GLS UNDER MONOTONE HETEROSKEDASTICITY

YOICHI ARAI, TAISUKE OTSU, AND MENGSHAN XU

Abstract. The generalized least square (GLS) is one of the most basic tools in regression analyses. A major issue in implementing the GLS is estimation of the conditional variance function of the error term, which typically requires a restrictive functional form assumption for parametric estimation or smoothing parameters for nonparametric estimation. In this paper, we propose an alternative approach to estimate the conditional variance function under nonparametric monotonicity constraints by utilizing the isotonic regression method. Our GLS estimator is shown to be asymptotically equivalent to the infeasible GLS estimator with knowledge of the conditional error variance, and involves only some tuning to trim boundary observations, not only for point estimation but also for interval estimation or hypothesis testing. Our analysis extends the scope of the isotonic regression method by showing that the isotonic estimates, possibly with generated variables, can be employed as first stage estimates to be plugged in for semiparametric objects. Simulation studies illustrate excellent finite sample performances of the proposed method. As an empirical example, we revisit Acemoglu and Restrepo's (2017) study on the relationship between an aging population and economic growth to illustrate how our GLS estimator effectively reduces estimation errors.

1. Introduction

The generalized least square (GLS) is one of the most basic tools in regression analyses. It yields the best linear unbiased estimator in the classical linear regression model, and has been studied extensively in econometrics and statistics literature; see e.g., Wooldridge (2010, Chapter 7) for a review. A major issue in implementing the GLS is that the optimal weights given by the conditional error variance function (say, $\sigma^2(\cdot)$) are typically unknown to researchers and need to be estimated. One way to estimate $\sigma^2(\cdot)$ is to specify its parametric functional form and estimate it by a parametric regression for the squared OLS residuals of the original regression on the specified covariates. However, economic theory rarely provides exact functional forms of $\sigma^2(\cdot)$, and the feasible GLS using misspecified $\sigma^2(\cdot)$ is no longer asymptotically efficient (Cragg, 1983). To address this issue, Carroll (1982) and Robinson (1987) proposed to estimate $\sigma^2(\cdot)$ nonparametrically and established the asymptotic equivalence of the resulting feasible GLS estimator with the infeasible one under certain regularity conditions. This is a remarkable result, but it requires theoretically and practically judicious choices of smoothing parameters, such as bandwidths, series lengths, or numbers of neighbors. It should be noted that such smoothing parameters appear in not only the point estimator but also its standard error for inference, and their choices typically require some assumption or knowledge of the smoothness of the conditional variance and associated density functions, such as their differentiability orders.

In this paper, we propose an alternative approach to estimate the conditional error variance function to implement the GLS by exploring a shape constraint of $\sigma^2(\cdot)$ instead of its smoothness as in Robinson (1987). As argued by Matzkin (1994), economic theory often provides shape constraints for functions of economic variables, such as monotonicity, concavity, or symmetry. In particular, we focus on situations where $\sigma^2(\cdot)$ is known to be monotone in its argument even though its exact functional form is unspecified, and propose to estimate $\sigma^2(\cdot)$ by utilizing the method of isotonic regression (see a review by Groeneboom and Jongbloed, 2014). It is known that the conventional isotonic regression estimator typically yields piecewise constant function estimates and does not involve any tuning parameters. Although the limiting behavior of the isotonic regression estimator is less tractable (such as the $n^{1/3}$-consistency and complicated limiting distribution), we show that our feasible GLS estimator using the optimal weights by the isotonic estimator with some trimming for boundary observations is asymptotically equivalent to the infeasible GLS estimator. Furthermore, we can plug in this isotonic estimator to estimate the asymptotic variance of the GLS estimator for statistical inference.

For the linear model $Y = X'\beta + U$ in the presence of heteroskedasticity $\sigma^2(X) = E[U^2|X]$, using feasible GLS to improve the estimation efficiency has a long history. On the one hand, several parametric models have been proposed to estimate conditional error variance function $\sigma^2(\cdot)$. See Remark 5 below. On the other hand, Carroll (1982) and Robinson (1987) estimated $\sigma^2(\cdot)$ with kernel and nearest neighbor estimator, respectively, and they showed their semiparametric GLS estimators are asymptotically equivalent to the infeasible GLS estimator and thus efficient. Compared to existing parametric methods, our proposed method imposes monotonicity, a feature implied by many parametric models, but it is nonparametric and does not rely on any specific parametric function form.[1] Compared to existing nonparametric methods, our proposed method involves only some tuning to trim boundary observations which does not require knowledge of the smoothness of the conditional variance and associated density functions. In the Monte Carlo simulations, we show that our proposed method outperforms the above-mentioned nonparametric methods at almost every choice of smoothing parameters, while it performs as well as parametric feasible GLS estimators with correctly specified conditional error variance function.

The isotonic estimator can date back to the middle of the last century. Earlier work includes Ayer *et al.* (1955), Grenander (1956), Rao (1969, 1970), and Barlow and Brunk (1972), among others. The isotonic estimator of a regression function can be formulated as a least square estimation with monotonicity constraints. Suppose that the conditional expectation $E[Y|X] = m(X)$ is monotone increasing, for an iid random sample $\{Y_i, X_i\}_{i=1}^n$, the isotonic estimator is the minimizer of the sum of squared errors, $\min_{m\in\mathcal{M}} \sum_{i=1}^n \{Y_i - m(X_i)\}^2$, where $\mathcal{M}$ is the class of monotone increasing functions. The minimizer can be calculated with the pool adjacent violators algorithm (Barlow and Brunk, 1972), or equivalently by solving the greatest convex minorant of the cumulative sum diagram $\{(0,0), (i, \sum_{j=1}^i Y_j), i = 1, \ldots, n\}$, where the corresponding $\{X_i\}_{i=1}^n$

[1] Monotone heteroskedasticity is often observed in economic literature. For example, Mincer (1974) argued that the variance of wages, when conditioned on education, should increase with the level of education because individuals with higher education have a broader array of job choices. Ruud (2000) cited this argument and provided empirical evidence in his Figure 18.1 based on the CPS data from March 1995. Another example can be found in Example 8.6 of Wooldridge (2013, pp. 283-284), where he employed a univariate conditional variance function of log income to explain the heteroskedasticity observed in net total financial wealth of people in the United States.

are ordered sequence; see Groeneboom and Jongbloed (2014) for a comprehensive discussion of different aspects of isotonic regression. Moreover, recent developments in the monotone single index model provide convenient and flexible tools for combining monotonicity and multi-dimensional covariates. In a monotone single index model, the conditional mean of Y is modeled as $E[Y|X] = m(X'\alpha)$, and the monotone link function $m(\cdot)$ is solved with isotonic regression. Balabdaoui, Durot and Jankowski (2019) studied the monotone single index model with the monotone least square method. Groeneboom and Hendrickx (2018), Balabdaoui, Groeneboom and Hendrickx (2019), and Balabdaoui and Groeneboom (2021) developed a score-type approach for the monotone single index model. Their approach can estimate the single index parameter α and the link function $m(\cdot)$ at $n^{-1/2}$-rate and $n^{-1/3}$-rate respectively. We employ their approach for the estimation of the conditional variance function in the multivariate case. Recently, Babii and Kumar (2023) applied the isotonic regression to their analysis of regression discontinuity designs. To this end, Babii and Kumar (2023) extended existing results concerning the boundary properties of Grenander's estimator (e.g., those from Woodroofe and Sun, 1993, and Kulikov and Lopuhaä, 2006) to derive the asymptotic distribution of their trimmed isotonic regression discontinuity estimator. To regularize the isotonic estimator in the weights of our proposed GLS estimator, we employ a similar trimming strategy while adapting the theory of Babii and Kumar (2023) to our context of the conditional variance estimation. We contribute to this literature on isotonic regression by showing that the isotonic estimates can be employed as first stage estimates to be plugged in for semiparametric objects. Furthermore, we note that our isotonic estimator involves generated variables (i.e., OLS residuals), which make theoretical developments substantially different from the existing ones.

This paper is organized as follows. In Section 2, we consider the case where $\sigma^2(\cdot)$ is monotone in one covariate, present our GLS estimator, and study its asymptotic properties. Section 3 extends our GLS approach to the case where $\sigma^2(\cdot)$ is specified by a monotone single index function. Section 4 illustrates the proposed method by a simulation study and empirical example.

2. Heteroskedasticity by univariate covariate

We first consider the case where monotone heteroskedasticity is caused by a single covariate. In particular, consider the following multiple linear regression model

$$Y = \alpha + \beta X + Z'\gamma + U, \qquad E[U|X, Z] = 0, \tag{2.1}$$

where $X \in \mathscr{X} = [x_L, x_U]$ is a scalar covariate with compact support and Z is a vector of other covariates. In this section, we focus on the case where heteroskedasticity is caused by the covariate X, i.e.,

$$E[U^2|X, Z] = E[U^2|X] =: \sigma^2(X), \tag{2.2}$$

and $\sigma^2(\cdot)$ is a monotone increasing function. The case of monotone decreasing $\sigma^2(\cdot)$ is analyzed analogously (by setting U^2 as $-U^2$). In the setup (2.2), we assume that the researcher knows which covariate should be included in $\sigma^2(\cdot)$ based on economic theory or other prior information. This setup should be considered as a useful benchmark to provide a clear exposition of the main concept and the asymptotic properties of the proposed monotone GLS estimator. Without the covariates Z, the above model covers a bivariate regression model, and our approach is new even in such a fundamental setup. Furthermore, this setup covers the case where X contained in (2.2) does not enter the regression model (2.1) by setting $\beta = 0$ (such a situation is considered in our empirical illustration in Section 4.2). Extensions to relax the assumption in (2.2) will be discussed in Remark 1 and Section 3.

Let $\theta = (\alpha, \beta, \gamma')'$ be a vector of the slope parameters and $W := (1, X, Z')'$ so that the model in (2.1) can be written as $Y = W'\theta + U$. Based on an iid sample $\{Y_i, X_i, Z_i\}_{i=1}^n$, the infeasible GLS estimator for θ is written as

$$\hat{\theta}_{\text{IGLS}} = \left(\sum_{i=1}^n \sigma_i^{-2} W_i W_i' \right)^{-1} \left(\sum_{i=1}^n \sigma_i^{-2} W_i Y_i \right), \tag{2.3}$$

where $\sigma_i^2 = \sigma^2(X_i)$. In order to make this estimator feasible, various approaches have been proposed in the literature.

In this paper, we are concerned with the situation where the researcher knows $\sigma^2(\cdot)$ is monotone in a particular regressor X but its exact functional form is unspecified. In particular, by utilizing knowledge of the monotonicity of $\sigma^2(\cdot)$, we propose to estimate $\sigma^2(\cdot)$ by the isotonic regression from the squared OLS residual on the regressor X. More precisely, let $\hat{\theta}_{\text{OLS}} = (\sum_{i=1}^n W_i W_i')^{-1} (\sum_{i=1}^n W_i Y_i)$ be the OLS estimator for (2.1), and $\hat{U}_j = Y_j - W_j'\hat{\theta}_{\text{OLS}}$ be its residual. Then we estimate $\sigma^2(\cdot)$ by

$$\hat{\sigma}^2(\cdot) = \text{isotonic regression function from } \{\hat{U}_j^2\}_{j=1}^n \text{ on } \{X_j\}_{j=1}^n. \tag{2.4}$$

Although this estimator is shown to be consistent for $\sigma^2(\cdot)$ in the interior of support $[x_L, x_U]$ of X, it is generally biased at the lower boundary x_L, which may cause inconsistency of the resulting GLS estimator. Therefore, we propose to trim observations whose X_i's are too close to x_L, and develop the following feasible GLS estimator

$$\hat{\theta} = \left(\sum_{i=1}^n \mathbb{I}\{X_i \geq q_n\} \hat{\sigma}_i^{-2} W_i W_i' \right)^{-1} \left(\sum_{i=1}^n \mathbb{I}\{X_i \geq q_n\} \hat{\sigma}_i^{-2} W_i Y_i \right), \tag{2.5}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, and the trimming term q_n is set as the $(n^{-1/3})$-th sample quantile of $\{X_i\}_{i=1}^n$.

Let $\mathcal{B}(a, R)$ be a ball around a with radius R; for $\varepsilon = U^2 - \sigma^2(X)$, define $\sigma_\varepsilon^2(x) = E[\varepsilon^2 | X = x]$. To study the asymptotic properties of the proposed estimator $\hat{\theta}$, we impose the following assumptions.

Assumption.

- **A1:** $\{Y_i, X_i, Z_i\}_{i=1}^n$ *is an iid sample of* (Y, X, Z). *The support of* (X, Z) *is convex with non-empty interiors and is a subset of* $\mathcal{B}(0, R)$ *for some* $R > 0$. *The support of* X *is a compact interval* $\mathscr{X} = [x_L, x_U]$.
- **A2:** $\sigma^2 : \mathscr{X} \to \mathbb{R}$ *is a monotone increasing function defined on* $\mathscr{X}$, *and* $0 < \sigma^2(x_L) < \sigma^2(x_U) < \infty$. *There exist positive constants* a_0 *and* M *such that* $E[|U|^{2s}|X = x] \leq a_0 s! M^{s-2}$ *for all integers* $s \geq 2$ *and* $x \in \mathscr{X}$. *For some positive constant* δ, $\sigma^2(\cdot)$ *is continuously differentiable on* $(x_L, x_L + \delta)$, *and* $\sigma_\varepsilon^2(\cdot)$ *is continuous on* $(x_L, x_L + \delta)$.

A3: *X has a continuous density function $f_X(\cdot)$ on $\mathscr{X}$, and there exists a positive constant b such that $b < f_X(x) < \infty$ for all $x \in \mathscr{X}$.*

Assumption A1 is standard. As pointed out in Balabdaoui, Groeneboom and Hendrickx (2019, p.13), the compact support assumption can be relaxed when X follows a sub-Gaussian distribution. In this case, the L^2-convergence rate of the isotonic estimator will decrease from $O_p(n^{-1/3}\log n)$ to $O_p(n^{-1/3}(\log n)^{5/4})$. Another impact of relaxing the distribution of X (and Z) to a sub-Gaussian one is on the concentration rate of $\max_j |\hat{U}_j^2 - U_j^2|$ (see Appendix A for more details). This rate, used in proving Lemma 1 and explaining the concentration of T_1 and T_2 in Appendix A.2, will inflate by a factor of $\log n$. However, even with this change, we still have $\max_j |\hat{U}_j^2 - U_j^2| = o_p(n^{-1/3})$, which is the key to show that the impact of substituting infeasible U^2 with estimated $\hat{U}^2$ on isotonic estimators is asymptotically negligible. Considering that the convergence rates of these aforementioned terms are slowed down by a factor of $\log n$ at most, the validity of the main results in this paper is preserved with sub-Gaussian covariates, but the analytical derivation would become more cumbersome. For a clearer and more concise exposition, we maintain the compact support assumption on X. Assumption A2 is on the error term. The monotonicity of $\sigma^2(\cdot)$ is the main assumption. The assumption on arbitrary higher moments, which rules out some fat-tailed distributions, is commonly used to obtain some maximal inequalities (cf. van der Vaart and Wellner, 1996, Lemma 2.2.11, for a similar assumption). Assumption A3 contains additional mild conditions on the density of X.

