Kernel Regression Utilizing External Information as Constraints

Chi-Shian Dai, Advisor: Jun Shao

1 Introduction

In modern statistical analysis we have not only individual-level data carefully collected from a population of interest, but also information from some external datasets, which typically have large or huge sizes but often contain relatively crude information such as summary statistics, due to practical and ethical reasons. Sources of external datasets include, for example, population-based censuses, administrative datasets, and databases from past investigations. Since the main individual-level dataset, called the internal dataset in what follows, is obtained to address specific scientific questions, it may contain more measured covariates from each sampled subject and, consequently, its size may be much smaller than those of the external datasets due to cost considerations. Thus, there is a growing need for individual-level data analysis utilizing summary or individual-level information from external datasets. This problem is different from the traditional meta-analysis in which the analysis is based on multiple datasets with summary statistics, without an internal individual-level dataset possibly containing more covariates.

In this paper we primarily focus on regression between a response variable Y and a covariate vector $\boldsymbol{U}$, based on an internal individual-level dataset in which both Y and $\boldsymbol{U}$ are measured, and a summary or individual-level external dataset on Y and $\boldsymbol{X}$, a part of the vector $\boldsymbol{U}$. The part of $\boldsymbol{U}$ measured in the internal but not external dataset is denoted by $\boldsymbol{Z}$, i.e., $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Z})$. As we indicated earlier, one reason why $\boldsymbol{Z}$ is not measured in the external dataset is that $\boldsymbol{Z}$ is relatively expensive to measure; another possible reason is due to the progress of new technology and/or new scientific relevance for measuring $\boldsymbol{Z}$.

Under the same setting and a parametric model between the response Y and covariate vector $\boldsymbol{U}$, [2] proposed a constrained maximum likelihood estimation by utilizing the summary information from an external dataset in the form of constraints to the observed likelihood. Other parametric or semiparametric approaches on using information from external datasets can be found, for example, in [4], [1], [11], [8], [3], [14], and [10].

A well-established nonparametric regression approach is the kernel regression described for example in [13], which does not require any assumption except for some smoothness condition on the regression function between Y and $\boldsymbol{U}$. To make use of summary or individual-level information from an external dataset, in this paper we propose a two-step kernel regression method. In the first step, we apply an optimization to obtain fitted values $\widehat{\mu}_1, ..., \widehat{\mu}_n$ at $\boldsymbol{U}_1, ..., \boldsymbol{U}_n$, respectively, subject to some constraints constructed using summary or individual-level information from the external dataset. As a predict, $\widehat{\mu}_i$ is better than the fitted value at $\boldsymbol{U}_i$ from the standard kernel regression, as it utilizes external information. In the second step, we apply the standard kernel regression treating $\widehat{\mu}_i$'s as the observed Y-values to obtain the estimated regression function.

We organize this paper as follows. Section 2 describes the notation and methodology. In Section 3, we establish the asymptotic normality of constrained kernel regression estimator at each fixed $\boldsymbol{U} = \boldsymbol{u}$, and provide a theoretical comparison between our proposed estimator and the standard kernel estimator, in terms of asymptotic mean integral squared error (AMISE), a measure often used for comparison [5]. We discuss how to choose the constraints in Section 4. Section 5 presents some simulation results as complements to asymptotic results. All technical proofs are given in the Appendix.

2 The Use of External Summary Statistics

2.1 Methodology

The internal dataset contains individual-level observations of an independent and identically distributed sample of size n, $\{(Y_i, \boldsymbol{U}_i), i = 1, ..., n\}$, from the population of $(Y, \boldsymbol{U})$, where Y is a response of interest and $\boldsymbol{U}$ is a p-dimensional vector of continuous covariates associated with Y, and p is a fixed integer smaller than n. We are interested in the estimation of regression function

$$\mu(\boldsymbol{u}) = E(Y \mid \boldsymbol{U} = \boldsymbol{u}),$$

the conditional expectation of Y given $\boldsymbol{U} = \boldsymbol{u}$, for any $\boldsymbol{u}$ in the range of $\boldsymbol{U}$.

Let $\kappa(\boldsymbol{u})$ be a given kernel function on the p-dimensional Euclidean space. We assume that $\boldsymbol{U}$ is standardized so that the same bandwidth $b > 0$ is used for every component of $\boldsymbol{U}$. The standard kernel regression estimator of $\mu(\boldsymbol{u})$ for any fixed $\boldsymbol{u}$ based on the internal dataset is

$$\begin{aligned}\widehat{\mu}_K(\boldsymbol{u}) &= \arg\min_{\mu} \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)(Y_i - \mu)^2 \\ &= \sum_{i=1}^{n} Y_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\end{aligned} \qquad (1)$$

where $\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) = b^{-p}\kappa\left(b^{-1}(\boldsymbol{u} - \boldsymbol{U}_i)\right)$.

The external dataset is another sample of size m from the population of $(Y, \boldsymbol{X})$, independent of the internal sample, where $\boldsymbol{X}$ is a q-dimensional sub-vector of $\boldsymbol{U}$, $q \leq p$. In this section, we consider the scenario where only some summary statistics are available from the external dataset. Specifically, the external dataset provides the vector $\widehat{\boldsymbol{\beta}}$ of least squares estimate of $\boldsymbol{\beta}$ based on external data and a working model $E(Y|\boldsymbol{X}) = \boldsymbol{\beta}^\top \boldsymbol{X}$, where $\boldsymbol{a}^\top$ is the transpose of vector $\boldsymbol{a}$. Regardless of whether the working model is correct or not, the asymptotic limit of $\widehat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}_X^{-1} E\{\boldsymbol{X} E(Y|\boldsymbol{X})\}$ under some moment conditions, where $\boldsymbol{\Sigma}_X = E(\boldsymbol{X}\boldsymbol{X}^T)$ is assumed to be positive definite.

By the definition of $\boldsymbol{\beta}_0$ and $E(Y|\boldsymbol{X}) = E\{E(Y|\boldsymbol{U})|\boldsymbol{X}\} = E\{\mu(\boldsymbol{U})|\boldsymbol{X}\}$,

$$\begin{aligned} E(\boldsymbol{\beta}_0^\top \boldsymbol{X}\boldsymbol{X}^\top) &= E\{E(Y|\boldsymbol{X})\boldsymbol{X}^\top\}\boldsymbol{\Sigma}_X^{-1} E(\boldsymbol{X}\boldsymbol{X}^\top) \\ &= E[E\{\mu(\boldsymbol{U})|\boldsymbol{X}\}\boldsymbol{X}^\top] \\ &= E[E\{\mu(\boldsymbol{U})\boldsymbol{X}^\top|\boldsymbol{X}\}] \\ &= E\{\mu(\boldsymbol{U})\boldsymbol{X}^\top\} \end{aligned}$$

and, consequently, the summary information from external dataset can be utilized through the constraint

$$E[\{\boldsymbol{\beta}_0^\top \boldsymbol{X} - \mu(\boldsymbol{U})\}\boldsymbol{X}] = 0. \tag{2}$$

We propose a two-step procedure. In the first step, we make use of (2) and the external information to obtain estimated values $\widehat{\mu}_1, ..., \widehat{\mu}_n$ of $\mu(\boldsymbol{U}_1), ..., \mu(\boldsymbol{U}_n)$, respectively. To achieve this, we solve the minimization in (1) subject to constraint (2). Specifically, we obtain the n-dimensional vector $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, ..., \widehat{\mu}_n)^\top$ by solving the following constrained minimization:

$$\begin{aligned} \widehat{\boldsymbol{\mu}} = \arg \min_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^{n} \kappa_l(\boldsymbol{U}_k - \boldsymbol{U}_j)}(Y_j - \mu_i)^2 \\ \text{subject to} \quad \sum_{i=1}^{n}(\mu_i - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)\boldsymbol{X}_i = 0, \end{aligned} \tag{3}$$

where l is a bandwidth that may be different from b in (1). More discussion about bandwidths is provided later.

To motivate the objective function in (3) being minimized, note that

$$\sum_{j=1}^{n} \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^{n} \kappa_l(\boldsymbol{U}_k - \boldsymbol{U}_j)} (Y_j - \mu_i)^2 \approx E[\{Y - \mu(\boldsymbol{U})\}^2 | \boldsymbol{U} = \boldsymbol{U}_i]$$

for each i and, hence, the objective function in (3) divided by n approximates

$$\frac{1}{n} \sum_{i=1}^{n} E[\{Y - \mu(\boldsymbol{U})\}^2 | \boldsymbol{U} = \boldsymbol{U}_i] \approx E[\{Y - \mu(\boldsymbol{U})\}^2].$$

To derive an explicit form of $\widehat{\boldsymbol{\mu}}$ in (3), let $\boldsymbol{G}$ be the $n \times n$ matrix whose ith row is $\boldsymbol{X}_i$, and $\boldsymbol{\mu}$, $\widehat{\boldsymbol{h}}$, and $\widehat{\boldsymbol{\mu}}_K$ be the n-dimensional vectors whose ith components are μ_i, $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i$, and $\widehat{\mu}_K(\boldsymbol{U}_i)$, respectively, with $\widehat{\mu}_K$ being defined by (1). Then (3) is the same as

$$\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} (\boldsymbol{\mu}^\top \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \widehat{\boldsymbol{\mu}}_K) \quad \text{subject to} \quad \boldsymbol{G}^\top (\boldsymbol{\mu} - \widehat{\boldsymbol{h}}) = 0.$$

From the Lagrange multiplier $L(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \boldsymbol{\mu}^\top \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \widehat{\boldsymbol{\mu}}_K + 2\boldsymbol{\lambda}^\top \boldsymbol{G}^\top (\boldsymbol{\mu} - \widehat{\boldsymbol{h}})$ and $\nabla_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\lambda}) = 2\boldsymbol{\mu} - 2\widehat{\boldsymbol{\mu}}_K + 2\boldsymbol{G}\boldsymbol{\lambda}$, we obtain that $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_K - \boldsymbol{G}\boldsymbol{\lambda}$. From the constraint, $\boldsymbol{G}^\top \widehat{\boldsymbol{h}} = \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}} = \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}}_K - \boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{\lambda}$. Solving for $\boldsymbol{\lambda}$, we obtain that $\boldsymbol{\lambda} = (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top \widehat{\boldsymbol{\mu}}_K - (\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top \widehat{\boldsymbol{h}}$. Hence, $\widehat{\boldsymbol{\mu}}$ has an explicit form

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_K + \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top (\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\mu}}_K). \tag{4}$$

This estimator adds an adjustment term to $\widehat{\boldsymbol{\mu}}_K$, the estimator from the standard kernel regression. The adjustment involves the difference $\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\mu}}_K$ and the projection matrix $\boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top$.

Since the additional information from the external dataset is used in (3), $\widehat{\boldsymbol{\mu}}$ in (3) is expected to be better than $\widehat{\boldsymbol{\mu}}_K$ that does not use external information. In the Appendix we show that $\lim_{n\to\infty} \sum_{i=1}^{n} E\{\widehat{\mu}_i - \mu(U_i)\}^2/n \leq \lim_{n\to\infty} \sum_{i=1}^{n} E\{\widehat{\mu}_K(U_i) - \mu(U_i)\}^2/n$.

To obtain an improved estimator of the entire function $\mu(\boldsymbol{u})$, not just the function $\mu(\boldsymbol{u})$ at $\boldsymbol{U}_1, ..., \boldsymbol{U}_n$, we propose a second step to apply the standard kernel regression with responses $Y_1, ..., Y_n$ replaced by $\widehat{\mu}_1, ..., \widehat{\mu}_n$. Specifically, our proposed estimator of $\mu(\boldsymbol{u})$ is

$$\widehat{\mu}_{CK}(\boldsymbol{u}) = \sum_{i=1}^{n} \widehat{\mu}_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Bigg/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i), \tag{5}$$

where b is the same bandwidth in (1).

If we apply kernel regression with $\widehat{\mu}_1, ..., \widehat{\mu}_n$ in (5) replaced by $\widehat{\mu}_1^{(0)}, ..., \widehat{\mu}_n^{(0)}$, defined as the solution to minimization in (3) without applying the constraint $\boldsymbol{G}^\top(\boldsymbol{\mu} - \widehat{\boldsymbol{h}}) = 0$, i.e., $\widehat{\mu}_i^{(0)}$ is the estimator (1) at $\boldsymbol{u} = \boldsymbol{U}_i$ but with b replaced by l, then we obtain another estimator

$$\widehat{\mu}_{DK}(\boldsymbol{u}) = \sum_{i=1}^n \widehat{\mu}_i^{(0)} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Big/ \sum_{i=1}^n \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i). \tag{6}$$

It does not use external information, but is obtained by applying kernel regression twice, and will be referred to as double kernel regression estimator. Intuitively, $\widehat{\mu}_{DK}$ should not be better or worse than the standard $\widehat{\mu}_K$, as no additional information is utilized in (6). In the asymptotic theory presented next, we show that $\widehat{\mu}_{DK}$ is asymptotically equivalent to the standard kernel regression using a kernel different from κ.

2.2 Asymptotic Theory

We now establish some asymptotic results (as the sample size of the internal dataset n increases to infinity), which enables us to compare three estimators in (1), (5), and (6). The first result is the asymptotic normality of $\widehat{\mu}_{CK}(\boldsymbol{u})$ and $\widehat{\mu}_{DK}(\boldsymbol{u})$ in (5)-(6) for a fixed $\boldsymbol{u}$, under regularity conditions (A1)-(A5).

Theorem 2.1. *Assume the following conditions.*

(A1) The response Y has a finite $E(|Y|^s)$ with $s > 2 + p/2$. The covariate vector $\boldsymbol{U}$ has compact support and has a positive definite covariance matrix. The density of $\boldsymbol{U}$ is bounded away from infinity and zero, and has bounded second-order derivatives.

(A2) Functions $\mu(\boldsymbol{u}) = E(Y|\boldsymbol{U} = \boldsymbol{u})$ and $\sigma^2(\boldsymbol{u}) = E[\{Y - \mu(\boldsymbol{U})\}^2|\boldsymbol{U} = \boldsymbol{u}]$ are Lipschitz continuous. The function $\mu(\boldsymbol{u})$ has bounded third-order derivatives.

(A3) The kernel κ is a positive bounded density with mean zero and finite sixth moments.

(A4) The bandwidths b in (1) and l in (3) are polynomial rate of n, satisfying $b \to 0$, $l \to 0$, $l/b \to \gamma \in (0, \infty)$, and $nb^{4+p} \to c \in [0, \infty)$ as the internal sample size $n \to \infty$.

(A5) The external sample size m satisfies $n = O(m)$, i.e., n/m is bounded by a fixed constant.

Then, as $n \to \infty$,

$$\sqrt{nb}\{\widehat{\boldsymbol{\mu}}_t(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \to N\big(B_t(\boldsymbol{u}), V_t(\boldsymbol{u})\big) \quad \textit{in distribution}, \tag{7}$$

where $t = DK$ or CK,

$$\begin{aligned}
B_{DK}(\boldsymbol{u}) &= c^{1/2}(1+\gamma^2)A(\boldsymbol{u}),\\
B_{CK}(\boldsymbol{u}) &= c^{1/2}\{(1+\gamma^2)A(\boldsymbol{u}) - \gamma^2\boldsymbol{x}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\},\\
V_{DK}(\boldsymbol{u}) &= \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})}\int\left\{\int\kappa(\boldsymbol{w}-\boldsymbol{v}\gamma)\kappa(\boldsymbol{v})d\boldsymbol{v}\right\}^2 d\boldsymbol{w},\\
V_{CK}(\boldsymbol{u}) &= V_{DK}(\boldsymbol{u}),\\
A(\boldsymbol{u}) &= \int\kappa(\boldsymbol{v})\left\{\tfrac{1}{2}\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v} + \nabla\mu(\boldsymbol{u})^T\boldsymbol{v}\boldsymbol{v}^T\nabla\log f_U(\boldsymbol{u})\right\}d\boldsymbol{v},
\end{aligned} \tag{8}$$

and f_U is the density of $\boldsymbol{U}$.

The proof is given in the Appendix.

Note that $B_t(\boldsymbol{u})$ and $V_t(\boldsymbol{u})$ are asymptotic bias and variance, respectively. The results in Theorem 2.1 indicate that the use of external information affects the asymptotic bias but not the asymptotic variance when (A5) holds, i.e., m is at least comparable with n. This effect makes $\widehat{\mu}_{CK}(\boldsymbol{u})$ less biased than $\widehat{\mu}_{DK}(\boldsymbol{u})$, because

$$\begin{aligned}
E\{B_{CK}(\boldsymbol{U})\}^2 &= c(1+\gamma^2)^2E[A(\boldsymbol{U}) - \boldsymbol{X}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\}]^2\\
&\quad + cE[\boldsymbol{X}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\}]^2\\
&\le c(1+\gamma^2)^2E[A(\boldsymbol{U}) - \boldsymbol{X}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\}]^2\\
&\quad + c(1+\gamma^2)^2E[\boldsymbol{X}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\}]^2\\
&= c(1+\gamma^2)^2E\{A(\boldsymbol{U})\}^2\\
&= E\{B_{DK}(\boldsymbol{U})\}^2
\end{aligned}$$

with equality holds if and only if $c = 0$. Since they have the same asymptotic variance, if $c > 0$, $\widehat{\mu}_{CK}(\boldsymbol{u})$ is asymptotically better than $\widehat{\mu}_{DK}(\boldsymbol{u})$ in terms of the asymptotic mean integrated square error (AMISE) defined as

$$\text{AMISE}(\widehat{\mu}_t) = E[\{B_t(\boldsymbol{U})\}^2 + V_t(\boldsymbol{U})], \quad t = CK \text{ or } DK,$$

a globe accuracy measure considered in the literature [5].

From the theory of kernel regression [9], under (A1)-(A4), the kernel estimator $\widehat{\mu}_K(\boldsymbol{u})$ in (1) also satisfies (7) with $t = K$,

$$B_K(\boldsymbol{u}) = c^{1/2} A(\boldsymbol{u}) \quad \text{and} \quad V_K(\boldsymbol{u}) = \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})} \int \{\kappa(\boldsymbol{v})\}^2 d\boldsymbol{v}.$$

It is shown in the Appendix that if we replace the kernel $\kappa(\boldsymbol{u})$ in $B_K(\boldsymbol{u})$ and $V_K(\boldsymbol{u})$ with the convolution kernel $\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{w})d\boldsymbol{w}$, then they become $B_{DK}(\boldsymbol{u})$ and $V_{DK}(\boldsymbol{u})$, respectively, i.e., $\widehat{\mu}_{DK}(\boldsymbol{u})$ is asymptotically equivalent to the standard kernel regression estimator with the convolution kernel, in terms of the AMISE.

A comparison between $\widehat{\mu}_{CK}$ in (5) and $\widehat{\mu}_K$ in (1) is given as follows.

Theorem 2.2. *Under the conditions in Theorem 2.1 and an additional condition that $\int \nabla^2 \kappa(\boldsymbol{u})\kappa(\boldsymbol{u})d\boldsymbol{u}$ being strictly negative definite,* AMISE($\widehat{\mu}_{CK}$) < AMISE($\widehat{\mu}_K$) *for c and γ in a neighborhood of 0.*

Example 2.1 (Gaussian Kernels)**.** *If we use the Gaussian kernel $\kappa(\boldsymbol{u}) = (2\pi)^{-p/2} e^{-\|\boldsymbol{u}\|^2/2}$, the density of $N(0, \boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the identity matrix of order p, then*

$$\int \nabla^2 \kappa(\boldsymbol{u})\kappa(\boldsymbol{u})d\boldsymbol{u} = \frac{1}{(2\pi)^p} \int \left(\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{I}_p\right) e^{-\|\boldsymbol{u}\|^2} d\boldsymbol{u} = -\frac{1}{2^{1+p/2}(2\pi)^{p/2}} \boldsymbol{I}_p$$

which shows that the Gaussian kernel satisfies the condition in Theorem 2.2. In this case, the convolution kernel $\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{w})d\boldsymbol{w}$ is the density of $N(0, (1+\gamma^2)\boldsymbol{I}_p)$.

