A SHARPER RATE OF CONVERGENCE OF
GENERALIZED EMPIRICAL LIKELIHOOD WEIGHTS
FOR INCORPORATING AUXILIARY DATA
INFORMATION *

BY K. C. G. CHAN†, F. LIU‡, S. C. P. YAM†,++ AND S. P. YUNG§,||

University of Washington†, The University of Hong Kong‡, Chinese
University of Hong Kong+++.

Due to practical constraints in data collection, certain variables are only observed in a subsample while other variables are available in full. An example is two phase sampling design, where an investigator selects a subset of Phase I subjects for detailed measurement of certain variables in Phase II, and this can dramatically reduce the cost of studies. Analysis using only Phase II data are inefficient and auxiliary information from Phase I can be incorporated by empirical likelihood weighting as in Hellerstein and Imbens (1999). We studied extensions of this method and the rate of convergence of the estimated weights. We found that the estimated weights has a uniform convergence rate of $O_p(n^{-3/2 + 1/2})$, for some $\alpha \geq 2$, which is sharper than the usual convergence rate of weights in generalized empirical likelihood estimator, which is only $O_p(n^{-1})$. While the rate of convergence for weights is faster, the associated weighted estimating equation is still having a nontrivial improvement in efficiency by incorporating Phase I auxiliary data.

1. Introduction. When a finite dimensional parameter of interest $\theta$ satisfies a moment condition $E[g(Y_i, X_i; \theta)] = 0$, where the dimension of $g$ is the same as $\theta$ and i.i.d. observations $(Y_i, X_i), i = 1, \ldots, N$ are observed, then $\theta$ can be conveniently estimated by solving the estimating equation $N^{-1} \sum_{i=1}^{N} g(Y_i, X_i; \theta) = 0$.

In many practical situations, certain variables of interest are only observed

---

*E-mail of corresponding author: kgchan@uw.edu
†The first author was partially supported by the National Institutes of Health grant RO1 AI089341.
‡The third author-Phillip Yam acknowledges financial support from The Hong Kong RGC GRF 502909, The Hong Kong RGC GRF 500111, The Hong Kong Polytechnic University Internal Grant APC0D, and The Hong Kong Polytechnic University Collaborative Research Grant G-YH96, and The Chinese University of Hong Kong Direct Grant 2010/2011 Project ID: 2060422.
§The forth author is supported by an HKU internal grant of code 200807176228.

Keywords and phrases: Auxiliary data, Data linkage, Estimating equation, Generalized empirical likelihood, Rate of convergence, Two phase sampling
in a subsample while other variables are available in full. An example is two
phase sampling design, where a random sample is drawn in Phase I and
variables $X$ are collected for the Phase I sample, and a Phase II subsample
is drawn from the Phase I sample and variables $Y$ are collected only from
the Phase II subsample. The two phase sampling design was first introduced
by Neyman (1938) and is often employed in practice to reduce the cost
of studies. Recently, large national-based data sets and registry data are
becoming increasingly available for researchers. While sample size can be
huge, those data sets may not be tailor made for answering specific scientific
questions and important variables may not be collected. Researchers usually
draw a much smaller sample and collect relevant variables for answering
specific scientific questions. When the larger population based data set is
available to the researchers, performing an analysis only using the smaller
data set correspond to an inefficient use of data. In these situations, the
observed data can be represented as $(R_i, Y_i, X_i), i = 1, \ldots, N$ where $R_i =
1$ if the subject is included in the Phase II sample and $R_i = 0$ otherwise.
Without loss of generality, we assume that $R_i = 1$ for $i = 1, \ldots, n$ where
$n < N$ and $R_i = 0$ otherwise. We also assume that the selection probability
$\mathbb{P}(R = 1|Y, X) = h(X)$ is chosen by the investigator and depends only on
Phase I data and $0 < h(X) \leq 1$.

To incorporate information from all Phase I data in linear regression anal-
ysis, Hellerstein and Imbens (1999) studied a weighted complete-case least
square method where weights are found by maximizing an empirical like-
lihood function incorporating information from Phase I data. Chen, Sitter
and Wu (2002), and Qin and Zhang (2007) also considered incorporating
auxiliary information using empirical likelihood for estimating the mean of
$Y$. In both settings, efficiency is usually gained by incorporating auxiliary
information from Phase I and the weights are converging to the inverse of
selection probability as sample size grows. A remaining important theoretical
question to be answered is how fast the estimated weights converges.
Intuitively, the weights cannot converge too fast since otherwise there will
be no asymptotic efficiency gain in using the weights over a complete-case
analysis. On the other hand, if the rate of convergence is too slow, then the
estimation by solving a weighted estimating equation can have sizable small
sample bias since the weighted estimating equation is only asymptotically
unbiased.

Rate of convergence for estimated weights in generalized empirical like-
lihood was studied in Newey and Smith (2004). They showed that when
generalized empirical likelihood is used to solve an over-identified estimat-
ing equations, the estimated weight has a rate of convergence in an order of
Our setting here is different in the sense that we are not solving the estimating equations from the model and the auxiliary data simultaneously but instead performs the estimation in a two step manner, each of the procedures solves a just-identified system of estimating equations. We show that the estimated weights in this case convergence in a sharper rate in the order of $O_p(n^{-3/2 + 1/\alpha})$, for an $\alpha \geq 2$. The theoretical results are presented in Section 2, and are confirmed with a simulation study in Section 4 where the rate of convergence can be very close to $O_p(n^{-3/2})$ in some cases. We also study the asymptotic properties of the solution of a corresponding weighted estimating equation in Section 3. We found that while the rate of convergence for weights is fast, the associated weighted estimating equation is still having a nontrivial improvement in efficiency by incorporating Phase I auxiliary data.

2. GEL weights and its rate of convergence. To incorporate information from $X$ in Phase I data, we note that for a real-valued vector function $u : \mathbb{R}^d \to \mathbb{R}^d$ we could estimate $\mu = \mathbb{E}[u(X)]$ by sample mean from Phase I data or a weighted average from the Phase II complete-case data. We shall denote probability measures on Phase I data and Phase II data, which are two equivalent probability measures, by $\mathbb{P}$ and $\mathbb{P}_s$ respectively. It is reasonable to assume that, firstly, as long as one data set goes to infinity, the other one will go to infinity as well; secondly, the proportion of Phase II data in Phase I data will also converge to a nonzero number. Since the sample mean incorporates all $X$ information, we would like to find weights for the Phase II complete-case data that satisfies

$$\bar{u}_N \triangleq \frac{1}{N} \sum_{i=1}^{N} u(x_i) = \sum_{i=1}^{n} \bar{P}_i u(x_i).$$

Here, $N$ and $n$ are sample size of Phase I data and Phase II data respectively. In this section, we shall focus on the construction of the weights $P_i$ and establish a uniform convergence rate for $P_i$. While in the next section, we shall study the asymptotic properties of the $\bar{P}_i$-weighted estimating equations.

To construct $\bar{P}_i$, let $\rho(x)$ be a function of scalar $x$ that is concave. Following the discussion in the work of Newey and Smith (2004), we shall impose a normalization assumption on $\rho(x)$: $\rho'(0) = -1$ and $\rho''(0) = -1$, which will not affect parameter estimation.

We measure the size of a compact set $E$ by

$$\|E\| = \max \{\|e\|, e \in E\}.$$
Let $\Lambda$ be the parameter space which is a compact subset in $\mathbb{R}^d$ and $0 \in \Lambda^0$. Consider a pair of functions in $\lambda \in \Lambda$:

$$f(\lambda) \triangleq \mathbb{E} \left[ \rho \left( \lambda^T (u(X) - \mu) \right) \right]$$

and

$$\tilde{f}_n(\lambda) \triangleq \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho \left( \lambda^T (u(x_i) - \tilde{u}_N) \right),$$

where the selection probability $h(X) \triangleq \mathbb{P}(R = 1|Y, X)$ is chosen by the investigator and depends only on Phase I data and $0 < h(X) \leq 1$. Their respective gradients in the interior $\Lambda^0$ are

$$\nabla f(\lambda) = \mathbb{E} \left[ \rho' \left( \lambda^T (u(X) - \mu) \right) (u(X) - \mu) \right]$$

and

$$\nabla \tilde{f}_n(\lambda) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T (u(x_i) - \tilde{u}_N) \right) (u(x_i) - \tilde{u}_N).$$

The function $f(\lambda)$ inherits strict concavity from $\rho(x)$. For each $u$, it is easy to see that since $\rho''$ is negative, and the Hessian matrix of $\tilde{f}_n(\lambda)$

$$H_n(\lambda) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(u(x_i) - \tilde{u}_N)(u(x_i) - \tilde{u}_N)(u(x_i) - \tilde{u}_N)^T$$

is semi-negative definite, and thus each function $\tilde{f}_n(\lambda)$ is concave. Since $\Lambda$ is compact, $\tilde{f}_n(\lambda)$ can achieve its maximum on $\Lambda$. For if, furthermore, the maximum value is obtained at an interior point $\tilde{\lambda}_n$ of $\Lambda$, then the first-order condition $R_n(\tilde{\lambda}_n) = 0$ implies Equation 2.1 with coefficients

$$\tilde{p}_i \triangleq \frac{1}{h(x_i)} \frac{1}{\sum_{j=1}^{n} \frac{1}{h(x_j)} \rho' \left( \lambda^T (u(x_j) - \tilde{u}_N) \right), \text{ for } i = 1, \ldots, n.}$$

It should be noted that, for any summation depending only on Phase II data, it could be extended to the summation on Phase I data with the help of the indicator $R$. Thus, for any function $F(Y, X)$ of random variables $(Y, X)$, we have

$$\frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} F(y_i, x_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{h(x_i)} F(y_i, x_i).$$
On the other hand, given \((Y, X)\), the indicator \(R\) follows Bernoulli distribution with \(\mathbb{E}[R|Y, X] = h(X)\) and \(\text{Var}[R|Y, X] = h(X)(1 - h(X))\). It follows that, by applying Tower property,

\[
\mathbb{E} \left[ \frac{R}{h(X)} F(Y, X) \right] = \mathbb{E} \left[ \frac{\mathbb{E}[R|Y, X]}{h(X)} F(Y, X)|Y, X \right] = \frac{\mathbb{E}[R|Y, X]}{h(X)} F(Y, X) = \mathbb{E} \left[ F(Y, X) \right].
\]

In the rest of this section, we shall establish the rate of convergence of weights \(\hat{P}_t\), under the following five mild technical conditions.