We first present asymptotic properties of the conditional error variance estimator $\hat{\sigma}^2(\cdot)$ in (2.4). Let q_n^* be the $(n^{-1/3})$-th population quantile of X, $D_A^L[f](a)$ be the left derivative of the greatest convex minorant of a function $f(\cdot)$ evaluated at $a \in A$, and $\{\mathcal{W}_t\}$ be the standard Brownian motion. Also define $c^* = \lim_{n\to\infty} n^{1/3}(q_n^* - x_L)$. Assumption A3 guarantees $0 < c^* < \infty$. Then we obtain the following lemma for the behavior of $\hat{\sigma}^2(\cdot)$ around the boundary x_L, which extends the result by Babii and Kumar (2023, Theorem 2.1(ii)) by allowing the generated variable $\hat{U}_i^2$ as a regressand for $\hat{\sigma}^2(\cdot)$.

Lemma 1. *Under Assumptions A1-A3 and $\lim_{x\downarrow x_L} \frac{d\sigma^2(x)}{dx} > 0$, it holds*

$$n^{1/3}\{\hat{\sigma}^2(q_n) - \sigma^2(q_n)\} \xrightarrow{d} D^L_{[0,\infty)}\left[\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \left(\lim_{x\downarrow x_L}\frac{d\sigma^2(x)}{dx}\right)c^*\left(\frac{1}{2}t^2 - t\right)\right](1). \quad (2.6)$$

Based on this lemma, the asymptotic distribution of our feasible GLS estimator $\hat{\theta}$ is obtained as follows.

Theorem 1. *Under Assumptions A1-A3, it holds*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, E[\sigma^{-2}(X)WW']^{-1}),$$

and the asymptotic variance matrix is consistently estimated by $\left(\frac{1}{n}\sum_{i=1}^n \hat{\sigma}_i^{-2} W_i W_i'\right)^{-1}$.

This theorem implies that our estimator $\hat{\theta}$ has the same limiting distribution as the infeasible GLS estimator $\hat{\theta}_{\text{IGLS}}$ and thus achieves the semiparametric efficiency bound. This result extends the scope of the isotonic regression method by showing that the isotonic estimates, possibly with generated variables, can be employed as first stage estimates to be plugged in for semiparametric objects. We re-emphasize that $\hat{\theta}$ involves only a trimming term q_n, the $(n^{-1/3})$-th sample quantile of $\{X_i\}_{i=1}^n$.[2]

Remark 1. [Extensions of (2.2)] The benchmark setup $E[U^2|X,Z] = \sigma^2(X)$ considered in this section can be extended in various ways. First, an extension to a single index model (say, $E[U^2|X,Z] = \sigma^2(X\eta_x + Z'\eta_z)$) will be discussed in the next section. Second, the model in (2.1)-(2.2) can be extended to the case where the conditional variance varies with discrete covariates Z (or its subvector), say $E[U^2|X,Z=z] = \sigma_z^2(X)$ with monotone functions $\sigma_z^2(\cdot)$ for $z \in \{z^{(1)}, \dots, z^{(D)}\}$. In this case, we can implement the isotonic regression for each group categorized by z, and construct the feasible GLS estimator in an analogous way as (2.5). Third, our approach may be extended to the additive monotone heteroskedasticity, say $E[U^2|X,Z] = \sigma_x^2(X)+\sigma_z^2(Z)$ with monotone functions $\sigma_x^2(\cdot)$ and $\sigma_z^2(\cdot)$. Although formal analysis is beyond the scope of this paper, the results in Mammen and Yu (2007) suggest that the isotonic estimators for additive functions converge at similar rates as the univariate case, and we conjecture that a similar result as Theorem 1 can be obtained. Finally, when the conditional error variance function is multiplicative, say $E[U^2|X,Z] = \sigma_x^2(X)\sigma_z^2(Z)$, and the researcher knows the form of $\sigma_z^2(\cdot)$ (e.g., Z is household size and $\sigma_z^2(Z) = Z^2$), then our feasible GLS estimator can be applied to observations reweighted by $1/\sigma_z(Z)$.

Remark 2. [Monotonicity testing] Monotonicity is an assumption that can be tested. For observable random variables (Y,X), several methods have been developed to test whether $E[Y|X]$ is monotone increasing in X; see, e.g., Ghosal, Sen and van der Vaart (2000), Hall and Heckman (2000), Dümbgen and Spokoiny (2001), Chetverikov (2019), and Hsu, Liu and Shi (2019), among others. All these tests can be adapted for our case, testing the monotonicity of $\sigma^2(\cdot)$ with generated $\{\hat{U}_j^2\}_{j=1}^n$ and observed $\{X_j\}_{j=1}^n$. Since Assumptions A1-A2 and $\hat{\theta}_{\text{OLS}} - \theta = O_p(n^{-1/2})$ imply $\hat{U}_j^2 - U_j^2 = O_p(n^{-1/2}\log n)$ uniformly over $j = 1, \dots, n$, the critical values of these tests can be adjusted accordingly to maintain a proper asymptotic size.

Remark 3. [Misspecification of $E[U^2|X,Z]$] We want to note that even if the assumption in (2.2) is violated (e.g., $E[U^2|X,Z]$ varies with Z or $E[U^2|X,Z] = \sigma^2(X)$ with non-monotone $\sigma^2(\cdot)$), our feasible GLS estimator $\hat{\theta}$ in (2.5) is still consistent for θ due to $E[U|X,Z] = 0$, and asymptotically normal at the $\sqrt{n}$-rate with the limiting distribution

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, E[\rho(X)^{-1}WW']^{-1}E[\rho(X)^{-2}E[U^2|X,Z]WW']E[\rho(X)^{-1}WW']^{-1}),$$

where $\rho(\cdot) = \arg\min_{m \in \mathcal{M}} E[\{U^2 - m(X)\}^2]$ for the class of monotone increasing functions $\mathcal{M}$. Since $\hat{\sigma}^2(\cdot)$ can estimate $\rho(\cdot)$, then the asymptotic variance matrix can be consistently estimated

[2] Although our estimator $\hat{\theta}$ in (2.5) does not involve any tuning constant, the trimming term q_n should be understood as the $c \cdot (n^{-1/3})$-th sample quantile of $\{X_i\}_{i=1}^n$, where the tuning constant is set as $c = 1$. Indeed Theorem 1 holds true with any $c > 0$. If we compare with other nonparametric methods, smoothing parameters, such as bandwidths, series lengths, and neighbors, typically require two constants to implement. For example, for the bandwidth parameter $b = c_1 n^{-c_2}$, researchers need to choose c_1 and c_2. The constant c_1, which is analogous to c above, can be any positive number. However, they also need to choose a positive constant c_2 whose upper bound typically depends on (unknown) smoothness of underlying functions.

by

$$\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^{-2}(X_i)W_iW_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^{-4}(X_i)\hat{U}_i^2W_iW_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^{-2}(X_i)W_iW_i'\right)^{-1}. \quad (2.7)$$

This misspecification robust variance estimator is analogous to the one proposed by Cragg (1992) for the feasible GLS estimator with parametrically specified models for the conditional error variance $E[U^2|X,Z]$.[3]

Remark 4. [Endogenous regressor] The result of Theorem 1 can also be extended to some linear instrumental variable (IV) regression model. For notational simplicity, consider the following univariate IV regression:

$$Y=\alpha+\beta X+U, \qquad E[U|Z]=0,$$

where X is a scalar endogenous regressor and Z is a scalar IV, and we further assume $E[X|Z]=\eta+\gamma Z$ for some parameters (η,γ). This linearity assumption on $E[X|Z]$ is not essential, and may be relaxed by some nonparametric estimator of $E[X|Z]$. In this setup, the optimal instrument for estimating $(\alpha,\beta)'$ is given by (see, e.g., Newey, 1993)

$$E\left[\frac{\partial(Y-\alpha-\beta X)}{\partial(\alpha,\beta)'}\middle|Z\right]E[U^2|Z]^{-1}=-\begin{pmatrix}1 & 0\\ \eta & \gamma\end{pmatrix}\begin{pmatrix}1\\ Z\end{pmatrix}v^{-2}(Z),$$

where $v^2(\cdot)=E[U^2|Z=\cdot]$. Under the assumption of $\gamma\neq 0$ (i.e., the IV is relevant), the optimal IV estimator is obtained by the method of moments estimator of the following moment condition:

$$E\left[\begin{pmatrix}1\\ Z\end{pmatrix}v^{-2}(Z)(Y-\alpha-\beta X)\right]=0. \quad (2.8)$$

Under the monotonicity assumption of $v^2(\cdot)$, we can obtain the isotonic estimator $\hat{v}^2(\cdot)$ for $v^2(\cdot)$ by regressing the squared residuals $\hat{e}^2=(Y-\tilde{\alpha}-\tilde{\beta}X)^2$ for an initial estimator $(\tilde{\alpha},\tilde{\beta})$ (e.g., the two-stage least squares estimator) on Z. The resulting estimator, $\hat{v}^2(\cdot)$, should have the same properties as those of $\hat{\sigma}^2(\cdot)$ presented in Lemma 1, where q_n is replaced with the $(n^{-1/3})$-th sample quantile of $\{Z_i\}_{i=1}^n$. Based on this isotonic estimator, a feasible optimal IV estimator $\hat{\theta}_{\mathrm{IV}}=(\hat{\alpha}_{\mathrm{IV}},\hat{\beta}_{\mathrm{IV}})'$ is given by

$$\hat{\theta}_{\mathrm{IV}}=\left(\sum_{i=1}^{n}\mathbb{I}\{Z_i\geq q_n\}\hat{v}^{-2}(Z_i)\begin{pmatrix}1\\ Z_i\end{pmatrix}(1,X_i)\right)^{-1}\left(\sum_{i=1}^{n}\mathbb{I}\{Z_i\geq q_n\}\hat{v}^{-2}(Z_i)\begin{pmatrix}1\\ Z_i\end{pmatrix}Y_i\right).$$

By applying the same arguments for Theorem 1, we can show that $\hat{\theta}_{\mathrm{IV}}$ is asymptotically equivalent to the infeasible optimal IV estimator based on (2.8) with known $v^2(\cdot)$.

[3]Based on simulation studies, Cragg (1992) recommended to use his misspecification robust variance estimator even when the parametric form of heteroskedasticity is correctly specified. Although a similar analysis is beyond the scope of this paper, we also recommend to employ the variance estimator (2.7) in practice due to its consistency regardless of the assumption in (2.2).

3. Heteroskedasticity by multivariate covariates

We now consider the model

$$Y = \alpha + X'\beta + Z'\gamma + U, \qquad E[U|X, Z] = 0, \tag{3.1}$$

where X is a vector of covariates. This section focuses on the case where heteroskedasticity takes the form of a monotone single index function of X with unknown parameters η_0, i.e., $E[U^2|X, Z] = E[U^2|X] = \sigma^2(X'\eta_0)$ for a monotone increasing function $\sigma^2(\cdot)$. Single index models are known to be more flexible than parametric models and achieve dimension reduction relative to nonparametric models.

Remark 5. First, the monotone index model $\sigma^2(X'\eta_0)$ covers several existing parametric models. Popular examples include $\sigma^2(X) = C(X'\eta_0)^{2-2\lambda}$ (Box and Hill, 1974), $\sigma^2(X) = C\exp(\lambda(X'\eta_0))$ (Bickel, 1978), $\sigma^2(X) = C\{1 + \lambda(X'\eta_0)^2\}$ (Fuller, 1980) for some constants $C > 0$ and λ; interestingly, all these parametric functions are monotone increasing (or decreasing) in the index of X. Second, although the setup $E[U^2|X, Z] = \sigma^2(X'\eta_0)$ assumes that the researcher knows which (sub-)vector of covariates should be included in $\sigma^2(\cdot)$, researchers do *not* have to select those covariates in the case where such prior information is unavailable. They can simply re-define the model in (3.1) without covariates Z (or equivalently specify as $E[U^2|X, Z] = \sigma^2(X'\eta_0 + Z'\eta_{z0})$). Our asymptotic theory below applies even if some covariates are irrelevant for $E[U^2|X, Z]$.

For identification, η_0 is normalized as $||\eta_0|| = 1$. Define

$$\sigma^2_\eta(a) = E[\sigma^2(X'\eta_0)|X'\eta = a]. \tag{3.2}$$

We show in Lemma 4 that $\sigma^2(\cdot)$ and η_0 can be consistently estimated by extending the method proposed in Balabdaoui, Groeneboom and Hendrickx (2019) (BGH hereafter) and Balabdaoui and Groeneboom (2021) to allow generated variables. In particular, for a given η, define the isotonic regression of $\{\hat{U}_i^2\}_{i=1}^n$ on $\{X_i'\eta\}_{i=1}^n$ as

$$\hat{\sigma}^2_\eta = \arg\min_{m\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \{\hat{U}_i^2 - m(X_i'\eta)\}^2, \tag{3.3}$$

where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$. Based on this, $\hat{\eta}$ can be estimated by minimizing the square sum of a score function. For example, the simple score estimator in the spirit of BGH and Balabdaoui and Groeneboom (2021) is given by

$$\hat{\eta} = \arg\min_\eta \left\| \frac{1}{n}\sum_{i=1}^n X_i\{\hat{U}_i^2 - \hat{\sigma}^2_\eta(X_i'\eta)\} \right\|^2, \tag{3.4}$$

where $\|\cdot\|$ is the Euclidean norm: $\|a\| = \sqrt{\sum_{j=1}^k a_j^2}$ for $a = (a_1, \ldots, a_k)' \in \mathbb{R}^k$.

Letting $\hat{\sigma}_i^2 = \hat{\sigma}^2_{\hat{\eta}}(X_i'\hat{\eta})$ and $W = (1, X', Z')'$, we propose the following GLS estimator

$$\hat{\theta} = \left(\sum_{i=1}^n \mathbb{I}\{X_i'\hat{\eta} \geq q_n\}\hat{\sigma}_i^{-2} W_i W_i'\right)^{-1} \left(\sum_{i=1}^n \mathbb{I}\{X_i'\hat{\eta} \geq q_n\}\hat{\sigma}_i^{-2} W_i Y_i\right), \tag{3.5}$$

where q_n is the $(n^{-1/3})$-th sample quantile of $\{X_i'\hat{\eta}\}_{i=1}^n$.

To avoid unnecessarily heavy notations, in the multivariate case, we redefine some notations, which have similar meanings to those used in Section 2. Define $\varepsilon = U^2 - \sigma^2(X'\eta_0)$, $\sigma_\varepsilon^2(\cdot) = E[\varepsilon^2|X'\eta_0 = \cdot]$, $x_L = \inf_{x\in\mathscr{X}}(x'\eta_0)$, and $x_U = \sup_{x\in\mathscr{X}}(x'\eta_0)$. Let $f_X(\cdot)$ be the density function of the random variable $X'\eta_0$. Let q_n^* be the $(n^{-1/3})$-th population quantile of $X'\eta_0$, q_n be the $(n^{-1/3})$-th sample quantile of $\{X_i'\hat{\eta}\}_{i=1}^n$, $c^* = \lim_{n\to\infty} n^{1/3}(q_n^* - x_L)$, and $D_A^L[f](a)$ be the left derivative of the greatest convex minorant of function $f(\cdot)$ evaluated at $a \in A$. Let $\dim(w)$ be the dimension of a vector w.