2.3 Bandwidth

(A4) in Theorem 2.1 provides the rates of the bandwidths l and b for $\widehat{\mu}_{CK}$. In an application, however, we need to choose l and b in terms of a finite sample. Following the idea in [6], we choose the bandwidths l and b by minimizing the following leave-one-out cross-validation,

$$\text{CV}(l, b) = \frac{1}{n} \sum_{i=1}^{n} \{\widehat{\mu}_{CK}^{(-i)}(\boldsymbol{U}_i) - Y_i\}^2,$$

where $\widehat{\mu}_{CK}^{(-i)}(\boldsymbol{u})$ is the kernel regression estimator (5) with bandwidths l and b but without using the ith point $(Y_i, \boldsymbol{U}_i)$ in the internal dataset.

3 The Use of External Individual-Level Data

3.1 Methodology and Asymptotic Theory

In some applications the individual-level data, $\{Y_{n+j}, \boldsymbol{X}_{n+j}, j = 1, ..., m\}$, are available from the external dataset, which are from an independent and identically distributed sample, independent of the internal data. We assume that the populations of external and internal data are the same, except that only $\boldsymbol{X}$, not $\boldsymbol{U}$, is measured in the external dataset. Although we consider the frame work of internal and external datasets, our result also covers the scenario where we have a single dataset with n observations $(Y_1, \boldsymbol{U}_1), ..., (Y_n, \boldsymbol{U}_n)$ and m observations $(Y_{n+1}, \boldsymbol{X}_{n+1}), ..., (Y_{n+m}, \boldsymbol{X}_{n+m})$ without $\boldsymbol{Z}$-values due to the fact that the measurement of $\boldsymbol{Z}$ is difficult or expensive.

Let $g(\boldsymbol{x})$ be a vector-valued function on the range of $\boldsymbol{X}$ such that all components of $\mu(\boldsymbol{U})g(\boldsymbol{X})$ are integrable. The choice of g is discussed in Section 3.2. From the property of conditional expectation,

$$E[\{E(Y|\boldsymbol{X}) - \mu(\boldsymbol{U})\}g(\boldsymbol{X})] = 0. \tag{9}$$

We use identity (9) as a constraint to incorporate the information from the external dataset, with the function $h(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ being estimated by $\widehat{h}(\boldsymbol{x})$ based on the standard kernel regression using the external individual-level data.

In the first step we obtain $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, ..., \widehat{\mu}_n)^\top$ through the constrained minimization (3) with the constraint in (3) replaced by the constraint constructed using (9),

$$\sum_{i=1}^{n} \{\widehat{\mu}_i - \widehat{h}(\boldsymbol{X}_i)\}g(\boldsymbol{X}_i) = 0. \tag{10}$$

The solution $\widehat{\boldsymbol{\mu}}$ is still given by (4) with $\widehat{\boldsymbol{h}}$ being the n-dimensional vector whose ith component is $\widehat{h}(\boldsymbol{X}_i)$ and $\boldsymbol{G}$ being the matrix whose ith row is $g(\boldsymbol{X}_i)$. In the second step, our proposed estimator of $\mu(\boldsymbol{u})$ is still $\widehat{\mu}_{CK}(\boldsymbol{u})$ given by (5).

For the asymptotic normality of $\widehat{\mu}_{CK}(\boldsymbol{u})$, we have the following result similar to Theorem 2.1.

Theorem 3.1. *Assume (A1)-(A5) in Theorem 2.1 and*

(A1′) The matrix $\boldsymbol{\Sigma}_g = E\{g(\boldsymbol{X})g(\boldsymbol{X})^\top\}$ is positive definite.

(A2′) The functions $h(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are Lipschitz continuous. The function $h(\boldsymbol{x})$ has bounded third-order derivatives.

(A3′) The kernel function used in the kernel regression based on the external dataset satisfies condition (A3).

(A4′) The bandwidth used in the kernel regression based on the external dataset is of the order $m^{-1/(4+q)}$ as $m \to \infty$.

Then, as $n \to \infty$,

$$\sqrt{nb}\{\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \to N\big(B_{CK}(\boldsymbol{u}), V_{CK}(\boldsymbol{u})\big) \quad \textit{in distribution,}$$

where

$$B_{CK}(\boldsymbol{u}) = c^{1/2}[(1+\gamma^2)A(\boldsymbol{u}) - \gamma^2 g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}]$$

and $V_{CK}(\boldsymbol{u})$ and $A(\boldsymbol{u})$ are the same as those in Theorem 2.1.

Based on the result in Theorem 3.1, the conclusion in Theorem 2.2 still holds.

3.2 Choice of g in Constraint (9)

In the rest of this section we consider the choice of the function g in constraint (9). Note that g does not affect the asymptotic variance V_{CK}, but it affects the asymptotic bias B_{CK} through the term $g(\boldsymbol{x})^\top \boldsymbol{\xi}$, where $\boldsymbol{\xi} = \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}$. Since

$$E\{B_{CK}(\boldsymbol{U})\}^2 = cE\{A(\boldsymbol{U})\}^2 + c\gamma^4 E\{A(U) - g(\boldsymbol{X})^\top \boldsymbol{\xi}\}^2, \tag{11}$$

the best g is the one minimizing $E\{A(U) - g(\boldsymbol{X})^\top \boldsymbol{\xi}\}^2$. It is well-known that the solution to

$$\min_{\text{all function } \psi} E\{A(\boldsymbol{U}) - \psi(\boldsymbol{X})\}^2$$

is $\psi(\boldsymbol{X}) = E\{A(\boldsymbol{U})|\boldsymbol{X}\}$. Hence, the best g is the one-dimensional function $g^*(\boldsymbol{X}) = E\{A(\boldsymbol{U})|\boldsymbol{X}\}$ with $\boldsymbol{\xi} = 1$.

Unfortunately, the function $g^*(\boldsymbol{x}) = E\{A(\boldsymbol{U})|\boldsymbol{X} = \boldsymbol{x}\}$ is typically unknown. In the following we propose an estimator of $g^*(\boldsymbol{x})$ and study the asymptotic property of $\widehat{\mu}_{CK}$ with the estimated function g^*.

First, we construct an estimator $\widehat{A}(\boldsymbol{u})$ of $A(\boldsymbol{u})$. Suppose that the kernel κ has the property that, for any components u_k and u_j of $\boldsymbol{u}$, $\int u_k u_j \kappa(\boldsymbol{u}) d\boldsymbol{u} = 0$

when $k \neq j$, and $\int u_k^2 \kappa(\boldsymbol{u}) d\boldsymbol{u} = 1$. Then the function $A(\boldsymbol{u})$ in (8) has the form

$$\begin{aligned} A(\boldsymbol{u}) &= \frac{1}{2} \sum_{k=1}^{p} \left\{ \nabla_{kk}^2 \mu(\boldsymbol{u}) + \frac{2\nabla_k \mu(\boldsymbol{u}) \nabla_k f_U(\boldsymbol{u})}{f_U(\boldsymbol{u})} \right\} \\ &= \frac{1}{2} \sum_{k=1}^{p} \left\{ \frac{\nu_k(\boldsymbol{u})}{f_U(\boldsymbol{u})} - \frac{\nu_0(\boldsymbol{u}) \nabla_{kk}^2 f_U(\boldsymbol{u})}{f_U^2(\boldsymbol{u})} \right\}, \end{aligned}$$

where $\nu_k(\boldsymbol{u}) = \nabla_{kk}^2 \mu(\boldsymbol{u}) f_U(\boldsymbol{u}) + 2\nabla_k \mu(\boldsymbol{u}) \nabla_k f_U(\boldsymbol{u}) + \mu(\boldsymbol{u}) \nabla_{kk}^2 f_U(\boldsymbol{u})$, $\nu_0(\boldsymbol{u}) = \mu(\boldsymbol{u}) f_U(\boldsymbol{u})$, ∇_k denotes the kth component of ∇, and ∇_{kk}^2 denotes the kth diagonal element of ∇^2. We then obtain an estimator $\widehat{A}(\boldsymbol{u})$ by estimating $f_U(\boldsymbol{u})$, $\nu_0(\boldsymbol{u})$, $\nu_k(\boldsymbol{u})$, and $\nabla_{kk}^2 f_U(\boldsymbol{u})$, $k = 1, \ldots, p$, with

$$\begin{aligned} \widehat{f}_U(\boldsymbol{u}) &= \frac{1}{n\lambda_1^p} \sum_{i=1}^{n} \kappa \left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{\lambda_1} \right) \\ \widehat{\nu}_0(\boldsymbol{u}) &= \frac{1}{n\lambda_1^p} \sum_{i=1}^{n} \kappa \left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{\lambda_1} \right) Y_i \\ \widehat{\nu}_k(\boldsymbol{u}) &= \frac{1}{n\lambda_2^{p+2}} \sum_{i=1}^{n} \nabla_{kk}^2 \widetilde{\kappa} \left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{\lambda_2} \right) Y_i \\ \nabla_{kk}^2 \widehat{f}_U(\boldsymbol{u}) &= \frac{1}{n\lambda_2^{p+2}} \sum_{i=1}^{n} \nabla_{kk}^2 \widetilde{\kappa} \left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{\lambda_2} \right), \end{aligned}$$

respectively, where $\widetilde{\kappa}$ is a kernel twice differentiable and λ_1 and λ_2 are bandwidths. Then, we apply kernel regression to estimate $g^*(\boldsymbol{x})$ by

$$\widehat{g}^*(\boldsymbol{x}) = \frac{\sum_{k=1}^{n} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k) \widehat{A}(\boldsymbol{U}_k)}{\sum_{k=1}^{n} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k)} \tag{12}$$

and use this $\widehat{g}^*$ as the function g in the constraint (9) to obtain $\widehat{\mu}_{CK}$ in (5).

We establish the following theorem for $\widehat{\mu}_{CK}$ based on estimated constraint $\widehat{g}^*$.

Theorem 3.2. *Assume the conditions in Theorem 3.1 and the following additional conditions.*

(C1) The kernel κ in (A3) satisfies $\int u_k^2 \kappa(\boldsymbol{u}) d\boldsymbol{u} = 1$ and $\int u_k u_j \kappa(\boldsymbol{u}) d\boldsymbol{u} = 0$ when $k \neq j$. The kernel $\widetilde{\kappa}$ in the estimators $\widehat{\nu}_k$ and $\nabla_{kk}^2 \widehat{f}_U$, $k = 1, ..., p$,

has finite second-order moments, bounded $\nabla^2_{kk}\widetilde{\kappa}$, *finite* $\int |\nabla^2_{kk}\widetilde{\kappa}(\boldsymbol{u})|d\boldsymbol{u}$, *and bounded* $\sup_{\boldsymbol{u}} \lambda^{-2}|\widetilde{\kappa}(\boldsymbol{u}/\lambda)|$ *and* $\sup_{\boldsymbol{u}} \lambda^{-3}|\nabla_k\widetilde{\kappa}(\boldsymbol{u}/\lambda)|$ *as* $\lambda \to 0$, $k = 1, ..., p$.

(C2) *The bandwidth* λ_1 *for* $\widehat{\nu}_0$ *and* $\widehat{f}_U$ *has order* $n^{-1/(p+4)}$, *the bandwidth* λ_2 *for* $\widehat{\nu}_k$ *and* $\nabla^2_{kk}\widehat{f}_U$ *has order* $n^{-1/(p+8)}$, *and the bandwidth* δ *in* (12) *has order* $n^{-1/(q+4)}$.

Then, the result in Theorem 3.1 with $g = g^*$ *holds for* $\widehat{\mu}_{CK}$ *using the estimated constraints* $\widehat{g}^*$ *in (12).*

Although g^* is the best choice for the constraint, the estimator $\widehat{g}^*$ is complicated as it involves the estimation of second-order gradient. Furthermore, the estimation of g^* has to use $\boldsymbol{U}$-data from the internal dataset with a sample size that may be smaller than the size of external dataset. Thus, the estimator $\widehat{\mu}_{CK}$ using $\widehat{g}^*$ may not perform well for finite sample size n although it is asymptotically optimal. For this reason, we consider an alternative in the next section.

3.3 The Choice of $g = (1, \widehat{h})$ in Constraint (9)

We consider $g(\boldsymbol{x}) = (1, \widehat{h}(\boldsymbol{x}))$ in constraint (9). First, this choice is asymptotically justified.

Theorem 3.3. *Assume the conditions in Theorem 3.1. Then, the result in Theorem 3.1 holds for* $\widehat{\mu}_{CK}$ *using the estimated constraint* $g = (1, \widehat{h})$ *in (9).*

Another g can be used in constraint (9) is $g = (1, \boldsymbol{X})$, as it is actually used in the case where we have summary statistics from external dataset (Section 2), not individual level data. In the rest of this section we argue that $(1, \widehat{h})$ is a better choice than $(1, \boldsymbol{X})$.

Since the asymptotic variance is the same for $\widehat{\mu}_{CK}$ with different g, it suffices to consider the asymptotic bias to compare the performance. From (11), the effect of different g is in the term

$$E\{A(\boldsymbol{U}) - g(\boldsymbol{X})^\top \boldsymbol{\xi}\}^2 = E\{A(\boldsymbol{U}) - g^*(\boldsymbol{X})\}^2 + E\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})^\top \boldsymbol{\xi}\}^2,$$

where $g^*(\boldsymbol{X}) = E\{A(\boldsymbol{U})|\boldsymbol{X}\}$. Since $E\{A(\boldsymbol{U}) - g^*(\boldsymbol{X})\}^2$ does not involve g, we only need to compare

$$E\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})^\top \boldsymbol{\xi}\}^2 = \min_{\boldsymbol{\beta}} E\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})^\top \boldsymbol{\beta}\}^2. \qquad (13)$$

In the following examples, we consider $\boldsymbol{X} = X$, $\boldsymbol{Z} = Z$, and $\boldsymbol{U} = (X, Z)$ is bivariate normal with zero means, unit variances, and correlation ρ.

Example 1: $\mu(\boldsymbol{U}) = X^3 + Z^2$.

In this case, for all $\rho \in [0, 1)$, it can be calculated that

$$A(\boldsymbol{U}) = 3X + 1 - \{3X^2(X - \rho Z) + 2Z(Z - \rho X)\}/(1 - \rho^2)$$

$$g^*(X) = 3X - 3X^3 - 1, \qquad h(X) = X^3 + \rho^2 X^2 + 1 - \rho^2.$$

First, we would calculate the common term $E\{A(\boldsymbol{U}) - g^*(X)\}^2$.

$$\begin{aligned}
&E\{A(\boldsymbol{U}) - g^*(X)\}^2\\
=&E\left[\frac{3\rho X^2(Z - \rho X) + 2\rho Z(X - \rho Z)}{1 - \rho^2} + 2 - 2Z^2\right]^2\\
=&\frac{1}{(1 - \rho^2)^2} E\{9\rho^2 X^4 (Z - \rho X)^2\}\\
+&\frac{1}{(1 - \rho^2)^2} E\{4\rho^2 Z^2 (X - \rho Z)^2\}\\
+&\frac{1}{(1 - \rho^2)^2} E\{12\rho^2 Z X^2 (Z - \rho X)(X - \rho Z)\}\\
+&\frac{1}{(1 - \rho^2)} E\{8\rho(1 - Z^2) Z (X - \rho Z)\}\\
+&\frac{1}{(1 - \rho^2)} E\{12\rho(1 - Z^2) X (Z - \rho X)\}\\
+&E(1 - Z^2)^2
\end{aligned}$$

Using conditional expectation, only the first two terms and the last term are not zero.

$$\begin{aligned}
=&9\frac{\rho^2}{1 - \rho^2}\mu_4 + 4\frac{\rho^2}{1 - \rho^2}\mu_2 + 4(\mu_4 - 2\mu_2 + 1)\\
=&31\frac{\rho^2}{1 - \rho^2} + 8.
\end{aligned}$$

For the different term (13), we compare two constraints (1,X), and $(1, h(X))$ If $g = (1, X)$, then

$$\min_{\beta_0, \beta_1} E\{g^*(X) - \beta_0 - \beta_1 X\}^2.$$

Since $E(X) = 0$, the standard result of linear regression shows that the minimizer is $\beta_0^* = E[g^*(X)] = -1$ and $\beta_1^* = E[g^*(X)X] = -6$. Then, the quantity in (13) is

$$E\{g^*(X) + 1 + 6X\}^2 = E\{9X - 3X^3\}^2 = 9\mu_6 + 81\mu_2 - 54\mu_4 = 54,$$

where μ_k indicate the moments of standard normal.

If $g = (1, h)$, then the quantity in (13) equals

$$\min_{\beta_0,\beta_1} E\{g^*(X) - \beta_0 - \beta_1[h(X) - 1]\}^2.$$

Since $E[h(X) - 1] = 0$, the standard result of linear regression shows that the minimizer is $\beta_0^* = E[g^*(X)] = -1$ and

$$\beta_1^* = E\{g^*(X)[h(X) - 1]/E[h(X) - 1]^2\} = -36/(15 + 2\rho^4).$$

The quantity in (13) is equal to

$$\begin{aligned}
& E\{g^*(X) + 1 - \beta_1^*[h(X) - 1]\}^2 \\
&= E\{3X - 3X^3 - \beta_1^*(X^3 + \rho^2X^2 - \rho^2)\}^2 \\
&= E\{(-\beta_1^* - 3)X^3 - \beta_1^*\rho^2X^2 + 3X + \beta_1^*\rho^2\}^2 \\
&= (-\beta_1^* - 3)^2\mu_6 + \beta_1^{*2}\rho^4\mu_3 + 9\mu_2 + \beta_1^{*2}\rho^4 + 3(-\beta_1^* - 3)\mu_4 - \beta_1^{*2}\rho^4\mu_2 \\
&= 15(\beta_1^* + 3)^2 + 9(-\beta_1^* - 3) + 3\beta_1^{*2}\rho^4 + 9 \\
&\le 15^3/17^2 + 297/25 + 9 \qquad \text{since } -\tfrac{36}{15} < \beta_1^* < -\tfrac{36}{17} \\
&= 32.56.
\end{aligned}$$

Then, we can see (1, h(X)) is better than (1, X) for all ρ. If given ρ, it can combine the common term $E\{A(\boldsymbol{U}) - g^*(X)\}^2$ to get the improvement. For example, let $\rho = 1/2$. For (1,X) we have $18.33 + 54 = 72.33$. On the orther hand, $(1, h(X))$ have $18.33 + 10.224 = 28.55$. So, the improvement might be $(72.33 - 28.55)/72.33 = 60.5\%$.

Example 2: Polynomial Models.