**Assumption 2.1** Fix \(\alpha \geq 2\). The sequence of random variables \(\{X_i\}_{i=1}^{\infty}\) is i.i.d. and uniformly bounded in \(L^\alpha\) norm, i.e. \(\mathbb{E}[\|X_i\|^{\alpha}] < \infty\).

**Assumption 2.2** \(u\) is a Lipschitz continuous function with Lipschitz-constant \(L\). The support of \(u(X)\) contains a neighborhood around its own theoretical mean \(\mu\). Moreover, \(\text{Var}[u(X)]\) is nonsingular.

**Assumption 2.3** The function \(\rho\) is thrice differentiable, strictly concave and normalized so that \(\rho(0) = 0\) and \(\rho'(0) = \rho''(0) = -1\). The set of all critical points of \(\rho''\), denoted by \(\mathcal{A}\), is compact.

Common examples of the function \(\rho\) that can satisfy Assumption 2.3 are power, logarithmic and exponential functions.

Define two random variables as follows:

\[
Z_1 \triangleq \left| \rho'(\|\Lambda\|\|u(X) - \mu\|) \right| \quad \text{and} \quad Z_2 \triangleq \left| \rho'(\frac{\|\Lambda\|\|u(X) - \mu\|}{\|\Lambda\|\|u(X) - \mu\|}) \right|.
\]

**Assumption 2.4** There exists a neighborhood \(\mathcal{V}_0 \subseteq \Lambda\) containing the zero such that:

\[
\mathbb{E} \left[ \sup_{\Lambda \in \mathcal{V}_0} \left| \rho''\left(\lambda^T (u(X) - \mu)\right) \right|^{\eta} \right] < \infty, \text{ for } \eta = 1, 2.
\]

Also, \(\mathbb{E}[Z_1^{\eta}] < \infty\) and \(\mathbb{E}[Z_2^{\eta}] < \infty\), for \(\eta = 1, 2\).

**Assumption 2.5** There exists \(0 < K < 1\) such that \(\lim_{N \to \infty} \frac{1}{N} K = K\).

Assumption 2.5 guarantees that \(\frac{1}{N} = O(1)\) and therefore whenever we talk about the convergence rate given by \(n\) or \(N\), we could replace one by the another. For the sake of convenience, we shall indicate all convergence rate in term of \(n\).

**Theorem 2.1.** If assumptions 2.1-2.5 are satisfied,

\[
\max_{i=1,\ldots,n} \left| Nh(x_i) \hat{P}_t - 1 \right| = O_p \left( n^{-\frac{1}{2} - \frac{1}{\eta}} \right).
\]
Therefore, for each $i = 1, \ldots, n$, we have
\[ \tilde{P}_i = \frac{1}{N h(x_i)} \left( 1 + O_p \left( n^{-\left(\frac{1}{2} - \frac{1}{2}\alpha \right)} \right) \right). \]

Before we prove the main theorem, we need the following lemmas.

**Lemma 2.1.** Under assumptions 2.1, 2.2 and 2.5,
\[ \max_{i=1,\ldots,n} \| u(x_i) - \mu \| = O_p(n^{\frac{1}{2}}) \]
and
\[ \max_{i=1,\ldots,n} \| u(x_i) - \tilde{u}_N \| = O_p(n^{\frac{1}{2}}). \]

**Proof.** Since $X_i$ are i.i.d. and $\mu$ is the common theoretical mean of $u(X_i)$, then $u(X_i) - \mu$ are also i.i.d. random variables. Thus
\[ P \left( \max_{i=1,\ldots,n} \| u(X_i) - \mu \| \leq cn^{\frac{1}{2}} \right) = \prod_{i=1}^{n} P \left( \| u(X_i) - \mu \| \leq cn^{\frac{1}{2}} \right) \]
\[ = \prod_{i=1}^{n} \left( 1 - P \left( \| u(X_i) - \mu \| \geq cn^{\frac{1}{2}} \right) \right) \]
\[ \geq \prod_{i=1}^{n} \left( 1 - 1 - \frac{\tilde{L}}{c^\alpha n} \right) \]
\[ \geq \prod_{i=1}^{n} \left( 1 - \frac{2^\alpha - 1}{c^\alpha n} \left( \mathbb{E} \| u(X_i) \|^\alpha + \| \mu \|^\alpha \right) \right) \]
\[ \geq \prod_{i=1}^{n} \left( 1 - \frac{2^\alpha - 1}{c^\alpha n} \left( \mathbb{L} \| X_i \|^\alpha + \| \mu \|^\alpha \right) \right) \]
\[ = \left( 1 - \frac{\tilde{L}}{c^\alpha n} \right)^n \]
where $\tilde{L} = 2^\alpha - 1 \left( \mathbb{L} \mathbb{E} \| X \|^\alpha + \| \mu \|^\alpha \right)$. The last term will converge to $e^{-\tilde{L}/c^\alpha}$ as $n$ goes to infinity. Thus, we have
\[ \max_{i=1,\ldots,n} \| u(x_i) - \mu \| = O_p(n^{\frac{1}{2}}). \]

Due to the central limit theorem, $\| \mu - \tilde{u}_N \| = O_p(n^{-\frac{1}{2}})$ and the second
conclusion could be deduced by triangle inequality as follows:
\[
\max_{i=1,\ldots,n} \|u(x_i) - \bar{u}_N\| \leq \max_{i=1,\ldots,n} \|u(x_i) - \mu\| + \|\mu - \bar{u}_N\| \\
= O_p(u^{\frac{1}{2}}) + O_p(u^{-\frac{1}{2}}) \\
= O_p(u^{\frac{1}{2}}).
\]

We next define a point \(\lambda^*\) and a sequence of subsets \(\{\tilde{\Lambda}_n\}_{n=1}^{\infty}\) in \(\mathbb{R}^d\) as:

1. \(\lambda^* \triangleq \arg \max_{\lambda \in \Lambda} \mathbb{E} \left[ \rho \left( \lambda^T (u(X) - \mu) \right) \right] \)
2. \(\tilde{\Lambda}_n \triangleq \left\{ \lambda_n : \lambda_n = \arg \max_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho \left( \lambda^T (u(x_i) - \bar{u}_N) \right) \right\} \)

Both \(\lambda^*\) and \(\tilde{\Lambda}_n\) are well-defined due to the concavity of functions \(f(\lambda)\) and \(\tilde{f}_n(\lambda)\) respectively. Since the strictly concave function \(f(\lambda)\) has unique maximum and its first order condition is satisfied at zero, i.e. \(\nabla f(0) = 0\), then \(\lambda^* = 0\) which is an interior point in \(\Lambda\). It is natural to expect that there exists \(\lambda_n \in \tilde{\Lambda}_n\) such that \(\lambda_n\) is an interior point in \(\Lambda\). Hence each estimated weight \(\tilde{P}_i\) as defined in Equation 2.2 is well-posed, and altogether satisfy Equation 2.1; this claim can be promised by the next three lemmas 2.2 – 2.4.

**Lemma 2.2.** Under assumptions 2.1-2.5, if \(\nabla \tilde{f}_n(\lambda)\) uniformly converges to \(\nabla f(\lambda)\) in probability on \(\nu_0\), then \(\lim_{n \to \infty} \mathbb{P} \left( \tilde{\Lambda}_n \cap \nu_0 \neq \emptyset \right) = 1\).

**Proof.** For any fixed value \(\tilde{\lambda} \in \nu_0\), we integrate the gradient fields \(\nabla f(\lambda)\) and \(\nabla \tilde{f}_n(\lambda)\) respectively along the directed line connecting from 0 to \(\tilde{\lambda}\). Then
\[
\int_0^{\tilde{\lambda}} \nabla f(\lambda) d\lambda = f(\tilde{\lambda}) - f(0) = f(\tilde{\lambda}),
\]
\[
\int_0^{\tilde{\lambda}} \nabla \tilde{f}_n(\lambda) d\lambda = \tilde{f}_n(\tilde{\lambda}) - \tilde{f}_n(0) = \tilde{f}_n(\tilde{\lambda}),
\]
where \(f(0) = 0 = \tilde{f}_n(0)\). Take a compact set \(\mathcal{W} \subseteq \nu_0\) such that zero is an interior point of \(\mathcal{W}\). The uniform convergence of \(\nabla \tilde{f}_n\) to \(\nabla f(\lambda)\) on \(\mathcal{W}\) guarantees the uniform convergence of \(\tilde{f}_n\) to \(f\) on \(\mathcal{W}\). Since \(f(\lambda)\) is strictly concave, \(\nabla f(\lambda)\) has exactly one root at zero; therefore, \(f(\lambda)\) attains its maximum at zero and is non-positive on set \(\mathcal{W}\). Moreover \(\max_{\lambda \in \partial \mathcal{W}} f(\lambda) < 0;\)
indeed, suppose \( f(\lambda) = \max_{\lambda \in \partial W} f(\lambda) \geq 0 \) where \( \lambda \in \partial W \). Since \( f \) is strictly concave, there exists \( \alpha \in (0, 1) \) such that
\[
f\left(\alpha \lambda + (1 - \alpha) 0\right) > \alpha f(\lambda) + (1 - \alpha) f(0) \geq 0.
\]
This contradicts with the fact that \( f(\lambda) \) has a unique maximum at zero, and hence \( \max_{\lambda \in \partial W} f(\lambda) < 0 \). On the other hand, uniform convergence in probability of \( \{\hat{f}_n\}_{n=1}^\infty \) implies that
\[
\lim_{n \to \infty} P\left( \max_{\lambda \in \partial W} \hat{f}_n(\lambda) < 0 - \hat{f}_n(0) \right) = 1.
\]
It follows that every point \( \lambda_n \) at which function \( \hat{f}_n(\lambda) \) achieves its maximum on \( W \) is an interior point of \( W \subset V_0 \) with probability approaching to one. Thus, by applying first order condition \( \nabla \hat{f}_n(\lambda_n) = 0 \), \( f_n(\lambda) \) achieves its maximum on \( V_0 \) at \( \lambda_n \) as well, i.e. there exists at least one element in the intersection of \( \hat{\Lambda}_n \) and \( V_0 \) with probability approaching to one. In summary,
\[
\lim_{n \to \infty} P\left( \hat{\Lambda}_n \cap V_0 \neq \emptyset \right) = 1
\]
and for each \( \lambda_n \in \hat{\Lambda}_n \cap V_0 \), the first order condition is satisfied, i.e. \( \nabla \hat{f}_n(\lambda_n) = 0 \). \( \square \)

We now choose a representative in each \( \hat{\Lambda}_n \). We claim that there is a unique point \( \hat{\lambda}_n \in \hat{\Lambda}_n \) such that \( \|\hat{\lambda}_n\| = \inf_{\lambda_n \in \hat{\Lambda}_n} \|\lambda_n\| \) with probability approaching to one. Based on Lemma 2.2, \( \hat{\Lambda}_n \cap W = \{ \lambda \in W, \nabla \hat{f}_n(\lambda) = 0 \} \) is non-empty with probability approaching to one. Since \( \hat{\Lambda}_n \cap W \) is clearly a closed subset of compact set \( W \), it is also compact. Therefore, the norm function \( \| \cdot \| : \hat{\Lambda}_n \cap W \to \mathbb{R} \), being strictly convex, can achieve its unique minimum at \( \hat{\lambda}_n \in \hat{\Lambda}_n \cap W \).