Assumption.

M1: *$\{Y_i, X_i, Z_i\}_{i=1}^n$ is an iid sample of (Y, X, Z). The support of (X, Z), $\mathscr{X}\times\mathscr{Z}$, is convex with non-empty interiors and is a subset of $\mathcal{B}(0, R)$ for some $R > 0$.*

M2: *(i) There exists $\delta_0 > 0$ such that the function $a \mapsto \sigma_\eta^2(a)$ defined in (3.2) is monotone increasing on $I_\eta = \{x'\eta, x \in \mathscr{X}\}$ for each $\eta \in \mathcal{B}(\eta_0, \delta_0)$. (ii) $0 < \inf_{a\in I_\eta} \sigma_\eta^2(a) < \sup_{a\in I_\eta} \sigma_\eta^2(a) < \infty$ for each $\eta \in \mathcal{B}(\eta_0, \delta_0)$. (iii) There exist positive constants a_0 and M such that $E[|U|^{2s}|X = x] \leq a_0 s! M^{s-2}$ for all integers $s \geq 2$ and $x \in \mathscr{X}$. (iv) $\sigma_\eta^2(\cdot)$ is continuously differentiable on I_η for each $\eta \in \mathcal{B}(\eta_0, \delta_0)$. (v) $\sigma_\varepsilon^2(\cdot)$ is continuous on $(x_L, x_L + \delta_1)$ for some $\delta_1 > 0$.*

M3: *The random variable $X'\eta_0$ has a density function $f_X(\cdot)$ that is continuous on I_{η_0}. There exists some real positive numbers $\underline{b}$ and $\overline{b}$, such that $0 < \underline{b} < f_X(a) < \overline{b} < \infty$ holds for all $a \in I_{\eta_0}$.*

M4: *For each $\eta \in \mathcal{B}(\eta_0, \delta_0)$, the mapping $a \mapsto E[X|X'\eta = a]$ defined on I_η is bounded and has a finite total variation.*

M5: *$\mathrm{Cov}[X'(\eta_0 - \eta),\ \sigma^2(X'\eta_0)|X'\eta] \neq 0$ almost surely for each $\eta \neq \eta_0$.*

M6: *$B := \int (x - E[X|x'\eta_0])(x - E[X|x'\eta_0])' \left.\frac{d\sigma^2(a)}{da}\right|_{a=x'\eta_0} dP(x)$ has rank $\dim(\eta_0) - 1$.*

Assumptions M1-M3 are analogs of Assumptions A1-A3, respectively. The main assumption is the monotonicity of $\sigma_\eta^2(\cdot)$. Assumptions M4-M6 are additional regularity conditions for the monotone index model. By Assumption M1, we have $-\infty < x_L < x_U < \infty$. Then similar to Lemma 1, we obtain the following lemma for the behavior of $\hat{\sigma}_{\hat{\eta}}^2(\cdot)$ around x_L.

Lemma 2. *Under Assumptions M1-M6 and $\lim_{a\downarrow x_L} \frac{d\sigma^2(a)}{da} > 0$, it holds*

$$n^{1/3}\{\hat{\sigma}_{\hat{\eta}}^2(q_n) - \sigma^2(q_n)\} \xrightarrow{d} D_{[0,\infty)}^L \left[\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}} \mathcal{W}_t + \left(\lim_{a\downarrow x_L} \frac{d\sigma^2(a)}{da} \right) c^* \left(\frac{1}{2}t^2 - t \right) \right] (1).$$

Based on this lemma, the asymptotic distribution of the GLS estimator $\hat{\theta}$ in (3.5) is obtained as follows. Let $\sigma_i^2 = \sigma^2(X_i'\eta_0)$.

Theorem 2. *Under Assumptions M1-M6, it holds*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, E[\sigma^{-2}(X'\eta_0)WW']^{-1}),$$

and the asymptotic variance matrix is consistently estimated by $\left(\frac{1}{n}\sum_{i=1}^n \hat{\sigma}_i^{-2} W_i W_i'\right)^{-1}$.

Similar comments to Theorem 1 apply here. Our estimator $\hat{\theta}$ is asymptotically equivalent to the infeasible GLS estimator $\hat{\theta}_{\text{IGLS}}$. In terms of technical contribution, our theoretical analysis generalizes existing ones in, e.g., Babii and Kumar (2023), BGH, and Balabdaoui and Groeneboom (2021) to accommodate generated variables. Similar to Remark 3, even when the monotonicity assumption of $\sigma_\eta^2(\cdot)$ is violated, $\hat{\theta}$ is still consistent for θ and asymptotically normal at the $\sqrt{n}$-rate with certain robust asymptotic variance. Furthermore, endogenous regressors can be accommodated as in Remark 4.

Remark 6. We can suggest two informal robustness checks for the monotone index assumption in (3.2). One is to compute the standard errors robust to possible misspecification obtained in the same manner as Remark 3 and compare them to those in Theorem 2. This can serve as a robustness check for the monotone specification given variables of the conditional error variance functions. Another is to report the results for the specification where all exogenous variables are included to $\sigma^2(\cdot)$ in addition to those for the chosen specifications. A large difference between these results can be a sign of the misspecification of the chosen ones. See Section 4.2 for illustration.

Remark 7. In this section, we employ the monotone single index structure to model the multivariate conditional variance function. This strategy allows us to strike a balance between robustness and mitigating the curse of dimensionality. Indeed, the current specification can be extended to the multiple index model $E[U^2|X = x] = x_0'\eta_0 + \sum_{i=1}^M G_i(x_i'\eta_i)$, for $X = (X_0', X_1' \dots, X_M')'$, where $\{G_i(\cdot)\}_{i=1}^M$ are unknown monotone increasing functions. For the case of $M = 1$, this model simplifies to a monotone partially linear single index model whose properties have been studied by Xu and Otsu (2020). We are optimistic that, under certain regularity conditions, similar results as in this section can be obtained. To the best of our knowledge, we have not come across any works that discuss the multiple monotone index model with $M > 1$ even for the conventional regression setup for $E[Y|X = x]$. A possible solution could be derived by combining the existing literature on the monotone single index model (as cited in Section 1) with the literature on the monotone additive model (for instance, Mammen and Yu, 2007). Another potential extension involves employing the nonparametric framework of Fang, Guntuboyina and Sen (2021) to model the multivariate conditional variance function. This framework is free of parametric structure, and it requires the true conditional variance to be entirely monotone increasing in its arguments, i.e., $\sigma^2(x_1, z_1) \le \sigma^2(x_2, z_2)$ if only if $x_1 \le x_2$ and $z_1 \le z_2$. Explorations of these extensions exceed the scope of this paper, and we leave them for future research.

4. Numerical illustrations

4.1. **Simulation.** We now investigate the finite sample properties of the proposed GLS estimator by a Monte Carlo experiment. We follow the simulation design by Cragg (1983) and Newey (1993). The first data generating process, denoted by DGP1, is the heteroskedastic linear model with a univariate covariate and normally distributed disturbance:[4]

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + u_i, \qquad u_i = \sigma_i \varepsilon_i, \qquad \varepsilon_i \sim N(0,1), \\
\beta_0 &= \beta_1 = 1, \qquad \log(X_i) \sim N(0,1), \qquad X_i \text{ and } \varepsilon_i \text{ are independent}, \\
\sigma_i^2 &= .1 + .2X_i + .3X_i^2. \qquad (4.1)
\end{aligned}
$$

We consider three sample sizes, $n = 50$, 100, and 500. The number of replications is set to 1,000.

In addition to the feasible GLS estimator with monotone heteroskedasticity (MGLS), we consider the ordinary least squares (OLS), infeasible generalized least squares (GLS), feasible GLS (FGLS), and nearest neighbor estimators (k-NN). GLS requires knowledge of the conditional error variance function (4.1), including the values of the coefficients. In contrast, FGLS proceeds with the known functional form, but the coefficients are estimated. The "k-NN automatic" chooses the number of neighbors by a cross-validation procedure suggested by Newey (1993). All the estimators except OLS are the weighted least squares estimators, and their differences come from how the weights are calculated. Following Newey (1993), we calculate the weights for each method by taking a ratio of the predicted squared residual to the estimated variance of the disturbance, censoring the result below 0.04.

Table 4.1 presents the simulation results for estimation. The first column shows the estimation methods, and the following two columns show the root mean-squared error (RMSE) and mean absolute error (MAE) for DGP1 with $n = 50$. The results for GLS report the levels of the RMSE and MAE, and those for others are their ratios relative to GLS. The next two columns give the corresponding results with $n = 100$ and the last two columns with $n = 500$. Two rows for each estimator show the results for β_0 and β_1, respectively. The inefficiency and inaccuracy of OLS are apparent. FGLS performs quite well, and this is natural when the conditional error variance functions are correctly specified. The performance of k-NN varies with the choice of k and is in between OLS and FGLS. We observe that the performance of MGLS is better than k-NN in every choice of smoothing parameters. The result of MGLS is comparable to that of FGLS if not better. MGLS's independence of a smoothing parameter is clearly desirable. We also note that MGLS performs well even for $n = 50$.

The last four columns of Table 4.1 present the results for DGP2 with a homoskedastic error:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + u_i, \qquad u_i \sim N(0,1), \\
\beta_0 &= \beta_1 = 1, \qquad \log(X_i) \sim N(0,1), \qquad X_i \text{ and } u_i \text{ are independent.}
\end{aligned}
$$

For DGP2, all estimators work reasonably well although the performance of k-NN with $k = 6$ is worse than others.

[4]Normal random variables are not compactly supported, and hence it violates Assumption A1. However, as discussed in the remark on Assumption A1, this assumption can be relaxed.

Table 4.1. Simulation: Estimation with univariate covariate

	DGP1						DGP2					
	$n = 50$		$n = 100$		$n = 500$		$n = 50$		$n = 100$		$n = 500$	
Estimator	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
GLS (infeasible)	0.132	0.085	0.093	0.059	0.041	0.028	0.194	0.122	0.133	0.088	0.057	0.039
	0.157	0.100	0.108	0.073	0.048	0.032	0.083	0.046	0.055	0.034	0.021	0.014
OLS	3.103	2.856	3.831	3.479	5.574	4.495	1.000	1.000	1.000	1.000	1.000	1.000
	2.072	2.098	2.543	2.370	3.377	2.971	1.000	1.000	1.000	1.000	1.000	1.000
FGLS	1.279	1.210	1.245	1.233	1.598	1.152	1.032	1.041	1.026	1.033	1.024	1.067
	1.427	1.268	1.406	1.280	1.271	1.242	1.090	1.092	1.075	1.036	1.088	1.090
k-NN (Automatic)	1.630	1.373	1.633	1.511	1.355	1.167	1.123	1.081	1.130	1.081	1.181	1.138
	1.535	1.427	1.606	1.431	1.424	1.267	1.092	1.074	1.065	1.006	1.197	1.097
k-NN ($k = 6$)	1.554	1.361	1.525	1.498	1.474	1.417	1.274	1.243	1.253	1.155	1.359	1.276
	1.466	1.421	1.472	1.462	1.454	1.459	1.178	1.143	1.177	1.114	1.350	1.344
k-NN ($k = 15$)	1.600	1.386	1.566	1.365	1.251	1.108	1.037	1.076	1.079	1.059	1.081	1.140
	1.520	1.398	1.546	1.408	1.247	1.197	1.003	1.046	1.037	1.012	1.066	1.053
k-NN ($k = 24$)	1.781	1.568	1.685	1.457	1.291	1.160	1.011	1.039	1.039	0.980	1.044	1.098
	1.630	1.560	1.673	1.471	1.312	1.246	1.002	1.026	1.015	0.994	1.038	1.025
MGLS	1.379	1.285	1.326	1.279	1.113	1.129	1.039	1.091	1.049	1.075	1.027	1.075
	1.327	1.214	1.332	1.249	1.113	1.144	1.043	1.051	1.051	1.058	1.055	1.066

Note: "RMSE" and "MAE" stand for the root mean squared error and mean absolute error, respectively. The results for GLS report the levels of the RMSE and MAE, and those for others are their ratios relative to GLS.

Next, we consider the heteroskedastic linear models with multivariate covariates, denoted by DGP3:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \qquad u_i = \sigma_i \varepsilon_i, \qquad \varepsilon_i \sim N(0,1), \\ \beta_0 &= \beta_1 = \beta_2 = 1, \qquad \log(X_{1i}),\ \log(X_{2i}) \sim N(0,1), \qquad X_{1i},\ X_{2i} \text{ and } \varepsilon_i \text{ are independent}, \\ \sigma_i^2 &= .2(X_{1i} + X_{2i})^2. \end{aligned} \tag{4.2}$$

The conditional error variance function of DPG3 is of a monotone single index structure. Using the notation in (3.2), DGP3 corresponds to the structure with $\sigma^2(a) = a^2$, $X' = (X_1, X_2)$, and $\eta_0 = (\sqrt{.2}, \sqrt{.2})'$. The left panel of Table 4.2 shows the results of DGP3 in the same manner as Table 4.1. For each estimation method, two rows show the results for β_0 and β_1, and those for β_2 are omitted to avoid redundancy. k-NNs and MGLS perform better than FGLS, and this is in contrast to the performance of DGP1. In general, MGLS works better than k-NNs except for a few cases.

To see the potential applicability of MGLS to a non-single index structure, we consider another heteroskedastic linear model denoted by DGP4:

$$\begin{aligned} &Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad u_i = \sigma_i \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \\ &\beta_0 = \beta_1 = \beta_2 = 1, \quad \log(X_{1i}),\ \log(X_{2i}) \sim N(0,1), \qquad X_{1i},\ X_{2i} \text{ and } \varepsilon_i \text{ are independent}, \\ &\sigma_i^2 = .1 + .2\tilde{X}_i + .3\tilde{X}_i^2, \quad \log(\tilde{X}_i) = \frac{\log(X_{1i}) + \log(X_{2i})}{\sqrt{2}}. \end{aligned} \tag{4.3}$$

The right panel of Table 4.2 shows the results. The results for DGP 4 are overall similar to those of DGP3. An exception is FGLS, which performs poorly for DGP3. MGLS works remarkably well for the heteroskedasticity of a non-single index structure.