An extension of the result in Example 1 is to consider

$$\mu(\boldsymbol{U}) = \Theta(X) + \Phi(Z),$$

where Θ and Φ are polynomials with degrees d_x and d_z, respectively. Then,

$$A(\boldsymbol{U}) = \frac{1}{2}(\Theta''(X) + \Phi''(Z)) - \frac{1}{1-\rho^2}\{\Theta'(X)(X - \rho Z) + \Phi'(Z)(Z - \rho X)\},$$

which is a polynomial with the highest order terms $X^{d_x}, Z^{d_z}, X^{d_x-1}Z, Z^{d_z-1}X$. Since $Z|X \sim N(\rho X, (1-\rho^2))$, which can make sure $deg(E[Z^k|X]) = k$,? $g^*(X)$ is a polynomial with order $\max(d_x, d_z)$. On the orther hand,

$$h(X) = \Theta(X) + E[\Phi(Z)|X],$$

which is a polynomial with degree $\max(d_x, d_z)$. Now, we can see $h(X)$ and $g^*(X)$ have the same degree, and hence approximating $g^*(X)$ with the polynomial h having the same degree is better than approximating g^* by the lower degree polynomial such as X.

Example 3: Cosine Basis.

Assume that

$$\mu(\boldsymbol{U}) = \sum_k \{c_k \cos(kX) + d_k \cos(kZ)\},$$

where c_k's and d_k's are constants. Note that

$$\begin{aligned}
E[\nabla_x \mu(\boldsymbol{U}) \nabla_x \log f(\boldsymbol{U})|X] &= -\sum_k c_k k \sin(kX) \nabla_x \log f_X(X) \\
&= \sum_k c_k k X \sin(kX) \\
E[\nabla_z \mu(\boldsymbol{U}) \nabla_z \log f(\boldsymbol{U})|X] &= \sum_k \int -d_k k \sin(kz) \nabla_z f(z|X) dz \\
&= \sum_k d_k k^2 e^{-\frac{k^2(1-\rho^2)}{2}} \cos(k\rho X) \\
\frac{1}{2} E[\nabla^2_{xx} \mu(\boldsymbol{U})|X] &= -\sum_k \frac{c_k k^2}{2} \cos(kX) \\
\frac{1}{2} E[\nabla^2_{zz} \mu(\boldsymbol{U})|X] &= -\sum_k \frac{d_k k^2}{2} e^{-\frac{k^2(1-\rho^2)}{2}} \cos(k\rho X)
\end{aligned}$$

Therefore,

$$h(X) = E[\mu(\boldsymbol{U})|X] = \sum_k \left\{ c_k \cos(kX) + d_k e^{-\frac{k^2(1-\rho^2)}{2}} \cos(k\rho X) \right\}$$

and

$$
\begin{aligned}
g^*(X) &= E[A(\boldsymbol{U})|X] \\
&= \sum_k \left\{ c_k k X \sin(kX) - \frac{c_k k^2}{2}\cos(kX) + \frac{d_k k^2}{2} e^{-\frac{k^2(1-\rho^2)}{2}} \cos(k\rho X) \right\}.
\end{aligned}
$$

In this case, $E[Xg^*(X)] = 0$; hence, we conclude that X is useless in approximating g^* and using $g = 1$ in constraint (9) is the same as using $g = (1, X)$. It can be shown that $E[h(X)g^*(X)] \neq 0$. Thus, we conclude that using $g = (1, h)$ in constraint (9) is better than using $g = (1, X)$.

4 Under Missing At Random Scenery

In this section, we consider the external data having missing covariate $\boldsymbol{Z}$. We observe $(Y_i, \boldsymbol{X}_i, R_i\boldsymbol{Z}_i, R_i)$, where R_i stand for $\boldsymbol{Z}$ is observed or not. Let the propensity score $\pi(Y, X) := P(R = 1|Y, X)$. In the general MAR setting, the most chanllenge part is the distribution of internal date $(X, Z)|R = 1$ and external date $(X, Z)|R = 0$ maynot be the same distribution. To make sure the identibility, we further assume $Y = \mu(\boldsymbol{X}, \boldsymbol{Z}) + \sigma(\boldsymbol{X})\epsilon$, where ϵ is error with mean 0 variance 1 and independent to $\boldsymbol{U}$. This can make sure $\mu(\boldsymbol{U}) = E[Y|\boldsymbol{U}; R = 1] = E[Y|\boldsymbol{U}; R = 0]$.

In order to make sure CKR hold, we have to figure out whether the constraints equation (9) work. Written it as MAR setting

$$(9) = E[\{E(Y|\boldsymbol{X}; R = 0) - \mu(\boldsymbol{U})\}g(\boldsymbol{X})|R = 1].$$

Note that $h = \{E(Y|\boldsymbol{X}; R = 0)$, which can only secure

$$E[\{E(Y|\boldsymbol{X}; R = 0) - \mu(\boldsymbol{U})\}g(\boldsymbol{X})|R = 0] = 0. \tag{14}$$

Hence, in the MAR scenery, (9) may not be zero.

There are three insteresting case we want to discuss. The first one is missing completely at random (MCAR). In this case, the distribution between internal data set $(X, Z)|R = 1$ and external data set $(X, Z)|R = 0$ is the same. Hence, it is the case we discuss in previous section.

Another case is $R \perp Z|X$. That mean the propensity score is only depend on $\boldsymbol{X}$. We can also get the relationship of the density,

$$\frac{f(\boldsymbol{U}|R=1)}{f(\boldsymbol{U}|R=0)} = \frac{\pi(\boldsymbol{X})P(R=0)}{(1-\pi(\boldsymbol{X}))P(R=1)}.$$

Untilize this relationship we can get

$$\begin{aligned}
(9) &= \int \{h(\boldsymbol{x}) - \mu(\boldsymbol{u})\}g(\boldsymbol{x})f(\boldsymbol{u}|R=1)d\boldsymbol{u} \\
&= \int \{h(\boldsymbol{x}) - \mu(\boldsymbol{u})\}g(\boldsymbol{x})\frac{f(\boldsymbol{u}|R=1)}{f(\boldsymbol{u}|R=0)}f(\boldsymbol{u}|R=0)d\boldsymbol{u} \\
&= \int \{h(\boldsymbol{x}) - \mu(\boldsymbol{u})\}g(\boldsymbol{x})\frac{\pi(\boldsymbol{x})P(R=0)}{(1-\pi(\boldsymbol{x}))P(R=1)}f(\boldsymbol{u}|R=0)d\boldsymbol{u} \\
&= E\left[\{h(\boldsymbol{X}) - \mu(\boldsymbol{U})\}g(\boldsymbol{X})\frac{\pi(\boldsymbol{X})P(R=0)}{(1-\pi(\boldsymbol{X}))P(R=1)}\middle| R=0\right].
\end{aligned}$$

In the scenery of Section 2, h is a given summary statistics, $g(\boldsymbol{X}) = \boldsymbol{X}$. (9) may not be zero since (14) cannot hold for general constraints. However, under scenery of Section 3, (14) can hold with $g = g(\boldsymbol{X})\frac{\pi(\boldsymbol{X})}{(1-\pi(\boldsymbol{X}))}$, get $(9) = 0$. From this we can see even the distribution of internal and external data are not the same, (9) still holds. Hence, Theorem 3.1 can holds without any motification. Furthermore, we even don't need to put effert on estimating the propensity score $\pi(\boldsymbol{X})$.

The last one is the general case. The propensity score is depend on both $\boldsymbol{X}$ and Y. The same procedure can get

$$(9) = E\left[\{h(\boldsymbol{X}) - \mu(\boldsymbol{U})\}g(\boldsymbol{X})\frac{\pi(\boldsymbol{U})P(R=0)}{(1-\pi(\boldsymbol{U}))P(R=1)}\middle| R=0\right],$$

which is not zero for sure. Hence, we have no choose but estimate the propensity score π. Here, we propose the modification. Let constraint funtion (10) be

$$\sum_{i=1}^{n}\{\widehat{\mu}_i - \widehat{h}(\boldsymbol{X}_i)\}g(\boldsymbol{X}_i)\frac{(1-\widehat{\pi}(\boldsymbol{U}_i))}{\widehat{\pi}(\boldsymbol{U}_i)} = 0, \tag{15}$$

where the $\widehat{\pi}$ is estimated propensity score. Then, we can get rid off the extra term in the above equation. Then, get (9) is zero. Here, we state the modification theorem with having exact propensity score π.

Theorem 4.1. *Assume the conditions in Theorem 3.1. Under MAR scenery, given propensity score* π*, derive CKR with modification constraints*(15)*. Then, as* $n \to \infty$*,*

$$\sqrt{nb}\{\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \to N\big(B_{CK}(\boldsymbol{u}), V_{CK}(\boldsymbol{u})\big) \quad \textit{in distribution,}$$

where

$$B_{CK}(\boldsymbol{u}) = c^{1/2}[(1+\gamma^2)A(\boldsymbol{u})-\gamma^2 frac(1-\pi(\boldsymbol{u}^*))\pi(\boldsymbol{u}^*)g(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}|R=0],$$

$\widetilde{\boldsymbol{\Sigma}}_g$ *is* $E\{g(\boldsymbol{X})g(\boldsymbol{X})^\top|R=0\}$*, and* $V_{CK}(\boldsymbol{u})$ *and* $A(\boldsymbol{u})$ *are the same as those in Theorem 2.1.*

5 Simulation Results

In this section, we provide the simulation result. We generate the internal data $(Y, \boldsymbol{X}, \boldsymbol{Z})$ with sample size n and the external data with sampe size 1000 but the variables Z is missing.

The data have the relationship

$$Y = \mu(\boldsymbol{X}, \boldsymbol{Z}) + \epsilon,$$

where $\epsilon \sim N(0, V)$, and $\mu(\cdot)$ would be specified latter. And, we consider two different covariate generating method. First one is unbounded random variables. X and Z are bivariate normal distribution with mean 0, variance 1, and covariance 0.5. Second one is bounded varaibles. X and Z have following relationship.

$$\begin{aligned} X =& BW_1 + (1-B)W_2 \\ Z =& BW_1 + (1-B)W_3, \end{aligned}$$

where W_1, W_2, W_3 are i.i.d uniform $[-1, 1]$, and B is uniform $[0, 1]$ independent to them. From, this setting X, Z are uniform $[-1, 1]$ with corelation 0.5.

We apply sevaral methods to get the estimator $\widehat{\mu}$, including standard Kernel Regression(KR), Double Kernel Regression (DKR), and Constrainted Kernel Regression (CKR). All these kernel type methods we use Gaussian kernels. In Constrainted Kernel Regression, we have six different methods CKR(1), CKR(1,X), CKR(1,$\widehat{h}$), CKR($\widehat{g^*}$), CKR(g^*), and CKRm(1,X). The function in parenthesis of CKR($\cdot$) and CKRm($\cdot$) indicate which constraints we use. In CKR($\cdot$) we use standard kernel regression to estimate the submodel h; hence, all the continuous function of $\boldsymbol{X}$ can be the constraint. In simulation study, we consider the constraints g^* having theoretical best performance in global, and its estimator $\widehat{g}^*$. Also, CKR(1), CKR(1,X), CKR(1,$\widehat{h}$) are easily derived and usually have decent performance locally. For CKRs($\cdot$), instead of applying kernel regression, we use the information of summary statistics derived from the external data set via linear model(see Section 2). Hence, in this case, we consider the only reasonable constraint which is $(1, X)$.

The way to evaluate the performace of method is mean intergral square error(MISE). The formula like this

$$\frac{1}{\sharp\text{Repetition}} \sum_{Repetition} \frac{\sum_{u\in \text{ test data}}(\widehat{\mu}(u) - \mu(u))^2}{\sharp\text{test data}}.$$

In each repetition, we would generate data agian and $\widehat{\mu}$ would be re-estimated for each methods. And the number of repetition is 200. The next question is how to choose the test data. We have two way to approach it. First one is fixed grid points on $[-1, 1]^2$. We divid $[-1, 1]$ into 11 points with equal space. Hence, we have 121 fixed grid points in total. We refer this evaluation as local evaluation since the evalued points are in the middle. For another one, we sample 121 data from the covariate from internal data set without replacement. In this case, the test data have the same distribution as covariate. Hence, it approximates AMISE in section 2.2. So, we would expect CKR(g^*) would have the best performance. We refer this evaluation as global evaluation since the evalued points are at whole space.

Tuning bandwidths is a key in kernel types method. We consider two way to select bandwidths. The first one is call "Best bandwidth". In this case, we evaluate MISE in a pool of bandwidths and we display the one have the best performance (minimal MISE) for each methods. The second one we select bandwidths from a pool of bandwidths vis 10-fold cross validation. The first approach show that there exist the best bandwidth, and the second one says even in the real data, it is possible to have decent bandwidths. In CKR, we have several bandwidths; however, in the simulation we only tune the bandwidth b and l in (5). For saving the computation times, we don't tune bandwidths of $\widehat{h}$, and $\widehat{g}^*$. The reason is after tuning $\widehat{h}$, and $\widehat{g}^*$, they would be closer to h, and g^*. And it is obviously that the closer they are, the better performance they have. And, the simulations can show that even under these unfair situation CKR can have decent performance. Furthermore, in order to enlight the benefit of external information, we calculate the percentage of improvement via the following formula.

$$\text{Improve} = \frac{\min\{MISE(DKR), MISE(KR)\} - \min\{MISE(CKR(\cdot))|\text{for all CKR methods}\}}{\min\{MISE(DKR), MISE(KR)\}}.$$

Table 1:

	DKR	CKR(1)	CKR(1,X)	CKR(1,$\widehat{h}$)	CKR($\widehat{g}^*$)	CKR(g^*)	CKRs(1,X)	KR
Model: $Y = X^3 + Z^2$; $n = 200$; $sd = 3$; Best bandwidth								
Test data: Fixed grid points on $[-1, 1]^2$; Improve: 0.1763; Improve-CV:0.1140								
MISE	0.208	0.175	0.171	0.181	0.205	0.206	0.243	0.21
MISE-CV	0.384	0.382	0.365	0.341	0.388	0.355	0.432	0.412
Test data: $\boldsymbol{U}$; Improve: 0.2011; Improve-CV: 0.0918								
MISE	1.056	1.045	0.924	0.912	1.055	0.843	1.152	1.081
MISE-CV	1.181	1.165	1.148	1.073	1.19	1.176	1.409	1.239
Model: $Y = 1/2Cos(2X) + Cos(Z)$; $n = 200$; $sd = 3$; Best bandwidth								
Test data: Fixed grid points on $[-1, 1]^2$; Improve: 0.4998; Improve-CV: 0.265								
MISE	0.153	0.111	0.108	0.076	0.148	0.132	0.100	0.153
MISE-CV	0.162	0.141	0.138	0.117	0.164	0.151	0.134	0.160
Test data: $\boldsymbol{U}$; Improve: 0.2261; Improve-CV: 0.1823								
MISE	0.260	0.220	0.225	0.222	0.259	0.210	0.201	0.266
MISE-CV	0.343	0.297	0.290	0.323	0.341	0.282	0.274	0.335
Model: $Y = Cos(X) + Cos(Z)$; $n = 200$; $sd = 3$; Best bandwidth								
Test data: Fixed grid points on $[-1, 1]^2$; Improve: 0.2175; Improve-CV: 0.3011								
MISE	0.132	0.102	0.100	0.070	0.129	0.131	0.089	0.135
MISE-CV	0.142	0.123	0.122	0.099	0.145	0.151	0.121	0.142
Test data: $\boldsymbol{U}$; Improve: 0.1726; Improve-CV:0.2256								
MISE	0.258	0.213	0.216	0.194	0.259	0.176	0.203	0.259
MISE-CV	0.346	0.300	0.300	0.305	0.345	0.260	0.281	0.335

From table 1, while the test data having distribution $\boldsymbol{U}$, the g^* is the best one among all the orther constraints. This result can be expected from our derivation in Section 3.2. Futhermore, the estimator constraint $\widehat{g}^*$ has the terrible performance. That is also predictable since the algorithm depends on the estimate of second derivatives which usually have slow convergence rate. Another interesting thing is CKRs(1,X) sometimes has better performance than CKR(g^*). That is because, summary statistics is derived from parametric linear models which convergence faster than the one derived from kernel regression which is non-parametric. On the contrary, while the testing data are fixed grim points in $[-1,1]^2$, CKR(1,X) and CKR(1,$\widehat{h}$) have better performance. This result show that if we want to focus on locally estimations, it is better not to use the constaint g^*.

Table 2: Impact of Variance

	DKR	CKR(1)	CKR(1,X)	CKR(1,$\widehat{h}$)	CKR($\widehat{g^*}$)	CKR(g^*)	CKRs(1,X)	KR
Model: $Y = X^3 + XZ + Z^2$; Best bandwidth; Test data: Fixed grid points on $[-1,1]^2$								
$n = 200$; $sd = 5$; Improve: 0.217; Improve-CV: 0.156								
MISE	0.472	0.377	0.383	0.369	0.464	0.472	0.458	0.483
MISE-CV	0.734	0.718	0.723	0.598	0.735	0.648	0.735	0.708
$n = 200$; $sd = 3$; Improve: 0.088; Improve-CV:0.0212								
MISE	0.251	0.231	0.244	0.229	0.250	0.251	0.338	0.251
MISE-CV	0.397	0.414	0.426	0.359	0.400	0.369	0.486	0.367
Model: $Y = X^3 + XZ + Z^2$; Test data: $\boldsymbol{U}$								
$n = 200$; $sd = 5$; Improve: 0.1999; Improve-CV: 0.0714								
MISE	2.068	2.021	1.888	1. 698	2.075	1.654	2.214	2.068
MISE-CV	2.553	2.520	2.350	2.243	2.562	2.316	2.662	2.416
$n = 200$; $sd = 3$; Improve: 0.0733; Improve-CV: 0								
MISE	1.099	1.102	1.066	1.018	1.102	1.020	1.437	1.117
MISE-CV	1.270	1.276	1.294	1.329	1.262	1.410	1.624	1.208

From Table 2, we can see as the standard deviation equal to 3, the improvement is not significnat. However, as the standard deviation equal to 5, we have huge improvement. Theorem 2.1 can explain this phenomenon. If we increase $\sigma^2(\boldsymbol{u})$, the kernels type methods tends to choose the bandwidth such that B_t^2 becomes larger and $V_t^2/\sigma^2(\boldsymbol{u})$ becomes smaller. Because while $\sigma^2(\boldsymbol{u})$ is large, the weighted of variance V_t is relative large. Since the improvement of CKR is on having smaller bias, we can say larger σ^2 implies larger Bias B_t^2, and larger Bias B_t^2 implies better performance of CKR.