The next lemma summarizes the nature of convergence and its rate of the sequence \( \{\hat{\lambda}_n\}_{n=1}^\infty \).

**Lemma 2.3.** Under assumptions 2.1-2.5, suppose that we also have
\[
\left\| \nabla \hat{f}_n(\lambda) - \frac{1}{N} \sum_{i=1}^N \frac{1}{h(x_i)} \rho'(\lambda^T (u(x_i) - \mu))(u(x_i) - \mu) \right\|
\]
uniformly converges to zero on \( V_0 \) with order \( O_p(n^{-\frac{1}{2}}) \), then \( \hat{\lambda}_n = O_p(n^{-\frac{1}{2}}) \) and
\[
\max_{i=1, \ldots, n} \left| \hat{\lambda}^T_n (u(x_i) - \bar{u}_N) \right| = O_p(n^{-\frac{1}{2}}).
\]
Proof. Lemma 2.2 implies the first order condition $\nabla \tilde{f}_n(\tilde{\lambda}_n) = 0$. Then by invoking the additional assumption, we have

$$O_p(n^{-\frac{1}{2}}) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'(\tilde{\lambda}^T_n (u(x_i) - \mu)) (u(x_i) - \mu).$$

By using Taylor series expansion at point zero, there exists $\phi_i$ between zero and $\tilde{\lambda}^T_n (u(x_i) - \mu)$ such that

$$\frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \tilde{\lambda}^T_n (u(x_i) - \mu) \right) (u(x_i) - \mu)$$

$$= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \left( -1 + \rho''(\phi_i) \left( \tilde{\lambda}^T_n (u(x_i) - \mu) \right) \right) (u(x_i) - \mu)$$

$$= O_p(n^{-\frac{1}{2}}) + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\phi_i) \left( \tilde{\lambda}^T_n (u(x_i) - \mu) \right) (u(x_i) - \mu).$$

It follows that

$$-\frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\phi_i) \left( \tilde{\lambda}^T_n (u(x_i) - \mu) \right) (u(x_i) - \mu) = O_p(n^{-\frac{1}{2}}).$$

We now multiply $\tilde{\lambda}^T_n$ on both sides of the previous equation and obtain:

$$-\frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\phi_i) \left( \tilde{\lambda}^T_n (u(x_i) - \mu) \right)^2 \leq ||\tilde{\lambda}_n|| O_p(n^{-\frac{1}{2}}).$$

Note that $\mathbb{E} [u(X) - \mu] = 0$. Under Assumption 2.2, the support of random variable $u(X)$ contains a neighborhood of $\mu$, and so for small enough $\varepsilon > 0$,

$$P_\varepsilon \triangleq \mathbb{P} \left( ||u(X) - \mu|| \leq \frac{\varepsilon}{||\Lambda||} \right) > 0.$$ 

Now, choose small enough $\varepsilon > 0$ so that $\rho''(x) \leq -\frac{1}{2}$ for all $0 < x \leq \varepsilon$. This could be done because $\rho''$ is continuous and $\rho''(0) = -1$. Define a sequence of random times $\{\tau_i\}_{i=0}^\infty$ such that $\tau_0 = 0$ and

$$\tau_{i+1} = \min \left\{ n > \tau_i : ||u(X_n) - \mu|| \leq \frac{\varepsilon}{||\Lambda||} \right\},$$

with convention $\tau_i = \infty$ if the corresponding set is empty. Due to Borel’s Law of Large Number,

$$\mathbb{P} \left( \lim_{n \to \infty} \frac{M_n}{n} = P_\varepsilon \right) = 1,$$
where \( M_n = \# \{ i : \tau_i \leq n \} \). For each \( \tau_i \leq n \),

\[
|\phi_{\tau_i}| \leq \left| \frac{\hat{\lambda}_n^T}{n} (u(x_{\tau_i}) - \mu) \right| \leq \varepsilon.
\]

Due to the choice of \( \varepsilon \), \( \rho''(\phi_{\tau_i}) \leq -\frac{1}{2} \) and

\[
- \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\phi_{\tau_i}) \left( \frac{\hat{\lambda}_n^T}{n} (u(x_i) - \mu) \right)^2 \\
\geq - \frac{1}{N} \sum_{\tau_i \leq n} \frac{1}{h(x_{\tau_i})} \rho''(\phi_{\tau_i}) \left( \frac{\hat{\lambda}_n^T}{n} (u(x_{\tau_i}) - \mu) \right)^2 \\
\geq \frac{n}{N} \frac{M_n}{2n} \frac{1}{M_n} \sum_{\tau_i \leq n} \frac{1}{h(x_{\tau_i})} \left( \frac{\hat{\lambda}_n^T}{n} (u(x_{\tau_i}) - \mu) \right)^2,
\]

where

\[
\frac{n}{N} \frac{M_n}{2n} \frac{1}{M_n} \sum_{\tau_i \leq n} \frac{1}{h(x_{\tau_i})} (u(x_{\tau_i}) - \mu)(u(x_{\tau_i}) - \mu)^T
\]

converges to \( \frac{1}{2} K P \varepsilon \text{Var}_{\varepsilon}[u(X_{\tau_i}) - \mu] \) as \( n \) goes to infinity. We claim that \( \text{Var}_{\varepsilon}[u(X_{\tau_i}) - \mu] \) is non-degenerate. Indeed, the distribution of \( u(X_{\tau_i}) \) is a conditional distribution of \( u(X) \) given the event that \( \|u(X) - \mu\| \leq \|\varepsilon\| \).

Under Assumption 2.2, \( u(X) \) has a support around \( \mu \), therefore \( u(X_{\tau_i}) \) also has a support of full dimension \( d \) around \( \mu \); in particular, each projection of \( u(X) \) on any hyperplane also has a non-degenerated distribution, and hence \( u(X_{\tau_i}) \) has a non-degenerated variance. Thus, there exists a positive constant \( C \) so that with probability approaching to one,

\[
- \frac{1}{2} K P \varepsilon \text{Var}_{\varepsilon}[u(X_{\tau_i}) - \mu] \leq \|\hat{\lambda}_n\| O_p(n^{-\frac{1}{2}}),
\]

and moreover \( \|\hat{\lambda}_n\| = O_p(n^{-\frac{1}{2}}) \), i.e., \( \hat{\lambda}_n = O_p(n^{-\frac{1}{2}}) \). The second conclusion is an immediate consequence of Lemma 2.1 and \( \hat{\lambda}_n = O_p(n^{-\frac{1}{2}}) \).

The uniform convergence of \( \nabla \hat{f}_n(\lambda) \) plays an indispensable role in the proofs of lemmas 2.2 and 2.3, we now establish this in the next lemma:

**Lemma 2.4.** Under assumptions 2.1-2.5, we have

(i) \( \nabla \hat{f}_n(\lambda) \) uniformly converges to \( \nabla f(\lambda) \) on \( \mathcal{V}_0 \);

(ii) \[
\left\| \nabla \hat{f}_n(\lambda) - \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'(\lambda^T(u(x_i) - \mu)) (u(x_i) - \mu) \right\|
\]

uniformly converges to zero on \( \mathcal{V}_0 \) with order \( O_p(n^{-\frac{1}{2}}) \).
RATE OF CONVERGENCE FOR GEL WEIGHTS

PROOF. Recall the definition of

$$\nabla \hat{f}_n(\lambda) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T (u(x_i) - \bar{u}_N) \right) (u(x_i) - \bar{u}_N),$$

and in accordance with the first order condition, we have $\nabla \hat{f}_n(\lambda_0) = 0$. We now expand each summand up to the first order term around the real number $\lambda^T (u(x_i) - \mu)$, for each $i = 1, \ldots, n$, and there exists a real number $\xi_i$ lying between $\lambda^T (u(x_i) - \mu)$ and $\lambda^T (u(x_i) - \bar{u}_N)$ such that

$$\begin{align*}
(2.3) \quad \nabla \hat{f}_n(\lambda) &= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T (u(x_i) - \mu) \right) (u(x_i) - \bar{u}_N) \\
&\quad + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i) \left( \lambda^T (\mu - \bar{u}_N) \right) (u(x_i) - \bar{u}_N)
\end{align*}$$

$$= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T (u(x_i) - \mu) \right) (u(x_i) - \mu)$$

$$\quad + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T (u(x_i) - \mu) \right) (u(x_i) - \bar{u}_N)$$

$$\quad + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i) \left( \lambda^T (\mu - \bar{u}_N) \right) (u(x_i) - \mu)$$

$$\quad + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i) \left( \lambda^T (\mu - \bar{u}_N) \right) (\mu - \bar{u}_N).$$

In order to establish the second claim, it is sufficient to show that the last three summands in $\nabla \hat{f}_n(\lambda)$ uniformly converge to zero in probability with order $O_p(n^{-\frac{1}{2}})$.