Table 4.2. Simulation: Estimation with multivariate covariates

	DGP 3						DGP 4					
	$n = 50$		$n = 100$		$n = 500$		$n = 50$		$n = 100$		$n = 500$	
Estimator	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
GLS (infeasible)	0.162	0.103	0.110	0.071	0.045	0.028	0.165	0.108	0.115	0.076	0.049	0.033
	0.163	0.107	0.109	0.072	0.048	0.033	0.108	0.067	0.071	0.046	0.029	0.020
OLS	3.401	3.589	4.255	4.168	6.650	6.695	3.069	2.653	3.792	2.980	4.897	4.186
	1.942	1.950	2.317	2.260	3.051	2.610	2.318	2.170	2.914	2.338	3.809	3.198
FGLS	2.531	2.239	2.516	2.141	2.731	2.037	1.381	1.189	1.427	1.233	1.699	1.219
	1.606	1.441	1.709	1.486	1.732	1.358	1.359	1.227	1.344	1.281	1.326	1.271
k-NN (Automatic)	1.952	1.925	2.108	1.709	1.786	1.537	1.868	1.638	1.778	1.488	1.709	1.390
	1.546	1.429	1.680	1.489	1.516	1.318	1.766	1.771	1.865	1.763	2.009	1.626
k-NN ($k = 6$)	1.827	1.766	1.787	1.587	1.670	1.666	1.719	1.541	1.674	1.521	1.594	1.537
	1.458	1.397	1.514	1.486	1.497	1.362	1.717	1.704	1.769	1.764	1.813	1.654
k-NN ($k = 15$)	1.914	1.957	1.850	1.669	1.490	1.385	1.769	1.611	1.669	1.491	1.373	1.246
	1.468	1.401	1.511	1.428	1.313	1.248	1.712	1.727	1.769	1.639	1.588	1.517
k-NN ($k = 24$)	2.182	2.203	2.008	1.816	1.570	1.510	1.952	1.888	1.799	1.581	1.392	1.254
	1.562	1.455	1.611	1.571	1.371	1.267	1.825	1.807	1.890	1.729	1.626	1.511
MGLS	2.144	1.977	1.993	1.659	1.667	1.481	1.839	1.549	1.647	1.422	1.320	1.251
	1.486	1.467	1.477	1.401	1.238	1.186	1.670	1.533	1.604	1.451	1.448	1.360

Note: "RMSE" and "MAE" stand for the root mean squared error and mean absolute error, respectively. The results for GLS report the levels of the RMSE and MAE, and those for others are their ratios relative to GLS.

Next, we turn to the simulation results on inference. Tables 4.3 and 4.4 show empirical coverages (EC) and average lengths (AL) for the 95% confidence intervals under DGPs 1-4. Again we consider GLS, OLS, FGLS, k-NN, and MGLS. For OLS, three types of confidence intervals are considered. They are based on the usual OLS standard error (OLS-U), the heteroskedasticity-robust standard error (OLS-R), and the wild bootstrap standard error (OLS-boot). For MGLS, we also present the results for its robust version. We observe that the empirical coverages are smaller than the nominal coverage 0.95 for all DGPs and all methods except GLS. It is natural that OLS-U performs poorly since it is invalid except for DGP2. The performance of k-NN is worse than others for all DGPSs in terms of empirical coverage. OLS-R, OLS-boot, FGLS, and MGLS work similarly in terms of empirical coverage, however, we note that the average length of OLS-R is much larger than those of FGLS and MGLS except for DGP2. While the empirical coverages of OLS-Boot are similar to those of OLS-R, the average lengths of OLS-Boot are smaller than those of OLS-R but still larger than those of MGLS. MGLS works quite well for all DGPs, especially for $n = 500$. The results of MGLS (Robust) are similar to those of MGLS especially when $n = 100$ and 500. Finally, we note that the empirical coverages tend to be lower when $n = 50$ than when $n = 100$ and 500. Careful interpretation of results is recommended when the sample size is small.

Table 4.3. Simulation: Inference with univariate covariate

	DGP 1						DGP 2					
	$n = 50$		$n = 100$		$n = 500$		$n = 50$		$n = 100$		$n = 500$	
Estimator	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
GLS (infeasible)	0.956	0.528	0.955	0.370	0.956	0.164	0.939	0.749	0.947	0.516	0.955	0.224
	0.946	0.627	0.962	0.441	0.960	0.196	0.948	0.316	0.952	0.207	0.970	0.085
OLS-U	0.798	1.008	0.742	0.740	0.636	0.349	0.939	0.749	0.947	0.516	0.955	0.224
	0.492	0.409	0.421	0.290	0.348	0.131	0.948	0.316	0.952	0.207	0.970	0.085
OLS-R	0.766	0.962	0.805	0.862	0.884	0.648	0.933	0.733	0.941	0.507	0.949	0.222
	0.730	0.761	0.772	0.689	0.880	0.488	0.874	0.272	0.881	0.185	0.935	0.081
OLS-Boot	0.740	0.885	0.845	0.517	0.907	0.356	0.917	0.718	0.947	0.516	0.955	0.224
	0.690	0.681	0.856	0.527	0.894	0.345	0.846	0.270	0.952	0.207	0.970	0.085
FGLS	0.800	0.451	0.847	0.328	0.872	0.162	0.916	0.709	0.925	0.493	0.935	0.216
	0.737	0.504	0.812	0.395	0.885	0.195	0.761	0.231	0.758	0.159	0.844	0.072
k-NN (Automatic)	0.708	0.410	0.659	0.258	0.701	0.102	0.902	0.711	0.884	0.483	0.883	0.205
	0.576	0.351	0.574	0.251	0.650	0.115	0.927	0.306	0.917	0.197	0.881	0.079
k-NN ($k = 6$)	0.732	0.410	0.666	0.258	0.621	0.102	0.845	0.711	0.845	0.483	0.819	0.205
	0.576	0.351	0.574	0.251	0.650	0.115	0.927	0.306	0.917	0.197	0.881	0.079
k-NN ($k = 15$)	0.735	0.418	0.704	0.266	0.717	0.105	0.929	0.725	0.907	0.492	0.914	0.210
	0.582	0.353	0.592	0.258	0.677	0.118	0.944	0.310	0.931	0.200	0.919	0.081
k-NN ($k = 24$)	0.711	0.440	0.688	0.269	0.721	0.107	0.945	0.744	0.921	0.504	0.935	0.216
	0.512	0.324	0.537	0.244	0.668	0.118	0.953	0.316	0.942	0.204	0.939	0.083
MGLS	0.779	0.499	0.812	0.363	0.905	0.165	0.885	0.640	0.907	0.468	0.937	0.219
	0.725	0.523	0.744	0.392	0.888	0.188	0.951	0.333	0.968	0.222	0.972	0.092
MGLS (Robust)	0.762	0.483	0.791	0.354	0.903	0.163	0.879	0.635	0.902	0.463	0.933	0.216
	0.725	0.465	0.744	0.359	0.888	0.181	0.951	0.258	0.968	0.177	0.972	0.079

Note: "EC" and "AL" stand for the empirical coverage probability and average length, respectively. "OLS-U", "OLS-R", and "OLS-Boot" use the normal approximation with the usual OLS standard error, the heteroskedasticity robust standard error, and the percentile bootstrap interval, respectively. "MGLS (Robust)" is based on the variance formula presented in Remark 2.

Table 4.4. Simulation: Inference with multivariate covariates

	DGP 3						DGP 4					
	$n = 50$		$n = 100$		$n = 500$		$n = 50$		$n = 100$		$n = 500$	
Estimator	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
GLS (infeasible)	0.944	0.611	0.951	0.413	0.946	0.175	0.944	0.636	0.943	0.440	0.956	0.192
	0.951	0.632	0.961	0.439	0.960	0.194	0.949	0.414	0.950	0.282	0.968	0.123
OLS-U	0.824	1.535	0.786	1.108	0.632	0.511	0.819	1.222	0.780	0.893	0.675	0.412
	0.589	0.526	0.549	0.369	0.491	0.164	0.639	0.420	0.611	0.298	0.521	0.133
OLS-R	0.787	1.411	0.815	1.232	0.869	0.873	0.797	1.197	0.843	1.068	0.906	0.718
	0.729	0.767	0.782	0.660	0.891	0.441	0.762	0.596	0.810	0.517	0.914	0.334
OLS-Boot	0.756	1.319	0.781	1.133	0.839	0.797	0.752	1.115	0.785	0.951	0.860	0.654
	0.688	0.708	0.749	0.593	0.826	0.400	0.719	0.569	0.780	0.460	0.866	0.301
FGLS	0.831	1.069	0.845	0.759	0.897	0.336	0.801	0.596	0.823	0.424	0.834	0.191
	0.658	0.517	0.722	0.395	0.826	0.198	0.797	0.382	0.862	0.262	0.855	0.112
k-NN (Automatic)	0.571	0.481	0.557	0.289	0.587	0.105	0.672	0.526	0.646	0.323	0.609	0.122
	0.471	0.296	0.472	0.205	0.534	0.091	0.596	0.298	0.599	0.200	0.605	0.082
k-NN ($k = 6$)	0.597	0.481	0.574	0.289	0.549	0.105	0.670	0.526	0.639	0.323	0.571	0.122
	0.471	0.296	0.472	0.205	0.534	0.091	0.596	0.298	0.599	0.200	0.605	0.082
k-NN ($k = 15$)	0.592	0.504	0.590	0.299	0.639	0.108	0.690	0.551	0.675	0.338	0.689	0.129
	0.491	0.301	0.484	0.212	0.570	0.096	0.607	0.309	0.629	0.210	0.668	0.088
k-NN ($k = 24$)	0.582	0.557	0.561	0.313	0.618	0.111	0.681	0.590	0.664	0.347	0.702	0.132
	0.459	0.287	0.450	0.204	0.562	0.096	0.598	0.297	0.608	0.207	0.662	0.091
MGLS	0.803	0.956	0.863	0.623	0.938	0.248	0.801	0.805	0.844	0.526	0.902	0.216
	0.687	0.524	0.756	0.401	0.897	0.198	0.707	0.404	0.776	0.305	0.908	0.154
MGLS (Robust)	0.755	0.855	0.833	0.573	0.920	0.234	0.762	0.750	0.833	0.511	0.902	0.219
	0.687	0.548	0.756	0.415	0.897	0.199	0.707	0.422	0.776	0.308	0.908	0.148

Note: "EC" and "AL" stand for the empirical coverage probability and average length, respectively. "OLS-U", "OLS-R", and "OLS-Boot" use the normal approximation with the usual OLS standard error, the heteroskedasticity robust standard error, and the percentile bootstrap interval, respectively. "MGLS (Robust)" is based on the variance formula presented in Remark 2.

4.2. **Empirical example.** We illustrate how the proposed method in this paper can improve the precision of the traditional OLS approach. In doing so, we revisit Acemoglu and Restrepo (2017) that investigate the relationship between an aging population and economic growth. After Hansen (1939), a popular perspective is that countries undergoing faster aging suffer more economically partly because of excessive savings by an aging population. In contrast to the perspective, Acemoglu and Restrepo (2017) find no evidence of a negative relationship between aging and GDP per capita after controlling for initial GDP per capita, initial demographic composition, and trends by region.

Acemoglu and Restrepo (2017) estimated eight specifications for the regression of the change in (log) GDP per capita from 1990 to 2015 (denoted by GDP) on the population aging measured by the change in the ratio of the population above 50 to those between the ages of 20 and 49 (denoted by Aging). The results are reproduced in Panel A of Table 4.5. Those in columns 1-5 are based on the sample including 169 countries. Column 1 shows the result of the simple regression. Standard errors robust to heteroskedasticity are reported in square brackets. Column 2 shows the result with an additional regressor, the initial log GDP per worker in 1990. Column 3 in addition includes the initial demographic information, the ratio of the population above 50 to those between 20 and 49 in 1990 (denoted by Initial Ratio), and the population in 1990. Column 4 additionally uses dummies for seven regions, Latin America, East Asia, South Asia, Africa, North Africa and Middle East, Eastern Europe and Central Asia, and Developed Countries. Column 5 estimates the same specification as Column 4 with instruments of birthrates for the 1960, 1965, 1970, 1975, and 1980 cohorts. Columns 6 to 8 report the result for OECD countries using specifications of Columns 1, 3, and 5, respectively. The number of observations for the first five columns is 169, and that for the last three columns is 35. Seven out of eight OLS estimates indicate positive relationships and five of them are statistically significant at the 5 percent level. Acemoglu and Restrepo (2017) discuss that these findings can be explained by the adoption of automation technologies based on a theoretical model.

We estimate the same specifications by MGLS proposed in this paper. Acemoglu and Restrepo (2017) show that the negative effect of aging can be mitigated or reversed by adopting new automation technologies given abundant capital. This also implies that the effect of aging can be negative without sufficient capital. Hence it would be reasonable to consider Aging as a source of heteroskedasticity. The upper panel of Figure 4.2 shows the relationship between the residual from the simple regression of column 1 in Panel A and Aging. Heteroskedasticity due to Aging is not easily confirmed visually. We consider Initial Ratio as another source of heteroskedasticity since the low ratio of old to young in 1990 is likely correlated with more aging in 2015, leading to larger variability in GDP per capita by the same reasoning discussed above. The lower panel of Figure 4.2 presents the relationship between the residual from the simple regression of column 1 in Panel A and Initial Ratio, and we see that the variability decreases with the growing ratio.

Panels B, C, and D of Table 4.5 show the results of MGLS. Panels B and C present the results for cases where the conditional error variance functions depend on Aging and Initial Ratio, respectively. Panel D reports the results where the conditional error variance functions depend on all exogenous regressors except the regional dummies. Standard errors based on

Theorems 1-2 and their analogous versions for IV estimators are reported in parentheses, while robust standard errors are reported in square brackets. First, we observe reductions in standard errors for all MGLSs relative to OLS. The differences stand out when $n = 169$. Second, the two standard errors are similar for the MGLS estimates under exogeneity while they differ for the IV estimates. These are the supporting evidence for the monotone specification of the conditional error variance function for the MGLS method with exogenous regressors but not for the IV method. Third, the results given in Columns 2, 3, and 4 are stable, while the results of IV estimates and OECD countries contain a lot of variations. Those variations can be due to non-monotone conditional error variance functions and/or small sample sizes, and further investigations will be required. Overall, the standard errors of MGLS tend to be smaller or no larger than those of OLS, which demonstrates the increased precision of MGLS.

