estimation.

From a function estimation perspective, Theorem 7 enables us to derive the asymptotically optimal smoothing parameter $h$ under the general BD. This result includes Fan and Zhang (2004), which focuses on the local-likelihood estimator (based on the deviance loss for responses having distributions in the exponential family), as a special case, and continues to be applicable to the exponential loss in AdaBoost, where the local-likelihood estimator will clearly not be applicable.

From a classification point of view, since the classifier only depends on the sign of $F(\hat{m}(x))$, 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. Indeed, Bühlmann and Yu (2003) found comparable performances between $L_2$Boost and LogitBoost. We will revisit this issue in more detail in the next section.
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 $\mathbf{x} = (X_1, \ldots, X_d)^T$. For the response variable $Y$, define by $m(u, x) = E(Y \mid U = u, \mathbf{x} = x)$ the conditional mean regression function, where $\mathbf{x} = (x_1, \ldots, x_d)^T$. The varying-coefficient model assumes that

$$F(m(u, x)) = \eta(u, x) = \sum_{k=1}^{d} a_k(u) x_k = x^T \mathbf{A}(u),$$

(5.2)

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

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

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

in which the symbol $\otimes$ denotes the Kronecker product, and thus,

$$\eta(U_i, \mathbf{x}_i) = \{\mathbf{u}_i(u) \otimes \mathbf{x}_i\}^T \mathbf{B}(u).$$

The local minimum BD estimator $\hat{\beta}(u)$ minimizes the criterion function,

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

(5.3)

The first $d$ entries of $\hat{\beta}(u)$ supply the local minimum BD estimates $\hat{\mathbf{A}}(u)$ of $\mathbf{A}(u)$, and the local minimum BD estimates of $\eta(u, x)$ and $m(u, x)$ are given by $\hat{\eta}(u, x) = x^T \hat{\mathbf{A}}(u)$ and $\hat{m}(u, x) = F^{-1}(\hat{\eta}(u, x))$, respectively. Proofs for the varying-coefficient regression model will be added.

Theorem 8 Define $H = \text{diag}\{1, h, \ldots, h^p\}$ and $\mathbf{H} = H \otimes \mathbf{I}_d$. Assume that the degree $p$ is odd. Assume that $Q$ is a BD. Suppose that Condition E7 in the Appendix holds. If $n \to \infty$, $nh \to \infty$ and $n h^{p+3} = O(1)$, then for the local minimum BD estimator $\hat{\beta}(u)$ at an interior point $u$ of the design density, we have

$$\sqrt{nh} \left[ H \hat{\beta}(u) - \beta(u) \right] - \left\{ S^{-1} c_p \otimes \mathbf{A}^{(p+1)}(u) \right\}_{(p+1)!} \xrightarrow{D} \mathcal{N}(0, \left\{ S^{-1} S^* S^{-1} \otimes \{\Gamma(u)^{-1} \Delta(u) \Gamma(u)^{-1}\} \right\}/f_U(u)),$$

where $\Gamma(u) = -E\{q''(m(u, x))F'(m(u, x))\}^{-2}XX^T|U = u$ and 

$$\Delta(u) = E\{\text{var}(Y|U = u, x)\} q''(m(u, x))^{-2} \{F'(m(u, x))\}^{-2} XX^T|U = u.$$
Similar to Theorem 6, if the $q$-function satisfies $q''(m(u, x)) = -c/\text{var}(Y | U = u, X = x)$ for a constant $c > 0$, then the asymptotic covariance matrix of $\hat{\beta}(u)$ in Theorem 8 achieves the lowest bound

$\left( \frac{1}{\text{var}(Y | U = u, X)} \{F'(m(u, x))\}^{-2} XX^T | U = u \right)^{-1}$.

6 Applications to Classification

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

6.1 Function estimation error and classification error

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

We consider the nonstandard situation with unequal misclassification cost. Indeed, Hand and Vinciotti (2003) provide insightful examples that serve to stress the importance of accounting for unequal classification costs. We now demonstrate that B&L provides a much flexible and convenient tool for the prediction error associated with the misclassification loss function and for studying the effect of estimation error of the local estimators on the classification error. It is easy to show that the Bayes classifier $Y_B(x) = I[m(x) > c]$ minimizes the misclassification risk $E[L(Y, \hat{Y}(X))]$ with respect to $\hat{Y}$. We now study the classification performance of $\hat{Y}(x) = I[\hat{m}(x) > c]$, which substitutes the true function in the Bayes rule by the function estimate. The corresponding misclassification loss is $L(Y, \hat{Y}(X))$, which agrees with $Q(Y, \hat{m}(X))$ in (3.12) associated with $q$ in (3.13). Applying (2.5) in Theorem 2, we deduce

$$r_B(x) = \min\{m(x)(1 - c), \{1 - m(x)\}c\}.$$  

Applying (3.15) to (2.6) yields

$$Q(m(x), \hat{m}(x)) = |m(x) - c| I\{Y_B(x) \neq \hat{m}(X) > c\}.$$
Thus the estimation error of $\hat{m}(x)$ in (2.7) is
\[ E\{Q(m(x), \hat{m}(x))\} = |m(x) - c|P\{Y_B(x) \neq \mathbb{1}[\hat{m}(x) > c]\}, \]
in which
\[ P\{Y_B(x) \neq \mathbb{1}[\hat{m}(x) > c]\} = \mathbb{1}[m(x) < c]P\{\hat{m}(x) > c\} + \mathbb{1}[m(x) > c]P\{\hat{m}(x) < c\}. \]

According to Theorem 7, the local estimator $\hat{m}(x)$ has an asymptotically normal distribution. Thus, by using the idea of normal approximation similar to that of Friedman (1997), an asymptotic approximation of the above probability is provided by
\[ \Phi\left[\text{sgn}\{c - m(x)\} \text{sgn}\{E(\hat{m}(x)) - c\} \frac{|E\{\hat{m}(x)\} - c|}{\sqrt{\text{var}\{\hat{m}(x)\}}} \right], \]
where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution. Thus, when $m(x)$ and the aggregated predictor $E\{\hat{m}(x)\}$ are on the same side of the classification boundary $m(x) = c$, the misclassification risk will decrease as $E\{\hat{m}(x)\}$ is farther away from $c$ irrespective of the function estimation bias $m(x) - E\{\hat{m}(x)\}$; when $m(x)$ and $E\{\hat{m}(x)\}$ are on opposite sides of the classification boundary, the misclassification risk will increase with the distance between $E\{\hat{m}(x)\}$ and $c$.

In the next two subsections, we will investigate two issues related to the misclassification risk.

### 6.2 Relation between margin-based loss function and Bregman divergence

Intuitively, the misclassification loss (3.12) should be used as the training loss, since it is the loss function used to evaluate the performances of classifiers. However, the function is not convex, not continuous, 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 function, such as the exponential loss function (Freund and Schapire, 1997; Friedman, Hastie, and Tibshirani,
is to show the convergence rate of boosting procedures (Schapire, 2002) and in turn indirectly bound the misclassification error.

One of the interests is to study 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. In the following, we first show that for a given loss function \( L(Y, \mu) \), the way of finding its margin-based loss function written in terms of \( V(Y^*F(\mu)) \).

**Lemma 2** Consider a loss function \( L(Y, \mu) \) for a binary variable \( Y \). Suppose

**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 2 actually supplies a simple method to represent \( L \) in the form of \( V(Y^*F) \). Conditions **F** hold 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 Bregman divergence, its alternative form \( V(Y^*F) \) in Lemma 2 is not necessarily either. Recall that we have shown in Section 3.3 the misclassification loss is BD. We now show that under very mild smoothness conditions on \( V \) and \( F \), the corresponding margin-based loss function is indeed a Bregman divergence.

**Theorem 9** Consider a margin-based loss function \( V(Y^*F) \). Define \( \mathcal{F}_{k+1} = \{ s : V^{(k)}(s) \text{ exists} \} \). Assume that

**H.** There exists \( F_B \) such that \( \pm F_B \in \mathcal{F}_2 \) and \( \frac{\partial V(Y^*F_B(\mu))}{\partial \mu} \) exists almost everywhere. Define \( N_2 = \{ \mu : \frac{\partial V(Y^*F_B(\mu))}{\partial \mu} \text{ exists} \} \), and for all \( \mu \in N_2 \), we have

\[
F_B'(\mu) \geq 0; \tag{6.1}
\]

**I.** \( V'(F_B(\mu)) \leq 0 \) and

\[
\mu V'(F_B(\mu)) = (1 - \mu)V'(-F_B(\mu)). \tag{6.2}
\]

Then \( V(Y^*F_B(\mu)) \) is a Bregman divergence written in the form,

\[
Q(Y, \mu) = V(Y^*F_B(\mu)) - V(Y^*F_B(Y)), \quad \mu \in N_2, \tag{6.3}
\]

for which the generating function \( q \) is given by

\[
q(\mu) = \mu V(F_B(\mu)) + (1 - \mu)V(-F_B(\mu)), \quad \mu \in N_2. \tag{6.4}
\]

20
The applicability of Theorem 9 relies on finding \( F_B \). In the case of convex \( V \), Lemma 3 below draws connections between \( F_B \) and the Bayes rule, whereas Lemma 4 indicates that Condition \( \mathbf{H} \) is also necessary in certain situations.

**Lemma 3** 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

\[
F_B(m(x)) = \arg \min_{F \in \mathcal{F}_2} E\{V(Y^*F)|X = x\} \tag{6.5}
\]

implies (6.2).

**Lemma 4** Assume \( V''(s) \geq 0 \) for all \( s \in \mathcal{F}_2 \), and \( F_B \) satisfies Condition \( \mathbf{I} \). Then (6.1) must hold.