a) Due to the central limit theorem and $\rho'$ is a decreasing function, we can deduce the uniform convergence to zero in probability of the second summand in $\nabla \hat{f}_n(\lambda)$ from Assumption 2.4 based on the following argu-
\[
\sup_{\lambda \in \Lambda} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'(\lambda^T(u(x_i) - \mu))(\mu - \bar{u}_N) \right\| \\
\leq \|\mu - \bar{u}_N\| \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \sup_{\lambda \in \Lambda} \left| \rho'(\lambda^T(u(x_i) - \mu)) \right| \right\| \\
\leq \|\mu - \bar{u}_N\| \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \left| \rho'(||\Lambda|| \|u(x_i) - \mu\|) \right| \right\| \\
+ \|\mu - \bar{u}_N\| \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \left| \rho'(-||\Lambda|| \|u(x_i) - \mu\|) \right| \right\| \\
= \|\mu - \bar{u}_N\| \left( \mathbb{E}[Z_1] + \mathbb{E}[Z_2] + O_p(n^{-\frac{1}{2}}) \right) \\
= O_p(n^{-\frac{1}{2}}).
\]

b) Fix \(0 < \gamma < \frac{1}{2}\) and \(\Omega_n\) represents the event set that \(\|\mu - \bar{u}_N\| \leq n^{-\gamma}\). On the set \(\Omega_n\), since \(\xi_i\) is located between \(\lambda^T(u(x_i) - \mu)\) and \(\lambda^T(u(x_i) - \bar{u}_N)\) and depends on \(\lambda\),

\[
|\xi_i - \lambda^T(u(x_i) - \mu)| \leq |\lambda^T(\mu - \bar{u}_N)| \leq ||\Lambda|| n^{-\gamma}\]

Therefore, \(|\rho''(\xi_i)|\) is bounded by three parts and we could deduce the following inequality:

\[
\sup_{\lambda \in V_0} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i)(u(x_i) - \mu) \right\| \\
\leq \sup_{\lambda \in V_0} \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \left| \rho''(\lambda^T(u(x_i) - \mu) + ||\Lambda|| n^{-\gamma}) \|u(x_i) - \mu\| \right| \\
+ \sup_{\lambda \in V_0} \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \left| \rho''(\lambda^T(u(x_i) - \mu) - ||\Lambda|| n^{-\gamma}) \|u(x_i) - \mu\| \right| \\
+ \max_{x \in \mathcal{A}} |\rho''(x)| \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \|u(x_i) - \mu\| \right).
\]
Under Assumption 2.4, we deduce that
\[
\mathbb{E} \left[ \sup_{\lambda \in \mathcal{V}_0} \left\| \rho'' (\lambda^T (u(X) - \mu)) (u(X) - \mu) \right\| \right] 
\leq \mathbb{E} \left[ \sup_{\lambda \in \mathcal{V}_0} \left| \rho'' (\lambda^T (u(X) - \mu)) \right| \left\| (u(X) - \mu) \right\| \right] 
+ \text{Con}_{\mathcal{I}} \left[ \sup_{\lambda \in \mathcal{V}_0} \left| \rho'' (\lambda^T (u(X) - \mu)) \right| \left\| (u(X) - \mu) \right\| \right] 
\leq \mathbb{E} \left[ \sup_{\lambda \in \mathcal{V}_0} \left| \rho'' (\lambda^T (u(X) - \mu)) \right| \right] \mathbb{E} \left[ \left\| (u(X) - \mu) \right\| \right] 
+ \left( \text{Var}_{\mathcal{I}} \left[ \sup_{\lambda \in \mathcal{V}_0} \left| \rho'' (\lambda^T (u(X) - \mu)) \right| \right] \text{Var}_{\mathcal{I}} \left[ \left\| (u(X) - \mu) \right\| \right] \right)^{\frac{1}{2}} 
< \infty.
\]

Since \( \mathcal{A} \) is compact and \( \max_{x \in \mathcal{A}} |\rho''(x)| < \infty \), and by applying the law of large number, we deduce that
\[
\lim_{n \to \infty} \sup_{\lambda \in \mathcal{V}_0} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' (\xi_i) (u(x_i) - \mu) \right\| 
\leq 2 \mathbb{E} \left[ \sup_{\lambda \in \mathcal{V}_0} \left| \rho'' (\lambda^T (u(X) - \mu)) \right| \left\| (u(X) - \mu) \right\| \right] 
+ \max_{x \in \mathcal{A}} |\rho''(x)| \mathbb{E} \left[ \left\| (u(X) - \mu) \right\| \right] 
< \infty.
\]

Since \( u^\gamma \|\mu - \bar{u}_N\| = O_p(n^{-\gamma}) \), for any \( \delta > 0 \), there exist \( N \in \mathbb{N} \) and a constant \( C \) such that whenever \( n > N \), one can have
\[
\mathbb{P}(\Omega_n^c) = 1 - \mathbb{P} \left( \|\mu - \bar{u}_N\| \leq n^{-\gamma} \right) < \frac{\delta}{2},
\]
and
\[
\mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_0} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(X_i)} \rho'' (\xi_i) (u(X_i) - \mu) \right\| > C; \Omega_n \right) < \frac{\delta}{2}.
\]

Therefore, whenever \( n > N \), we can have
\[
\mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_0} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(X_i)} \rho'' (\xi_i) (u(X_i) - \mu) \right\| > C \right) 
\leq \mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_0} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(X_i)} \rho'' (\xi_i) (u(X_i) - \mu) \right\| > C; \Omega_n \right) + \mathbb{P}(\Omega_n^c) 
< \delta.
\]
This implies that \( \sup_{\lambda \in \mathcal{V}_h} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i)(u(x_i) - \mu) \left( \lambda^T (\mu - \bar{u}_N) \right) \right\| \) is bounded in probability. Therefore, we can now deduce the order of convergence of:

\[
\sup_{\lambda \in \mathcal{V}_h} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i)(u(x_i) - \mu) \left( \lambda^T (\mu - \bar{u}_N) \right) \right\| \leq \sup_{\lambda \in \mathcal{V}_h} \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho''(\xi_i)(u(x_i) - \mu) \right\| \sup_{\lambda \in \mathcal{V}_h} |\lambda^T (\mu - \bar{u}_N)| = O_p(n^{-\frac{1}{2}})
\]

which is a uniform convergence since

\[ \sup_{\lambda \in \mathcal{V}_h} |\lambda^T (\mu - \bar{u}_N)| \leq \|\lambda\| O_p(n^{-\frac{1}{2}}). \]

\[ c) \] For the last summand in (2.3), by using the similar argument as presented in the part \( b) \), on the set \( \Omega_n \), we have

\[
\sup_{\lambda \in \mathcal{V}_h} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho''(\xi_i)}{h(x_i)} \right| \leq \sup_{\lambda \in \mathcal{V}_h} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \right| \rho'' \left( \lambda^T (u(x_i) - \mu) + \|\lambda\| n^{-\gamma} \right) \]

\[ + \sup_{\lambda \in \mathcal{V}_h} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \right| \rho'' \left( \lambda^T (u(x_i) - \mu) - \|\lambda\| n^{-\gamma} \right) \]

\[ + \max_{x \in \mathcal{A}} \left| \rho''(x) \right| \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \right) \]

\[ \to 2E \sup_{\lambda \in \mathcal{V}_h} \left| \rho'' \left( \lambda^T (u(X) - \mu) \right) \right| + \max_{x \in \mathcal{A}} \left| \rho''(x) \right| < \infty. \]

For any \( \delta > 0 \), there exist \( N \in \mathbb{N} \) and a constant \( C \) such that whenever \( n > N \),

\[ \mathbb{P}(\Omega_n^C) < \frac{\delta}{2} \]

and

\[ \mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_h} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho''(\xi_i)}{h(x_i)} \right| > C; \, \Omega_n \right) < \frac{\delta}{2}. \]
Therefore, whenever \( n > N \), we also have:

\[
\mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_0} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho'(\xi_i)}{h(x_i)} \right| > C \right) 
\leq \mathbb{P} \left( \sup_{\lambda \in \mathcal{V}_0} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho''(\xi_i)}{h(x_i)} \right| > C; \Omega_n \right) + \mathbb{P}(\Omega_n^c)
\leq \delta,
\]

and hence \( \sup_{\lambda \in \mathcal{V}_0} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho'(\xi_i)}{h(x_i)} \right| = O_p(1) \). Similar to the derivation in the part (b), we can also conclude that the last summand in (2.3) uniformly converges to zero on \( \mathcal{V}_0 \) in probability based on the fact that

\[
\sup_{\lambda \in \mathcal{V}_0} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho''(\xi_i)}{h(x_i)} \right| (\lambda^T (\mu - \bar{u}_N)) (\mu - \bar{u}_N) \leq \|\mu - \bar{u}_N\| \sup_{\lambda \in \mathcal{V}_0} \left| \frac{1}{N} \sum_{i=1}^{n} \frac{\rho'(\xi_i)}{h(x_i)} \right| \sup_{\lambda \in \mathcal{V}_0} |\lambda^T (\mu - \bar{u}_N)|
\leq O_p(n^{-1}).
\]

Finally, for the first claim, it could be deduced by augmenting the parts (a) to (c) with the weak uniform convergence of the first summand in (2.3) to \( \nabla f(\lambda) \). For each \( \lambda \in \Lambda \), \( \rho'(\lambda^T (u(X_i) - \mu)) (u(X_i) - \mu) \) are i.i.d. random variables. Under assumptions 2.2 and 2.4, for \( i = 1, 2 \),

\[
\mathbb{E} \left[ \|u(X) - \mu\| Z_i \right] = \mathbb{E} \left[ \|u(X) - \mu\| \mathbb{E}[Z_i] + \text{Cov} \left[ \|u(X) - \mu\|, Z_i \right] \right] 
\leq \mathbb{E} \left[ \|u(X) - \mu\| \mathbb{E}[Z_i] + \text{Var} \left[ \|u(X) - \mu\| \right] \text{Var} \left[ Z_i \right] \right]^{1/2} 
\leq \infty.
\]