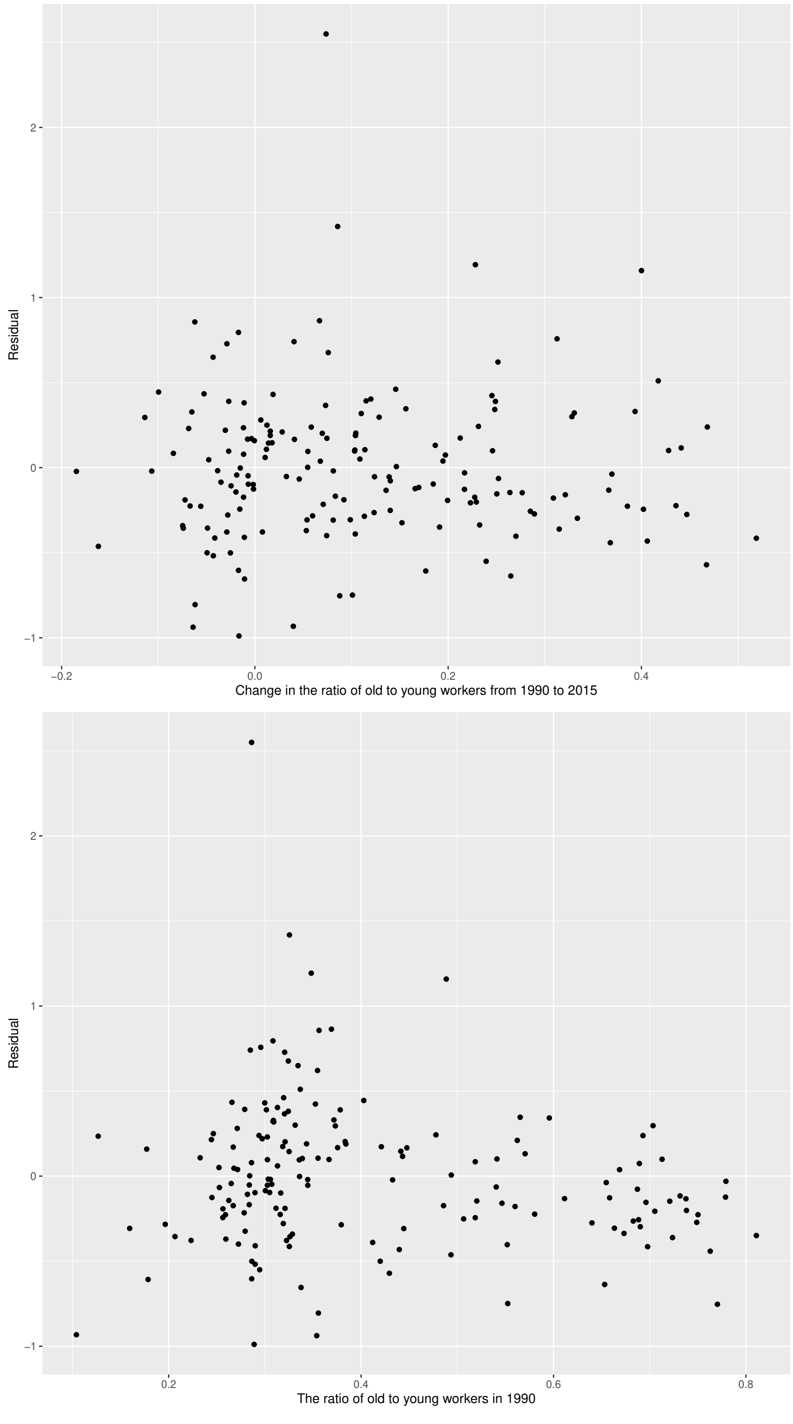

FIGURE 4.1. Plots for residual and aging (upper) and residual and ratio of old to young workers in 1990 (lower)

Note: For both panels, the residuals are obtained from the regression of the change in GDP per capita from 1990 to 2015 (GDP) on the population aging measured by the ratio of the population above 50 to those between the ages of 20 and 49 (Aging). For the upper panel, the variable on the X-axis represents the change in the ratio of old to young workers from 1990 to 2015. For the lower panel, it represents the ratio of old to young workers in 1990.

Table 4.5. Effects of the Aging on GDP by OLS and MGLS

	Sample of all countries ($n = 169$)					OECD countries ($n = 35$)		
Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS								
Aging	0.335	1.036	1.162	0.773	1.703	-0.262	0.042	1.186
	(0.210)	(0.257)	(0.276)	(0.322)	(0.411)	(0.352)	(0.346)	(0.458)
Initial GDP		-0.153	-0.138	-0.156	-0.190		-0.205	-0.260
		(0.039)	(0.042)	(0.046)	(0.045)		(0.072)	(0.092)
Panel B: MGLS (covariate of $\sigma^2(\cdot)$ =Aging)								
Aging	0.387	1.098	1.191	0.751	0.414	-0.391	-0.029	-0.458
	(0.189)	(0.187)	(0.205)	(0.267)	(0.101)	(0.247)	(0.284)	(0.092)
	[0.150]	[0.179]	[0.198]	[0.310]	[0.472]	[0.190]	[0.340]	[0.451]
Initial GDP		-0.164	-0.155	-0.168	-0.079		-0.190	-0.297
		(0.027)	(0.029)	(0.030)	(0.009)		(0.069)	(0.025)
		[0.031]	[0.032]	[0.029]	[0.046]		[0.069]	[0.141]
Panel C: MGLS (covariate of $\sigma^2(\cdot)$ =Initial Ratio)								
Aging	0.065	0.771	0.894	0.574	0.483	-0.501	-0.344	-0.585
	(0.196)	(0.223)	(0.231)	(0.235)	(0.142)	(0.270)	(0.219)	(0.226)
	[0.196]	[0.249]	[0.262]	[0.272]	[0.603]	[0.231]	[0.213]	[0.758]
Initial GDP		-0.164	-0.141	-0.159	-0.080		-0.148	-0.379
		(0.031)	(0.035)	(0.041)	(0.012)		(0.056)	(0.096)
		[0.035]	[0.037]	[0.046]	[0.055]		[0.065]	[0.288]
Panel D: MGLS (covariates of $\sigma^2(\cdot)$ =All)								
Aging	0.285	1.064	1.188	0.810	0.494	-0.391	0.062	-0.268
	(0.221)	(0.265)	(0.281)	(0.289)	(0.100)	(0.247)	(0.274)	(0.186)
	[0.206]	[0.249]	[0.271]	[0.323]	[0.442]	[0.190]	[0.340]	[0.972]
Initial GDP		-0.152	-0.136	-0.146	-0.079		-0.203	-0.250
		(0.030)	(0.033)	(0.041)	(0.009)		(0.072)	(0.049)
		[0.033]	[0.036]	[0.044]	[0.051]		[0.072]	[0.197]

Note: For all specifications from (1) to (8), GDP is the dependent variable. Column 1 shows the result of the simple regression of Aging on GDP. Column 2 shows the result with an additional regressor, the initial log GDP per worker in 1990. Column 3, in addition, includes the initial demographic information, the ratio of the population above 50 to those between 20 and 49 in 1990 (denoted by Initial Ratio), and the population in 1990. Column 4 additionally uses dummies for seven regions, Latin America, East Asia, South Asia, Africa, North Africa and Middle East, Eastern Europe and Central Asia, and Developed Countries. Columns (6), (7) and (8) report the result for OECD countries using specifications (1), (3) and (5), respectively. Panel A reproduces the results by Acemoglu and Restrepo (2017). For Panel A, heteroskedasticity robust standard errors are presented in parentheses. Panels B, C, and D present the results by MGLS. Panels B and C show the results where the conditional error variance functions depend on Aging and Initial Ratio, respectively. Panel D reports the results where the conditional error variance functions depend on all exogenous variables except the regional dummies. For Columns (1)-(4) and (6)-(7) of Panels B and C, standard errors based on the formula in Theorem 1 are presented in parentheses, while those based on the formula in Remark 3 are presented in square brackets. For Columns (1)-(4) and (6)-(7) of Panel D, standard errors based on the formula in Theorem 2 are presented in parentheses, while those based on the formula analogous to Remark 3 are presented in square brackets. For Columns (5) and (8) of Panels B, C, and D, standard errors are based on the formulae analogous to Remark 3.

Appendix A. Proof of lemma and theorem in Section 2

Notation. In this section, we use the following notation. For a function $f(\cdot)$, we let $\|f\|_\infty = \sup_{x\in\mathscr{X}} |f(x)|$ be the sup-norm and $\|f\|_{2,P} = \sqrt{\int |f(x)|^2 dP}$ be the $L_2(P)$ norm; given there is no confusion in the context, we use the same set of notations for a vector $a = (a_1, \ldots, a_k)'$: we let $\|a\|_\infty = \max_{j\in\{1,\ldots,k\}} |a_k|$ be the sup-norm and $\|a\| = \sqrt{\sum_{j=1}^k a_j^2}$ be the Euclidean norm. $D_A^L[f](a)$ be the left derivative of the greatest convex minorant of a function f evaluated at $a \in A$, $\mathbb{P}_n$ be the empirical measure of $\{Y_i, X_i, Z_i\}_{i=1}^n$, $\mathbb{G}_n$ be the empirical process, i.e., $\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{f(X_i) - E[f(X_i)]\}$, $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$, and $\mathbb{I}_A(x) = \mathbb{I}\{x \in A\}$. Let $\tau_0(x) = \sigma^2(x)$, $\tau_0'(x_L)$ be the right derivative of τ_0 at x_L, $\hat{\tau}(x) = \hat{\sigma}^2(x)$, $\mathscr{W}$ be the support of $W := (1, X, Z')'$, $F(x)$ be the distribution function of X, $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \le x\}$, and $M_n(x) = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 \mathbb{I}\{X_i \le x\}$. For $a, b \in \mathbb{R}$, let $a \wedge b$ denote $\min\{a, b\}$, and $a \lesssim b$ denote that there exists a positive constant C such that $a \le C \cdot b$. Let $\dim(w)$ be the dimension of a vector w.

A.1. **Proof of Lemma 1.** Since $\hat{U}_j = Y_j - W_j'\hat{\theta}_{\text{OLS}}$ is the OLS residual, Assumptions A1-A2 and $\hat{\theta}_{\text{OLS}} - \theta = O_p(n^{-1/2})$ imply $\hat{U}_j^2 - U_j^2 = O_p(n^{-1/2}\log n) = o_p(n^{-1/3})$ uniformly over $j = 1, \ldots, n$. To see this, decompose

$$\begin{aligned}
\hat{U}_j^2 - U_j^2 &= (Y_j - W_j'\hat{\theta}_{\text{OLS}})^2 - (Y_j - W_j'\theta)^2 \\
&= W_j'(\hat{\theta}_{\text{OLS}} + \theta) \cdot W_j'(\hat{\theta}_{\text{OLS}} - \theta) - 2W_j'\theta \cdot W_j'(\hat{\theta}_{\text{OLS}} - \theta) - 2U_j W_j'(\hat{\theta}_{\text{OLS}} - \theta) \\
&=: I_j + II_j + III_j.
\end{aligned}$$

For I_j, note that

$$\begin{aligned}
I_j &= [W_j'(\hat{\theta}_{\text{OLS}} - \theta)]^2 + 2W_j'\theta \cdot W_j'(\hat{\theta}_{\text{OLS}} - \theta) \\
&\le \|W_j\|^2 \left\|\hat{\theta}_{\text{OLS}} - \theta\right\|^2 + 2\|W_j\| \cdot \|\theta\| \cdot \|W_j\| \cdot \left\|\hat{\theta}_{\text{OLS}} - \theta\right\| \\
&\le R^2 \left\|\hat{\theta}_{\text{OLS}} - \theta\right\|^2 + 2R^2 \|\theta\| \cdot \left\|\hat{\theta}_{\text{OLS}} - \theta\right\| = O_p(n^{-1/2}),
\end{aligned} \tag{A.1}$$

where R is the constant defined in Assumption A1. The first inequality follows from the Cauchy-Schwarz inequality, the second inequality follows from $\|W_j\| \le R$ (by Assumption A1), and the last equality follows from $\hat{\theta}_{\text{OLS}} - \theta = O_p(n^{-1/2})$. Note that in the second inequality, the upper bound no longer depends on the index j, so we have $\max_j |I_j| = O_p(n^{-1/2})$. For II_j, using the same reasoning as for the first inequality in (A.1), we have $\max_j |II_j| = O_p(n^{-1/2})$. Note that here we only consider the second term following the first inequality of (A.1). For III_j, the same argument as above yields $\max_j |W_j'(\hat{\theta}_{\text{OLS}} - \theta)| = O_p(n^{-1/2})$. Furthermore, by Assumption A2 and a similar argument after equation (7.11) on p.3297 of Balabdaoui, Durot and Jankowski (2019) (BDJ hereafter), we have $\max_{1\le j\le n} |U_j^2| = O_p(\log n)$. By the fact that $\max_{1\le j\le n} |U_j| \le \max_{1\le j\le n} |U_j^2$— if $\max_{1\le j\le n} |U_j| \ge 1$, we have

$$\max_{1\le j\le n} |U_j| = O_p(\log n). \tag{A.2}$$

In the case of $\max_{1\le j\le n} |U_j| < 1$, $\max_{1\le j\le n} |U_j| = O_p(\log n)$ holds trivially. Combining (A.2) and $\max_j |W_j'(\hat{\theta}_{\text{OLS}} - \theta)| = O_p(n^{-1/2})$, we have $\max_j |III_j| = O_p(n^{-1/2}\log n)$. Consequently, we have

$$\max_j |\hat{U}_j^2 - U_j^2| \leq \max_j |I_j| + \max_j |II_j| + \max_j |III_j|$$
$$= O_p(n^{-1/2}\log n) = o_p(n^{-1/3}). \qquad \text{(A.3)}$$

Furthermore, Assumption A3 guarantees $q_n^* - x_L = O(n^{-1/3})$ (by an expansion of $q_n^* = F^{-1}(n^{-1/3})$ for the quantile function $F^{-1}(\cdot)$ of X), and we can define $c^* = \lim_{n\to\infty} n^{1/3}(q_n^* - x_L) = \left.\frac{dF^{-1}(q)}{dq}\right|_{q\downarrow 0} \in (0,\infty)$.

Now, we analyze $n^{1/3}\{\hat{\tau}(q_n^*) - \tau_0(x_L)\}$. The term $n^{1/3}\{\hat{\tau}(q_n) - \tau_0(q_n)\}$ will be addressed in the final step of this subsection. Pick any $m > 0$. Let

$$Z_{n1}(t) = n^{2/3}[\{n^{-1/3}m + \tau_0(x_L)\}F_n(x_L + t(q_n^* - x_L)) - M_n(x_L + t(q_n^* - x_L))].$$

Observe that

$$P\left(n^{1/3}\{\hat{\tau}(q_n^*) - \tau_0(x_L)\} \leq m\right) = P\left(\arg\max_{s\in[x_L,x_U]}[\{n^{-1/3}m + \tau_0(x_L)\}F_n(s) - M_n(s)] \geq q_n^*\right)$$
$$= P\left(\arg\max_{t\in[0,(x_U-x_L)/(q_n^*-x_L)]} n^{-2/3}Z_{n1}(t) \geq 1\right), \qquad \text{(A.4)}$$

where the first equality follows from the switch relation (see a review by Groeneboom and Jongbloed, 2014), and the second equality follows from a change of variables $s = x_L + t(q_n^* - x_L)$ and its implication, $s \geq q_n^* \Leftrightarrow t \geq 1$. Let $\hat{U}(y,w) = y - w'\hat{\theta}_{\text{OLS}}$ and

$$g_{n,t}(y,w) = n^{1/6}\{\tau_0(x_L) - \hat{U}(y,w)^2\}\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(x).$$

We decompose

$$\begin{aligned} Z_{n1}(t) &= \sqrt{n}(\mathbb{P}_n - P)g_{n,t} + n^{2/3}E[\{\tau_0(x_L) - \hat{U}(Y,W)^2\}\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(X)] \\ &\quad + n^{1/3}m\{F_n(x_L + t(q_n^* - x_L)) - F(x_L + t(q_n^* - x_L))\} + n^{1/3}mF(x_L + t(q_n^* - x_L)) \\ &=: Z_{n1}^a(t) + Z_{n1}^b(t) + Z_{n1}^c(t) + Z_{n1}^d(t). \end{aligned}$$

Analysis of $Z_{n1}^a(t)$. We verify the conditions of van der Vaart (2000, Theorem 19.28). Define the class of random functions (depending on $\hat{\theta}_{\text{OLS}}$):

$$\mathcal{G}_{n1} = \{g_{n,t}(y,w) = n^{1/6}(\tau_0(x_L) - \hat{U}(y,w)^2)\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(x) : t \in [0,K]\},$$

for $K \in (0,\infty)$, where n in the subscript indicates the dependence on both the scaling parameter $n^{1/6}$ and $\hat{\theta}_{\text{OLS}}$. By van der Vaart (2000, Example 19.6) we know that for a bracket size ϵ, $\mathcal{G}_{n1}$ has the entropy with bracketing of order $\log(1/\epsilon)$. Thus, $\mathcal{G}_{n1}$ satisfies the entropy condition for van der Vaart (2000, Theorem 19.28).