Table 3: Bounded Design

	DKR	CKR(1)	CKR(1,X)	CKR(1,$\widehat{h}$)	CKR($\widehat{g^*}$)	CKRs(1,X)	KR
Model: $Y = X^3 + Z^2$; $n = 200$; $sd = 3$; Best bandwidth							
Test data: Fixed grid points on $[-1,1]^2$; Improve: 0.3080; Improve-CV:0.2025							
MISE	0.37	0.336	0.268	0.271	0.370	0.256	0.37
MISE-CV	0.462	0.420	0.380	0.399	0.444	0.369	0.511
Test data: $\boldsymbol{U}$; Improve: 0.4337; Improve-CV:0.2883							
MISE	0.143	0.102	0.086	0.085	0.139	0.081	0.143
MISE-CV	0.182	0.154	0.125	0.152	0.174	0.119	0.175
Model: $Y = X^3 + XZ + Z^2$; $n = 200$; $sd = 3$; Best bandwidth							
Test data: Fixed grid points on $[-1,1]^2$; Improve: 0.2364; Improve-CV:0.2661							
MISE	0.409	0.375	0.319	0.314	0.405	0.308	0.404
MISE-CV	0.226	0.202	0.178	0.195	0.219	0.164	0.224
Test data: $\boldsymbol{U}$; Improve: 0.3419; Improve-CV:0.17387							
MISE	0.194	0.152	0.130	0.135	0.189	0.125	0.190
MISE-CV	0.51	0.468	0.432	0.465	0.502	0.422	0.555
Model: $Y = 1/2Cos(2X) + Cos(Z)$; $n = 200$; $sd = 1$; Best bandwidth							
Test data: Fixed grid points on $[-1,1]^2$; Improve: 0.4053; Improve-CV:0.3128							
MISE	0.089	0.083	0.083	0.053	0.090	0.081	0.089
MISE-CV	0.108	0.102	0.100	0.074	0.108	0.098	0.108
Test data: $\boldsymbol{U}$; Improve: 0.5493; Improve-CV: 0.3957							
MISE	0.032	0.027	0.025	0.014	0.03	0.025	0.032
MISE-CV	0.037	0.034	0.032	0.022	0.036	0.032	0.037
Model: $Y = Cos(X) + Cos(Z)$; $n = 200$; $sd = 1$; Best bandwidth							
Test data: Fixed grid points on $[-1,1]^2$; Improve:0.3137; Improve-CV:0.1915							
MISE	0.067	0.061	0.061	0.046	0.067	0.059	0.069
MISE-CV	0.084	0.078	0.075	0.067	0.082	0.074	0.083
Test data: $\boldsymbol{U}$; Improve: 0.4639; Improve-CV:0.30701							
MISE	0.026	0.021	0.020	0.014	0.027	0.019	0.027
MISE-CV	0.030	0.026	0.024	0.021	0.028	0.024	0.030

6 Appendix

6.1 Proof of Theorem 2.1 and Theorem 3.1

Proof of Theorem 3.1. Before we start to begin the proof. We would introduce some notation. Define $\widehat{f}_l(\boldsymbol{u}^*) := \frac{1}{n}\sum_i^n \kappa_l(\boldsymbol{u}^* - \boldsymbol{U}_i)$. B_l is a $n \times n$ diagonal matrix with $i, i-th$ element being $\widehat{f}_l(\boldsymbol{U}_i)$, Δ is a $n \times n$ matrix with $i, j-th$ element being $\frac{1}{n}\kappa(\boldsymbol{U}_i - \boldsymbol{U}_j)$. Then $\widehat{\boldsymbol{\mu}}_k = B_l^{-1}\Delta\boldsymbol{Y}$. $\boldsymbol{\delta}_b(\boldsymbol{u}^*) = (\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_1), \ldots, \kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_n))^\top$. For convenience we also define $\mathcal{U}_n = \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n\}$. Let $\boldsymbol{G}$ be an $n \times n$ matrix with i-th row being $g(\boldsymbol{X}_i)$. And $\boldsymbol{P}$ is the projection matrix $\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1}\boldsymbol{G}^\top$.

Consider following decomposition at a fixed point $\boldsymbol{u}^*$.

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_{CK}(\boldsymbol{u}^*) - \mu(\boldsymbol{u}^*) =& \frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{B}_l^{-1}\boldsymbol{\Delta}_l(\boldsymbol{Y}-\boldsymbol{\mu}) && (16)\\ &+\frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top\boldsymbol{P}(\boldsymbol{h}-\mu) && (17)\\ &+\frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top(\boldsymbol{\mu}-\mu(\boldsymbol{u}^*)) && (18)\\ &+\frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top(\boldsymbol{B}_l^{-1}\boldsymbol{\Delta}_l\boldsymbol{\mu}-\boldsymbol{\mu}) && (19)\\ &-\frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top\boldsymbol{P}(\boldsymbol{B}_l^{-1}\boldsymbol{\Delta}_l\boldsymbol{\mu}-\boldsymbol{\mu}). && (20)\\ &+\frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top\boldsymbol{P}(\widehat{\boldsymbol{h}}-\boldsymbol{h}) && (21)\end{aligned}$$

(16) is the variance term. And (17) − (21) are the bias terms. Our goal is show that (16) convergence in distribution to Gaussian via Liderberg Central limit theorem. And find the asymptotic bias of the rest terms.

Step 1: Variance

$$(16) = \frac{1}{n\widehat{f}_b(\boldsymbol{u}^*)}\boldsymbol{\delta}_b(\boldsymbol{u}^*)^\top(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{B}_l^{-1}\boldsymbol{\Delta}_l\boldsymbol{\epsilon} := \frac{1}{n}\sum_i^n W_j\epsilon_j.$$

Note that conditional on $\mathcal{U}_n$, W_j are constants. Hence, we only have to proof the following two claims.

Claim 1:

$$\frac{b}{n}\sum_{j=1}^n W_j^2\sigma^2(\boldsymbol{U}_j) \to V_{CK}(\boldsymbol{u}^*),$$

where $\sigma^2(\boldsymbol{U}_j) = E[\epsilon^2|\boldsymbol{U}=\boldsymbol{U}_j]$.

Claim 2:

$$\max_{i\in\mathcal{U}}\frac{\max_j W_j^2}{\sum_{j=1}^n W_j^2} \to 0.$$

These two claims together with Linderberg's CLT (Corollary in [12]) can imply

$$\sqrt{nb}\frac{1}{n}\sum_i^n W_j\epsilon_j|\mathcal{U}_n \xrightarrow{d} N\{0, V_{CK}(\boldsymbol{u}^*)\}.$$

Since the convergence is independent to $\boldsymbol{U}_n$, the unconditional sequence are also convergence weakly to that normal distribution.

Now, we check Claim 1. Let $W_j = W_{1j} + W_{2j}$, where W_{1j} and W_{2j} are defined in below.

$$W_{1j} = \frac{1}{\widehat{f}_b(\boldsymbol{u}^*)} \frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\widehat{f}_l(\boldsymbol{U}_i)}$$

$$W_{2j} = \frac{1}{\widehat{f}_b(\boldsymbol{u}^*)} \left[\frac{1}{n}\sum_{i=1}^{n} \kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i) g(\boldsymbol{X}_i)^\top\right] \left[\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) g(\boldsymbol{X}_i)^\top\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)\frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\widehat{f}_l(\boldsymbol{U}_i)}\right]$$

From Lemma 7.4, we have following unifomly convergence.

$$(W_{1j} + W_{2j})^2 = \left[\frac{1}{b^p}\frac{1}{f(\boldsymbol{u}^*)}\int \kappa(\boldsymbol{v})\kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{U}_j}{b} - \frac{l}{b}\right) d\boldsymbol{v}\right]^2 (1 + o_p(1)),$$

where $o_p(1)$ is uniform in $j = 1, \ldots, n$. Hence,

$$\frac{b}{n}\sum_{j=1}^{n} W_j^2 \sigma^2(\boldsymbol{U}_j) = \left[\frac{1}{n}\sum_{j}^{n} \frac{1}{b^p}\left\{\frac{1}{f(\boldsymbol{u}^*)}\int \kappa(\boldsymbol{v})\kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{U}_j}{b} - \boldsymbol{v}\frac{l}{b}\right) d\boldsymbol{v}\right\}^2 \sigma^2(\boldsymbol{U}_j)\right](1 + o_p(1))$$

From Law of large number and the Lipschitz's continuity of $\sigma^2(\cdot)$, $f_U(\cdot)$,

$$\frac{b}{n}\sum_{j=1}^{n} W_j^2 \sigma^2(\boldsymbol{U}_j) = V_{CK}(\boldsymbol{u}^*) + o_p(1).$$

We have claimed Claim 1. From Lemma 7.4, we know that $\sup\limits_{j=1,\ldots,n} W_j^2 = O_p(\frac{1}{b^2})$. From the proof of Claim 1, $\sum_{j=1}^{n} W_j^2 = O_p(n/b)$. Hence, Claim 2 is satisfied.

Step 2: Bias From (39)(40) in Lemma 7.3, Condition(C1), and Central Limit Theorm, we have

$$\begin{aligned}(17) =& \frac{1}{\widehat{f}_b(\boldsymbol{u}^*)} \left[\frac{1}{n}\sum_{i}^{n} \kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i) g(\boldsymbol{X}_i)\right] \left[\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) g(\boldsymbol{X}_i)^\top\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)(h(X_i) - \mu(\boldsymbol{U}_i))\right] \\ =& O_p(1) * O_p(1) * O_p(1) * O_p(1/\sqrt{n}) = O_p(1/\sqrt{n}),\end{aligned}$$

Hence, $\sqrt{nb}(17) = O_p(b) = o_p(1)$, which imply (17) is eligiable. From Lemma 7.5, and (A4),

$$\sqrt{nb}(18) = \sqrt{nb}b^{2/p}A(\boldsymbol{u}^*)(1+o_p(1)) = \sqrt{c}A(\boldsymbol{u}^*)(1+o_p(1)).$$

$$\sqrt{nb}(19) = \frac{\sqrt{nb}}{n\widehat{f}_b(\boldsymbol{u}^*)}\sum_{j=1}^{n}\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_j)\left[\frac{1}{n\widehat{f}(\boldsymbol{U}_j)}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)(\mu(\boldsymbol{U}_i) - \mu(\boldsymbol{U}_j))\right]$$

From (39) in Lemma 7.3, Lemma 7.5 and Assumption (A4),

$$= \left[\frac{\sqrt{c}\gamma^2}{nf_b(\boldsymbol{u}^*)}\sum_{j=1}^{n}\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_j)A(\boldsymbol{U}_j)\right](1+o_p(1))$$

From Lemma 7.2 and continuity of $A(\cdot)$,

$$= \sqrt{c}\gamma^2 A(\boldsymbol{u}^*)(1+o_p(1)).$$

(20) is equal to

$$\sum_{i}^{n}\frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)g(\boldsymbol{X}_i)^\top}{n\widehat{f}_b(\boldsymbol{u}^*)}\left[\frac{1}{n}\sum_{i=1}^{n}g(\boldsymbol{X}_i)g(\boldsymbol{X}_i)^\top\right]^{-1}\frac{1}{n}\sum_{j=1}^{n}\frac{g(\boldsymbol{X}_j)}{n\widehat{f}(\boldsymbol{U}_j)}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)(\mu(\boldsymbol{U}_i) - \mu(\boldsymbol{U}_j))$$

From (39), (40), and Law of Large Number

$$= \left[g(\boldsymbol{x}^*)^\top\Sigma_g^{-1}\frac{1}{n}\sum_{j=1}^{n}\frac{g(\boldsymbol{X}_j)}{nf(\boldsymbol{U}_j)}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)(\mu(\boldsymbol{U}_i) - \mu(\boldsymbol{U}_j))\right](1+o_p(1))$$

From Lemma 7.5

$$= \left[g(\boldsymbol{x}^*)^\top\Sigma_g^{-1}\frac{l^{2/p}}{n}\sum_{j=1}^{n}g(\boldsymbol{X}_j)A(\boldsymbol{U}_j)\right](1+o_p(1))$$

From Law of Large Number

$$= l^{2/p}\left[g(\boldsymbol{x}^*)^\top\Sigma_g^{-1}E[g(\boldsymbol{X})A(\boldsymbol{U})]\right](1+o_p(1)) = l^{2/p}g(\boldsymbol{x}^*)^\top\boldsymbol{\zeta}(1+o_p(1))$$

Hence, $\sqrt{nb}(20) = \sqrt{c}\gamma^2 g(\boldsymbol{x}^*)^\top \boldsymbol{\zeta}(1 + o_p(1))$. (21) is equal to

$$\sum_{i}^{n} \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i) g(\boldsymbol{X}_i)^\top}{n\widehat{f}_b(\boldsymbol{u}^*)} \left[\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) g(\boldsymbol{X}_i)^\top\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)(\widehat{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i))$$

From (39), (40),

$$= \left[g(\boldsymbol{x}^*)^\top \Sigma_g^{-1} \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)(\widehat{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i))\right](1 + o_p(1))$$

There is a constant M, such that

$$\leq (1 + o_p(1)) M \sup_{j=1,\ldots,n} |\widehat{h}(\boldsymbol{U}_j) - h(\boldsymbol{U}_j)|.$$

From Assumption (A4), the sufficient condition for $\sqrt{nb}(21) = o_p(1)$ is $\sup_{j=1,\ldots,n} |\widehat{h}(\boldsymbol{U}_j) - h(\boldsymbol{U}_j)| = o_p(n^{-2/(p+4)})$. From the assumption (A1)(A2')(A3')(A4') we would be able to apply Lemma 7.2 which claim $\|\widehat{h} - h\|_\infty = (\frac{\log n}{n})^{2/(q+4)}$. Then, assumption (A5) can make sure the unifom convergence rate of $\widehat{h}$ is of the order $o_p(n^{-2/(p+4)})$. Hence, $\sqrt{nb}(21) = o_p(1)$. We complete the proof.

□

Proof of Theorem 2.1. The only difference between Theorem 3.1 and Theorem 2.1 is (21). From the proof in Theorem 3.1, it is suffice to claim $\sup_{j=1,\ldots,n} |\widehat{h}(\boldsymbol{U}_j) - h(\boldsymbol{U}_j)| = o_p(n^{-2/(p+4)})$. Since $\widehat{h}$ is the linear estimator,

$$\widehat{h}(\boldsymbol{X}) - h(\boldsymbol{X}) = \boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

The well-known result is $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is of order $O_p(1/\sqrt{m})$. Hence, from the boundness of $\boldsymbol{X}$ (A1) and assumption (A5), the unifom convergence rate of $\widehat{h}$ is of the order $o_p(n^{-2/(p+4)})$. Hence, $\sqrt{nb}(21) = o_p(1)$. We complete the proof.

□

6.2 Proof of Theorem 2.2

Let $AMISE(\widehat{\mu}_{CK})(\gamma)$ be the AMISE for a Constrainted Kernel Regression with parametor γ. From the result in Theorem 2.1, it is obvious that

$$\lim_{\gamma \to 0} AMISE(\widehat{\mu}_{CK})(\gamma) = AMISE(\widehat{\mu}_{KR}).$$

Hence,

$$\inf_{\gamma>0} AMISE(\widehat{\mu}_{CK})(\gamma) \leq AMISE(\widehat{\mu}_{KR}).$$

The derivative of asymptotic $AMISE(\widehat{\mu}_{CK})(\cdot)$ is

$$\begin{aligned}
&2c(1+\gamma^2)\gamma E\left\{A(\boldsymbol{U}) - \boldsymbol{X}^\top \boldsymbol{\zeta}\right\}^2 \\
&+ E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) \left(\int -\nabla\kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)^\top \boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) d\boldsymbol{w},
\end{aligned}$$

where $\zeta := \boldsymbol{\Sigma}_X^{-1} E\{\boldsymbol{X}A(\boldsymbol{U})$. Since the mean of $\kappa(\cdot)$ is zero, the derivative is zero taking value at $\gamma = 0$.

$$0 + E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \kappa(\boldsymbol{w})\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) \left(\int -\nabla\kappa(\boldsymbol{w})^\top \boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) d\boldsymbol{w} = 0.$$

If the second derivative is negative, then we can say there is a γ such the CKR is better than the original KR with respect to AMISE. The second derivative of asymptotic $AMISE(\widehat{\mu}_{CK})(\cdot)$ is

$$\begin{aligned}
&2cE\{A(\boldsymbol{U}) - g(\boldsymbol{X})^\top \boldsymbol{\zeta}\}^2(1+3\gamma^2) \\
&+E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \nabla\kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)^\top \boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right)^2 d\boldsymbol{w} \\
&+E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) \left(\int \boldsymbol{v}^\top \nabla^2\kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) d\boldsymbol{w}.
\end{aligned}$$

Taking value at $\gamma = 0$

$$\begin{aligned}
&2cE\{A(\boldsymbol{U}) - g(\boldsymbol{X})^\top \boldsymbol{\zeta}\}^2 + E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \nabla\kappa(\boldsymbol{w})^\top \boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right)^2 d\boldsymbol{w} \\
&+E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \left(\int \kappa(\boldsymbol{w})\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) \left(\int \boldsymbol{v}^\top \nabla^2\kappa(\boldsymbol{w})\boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v} \right) d\boldsymbol{w} \\
&=2cE\{A(\boldsymbol{U}) - g(\boldsymbol{X})^\top \boldsymbol{\zeta}\}^2 + 0 + E\left\{\frac{2\sigma^2(\boldsymbol{U})}{f_U(\boldsymbol{U})}\right\} \int \boldsymbol{v}^\top \left[\int \nabla^2\kappa(\boldsymbol{w})\kappa(\boldsymbol{w}) d\boldsymbol{w} \right] \boldsymbol{v}\kappa(\boldsymbol{v}) d\boldsymbol{v}
\end{aligned}$$

Hence, as long as $\int \nabla^2\kappa(\boldsymbol{w})\kappa(\boldsymbol{w})d\boldsymbol{w}$ is negative definite there exsit a constant $c^* >$ such that for all $c < c^*$ the second derivative of asymptotic $AMISE(\widehat{\mu}_{CK})$ is negative. That mean $\gamma = 0$ is local maximal. So, there exist a constant $\gamma > 0$ such that

$$AMISE(\widehat{\mu}_{CK})(\gamma) < AMISE(\widehat{\mu}_{KR}).$$

6.3 Proof of Theorem 3.2

It is suffice to show that

$$\sqrt{nb}\frac{1}{n}\sum_{i}^{n} g_n(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\} = o_p(1).$$

From (A4) it is equivalent to

$$\frac{1}{n}\sum_{i}^{n} g_n(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\} = o_p(n^{-2/(p+4)}).$$

Hence, it is suffice to claim the condition (49)-(51) in Lemma 7.6. Define the leave-$i-$out estimator would be

$$g_n^{-i}(\boldsymbol{x}) := \frac{\sum_{k\neq i} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k)\widehat{A}^{-i}(\boldsymbol{U}_k)}{\sum_{k\neq i} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k)},$$

where $\widehat{A}^{-i}(\cdot)$ is a estimator of $A(\cdot)$ without using sample i. In this case, we have the upperbound,

$$|g_n(\boldsymbol{X}_i) - g_n^{-i}(\boldsymbol{X}_i)| \leq \frac{\sup \kappa(\cdot)}{n\delta^q}\left|\frac{\widehat{A}(\boldsymbol{U}_i) - g_n^{-i}(\boldsymbol{X}_i)}{\widehat{f}_x(\boldsymbol{X}_i)}\right| + \max_{j=1,\ldots,n} |\widehat{A}(\boldsymbol{U}_j) - \widehat{A}^{-i}(\boldsymbol{U}_j)|.$$

From (71) in Lemma 7.8, the second term is of order $O_p(n^{-6/(p+8)})$, which is $o_p(n^{-2/(p+4)})$. And it is obviously that $\max\limits_{i=1,\ldots,n}\left|\frac{\widehat{A}(\boldsymbol{U}_i)-g_n^{-i}(\boldsymbol{X}_i)}{\widehat{f}_x(\boldsymbol{X}_i)}\right|$ is $O_p(1)$. And $1/n\delta^q = O(n^{4/(4+q)})$, which is also $o(n^{-2/(p+4)})$. Hence, (49) is satisfid. from (72) in Lemma 7.8, $E\{\max\limits_{j=1,\ldots,n} |\widehat{A}(\boldsymbol{U}_j) - \widehat{A}^{-i}(\boldsymbol{U}_j)|\}^2$ is order $O_p(n^{-6/(p+8)})$. From (73) in Lemma 7.8, and (34) in Lemma 7.2 $\max\limits_{j=1,\ldots,n} | \left|\frac{\widehat{A}(\boldsymbol{U}_i)-g_n^{-i}(\boldsymbol{X}_i)}{\widehat{f}_x(\boldsymbol{X}_i)}\right|$ are L_2. Then, we can conclude that (51) is satisfied. For (50), we need to define

$$\widetilde{g}_n(\boldsymbol{x}) := \frac{\sum_{k=1}^{n} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k)A(\boldsymbol{U}_k)}{\sum_{k=1}^{n} \kappa_\delta(\boldsymbol{x} - \boldsymbol{X}_k)}.$$

And it is equivelant to show that

$$E\int\{g_n(\boldsymbol{x}) - g(\boldsymbol{x})\}^2 f_x(\boldsymbol{x})d\boldsymbol{x} = o(1).$$

$$E\int\{g_n(\boldsymbol{x})-g(\boldsymbol{x})\}^2 f_x(\boldsymbol{x})d\boldsymbol{x} = E\int\{\widetilde{g}_n(\boldsymbol{x})-g(\boldsymbol{x})\}^2 f_x(\boldsymbol{x})d\boldsymbol{x} \tag{22}$$

$$+E\int\{g_n(\boldsymbol{x})-\widetilde{g}_n(\boldsymbol{x})\}^2 f_x(\boldsymbol{x})d\boldsymbol{x} \tag{23}$$

(22) is stadard mean integral square error which is $o(1)$(Theorem 2 in [5])

.

$$(23) \leq E \max_{j=1,\ldots,n} |\widehat{A}(\boldsymbol{U}_j) - A(\boldsymbol{U}_j)|^2;$$

hence, the desired result follows (73) in Lemma 7.8.