In the following, we demonstrate that five margin-based loss functions are Bregman divergence. Figure 3 presents an illustration. In all cases, it can be seen that \( \text{sgn}\{F_B(\mu)\} = \text{sgn}(\mu - 1/2) \).

\[
\text{[Put Figure 3 about here]}
\]

- \( V(s) = (1 - s)^2 \) for the quadratic loss: We have that \( V'(s) = -2(1 - s) \). From (6.2), setting \( \mu(1 - F) = (1 - \mu)(1 + F) \) leads to \( F_B(\mu) = 2\mu - 1 \) with \( F_B'(\mu) = 2 \) and \( V'(F_B(\mu)) = -4(1 - \mu) \leq 0 \). Thus Conditions \( \mathbf{H} \) and \( \mathbf{I} \) are satisfied and

\[
V(Y^*F_B(\mu)) = \{1 - Y^*F_B(\mu)\}^2 = Y\{1 - F_B(\mu)\}^2 + (1 - Y)\{1 + F_B(\mu)\}^2 = Y(2 - 2\mu)^2 + (1 - Y)(2\mu)^2,
\]

\[
V(Y^*F_B(Y)) = 0,
\]

\[
Q(Y, \mu) = 4\{Y(1 - \mu)^2 + (1 - Y)\mu^2\} = 4(Y - \mu)^2,
\]

\[
q(\mu) = 4\{(1 - \mu)^2 + (1 - \mu)\mu^2\} = 4\mu(1 - \mu).
\]

- \( V(s) = (1 - s)^5 \) given in Breiman (1998): We have that \( V'(s) = -5(1 - s)^4 \leq 0 \). From (6.2), setting \( \mu(1 - F)^4 = (1 - \mu)(1 + F)^4 \) leads to \( F_B(\mu) = \frac{\mu^{1/4} - (1 - \mu)^{1/4}}{(1 - \mu)^{1/4} + (1 - \mu)^{1/4}} \) with \( F_B'(\mu) = \frac{\mu^{3/4} - (1 - \mu)^{3/4}}{(1 - \mu)^{3/4} + (1 - \mu)^{3/4}} \geq 0 \). Thus Conditions \( \mathbf{H} \) and \( \mathbf{I} \) are satisfied and

\[
V(Y^*F_B(\mu)) = Y\{1 - F_B(\mu)\}^5 + (1 - Y)\{1 + F_B(\mu)\}^5 = Y\left\{ \frac{2(1 - \mu)^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5 + (1 - Y)\left\{ \frac{2\mu^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5,
\]

\[
V(Y^*F_B(Y)) = 0,
\]

\[
Q(Y, \mu) = Y\left\{ \frac{2(1 - \mu)^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5 + (1 - Y)\left\{ \frac{2\mu^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5,
\]

\[
q(\mu) = \mu\left\{ \frac{2(1 - \mu)^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5 + (1 - \mu)\left\{ \frac{2\mu^{1/4}}{\mu^{1/4} + (1 - \mu)^{1/4}} \right\}^5.
\]
\( V(s) = \ln(1 + e^{-s}) \) for negative log-likelihood: We have that \( V'(s) = -e^{-s} = -\frac{1}{1 + e^{-s}} \leq 0 \). From (6.2), setting \( \mu \frac{1}{1+e^{-s}} = (1 - \mu) \frac{1}{1+e^{-s}} = (1 - \mu) \frac{e^{s}}{1+e^{s}} \) leads to \( F_B(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) \) with \( F_B(\mu) = \frac{1}{\mu(1-\mu)} \). Thus Conditions \( \text{H} \) and \( \text{I} \) are satisfied and

\[
V(Y^*F_B(\mu)) = \ln\left(1 + e^{-Y^*F_B(\mu)}\right) = Y \ln\left(1 + e^{-F_B(\mu)}\right) + (1 - Y) \ln\left(1 + e^{F_B(\mu)}\right) = -Y \ln(\mu) + (1 - Y) \ln(1 - \mu),
\]

\[
V(Y^*F_B(Y)) = 0,
\]

\[
Q(Y, \mu) = -Y \ln(\mu) + (1 - Y) \ln(1 - \mu),
\]

\[
q(\mu) = -\mu \ln(\mu) + (1 - \mu) \ln(1 - \mu).
\]


\( V(s) = e^{-s} \) for the exponential loss: We have that \( V'(s) = -e^{-s} \leq 0 \). From (6.2), setting \( \mu e^{-F} = (1 - \mu)e^{F} \) leads to \( F_B(\mu) = \frac{1}{\mu(1-\mu)} \) with \( F_B(\mu) = \frac{1}{2\mu(1-\mu)} \). Thus Conditions \( \text{H} \) and \( \text{I} \) are satisfied and

\[
V(Y^*F_B(\mu)) = e^{-Y^*F_B(\mu)} = Ye^{-F_B(\mu)} + (1 - Y) e^{F_B(\mu)} = Y \left\{ \left(1 - \mu\right)/\mu \right\}^{1/2} + (1 - Y) \left\{ \mu/(1 - \mu) \right\}^{1/2},
\]

\[
V(Y^*F_B(Y)) = 0,
\]

\[
Q(Y, \mu) = Y \left\{ \left(1 - \mu\right)/\mu \right\}^{1/2} + (1 - Y) \left\{ \mu/(1 - \mu) \right\}^{1/2},
\]

\[
q(\mu) = \mu \left\{ \left(1 - \mu\right)/\mu \right\}^{1/2} + (1 - \mu) \left\{ \mu/(1 - \mu) \right\}^{1/2} = 2 \left\{ \mu(1 - \mu) \right\}^{1/2}.
\]

\( V(s) = \max(1 - s, 0) \) for the hinge loss used in support vector machine: We have that \( V \) is continuous, convex and \( V'(s) = -I[s < 1] \) for \( s \in \mathcal{F}_2 = \{ s : s \neq 1 \} \). We see that

\[
E\{V(Y^*F|x = x)\} = m(x)(1 - F)I[F \leq 1] + \{1 - m(x)\}(1 + F)I[F \geq -1]
\]

\[
= \begin{cases} 
  m(x)(1 - F), & \text{if } F \leq -1, \\
  1 - \{2m(x) - 1\}F, & \text{if } -1 < F < 1, \\
  \{1 - m(x)\}(1 + F), & \text{if } F \geq 1.
\end{cases}
\]

By a graphical approach, we observe that the desired \( F_B \) to minimize \( E\{V(Y^*F|x = x)\} \) is \( F_B(m(x)) = \text{sgn}\{2m(x) - 1\} \). Thus

\[
V(Y^*F_B(\mu)) = Y \{1 - F_B(\mu)\} I[F_B(\mu) \leq 1] + (1 - Y) \{1 + F_B(\mu)\} I[F_B(\mu) \geq -1]
\]

\[
= Y \{1 - \text{sgn}(2\mu - 1)\} I[\text{sgn}(2\mu - 1) \leq 1] + (1 - Y) \{1 + \text{sgn}(2\mu - 1)\} I[\text{sgn}(2\mu - 1) \geq -1]
\]

\[
= 2\{Y I[2\mu < 1] + (1 - Y) I[2\mu > 1]\},
\]

\[
V(Y^*F_B(Y)) = 0.
\]
Define $Q(Y, \mu) = V(Y^*F_B(\mu))$. Since Condition I does not hold, we can not use Theorem 9. Nonetheless, we can directly check Condition $B'$ in Theorem 3 that $V(Y^*F_B(\mu))$ is a BD with

$$Q(Y, \mu) = 2\{Y 1[2\mu < 1]+(1-Y) 1[2\mu > 1]\}, \quad q(\mu) = 2\{\mu 1[2\mu < 1]+(1-\mu) 1[2\mu > 1]\}.$$

Theorem 7 implies that for the margin-based loss functions $V(Y^*F_B(\mu))$, the asymptotic distribution of $\hat{m}(X)$ 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 $m(x)$ is known.

### 6.3 Optimal smoothing parameter of local estimation under misclassification loss

Classification performance is commonly measured by the misclassification rate. Ideally, we would like to choose the optimal smoothing parameter $h$ for $\hat{m}(x)$ to minimize the true misclassification risk. However, the optimal bandwidth established in Section 5 does not work for the misclassification loss function (3.12), due to the fact that the corresponding $q$ function does not satisfy Condition E in Theorem 7.

We now examine the covariance penalty expression (2.8) for the estimation of the prediction error. Since the misclassification loss function (3.12) is bounded below and above, we first present a general result on the covariance penalty term under bounded BD. Clearly, this result will also be applicable to other margin-based loss functions that are BD and bounded above.

**Definition 2** We say two functions $F$ and $G$ are similarly ordered if $\{F(x_1)-F(x_2)\} \{G(x_1)-G(x_2)\} \geq 0$ for all $x_1$ and $x_2$, and oppositely ordered if the inequality is reversed.

**Theorem 10** For $Y|X = x \sim$ Bernoulli$(m(x))$, suppose that the loss function $Q(Y, \mu)$ is a Bregman divergence. Assume that

- J. $q'(\hat{m}(X_i))$ and $Y_i$ are oppositely ordered;
- K. There exist constants $a$ and $b$ such that $a \leq Q(Y, \mu) \leq b$ holds for all $Y$ and $\mu$.

Then

$$0 \leq \text{cov}(\hat{X}_i, Y_i|X) \leq (b-a)/4.$$

**Remark 1** If $\partial Q(Y, \mu)/\partial \mu$ exists, then a sufficient condition for Condition J is that $\frac{\partial \hat{m}(X_i)}{\partial Y_i} \geq 0$, which is satisfied in almost all the model fitting methods encountered in practice.
For the misclassification loss (3.12), we observe from (3.14) that \( \hat{\lambda}_i = \frac{[\text{sgn}\{\hat{m}(x_i) - c\} - (1 - 2c)]}{4} \), and thus
\[
\text{cov}(\hat{\lambda}_i, Y_i | \mathcal{X}) = \text{cov}[4^{-1}\text{sgn}\{\hat{m}(x_i) - c\}, Y_i | \mathcal{X}] = E[4^{-1}\text{sgn}\{\hat{m}(x_i) - c\}\{Y_i - m(x_i)\}] | \mathcal{X}.
\]
For any function estimates \( \hat{m} \) satisfying Condition J, it follows that \( \hat{m}(x_i) > c \) for \( Y_i = 1 \) and \( \hat{m}(x_i) < c \) for \( Y_i = 0 \), and hence
\[
\text{cov}(\hat{\lambda}_i, Y_i | \mathcal{X}) = \frac{m(x_i)\{1 - m(x_i)\}}{4} - \frac{m(x_i)\{0 - m(x_i)\}}{1/4} = \frac{m(x_i)\{1 - m(x_i)\}}{2}.
\]
Since this part is free of the bandwidth parameter, the optimal bandwidth that minimizes (2.8) will be possibly much larger. Our analysis lends support to the conjecture of Friedman (1997, p. 63). As a consequence, we will not use the misclassification loss function for function estimation.