For each $t, s \in [0, K]$, note that

$$\begin{aligned}
\mathrm{Cov}(g_{n,t}, g_{n,s}) &= n^{1/3}E[\{\hat{U}(Y,W)^2 - \tau_0(x_L)\}^2\mathbb{I}_{[x_L, x_L+(t\wedge s)(q_n^*-x_L)]}(X)] + o_p(1) \\
&= n^{1/3}E[\{U^2 - \tau_0(x_L)\}^2\mathbb{I}_{[x_L, x_L+(t\wedge s)(q_n^*-x_L)]}(X)] + o_p(1) \\
&= n^{1/3}E[[\varepsilon^2 + \{\tau_0(X) - \tau_0(x_L)\}^2]\mathbb{I}_{[x_L, x_L+(t\wedge s)(q_n^*-x_L)]}(X)] + o_p(1) \\
&= n^{1/3}\int_{x_L}^{x_L+(t\wedge s)(q_n^*-x_L)} [\sigma_\varepsilon^2(x) + \{\tau_0(x) - \tau_0(x_L)\}^2]f_X(x)dx + o_p(1) \\
&= [\sigma_\varepsilon^2(\xi_n) + \{\tau_0(\xi_n) - \tau_0(x_L)\}^2]f_X(\xi_n)c^*(t\wedge s) + o_p(1) \\
&= \sigma_\varepsilon^2(x_L)f_X(x_L)c^*(t\wedge s) + o_p(1), \qquad (A.5)
\end{aligned}$$

for $\xi_n \in (x_L, x_L + (t \wedge s)q_n^*)$. The first equality follows from $q_n^* - x_L = O(n^{-1/3})$. In the second equality, we replace the estimated $\hat{U}^2$ with the unobservable U^2. By (A.3), the discrepancy between $\hat{U}^2$ and U^2 converges more rapidly than $n^{-1/3}$, and the factor $\mathbb{I}_{[x_L, x_L+(t\wedge s)(q_n^*-x_L)]}(X)$ further refines this rate. Consequently, under Assumptions A1 and A2, the impact of substituting $\hat{U}^2$ with U^2 in the second line is $o_p(1)$. The third equality follows from the definition $\varepsilon = U^2 - \tau_0(X)$ and $E[\varepsilon|X] = 0$, the fourth equality follows from the law of iterated expectations, the fifth equality follows from a Taylor expansion, and the last equality follows from $c^* = \lim_{n\to\infty} n^{1/3}(q_n^* - x_L)$ and the continuity of $\sigma_\varepsilon^2(\cdot)$ and $\tau_0(\cdot)$ at x_L from right. Similarly, we have $\mathrm{Var}(g_{n,t}) = \sigma_\varepsilon^2(x_L)f_X(x_L)c^*t + o_p(1)$.

We next consider the envelop function of the class $\mathcal{G}_{n1}$, that is

$$G_{n1}(y, w) = n^{1/6}|\tau_0(x_L) - \hat{U}(y, w)^2| \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(x).$$

We can see that G_{n1} is square integrable since similar arguments to (A.5) yield

$$\begin{aligned}
E[G_{n1}^2(Y,W)] &= n^{1/3}E[|\tau_0(x_L) - \hat{U}(Y,W)^2| \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(X)] \\
&= n^{1/3}E[|\tau_0(x_L) - U^2| \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(X)] + o_p(1) \\
&= n^{1/3}E[[\varepsilon^2 + \{\tau_0(X) - \tau_0(x_L)\}^2] \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(X)] + o_p(1) \\
&= n^{1/3}\int_{x_L}^{x_L+K(q_n^*-x_L)} [\sigma_\varepsilon^2(x) + \{\tau_0(x) - \tau_0(x_L)\}^2]f_X(x)dx + o_p(1) \\
&= O_p(1), \qquad (A.6)
\end{aligned}$$

and thus the Lindeberg condition can be verified by Assumption A2: for any $\zeta > 0$ and some $\delta > 0$,

$$\begin{aligned}
E[G_{n1}^2\mathbb{I}\{G_{n1} > \zeta\sqrt{n}\}] &\leq \frac{n^{(2+\delta)1/6}}{\zeta^\delta n^{\delta/2}}E[|\tau_0(x_L) - \hat{U}(Y,W)^2|^{2+\delta} \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(X)] \\
&= \frac{n^{(2+\delta)1/6}}{\zeta^\delta n^{\delta/2}}E[|\tau_0(x_L) - U^2|^{2+\delta} \cdot \mathbb{I}_{[x_L, x_L+K(q_n^*-x_L)]}(X)] + o_p(1) \\
&= O(n^{-\delta/3}) + o_p(1) = o_p(1), \qquad (A.7)
\end{aligned}$$

where the inequality follows from the same arguments that are used in the proof of Markov's inequality, the first equality follows from $\hat{\theta}_{\mathrm{OLS}} - \theta = O_p(n^{-1/2})$ and Assumptions A1-A2, and the second equality follows from a similar argument to (A.6).

Furthermore, as $\delta_n \to 0$, we obtain

$$\begin{aligned}
\sup_{|t-s|\le\delta_n} E|g_{n,t}-g_{n,s}|^2 &= n^{1/3}\sup_{|t-s|\le\delta_n} E[\{\hat{U}(Y,W)^2-\tau_0(x_L)\}^2\mathbb{I}_{[x_L,x_L+|t-s|q_n^*]}(X)] \\
&= n^{1/3}\sup_{|t-s|\le\delta_n} E[[\varepsilon^2+\{\tau_0(X)-\tau_0(x_L)\}^2]\cdot\mathbb{I}_{[x_L,x_L+|t-s|q_n^*]}(X)]+o_p(\delta_n) \\
&= O_p(\delta_n)=o_p(1). \qquad (A.8)
\end{aligned}$$

By (A.5)-(A.8), we can apply van der Vaart (2000, Theorem 19.28), which implies for each $K\in(0,\infty)$,

$$Z_{n1}^a(t)\xrightarrow{d}\sqrt{\sigma_\varepsilon^2(x_L)f_X(x_L)c^*}\mathcal{W}_t \text{ in } l^\infty[0,K]. \qquad (A.9)$$

Analysis of $Z_{n1}^b(t)$. Observe that

$$\begin{aligned}
Z_{n1}^b(t) &= n^{2/3}E[\{\tau_0(x_L)-U^2\}\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(X)]+n^{2/3}E[(U^2-\hat{U}(Y,W)^2)\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(X)] \\
&= n^{2/3}\int_{x_L}^{x_L+t(q_n^*-x_L)}\{\tau_0(x_L)-\tau_0(F^{-1}(F(x)))\}dF(x)+o_p(1) \\
&= n^{2/3}\int_{F(x_L)}^{F(x_L+t(q_n^*-x_L))}\{\tau_0(x_L)-\tau_0(F^{-1}(v))\}dv+o_p(1) \\
&= -n^{2/3}\int_{F(x_L)}^{F(x_L+t(q_n^*-x_L))}\tau_0'(x_L)\{F^{-1}(v)-F^{-1}(F(x_L))\}dv+o_p(1) \\
&= -n^{2/3}\int_{F(x_L)}^{F(x_L+t(q_n^*-x_L))}\tau_0'(x_L)\frac{v-F(x_L)}{f_X(x_L)}dv+o_p(1) \\
&= -n^{2/3}\tau_0'(x_L)\frac{\{F(x_L+t(q_n^*-x_L))-F(x_L)\}^2}{2f_X(x_L)}+o_p(1) \\
&= -\tau_0'(x_L)\frac{t^2(c^*)^2}{2}f_X(x_L)+o_p(1) \qquad (A.10)
\end{aligned}$$

holds uniformly over $t\in[0,K]$, where the second equality follows from $E[\{U^2-\hat{U}(Y,W)^2\}\cdot\mathbb{I}_{[x_L,x_L+t(q_n^*-x_L)]}(X)]=o_p(n^{-2/3})$, the third equality follows from a change of variables $v=F(x)$, the fourth equality follows from a Taylor expansion, the fifth equality follows from $F^{-1}(v)-x_L=\frac{1}{f_X(x_L)}(v-F(x_L))+o(v-F(x_L))$, the sixth equality follows from evaluating the integral, and the last equality follows from a Taylor expansion and $c^*=\lim_{n\to\infty}n^{1/3}(q_n^*-x_L)$.

Analysis of $Z_{n1}^c(t)$. By Kim and Pollard (1990, Maximal inequality 3.1),

$$E\left[\sup_{t\in[0,K]}|F_n(x_L+t(q_n^*-x_L))-F(x_L+t(q_n^*-x_L))|\right]\le n^{-1/2}J\sqrt{PG_n^2}$$

holds for some constant $J\in(0,\infty)$. Here G_n is the envelope of the set of indicator functions, thus $PG_n^2\le 1$. As a result,

$$Z_{n1}^c(t)\le n^{1/3}n^{-1/2}mJ\sqrt{PG_n^2}=o(1), \qquad (A.11)$$

uniformly over $t\in[0,K]$.

Analysis of $Z_{n1}^d(t)$. A Taylor expansion yields

$$Z_{n1}^d(t)=n^{1/3}mF(x_L+t(q_n^*-x_L))=m\cdot t\cdot f_X(x_L)c^*+o(1), \qquad (A.12)$$

uniformly over $t \in [0, K]$, for every $K < \infty$.

Combining (A.9)-(A.12), it holds that for each $0 < K < \infty$,

$$Z_{n1}(t) \xrightarrow{d} Z_1(t) := \sqrt{\sigma_\varepsilon^2(x_L) f_X(x_L) c^*}\mathcal{W}_t - \tau_0'(x_L)\frac{t^2(c^*)^2}{2} f_X(x_L) + m \cdot t \cdot f_X(x_L) c^* \text{ in } l^\infty[0, K]. \tag{A.13}$$

We now verify the conditions of the argmax continuous mapping theorem (Kim and Pollard, 1990). Note that for each $t \neq s$,

$$\text{Var}(Z_1(s) - Z_1(t)) = \sigma_\varepsilon^2(x_L) f_X(x_L) c^* |t - s| \neq 0.$$

By Kim and Pollard (1990), the process $t \to Z_1(t)$ achieves its maximum a.s. at a unique point. Consider extended versions of Z_{n1} and Z_1 to the real line:

$$\tilde{Z}_{n1}(t) = \begin{cases} Z_{n1}(t), & t \geq 0 \\ t & t < 0 \end{cases}, \qquad \tilde{Z}_1(t) = \begin{cases} Z_1(t), & t \geq 0 \\ t & t < 0 \end{cases}.$$

It holds $\tilde{Z}_{n1}(t) \xrightarrow{d} \tilde{Z}_1(t)$, and the similar argument to Lemma SM.2.1 (ii) in Babii and Kumar (2023) yields that the maximum of $\tilde{Z}_{n1}(t)$ is uniformly tight. Therefore, by Kim and Pollard (1990, Theorem 2.7),

$$\begin{aligned} & P\left(n^{1/3}\{\hat{\tau}(q_n^*) - \tau_0(x_L)\} \leq m\right) \to P\left(\left[\arg\max_{t \geq 0} Z_1(t)\right] \geq 1\right) \\ = \;& P\left(\left[\arg\max_{t \geq 0} \sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t - \tau_0'(x_L)\frac{t^2 c^*}{2} + mt\right] \geq 1\right) \\ = \;& P\left(\left[D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2}\right)(1)\right] \leq m\right), \end{aligned}$$

where the second equality follows from the switch relation and symmetry of the process $\mathcal{W}_t$. Thus, we have

$$n^{1/3}\{\hat{\tau}(q_n^*) - \tau_0(x_L)\} \xrightarrow{d} D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2}\right)(1), \tag{A.14}$$

which also implies

$$\begin{aligned} & n^{1/3}\{\hat{\tau}(q_n^*) - \tau_0(q_n^*)\} \\ \xrightarrow{d} \;& D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2}\right)(1) - \lim_{n \to \infty} n^{1/3}\{\tau_0(q_n^*) - \tau_0(x_L)\} \\ \overset{d}{\sim} \;& D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2} - \tau_0'(x_L) c^* t\right)(1), \end{aligned} \tag{A.15}$$

where the distribution relation follows from the fact that the $D^L_{[0,\infty)}$ is a linear operator for a linear function of t.

Finally, we analyze $n^{1/3}\{\hat{\tau}(q_n) - \tau_0(q_n)\}$. Recall q_n is the $(n^{-1/3})$-th sample quantile of X. Assumption A3 guarantees $q_n - q_n^* = O_p(n^{-1/2}) = o_p(n^{-1/3})$, which also implies $\text{plim}_{n \to \infty} n^{1/3}(q_n - x_L) = \lim_{n \to \infty} n^{1/3}(q_n^* - x_L) = c^*$. Thus, the same argument for (A.14) can be applied to show that the result in (A.14) holds true even if we replace q_n^* with q_n. Therefore, the conclusion follows.

A.2. **Proof of Theorem 1.** By the definitions of the estimators, it holds that

$$\sqrt{n}(\hat{\theta}-\theta) = \left(\frac{1}{n}\sum_{i:x_i>q_n}\hat{\sigma}_i^{-2}W_iW_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i:x_i>q_n}\hat{\sigma}_i^{-2}W_iU_i\right),$$

$$\sqrt{n}(\hat{\theta}_{\text{IGLS}}-\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^{-2}W_iW_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sigma_i^{-2}W_iU_i\right).$$

Thus it is sufficient for the conclusion to show

$$T_1 := \frac{1}{\sqrt{n}}\sum_{i:x_i>q_n}\hat{\sigma}_i^{-2}W_iU_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sigma_i^{-2}W_iU_i \xrightarrow{p} 0,$$

$$T_2 := \frac{1}{n}\sum_{i:x_i>q_n}\hat{\sigma}_i^{-2}W_iW_i' - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^{-2}W_iW_i' \xrightarrow{p} 0.$$

A.2.1. *The concentration of* T_1. Decompose

$$T_1 = \frac{1}{\sqrt{n}}\sum_{i:x_i>q_n}(\hat{\sigma}_i^{-2}-\sigma_i^{-2})W_iU_i - \frac{1}{\sqrt{n}}\sum_{i:x_i\le q_n}\sigma_i^{-2}W_iU_i =: T_{11}-T_{12}.$$

We first consider T_{12}. For any $h \in \{1:\dim(W)\}$, let W_i^h and T_{12}^h be the h-th element of W_i and T_{12}, respectively. Note that $E[T_{12}^h|q_n]=0$ by $E[U|W]=0$. Also we have $\text{Var}(T_{12}^h|q_n)\xrightarrow{p} 0$. To see this, decompose

$$\text{Var}(T_{12}^h|q_n) = I_h - n\cdot(II_h)^2,$$

where $I_h = \frac{1}{n}E\left[\left(\sum_{i=1}^n \mathbb{I}\{X_i\le q_n\}\sigma_i^{-2}W_i^hU_i\right)^2\Big| q_n\right]$ and $II_h = E[\mathbb{I}\{X_i\le q_n\}\sigma_i^{-2}W_i^hU_i|q_n]$. For I_h, note that

$$\begin{aligned} I_h &= \frac{1}{n}E\left[E\left[\left(\sum_{i=1}^n \mathbb{I}\{X_i\le q_n\}\sigma_i^{-2}W_i^hU_i\right)^2\Bigg|\mathbf{W}\right]\Bigg| q_n\right] \\ &= E[E[(\mathbb{I}\{X_i\le q_n\}\sigma_i^{-2}W_i^hU_i)^2|\mathbf{W}]|q_n] = E[\mathbb{I}\{X\le q_n\}\sigma^{-2}(X)(W^h)^2|q_n] \\ &\le R^2\sigma^{-2}(x_L)E[\mathbb{I}\{X\le q_n\}|q_n]\xrightarrow{p} 0, \end{aligned}$$

where $\mathbf{W}=(W_1,\dots,W_n)'$. The first equality follows from the law of iterated expectation and the fact that q_n is a function of $\mathbf{W}$, the second equality follows from $E[U|W]=0$ and $\{U_i\}_{i=1}^n$ being iid, the third equality follows because conditional on $\mathbf{W}$, $\mathbb{I}\{X_i\le q_n\}(\sigma_i^{-2}W_i^h)^2$ is treated as fixed, the inequality follows from Assumptions A1 and A2, and the convergence follows from $q_n\xrightarrow{p} x_L$. For II_h, note that

$$II_h = E[\mathbb{I}\{X_i\le q_n\}\sigma_i^{-2}W_i^hE[U_i|\mathbf{W}]|q_n]=0,$$

where the first equality follows from the law of iterated expectation and the fact that q_n is a function of $\mathbf{W}$, and the second equality follows from $E[U_i|\mathbf{W}]=E[U_i|W_i]=0$. Since $E[T_{12}^h|q_n]=0$ and $\text{Var}(T_{12}^h|q_n)\xrightarrow{p} 0$ hold for every h, we can conclude that $T_{12}\xrightarrow{p} 0$.