Since \( \rho' \) is a monotonic function, then we have

\[
\sup_{\lambda \in \Lambda} \|\nabla f(\lambda)\| = \sup_{\lambda \in \Lambda} \left\| \mathbb{E} \left[ \rho'(\lambda^T (u(X) - \mu)) (u(X) - \mu) \right] \right\|
\leq \mathbb{E} \left[ \sup_{\lambda \in \Lambda} \left\| \rho'(\lambda^T (u(X) - \mu)) (u(X) - \mu) \right\| \right]
\leq \mathbb{E} \left[ \|u(X) - \mu\| \max \{Z_1, Z_2 \} \right]
\leq \mathbb{E} \left[ \|u(X) - \mu\| Z_1 \right] + \mathbb{E} \left[ \|u(X) - \mu\| Z_2 \right]
\leq \infty.
\]
In order to establish the uniform convergence of the first summand in (2.3), we shall apply the uniform weak law of large number (see, e.g., Bierens (2004), Newey and MacFadden (1994)) with which certain regularity conditions have to be verified: (i) Firstly, the first summand in (2.3) converges to $\nabla f(\lambda)$ pointwisely in accordance with the standard weak law of large number; (ii) secondly, there exists a dominating function

$$d(X) \overset{\triangle}{=} \frac{R}{h(X)} \|u(X) - \mu\| \max\{Z_1, Z_2\}$$

such that $\mathbb{E}[d(X)] < \infty$, and for every $x \in \mathbb{R}^d$,

$$\left\| \frac{R}{h(x)} \rho'(\lambda^T(u(x) - \mu))(u(x) - \mu) \right\| \leq \|d(x)\|$$

(iii) thirdly, $\frac{R}{h(x)} \rho'(\lambda^T(u(x) - \mu))(u(x) - \mu)$ is continuous at each $\lambda \in \Lambda$ for all $x \in \mathbb{R}^d$ and the parameter space $\Lambda$ is compact. Therefore, by applying the uniform weak law of large number, the first summand in (2.3) uniformly converges to $\nabla f(\lambda)$ in probability and the proof is completed. \qed

Up to this point, we have shown that, firstly, $\lambda_n$ is an interior point of $\mathcal{V}_0$ with probability approaching to one. This fact implies the first order condition $\nabla f_n(\lambda_n) = 0$ which is used to define calibrated weights $\hat{P}_i$ by Equation (2.2) such that $\sum_{i=1}^n \hat{P}_i u(x_i) = \bar{u}_N$ holds for a given calibration function $u(\cdot)$. Secondly, $\lambda_n = O_p(n^{-\frac{1}{2}})$ and

$$\max_{i=1, \ldots, n} \left| \lambda_n^T(u(x_i) - \bar{u}_N) \right| = O_p(n^{\frac{1}{2} - \frac{1}{2}}),$$

which will be useful in the main theorem. We now prove the main theorem by applying the lemmas 2.1-2.4:

**Proof of Main Theorem.** The first-order condition for $\lambda_n$ implies

$$\sum_{i=1}^n \hat{P}_i u(x_i) = \bar{u}_N,$$

where

$$\hat{P}_i \overset{\triangle}{=} \frac{1}{h(x_i)} \frac{\rho'(\lambda_n^T(u(x_i) - \bar{u}_N))}{\sum_{j=1}^n \frac{1}{h(x_j)} \rho'(\lambda_n^T(u(x_j) - \bar{u}_N))}, \text{ for } i = 1, \ldots, n.$$
By applying the mean value theorem twice, for each \( i = 1, \ldots, n \), there exists \( \xi_i \) lying between \( \lambda_n^T(u(x_i) - \mu) \) and \( \lambda_n^T(u(x_i) - \bar{u}_N) \), and \( \phi_i \) lying between zero and \( \lambda_n^T(u(x_i) - \mu) \) such that

\[
(2.4) \quad \rho' \left( \lambda_n^T(u(x_i) - \bar{u}_N) \right) = -1 + \rho''(\phi_i) \left( \lambda_n^T(u(x_i) - \mu) \right) + \rho''(\xi_i) \left( \lambda_n^T(u(x_i) - \mu) \right),
\]

and then

\[
(2.5) \quad \sum_{j=1}^{n} \frac{1}{Nh(x_j)} \rho' \left( \lambda_n^T(u(x_j) - \bar{u}_N) \right)

= \sum_{j=1}^{n} \frac{-1}{Nh(x_j)} + \lambda_n^T(n) \sum_{j=1}^{n} \frac{\rho''(\phi_j)}{Nh(x_j)} (u(x_j) - \mu)

+ \lambda_n^T(\mu - \bar{u}_N) \sum_{j=1}^{n} \frac{\rho''(\xi_j)}{Nh(x_j)}.
\]

Due to the central limit theorem, the first summand on right hand side of (2.5) converges to \(-1\) in probability with order \( O_p(n^{-\frac{1}{2}}) \). For the other two summands, similar to the derivations as in parts (b) and (c) in Lemma 2.4, one can show that:

\[
\left\| \sum_{j=1}^{n} \frac{\rho''(\phi_j)}{Nh(x_j)} (u(x_i) - \mu) \right\|

\leq \sup_{\lambda \in V_0} \sum_{j=1}^{n} \frac{1}{Nh(x_j)} \left| \rho''(\lambda^T(u(x_j) - \mu)) \right| \|u(x_j) - \mu\|

= \mathbb{E} \left[ \sup_{\lambda \in V_0} \left| \rho''(\lambda^T(u(X) - \mu)) \right| \|u(X) - \mu\| \right] + O_p(n^{-\frac{1}{2}})

= O_p(1)
\]

and

\[
\left\| \sum_{j=1}^{n} \frac{\rho''(\xi_j)}{Nh(x_j)} \right\| = O_p(1)
\]

respectively with probability approaching to one. It follows that

\[
\left| 1 + \sum_{j=1}^{n} \frac{1}{Nh(x_j)} \rho' \left( \lambda_n^T(u(x_j) - \bar{u}_N) \right) \right| = O_p(n^{-\frac{1}{2}}).
\]
Next, we study the convergence of each term on the right hand side of Equation (2.4). By using Lemma 2.3, since $\alpha \geq 2$,

$$\max_{i=1,\ldots,n} |\phi_i| \leq \max_{i=1,\ldots,n} \left| \tilde{\lambda}_n^T (u(x_i) - \mu) \right| = O_p(n^{-\frac{1}{2}}) \leq O_p(1)$$

and

$$\max_{i=1,\ldots,n} |\xi_i| \leq \max_{i=1,\ldots,n} \left| \tilde{\lambda}_n^T (u(x_i) - \mu) \right| + \max_{i=1,\ldots,n} \left| \tilde{\lambda}_n^T (u(x_i) - \bar{u}_N) \right|$$

$$= O_p(n^{-\frac{1}{2} - \frac{1}{2}}) \leq O_p(1),$$

since $\rho''$ is uniformly continuous on compacta, we have:

$$\max_{i=1,\ldots,n} |\rho''(\phi_i)| \leq O_p(1) \text{ and } \max_{i=1,\ldots,n} |\rho''(\xi_i)| \leq O_p(1).$$

It follows that

$$\max_{i=1,\ldots,n} \left| 1 + \rho' \left( \tilde{\lambda}_n^T (u(x_i) - \bar{u}_N) \right) \right|$$

$$\leq \max_{i=1,\ldots,n} \left| \tilde{\lambda}_n^T (u(x_i) - \mu) \right| \max_{i=1,\ldots,n} |\rho''(\phi_i)|$$

$$+ \left| \tilde{\lambda}_n^T (\mu - \bar{u}_N) \right| \max_{i=1,\ldots,n} |\rho''(\xi_i)|$$

$$\leq O_p(n^{-\frac{1}{2} - \frac{1}{2}}),$$

and we can now establish the desired claim:

$$\max_{i=1,\ldots,n} \left| 1 + Nh(x_i) \hat{p}_i \right| = \frac{O_p(n^{-\frac{1}{2} - \frac{1}{2}})}{1 + O_p(n^{-\frac{1}{2}})} = O_p(n^{-\frac{1}{2} - \frac{1}{2} - \frac{1}{2} - \frac{1}{2}}).$$

$$\square$$

3. Asymptotic properties of weighted estimating equation. In Section 2, we have used generalized empirical likelihood (GEL) to find weights that also satisfy calibration constraints (2.1). The true parameter $\theta_0 \in \Theta$ satisfies $\mathbb{E}[g(Y, X, \theta_0)] = 0$ where $\Theta$ denotes the underlying parameter space. Before we find the estimator of $\theta_0$ based on GEL method, we first list out some mild technical conditions as follows:

Assumption 3.1 Parameter space $\Theta$ is a compact subset of $\mathbb{R}^d$.

Assumption 3.2 The function $g(Y, X, \cdot) = (g_1(Y, X, \cdot), \ldots, g_d(Y, X, \cdot))^T$ is continuously differentiable on $\Theta$ and $\mathbb{E} \left[ \sup_{\theta \in \Theta} \| g(Y, X, \theta) \|^2 \right] < \infty$. 

The equation $\mathbb{E} \left[ g(Y, X, \theta) \right] = 0$ has a unique root at the true parameter $\theta_0$ on $\Theta$. There exists a neighborhood $\Theta_0 \subset \Theta$ containing $\theta_0$ such that the matrix $\mathbb{E} \left[ \partial g (Y, X, \theta) / \partial \theta \right]$ is nonsingular on $\Theta_0$ and

$$\mathbb{E} \left[ \sup_{\theta \in \Theta_0} \| \partial g (Y, X, \theta) / \partial \theta \|_F \right] < \infty$$

where $\| \cdot \|_F$ denotes the Frobenius norm for matrices.

**Assumption 3.3** The following three expectations are finite:

(i) $\mathbb{E} \left[ \sup_{\lambda \in V_0} \rho'' \left( \lambda^T (u(X) - \mu) \right) \| u(X) - \mu \|^2 \right]$;

(ii) $\mathbb{E} \left[ \sup_{\lambda \in V_0} \rho' \left( \lambda^T (u(X) - \mu) \right) \sup_{\theta \in \Theta_0} \| \partial g (Y, X, \theta) / \partial \theta \|_F \right]$;

(iii) $\mathbb{E} \left[ \sup_{\lambda \in V_0} \rho'' \left( \lambda^T (u(X) - \mu) \right) \sup_{\theta \in \Theta} \| g(Y, X, \theta) \| \| u(X) - \mu \| \right]$.

**Definition 3.1.** The GEL based estimator $\{ \hat{\theta}_n, n \geq 1 \}$ of the true parameter $\theta_0 \in \Theta$ is a sequence of random vectors such that $\sum_{i=1}^n P_i g(y_i, x_i, \hat{\theta}_n) = 0$ holds.

The following theorem reveals the consistency of the estimator as defined above.

**Theorem 3.1.** Under assumptions 2.1-2.5 and 3.1-3.2, for the estimator $\{ \hat{\theta}_n, n \geq 1 \}$ as defined in Definition 3.1, these $\hat{\theta}_n$’s converge to the true parameter $\theta_0$ in probability.