7 Simulations and Data Analysis

We now consider the varying-coefficient logistic regression for Bernoulli responses, in which varying-coefficient functions in \( \logit\{P(Y = 1|U = u, X = x)\} \) are modeled as

Example 1: \( d = 2, a_1(u) = 1.3\{\exp(2u - 1) - 1.5\}, a_2(u) = 1.2\{8u(1 - u) - 1\} \),

Example 2: \( d = 3, a_1(u) = \exp(2u - 1) - 1.5, a_2(u) = 0.8\{8u(1 - u) - 1\}, a_3(u) = 0.9\{2\sin(\pi u) - 1\} \)

We assume that \( X_1 = 1; X_2 \) and \( X_3 \) are uncorrelated standard normal variables, and are independent of \( U \sim U(0, 1) \). Figure 4 displays the data generated from the model, along with the true decision boundary (black curves). The decision boundary obtained from the local likelihood estimates are denoted by the green curves. The bandwidth is determined by minimizing the cross-validated prediction error shown in Figure 5.

[ Put Figures 4-6 about here ]

8 Discussion

The work can be extended to semiparametric regression model, additive regression model. Statistical inference for parametric and non-parametric estimation under the Bregman divergence can also be conducted.

References


**Appendix: Proofs of Main Results**

We first impose some technical conditions. They are not the weakest possible.

**Condition C:**

C1. Θ is compact in R^{d+1}.

C2. χ is on a compact support Λ, and χ has the design density fχ(·) with fχ(·) > 0.

C3. F is a bijection, with F^{-1}(·) continuously differentiable.

C4. q is twice continuously differentiable with q''(·) < 0.

C5. There does not exist a vector c ≠ 0 such that x^Tc = 0 almost surely.
Condition D:

D1. $\Theta$ is compact in $\mathbb{R}^{d+1}$.

D2. $x$ is on a compact support $\Lambda$, and $x$ has the design density $f_X(\cdot)$ with $f_X(\cdot) > 0$.

D3. $F$ is a bijection, with $F^{-1}(\cdot)$ twice continuously differentiable.

D4. $q''(\cdot)$ is continuous and $q''(\cdot) < 0$.

D5. There does not exist a vector $\overline{c} \neq 0$ such that $\overline{x}^T \overline{c} = 0$ almost surely.

D6. $E(Y^2) < \infty$.

Condition E:

E1. The function $q$ is concave, $q''(m(x)) < 0$ and $q'''(\cdot)$ is continuous at $m(x)$.

E2. Let $q_j(y; \theta) = (\partial^j / \partial \theta^j)Q(y, F^{-1}(\theta))$. Assume that $q_2(y; \theta) > 0$ 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 neighborhood 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 at $x$, and $f_X(x) > 0$.

E6. Both $m(\cdot)$ and $\text{var}(Y|X = \cdot)$ are continuous at $x$, and $\text{var}(Y|X = x) > 0$.

E7. $F$ is a bijection. $F'(m(x)) > 0$ and $F''(\cdot)$ is continuous at $m(x)$. $F^{-1}(\cdot)$ is continuous at $\eta(x)$.

E8. $\eta^{(p+1)}(\cdot)$ is continuous at $x$.

Condition E':

E1'. The function $q$ is concave, $q''(m(u, x)) < 0$ and $q'''(\cdot)$ is continuous at $m(u, x)$.

E2'. Let $q_j(y; \theta) = (\partial^j / \partial \theta^j)Q(y, F^{-1}(\theta))$. Assume that $q_3(y; \theta) > 0$ for $\theta \in \mathbb{R}$ and $y$ in the range of the response variable.

E3'. There exists some $\delta > 0$ such that $E(|Y|^{2+\delta}|U = \cdot, X = x)$ is bounded in a neighborhood of $u$. 

27
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 at $u$, and $f_U(u) > 0$.

E6'. Both $m(\cdot, x)$ and $\text{var}(Y|U = u, X = x)$ are continuous at $u$, and $\text{var}(Y|U = u, X = x) > 0$.

E7'. $F$ is a bijection. $F'(m(u, x)) > 0$ and $F''(\cdot)$ is continuous at $m(u, x)$. $F^{-1}(\cdot)$ is continuous at $\eta(u, x)$.

E8'. $a_j^{(p+1)}(\cdot)$, $j = 1, \ldots, d$, are continuous at $u$.

E9'. $\Gamma(\cdot)$ is continuous at $u$ and $\Gamma(u) > 0$. $\Delta(\cdot)$ is continuous at $u$ and $\Delta(u) > 0$.

**Proof of Theorem 1**

It suffices to show that

$$E\{Q(Y, \mu(X))|X\} \geq E\{Q(Y, m(X))|X\}.$$  

To this end, we perform for a suitably chosen sigma-field $F$ the following decomposition,

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

By setting $F = \sigma(X)$, the sigma-field generated by $X$, and taking the expectation of $Q(Y, \mu(X))$ conditional on $F$, we obtain

$$E\{Q(Y, \mu(X))|X\} = [q(m(X)) - E\{q(Y)|X\}] - q(m(X)) + q(\mu(X)) + \{m(X) - \mu(X)\}q'(\mu(X))$$

$$= E\{Q(Y, m(X))|X\} + Q(m(X), \mu(X))$$

$$\geq E\{Q(Y, m(X))|X\}.$$  

This completes the proof.

**Proof of Theorem 2**

Applying the decomposition strategy in Theorem 1 and setting $F = \sigma\{T, X^o\}$, we have that

$$Q(Y^o, \widehat{m}(X^o)) = [q(m(X^o)) - E\{q(Y^o)|T, X^o\}] + [E\{q(Y^o)|T, X^o\} - q(Y^o)] - q(m(X^o)) + q(\widehat{m}(X^o)) + \{Y^o - \widehat{m}(X^o)\}q'(\widehat{m}(X^o)).$$  

(A.1)
Since \((X^o, Y^o)\) is independent of \(T\), we deduce from Chow and Teicher (1989, Corollary 3, p. 223) that

\[
E\{q(Y^o)|T, X^o\} = E\{q(Y^o)|X^o\}. \tag{A.2}
\]

Similarly,

\[
E\{Y^o q'(\hat{m}(X^o))|T, X^o\} = E\{Y^o|T, X^o\}q'(\hat{m}(X^o)) = E\{Y^o|X^o\}q'(\hat{m}(X^o)) = m(X^o)q'(\hat{m}(X^o)). \tag{A.3}
\]

Applying (A.2) and (A.3) to (A.1) results in

\[
E\{Q(Y^o, \hat{m}(X^o))|T, X^o\} = \left[q(m(X^o)) - E\{q(Y^o)|X^o\}\right] + Q(m(X^o), \hat{m}(X^o))
\]

and thus (2.6), which in turn shows (2.7). The proof is completed.

**Proof of Lemma 1**

If Condition A holds, then \(\mathcal{N}_1 = \mathcal{M}_1, \mathcal{N}_2 = \mathcal{M}_2\), and Condition B will directly follow from (2.2).

Conversely, assume that Condition B holds. Define a function \(q\) by

\[
q(\mu) = \int_a^\mu \left\{ \int_a^t \frac{\partial Q(\nu, s)}{\partial s} \, ds \right\} dt + C \tag{A.4}
\]

for constants \(a\) and \(C\). It follows that \(q(a) = C\) and

\[
q'(\mu) = \int_a^\mu \frac{\partial Q(\nu, s)}{\partial s} \, ds.
\]

According to Condition B, the integrand is free of \(\nu\), so \(q'(\mu)\) is monotone decreasing for \(\mu \in \mathcal{N}_2\). Since \(\mathcal{N}_2\) is an interval except at most countably many points, the concaveness of \(q\) is implied. Note that (3.1) in Condition B guarantees the use of Fubini theorem in (A.4), and thus

\[
q(\mu) = \int_a^\mu \left\{ \int_a^\mu \frac{\partial Q(\nu, s)}{\partial s} \, ds \right\} dt + C = \int_a^\mu \frac{\nu - s - (\nu - \mu)}{\nu - s} \, ds + C
\]

\[
= \int_a^\mu \frac{\partial Q(\nu, s)}{\partial s} \, ds - (\nu - \mu) \int_a^\mu \frac{\partial Q(\nu, s)}{\partial s} \, ds + C
\]

\[
= Q(\nu, \mu) - Q(\nu, a) - (\nu - \mu)q'(\mu) + C;
\]

\[29\]
which is equivalent to
\[ Q(\nu, \mu) = q(\mu) + (\nu - \mu)q'(\mu) + Q(\nu, \mu) - C, \]  
(A.5)

and also indicates, using \( Q(\nu, \nu) = 0 \), the identity
\[ q(\nu) = -Q(\nu, \mu) + C. \]  
(A.6)

Combining (A.5) and (A.6) implies the result desired in Condition A. The proof completes.

**Proof of Theorem 3**

The proofs are similar to those used in Lemma 1, and thus are omitted.

**Proof of Corollary 1**

If Condition A' holds, it follows from (3.4) that for \( \mu \in \mathcal{N}_1 \),
\[ E\{Q(Y, \mu) | X = x\} = -E\{q(Y) | X = x\} + q(\mu) + \{m(x) - \mu\}q'(\mu). \]

Note \( E\{q(Y) | X = x\} \) is independent of \( \mu \). According to the definition, \( \mathcal{E}(\mu; m(x)) = q(\mu) + \{m(x) - \mu\}q'(\mu) \), which implies \( q(\mu) = \mathcal{E}(\mu; \mu) \).

**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
\[ E\{Q(Y, \mu) | X = x\} = m(x)Q(1, \mu) + \{1 - m(x)\}Q(0, \mu), \]

which depends on the conditional distribution of \( Y \) only through \( m(x) \). An application of Corollary 1 yields \( \mathcal{E}(\mu; m(x)) = m(x)Q(1, \mu) + \{1 - m(x)\}Q(0, \mu) \) and hence \( \mathcal{E}(\mu; \mu) = \mu Q(1, \mu) + (1 - \mu)Q(0, \mu) \). The conclusion (3.8) follows from applying (3.7). From (3.8), it follows immediately that \( q(Y) = Q(Y, Y) \) and thus \( q(Y) \equiv 0 \).