6.4 Proof of Theorem 3.3

It is suffice to show that

$$\frac{1}{n}\sum_i^n \widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\} = o_p(\frac{1}{\sqrt{nb}}).$$

Since the external data set is independent to the internal data set,

$E[\widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}] = E\{E[\widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}|\boldsymbol{X}_{n+1},\ldots,\boldsymbol{X}_{n+m}]\} = 0$

Now, check the variance.

$$\begin{aligned}
&Var\left[\frac{1}{n}\sum_i^n \widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\right] \\
=&E\left(Var\left[\frac{1}{n}\sum_i^n \widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}|\boldsymbol{X}_{n+1},\ldots,\boldsymbol{X}_{n+m}\right]\right) \\
+&Var\left(E\left[\frac{1}{n}\sum_i^n \widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}|\boldsymbol{X}_{n+1},\ldots,\boldsymbol{X}_{n+m}\right]\right) \\
=&E\left(\frac{1}{n}Var\left[\widehat{h}(\boldsymbol{X})\{\mu(\boldsymbol{U})-h(\boldsymbol{X})\}|\boldsymbol{X}_{n+1},\ldots,\boldsymbol{X}_{n+m}\right]\right)+0 \\
\leq&\frac{\{\sup\mu(\cdot)+\sup h(\cdot)\}^2}{n}E\left[E\left\{\widehat{h}(\boldsymbol{X})|\boldsymbol{X}_{n+1},\ldots,\boldsymbol{X}_{n+m}\right\}^2\right]
\end{aligned}$$

Now we only need to show the expectation is bounded.

$$E\left[E\left\{\widehat{h}(\boldsymbol{X})|\boldsymbol{X}_{n+1},\dots,\boldsymbol{X}_{n+m}\right\}^2\right] = E\left\{\frac{\sum_{i=n+1}^{n+m}\kappa\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{b}\right)h(\boldsymbol{X}_i)}{\sum_{i=n+1}^{n+m}\kappa\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{b}\right)}\right\}^2 = E(\boldsymbol{W}(\boldsymbol{X})^\top\boldsymbol{h})^2,$$

where h is a m-vector $(h(X_{n+1}),\cdots,h(X_{n+m}))^\top$, and $\boldsymbol{W}(\boldsymbol{X})$ is a corresponded weighted satisfying $\|\boldsymbol{W}(\boldsymbol{X})\|_1 = 1$. Hence,

$$E(\boldsymbol{W}(\boldsymbol{X})^\top\boldsymbol{h})^2 \leq E\|\boldsymbol{W}(\boldsymbol{X})\|_1^2\|\boldsymbol{h}\|_\infty^2 \leq \sup h(\cdot)^2 < \infty.$$

Hence,

$$\frac{1}{n}\sum_i^n \widehat{h}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\} = O_p(1/\sqrt{n}) = o_p(1/\sqrt{nb}).$$

6.5 $\widehat{\boldsymbol{\mu}}$ is better than $\widehat{\boldsymbol{\mu}}_K$

We use notation $\sigma^2_{Y|\boldsymbol{U}}$ to represent the conditional variance $E[(Y-E[Y|\boldsymbol{U}])^2]$. So, $\sigma^2_{Y|\boldsymbol{U}} = E[\epsilon^2] = \sigma^2$, $\sigma^2_{Y|\boldsymbol{X}} = E[(Y-h(\boldsymbol{X}))^2]$, and $\sigma^2_{\mu|\boldsymbol{X}} = E[(\mu(\boldsymbol{U}) - h(\boldsymbol{X}))^2]$. Note that $\sigma^2_{Y|\boldsymbol{X}} = \sigma^2_{\mu|\boldsymbol{X}} + \sigma^2_{Y|\boldsymbol{U}}$.

In this subsection, we want to compare $E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ with $E\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2$.

Theorem 6.1. *Under assumption (A1)-(A6)(A8), then*

$$E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq E\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2 + q(\sigma^2_{Y|\boldsymbol{X}} - 2\sigma^2_{Y|\boldsymbol{U}}),$$

when n goes to infinity.

Proof.

$$\begin{aligned} E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = E\|&(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{B}^{-1}\boldsymbol{\Delta Y} - \boldsymbol{\mu})\|^2 \\ &+ E\|\boldsymbol{P}(\boldsymbol{h} - \boldsymbol{\mu})\|^2, \end{aligned} \tag{24}$$

and

$$\begin{aligned} E\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2 = E\|&(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{B}^{-1}\boldsymbol{\Delta Y} - \boldsymbol{\mu})\|^2 \\ &+ E\|\boldsymbol{P}(\boldsymbol{B}^{-1}\boldsymbol{\Delta Y} - \boldsymbol{\mu}))\|^2,. \end{aligned} \tag{25}$$

Since the first term of above two equality are the same, we only have to compare (24) to (25).

$$\begin{align*}
(24) =& tr\left(E[\boldsymbol{P}(\boldsymbol{h}-\boldsymbol{\mu})(\boldsymbol{h}-\boldsymbol{\mu})^\top]\right) \\
=& tr\left(E[\boldsymbol{P}E[(\boldsymbol{h}-\boldsymbol{\mu})(\boldsymbol{h}-\boldsymbol{\mu})^\top|\boldsymbol{X}]]\right) \\
=& tr\left(E[\boldsymbol{P}E[(h_i-\mu_i)^2|\boldsymbol{X}]]\right) \\
=& E[tr(\boldsymbol{P})E[(h_i-\mu_i)^2|\boldsymbol{X}]] = qE[(h_i-\mu_i)^2] \tag{26} \\
=& q\sigma^2_{\mu|\boldsymbol{X}} = q(\sigma^2_{Y|\boldsymbol{X}} - \sigma^2_{Y|\boldsymbol{U}}), \tag{27}
\end{align*}$$

$$\begin{align*}
(25) =& E\|\boldsymbol{P}\boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{Y}-\boldsymbol{\mu})\|^2 \tag{28} \\
&+E\|\boldsymbol{P}(\boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{\mu}-\boldsymbol{\mu}))\|^2 \tag{29}
\end{align*}$$

$$\begin{align*}
(28) =& tr\left(E[\boldsymbol{P}\boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^\top]\right) \\
=& tr\left(E[\boldsymbol{P}\boldsymbol{B}^{-1}\boldsymbol{\Delta}E[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^\top|\boldsymbol{U}]]\right) \\
=& tr\left(E[\boldsymbol{P}\boldsymbol{B}^{-1}\boldsymbol{\Delta}E[(Y_i-\mu_i)^2|\boldsymbol{U}]]\right) \\
=& tr\left(E[\boldsymbol{X}^\top\boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{\Delta}\boldsymbol{B}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}]\right) * E[(Y_i-\mu_i)^2] \\
\to& pE[(Y_i-\mu_i)^2] = q\sigma^2_{Y|\boldsymbol{U}}, \tag{30}
\end{align*}$$

The last convergence comes from following Lemma.

Lemma 6.2.

$$tr\left((\boldsymbol{X}^\top\boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{\Delta}\boldsymbol{B}^{-1}\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)\xrightarrow{p}0$$

Furthere assume the covariate $\boldsymbol{X}$ is bounded variables, then

$$E[tr\left((\boldsymbol{X}^\top\boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{\Delta}\boldsymbol{B}^{-1}\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)]\to 0$$

Proof. Let $B_n := \frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$, which convergence to $\Sigma_{\boldsymbol{X}}$ in probability.
Let $A_n := \frac{\boldsymbol{X}^\top\boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{\Delta}\boldsymbol{B}^{-1}\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{X}}{n} - \frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$. Apply Lemma **??**, $\|A_n\|_\infty \xrightarrow{p} 0$, where $\|\cdot\|_\infty$ is the infinity norm in terms of all the elements of the matrix.

$$tr(A_n^\top B_n) = \sum_{i=1}^{p}\boldsymbol{a}_i^\top\boldsymbol{b}_i,$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the $i-$th columns of matrixs A_n and B_n respectively.

$$\begin{aligned}&\leq \sum_{i=1}^{p} \|\boldsymbol{a}_i\|_\infty \|\boldsymbol{b}_i\|_1 \\ &\leq \|\boldsymbol{A}_n\|_\infty \sum_{i=1}^{p} \|\boldsymbol{b}_i\|_1\end{aligned}$$

the last line convergence in probabiltiy to zero since the convergenceness of $\|\boldsymbol{A}_n\|_\infty$, and the boundness of the term relative to B_n. Furthermore, we can apply Bounded Convergence Theorem to get the L_1 convergenceness. □

So, the difference are

$$E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq E\|\widehat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}\|^2 + q(\sigma^2_{Y|\boldsymbol{X}} - 2\sigma^2_{Y|\boldsymbol{U}}) - E\|\boldsymbol{P}(\boldsymbol{B}^{-1}\boldsymbol{\Delta\mu} - \boldsymbol{\mu})\|^2.$$

The last term is positive, so we get our desired result. □

6.6 Property of Double Kernel Regression

$$\begin{aligned}&\int\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{w})\boldsymbol{u}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{u} d\boldsymbol{w} d\boldsymbol{u} \\ =&\int \kappa(\boldsymbol{w}) \int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\boldsymbol{u}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{u} d\boldsymbol{u} d\boldsymbol{w} \\ =&\int \kappa(\boldsymbol{w}) \int \kappa(\boldsymbol{w})(\boldsymbol{w} + \boldsymbol{w}\gamma)^\top \nabla^2 \mu(\boldsymbol{u})(\boldsymbol{w} + \boldsymbol{w}\gamma) d\boldsymbol{w} d\boldsymbol{w} \\ =&\int \kappa(\boldsymbol{w}) \int \kappa(\boldsymbol{w}) \left(\boldsymbol{w}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{w} + 2\gamma \boldsymbol{w}^\top \nabla^2 \boldsymbol{w} + \gamma^2 \boldsymbol{w}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{w}\right) d\boldsymbol{w} d\boldsymbol{w} \\ =&\int \kappa(\boldsymbol{w}) \left(A(\boldsymbol{u}) + 0 + \gamma^2 \boldsymbol{w}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{w}\right) d\boldsymbol{w} = (1 + \gamma^2)A(\boldsymbol{u}).\end{aligned}$$

Thus, if we use the standard kernel regression with kernel $\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{w})d\boldsymbol{w}$, then the asymptotic bias is $B_{DK}(\boldsymbol{u})$ and the asymptotic variance is $V_{DK}(\boldsymbol{u})$.

7 Some Lemma

Lemma 7.1 (Bernstein's inequalities)**.** *Let $X_1, \ldots, X_n$ be i.i.d random variables, such that $|X| \leq C$. Then,*

$$P(|\bar{X} - E[X]| > t) \leq 2\exp\left(-\frac{nt^2/2}{Var(X) + Ct/3}\right).$$

Lemma 7.2. *Assume that $E|Y|^s < \infty$ for some $s > 2$, $\sup f_U(\cdot)$, $\mu(\boldsymbol{u}) := E[Y|\boldsymbol{U} = \boldsymbol{u}]$ are bounded and Lipschitz , $\sigma^2(\boldsymbol{u}) := E[\{Y - \mu(\boldsymbol{U})\}^2|\boldsymbol{U} = \boldsymbol{u}]$ are bounded. The kernel κ is bounded, $\int \kappa(\boldsymbol{v})d\boldsymbol{v} = 1$, and $\int |\kappa(\boldsymbol{v})|d\boldsymbol{v} < \infty$, $\int |\kappa(\boldsymbol{v})|\|\boldsymbol{v}\|_2 d\boldsymbol{v} < \infty$. The bandwidth b is polynomial order of n with the properties $n^{1-2/s}b^{p+\theta} \to \infty$, for some constant $\theta > 0$. Define*

$$\widehat{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^p}\sum \kappa\left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{b}\right) Y_i,$$

$$a_n = \sqrt{\frac{\log n}{nb^p}} \to 0.$$

Then,

$$\max_{i=1,\ldots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}| = O_p(a_n), \tag{31}$$

and

$$E\max_{i=1,\ldots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}|^2 = o(1). \tag{32}$$

Define $\nu(\boldsymbol{u}) = \mu(\boldsymbol{u})f_U(\boldsymbol{u})$. Then,

$$\max_{i=1,\ldots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - \nu(\boldsymbol{U}_i)\}| = O_p(a_n) + O(b), \tag{33}$$

and

$$E\max_{i=1,\ldots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - \nu(\boldsymbol{U}_i)\}|^2 = o(1). \tag{34}$$

Proof. Define $\tau_n = a_n^{-1}$.

$$\widetilde{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^p}\sum \kappa\left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{b}\right) Y_i 1(|Y_i| \leq \tau_n),$$

$$R_n(\boldsymbol{u}) := \widehat{\nu}_n(\boldsymbol{u}) - \widetilde{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^p}\sum \kappa\left(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{b}\right) Y_i 1(|Y_i| > \tau_n).$$

Then, (31) have upper bound

$$\begin{aligned}
&\max_{i=1,\dots,n}|\widehat{\nu}_n(\boldsymbol{U}_i)-E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}| \\
\leq &\max_{i=1,\dots,n}|R_n(\boldsymbol{U}_i)| && (35)\\
+ &\max_{i=1,\dots,n}E|R_n(\boldsymbol{U}_i)| && (36)\\
+ &\max_{i=1,\dots,n}|\widetilde{\nu}_n(\boldsymbol{U}_i)-E\{\widetilde{\nu}_n(\boldsymbol{U}_i)\}|. && (37)
\end{aligned}$$

For (35), from Chebyshev's inequality, boundness of $s-th$ moments of Y , and the rates of n, b, we have

$$\begin{aligned}
P(\max_{i=1,\dots,n}|R_n(\boldsymbol{U}_i)|>0)=&P(\max_{i=1,\dots,n}|Y_i|>\tau_n)\leq nP(|Y_i|>\tau_n)\\
\leq& n\frac{E|Y|^s}{\tau_n^s}=(\frac{\log n}{n^{1-2/s}b^p})^{s/2}\to 0.
\end{aligned}$$

Hence, (35) is zero with probability goes to one. For (36), from the condition on κ, and the boundness of f_U, we have

$$\begin{aligned}
E|R_n(\boldsymbol{u}^*)|=&\frac{1}{b^p}\int\left|\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{u}}{b}\right)\right||y|1(|y|>\tau_n)f(y|\boldsymbol{u})f_U(\boldsymbol{u})dyd\boldsymbol{u}\\
=&\int|\kappa\left(\boldsymbol{v}\right)||y|1(|y|>\tau_n)f(y|\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})dyd\boldsymbol{v}\\
\leq&\sup f_U(\cdot)\int|\kappa(\boldsymbol{v})|d\boldsymbol{v}\frac{\sup_{\boldsymbol{u}}E[y^2|\boldsymbol{U}=\boldsymbol{u}]}{\tau_n}\\
\leq& a_n\left[\sup f_U(\cdot)\left\{\int|\kappa(\boldsymbol{v})|d\boldsymbol{v}\right\}\sup_{\boldsymbol{u}}\left\{\sigma^2(\boldsymbol{u})+\mu^2(\boldsymbol{u})\right\}\right]=O(a_n).
\end{aligned}$$

Note that the upperbound is uniform for all $\boldsymbol{u}^*$. Hence, (36) is of order a_n. For (37), we would like to apply Lemma 7.1. Hence, we have to calculate following quantity. From the condition on κ,

$$\sup_{\boldsymbol{U}}\frac{1}{b^p}\kappa\left(\frac{\boldsymbol{u}-\boldsymbol{U}}{b}\right)YI(|Y|\leq\tau_n)\leq\sup\kappa(\cdot)\frac{\tau_n}{b^p}.$$

$$
\begin{aligned}
&Var\left\{\sup_{\boldsymbol{U}}\frac{1}{b^p}\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{U}}{b}\right)YI(|Y|\leq\tau_n)\right\}\leq E\left\{\sup_{\boldsymbol{U}}\frac{1}{b^p}\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{U}}{b}\right)YI(|Y|\leq\tau_n)\right\}^2\\
=&\frac{1}{b^{2p}}\int\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{u}}{b}\right)^2|y^2|I(|Y|\leq\tau_n)f(y|\boldsymbol{u})f_U(\boldsymbol{u})dyd\boldsymbol{u}\\
=&\frac{1}{b^p}\int\kappa\left(\boldsymbol{v}\right)^2|y^2|I(|Y|\leq\tau_n)f(y|\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})dyd\boldsymbol{v}\\
\leq&\frac{1}{b^p}\sup f_U(\cdot)\sup\kappa(\cdot)\{\int|\kappa(\boldsymbol{v})|d\boldsymbol{v}\}\sup\{\sigma^2(\cdot)+\mu^2(\cdot)\}.
\end{aligned}
$$

Apply Lemma 7.1 with $t=ca_n$, and there is constant C^* such that

$$
\begin{aligned}
P\left(\max_{i=1,\ldots,n}|\widetilde{\nu}_n(\boldsymbol{U}_i)-E\{\widetilde{\nu}_n(\boldsymbol{U}_i)\}|>ca_n\right)\leq&2n\exp\left\{-C^*\min(c^2nb^pa_n^2,cnb^p\frac{a_n}{\tau_n})\right\}\\
\leq&2n\exp\left\{-C^*\log n\min(c^2,c)\right\}\\
\leq&2\exp\left[\log n\{1-C^*\min(c^2,c)\}\right]\to 0\\
&\text{for } c \text{ is large enough.} \qquad (38)
\end{aligned}
$$

Hence, (37) is of order a_n. (31) is proved. Next, we have to claim (35) - (37) convergence to zero in L_2. First, with the same argument in (38), choose $t=c$, we can see (37) has exponetial tail probability. Hence, it can convergence in L_2. And (36) is non-random with order $o(1)$; hence, it is also convergence in L_2. For (35)

$$
\max_{i=1,\ldots,n}|R_n(\boldsymbol{U}_i)|\leq\frac{\sup\kappa(\cdot)}{nb^p}\sum|Y_i|^2 1(|Y_i|>\tau_n):=\sup\kappa(\cdot)\times M_n
$$

$$
\begin{aligned}
E[M_n^2]=&\frac{1}{nb^{2p}}E\{|Y_i|^2 1(|Y_i|>\tau_n)\}\\
&\text{Since } E|Y|^s<\infty \text{ for all } d>0\\
\leq&\frac{1}{nb^{2p}\tau^{2d}}E|Y_i|^s\\
=&E|Y_i|^s(\log n)^{s/2-1}n^{-s/2}b^{-p(s/2+1)}\\
=&E|Y_i|^s(\log n)^{s/2-1}\{n^{1-2/(s+2)}b^p\}^{-s/2-1}\\
\leq&E|Y_i|^s(\frac{\log n}{n^{1-2/s}b^p})^{1+s/2},
\end{aligned}
$$

which convergence to zero with the condition on n, b. (32) is proved. For (33) and (34), it is sufficient to calculate the bias.