**Proof.** By definition, $\hat{\theta}_n$ is the root of the equation $\sum_{i=1}^n P_i g(y_i, x_i, \hat{\theta}_n) = 0$ which is equivalent to say that

$$\sum_{i=1}^n \frac{\rho' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right)}{N h(x_i)} g(y_i, x_i, \hat{\theta}_n) = 0. \tag{3.1}$$

We next establish the uniform convergence of the following:

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \frac{\rho' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right)}{N h(x_i)} g(y_i, x_i, \theta) - \mathbb{E} \left[ g(Y, X, \theta) \right] \right\|_p \to 0. \tag{3.2}$$
Since $\mathbb{E} [g(Y, X, \theta)]$ has a unique root at the true parameter $\theta_0$ on $\Theta$, (3.2) is a sufficient condition for the consistency of estimator $\hat{\theta}_n$. To this end, we substitute (2.4) into (3.1), for each $\theta \in \Theta$,

$$
\sum_{i=1}^{n} \frac{\rho' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right)}{N h(x_i)} g(y_i, x_i, \theta) \\
= \sum_{i=1}^{n} \frac{-1}{N h(x_i)} g(y_i, x_i, \theta) + \sum_{i=1}^{n} \frac{\rho''(\phi_i)}{N h(x_i)} \lambda_n^T (u(x_i) - \mu) g(y_i, x_i, \theta) \\
+ \lambda_n^T (\mu - \bar{u}_N) \sum_{i=1}^{n} \frac{\rho''(\phi_i)}{N h(x_i)} g(y_i, x_i, \theta).
$$

(3.3)

Under Assumption 3.2, the first summand on the right hand side of (3.3) uniformly converges to $\mathbb{E} [g(Y, X, \theta)]$ on $\Theta$ in accordance with the uniform weak law of large number. In the proof of Main Theorem 2.1 in Section 2, we have shown that $\max_{i=1,\ldots,n} |\rho''(\phi_i)| \leq O_p(1)$, we can now use this fact to establish the order of convergence of the second summand in (3.3) as follows:

$$
\sup_{\theta \in \Theta} \left\| \sum_{i=1}^{n} \frac{\rho''(\phi_i)}{N h(x_i)} \lambda_n^T (u(x_i) - \mu) g(y_i, x_i, \theta) \right\| \\
\leq \sum_{i=1}^{n} \frac{|\rho''(\phi_i)|}{N h(x_i)} \left\| \lambda_n^T (u(x_i) - \mu) \right\| \sup_{\theta \in \Theta} \| g(y_i, x_i, \theta) \|
$$

$$
\leq \max_{i=1,\ldots,n} |\rho''(\phi_i)| \left\| \lambda_n \right\| \sum_{i=1}^{n} \frac{1}{N h(x_i)} \| u(x_i) - \mu \| \sup_{\theta \in \Theta} \| g(y_i, x_i, \theta) \|
$$

$$
= O_p(n^{-\frac{1}{2}}).
$$

The last equality comes from the assumption that $\mathbb{E} \left[ \sup_{\theta \in \Theta} \| g(Y, X, \theta) \|^2 \right]$ and $\mathbb{E} \left[ \| u(X) \|^2 \right]$ are both finite. Similar arguments could be used to deduce the uniform convergence of the last summand. Therefore, (3.2) holds which also implies the consistency of $\hat{\theta}_n$ to $\theta_0$. \hfill \square

Our next goal is to establish the asymptotic normality of the estimator $\hat{\theta}_n$. Before that, it is helpful to construct a representation for $\lambda_n$. We shall use $\otimes^2$ to denote the tensor product of vectors. For instance,

$$(u(x_i) - \mu) \otimes^2 = (u(x_i) - \mu) (u(x_i) - \mu)^T.$$
**Lemma 3.1.** Under assumptions 2.1-2.5, for each \( n \), \( \hat{\lambda}_n \) could be expressed as follows:

\[
\hat{\lambda}_n = -E\left[(u(X) - \mu)^2\right]^{-1} \frac{1}{N} \sum_{i=1}^{n} \frac{u(x_i) - \bar{u}_N}{h(x_i)} + o_p(n^{-\frac{1}{2}}).
\]

**Proof.** In accordance with Lemma 2.3, with probability approaching to one, \( \hat{\lambda}_n \in V_0 \) satisfies that first order condition \( \nabla f_n(\hat{\lambda}_n) = 0 \), i.e.

\[
0 = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right) (u(x_i) - \bar{u}_N)
\]

\[
= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda_n^T (u(x_i) - \mu) \right) (u(x_i) - \bar{u}_N)
\]

\[
+ \lambda_n^T (\mu - \bar{u}_N) \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' (\xi_i) (u(x_i) - \bar{u}_N).
\]

The second equality comes from Equation (2.3) and for each \( i = 1, \ldots, n \), \( \xi_i \) lies between \( \lambda_n^T (u(x_i) - \mu) \) and \( \lambda_n^T (u(x_i) - \bar{u}_N) \). From the parts (b) and (c) of the proof of Lemma 2.4, it is already known that

\[
\left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' (\xi_i) (u(x_i) - \bar{u}_N) \right\|
\]

\[
\leq \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' (\xi_i) (u(x_i) - \mu) \right\| + \left\| \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' (\xi_i) (\mu - \bar{u}_N) \right\|
\]

\[
= O_p(1).
\]

Therefore the first order condition for \( \hat{\lambda}_n \) is simplified as follows:

\[
0 = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda_n^T (u(x_i) - \mu) \right) (u(x_i) - \bar{u}_N) + o_p(n^{-\frac{1}{2}}).
\]

By using Taylor series expansion again for each \( \rho' \left( \lambda_n^T (u(x_i) - \mu) \right) \) at zero,
\[ i = 1, \ldots, n, \text{ there exists } \hat{\lambda}_i \text{ on the line joining zero and } \hat{\lambda}_n \text{ such that} \]
\[ 0 = O_p(n^{-1}) + \frac{1}{N} \sum_{i=1}^{n} \frac{-1}{h(x_i)} (u(x_i) - \bar{u}_N) \]
\[ + \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' \left( \hat{\lambda}_i^T (u(x_i) - \mu) \right) \hat{\lambda}_n^T (u(x_i) - \mu) (u(x_i) - \bar{u}_N) \]
\[ = O_p(n^{-1}) + \frac{1}{N} \sum_{i=1}^{n} \frac{-1}{h(x_i)} (u(x_i) - \bar{u}_N) \]
\[ + \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' \left( \hat{\lambda}_i^T (u(x_i) - \mu) \right) (u(x_i) - \mu)^{\otimes 2} \right) \hat{\lambda}_n. \]

Next we are going to prove the convergence of following sequence of semi-positive definite matrices:

\[ M_n \triangleq \sum_{i=1}^{n} \frac{-1}{N h(x_i)} \rho'' \left( \hat{\lambda}_i^T (u(x_i) - \mu) \right) (u(x_i) - \mu)^{\otimes 2}. \]

By Assumption 3.2, \( \| \mathbb{E}[M_{Y_0}] \|_F \) is finite where

\[ M_{Y_0} \triangleq \sup_{\lambda \in \Lambda_0} \rho'' \left( \lambda^T (u(X) - \mu) \right) (u(X) - \mu)^{\otimes 2}. \]

For every small enough \( \varepsilon > 0 \) and \( \delta > 0 \), since \( \max_{i=1,\ldots,n} \| \hat{\lambda}_i \| = O_p(n^{-\frac{1}{2}}) \), there exists a neighborhood of zero \( V_{\varepsilon, \delta} \subseteq \mathcal{V}_0 \) such that, with probability approaching to one,

\[ \limsup_{n \to \infty} P \left( \| M_n - \mathbb{E} [M_{Y_{\varepsilon, \delta}}] \|_F > \delta \right) < \varepsilon, \]

and

\[ \| \mathbb{E} [M_{Y_{\varepsilon, \delta}}] - \mathbb{E} [(u(X) - \mu)^{\otimes 2}] \|_F < \delta, \]

due to the dominated convergence theorem. Therefore, by applying triangle inequality,

\[ \limsup_{n \to \infty} P \left( \| M_n - \mathbb{E} [(u(X) - \mu)^{\otimes 2}] \|_F > \delta \right) \]
\[ \leq \limsup_{n \to \infty} P \left( \| M_n - \mathbb{E} [M_{Y_{\varepsilon, \delta}}] \|_F > \delta \right) \]
\[ + \limsup_{n \to \infty} P \left( \| \mathbb{E} [M_{Y_{\varepsilon, \delta}}] - \mathbb{E} [(u(X) - \mu)^{\otimes 2}] \|_F > \delta \right) \]
\[ \leq \varepsilon. \]
Under Assumption 2.2, $\mathbb{E} \left[ (u(X) - \mu)^{\otimes 2} \right] = \text{Var} \left[ u(X) \right]$ is nonsingular, therefore, with probability approaching to one,

$$M_n^{-1} = \mathbb{E} \left[ (u(X) - \mu)^{\otimes 2} \right]^{-1} + o_p(1)$$

and $\tilde{\lambda}_n$ admits the form

$$\sqrt{n} \tilde{\lambda}_n = -\mathbb{E} \left[ (u(X) - \mu)^{\otimes 2} \right]^{-1} \left( \frac{\sqrt{n} \sum_{i=1}^{n} u(x_i) - u_N}{h(x_i)} \right) + o_p(1).$$

Our proof of Lemma 3.1 reveals how one can deduce the convergence of a sequence of matrices $(M_n)$; under Assumption 3.3, similar method can also be applied to establish the convergence of the next two sequences of matrices, with probability approaching to one:

1. For any vector $\theta_i$ located on the line joining $\tilde{\theta}_n$ and $\theta_0$,

$$G_n(\theta_i) \triangleq \frac{1}{N} \sum_{i=1}^{n} \frac{\rho' \left( \tilde{\lambda}_n^T (u(x_i) - \mu) \right)}{h(x_i)} (\partial g(y_i, x_i, \theta_i) / \partial \theta)$$

$$= -\mathbb{E} \left[ \partial g(Y, X, \theta_0) / \partial \theta \right] + o_p(1).$$