To show (3.9), we see from (3.4) that
\[
\begin{align*}
Q(1, \mu) &= -q(1) + q(\mu) + (1 - \mu)q'(\mu), \\
Q(0, \mu) &= -q(0) + q(\mu) + (-\mu)q'(\mu),
\end{align*}
\]

which implies that \( q'(\mu) = Q(1, \mu) - Q(0, \mu) + q(1) - q(0) \), namely, (3.9), due to the fact that \( q(Y) \equiv 0 \). The proof is completed.
Proof of Theorem 4

Define $\hat{x} = (1, x^T)^T$ and $l(y, x; \hat{\alpha}) = Q(y, F^{-1}(\hat{x}^T \hat{\alpha}))$. The target function in (4.1) can be rewritten as $\ell_n(\hat{\alpha}) = n^{-1} \sum_{i=1}^{n} l(Y_i, x_i; \hat{\alpha})$. Then $\hat{\alpha} = \arg \min_{\alpha} \ell_n(\alpha)$. Let $\ell(\hat{\alpha}) = E\{l(Y, x; \hat{\alpha})\}$. According to van der Vaart (1998, p. 45, Theorem 5.7), it suffices to prove two parts:

- Identifiability of $\ell(\hat{\alpha})$: For any $\epsilon > 0$, $\inf_{\alpha: \|\alpha - \alpha^{(0)}\| \geq \epsilon} \ell(\hat{\alpha}) > \ell(\alpha^{(0)})$.

- Uniform convergence of $\ell_n(\hat{\alpha})$: $\sup_{\alpha \in \Theta} |\ell_n(\hat{\alpha}) - \ell(\hat{\alpha})| \xrightarrow{P} 0$.

We first prove the identifiability. Under Condition C1, we only need to prove that for any $\hat{\alpha} \neq \alpha^{(0)}$,

$$\ell(\hat{\alpha}) > \ell(\alpha^{(0)}). \quad (A.7)$$

To see this, note that

$$\ell(\hat{\alpha}) - \ell(\alpha^{(0)}) = E\{l(Y, x; \hat{\alpha}) - l(Y, x; \alpha^{(0)})\}$$

$$= E\{E\{l(Y, x; \hat{\alpha}) - l(Y, x; \alpha^{(0)})|X\}\}. \quad (A.8)$$

From the definition of $Q(\cdot, \cdot)$ in (2.1), we have that

$$E\{l(Y, x; \hat{\alpha}) - l(Y, x; \alpha^{(0)})|X\} = E\{q(F^{-1}(\hat{x}^T \hat{\alpha})) - q(F^{-1}(\hat{x}^T \alpha^{(0)}))\}$$

$$+ \{Y - F^{-1}(\hat{x}^T \hat{\alpha})\}q'(F^{-1}(\hat{x}^T \hat{\alpha})) - \{Y - F^{-1}(\hat{x}^T \alpha^{(0)})\}q'(F^{-1}(\hat{x}^T \alpha^{(0)}))|X\}$$

$$= E\{q(F^{-1}(\hat{x}^T \alpha)) - q(F^{-1}(\hat{x}^T \alpha^{(0)}))\} + \{Y - F^{-1}(\hat{x}^T \alpha)\}q'(F^{-1}(\hat{x}^T \alpha))|X\}$$

$$= q(F^{-1}(\hat{x}^T \alpha)) - q(F^{-1}(\hat{x}^T \alpha^{(0)})) + \{F^{-1}(\hat{x}^T \alpha) - F^{-1}(\hat{x}^T \alpha^{(0)})\}q'(F^{-1}(\hat{x}^T \alpha))$$

$$= Q(F^{-1}(\hat{x}^T \alpha^{(0)}), F^{-1}(\hat{x}^T \alpha)), \quad (A.9)$$

in which the second and third equalities follow from the fact $E(Y|X) = F^{-1}(\hat{x}^T \alpha^{(0)})$. Putting (A.9) to (A.8), we observe that $\ell(\hat{\alpha}) - \ell(\alpha^{(0)}) = E\{Q(F^{-1}(\hat{x}^T \alpha^{(0)}), F^{-1}(\hat{x}^T \alpha))\}$. From Conditions C3, C4 and C5, which imply that $F$ is a bijection, $q$ is strictly concave, and $\hat{x}^T \alpha^{(0)} \neq \hat{x}^T \alpha$ on some positive probability set when $\hat{\alpha} \neq \alpha^{(0)}$, it follows that $Q(F^{-1}(\hat{x}^T \alpha^{(0)}), F^{-1}(\hat{x}^T \alpha)) > 0$ on some positive probability set, when $\hat{\alpha} \neq \alpha^{(0)}$. By Condition C2, we complete the proof of (A.7), and consequently the proof of identifiability.

Next, we show the uniform convergence. By the weak law of large numbers, it is trivially known that $\ell_n(\hat{\alpha}) \xrightarrow{P} \ell(\hat{\alpha})$. If in addition, we can prove that $\{\ell_n(\hat{\alpha})\}$ is stochastically equicontinuous, and $\{\ell(\hat{\alpha})\}$ is equicontinuous, then uniform convergence will readily hold (see, for example, Newey, 1991). To show this, it suffices to prove that there exists $b(y, x)$, which does not depend on $\hat{\alpha}$, such that $E\{b(Y, X)\} < \infty$, and

$$|l(y, x; \hat{\alpha}_1) - l(y, x; \hat{\alpha}_2)| \leq b(y, x)\|\hat{\alpha}_1 - \hat{\alpha}_2\|, \quad (A.10)$$
for any $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ in $\Theta$. Now we show (A.10). Using the definition,

$$
|l(y, x; \tilde{\alpha}_1) - l(y, x; \tilde{\alpha}_2)|
= |q(F^{-1}(\tilde{x}^T \tilde{\alpha}_1)) - q(F^{-1}(\tilde{x}^T \tilde{\alpha}_2)) + q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_1)\}
- q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\}|
\leq |q(F^{-1}(\tilde{x}^T \tilde{\alpha}_1)) - q(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))| + |q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_1)\}|
- q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\}| = I_1 + I_2.
$$

For $I_1$, there exists some $\lambda \in (0, 1)$, such that

$$
I_1 = |(q \circ F^{-1})(\lambda \tilde{x}^T \tilde{\alpha}_1 + (1 - \lambda)\tilde{x}^T \tilde{\alpha}_2)(\tilde{x}^T \tilde{\alpha}_1 - \tilde{x}^T \tilde{\alpha}_2)|
\leq |(q \circ F^{-1})(\lambda \tilde{x}^T \tilde{\alpha}_1 + (1 - \lambda)\tilde{x}^T \tilde{\alpha}_2)| \cdot \|\tilde{x}\| \cdot \|\tilde{\alpha}_1 - \tilde{\alpha}_2\|.
$$

By the continuity of $q'(\cdot)$ and $(F^{-1})'(\cdot)$, and the compactness of $\Theta$ and $\Lambda$, we conclude that

$$
I_1 \leq C_1\|\tilde{\alpha}_1 - \tilde{\alpha}_2\|,
$$

where $C_1$ is some constant. For $I_2$, it is observed that

$$
I_2 \leq |q(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_1)\} - q(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\}|
+ |q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\} - q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))\{y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\}|
= I_{2,1} + I_{2,2},
$$

in which there exists some $\theta \in (0, 1)$ such that

$$
I_{2,1} = |q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))\{F^{-1}(\tilde{x}^T \tilde{\alpha}_1) - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)\}|
\leq |q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1))| \cdot |(F^{-1})'(\theta \tilde{x}^T \tilde{\alpha}_1 + (1 - \theta)\tilde{x}^T \tilde{\alpha}_2)| \cdot \|\tilde{x}\| \cdot \|\tilde{\alpha}_1 - \tilde{\alpha}_2\|.
$$

Thus, we obtain

$$
I_{2,1} \leq C_2\|\tilde{\alpha}_1 - \tilde{\alpha}_2\|,
$$

where $C_2$ is some constant. For $I_{2,2}$, direct calculations give that

$$
I_{2,2} \leq |y - F^{-1}(\tilde{x}^T \tilde{\alpha}_2)| \cdot |q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_1)) - q'(F^{-1}(\tilde{x}^T \tilde{\alpha}_2))|.
$$

Applying arguments similar to those used for $I_1$, we conclude that

$$
I_{2,2} \leq (C_4|y| + C_5)\|\tilde{\alpha}_1 - \tilde{\alpha}_2\|,
$$

where both $C_3$ and $C_4$ are constants. By (A.12) and (A.13), it follows that

$$
I_2 \leq (C_4|y| + C_3 + C_2)\|\tilde{\alpha}_1 - \tilde{\alpha}_2\|.
$$

Combining (A.11) and (A.14) verifies (A.10), in which $b(y, x)$ equals $C_4|y| + C_3 + C_2 + C_1$ and satisfies $E\{b(Y, X)\} < \infty$ by Condition C6. The proof is completed.
Proof of Theorem 5

Following the notations used in the proof of Theorem 4, \( \hat{\alpha} \xrightarrow{P} \alpha^{(0)} \). To further establish the asymptotic normality of \( \hat{\alpha} \), we only need to prove the following three statements (van der Vaart, 1998, p. 53, Theorem 5.23):

1. \( (y, x) \mapsto l(y, x; \alpha) \) is measurable for each \( \alpha \in \Theta \); and \( \alpha \mapsto l(y, x; \alpha) \) is differentiable at \( \alpha^{(0)} \) for almost every \( (y, x) \).

2. There exists a measurable function \( b(y, x) \) with \( E\{b^2(Y, X)\} < \infty \), such that

\[
|l(y, x; \alpha_1) - l(y, x; \alpha_2)| \leq b(y, x)\|\alpha_1 - \alpha_2\|
\]

for every \( \alpha_1 \) and \( \alpha_2 \) in a neighborhood of \( \alpha^{(0)} \).

3. \( \alpha \mapsto E\{l(Y, X; \alpha)\} \) admits a second-order Taylor expansion at \( \alpha^{(0)} \) with a nonsingular symmetric second derivative matrix \( H_0 \).

The first statement is straightforward under Conditions D3 and D4. Recall that (A.10) has been shown in the proof of Theorem 4 with \( b(y, x) = C_1|y| + C_3 + C_2 + C_1 \), thus the second statement becomes trivial under Condition D6.