To proceed, we will utilize Lemma 3 below. Its proof can be found at the end of Appendix A.2. Recall that earlier in this appendix, we relabel $\sigma^2(\cdot)$ as $\tau_0(\cdot)$, and $\hat{\tau}$ is used to denote the isotonic estimator of $\sigma^2(\cdot)$. Additionally, with some abuse of notation, we use w_h to denote the h-th element of vector w.

Lemma 3. *Under Assumptions A1-A3,*

- **(i):** $\|\hat{\tau}\|_\infty = O_p(\log n)$,
- **(ii):** $\|\hat{\tau} - \tau_0\|_{2,P}^2 = O_p((\log n)^2 n^{-2/3})$,
- **(iii):** $E[\|\mathbb{G}_n\|_{\mathcal{F}_n}] \leq \frac{A\nu}{2}$ *holds for any constants* $A > 0$ *and* $\nu > 0$, *and all sufficiently large* n, *where* $\mathcal{F}_n$ *is the function class defined as*

$$\mathcal{F}_n = \left\{ f_n(w,u) = \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau(x)} - \frac{1}{\tau_0(x)}\right) w_h u : \begin{array}{c} \tau \geq 0 \text{ is monotone increasing on } \mathscr{X}, \\ \|\tau\|_\infty \leq C\log n, \ \|\tau - \tau_0\|_{2,P}^2 \leq Cr_n, \\ \mathbb{I}\{x > q_n\}/\tau(x) \leq 1/K_0, h \in \{1 : \dim(w)\} \end{array} \right\}, \tag{A.16}$$

with C *and* K_0 *being some positive constants, and* $r_n = (\log n)^2 n^{-2/3}$.

Now we focus on T_{11}. Since the proof is similar, we only present the proof for the h-th element of T_{11}, i.e., for any constant $A > 0$,

$$P\{|\mathbb{G}_n \hat{f}| \geq A\} \to 0, \tag{A.17}$$

where $\hat{f}(w,u) = \mathbb{I}\{x > q_n\}\left(\frac{1}{\hat{\tau}(x)} - \frac{1}{\tau_0(x)}\right) w_h u$. To this end, we set $\tau_0(x_L) = C_0 = 2K_0 > 0$. It holds that for any $A > 0$ and $\nu > 0$, there exists a positive constant C such that

$$\begin{aligned} P\{|\mathbb{G}_n \hat{f}| \geq A\} &\leq P\left\{|\mathbb{G}_n \hat{f}| \geq A, \ \|\hat{\tau}\|_\infty \leq C\log n, \ \|\hat{\tau} - \tau_0\|_{2,P}^2 \leq Cr_n, \ \frac{\mathbb{I}\{x > q_n\}}{\hat{\tau}(x)} \leq \frac{1}{K_0}\right\} + \frac{\nu}{2} \\ &\leq \frac{E\left[\|\mathbb{G}_n\|_{\mathcal{F}_n}\right]}{A} + \frac{\nu}{2} \leq \nu, \end{aligned} \tag{A.18}$$

for all sufficiently large n, where the first inequality follows from Lemma 1 and Lemma 3 (i)-(ii), and the fact that $\hat{\tau}$ is monotone increasing (so that the lower bound at the truncation point is the uniform lower bound). Specifically, for any $\nu > 0$, we can find $C > 0$ and a positive integer n_0 such that for any integer $n > n_0$, it holds that (a) $P\{\|\hat{\tau}\|_\infty > C\log n\} < \frac{\nu}{6}$, (b) $P\{\|\hat{\tau} - \tau_0\|_{2,P}^2 > Cr_n\} < \frac{\nu}{6}$, and (c) $P\left\{\frac{\mathbb{I}\{x>q_n\}}{\hat{\tau}(x)} > \frac{1}{K_0}\right\} < \frac{\nu}{6}$. Parts (a) and (b) are ensured by Lemma 3 (i) and (ii), respectively; part (c) is guaranteed by Lemma 1. As a result, $P\left(\{\|\hat{\tau}\|_\infty > C\log n\} \text{ or } \{\|\hat{\tau} - \tau_0\|_{2,P}^2 > Cr_n\} \text{ or } \left\{\frac{\mathbb{I}\{x>q_n\}}{\hat{\tau}(x)} > \frac{1}{K_0}\right\}\right) < \frac{\nu}{2}$. In the case of $\lim_{x \downarrow x_L} \frac{d\sigma^2(x)}{dx} = 0$, part (c) remains valid since $\hat{\tau}(x)$ will converge to τ_0 at a faster rate (the $\sqrt{n}$-rate), then the first inequality of (A.18) holds without invoking Lemma 1.

The second inequality of (A.18) follows from Markov's inequality and the definition of $\mathcal{F}_n$, which is given by (A.16). The last inequality follows from Lemma 3 (iii). Since ν can be arbitrarily small, we obtain (A.17) and the conclusion follows.

A.2.2. *Proof of $T_2 = o_p(1)$.* Note that

$$\begin{aligned} T_2 &= \frac{1}{n}\sum_{i:x_i>q_n} \hat{\sigma}_i^{-2} W_i W_i' - \frac{1}{n}\sum_{i=1}^{n} \sigma_i^{-2} W_i W_i' \\ &= \frac{1}{n}\sum_{i:x_i>q_n} (\hat{\sigma}_i^{-2} - \sigma_i^{-2}) W_i W_i' - \frac{1}{n}\sum_{i:x_i\le q_n} \sigma_i^{-2} W_i W_i' \\ &=: T_{21} - T_{22}. \end{aligned}$$

First, we have $T_{22} \xrightarrow{p} 0$ since $q_n \xrightarrow{p} x_L$. For T_{21}, let s_n be the $(1-n^{-1/3})$-th sample quantile of $\{X_i\}_{i=1}^n$. By employing arguments similar to those in the proof of Lemma 1, we have $\hat{\sigma}^2(s_n) - \sigma^2(s_n) = O_p(n^{-1/3})$. Using reasoning akin to, yet simpler than, those in the proof of Lemma 1, we can establish that for any $x \in (q_n, s_n)$, it holds that $\hat{\sigma}^2(x) - \sigma^2(x) = O_p(n^{-1/3})$. Combining the aforementioned results with the monotonicity of both $\hat{\sigma}^2(\cdot)$ and $\sigma^2(\cdot)$, we can conclude that $\sup_{x\in[q_n,s_n]} |\hat{\sigma}^2(x) - \sigma^2(x)| = O_p(n^{-1/3})$, i.e., $\hat{\sigma}^2(x)$ is uniformly consistent within trimmed domain $[q_n, s_n]$ (the proof here resembles the one given for the Glivenko-Cantelli Theorem regarding the uniform consistency of the empirical distribution function; see, for example, the proof of Theorem 19.1 in van der Vaart, 2000). Therefore, we have

$$T_{21} = \frac{1}{n}\sum_{i:q_n<x_i<s_n} (\hat{\sigma}_i^{-2} - \sigma_i^{-2}) W_i W_i' + \frac{1}{n}\sum_{i:x_i\ge s_n} (\hat{\sigma}_i^{-2} - \sigma_i^{-2}) W_i W_i' = o_p(1), \tag{A.19}$$

where the second equality follows from the preceding argument, $|s_n - x_U| = O_p(n^{-1/3})$, and Lemma 3 (i). Combining $T_{22} \xrightarrow{p} 0$ and (A.19), we have $T_2 \xrightarrow{p} 0$.

A.2.3. *Proof of Lemma 3 (i).* The min-max formula of the isotonic regression says

$$\min_{1\le k\le n} \frac{\sum_{j=1}^k \hat{U}_j^2}{k} \le \hat{\tau}(x) \le \max_{1\le k\le n} \frac{\sum_{j=k}^n \hat{U}_j^2}{n-k+1},$$

for each $x \in \mathscr{X}$, which implies $\min_{1\le j\le n} \hat{U}_j^2 \le \hat{\tau}(x) \le \max_{1\le j\le n} \hat{U}_j^2$ for each $x \in \mathscr{X}$. Thus, it is sufficient for the conclusion to show that

$$\max_{1\le j\le n} \hat{U}_j^2 = O_p(\log n). \tag{A.20}$$

Observe that

$$\max_{1\le j\le n} \hat{U}_j^2 \le \max_{1\le j\le n} U_j^2 + 2Rk||\hat{\theta}_{\text{OLS}} - \theta||_\infty \max_{1\le j\le n} |U_j| + R^2k^2||\hat{\theta}_{\text{OLS}} - \theta||_\infty^2.$$

From BDJ (2019, eq. (7.11) on p.3297), Assumption A2 guarantees $\max_{1\le j\le n} U_j^2 = O_p(\log n)$. Since $\hat{\theta}_{\text{OLS}}$ is the OLS estimator, it holds that $||\hat{\theta}_{\text{OLS}} - \theta||_\infty = O_p(n^{-1/2})$. By (A.2), we also have $\max_{1\le j\le n} |U_j| = O_p(\log n)$. Combining these results with Assumption A1, we have (A.20).

A.2.4. *Proof of Lemma 3 (ii).* The proof is based on that of Proposition 4 of BGH (p.8 of BGH-supp). Recall that $\hat{\tau}(\cdot)$ is the solution of $\min_{\tau\in\{\text{all monotone funcitons}\}} \sum_{j=1}^n \{\hat{U}_j^2 - \tau(X_j)\}^2$, or equivalently

$$\max_{\tau\in\{\text{all monotone funcitons}\}} \sum_{j=1}^n \{2\hat{U}_j^2 \tau(X_j) - \tau(X_j)\}. \tag{A.21}$$

On the other hand, $\tau_0(\cdot)$ is the solution of $\min_{\tau\in\{\text{all monotone funcitons}\}} E[\{U^2-\tau(X)\}^2]$, or equivalently

$$\max_{\tau\in\{\text{all monotone funcitons}\}} E[2U^2\tau(X)-\tau(X)^2]. \tag{A.22}$$

By (A.21), it holds

$$\sum_{j=1}^n\{2\hat{U}_j^2\hat{\tau}(X_j)-\hat{\tau}(X_j)^2\}\geq\sum_{j=1}^n\{2\hat{U}_j^2\tau_0(X_j)-\tau_0(X_j)^2\},$$

or equivalently (by plugging in $\hat{U}_j=U_j-W_j'(\hat{\theta}_{\text{OLS}}-\theta)$),

$$\begin{aligned}&\sum_{j=1}^n\{2U_j^2\hat{\tau}(X_j)-\hat{\tau}(X_j)^2\}+2\sum_{j=1}^n\left(-2U_jW_j'(\hat{\theta}_{\text{OLS}}-\theta)+\{W_j'(\hat{\theta}_{\text{OLS}}-\theta)\}^2\right)\{\hat{\tau}(X_j)-\tau_0(X_j)\}\\ \geq\;&\sum_{j=1}^n\{2U_j^2\tau_0(X_j)-\tau_0(X_j)^2\}.\end{aligned} \tag{A.23}$$

Define $d_2^2(\tau_1,\tau_2)=-E[2\tau_1\tau_2-\tau_1^2-\tau_2^2]$. Note that for any monotone function τ,

$$\begin{aligned}&E[2U^2\tau(X)-\tau(X)^2]-E[2U^2\tau_0(X)-\tau_0(X)^2]\\ =\;&E[2E[U^2|X]\tau(X)-\tau(X)^2-2E[U^2|X]\tau_0(X)+\tau_0(X)^2]\\ =\;&E[2\tau_0(X)\tau(X)-\tau(X)^2-\tau_0(X)^2]=-d_2^2(\tau,\tau_0),\end{aligned} \tag{A.24}$$

where the first equality follows from the law of iterated expectation, the second equality follows from the definition $\tau_0(x)=E[U^2|X=x]$, and the last equality follows from the definition of $d_2^2(\cdot,\cdot)$.

Define

$$\begin{aligned}g_\tau(u,x)&=\{2u^2\tau(x)-\tau(x)^2\}-\{2u^2\tau_0(x)-\tau_0(x)^2\},\\ R_n&=\frac{2}{n}\sum_{j=1}^n\left(-2U_jW_j'(\hat{\theta}_{\text{OLS}}-\theta)+\{W_j'(\hat{\theta}_{\text{OLS}}-\theta)\}^2\right)\{\hat{\tau}(X_j)-\tau_0(X_j)\}.\end{aligned}$$

From (A.23) and (A.24), it holds

$$\int g_{\hat{\tau}}(u,x)d(\mathbb{P}_n-P)(u,x)+R_n\geq d_2^2(\hat{\tau},\tau_0). \tag{A.25}$$

Note that R_n is bounded as

$$|R_n|\leq\left|-(\hat{\theta}_{\text{OLS}}-\theta)'\frac{4}{n}\sum_{j=1}^nW_jU_j\{\hat{\tau}(X_j)-\tau_0(X_j)\}\right|+\left|\frac{2}{n}\sum_{j=1}^n\{W_j'(\hat{\theta}_{\text{OLS}}-\theta)\}^2\{\hat{\tau}(X_j)-\tau_0(X_j)\}\right|.$$

The second term is of order $O_p(n^{-1}\log n)$ (because $\hat{\theta}_{\text{OLS}}-\theta=O_p(n^{-1/2})$ and Lemma 3 (i)). By similar arguments in p.22 of BGH-supp and in the proof of Lemma 3 (i), the first term is of order $O_p\left(n^{-1}(\log n)^2\right)$.