2. For any vector $\lambda_i$ located on the line joining $\tilde{\lambda}_n$ and zero,

$$Q_n(\lambda_i) \triangleq \frac{1}{N} \sum_{i=1}^{n} \frac{\rho'' \left( \tilde{\lambda}_n^T (u(x_i) - \mu) \right)}{h(x_i)} g(y_i, x_i, \theta_0) (u(x_i) - \mu)^T$$

$$= -\mathbb{E} \left[ g(Y, X, \theta_0) (u(X) - \mu)^T \right] + o_p(1).$$

Our next theorem reveals the asymptotic normality of estimators $\tilde{\theta}_n$. Denote

$$U \triangleq \mathbb{E} \left[ g(Y, X, \theta_0) (u(X) - \mu)^T \right] \mathbb{E} \left[ (u(X) - \mu)^{\otimes 2} \right]^{-1}$$

and

$$\Sigma \triangleq \mathbb{E} \left[ g(Y, X, \theta_0)^{\otimes 2} + \frac{1 - h(X)}{h(X)} (g(Y, X, \theta_0) - U (u(X) - \mu))^{\otimes 2} \right].$$

THEOREM 3.2. Under assumptions 2.1-2.5 and 3.1-3.3,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \sim N(0, \Sigma),$$
where
\[
\Sigma = \mathbb{E} \left[ \frac{\partial g(Y, X, \theta_0)}{\partial \theta} \right]^{-1} \mathbb{E} \left( \frac{\partial g(Y, X, \theta_0)}{\partial \theta} \right)^T
\]

Proof. Since $\hat{\theta}_n$ is the root of Equation (3.1), Equation (3.3) indicates that there exists $\xi_i$ lying between $\lambda^T_n (u(x_i) - \bar{u}_N)$ and $\tilde{\lambda}^T_n (u(x_i) - \mu)$ such that

\[
0 = \frac{1}{N} \sum_{i=1}^{n} \frac{g(y_i, x_i, \hat{\theta}_n)}{h(x_i)} \rho' \left( \lambda^T_n (u(x_i) - \bar{u}_N) \right)
= \frac{1}{N} \sum_{i=1}^{n} \frac{g(y_i, x_i, \hat{\theta}_n)}{h(x_i)} \left( \rho' \left( \lambda^T_n (u(x_i) - \mu) \right) + \rho''(\xi_i) \tilde{\lambda}^T_n (\mu - \bar{u}_N) \right)
= \frac{1}{N} \sum_{i=1}^{n} \frac{g(y_i, x_i, \hat{\theta}_n)}{h(x_i)} \rho' \left( \lambda^T_n (u(x_i) - \mu) \right) + O_p(n^{-1}).
\]

By applying Taylor series expansion to $\rho' \left( \lambda^T_n (u(x_i) - \mu) \right)$ at zero and $g(y_i, x_i, \hat{\theta}_n)$ at $\theta_0$ respectively, there exists $\bar{\theta}_i$ located on the line joining $\hat{\theta}_n$ and $\theta_0$, $\tilde{\lambda}_i$ lying in the line joining $\lambda_n$ and zero such that

\[
0 = O_p(n^{-1}) + \frac{1}{N} \sum_{i=1}^{n} \frac{-1}{h(x_i)} g(y_i, x_i, \theta_0)
+ \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' \left( \lambda^T_n (u(x_i) - \mu) \right) g(y_i, x_i, \theta_0) (u(x_i) - \mu)^T \tilde{\lambda}_n
+ \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho' \left( \lambda^T_n (u(x_i) - \mu) \right) \left( \frac{\partial g(y_i, x_i, \bar{\theta}_i)}{\partial \theta} \right) (\hat{\theta}_n - \theta_0).
\]

It follows that with probability approaching to one

\[
(3.4) \quad G_n(\bar{\theta}_i) \left( \sqrt{n}(\hat{\theta}_n - \theta_0) \right)
= O_p(n^{-\frac{1}{2}}) + \frac{\sqrt{n}}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} g(y_i, x_i, \theta_0) - \sqrt{n} Q_n(\tilde{\lambda}_i) \tilde{\lambda}_n
= o_n(1) + \frac{\sqrt{n}}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} g(y_i, x_i, \theta_0) - \frac{\sqrt{n}}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} (u(x_i) - \bar{u}_N)
= o_n(1) + \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left( \frac{R_i}{h(x_i)} g(y_i, x_i, \theta_0) - U \left( \frac{R_i}{h(x_i)} - 1 \right) (u(x_i) - \mu) \right),
\]
where
\[
U \triangleq \mathbb{E} \left[ g(Y, X, \theta_0)(u(X) - \mu)^T \right] \mathbb{E} \left[ (u(X) - \mu)^{\otimes 2} \right]^{-1}.
\]

Since all summands are i.i.d., the asymptotic normality of $\hat{\theta}_n$ could be deduced by applying the central limit theorem to the right hand side of the last equality in Equation (3.4). Given $Y$ and $X$, $g(Y, X, \theta_0)$, $u(X)$ and $h(X)$ are known and at the same time, the indicator function $R$ follows Bernoulli distribution with $\mathbb{E}[R|Y, X] = h(X)$ and $\text{Var}(R|Y, X) = h(X)(1 - h(X))$. It follows that
\[
\begin{align*}
\mathbb{E} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(X) - \mu) | Y, X \right] & = \frac{\mathbb{E}[R|Y, X]}{h(X)} g(Y, X, \theta_0) - U \left( \frac{\mathbb{E}[R|Y, X]}{h(X)} - 1 \right) (u(X) - \mu) \\
& = g(Y, X, \theta_0)
\end{align*}
\]
and
\[
\begin{align*}
\text{Var} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(X) - \mu) | Y, X \right] & = \text{Var} \left[ \frac{R}{h(X)} (g(Y, X, \theta_0) - U (u(X) - \mu)) | Y, X \right] \\
& = \frac{1}{h(X)^2} \text{Var}[R|Y, X] (g(Y, X, \theta_0) - U (u(X) - \mu))^{\otimes 2} \\
& = \frac{1 - h(X)}{h(X)} (g(Y, X, \theta_0) - U (u(X) - \mu))^{\otimes 2}.
\end{align*}
\]
Therefore,
\[
\begin{align*}
\mathbb{E} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(X) - \mu) \right] & = \mathbb{E} \left[ \mathbb{E} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(X) - \mu) | Y, X \right] \right] \\
& = \mathbb{E}[g(Y, X, \theta_0)] = 0
\end{align*}
\]
and
\[
\text{Var} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(x) - \mu) \right]
\]
\[
= \text{Var} \left[ E \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(x) - \mu) | Y, X \right] \right]
\]
\[
+ E \left[ \text{Var} \left[ \frac{R}{h(X)} g(Y, X, \theta_0) - U \left( \frac{R}{h(X)} - 1 \right) (u(x) - \mu) | Y, X \right] \right]
\]
\[
= \text{Var} \left[ g(Y, X, \theta_0) \right] + E \left[ \frac{1 - h(X)}{h(X)} \left( g(Y, X, \theta_0) - U (u(x) - \mu) \right)^2 \right]
\]
\[
= \Sigma.
\]

Finally, by denoting
\[
\Sigma = E \left[ \frac{\partial g(Y, X, \theta_0)}{\partial \theta} \right]^{-1} \Sigma \left( E \left[ \frac{\partial g(Y, X, \theta_0)}{\partial \theta} \right]^{-1} \right)^T,
\]

Equation (3.4) implies
\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \sim N(0, \Sigma).
\]

4. Simulation studies. We studied the rate of convergence of the GEL weights and the performance of the corresponding weighted estimator by simulation studies. We considered two scenarios. In Scenario I, we were interested in estimating the mean of a continuous variable Y which was only observed in a subsample, while an auxiliary variable X was observed in full. In Scenario II, we were interested in estimating the regression coefficient \( \beta_1 \) of a linear model \( E(Y|X) = \beta_0 + \beta_1 X \), in which Y was observed for the full sample but covariate X was observed only in a subsample. In both scenarios, X was generated from (a) standard normal distribution, (b) t-distribution with 3 degrees of freedom, (c) t-distribution with 2 degrees of freedom and (d) Cauchy distribution. We generate \( Y = -2 - X + \epsilon \) where \( \epsilon \) is generated from the standard normal distribution independent of Y and X. Sample sizes being considered were \( N = [10^j]^2 \), \( j = 5, 6, 7, 8 \) and \([x]\) denotes the integer part of x. In all cases, we generated the selection indicator R from a Bernoulli distribution with \( p = 0.25 \) and 1000 independent data sets were generated in each case. We examined the results for GEL estimators with different functions \( \rho(v) \), but the results for different choices of \( \rho \) are very similar as indicated by the theoretical properties. For clarity, we only report results for \( \rho \) being an exponential function, \( \rho(v) = -\exp(v) \).
To study the rate of convergence of the GEL weights, we calculated the maximal scaled difference \( \max_{i=1,...,n} |N \hat{P}_i - 1| \) for each simulated data set and reported the 80th, 90th and 95th percentile \( a_p, p = 0.8, 0.9, 0.95 \) for different sample sizes. When \( N \) is large,
\[
\log a_p \approx c + \gamma \log N \\
(4.1)
\]
where \( c \) is an arbitrary constant and \( \gamma = -(1/2 - 1/\alpha) \). Table 1 displays \( a_{0.8}, a_{0.9}, a_{0.95} \) and \( \gamma \) for different sample sizes. The empirical \( \gamma \) was found by ordinary least square estimates determined from (4.1) for data points corresponding to different sample sizes. The simulation results are shown in Table 1.

(a) When \( X \) was generated from the standard normal distribution, the rates of convergence of the scaled maximal difference were quite close to the theoretical value of \( n^{-0.5} \), which corresponds to \( \alpha \) being infinite. (b) When \( X \) was generated from a \( t \)-distribution with 3 degrees of freedom, \( E(|X|^\alpha) < \infty \) for \( \epsilon > 0 \) but \( E(|X|^\alpha) \) is infinite. The rate of convergence of the scaled maximal difference for the empirical study was close to \( n^{-1/2-1/3} \approx n^{-0.167} \).

(c) When \( X \) was generated from a \( t \)-distribution with 2 degrees of freedom, \( E(|X|^\alpha) < \infty \) for \( \epsilon > 0 \) but \( E(|X|^\alpha) \) is infinite. Strictly speaking, the conditions for the validity of Theorem 1 did not hold since \( \alpha = 2 - \epsilon \), but in any finite sample the data would behave as if \( \alpha = 2 \). The results show that the rate of convergence of the scaled maximal difference is close to \( n^0 = 1 \).