We then show the third statement. Since \( l(y, x; \alpha) \) is twice differentiable in \( \alpha \), if we can prove the interchangeability of differentiation and expectation, then

\[
\frac{\partial^2}{\partial \alpha \partial \alpha^T} E\{l(Y, X; \alpha)\} = E\left\{ \frac{\partial^2}{\partial \alpha \partial \alpha^T} l(Y, X; \alpha) \right\}.
\]

The interchangeability is valid if \( \frac{\partial}{\partial \alpha} l(y, x; \alpha) \) and \( \frac{\partial^2}{\partial \alpha \partial \alpha} l(y, x; \alpha) \) are bounded by some integrable functions independent of \( \alpha \). To see this, we first introduce some necessary notations. Define \( q_j(y; \theta) = (\partial^j / \partial \theta^j)Q(y, F^{-1}(\theta)) \). Let \( \theta = F(\mu) \), which implies \( \mu = F^{-1}(\theta) \) and \( d\theta / d\mu = F'(\mu) \). Direct calculations via (2.2)–(2.3) give that

\[
q_1(y; \theta) = \frac{\partial Q(y, \mu)}{\partial \theta} = \frac{\partial Q(y, \mu)}{\partial \mu} \frac{\partial \mu}{\partial \theta} = \frac{\partial Q(y, \mu)}{\partial \mu} \frac{1}{F'(\mu)}
\]

\[
= \frac{q''(\mu)(y - \mu)}{F'(\mu)}, \tag{A.16}
\]

and

\[
q_2(y; \theta) = \frac{d}{d\theta} \left\{ \frac{q''(\mu)(y - \mu)}{F'(\mu)} \right\} = \frac{d}{d\mu} \left\{ \frac{q''(\mu)(y - \mu)}{F'(\mu)} \right\} \frac{d\mu}{d\theta}
\]

\[
= \frac{\{q''(\mu)(y - \mu) - q''(\mu)(y - \mu)\} F'(\mu) - q''(\mu)(y - \mu) F''(\mu)}{\{F'(\mu)^2\}^2} \frac{1}{F'(\mu)}
\]

\[
= -q''(\mu)F'(\mu) + \frac{y - \mu}{\{F'(\mu)^3\}} \left\{ q''(\mu)F'(\mu) - q''(\mu)F''(\mu) \right\}.
\tag{A.17}
\]

33
Accordingly, \( q_j(y; \theta) \) is linear in \( y \) for fixed \( \theta \). Notice that from (A.16),

\[
\frac{\partial}{\partial \alpha} l(y, x; \tilde{\alpha}) = q_1(y; \tilde{x}^T \tilde{\alpha}) \tilde{x} = \{q''(F^{-1}(\tilde{x}^T \tilde{\alpha}))\} \{F'(F^{-1}(\tilde{x}^T \tilde{\alpha}))\}^{-1} \{y - F^{-1}(\tilde{x}^T \tilde{\alpha})\} \tilde{x},
\]

and from (A.17),

\[
\frac{\partial^2}{\partial \alpha \partial \alpha^T} l(y, x; \tilde{\alpha}) = q_2(y; \tilde{x}^T \tilde{\alpha}) \tilde{x} \tilde{x}^T = \{y - F^{-1}(\tilde{x}^T \tilde{\alpha})\} \{q'''(F^{-1}(\tilde{x}^T \tilde{\alpha}))F'(F^{-1}(\tilde{x}^T \tilde{\alpha}))\} - q''(F^{-1}(\tilde{x}^T \tilde{\alpha})) \{F'(F^{-1}(\tilde{x}^T \tilde{\alpha}))\}^{-3} \tilde{x} \tilde{x}^T - q''(F^{-1}(\tilde{x}^T \tilde{\alpha})) \{F'(F^{-1}(\tilde{x}^T \tilde{\alpha}))\}^{-2} \tilde{x} \tilde{x}^T.
\]

Applying similar arguments used in the proof of Theorem 4 to (A.18)–(A.19), we can see that both \( \frac{\partial}{\partial \alpha} l(y, x; \tilde{\alpha}) \) and \( \frac{\partial^2}{\partial \alpha \partial \alpha^T} l(y, x; \tilde{\alpha}) \) are bounded by some functions independent of \( \tilde{\alpha} \), which are integrable. Thus, the interchangeability is valid and (A.15) holds. By (A.15) and (A.19),

\[
H_0 = \left[ \frac{\partial^2}{\partial \alpha \partial \alpha^T} E\{l(Y, x; \tilde{\alpha})\} \right]_{\tilde{\alpha} = \tilde{\alpha}^{(0)}} = E\{\frac{\partial^2}{\partial \alpha \partial \alpha^T} l(Y, x; \tilde{\alpha})\}_{\tilde{\alpha} = \tilde{\alpha}^{(0)}} = - E\{q''(m(x))\} \{F'(m(x))\}^{-2} \tilde{x} \tilde{x}^T.
\]

The last step is to prove that \( H_0 \) is nonsingular. This can easily be verified by Conditions D2, D3 and D4. Thus the third statement is proved.

Now we are ready to deduce that \( \sqrt{n}\{\hat{\alpha} - \tilde{\alpha}^{(0)}\} \overset{d}{\rightarrow} N(0, H_0^{-1} \Omega_0 H_0^{-1}) \), where

\[
\Omega_0 = E\left[ \left\| \frac{\partial}{\partial \alpha} l(Y, x; \tilde{\alpha}) \right\|_{\tilde{\alpha} = \tilde{\alpha}^{(0)}}^2 \right].
\]

By (A.18), we get

\[
\Omega_0 = E\{q''(F^{-1}(\tilde{x}^T \tilde{\alpha}^{(0)}))\}^2 \{F'(F^{-1}(\tilde{x}^T \tilde{\alpha}^{(0)}))\}^{-2} \{y - F^{-1}(\tilde{x}^T \tilde{\alpha}^{(0)}\}^2 \tilde{x} \tilde{x}^T = E\{\text{var}(Y|x)q''(m(x))\} \{F'(m(x))\}^{-2} \tilde{x} \tilde{x}^T.
\]

This concludes the proof.

**Proof of Theorem 6**

Before showing Theorem 6, we need the following lemma.

**Lemma 5** For appropriately dimensioned random matrices \( A \) and \( B \), assume that \( E(BB^T) \) is positive definite. Then

\[
E(AA^T) \geq E(AB^T)\{E(BB^T)\}^{-1} E(BA^T).
\]

Moreover, if \( B = cA \) for a constant \( c \neq 0 \), then the inequality becomes an equality.
Proof: Let $C = A - E(AB^T)\{E(BB^T)\}^{-1}B$. Then $E(CB^T) = 0$. Thus

$$E(CC^T) = E([A - E(AB^T)\{E(BB^T)\}^{-1}B]A^T) = E(AA^T) - E(AB^T)\{E(BB^T)\}^{-1}E(BA^T),$$

which yields the matrix inequality and equality.

To show Theorem 6, define random matrices $A = \{\text{var}(Y|x)\}^{-1/2}\{F'(m(x))\}^{-1}\bar{x}$ and $B = -\text{var}(Y|x)q''(m(x))A$. Then $H_0 = E(AB^T) = E(BA^T)$ and $\Omega_0 = E(BB^T)$. Employing Lemma 5,

$$H_0^{-1}\Omega_0 H_0^{-1} \leq \left\{ E(AA^T) \right\}^{-1} = \left( E[1/\text{var}(Y|x)\{F'(m(x))\}^{-2}\bar{x}\bar{x}^T] \right)^{-1},$$

and the inequality becomes an equality when $q''(m(x)) = -c/\text{var}(Y|x = x)$ for a constant $c > 0$. This completes the proof.

Proof of Theorem 7

Define $\rho(x) = -q''(m(x))/\{F'(m(x))\}^2$. Notice that from (A.16) and (A.17),

$$q_1(m(x); \eta(x)) = 0, \quad q_2(m(x); \eta(x)) = -\frac{q''(m(x))}{\{F'(m(x))\}^2} = \rho(x).$$

(A.20)

It is clearly seen that

$$\frac{\partial}{\partial \beta} Q(y, F^{-1}(x^T\beta)) = q_1(y; x^T\beta)x,$$

$$\frac{\partial^2}{\partial \beta \partial \beta^T} Q(y, F^{-1}(x^T\beta)) = q_2(y; x^T\beta)xx^T.$$

These two identities along with (5.1) indicate that at the fitting point $x$,

$$\frac{\partial \ell_n(\beta; x)}{\partial \beta} = n^{-1}\sum_{i=1}^{n} q_1(Y_i; x_i(x)^T\beta)x_i(x)K_h(X_i - x),$$

$$\frac{\partial^2 \ell_n(\beta; x)}{\partial \beta \partial \beta^T} = n^{-1}\sum_{i=1}^{n} q_2(Y_i; x_i(x)^T\beta)x_i(x)x_i(x)^TK_h(X_i - x).$$

Next, we derive the asymptotic representation of $\sqrt{n}hH\{\hat{\beta}(x) - \beta(x)\} \equiv \hat{\beta}^*(x)$, so that each component has a nondegenerate asymptotic variance. Recall that $\hat{\beta}(x)$ minimizes $\ell_n(\beta; x)$ in (5.1) with respect to $\beta$. Set $\eta_i(x) = x_i(x)^T\beta(x), i = 1, \ldots, n$. Since

$$x_i(x)^T\hat{\beta}(x) = x_i(x)^T\beta(x) + x_i(x)^T\{\hat{\beta}(x) - \beta(x)\}$$

$$= \eta_i(x) + \frac{1}{\sqrt{nh}}\{H^{-1}x_i(x)\}^T[H^{1/2}\{\hat{\beta}(x) - \beta(x)\}]$$

$$\equiv \eta_i(x) + a_nZ_i^T\hat{\beta}^*(x),$$

35
where \( a_n = 1/\sqrt{n} \) and \( Z_i = H^{-1}x_i(x) \), it can easily be seen that \( \hat{\beta}(x) \) minimizes

\[
\ell_n(x; \beta^*) = \sum_{i=1}^{n} \left[ Q(Y_i, F^{-1}(\eta_i(x) + a_nZ_i^T\beta^*)) - Q(Y_i, F^{-1}(\eta_i(x))) \right] K((X_i - x)/h),
\]
as a function of \( \beta^* \). Condition E2 implies that \( \ell_n(\beta^*; x) \) is convex in \( \beta^* \). Using the Taylor expansion, for small \( \Delta \),

\[
Q(y, F^{-1}(\eta + \Delta)) - Q(y, F^{-1}(\eta)) = q_1(y; \eta)\Delta + q_2(y; \eta)\Delta^2/2 + q_3(y; \eta)\Delta^3/6,
\]
we find that

\[
\ell_n^0(\beta^*; x) = \sum_{i=1}^{n} \left[ q_1(Y_i; \eta_i(x))a_nZ_i^T\beta^* + q_2(Y_i; \eta_i(x))a_n^2\beta^*Z_i^Tz_i^T\beta^*/2 + q_3(Y_i; \eta_i^*(x))(Z_i^T\beta^*)^3/6 \right] K((X_i - x)/h)
\]

\[
= W_n^T\beta^* + \frac{1}{2}A_n^T\beta^* + \frac{a_n^3}{6}\sum_{i=1}^{n} q_3(Y_i; \eta_i^*(x))(Z_i^T\beta^*)^3K((X_i - x)/h), \tag{A.21}
\]

where \( \eta_i^*(x) \) is between \( \eta_i(x) \) and \( \eta_i(x) + a_nZ_i^T\beta^* \),

\[
W_n = a_n\sum_{i=1}^{n} q_1(Y_i; \eta_i(x))Z_iK((X_i - x)/h)
\]

\[
= \sqrt{nh}\frac{1}{n}\sum_{i=1}^{n} q_1(Y_i; \eta_i(x))Z_iK_h(x_i - x) = \sqrt{nh}H^{-1}\frac{\partial \ell_n(\beta; x)}{\partial \beta},
\]

and

\[
A_n = a_n^2\sum_{i=1}^{n} q_2(Y_i; \eta_i(x))Z_iZ_i^T\beta^* = \frac{1}{n}\sum_{i=1}^{n} q_2(Y_i; \eta_i(x))Z_iZ_i^T\beta^* = H^{-1}\frac{\partial^2 \ell_n(\beta; x)}{\partial \beta \partial \beta^T}H^{-1}.
\]