$$|E[\widehat{\nu}_n(\boldsymbol{u}^*)] - \nu(\boldsymbol{u}^*)| =| \int \kappa(\boldsymbol{v})\{\mu(\boldsymbol{u}^* - b\boldsymbol{v})f_U(\boldsymbol{u}^* - b\boldsymbol{v}) - \mu(bu^*)f_U(\boldsymbol{u}^*)\}dbv|$$

From (B5), there is a Lipschitz constant $L > 0$, such that

$$\leq bL \int |\kappa(\boldsymbol{v})| \|\boldsymbol{v}\|_2 dbv = O(b).$$

Note that the above bound is uniform for all $\boldsymbol{u}$. We complete the Lemma. □

Lemma 7.3. *Under assumptions (A1)-(A4),*

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\widehat{f_U}(\boldsymbol{u}) - f_U(\boldsymbol{u})| = o_p(1) \tag{39}$$

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\frac{1}{n}\sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)g(\boldsymbol{X}_i) - f_U(\boldsymbol{u})g(\boldsymbol{x})| = o_p(1) \tag{40}$$

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\frac{1}{n}\sum_{i=1}^{n} \frac{\kappa_l(\boldsymbol{u} - \boldsymbol{U}_i)g(\boldsymbol{X}_i)}{f_U(\boldsymbol{U}_i)} - g(\boldsymbol{x})| = o_p(1) \tag{41}$$

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} \left| \frac{b^p}{n}\sum_{i=1}^{n} \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)} - E\left[b^p \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)} \right] \right| = o_(1) \tag{42}$$

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\frac{1}{n}\sum_{i=1}^{n} \frac{1}{b^2}\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)(\boldsymbol{u} - \boldsymbol{U}_i)^\top \nabla^2\mu(\boldsymbol{u})(\boldsymbol{u} - \boldsymbol{U}_i) - f_U(\boldsymbol{u}) \int \kappa(\boldsymbol{v})\boldsymbol{v}^\top \nabla^2\mu(\boldsymbol{u})\boldsymbol{v}d\boldsymbol{v}| = o_p(1) \tag{43}$$

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\frac{1}{n}\sum_{i=1}^{n} \frac{1}{b}\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u} - \boldsymbol{U}_i) - b\int \kappa(\boldsymbol{v})\nabla^\top \boldsymbol{v}\boldsymbol{v}^\top \nabla f_U(\boldsymbol{u})d\boldsymbol{v}| = o_p(b) \tag{44}$$

$$\sup_{\boldsymbol{u}\in\boldsymbol{U}_n} |\frac{1}{n}\sum_{i=1}^{n} \kappa_l(u - \boldsymbol{U}_i)\|\boldsymbol{u} - \boldsymbol{U}_i\|^3| = o_p(l^2) \tag{45}$$

Proof. To claim (39)-(41), we would appy Lemma 7.2 with choosing proper Y. Note that assumetion (A1)-(A4) can make sure the condition on Lemma 7.2 hold in the following argument.

For (39), choose Y be 1, the result follows (33) in Lemma 7.2. For (40), choose Y be $g(\boldsymbol{X})$, the result follows (33) in Lemma 7.2. Since f_U is bounded away from zero, division is Lipschitz's continus. (39) and (40) can imply (41). For (42)-(45), we cannot apply Lemma 7.2. However, the similar argument can show the result. Key is to apply Lemma 7.1. Hence, we have to calculate the upper bound and variance of the random variables.

For (42), since boundness of κ, $\boldsymbol{U}$ (A1) (A3), and κ has finite first moment (A3), we have the following upper bounds which is uniform for every $\boldsymbol{u}$ and bU_i. For some constant C_1 and C_2,

$$|b^p \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)}| \leq \frac{C_1}{l^p}$$

$$\begin{aligned} &Var\{b^p \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)}\} \leq E\{b^p \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)}\}^2 \\ =&\frac{1}{l^p}\int \kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{u}}{b} - \frac{l}{b}\boldsymbol{v}\right)^2 \kappa(\boldsymbol{v})^2 d\boldsymbol{v} \leq \frac{C_2}{l^p}\end{aligned}$$

Hence, apply Lemma 7.1

$$\begin{aligned} &P(\sup_{\boldsymbol{u}\in\mathcal{U}_n}\left|\frac{b^p}{n}\sum_{i=1}^n \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)} - E\left[b^p \frac{\kappa_b(\boldsymbol{u}^* - \boldsymbol{U}_i)\kappa_l(\boldsymbol{U}_i - \boldsymbol{u})}{f_U(\boldsymbol{U}_i)}\right]\right| > t) \\ \leq&2n\exp\left(-\frac{nt^2/2}{C_2/l^p + tC_1/(3l^p)}\right),\end{aligned}$$

which goes to zero under assumption (A4). (42) is proved.

For (43), since boundness of κ, $\boldsymbol{U}$ (A1) (A3), and κ has finite forth moment (A3), we have the following upper bounds which is uniform for every $\boldsymbol{u}$ and bU_i. For some constant C_1 and C_2,

$$|\frac{1}{b^2}\kappa_b(\boldsymbol{u} - \boldsymbol{U}_i)(\boldsymbol{u} - \boldsymbol{U}_i)^\top \nabla^2\mu(\boldsymbol{u})(\boldsymbol{u} - \boldsymbol{U}_i)| \leq \frac{C_1}{b^{p+2}}$$

$$
\begin{aligned}
&Var\{\frac{1}{b^2}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)(\boldsymbol{u}-\boldsymbol{U}_i)^\top\nabla^2\mu(\boldsymbol{u})(\boldsymbol{u}-\boldsymbol{U}_i)\} \le E\{\frac{1}{b^2}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)(\boldsymbol{u}-\boldsymbol{U}_i)^\top\nabla^2\mu(\boldsymbol{u})(\boldsymbol{u}-\boldsymbol{U}_i)\}^2\\
=&\frac{1}{b^p}E\left[\frac{1}{b^p}\kappa\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)^2\left\{\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)^\top\nabla^2\mu(\boldsymbol{u})\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)\right\}^2\right]\\
=&\frac{1}{b^p}\int\kappa(\boldsymbol{v})^2\{\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v}\}^2 f_U(\boldsymbol{u}-b\boldsymbol{v})d\boldsymbol{v} \le \frac{C_2}{b^p}
\end{aligned}
$$

Hence, apply Lemma 7.1

$$
\begin{aligned}
&P(\sup_{\boldsymbol{u}\in\mathcal{U}_n}\left|\frac{1}{n}\sum_{i=1}^n\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)^\top\nabla^2\mu(\boldsymbol{u})\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)\right.\\
&\left.-E\left\{\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)^\top\nabla^2\mu(\boldsymbol{u})\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)\right\}\right| > t)\\
\le&2n\exp\left(-\frac{nt^2/2}{C_2/l^p+tC_1/(3l^{p+2})}\right),
\end{aligned}
$$

which goes to zero under assumption (A4). Now we have to calculate bias term. From Lipschitz's continunity of f_U, boundness of third moment of κ, and boundness of operator norm of second gradient of μ,

$$
\begin{aligned}
&\sup_{\boldsymbol{u}\in\mathcal{U}_n}\left|f_U(\boldsymbol{u})\int\kappa(\boldsymbol{v})\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v}d\boldsymbol{v}-E\left\{\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)^\top\nabla^2\mu(\boldsymbol{u})\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right)\right\}\right|\\
=&\sup_{\boldsymbol{u}\in\mathcal{U}_n}\left|f_U(\boldsymbol{u})\int\kappa(\boldsymbol{v})\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v}d\boldsymbol{v}-\int\kappa(\boldsymbol{v})\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v}f_U(\boldsymbol{u}-b\boldsymbol{v})d\boldsymbol{v}\right|\\
\le&b\sup\|\nabla^2\mu(\cdot)\|\sup_{\boldsymbol{u}\in\mathcal{U}_n}\int\kappa(\boldsymbol{v})\|\boldsymbol{v}\|_2^3d\boldsymbol{v}=o(1).
\end{aligned}
$$

(43) is proved. For (44), since boundness of κ, $\boldsymbol{U}$ (A1) (A3), and κ has finite fourth moment (A3), we have the following upper bounds which is uniform for every $\boldsymbol{u}$ and bU_i. For some constant C_1 and C_2,

$$
|\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)| \le \frac{C_1}{b^{1+p}}.
$$

$$
\begin{aligned}
&Var\{\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)\} \le E\{\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)\}^2\\
=&\frac{1}{b^p}\int\kappa(\boldsymbol{v})^2(\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v})^2 f(\boldsymbol{u}-\boldsymbol{v}b)d\boldsymbol{v}=\frac{C_2}{b^p}
\end{aligned}
$$

Hence, apply Lemma 7.1

$$
\begin{aligned}
&P(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)-E\{\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)\}\right|>t)\\
\leq&2\exp\left(-\frac{nt^2/2}{C_2/b^p+tC_1/(3b^{1+p})}\right).
\end{aligned}
$$

Let $t=b\epsilon$, for some $\epsilon>0$. Since $|\mathcal{U}_n|=n$,

$$
\begin{aligned}
&P(\sup_{\boldsymbol{u}\in\mathcal{U}_n}|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)-E[\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)]|>\epsilon b)\\
\leq&2n\exp\left\{-C_3\max(\epsilon,\epsilon^2)nb^{p+2}\right\},
\end{aligned}
$$

for some constant C_3. From (A4), the above probability goes to zero. Next we have to calculate the bias term. From Taylor expansion, boundness of the second derivative of density f, and the finite third moment of κ,

$$
\begin{aligned}
&\left|E[\frac{1}{b}\kappa_b(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{u}-\boldsymbol{U}_i)]-b\int\kappa(\boldsymbol{v})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}\boldsymbol{v}^\top\nabla f_U(\boldsymbol{u})d\boldsymbol{v}\right|\\
=&\left|\int\kappa(\boldsymbol{v})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}f_U(\boldsymbol{u}-b\boldsymbol{v})d\boldsymbol{v}-b\int\kappa(\boldsymbol{v})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}\boldsymbol{v}^\top\nabla f_U(\boldsymbol{u})d\boldsymbol{v}\right|
\end{aligned}
$$

$\boldsymbol{\xi}$ is given from Mean Value of Taylor expansion,

$$
=\left|f_U(\boldsymbol{u})\int\kappa(\boldsymbol{v})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}d\boldsymbol{v}+b^2\int\kappa(\boldsymbol{v})\nabla\mu(\boldsymbol{u})^\top\boldsymbol{v}\boldsymbol{v}^\top\nabla^2f_U(\boldsymbol{\xi})\boldsymbol{v}d\boldsymbol{v}\right|
$$

From κ is mean zero,

$$
\leq b^2\sup\|\nabla^2f_U(\cdot)\|\|\nabla\mu(\boldsymbol{u})\|_2\int\kappa(\boldsymbol{v})\|\boldsymbol{v}\|_2^3d\boldsymbol{v}.
$$

Hence the bias is of order $o(b)$. (44) is proved.

For (45), since boundness of κ, $\boldsymbol{U}$ (A1) (A3), and κ has finite six moment (A3), we have the following upper bounds which is uniform for every $\boldsymbol{u}$ and bU_i. For some constant C_1 and C_2,

$$
|\frac{1}{l^3}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|_2^3|\leq\frac{C_1}{l^{p+3}}
$$

$$
\begin{aligned}
&Var\{\frac{1}{l^3}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|_2^3\}\leq E\{\frac{1}{l^3}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|^3\}^2\\
=&\frac{1}{l^p}\int\kappa(\boldsymbol{v})^2\|\boldsymbol{v}\|_2^6f_U(\boldsymbol{u}-b\boldsymbol{v})d\boldsymbol{v}\leq\frac{C_2}{l^p}.
\end{aligned}
$$

Hence, apply Lemma 7.1

$$
\begin{aligned}
&P(\sup_{\boldsymbol{u}\in\boldsymbol{U}_n}|\frac{1}{nl^3}\sum_{i=1}^{n}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|^3-E\{\frac{1}{l^3}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|_2^3\}|>t)\\
\leq&2n\exp\left(-\frac{nt^2/2}{C_2/l^p+tC_1/(3l^{p+3})}\right),
\end{aligned}
$$

which goes to zero under assumption (A4). Now we have to calculate bias term.

$$
\begin{aligned}
\sup_{\boldsymbol{u}\in\boldsymbol{U}_n}\left|E\{\frac{1}{l^3}\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|_2^3\}\right|=&\sup_{\boldsymbol{u}\in\boldsymbol{U}_n}\left|\int\kappa(\boldsymbol{v})\|\boldsymbol{v}\|_2^3f_U(\boldsymbol{u}-b\boldsymbol{v})d\boldsymbol{v}\right|\\
\leq&\sup f_U(\cdot)\int\kappa(\boldsymbol{v})\|\boldsymbol{v}\|_2^3d\boldsymbol{v}<\infty.
\end{aligned}
$$

Hence $\sup_{\boldsymbol{u}\in\boldsymbol{U}_n}|\frac{1}{n}\sum_{i=1}^n\kappa_l(u-\boldsymbol{U}_i)\|\boldsymbol{u}-\boldsymbol{U}_i\|^3|=O_p(l^3)=o(l^2)$. (45) is proved. □

Lemma 7.4. *Under assumptions in (A1)-(A4),*

$$
(W_{1j}+W_{2j})^2=\left[\frac{1}{b^p}\frac{1}{f(\boldsymbol{u}^*)}\int\kappa(\boldsymbol{v})\kappa(\frac{\boldsymbol{u}^*-\boldsymbol{U}_j}{b}-\boldsymbol{v}(\frac{l}{b})^{1/p})d\boldsymbol{v}\right]^2(1+o_p(1)),
$$

where $o_p(1)$ is uniformly in $j=1,\ldots,n$.

Proof. Since f is bounded away from zero, and (39), (42) in Lemma 7.3,

$$
W_{1j}=\left[\frac{1}{b^p}\frac{1}{f(\boldsymbol{u}^*)}\int\kappa(\boldsymbol{v})\kappa(\frac{\boldsymbol{u}^*-\boldsymbol{U}_j}{b}-\boldsymbol{v}\frac{l}{b})d\boldsymbol{v}\right](1+o_p(1)),
$$

with $o_p(1)$ is uniformly in $j=1,\ldots,n$. Since f is bounded away from zero, and (39)-(41) in Lemma 7.3,

$$
W_{2j}=\left[g(\boldsymbol{x}^*)^\top\Sigma_g^{-1}g(\boldsymbol{X}_j)\right](1+o_p(1)),
$$

Hence, W_{1j} is dominated W_{2j}, which implys

$$
(W_{1j}+W_{2j})^2=\left[\frac{1}{b^p}\frac{1}{f(\boldsymbol{u}^*)}\int\kappa(\boldsymbol{v})\kappa(\frac{\boldsymbol{u}^*-\boldsymbol{U}_j}{b}-\boldsymbol{v}(\frac{l}{b})^{1/p})d\boldsymbol{v}\right]^2(1+o_p(1)).
$$

□

Lemma 7.5. *Under assumptions (A1)-(A4),*

$$\sup_{\boldsymbol{u}\in\mathcal{U}_n} |\frac{1}{n\widehat{f}_U(\boldsymbol{u})}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)(\mu(\boldsymbol{U}_i)-\mu(\boldsymbol{u}))-l^2A(\boldsymbol{u})| = o_p(l^2)$$

Proof. First, let's ingore the term $\widehat{f}_U(\boldsymbol{u})$. Apply taylor expansion,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)(\mu(\boldsymbol{U}_i)-\mu(\boldsymbol{u}))\\
=&\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)\nabla\mu(\boldsymbol{u})^\top(\boldsymbol{U}_i-\boldsymbol{u}) && (46)\\
+&\frac{1}{2n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)(\boldsymbol{U}_i-\boldsymbol{u})^\top\nabla^2\mu(\boldsymbol{u})(\boldsymbol{U}_i-\boldsymbol{u}) && (47)\\
+&\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)R(\boldsymbol{U}_i-\boldsymbol{u}), && (48)
\end{aligned}$$

where $R(\boldsymbol{U}_i-\boldsymbol{u})$ is the residual term, which is bounded by $M\|\boldsymbol{U}_i-\boldsymbol{u}\|^3$ for some constant M from the Assumption(A2) that the third moments of μ is bounded.

From (43), (44) in Lemma 7.3, (46) and (47), convergence to $l^{2/p}f_U(\boldsymbol{u})A(\boldsymbol{u})$ uniformly for all $\boldsymbol{u}\in\mathcal{U}_n$. And from (45) in Lemma 7.3, (48) is of order

$o_p(l^{2/p})$. Next consider $\widehat{f}$. We have to calculate the following difference.

$$\left|\left\{\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)(\mu(\boldsymbol{U}_i)-\mu(\boldsymbol{u}))\right\}\left\{\frac{1}{\widehat{f}_U(\boldsymbol{u})}-\frac{1}{f_U(\boldsymbol{u})}\right\}\right|$$

$$\leq\left|\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}-\boldsymbol{U}_i)(\mu(\boldsymbol{U}_i)-\mu(\boldsymbol{u}))\right|\left|\frac{1}{\widehat{f}_U(\boldsymbol{u})}-\frac{1}{f_U(\boldsymbol{u})}\right|$$

From (46)-(48)

$$=O_p(l^2)\left|\frac{1}{\widehat{f}_U(\boldsymbol{u})}-\frac{1}{f_U(\boldsymbol{u})}\right|$$

From Lemma 7.2 and f is bounded away from zero,

$$=O_p(l^2)O_p(\sqrt{\frac{\log n}{nl^p}}+l)$$

From Assumption(A4)

$$=o_p(l^2).$$

We got our desired result.

□

Lemma 7.6. *Assume that $\boldsymbol{U}$ is bounded R.V, μ, h are bounded function, and g_n satisfy following condition,*

$$\max_{i=1,\ldots,n}|g_n(\boldsymbol{X}_i)-g_n^{-i}(\boldsymbol{X}_i)|=o_p(n^{2/(p+4)}) \tag{49}$$

$$E[\{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}^2]=o(1) \tag{50}$$

$$E[\{g_n^{-i}(\boldsymbol{X}_i)-g_n^{-ij}(\boldsymbol{X}_i)\}^2]=o(n^{4/(p+4)}) \tag{51}$$

Then, $\frac{1}{n}\sum_i^n g_n(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}=o_p(n^{-2/(p+4)})$.