(d) When \( X \) was generated from a Cauchy distribution, the scaled maximal difference appeared to be unbounded.
Table 2

Comparison between sample mean and weighted mean. Sample sizes were \( N = [10^{j/2}] \), \( j=5,6,7,8 \). \( X \) was generated from (a) standard normal distribution, (b) \( t \)-distribution with 3 degrees of freedom, (c) \( t \)-distribution with 2 degrees of freedom and (d) Cauchy distribution. SSE represents sampling standard error.

<table>
<thead>
<tr>
<th></th>
<th>( j=5 )</th>
<th>( j=6 )</th>
<th>( j=7 )</th>
<th>( j=8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Sample mean</td>
<td>0.008</td>
<td>0.163</td>
<td>&lt;0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Weighted mean</td>
<td>&lt;0.001</td>
<td>0.131</td>
<td>-0.001</td>
<td>0.072</td>
</tr>
<tr>
<td>(b) Sample mean</td>
<td>-0.006</td>
<td>0.217</td>
<td>0.001</td>
<td>0.127</td>
</tr>
<tr>
<td>Weighted mean</td>
<td>-0.004</td>
<td>0.150</td>
<td>0.003</td>
<td>0.083</td>
</tr>
<tr>
<td>(c) Sample mean</td>
<td>0.016</td>
<td>0.417</td>
<td>0.008</td>
<td>0.275</td>
</tr>
<tr>
<td>Weighted mean</td>
<td>-0.001</td>
<td>0.245</td>
<td>0.003</td>
<td>0.137</td>
</tr>
<tr>
<td>(d) Sample mean</td>
<td>1.242</td>
<td>49.01</td>
<td>2.487</td>
<td>53.81</td>
</tr>
<tr>
<td>Weighted mean</td>
<td>0.383</td>
<td>12.95</td>
<td>0.792</td>
<td>16.19</td>
</tr>
</tbody>
</table>

Table 3

Comparison between ordinary least square (OLS) and weighted least square (WLS) estimates. Sample sizes were \( N = [10^{j/2}] \), \( j=5,6,7,8 \). \( X \) was generated from (a) standard normal distribution, (b) \( t \)-distribution with 3 degrees of freedom, (c) \( t \)-distribution with 2 degrees of freedom and (d) Cauchy distribution. SSE represents sampling standard error.

<table>
<thead>
<tr>
<th></th>
<th>( j=5 )</th>
<th>( j=6 )</th>
<th>( j=7 )</th>
<th>( j=8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) OLS</td>
<td>0.039</td>
<td>0.574</td>
<td>0.020</td>
<td>0.262</td>
</tr>
<tr>
<td>WLS</td>
<td>0.032</td>
<td>0.537</td>
<td>0.019</td>
<td>0.256</td>
</tr>
<tr>
<td>(b) OLS</td>
<td>0.018</td>
<td>0.754</td>
<td>0.010</td>
<td>0.410</td>
</tr>
<tr>
<td>WLS</td>
<td>0.019</td>
<td>0.738</td>
<td>0.010</td>
<td>0.408</td>
</tr>
<tr>
<td>(c) OLS</td>
<td>-0.059</td>
<td>1.440</td>
<td>0.017</td>
<td>0.641</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.057</td>
<td>1.123</td>
<td>0.017</td>
<td>0.640</td>
</tr>
<tr>
<td>(d) OLS</td>
<td>-0.002</td>
<td>0.278</td>
<td>0.001</td>
<td>0.081</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.002</td>
<td>0.166</td>
<td>0.001</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Table 2 shows the performance of the weighted mean in Scenario I and Table 3 shows the performance of the weighted least square estimator in Scenario II. In all situations when \( \alpha \geq 2 \), the weighted estimators was unbiased and had a smaller sampling error than the unweighted estimators from the reduced sample, as indicated by Theorem 2. Although our current results do not apply to situations where \( \alpha \leq 1 \), we found a rather surprising results that the bias and sampling error can still be reduced substantially by using the GEL weights.

5. Conclusion. In this paper we presented theoretical properties of GEL weighted estimation for incorporating information of auxiliary data. We showed that the rate of convergence of the GEL weights is sharper in general than the general theory of GEL suggested (Newey and Smith, 2004). Also, we showed the theoretical properties for the critical case \( \alpha = 2 \) whereas Newey and Smith (2004) only considered \( \alpha > 2 \). Question remains whether
the GEL weights are applicable when \( \alpha < 2 \). While simulation shows that the maximal scaled difference of weights is possibly unbounded, there are also encouraging signs that the weighted estimator may still be substantially improved over the unweighted estimators. Further theoretical investigation is needed to explain such a phenomenon.

6. Appendix: An alternative and simpler proof when \( \alpha > 2 \). In Assumption 2.1, \( \alpha \) is assumed to be larger than or equal to 2; in particular, our work still holds when \( \alpha \) exactly equals to 2. If one restricts the range of \( \alpha \) to be \((2, \infty)\), due to the higher order of uniform boundedness of the random sample \( X_i \)'s, the proof of convergence of \( \hat{\lambda}_n \) could be simplified. Fix \( \frac{1}{\sqrt{\alpha}} < \xi < \frac{1}{2} \), denote

\[
\Lambda_n \equiv \{ \lambda : \| \lambda \| \leq n^{-\xi} \}
\]

and

\[
\lambda_n \equiv \arg \max_{\lambda \in \Lambda_n} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho \left( \lambda^T (u(x_i) - \bar{u}_N) \right).
\]

Since \( \Lambda_n \) is compact for each \( n \), \( \lambda_n \in \Lambda_n \) is well-defined. We want to achieve the same result in Lemma 2.3 by proving that \( \lambda_n \) must be an interior point of \( \Lambda_n \).

**Lemma 6.1.** Under assumptions 2.1-2.3 and 2.5 with \( \alpha > 2 \), \( \hat{\lambda}_n = O_p(n^{-\frac{1}{2}}) \) with \( \hat{\lambda}_n \in \Lambda_n \) and \( \| \hat{\lambda}_n \| = \inf_{\lambda \in \Lambda_n} \| \lambda \| \).

**Proof.** From Lemma 2.1,

\[
\max_{i=1,\ldots,n} \| u(x_i) - \bar{u}_N \| = O_p(n^{-\frac{1}{2}})
\]

and then

\[
\max_{i=1,\ldots,n} \left| \lambda_n^T (u(x_i) - \bar{u}_N) \right| \leq \| \lambda_n \| \max_{i=1,\ldots,n} \| u(x_i) - \bar{u}_N \| = O_p(n^{\frac{1}{2} - \xi})
\]

which converges to zero in probability. It follows that for any \( \lambda_l \) located on the line joining zero and \( \lambda_n \), since \( \rho'' \) is continuous and \( \rho''(0) = -1 \), we have

\[
\lim_{n \to \infty} \mathbb{P} \left( \max_{i=1,\ldots,n} \rho'' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right) \leq -\frac{1}{2} \right) = 1.
\]

On the other hand, by Assumption 2.2, \( \text{Var} \left| u(X) \right| \) is nonsingular and therefore it is a positive definite matrix. It implies that the smallest eigenvalue
of \( \frac{1}{N} \sum_{i=1}^{n} (u(x_i) - \bar{u}_N)^{\otimes 2} / h(x_i) \) is bounded away from zero with probability approaching to one because of the following equality

\[
\frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} (u(x_i) - \bar{u}_N)^{\otimes 2} \\
= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} u(x_i)^{\otimes 2} - \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} u(x_i) \right) \bar{u}_N^{T} \\
- \bar{u}_N \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} u(x_i) \right)^{T} + \bar{u}_N^{\otimes 2} \left( \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \right) \\
= \text{Var}[u(X)] + O_p(n^{-\frac{1}{2}}).
\]

Thus, there exists a positive constant \( C \) so that

\[
\lim_{n \to \infty} \mathbb{P} \left( C \| \lambda_n \|^2 \leq \lambda_n^T \left( \frac{1}{4N} \sum_{i=1}^{n} \frac{1}{h(x_i)} (u(x_i) - \bar{u}_N)^{\otimes 2} \right) \lambda_n \right) = 1.
\]

In accordance with the definition of \( \lambda_{n,1} \) by using Taylor series expansion around zero for each term, there exists \( \lambda_i \) lying in the line joining zero and \( \lambda_n \) such that

\[
\tilde{f}_n (0) \leq \tilde{f}_n (\lambda_n) \\
= \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho (\lambda_n^T (u(x_i) - \bar{u}_N)) \\
= \tilde{f}_n (0) - \frac{1}{N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \lambda_n^T (u(x_i) - \bar{u}_N) \\
+ \frac{1}{2N} \sum_{i=1}^{n} \frac{1}{h(x_i)} \rho'' \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right) \left( \lambda_n^T (u(x_i) - \bar{u}_N) \right)^2.
\]

Therefore, with probability approaching to one, we could deduce the follow-
ing inequality from the previous one:

\[ C \| \lambda_n \|^2 \leq -\lambda_n^T \sum_{i=1}^{n} \frac{1}{N h(x_i)} (u(x_i) - \bar{u}_N) \]

\[ \leq \| \lambda_n \| \left| \sum_{i=1}^{n} \frac{1}{N h(x_i)} (u(x_i) - \bar{u}_N) \right| \]

\[ \leq \| \lambda_n \| \left( \sum_{i=1}^{n} \frac{1}{N h(x_i)} (u(x_i) - \mu) \right) + \left| \sum_{i=1}^{n} \frac{1}{N h(x_i)} \right| \]

\[ = \| \lambda_n \| O_p \left( n^{-\frac{1}{2}} \right). \]

That is to say, \( \| \lambda_n \| = O_p(n^{-\frac{1}{2}}) = o_p(n^{-\xi}) \) with probability approaching to one. Moreover, \( \lambda_n \) is an interior point of \( \Lambda_n \) and the first order condition \( \nabla \tilde{f}_n(\lambda_n) = 0 \) is satisfied with probability approaching to one. In view of concavity of \( \tilde{f}_n(\lambda) \), we could deduce that \( \lambda_n \in \Lambda_n \cap \Lambda^0 \neq \emptyset \) with probability approaching to one and therefore \( \lambda_n = O_p(n^{-\frac{1}{2}}) \) with \( \lambda_n \in \Lambda_n \) such that \( \| \lambda_n \| = \inf_{\lambda \in \Lambda_n} \| \lambda \| \). \qed

References.