Set \( A = \rho(x)f_X(x) > 0 \). It can easily be shown that

\[
A_n = A + o_P(1), \tag{A.22}
\]

using the mean-variance decomposition,

\[
(A_n)_{ij} = E(A_n)_{ij} + O_P(\{\text{var}(A_n)_{ij}\})^{1/2}.
\]

Applying (A.17) and (A.20), the mean and variance in the above expression equal

\[
E(A_n) = E(q_2(Y_i; \eta_i(x))Z_iZ_i^T\beta^*K_h(x_i - x))
\]

\[
= E(q_2(m(x_1); \eta_i(x))Z_iZ_i^T\beta^*K_h(x_i - x))
\]

\[
= \rho(x)f_X(x) + o(1),
\]

36
and

\[ \text{var}(A_n) = n^{-1} \text{var}\{q_2(Y_1; \eta_1(x))Z_1Z_1^TK_h(X_1 - x)\} \]
\[ \leq n^{-1} E[q_2^2(Y_1; \eta_1(x))Z_1Z_1^TK_h(X_1 - x)]^2 = O\left(\frac{1}{n^2h}\right) = o(1), \]

respectively. This verifies (A.22). Note that for each fixed \( \beta^* \) the last term in (A.21) is \( o_P(1) \), since

\[ O\left(\frac{n}{h^a_n}E\{q_3(Y_i; \eta_i(x))\beta^*^3K((X_i - x)/h)\}\right) = O(nh_a^3) = O(a_n). \]

This together with (A.21) and (A.22) leads to

\[ \ell_n^p(\beta^*; x) = W_n^T\beta^* + \frac{1}{2} \beta^*^T A\beta^* + o_P(1). \]

Moreover, we observe that

\[ E(W_n) = \sqrt{n}hE\{q_1(Y_1; \eta_1(x))Z_1K_h(X_1 - x)\} \]
\[ = \sqrt{n}hE\{q_1(m(X_1); x_1(x)^T\beta(x))H^{-1}x_1(x)K_h(X_1 - x)\} \]
\[ = \sqrt{n}h \int (1, \ldots, \{y - x\}/h)^p K_h(y - x)q_1 \left( m(y); \sum_{j=0}^p (y - x)^j \beta_j(x) \right) f_X(y)dy \]
\[ = \sqrt{n}h \int (1, \ldots, t^p) K(t)q_1 \left( m(x + ht); \sum_{j=0}^p (ht)^j \beta_j(x) \right) f_X(x + ht)dt. \]

Note \( \eta(x + ht) = \sum_{j=0}^p (ht)^j \beta_j(x) + (ht)^{p+1} \beta_{p+1}(x) + o(h^{p+1}) \). By (A.20) and Taylor expansion, for small \( h \),

\[ q_1 \left( m(x + ht); \sum_{j=0}^p (ht)^j \beta_j(x) \right) = q_1(m(x + ht); \eta(x + ht)) \]
\[ + \left\{ \sum_{j=0}^p (ht)^j \beta_j(x) - \eta(x + ht) \right\} q_2(m(x + ht); \eta(x + ht)) + o(h^{p+1}) \]
\[ = -(ht)^{p+1} \beta_{p+1}(x) \rho(x + ht) + o(h^{p+1}). \]

Thus we obtain

\[ E(W_n) = -\sqrt{n}h\rho(x)h^{p+1} \beta_{p+1}(x) f_X(x) \int t^{p+1}(1, \ldots, t^p) K(t)dt\{1 + o(1)\} \]
\[ = -\sqrt{n}h c_p\rho(x) \beta_{p+1}(x) f_X(x) h^{p+1}\{1 + o(1)\}, \]

(A.23)

and analogously,

\[ \text{var}(W_n) = \frac{nh}{n} \text{var}\{q_1(Y_1; \eta_1(x))Z_1K_h(X_1 - x)\} \]
\[ = h \text{var}\{q_1(Y_1; \eta_1(x))Z_1K_h(X_1 - x)\} \]
\[ = h[E\{q_1^2(Y_1; \eta_1(x))Z_1Z_1^T(K_h)^2(X_1 - x)\} - O(h^{2p+2})]. \]
Recall $\eta(x) = x_i(x)^T \beta(x)$. Setting $F^{-1}(\eta(x)) = m_i(x)$ leads to
\[
E\{q_1^2(Y_1; \eta_1(x))Z_iZ_i^T(K_h)^2(X_1 - x)\} \\
= E\left[ \left\{ \frac{q''(m_1(x))}{F'(m_1(x))} \right\} \frac{Y_1 - m_1(x)}{m_1(x)} \right]^2 Z_iZ_i^T(K_h)^2(X_1 - x) \\
= E\left[ \left\{ \frac{q''(m_1(x))}{F'(m_1(x))} \right\} \frac{Y_1 - m_1(x) + m(X_1) - m_1(x)}{m_1(x)} \right]^2 Z_iZ_i^T(K_h)^2(X_1 - x) \\
= E\left[ \left\{ \frac{q''(m_1(x))}{F'(m_1(x))} \right\} \frac{Y_1 - m_1(x) + m(X_1)}{m_1(x)} \right]^2 \left\{ \frac{m(X_1) - m_1(x)}{m_1(x)} \right\}^2 Z_iZ_i^T(K_h)^2(X_1 - x) \\
= \left\{ \frac{q''(m_1(x))}{F'(m_1(x))} \right\}^2 \text{var}(Y|X = x) \frac{1}{h} f_X(x) S^* \{1 + o(1)\} \\
= S^* \rho^2(x) v(x) f_X(x) \frac{1}{h} \{1 + o(1)\},
\]
which gives
\[
\text{var}(W_n) = S^* \rho^2(x) v(x) f_X(x) + o(1) + O(h^{2p+3}) \equiv B + o(1),
\]
(A.24)
where $B = S^* \rho^2(x) v(x) f_X(x) > 0$. Thus the assumption on $n$ and $h$ conclude $W_n = O_p(1)$. By the quadratic approximation lemma (Fan and Gijbels, 1996, p. 210, and Hjort and Pollard, 1996), it follows that
\[
\hat{\beta}^* (x) = -A^{-1} W_n + o_p(1).
\]
(A.25)

To demonstrate the asymptotic joint normality of $W_n$,
\[
\{\text{var}(W_n)\}^{-1/2} \{W_n - E(W_n)\} \xrightarrow{L} N(0, I_{p+1}),
\]
(A.26)
it suffices to show by the Cramer-Wold device that for any unit vector $a$,
\[
\frac{a^T \{W_n - E(W_n)\}}{\{a^T \text{var}(W_n) a\}^{1/2}} \xrightarrow{L} N(0, 1).
\]

We now check the Lyapounov condition which requires the existence of $\delta > 0$, such that
\[
\frac{n E\{\{a_n q_1(Y_1; \eta_1(x)) (a_i^T Z_i) K((X_1 - x)/h)^{2+\delta}\}^2\}}{\{\text{var}(a^T W_n)\}^{(2+\delta)/2}} \rightarrow 0.
\]
For $\delta$ assumed in Condition E3, the numerator becomes
\[
n(a_n h)^{2+\delta} E\{\{q_1^{2+\delta}(Y_1; \eta_1(x)) (a_i^T Z_i)^{2+\delta}(K_h)^{2+\delta}(X_1 - x)\}\}
\]
\[= n(a_n h)^{2+\delta} \frac{1}{h^{1+\delta}} = n a_n^{2+\delta} h = n h \left( \frac{1}{\sqrt{nh}} \right)^{2+\delta} = \frac{nh}{(nh)^{1+\delta/2}} \rightarrow 0,
\]
from which the Lyapounov condition is satisfied.
Finally, combining (A.25), (A.23), (A.24) and (A.26), we observe that

\[
\hat{\beta}^* (x) = -A^{-1} [\text{var}(W_n)^{1/2} \text{var}(W_n)^{-1/2} \{W_n - E(W_n)\} + E(W_n)] + o_P(1)
\]

\[
= -A^{-1} \{B^{1/2} N(0, I) + E(W_n)\} + o_P(1)
\]

\[
= N(0, A^{-1}BA^{-1}) - A^{-1}E(W_n) + o_P(1),
\]

in which the asymptotic covariance matrix is

\[
A^{-1}BA^{-1} = \frac{1}{\rho^2(x)f_X(x)} S^{-1} S^* \rho^2(x) v(x) f_X(x) S^{-1} = S^{-1} S^* S^{-1} v(x) / f_X(x),
\]

and the asymptotic bias vector is

\[
-A^{-1}E(W_n) = \frac{1}{\rho(x)f_X(x)} S^{-1} \sqrt{n} h c_p \rho(x) \beta_{p+1} (x) f_X(x) h^{p+1} \{1 + o(1)\}
\]

\[
= \sqrt{n} h S^{-1} c_p \beta_{p+1} (x) h^{p+1} \{1 + o(1)\}.
\]

This completes the proof.