Proof. First, we defind $\mathcal{U}_n$ be the set $\{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n\}$, $\mathcal{U}_n^{-i}$ be the set $\{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n\}\setminus\{\boldsymbol{U}_i\}$, $\mathcal{U}_n^{-ij}$ be the set $\{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n\}\setminus\{\boldsymbol{U}_i,\boldsymbol{U}_j\}$,. Let g_n, g_n^{-i}, g_n^{-ij} be the estimator of g via data set $\mathcal{U}$, $\mathcal{U}^{-i}$,$\mathcal{U}^{-ij}$ respectively.

$$
\begin{aligned}
\frac{1}{n}\sum_i^n g_n(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\} =&\frac{1}{n}\sum_i^n g(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\} && (52)\\
&+\frac{1}{n}\sum_i^n \{g_n(\boldsymbol{X}_i)-g_n^{-i}(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\} && (53)\\
&+\frac{1}{n}\sum_i^n \{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\} && (54)
\end{aligned}
$$

From CLT (52) is of order $O_p(1/\sqrt{n})$. On the orther hand, from Hölder's inequality and (G1)

$$
(53) \leq M \max_{i=1,\dots,n} |g_n(\boldsymbol{X}_i)-g_n^{-i}(\boldsymbol{X}_i)| = o_p(n^{2/(p+4)}).
$$

The last term, we would like to calculate it's variance.

$$
\begin{aligned}
Var((54)) \leq& \frac{1}{n}Var(\{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\})\\
&+|Cov(\{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\{g_n^{-j}(\boldsymbol{X}_j)-g(\boldsymbol{X}_j)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\})|
\end{aligned}
$$

For the first term, from the boundness of μ, h, $\boldsymbol{U}$, there is a constant M such that

$$
\frac{1}{n}Var(\{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}) \leq \frac{M}{n}E[\{g_n^{-i}(\boldsymbol{X}_i)-g(\boldsymbol{X}_i)\}^2] = o(\frac{1}{n}). \qquad \text{(G2)}
$$

Befor calculate the second term, there is some observations.

$$
\begin{aligned}
0 =&E\left[g_n^{-i}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\right] && (55)\\
=&E\left[g(\boldsymbol{X}_i)g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right] && (56)\\
=&E\left[g(\boldsymbol{X}_j)g_n^{-i}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right] && (57)\\
=&E\left[g(\boldsymbol{X}_j)g(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right] && (58)\\
=&E\left[g_n^{-ij}(\boldsymbol{X}_i)g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right] && (59)\\
=&E\left[g_n^{-ij}(\boldsymbol{X}_j)g_n^{-i}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right] && (60)\\
=&E\left[g_n^{-ij}(\boldsymbol{X}_j)g_n^{-ij}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i)-h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j)-h(\boldsymbol{X}_j)\}\right]. && (61)
\end{aligned}
$$

All of them use the same technique. For (55)

$$(55) = E\left(E\left[g_n^{-i}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}|\mathcal{U}_n^{-i}\right]\right)$$

Since condition on $\boldsymbol{U}_n^{-i}$ g_n^{-i} is constant function, and $\mu(\boldsymbol{U}) - h(\boldsymbol{X})$ is uncorrelative to $\boldsymbol{X}$, $E\left[g_n^{-i}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}|\mathcal{U}_n^{-i}\right]$ is zero. (56), (57), (58) all comes from same argument as below.

$$\begin{aligned}(56) =& E\left(E\left[g(\boldsymbol{X}_i)g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}|\mathcal{U}_n^{-j}\right]\right)\\ =& E\left(g(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}E\left[g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}|\mathcal{U}_n^{-j}\right]\right).\end{aligned}$$

We can see the conditional expectation are also zero. (59), (60), (61) all comes from same argument as below.

$$\begin{aligned}(59) =& E\left(E\left[g_n^{-ij}(\boldsymbol{X}_i)g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}|\mathcal{U}_n^{-j}\right]\right)\\ & \text{Condition on } \mathcal{U}_n^{-j},\ g_n^{-ij}(\boldsymbol{X}_i) \text{ is a constant}\\ & \text{And, } g_n^{-j} \text{ is a constant function; hence,}\\ =& E\left(g_n^{-ij}(\boldsymbol{X}_i)\{\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)\}E\left[g_n^{-j}(\boldsymbol{X}_j)\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}|\mathcal{U}_n^{-j}\right]\right).\end{aligned}$$

The conditional expectation are also zero same as the argument in above. From the observations we can see,

$$\begin{aligned}& Cov[\{g_n^{-i}(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}\{g_n^{-j}(\boldsymbol{X}_j) - g(\boldsymbol{X}_j)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}]\\ & \quad \text{From (55)}\\ =& E[\{g_n^{-i}(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}\{g_n^{-j}(\boldsymbol{X}_j) - g(\boldsymbol{X}_j)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}]\\ & \quad \text{From (56)(57)(58)}\\ =& E[\{g_n^{-i}(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}\{g_n^{-j}(\boldsymbol{X}_j)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}]\\ & \quad \text{From (59)(60)(61)}\\ =& E[\{g_n^{-i}(\boldsymbol{X}_i) - g_n^{-ij}(\boldsymbol{X}_i)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}\{g_n^{-j}(\boldsymbol{X}_j) - g_n^{-ij}(\boldsymbol{X}_j)\}\{\mu(\boldsymbol{U}_j) - h(\boldsymbol{X}_j)\}]\\ & \quad \text{From the boundness of } \mu,\ h,\ \boldsymbol{U}, \text{ and Cauchy inequaliy, there is a constant } M \text{ such that}\\ \leq& M\sqrt{E\{g_n^{-i}(\boldsymbol{X}_i) - g_n^{-ij}(\boldsymbol{X}_i)\}^2}\sqrt{E\{g_n^{-j}(\boldsymbol{X}_j) - g_n^{-ij}(\boldsymbol{X}_j)\}^2}\\ =& ME\{g_n^{-i}(\boldsymbol{X}_i) - g_n^{-ij}(\boldsymbol{X}_i)\}^2.\end{aligned}$$

From above argument and (G2) (G3), (54) is of order $o_p(\sqrt{n^{2/(p+4)}})$. □

Lemma 7.7. *Assume that* $E|Y|^s < \infty$ *for some* $s > 2$, *and* $\sigma^2(\boldsymbol{u}) := E[\{Y - \mu(\boldsymbol{U})\}^2|\boldsymbol{U} = \boldsymbol{u}]$ *are bounded.* μ, $\nabla_k\mu$, $\nabla_{kk}^2\mu$, f_U, $\nabla_k f_U$, $\nabla_{kk}^2 f_U$, *are*

Lipschitz continuous. The covariate $\boldsymbol{U}$ *has compact support and* $\sup f_U(\cdot)$*. The kernel satisfy* $\sup |\nabla^2_{kk}\kappa(\cdot)|$ *is bounded,* $\int \kappa(\boldsymbol{v})d\boldsymbol{v} = 1$*,* $\int |\nabla^2_{kk}\kappa(\boldsymbol{v})|d\boldsymbol{v} < \infty$*,* $\int |\kappa(\boldsymbol{v})|\|\boldsymbol{v}\|_2 d\boldsymbol{v} < \infty$*,* $\sup_{\boldsymbol{u}\neq 0} \frac{1}{b^2}|\nabla_k\kappa(\boldsymbol{u}/b)| = O(b)$*, and* $\sup_{\boldsymbol{u}\neq 0} \frac{1}{b}|\kappa(\boldsymbol{u}/b)| = O(b)$*. The bandwidth b is polynomial order of n with the following properties,* $nb^{p+4+\theta} \to \infty$*, and* $n^{1-2/s}b^{p+\theta} \to \infty$*, for some constant* $\theta > 0$*.*

Define

$$\widehat{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^{p+2}} \sum \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right) Y_i,$$

$$a_n = \sqrt{\frac{\log n}{nb^{p+4}}} \to 0$$

Then,

$$\max_{i=1,\dots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}| = O_p(a_n). \tag{62}$$

If

$$4/(s+2) < \theta, \tag{63}$$

then we also have

$$E \max_{i=1,\dots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}|^2 = o(1). \tag{64}$$

Define

$$\nu(\boldsymbol{u}) = \nabla^2_{kk}\{\mu(\boldsymbol{u})f_U(\boldsymbol{u})\} = \nabla^2_{kk}\mu(\boldsymbol{u})f_U(\boldsymbol{u}) + 2\nabla_k\mu(\boldsymbol{u})\nabla_k f_U(\boldsymbol{u}) + \mu(\boldsymbol{u})\nabla^2_{kk}f_U(\boldsymbol{u})$$

Then,

$$\max_{i=1,\dots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - \nu(\boldsymbol{U}_i)| = O_p(a_n) + O(b). \tag{65}$$

If (63) *hold, then we also have*

$$E \max_{i=1,\dots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - \nu(\boldsymbol{U}_i)|^2 = o(1). \tag{66}$$

Proof. This lemma follows the same arguments in Lemma 7.2. Define $\tau_n = \sqrt{\frac{nb^p}{\log n}}$.

$$\widetilde{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^{p+2}} \sum \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right) Y_i 1(|Y_i| \le \tau_n),$$

$$R_n(\boldsymbol{u}) := \widehat{\nu}_n(\boldsymbol{u}) - \widetilde{\nu}_n(\boldsymbol{u}) = \frac{1}{nb^{p+2}} \sum \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}-\boldsymbol{U}_i}{b}\right) Y_i 1(|Y_i| > \tau_n).$$

Then, (62) have upper bound

$$
\begin{aligned}
&\max_{i=1,\dots,n} |\widehat{\nu}_n(\boldsymbol{U}_i) - E\{\widehat{\nu}_n(\boldsymbol{U}_i)\}| \\
\leq &\max_{i=1,\dots,n} |R_n(\boldsymbol{U}_i)| \qquad (67)\\
+ &\max_{i=1,\dots,n} E|R_n(\boldsymbol{U}_i)| \qquad (68)\\
+ &\max_{i=1,\dots,n} |\widetilde{\nu}_n(\boldsymbol{U}_i) - E\{\widetilde{\nu}_n(\boldsymbol{U}_i)\}|. \qquad (69)
\end{aligned}
$$

For (67), from Chebyshev's inequality, the boundness of $s - th$ moment of Y, the rates of n, b

$$
\begin{aligned}
P(\max_{i=1,\dots,n} |R_n(\boldsymbol{U}_i)| > 0) =&P(\max_{i=1,\dots,n} |Y_i| > \tau_n) = nP(|Y_i| > \tau_n)\\
\leq& n\frac{E|Y|^s}{\tau_n^s} = n(nb^p)^{-s/2}(\log n)^{s/2} \to 0.
\end{aligned}
$$

Hence, (67) is zero with probability goes to one. For (68) from the condition on κ, and the boundness of f_U,

$$
\begin{aligned}
E|R_n(\boldsymbol{u}^*)| =&\frac{1}{b^{p+2}} \int \left| \nabla^2_{kk}\kappa \left(\frac{\boldsymbol{u}^* - \boldsymbol{u}}{b} \right) \right| |y| 1(|y| > \tau_n) f(y|\boldsymbol{u}) f_U(\boldsymbol{u}) dy d\boldsymbol{u} \\
=&\frac{1}{b^2} \int \left| \nabla^2_{kk}\kappa\left(\boldsymbol{v}\right) \right| |y| 1(|y| > \tau_n) f(y|\boldsymbol{u}^* - b\boldsymbol{v}) f_U(\boldsymbol{u}^* - b\boldsymbol{v}) dy d\boldsymbol{v} \\
\leq& \sup f_U(\cdot) \int |\nabla^2_{kk}\kappa(\boldsymbol{v})| d\boldsymbol{v} \frac{\sup_{\boldsymbol{u}} E[y^2|\boldsymbol{U} = \boldsymbol{u}]}{\tau_n b^2} \\
\leq& a_n \left[\sup f_U(\cdot) \left\{ \int |\nabla^2_{kk}\kappa(\boldsymbol{v})| d\boldsymbol{v} \right\} \sup_{\boldsymbol{u}} \left\{ \sigma^2(\boldsymbol{u}) + \mu^2(\boldsymbol{u}) \right\} \right] = O(a_n).
\end{aligned}
$$

Note that the upperbound is uniform for all $\boldsymbol{u}^*$. Hence, (68) is of order a_n. For (69), we would like to apply Lemma 7.1. Hence, we have to calculate following quantity.

$$
\sup_{\boldsymbol{U}} \frac{1}{b^{p+2}} \nabla^2_{kk}\kappa \left(\frac{\boldsymbol{u} - \boldsymbol{U}}{b} \right) Y I(|Y| \leq \tau_n) \leq \sup |\nabla^2_{kk}\kappa(\cdot)| \frac{\tau_n}{b^{p+2}} = O(\frac{\tau}{b^{p+2}}),
$$

$$
\begin{aligned}
&Var\left\{\sup_{\boldsymbol{U}} \frac{1}{b^{p+2}} \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{U}}{b}\right) Y I(|Y| \leq \tau_n)\right\} \leq E\left\{\sup_{\boldsymbol{U}} \frac{1}{b^{p+2}} \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{U}}{b}\right) Y I(|Y| \leq \tau_n)\right\}^2 \\
=&\frac{1}{b^{2p+4}} \int \nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}^* - \boldsymbol{u}}{b}\right)^2 |y^2| I(|Y| \leq \tau_n) f(y|\boldsymbol{u}) f_U(\boldsymbol{u}) dy d\boldsymbol{u} \\
=&\frac{1}{b^{p+4}} \int \nabla^2_{kk}\kappa\left(\boldsymbol{v}\right)^2 |y^2| I(|Y| \leq \tau_n) f(y|\boldsymbol{u}^* - b\boldsymbol{v}) f_U(\boldsymbol{u}^* - b\boldsymbol{v}) dy d\boldsymbol{v} \\
\leq&\frac{1}{b^{p+4}} \sup f_U(\cdot) \sup |\nabla^2_{kk}\kappa(\cdot)| \{ \int |\nabla^2_{kk}\kappa(\boldsymbol{v})| d\boldsymbol{v}\} \sup\{\sigma^2(\cdot) + \mu^2(\cdot)\} = O(\frac{1}{b^{p+4}}).
\end{aligned}
$$

Apply Lemma 7.1 with $t = ca_n$, and there is constant C^* such that

$$
\begin{aligned}
P\left(\max_{i=1,\ldots,n} |\widetilde{\nu}_n(\boldsymbol{U}_i) - E\{\widetilde{\nu}_n(\boldsymbol{U}_i)\}| > ca_n\right) \leq& 2n \exp\left\{-C^* \min(c^2 n b^{p+4} a_n^2, cnb^{p+2}\frac{a_n}{\tau_n})\right\} \\
=&2n \exp\left\{-C^* \log n \min(c^2, c)\right\} \\
\leq&2 \exp\left[\log n \{1 - C^* \min(c^2, c)\}\right] \to 0 \\
&\text{for } c \text{ is large enough.} \qquad (70)
\end{aligned}
$$

Hence, (69) is of order a_n. (62) is proved. Next, we have to claim (67)-(69) convergence to zero in L_2. First, with the same argument in (70), choose $t = c$, we can see (69) has exponetial tail probability. Hence, it can convergence in L_2. And (68) is non-random with order $o(1)$; hence, it is also convergence in L_2. For (67)

$$
\max_{i=1,\ldots,n} |R_n(\boldsymbol{U}_i)| \leq \frac{\sup |\nabla^2_{kk}\kappa(\cdot)|}{nb^{p+2}} \sum |Y_i|^2 1(|Y_i| > \tau_n) := \sup |\nabla^2_{kk}\kappa(\cdot)| \times M_n
$$

$$
\begin{aligned}
E[M_n^2] =& \frac{1}{nb^{2p+2}} E\{|Y_i|^2 1(|Y_i| > \tau_n)\} \\
&\text{Since } E|Y|^s < \infty \\
\leq& \frac{1}{nb^{2p+2}\tau^{s-2}} E|Y_i|^s \\
=&E|Y_i|^s (\log n)^{s/2-1} n^{-(s/2)} b^{-p(s/2+1)-2} \\
=&E|Y_i|^s (\log n)^{s/2-1} \{n^{1-2/(s+2)} b^p b^{4/(s+2)}\}^{-1-s/2} \\
\leq&E|Y_i|^s (\log n)^{s/2-1} \{n^{1-2/s} b^p b^{4/(s+2)}\}^{-1-s/2},
\end{aligned}
$$

which convergnece to zero since (63) and the rates of n, b. (64) is proved. For (65) and (66), it is sufficient to calculate bias term.

$$
\begin{aligned}
E\left\{\frac{1}{b^{p+2}}\nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{U}}{b}\right)Y\right\} &= \int \frac{1}{b^{p+2}}\nabla^2_{kk}\kappa\left(\frac{\boldsymbol{u}^*-\boldsymbol{u}}{b}\right)\mu(\boldsymbol{u})f_U(\boldsymbol{u})d\boldsymbol{u} \\
&= \int \frac{1}{b^2}\nabla^2_{kk}\kappa\left(\boldsymbol{v}\right)\mu(\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})d\boldsymbol{v} \\
&\quad \text{From Integration by Parts, and the condition on } \kappa. \\
&= \int \frac{1}{b}\nabla_k\kappa\left(\boldsymbol{v}\right)\nabla_k\{\mu(\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})\}d\boldsymbol{v} + O(b) \\
&\quad \text{From Integration by Parts, and and the condition on } \kappa. \\
&= \int \kappa\left(\boldsymbol{v}\right)\nabla^2_{kk}\{\mu(\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})\}d\boldsymbol{v} + O(b).
\end{aligned}
$$

Note that from the condition on κ the above equality holds uniformly for all $\boldsymbol{u}$. Then,

$$
\begin{aligned}
|E\widehat{\nu}_n(\boldsymbol{u})-\nu(\boldsymbol{u})| \leq & \int |\kappa\left(\boldsymbol{v}\right)|\left|\nabla^2_{kk}\{\mu(\boldsymbol{u}^*-b\boldsymbol{v})f_U(\boldsymbol{u}^*-b\boldsymbol{v})\}-\nu(\boldsymbol{u}^*)\right|d\boldsymbol{v}+O(b) \\
& \quad \text{From the Lipschitz assumptions, there is a Lipschitz constant } L\text{, such that} \\
\leq & bL\int |\kappa\left(\boldsymbol{v}\right)|\|\boldsymbol{v}\|_2 d\boldsymbol{v}+O(b)=O(b).
\end{aligned}
$$

The proof is completed. □

Lemma 7.8. *Under assumptions Theorem 3.2, then*

$$
\max_{k=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_k)-\widehat{A}^{-i}(\boldsymbol{U}_k)| = O_p(n^{-6/(p+8)}), \tag{71}
$$

also we can get a precisely upper bound,

$$
\max_{j=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_j)-\widehat{A}^{-i}(\boldsymbol{U}_j)| \leq L_n\max\{\frac{1}{n-1}\sum_{j\neq i}|Y_j|+|Y_i|,2\}a_n \tag{72}
$$

where L_n is bounded in probability and in L_2, a_n is a function of n and bandwidths using in estimating, which is of order $O(n^{-6/(p+8)})$. Furthermore,

$$
E\max_{j=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_j)-A(\boldsymbol{U}_j)|^2 = o(1) \tag{73}
$$

Proof. First, observe that from assumptions (A1)(A2)(A3)(A4), we can apply Lemma 7.2 on $\widehat{\nu_0}$ and $\widehat{f_U}$. From assumptoins (A1)(A2)(A3)(C1)(C2), we can apply Lemma 7.7 on $\{\nabla^2_{kk}\widehat{\nu_0}; k = 1, \dots, p\}$ and $\{\nabla^2_{kk}\widehat{f_U}; k = 1, \dots, p\}$. We will use the results in Lemma 7.2 and Lemma 7.7 in the following argument. Defince $\widehat{\boldsymbol{\theta}}^{-i}$ be the leave-i-out estimator of $\widehat{\boldsymbol{\theta}}$. Then, $\widehat{A}^{-i} = \eta \circ \widehat{\boldsymbol{\theta}}^{-i}$. Note that η is a rational funtions, and the denominator are $1/f_U^2$. From $f_U(\boldsymbol{u})$ is bounded away from zero and $\widehat{f_U}(\boldsymbol{u})$ is consistent estimator, we can conclude that $\inf_i \widehat{f_U}(\boldsymbol{U}_i) > \frac{1}{2} f_U$ with probability one; hence, η is a Lipschitz continuous. And we have,

$$|\widehat{A}(\boldsymbol{U}_j) - \widehat{A}^{-i}(\boldsymbol{U}_j)| = |\eta \circ \widehat{\theta}(\boldsymbol{U}_j) - \eta \circ \widehat{\theta}^{-i}(\boldsymbol{U}_j)| \leq L_n \|\widehat{\theta}(\boldsymbol{U}_j) - \widehat{\theta}^{-i}(\boldsymbol{U}_j)\|_\infty,$$

where L_n is

$$C\frac{\sum_{k=0}^p \|\nu_k\|_\infty + \max\limits_{i=1,\dots,n} |\widehat{\nu}_k(\boldsymbol{U}_i) - \nu_k(\boldsymbol{U}_i)| + \sum_{k=1}^p \|\nabla^2_{kk} f_U\| + \max\limits_{i=1,\dots,n} |\nabla^2_{kk}\widehat{f}_U(\boldsymbol{U}_i) - \nabla^2_{kk} f_U(\boldsymbol{U}_i)|}{\inf f_U - \max\limits_{i=1,\dots,n} |\widehat{f}_U(\boldsymbol{U}_i) - f_U(\boldsymbol{U}_i)|},$$

for some constant $C > 0$. And from (33) in Lemma 7.2, and (65) in Lemma 7.7, it is obviously that L_n is bounded in probability and in L_2. Let the function J can be κ, or $\nabla^2_{kk}\widetilde{\kappa}/\lambda_2^2$. D_j can be 1 or Y_j. λ is λ_1 if J is κ and λ is λ_2 if J is $\nabla^2_{kk}\widetilde{\kappa}/\lambda_2^2$.

$$\begin{aligned}
&\max_{j=1,\dots,n} |\frac{1}{n}\sum_{j=1}^n J(\frac{\boldsymbol{u} - \boldsymbol{U}_j}{\lambda})D_j - \frac{1}{n-1}\sum_{j\neq i} J(\frac{\boldsymbol{u} - \boldsymbol{U}_j}{\lambda})D_j| \\
=&\max_{j=1,\dots,n} |-\frac{1}{n(n-1)}\sum_{j\neq i} J(\frac{\boldsymbol{u} - \boldsymbol{U}_j}{\lambda})D_j + \frac{1}{n}J(\frac{\boldsymbol{u} - \boldsymbol{U}_i}{\lambda})D_i| \\
=&\frac{\sup_{\boldsymbol{u}} J(\boldsymbol{u})}{n}\left(\frac{1}{n-1}\sum_{j\neq i}|D_j| + |D_i|\right) \qquad (74)
\end{aligned}$$

From $\left(\frac{1}{n-1}\sum_{j\neq i}|D_j|+|D_i|\right)$ is $O_p(1)$, and condition (C2).

$$\max_{j=1,\ldots,n}|\widehat{f}_U(\boldsymbol{U}_j)-\widehat{f}_U^{-i}(\boldsymbol{U}_j)|=O_p(\frac{1}{n\lambda_1^p})=O_p(n^{-4/(p+4)})$$
$$\max_{j=1,\ldots,n}|\widehat{\nu}_0(\boldsymbol{U}_j)-\widehat{\nu}_0^{-i}(\boldsymbol{U}_j)|=O_p(\frac{1}{n\lambda_1^p})=O_p(n^{-4/(p+4)})$$
$$\max_{j=1,\ldots,n}|\nabla^2_{kk}\widehat{f}_U(\boldsymbol{U}_j)-\nabla_{kk}\widehat{f}_U^{-i}(\boldsymbol{U}_j)|=O_p(\frac{1}{nb^{p+2}})=O_p(n^{-6/(\lambda_2+8)})$$
$$\max_{j=1,\ldots,n}|\widehat{\nu}_k(\boldsymbol{U}_j)-\widehat{\nu}_k^{-i}(\boldsymbol{U}_j)|=O_p(\frac{1}{n\lambda_2^{p+2}})=O_p(n^{-6/(p+8)})$$

So,

$$\max_{j=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_j)-\widehat{A}^{-i}(\boldsymbol{U}_j)|\leq L_n\max_{j=1,\ldots,n}\|\widehat{\theta}(\boldsymbol{U}_j)-\widehat{\theta}^{-i}(\boldsymbol{U}_j)\|_\infty=O_p(n^{-6/(p+8)}),\tag{75}$$

we get our first result. From, (74) we can also get a precise upper bound. There is a constant M bounded κ, $\nabla^2_{kk}\widetilde{\kappa}$ for all k, such that

$$\max_{j=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_j)-\widehat{A}^{-i}(\boldsymbol{U}_j)|\leq ML_n\max\{\frac{1}{n-1}\sum_{j\neq i}|Y_j|+|Y_i|,2\}a_n,$$

where a_n is a function of n and the bandwidths using in estimating, here we use a_n to simplfy the result. From (75), we know that a_n is of order $O(n^{-6/(p+8)})$. Next, note that

$$\max_{j=1,\ldots,n}|\widehat{A}(\boldsymbol{U}_j)-A(\boldsymbol{U}_j)|=|\eta\circ\widehat{\theta}(\boldsymbol{U}_j)-\eta\circ\theta(\boldsymbol{U}_j)|\leq L_n\max_{j=1,\ldots,n}\|\widehat{\theta}(\boldsymbol{U}_k)-\theta(\boldsymbol{U}_k)\|_\infty.$$

Then (73) follows (34) in Lemma 7.2, and (66) in Lemma 7.7. □

8 Fixed $\boldsymbol{u}$

Now we consider fixed $\boldsymbol{u}$ rather than the average of bias. Simple algebra shows

$$\begin{aligned}B_{DK}(\boldsymbol{u})^2-B_{CK}(\boldsymbol{u})^2=&-c\gamma^4\left[g(\boldsymbol{x})^\top\boldsymbol{\Sigma}_g^{-1}E\{g(\boldsymbol{X})A(\boldsymbol{U})\}\right]^2\\&+2c(1+\gamma^2)\gamma^2A(\boldsymbol{u})\boldsymbol{x}^\top\boldsymbol{\Sigma}_g-1E\{g(\boldsymbol{X})A(\boldsymbol{U})\}.\end{aligned}$$

The sufficient condition that $B_{DK}(\boldsymbol{u})^2 \geq B_{CK}(\boldsymbol{u})^2$ can be

$$2c(1+\gamma^2)\gamma^2 A(\boldsymbol{u})g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\} \geq c\gamma^4 \left[g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}\right]^2.$$

Simple algebra get

$$\frac{A(\boldsymbol{u})}{g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}} \geq \frac{\gamma^2}{2(1+\gamma^2)}.$$

Now we can make a conclusion that if the sign of $A(\boldsymbol{u})$ is the same as its linear estimator $g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})$, then CK is better than DK for small γ. Also, if the ratio is greater than 1/2, CK is better than DK for every γ.

9 Multiply External Data Sets

In this section, we consider there are d external data sets, which provide the estimators $h_1, \ldots, h_d$, and the corresponded constraints $g_1, \ldots, g_d$. The modified optimaization equation (3) would be

$$\begin{aligned} \widehat{\boldsymbol{\mu}} = \arg \min_{\mu_1,\ldots,\mu_n} & \sum_{i=1}^n \sum_{j=1}^n \frac{\kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_i^n \kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_j)} (Y_j - \mu_i)^2 \\ \text{subject to} \quad & \sum_{j=1}^n \{\mu_j - \widehat{h}_k(\boldsymbol{X}_j)\} g_k(\boldsymbol{X}_j) = 0, \quad \forall k = 1, \ldots, d. \end{aligned} \tag{76}$$

(A6') The constraints g_k and the estimator h_k satisfy

$$E[g_k(\boldsymbol{X})(h_k(\boldsymbol{X}) - \mu(\boldsymbol{U}))] = 0, \quad \forall k = 1, \cdot, d.$$

Furthermore, g_k are a Lipshiszt Continuous. Define $g^*(\boldsymbol{X}) := (g_1(\boldsymbol{X}), \ldots, g_k(\boldsymbol{X}))^\top$, which covariance matirx is a positive definite. Denoted as $\boldsymbol{\Sigma}_{g^*(\boldsymbol{X})}$.

Theorem 9.1. *Under Assumption (A1)-(A5) (A6').*

$$\sqrt{nl}(\widehat{\boldsymbol{\mu}}_{CK}(\boldsymbol{u}^*) - \mu(\boldsymbol{u}^*)) \to N\left(Bias(\boldsymbol{u}^*), Var(\boldsymbol{u}^*)\right),$$

$$Bias(\boldsymbol{u}^*) = \frac{\sqrt{c}}{2}A(\boldsymbol{u}^*) + \frac{\sqrt{c}\gamma^2}{2}(A(\boldsymbol{u}^*) - g^*(\boldsymbol{x}^*)^\top \boldsymbol{\Sigma}_{g^*(\boldsymbol{X})}^{-1} E[g^*(\boldsymbol{X})A(\boldsymbol{U})]),$$

$$Var(\boldsymbol{u}^*) = \frac{\sigma^2}{f(\boldsymbol{u}^*)} \int (\int \kappa(\boldsymbol{w}\gamma^2 - \boldsymbol{u})\kappa(\boldsymbol{w})d\boldsymbol{w})^2 d\boldsymbol{u}$$

where $A(\boldsymbol{u}) = \int \kappa(\boldsymbol{u})\boldsymbol{u}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{u}d\boldsymbol{u}$, $\boldsymbol{\Sigma}_{g^*(\boldsymbol{X})}$ *is the covariance matrix with respend to* $g^*(\boldsymbol{X})$.

10 External and Internal Data are not the Same Distribution

In this section, we use f^E and f^I to represent the density of External and Internal Data. E_{F^E} and E_{F^I} denote the expectation under the distribution f^E and f^I respectively. We assume we can observe f^E and $h = E_{F^E}[Y|X]$; however, f^E is not equal to f^I. This would imply

$$0 = E_{f^E}[g(\boldsymbol{X})(\mu(\boldsymbol{U}) - h(\boldsymbol{X}))] \neq E_{f^I}[g(\boldsymbol{X})(\mu(\boldsymbol{U}) - h(\boldsymbol{X}))].$$

Hence, we have to modify the constraint in (3),

$$\begin{aligned} \widehat{\boldsymbol{\mu}} = \arg \min_{\mu_1,\ldots,\mu_n} & \sum_{i=1}^n \sum_{j=1}^n \frac{\kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_i^n \kappa_b(\boldsymbol{U}_i - \boldsymbol{U}_j)} (Y_j - \mu_i)^2 \\ \text{subject to} \quad & \sum_{j=1}^n \frac{f^E(\boldsymbol{X}_j)}{\widehat{f}^I(\boldsymbol{X}_j)} \{\mu_j - \widehat{h}(\boldsymbol{X}_j)\} g(\boldsymbol{X}_j) = 0. \end{aligned} \tag{77}$$

where $\widehat{f}^I$ is the density estimator satisfy the following assumption.

(A8) The density estimator $\widehat{f}^I$ should satisfy the following uniformly convergence rate.

$$\|\widehat{f}^I - f^I\|_\infty = o_p(n^{-2/(p+4)}).$$

Remark 10.1. *[7] provide the uniformly convergence rate $O_p((n/\log n)^{-1/(q+2)})$ with kernel density estimator. Hence, (A8) is not satisfied (A8) unless p is larger than $2q$. Hence, we recommend estimated f_I via parametric methods which generally can achieve $O_p(n^{-1/2})$ making (A8) satisfied.*

Example 10.1 (Rate of estimated density under exponetial family). *Consider the exponetial family*

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = h(x) \exp\{\boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta})\}.$$

Let $\widehat{\boldsymbol{\theta}}$ being MLE estimator. Then, the naive density estimator would be $\widehat{f}(\boldsymbol{x}) = f(\boldsymbol{x}|\widehat{\boldsymbol{\theta}})$. From Taylor's expansion,

$$\sup_{\boldsymbol{x}} |f(\boldsymbol{x}|\boldsymbol{\theta}) - f(\boldsymbol{x}|\widehat{\boldsymbol{\theta}})| \leq \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \sup_{\boldsymbol{x},\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta})\|_2.$$

From MLE, we got the unifrom rate $O_p(1/\sqrt{n})$, if the gradient of density is uniformly bounded. Inparticular, Gaussian density, exponetial density satisfy the condition.

Although we cannot observe $\boldsymbol{Z}$ in the externeal data set, we need the relationship between $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the same in both external and internal data set.

(A9)

$$f^E(\boldsymbol{U}|\boldsymbol{X}) = f^I(\boldsymbol{U}|\boldsymbol{X}).$$

Remark 10.2. *Under assumption (A9) we can make sure the weighted expectation can change the measure.*

$$\begin{aligned}
E_{F^I}[\frac{f^E(\boldsymbol{X})}{f^I(\boldsymbol{X})}g(\boldsymbol{X})(\mu(\boldsymbol{U}) - h(\boldsymbol{X}))] &= \int g(\boldsymbol{X})(\mu(\boldsymbol{u}) - h(\boldsymbol{x}))\frac{f^E(\boldsymbol{x})}{f^I(\boldsymbol{x})}f^I(\boldsymbol{x})f^I(\boldsymbol{u}|\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{u} \\
&= \int g(\boldsymbol{X})(\mu(\boldsymbol{u}) - h(\boldsymbol{x}))f^E(\boldsymbol{x})f^E(\boldsymbol{u}|\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{u} \qquad (A9) \\
&= E_{F^E}[g(\boldsymbol{X})(\mu(\boldsymbol{U}) - h(\boldsymbol{X}))] = 0.
\end{aligned}$$

Theorem 10.1. *Under Assumption (A1)-(A6), (A8)(A9). Using the modify optimalization eqaution* (77)*, we have*

$$\sqrt{nl}(\widehat{\boldsymbol{\mu}}_{CK}(\boldsymbol{u}^*) - \mu(\boldsymbol{u}^*)) \to N\left(Bias(\boldsymbol{u}^*), Var(\boldsymbol{u}^*)\right),$$

$$Bias(\boldsymbol{u}^*) = \frac{\sqrt{c}}{2}A(\boldsymbol{u}^*) + \frac{\sqrt{c}\gamma^2}{2}(A(\boldsymbol{u}^*) - \frac{f^E(\boldsymbol{x}^*)}{f^I(\boldsymbol{x}^*)}g(\boldsymbol{x}^*)^\top \widetilde{\boldsymbol{\Sigma}}^{-1}_{g(\boldsymbol{X})}E_{F^E}[g(\boldsymbol{X})A(\boldsymbol{U})]),$$

$$Var(\boldsymbol{u}^*) = \frac{\sigma^2}{f(\boldsymbol{u}^*)}\int(\int \kappa(\boldsymbol{w}\gamma^2 - \boldsymbol{u})\kappa(\boldsymbol{w})d\boldsymbol{w})^2 d\boldsymbol{u}$$

where $A(\boldsymbol{u}) = \int \kappa(\boldsymbol{u})\boldsymbol{u}^\top \nabla^2 \mu(\boldsymbol{u})\boldsymbol{u}d\boldsymbol{u}$, $\widetilde{\boldsymbol{\Sigma}}_{g(\boldsymbol{X})}$ *is the covariance matrix with respend to* $g(\boldsymbol{X})$ *under distribution* f^E.

10.1 Proof of Theorem 10.1

Solve (77), we get

$$\widehat{\boldsymbol{\mu}} = (\boldsymbol{I} - \widetilde{\boldsymbol{P}})\widehat{\boldsymbol{\mu}}_K + \widetilde{\boldsymbol{P}}\boldsymbol{h}, \tag{78}$$

where $\widetilde{\boldsymbol{P}}$ is the projection matrix $\widetilde{\boldsymbol{G}}(\boldsymbol{X})(\widetilde{\boldsymbol{G}}(\boldsymbol{X})^\top\widetilde{\boldsymbol{G}}(\boldsymbol{X}))^{-1}\widetilde{\boldsymbol{G}}(\boldsymbol{X})^\top$, $\widetilde{\boldsymbol{G}}(\boldsymbol{X})$ is a $n \times n$ matrix whose jth row is $\widetilde{g}(\boldsymbol{X}_j) = \frac{f^E(\boldsymbol{X}_j)}{\widehat{f}^I(\boldsymbol{X}_j)}g(\boldsymbol{X}_j)$. Consider the same decomposition as section 6.1. The same arguement and assumption (A8) can make sure (16) (18)-(20) hold. Only (17) should be discussed separately.

$$\begin{aligned}(17) =&\frac{\frac{1}{n}\sum_{i=1}^{n}\kappa_l(\boldsymbol{u}^* - \boldsymbol{U}_i)\widetilde{g}(\boldsymbol{X}_i)}{\frac{1}{n}\widehat{f}_b(\boldsymbol{u}^*)}(\frac{1}{n}\sum_{i=1}^{n}\widetilde{g}(\boldsymbol{X}_i)\widetilde{g}(\boldsymbol{X}_i)^\top)^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{g}(\boldsymbol{X}_i)(\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i))\\ &\text{(From (A8) and Lemma ??)}\\ =&(1 + o_p(1))g(\boldsymbol{x}^*)\frac{f^E(\boldsymbol{x}^*)}{f^I(\boldsymbol{x}^*)}\widetilde{\boldsymbol{\Sigma}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{g}(\boldsymbol{X}_i)(\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i))\end{aligned}$$

From (A1) and (A8), and the continuity of reciprocal,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f^E(\boldsymbol{X}_i)}{\widehat{f}^I(\boldsymbol{X}_i)}g(\boldsymbol{X}_i)(\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)) = \frac{1}{n}\sum_{i=1}^{n}\frac{f^E(\boldsymbol{X}_i)}{f^I(\boldsymbol{X}_i)}g(\boldsymbol{X}_i)(\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)) + o_p(n^{-2/(p+4)}).$$

From CLT, (A9) (Remark 10.2),

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f^E(\boldsymbol{X}_i)}{f^I(\boldsymbol{X}_i)}g(\boldsymbol{X}_i)(\mu(\boldsymbol{U}_i) - h(\boldsymbol{X}_i)) = O_p(\frac{1}{\sqrt{n}}).$$

Hence, (17) is of order $o_p(n^{-2/(p+4)})$, which is depend on the unifromly convergence rate of $\widehat{f^I}$. From (A4), $\sqrt{nl}$ is of order $O_p(n^{2/(p+4)})$. We get $\sqrt{nl}$(17) is of order $o_p(1)$. We get our desired property.

References

[1] Norman E Breslow and Richard Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461, 1997.

[2] Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

[3] Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):449–460, 2000.

[4] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

[5] Jianqing Fan and Irene Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036, 1992.

[6] Wolfgang Hardle and James Stephen Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, 13(4):1465–1481, 12 1985.

[7] Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703, 2017.

[8] JF Lawless, JD Kalbfleisch, and CJ Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

[9] Jean D Opsomer. Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73(2):166–179, 2000.

[10] Jing Qin, Han Zhang, Pengfei Li, Demetrius Albanes, and Kai Yu. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180, 2015.

[11] Alastair J Scott and Chris J Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

[12] Jun Shao. Mathematical statistics, 2003.

[13] MP Wand and MC Jones. Monographs on statistics and applied probability. *Kernel smoothing*, 1995.

[14] Changbao Wu and Randy R Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193, 2001.