\section*{Proof of Theorem 8}

Define \( \rho(u, x) = -q''(m(u, x))/\{F'(m(u, x))\}^2 \). Notice that from (A.16) and (A.17),

\[
q_1(m(u, x), \eta(u, x)) = 0, \quad q_2(m(u, x), \eta(u, x)) = -\frac{q''(m(u, x))}{\{F'(m(u, x))\}^2} = \rho(u, x). \quad (A.27)
\]

It is clearly seen that

\[
\frac{\partial}{\partial \beta} Q(Y_1, F^{-1}(\{u_1(u) \otimes x_1\}^T \beta)) = q_1(Y_1; \{u_1(u) \otimes x_1\}^T \beta) \{u_1(u) \otimes x_1\},
\]

\[
\frac{\partial^2}{\partial \beta \partial \beta^T} Q(Y_1, F^{-1}(\{u_1(u) \otimes x_1\}^T \beta)) = q_2(Y_1; \{u_1(u) \otimes x_1\}^T \beta) \{u_1(u) u_1(u)^T \otimes (x_1 x_1^T)\}.
\]

These two identities along with (5.3) indicate that at the fitting point \( u \),

\[
\frac{\partial \ell_n(\beta; u)}{\partial \beta} = n^{-1} \sum_{i=1}^n q_1(Y_i; \{u_i(u) \otimes x_i\}^T \beta) \{u_i(u) \otimes x_i\} K_h(U_i - u),
\]

\[
\frac{\partial^2 \ell_n(\beta; u)}{\partial \beta \partial \beta^T} = n^{-1} \sum_{i=1}^n q_2(Y_i; \{u_i(u) \otimes x_i\}^T \beta) \{u_i(u) u_i(u)^T \otimes (x_i x_i^T)\} K_h(U_i - u).
\]

Next, we derive the asymptotic representation of \( \sqrt{n} h H \{\hat{\beta}(u) - \beta(u)\} \equiv \hat{\beta}^*(u) \), so that each component has a nondegenerate asymptotic variance. Recall that \( \hat{\beta}(u) \) minimizes \( \ell_n(\beta; u) \) in (5.3) with respect to \( \beta \). Set \( \eta_i(u) = \{u_i(u) \otimes x_i\}^T \beta(u), \ i = 1, \ldots, n \). Since

\[
\{u_i(u) \otimes x_i\}^T \hat{\beta}(u) = \{u_i(u) \otimes x_i\}^T \beta(u) + \{u_i(u) \otimes x_i\}^T \{\hat{\beta}(u) - \beta(u)\}
\]

\[
= \eta_i(u) + \frac{1}{\sqrt{n} h} [H^{-1} u_i(u) \otimes x_i]^T [\sqrt{n} h H \{\hat{\beta}(u) - \beta(u)\}]
\]

\[
\equiv \eta_i(u) + a_i Z_i \hat{\beta}^*(u),
\]

39
where \( a_n = 1/\sqrt{n\hbar} \) and \( Z_i = \{H^{-1}u_i(u)\} \otimes x_i \), it can easily be seen that \( \hat{\beta}(u) \) minimizes

\[
E_n(\beta^*; u) = \sum_{i=1}^n \left[ Q\{Y_i, F^{-1}(\eta_i(u) + a_nZ_i^T\beta^*)\} - Q\{Y_i, F^{-1}(\eta_i(u))\} \right] K((U_i - u)/h),
\]
as a function of \( \beta^* \). Condition E2' implies that \( E_n(\beta^*; u) \) is convex in \( \beta^* \). Using the Taylor expansion, for small \( \Delta \),

\[
Q(y, F^{-1}(\eta + \Delta)) - Q(y, F^{-1}(\eta)) = q_1(y; \eta)\Delta + q_2(y; \eta)\Delta^2/2 + q_3(y; \eta^*)\Delta^3/6,
\]
we find that

\[
E_n(\beta^*; u) = \sum_{i=1}^n \left[ q_1(Y_i; \eta_i(u))a_nZ_i^T\beta^* + q_2(Y_i; \eta_i(u))a_n^2\beta^*Z_iZ_i^T\beta^*/2 
+ q_3(Y_i; \eta_i^*(u))a_n^3(Z_i^T\beta^*)^3/6 \right] K((U_i - u)/h)
= W_n^T\beta^* + \frac{1}{2}\beta^*A_n\beta^* + \frac{a_n^3}{6} \sum_{i=1}^n q_3(Y_i; \eta_i^*(u))(Z_i^T\beta^*)^3K((U_i - u)/h), \quad (A.28)
\]

where \( \eta_i^*(u) \) is between \( \eta_i(u) \) and \( \eta_i(u) + a_nZ_i^T\beta^* \),

\[
W_n = a_n \sum_{i=1}^n q_1(Y_i; \eta_i(u))Z_iK((U_i - u)/h)
= \sqrt{n\hbar} \times \frac{1}{n} \sum_{i=1}^n q_1(Y_i; \eta_i(u))Z_iK_h(U_i - u) = \sqrt{n\hbar}H^{-1}\partial E_n(\beta; u)/\partial \beta,
\]
and

\[
A_n = a_n^2 \sum_{i=1}^n q_2(Y_i; \eta_i(u))Z_iZ_i^TK((U_i - u)/h)
= \frac{1}{n} \sum_{i=1}^n q_2(Y_i; \eta_i(u))Z_iZ_i^TK_h(U_i - u) = H^{-1}\partial^2 E_n(\beta; u)/\partial \beta \partial \beta^T H^{-1}.
\]

Set \( A = \{S \otimes \Gamma(u)\} f_U(u) > 0 \). It can easily be shown that

\[
A_n = A + o_p(1), \quad (A.29)
\]
using the mean-variance decomposition,

\[
(A_n)_{ij} = E(A_n)_{ij} + O_p(\{\text{var}(A_n)_{ij}\}^{1/2}).
\]

Applying (A.17) and (A.27), the mean and variance in the above expression equal

\[
E(A_n) = E\{q_2(Y_i; \eta_i(u))Z_iZ_i^TK_h(U_i - u)\}
= E\{q_2(m(U, x); \eta_i(u))Z_iZ_i^TK_h(U_i - u)\}
= [S \otimes E\{\rho(u, x)x x^T|U = u\}] f_U(u) + o(1),
\]

40
and

\[
\text{var}(A_n) = \frac{1}{n} \text{var}\{q_2(Y; \eta(u))Z_1Z_1^T K_h(U_1 - u)\} \\
\leq \frac{1}{n} E\{q_2^2(Y; \eta(u))Z_1Z_1^T Z_1^T \{K_h(U_1 - u)\}^2\} = O\left(\frac{1}{nh}\right) = o(1),
\]

respectively. This verifies (A.29). Note that for each fixed \( \beta^* \) the last term in (A.28) is \( o_P(1) \), since

\[
O\left(\frac{n}{n} \text{var}\{q_3(Y; \eta^*(u))(Z_i^T \beta^*)^3 K((U_1 - u)/h)\}\right) = O(na_n^3) = O(a_n).
\]

This together with (A.28) and (A.29) leads to

\[
\ell_n^o(\beta^*; u) = \mathbf{w}_n^T \beta^* + \frac{1}{2} \beta^T A \beta^* + o_P(1).
\]

Moreover, we observe that

\[
E(W_n) = \sqrt{nh} E\{q_1(Y; \eta(u))Z_1K_h(U_1 - u)\} \\
= \sqrt{nh} E\{q_1(m(U_1, X_1); \{u_1(u) \otimes \chi_1\})^T \beta(u)\} \{H^{-1}u_1(u) \otimes \chi_1\}K_h(U_1 - u) \\
= \sqrt{nh} E\{[H^{-1}u_1(u)K_h(U_1 - u)] \otimes E\{X_1q_1(m(U_1, \chi_1); \{u_1(u) \otimes \chi_1\})^T \beta(u)|U_1\}\} \\
= \sqrt{nh} \int \{1, \ldots, (ht)^p\}^T K(t) \otimes \\
E\{X_1q_1(m(u + ht, \chi_1); \{1, \ldots, (ht)^p\} \otimes \chi_1)^T \beta(u)|U_1 = u + ht\} f_U(u + ht) dt.
\]

Note \( \eta(u + ht, \chi_1) = \{(1, \ldots, (ht)^p) \otimes \chi_1\}^T \beta(u) + (ht)^{p+1}x_1^T A^{(p+1)}(u)/(p + 1)! + o(h^{p+1}) \). By (A.27) and Taylor expansion, for small \( h \),

\[
q_1(m(u + ht, \chi_1); \{1, \ldots, (ht)^p\} \otimes \chi_1)^T \beta(u) = q_1(m(u + ht, \chi_1); \eta(u + ht, \chi_1)) \\
+ \{(1, \ldots, (ht)^p) \otimes \chi_1\}^T \beta(u) - \eta(u + ht, \chi_1)\}q_2(m(u + ht, \chi_1); \eta(u + ht, \chi_1) + o(h^{p+1}) \\
= -(ht)^{p+1}x_1^T A^{(p+1)}(u)/(p + 1)! \rho(u + ht, \chi_1) + o(h^{p+1}).
\]

Thus we obtain

\[
E(W_n) = -\sqrt{nh}\left\{c_p \otimes \frac{\Gamma(u)A^{(p+1)}(u)}{(p + 1)!}\right\} f_U(u)h^{p+1}\{1 + o(1)\},
\]

and analogously,

\[
\text{var}(W_n) = \frac{n\text{var}\{q_1(Y; \eta(u))Z_1K_h(U_1 - u)\}}{n} \\
= nh\text{var}\{q_1(Y; \eta(u))Z_1K_h(U_1 - u)\} \\
= h[E\{q_1^2(Y; \eta(u))Z_1Z_1^T (K_h)^2(U_1 - u)\} - O(h^{2p+2})].
\]
Recall $\eta_i(u) = \{u_i(u) \otimes x_i\}^T \beta(u)$. Setting $F^{-1}(\eta_i(u)) = m_i(u)$ leads to

$$E\{q''(Y_i; \eta_i(u))Z_iZ_i^T(K_h)^2(U_1 - u)\}$$

$$= E\left[\left\{q''(m_1(u)) \frac{Y_i - m_1(u)}{F'(m_1(u))}\right\}^2 Z_iZ_i^T(K_h)^2(U_1 - u)\right]$$

$$= E\left[\left\{q''(m_1(u)) \frac{Y_i - m_1(U_1,x_1) + m(U_1,x_1) - m_1(u)}{F'(m_1(u))}\right\}^2 Z_iZ_i^T(K_h)^2(U_1 - u)\right]$$

$$= E\left[\frac{q''(m_1(u))}{F'(m_1(u))} \left\{Y_i - m(U_1,x_1)\right\}^2 + \left\{m(U_1,x_1) - m_1(u)\right\}^2 Z_iZ_i^T(K_h)^2(U_1 - u)\right]$$

$$= E\left[\frac{q''(m_1(u))}{F'(m_1(u))} \text{var}(Y_i|U_1,x_1)Z_iZ_i^T(K_h)^2(U_1 - u)\right] \{1 + o(1)\}$$

$$= \{S^* \otimes \Delta(u)\} f_U(u) \frac{1}{h} \{1 + o(1)\},$$

which gives

$$\text{var}(W_n) = \{S^* \otimes \Delta(u)\} f_U(u) + o(1) + O(h^{2p+3}) \equiv B + o(1), \quad (A.31)$$

where $B = \{S^* \otimes \Delta(u)\} f_U(u) > 0$. Thus the assumption on $n$ and $h$ conclude $W_n = O_p(1)$. By the quadratic approximation lemma (Fan and Gijbels, 1996, p. 210, and Hjort and Pollard, 1996), it follows that

$$\hat{\beta}^*(u) = -A^{-1}W_n + o_P(1). \quad (A.32)$$

To demonstrate the asymptotic joint normality of $W_n$,

$$\{\text{var}(W_n)\}^{-1/2}\{W_n - E(W_n)\} \xrightarrow{\mathcal{L}} N(0, I_{d(p+1)}), \quad (A.33)$$

it suffices to show by the Cramer-Wold device that for any unit vector $a$,

$$\frac{a^T\{W_n - E(W_n)\}}{\{a^T\text{var}(W_n)a\}^{1/2}} \xrightarrow{\mathcal{L}} N(0, 1).$$

We now check the Lyapounov condition which requires the existence of $\delta > 0$, such that

$$nE\{|a_nq_1(Y_i; \eta_1(u))(a^TZ_i)K((U_1 - u)/h)^{2+\delta}\}/\{\text{var}(a^T W_n)^{(2+\delta)/2}\} \rightarrow 0.$$

For $\delta$ assumed in Condition E3', the numerator becomes

$$n(a_nh)^{2+\delta}E[|q_1^{2+\delta}(Y_i; \eta_1(u))(a^TZ_i)(K_h)^{2+\delta}(U_1 - u)|]$$

$$= n(a_nh)^{2+\delta} \frac{1}{h^{1+\delta}} = na_n^{2+\delta}h = nh\left(\frac{1}{n}h\right)^{2+\delta} = \frac{nh}{(nh)^{1+\delta/2}} \rightarrow 0,$$

from which the Lyapounov condition is satisfied.
Finally, combining (A.32), (A.30), (A.31) and (A.33), we observe that
\[
\hat{\beta}^*(u) = -A^{-1}|\text{var}(W_n)^{1/2}\text{var}(W_n)^{-1/2}\{W_n - E(W_n)\} + E(W_n)| + o_P(1)
= -A^{-1}\{B^{1/2}N(0, I) + E(W_n)\} + o_P(1)
= N(0, A^{-1}BA^{-1}) - A^{-1}E(W_n) + o_P(1),
\]
in which the asymptotic covariance matrix is
\[
A^{-1}BA^{-1} = [S^{-1}S^{-1} \otimes \{\Gamma(u)^{-1} \Delta(u) \Gamma(u)^{-1}\}]/f_U(u),
\]
and the asymptotic bias vector is
\[
-A^{-1}E(W_n) = \sqrt{n}h \{S^{-1}c_p \otimes \frac{A^{(p+1)}(u)}{(p+1)!}\} h^{p+1} \{1 + o(1)\}.
\]
This completes the proof.

**Proof of Lemma 2**

For a binary \( Y, L(Y, \mu) = Y L(1, \mu) + (1 - Y)L(0, \mu) \). Then Condition F implies
\[
L(Y, \mu) = Y L(1, \mu) + (1 - Y)L(1, 1 - \mu).
\]
Condition G combined with the continuity assumption on \( F^{-1} \) implies \( 1 - \mu = F^{-1}(-F(\mu)) \).
Thus
\[
L(Y, \mu) = Y L(1, F^{-1}(F(\mu))) + (1 - Y)L(1, F^{-1}(-F(\mu)))
= L(1, F^{-1}(Y^*F(\mu))).
\]
This completes the proof.

**Proof of Theorem 9**

We have that
\[
V(Y^*F) = Y V(F) + (1 - Y)V(-F), \quad \text{for } F \in \mathcal{F}_1, \quad \text{(A.34)}
\]
\[
\frac{\partial V(Y^*F)}{\partial F} = Y V'(F) - (1 - Y)V'(-F), \quad \text{for } F \in \mathcal{F}_2.
\]
Combining Condition I, we have that for all \( \mu \in \mathcal{N}_2, \)
\[
\frac{\partial V(Y^*F_B(\mu))}{\partial \mu} = \frac{\partial V(Y^*F)}{\partial F}\bigg|_{F=F_B(\mu)} F'_B(\mu)
= \left\{Y \frac{\mu V'(F_B(\mu)) - \frac{1 - Y}{1 - \mu}(1 - \mu)V'(-F_B(\mu))}{\mu - \frac{1 - Y}{1 - \mu}} F'_B(\mu)\right\} F'_B(\mu)
= \mu V'(F_B(\mu)) \left(Y - \frac{1 - Y}{1 - \mu}\right) F'_B(\mu)
= V'(F_B(\mu)) F'_B(\mu) \frac{Y - \mu}{1 - \mu},
\]
43
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 (6.3), satisfies $Q(Y, Y) = 0$ and fulfills (3.5), therefore is a BD. The choice $q$ in (6.4) follows from (3.7). This completes the proof.

**Proof of Lemma 3**

Define $L(F) = E\{V(Y^*F)|X = x\}$. From (A.34), it follows that $L(F) = m(x)V(F) + \{1 - m(x)\}V(-F)$ for $F \in \mathcal{F}_2$. Since $V$ is convex, we observed that $L(F)$ is convex in $F$, and

$$L'(F) = m(x)V'(F) - \{1 - m(x)\}V'(-F), \quad \text{for } \pm F \in \mathcal{F}_2.$$  

Thus (6.5) implies $L'(F_B(m(x))) = 0$ for any $m(x)$ such that $V'(F_B(m(x)))$ and $V'(-F_B(m(x)))$ exist, namely, (6.2). This completes the proof.

**Proof of Lemma 4**

Taking the derivatives with respect to $\mu$ on both sides of (6.2) 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

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

which concludes the proof.

**Proof of Theorem 10**

Before showing Theorem 10, we first show Lemma 6.

**Lemma 6** *For a random variable $X$, assume that $A_i \leq g_i(X) \leq B_i$ for finite constants $A_i$ and $B_i$, $i = 1, 2$. Then

$$|\text{cov}\{g_1(X), g_2(X)\}| \leq 4^{-1}(B_1 - A_1)(B_2 - A_2).$$

Proof: Define $\mu_i = E\{g_i(X)\}$, $i = 1, 2$, and $D(g_1, g_2) = \text{cov}\{g_1(X), g_2(X)\}$. Then

$$D(g_1, g_2) = E[\{g_1(X) - \mu_1\} \{g_2(X) - \mu_2\}] = \int (g_1 - \mu_1)(g_2 - \mu_2)dP(x).$$

By Cauchy-Schwarz inequality,

$$\{D(g_1, g_2)\}^2 \leq \int (g_1 - \mu_1)^2 dP(x) \int (g_2 - \mu_2)^2 dP(x) = D(g_1, g_1)D(g_2, g_2),$$

44
in which

\[
D(g_1, g_1) = \int (g_1 - \mu_1)^2 dP(x) = \int \{g_1(x) - A_1\} + (A_1 - \mu_1)\{g_1(x) - B_1\} + (B_1 - \mu_1)dP(x)
\]

\[
= \int \{g_1(x) - A_1\}\{g_1(x) - B_1\}dP(x) + (B_1 - \mu_1)\int \{g_1(x) - A_1\}dP(x)
\]

\[
+ (A_1 - \mu_1) + \int \{g_1(x) - B_1\}dP(x) + (A_1 - \mu_1)(B_1 - \mu_1)
\]

\[
= -\int \{g_1(x) - A_1\}\{B_1 - g_1(x)\}dP(x) + (\mu_1 - A_1)(B_1 - \mu_1)
\]

\[
+ (A_1 - \mu_1)(B_1 - B_1) + (A_1 - \mu_1)(B_1 - \mu_1)
\]

\[
\leq (\mu_1 - A_1)(B_1 - \mu_1) \leq \left( \frac{B_1 - A_1}{2} \right)^2.
\]

Similarly, \(D(g_2, g_2) \leq (\frac{B_2 - A_2}{2})^2\). This completes the proof of Lemma 6.

We now show Theorem 10. If Condition J holds, then \(\tilde{\lambda}_i\) and \(Y_i\) are similarly ordered, and Chebyshev inequality (Hardy, Littlewood, and Polya, 1988) implies \(\text{cov}(\tilde{\lambda}_i, Y_i|\mathcal{X}) \geq 0\). If Condition K holds, we apply Corollary 2. From (3.9), we observe \(-(b-a) \leq q'(\mu) \leq b-a\) for a.e. \(\mu\). Thus according to the definition of \(\tilde{\lambda}_i\), \(-(b-a)/2 \leq \tilde{\lambda}_i \leq (b-a)/2\). It is also noticed that \(0 \leq Y_i \leq 1\). An application of Lemma 6 gives the upper bound.

![Figure 1: Illustration of \(Q(\nu, \mu)\) as defined in (2.1). The concave curve is the \(q\)-function; the two dashed lines indicate locations of \(\nu\) and \(\mu\); the solid strict line is \(q(\mu) + (\nu - \mu)q'(\mu)\); the length of the vertical line with arrows at each end is \(Q(\nu, \mu)\).](image-url)
Figure 2: The plot of $q$-function as used in the loss functions for binary responses (in the left column) and three-class responses (in the right column).
Figure 3: Illustration of the margin-based loss functions $V(s)$. Line types are indicated in the legend box. Each function has been re-scaled to pass the point $(0,1)$.

Figure 4: The data generated from the model, blue dot—class 1; red circle—class 0. The black curves—true decision boundary; green curves—the decision boundary obtained from the local-likelihood estimation.
Figure 5: Cross-validated estimates of the prediction error as a function of bandwidth parameter.

Figure 6: The misclassification error rate. The red curve—training data; the blue dashed line—test data; horizontal line—Bayes rule.