Then

$$R_n=O_p\left(n^{-1}(\log n)^2\right). \tag{A.26}$$

Thus, for some constants $C, K > 0$ and a shrinking sequence ϵ_n, set inclusion relationships yield

$$\begin{aligned}
P(d_2^2(\hat{\tau}, \tau_0) \geq \epsilon_n^2) &= P\left(d_2(\hat{\tau}, \tau_0) \geq \epsilon_n, \ \int g_{\hat{\tau}}(u,x) d(\mathbb{P}_n - P)(u,x) + R_n \geq d_2^2(\hat{\tau}, \tau_0)\right) \\
&\leq P\left(\begin{array}{c} d_2(\hat{\tau}, \tau_0) \geq \epsilon_n, \ |R_n| \leq Cn^{-1}(\log n)^2, \ \|\hat{\tau}\|_\infty \leq K \log n \\ \int g_{\hat{\tau}}(u,x) d(\mathbb{P}_n - P)(u,x) + R_n - d_2^2(\hat{\tau}, \tau_0) \geq 0 \end{array}\right) \\
&\quad + P(|R_n| > Cn^{-1}(\log n)^2) + P(\|\hat{\tau}\|_\infty > K \log n) \\
&=: P_1 + P_2 + P_3,
\end{aligned}$$

where the first equality follows from (A.25). For P_2 and P_3, (A.26) and Lemma 3 (i) imply that we can choose C and K to make these terms arbitrarily small. Thus, we focus on the first term P_1.

Now let

$$\begin{aligned}
\mathcal{T} &= \{\tau : \tau \text{ is positive and monotone increasing on } \mathscr{X}, \ \|\tau\|_\infty \leq K \log n\}, \\
\mathcal{G} &= \{g_\tau(u,x) = \{2u^2\tau(x) - \tau(x)^2\} - \{2u^2\tau_0(x) - \tau_0(x)^2\} : \ \tau \in \mathcal{T}\}, \\
\mathcal{G}_v &= \{g \in \mathcal{G} : d_2(\tau, \tau_0) \leq v\}.
\end{aligned}$$

Set inclusion relationships and Markov's inequality yield

$$\begin{aligned}
P_1 &\leq P\left(\sup_{\tau \in \mathcal{T}, d_2(\tau,\tau_0) \geq \epsilon_n} \{\int g_\tau(u,x) d(\mathbb{P}_n - P)(u,x) - d_2^2(\tau, \tau_0)\} \geq -Cn^{-1}(\log n)^2\right) \\
&\leq \sum_{s=0}^\infty P\left(\sup_{\tau \in \mathcal{T}, 2^s\epsilon_n \leq d_2(\tau,\tau_0) \leq 2^{s+1}\epsilon_n} \sqrt{n}\{\int g_\tau(u,x) d(\mathbb{P}_n - P)(u,x)\} \geq \sqrt{n}\left(2^{2s}\epsilon_n^2 - Cn^{-1}(\log n)^2\right)\right) \\
&\leq \sum_{s=0}^\infty P\left(\|\mathbb{G}_n g\|_{\mathcal{G}_{2^{s+1}\epsilon_n}} \geq \sqrt{n}\left(2^{2s}\epsilon_n^2 - Cn^{-1}(\log n)^2\right)\right) \\
&\leq \sum_{s=0}^\infty E[\|\mathbb{G}_n g\|_{\mathcal{G}_{2^{s+1}\epsilon_n}}] / \{\sqrt{n}\left(2^{2s}\epsilon_n^2 - Cn^{-1}(\log n)^2\right)\}.
\end{aligned}$$

For a sufficiently large constant $\tilde{C} > 0$, the sequence $\epsilon_n^2 := \tilde{C}(\log n)^2 n^{-\frac{2}{3}}$ dominates $Cn^{-1}(\log n)^2$, so it holds $\sqrt{n}\left(2^{2s}\epsilon_n^2 - Cn^{-1}(\log n)^2\right) = \sqrt{n}2^{2s}\epsilon_n^2(1 + o(1))$. Therefore, the standard result for the L^2-convergence of the isotonic estimator under Assumption A2 (e.g., pp. 8-11 in BGH-supp) implies that the last term can be made arbitrarily small by appropriately selecting $\tilde{C}$. Thus, the proof is concluded.

A.2.5. *Proof of Lemma 3 (iii).* We show $E[\|\mathbb{G}_n\|_{\mathcal{F}_n}] \leq \frac{A\nu}{2}$ by using van der Vaart and Wellner (1996, Lemma 3.4.3). First we introduce some notation for this part. Let $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$ be the ε-bracketing number of the function class $\mathcal{F}$ under the norm $||\cdot||$, $H_B(\varepsilon, \mathcal{F}, ||\cdot||) = \log N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$ be the entropy, $J_n(\delta, \mathcal{F}, ||\cdot||) = \int_0^\delta \sqrt{1 + H_B(\varepsilon, \mathcal{F}, ||\cdot||)} d\varepsilon$, and $\|f\|_{B,P} = (2E[e^{|f|} - |f| - 1])^{1/2}$ be the Bernstein norm.

Lemma 3.4.3 in van der Vaart and Wellner (1996): Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{B,P}^2 \leq \delta$ for every f in $\mathcal{F}$. Then

$$E[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_n(\delta, \mathcal{F}, ||\cdot||_{B,P})\{1 + J_n(\delta, \mathcal{F}, ||\cdot||_{B,P})/(\sqrt{n}\delta^2)\}.$$

To apply this lemma, we need to compute $H_B(\epsilon, \tilde{\mathcal{F}}_n, \|\cdot\|_{B,P})$ and $\|\tilde{f}\|^2_{B,P}$, where $\tilde{\mathcal{F}}_n = \{\tilde{f} = D^{-1}f : f \in \mathcal{F}_n\}$, the function class $\mathcal{F}_n$ is defined below in (A.27), and the constant $D > 0$ will be chosen later to guarantee that the Bernstein norm of $\tilde{f}$ is finite. Moreover, let us define the following function class:

$$\mathcal{T}_{\mathcal{I},K_1} = \{\tau \text{ monotone non-decreasing on the interval } \mathcal{I} \text{ and } 0 < \tau < K_1\}.$$

Assumption A2 implies that there exist positive constants, $\underline{C}$ and $\overline{C}$, such that $0 < \underline{C} < \tau_0 < \overline{C} < \infty$. Also let

$$\mathcal{F}_n = \left\{ f_n(w,u) = \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau(x)} - \frac{1}{\tau_0(x)}\right) w_h u : \begin{array}{c} \tau \in \mathcal{T}_{\mathscr{X},K_1}, \ \|\tau - \tau_0\|^2_{2,P} \le v^2, \\ \mathbb{I}\{x > q_n\}/\tau(x) \le 1/K_0, h \in \{1 : \dim(w)\} \end{array} \right\}, \tag{A.27}$$

where w_h is the h-th component of vector w. We set $2K_0 = \underline{C}$, $K_1 = K_2 \log n$, and $v = K_3(\log n)n^{-1/3}$ for some constants $K_2, K_3 > 0$.

Consider ϵ-brackets (τ^L, τ^U) under the $L_2(P)$-norm for the functions in $\mathcal{T}_{\mathcal{I},K_1}$. According to van der Vaart and Wellner (1996, Theorem 2.7.5), there exists some constant $C' > 0$ such that

$$H_B(\epsilon, \mathcal{T}_{\mathscr{X},K_1}, \|\cdot\|_{2,P}) \le \frac{C'K_1}{\epsilon}, \quad \text{for each } \epsilon \in (0, K_1). \tag{A.28}$$

Without loss of generality, we can choose those bracket functions that satisfy $\mathbb{I}\{x > q_n\}/\tau^L(x) \le 1/K_0$.[5] Define

$$\begin{aligned} f^L(w,u) &= \begin{cases} \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau^U(x)} - \frac{1}{\tau_0(x)}\right) w_h u & \text{if } w_h u \ge 0, \\ \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau^L(x)} - \frac{1}{\tau_0(x)}\right) w_h u & \text{if } w_h u < 0, \end{cases} \\ f^U(w,u) &= \begin{cases} \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau^L(x)} - \frac{1}{\tau_0(x)}\right) w_h u & \text{if } w_h u \ge 0, \\ \mathbb{I}\{x > q_n\}\left(\frac{1}{\tau^U(x)} - \frac{1}{\tau_0(x)}\right) w_h u & \text{if } w_h u < 0. \end{cases} \end{aligned}$$

Note that (f^L, f^U) is a bracket of $f \in \mathcal{F}_n$ for every $q_n \in [x_L, x_U]$.

Now we compute the bracket size of $(\tilde{f}^L, \tilde{f}^U) := (D^{-1}f^L, D^{-1}f^U)$ with respect to the Bernstein norm. Note that

$$\begin{aligned} & \|\tilde{f}^U - \tilde{f}^L\|^2_{B,P} = \|D^{-1}f^U - D^{-1}f^L\|^2_{B,P} \\ \le \ & 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \int_{\mathscr{W}\times\mathbb{R}} \left| \frac{\tau^U(x) - \tau^L(x)}{\tau^L(x)\tau^U(x)} w_h u \right|^k dP(w,u) \\ \le \ & 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \left\{ \frac{R^k k! M_0^{k-2} a_0 (2K_1)^{k-2}}{K_0^{2k}} \|\tau^U - \tau^L\|^2_{2,P} \right\} \le 2a_0 \left(\frac{R}{DK_0^2}\right)^2 \sum_{k=0}^{\infty} \left(\frac{2RM_0K_1}{DK_0^2}\right)^k \epsilon^2, \end{aligned}$$

where the first inequality follows from the definition of $\|\cdot\|^2_{B,P}$ and $\mathbb{I}\{x > q_n\} \le 1$, the second inequality follows from Assumption A2 (where we can choose $a_0, M_0 > 1$) and $\frac{\mathbb{I}\{x>q_n\}}{\tau^L(x)} \le \frac{1}{K_0}$.

[5]By definition (A.27), the $\tau(\cdot)$ associated to $\mathcal{F}_n$ must satisfy $\mathbb{I}\{x > q_n\}/\tau(x) \le 1/K_0$. Since $\mathcal{T}_{\mathscr{X},K_1}$ is a class of monotone increasing function, any ϵ-brackets of $\mathcal{T}_{\mathscr{X},K_1}$ can be modified to be a ϵ-bracket of the "$\mathcal{F}_n$-subset" of $\mathcal{T}_{\mathscr{X},K_1}$, satisfying $\mathbb{I}\{x > q_n\}/\tau(x) \le 1/K_0$ by leveling-up certain part of lower bounds functions τ^L, without changing the bracket numbers, and the size of each modified bracket can only be smaller.

Thus, by setting $D = 4M_0RK_1/K_0^2$, we obtain$\|\tilde{f}^U - \tilde{f}^L\|_{B,P}^2 \leq \frac{a_0}{4M_0^2K_1^2}\epsilon^2$, which implies

$$\|\tilde{f}^U - \tilde{f}^L\|_{B,P} \leq \tilde{K}\epsilon,$$

for $\tilde{K} = \frac{a_0^{1/2}}{2M_0K_1}$. Note that $(\tilde{f}^L, \tilde{f}^U)$ is: (a) a set of brackets in $\tilde{\mathcal{F}}_n$, (b) one-to-one induced by (τ^L, τ^U), an ϵ-bracket in $\mathcal{T}_{\mathscr{X},K_1}$ with the entropy $H_B(\epsilon, \mathcal{T}_{\mathscr{X},K_1}, \|\cdot\|_{2,P})$, and (c) $\|\tilde{f}^U - \tilde{f}^L\|_{B,P} \leq \tilde{K}\epsilon$. Based on these facts, (A.28) yields

$$H_B(\tilde{K}\epsilon, \tilde{\mathcal{F}}_n, \|\cdot\|_{B,P}) \leq H_B(\epsilon, \mathcal{T}_{\mathscr{X},K_1}, \|\cdot\|_{2,P}) \leq \frac{C'K_1}{\epsilon},$$

which implies (by a change-of-variable argument)

$$H_B(\epsilon, \tilde{\mathcal{F}}_n, \|\cdot\|_{B,P}) \leq \frac{\tilde{K}C'K_1}{\epsilon} = \frac{\tilde{B}}{\epsilon}, \quad \text{for } \tilde{B} = \frac{C'a_0^{1/2}}{2M_0}. \tag{A.29}$$

We now characterize the Bernstein norm of $\tilde{f}$,

$$\begin{aligned} \|\tilde{f}\|_{B,P}^2 &\leq 2\sum_{k=2}^{\infty}\frac{1}{k!D^k}\int_{\mathscr{W}\times\mathbb{R}}\left|\frac{\tau(x)-\tau_0(x)}{\tau(x)\tau_0(x)}w_h u\right|^k dP(w,u) \\ &\leq 2\sum_{k=2}^{\infty}\frac{1}{k!D^k}\left\{\frac{R^k k! M_0^{k-2}a_0(2K_1)^{k-2}}{K_0^{2k}}\|\tau-\tau_0\|_{2,P}^2\right\} \\ &\leq 2a_0\left(\frac{R}{DK_0^2}\right)^2\sum_{k=0}^{\infty}\left(\frac{2RM_0K_1}{DK_0^2}\right)^k v^2 \leq \frac{a_0}{4M_0^2}\frac{1}{K_1^2}v^2, \end{aligned}$$

where the second inequality follows from $\frac{\mathbb{I}\{x>q_n\}}{\tau(x)} \leq \frac{1}{K_0}$, and the third inequality follows from (A.27) and some rearrangements. Then, we have

$$\|\tilde{f}\|_{B,P} \leq \frac{Bv}{K_1}, \quad \text{for } B = \frac{a_0^{1/2}}{2M_0}. \tag{A.30}$$

Combining (A.29) and (A.30), van der Vaart and Wellner (1996, Lemma 3.4.3) implies

$$E[\|\mathbb{G}_n\|_{\tilde{\mathcal{F}}_n}] \lesssim J_n(BK_1^{-1}v)\left(1 + \frac{J_n(BK_1^{-1}v)}{\sqrt{n}B^2v^2/K_1^2}\right),$$

where $J_n(\cdot)$ is the abbreviation of $J_n(\cdot, \tilde{\mathcal{F}}_n, \|\cdot\|_{B,P})$. By the arguments used in the proof of Proposition 7.9 of BDJ, it holds

$$J_n(BK_1^{-1}v) \leq BK_1^{-1}v + 2\tilde{B}^{1/2}B^{1/2}K_1^{-1/2}v^{1/2} \lesssim B_1K_1^{-1/2}v^{1/2},$$

for some $B_1 > 0$ and sufficiently small v. This implies

$$E[\|\mathbb{G}_n\|_{\tilde{\mathcal{F}}_n}] \lesssim B_1K_1^{-1/2}v^{1/2}\left(1 + K_1^2\frac{B_1K_1^{-1/2}v^{1/2}}{\sqrt{n}B^2v^2}\right) \lesssim B_1K_1^{-1/2}v^{1/2}\left(1 + \frac{B_2K_1^{3/2}}{\sqrt{n}v^{3/2}}\right),$$

for some $B_2 > 0$. By the definition of the class $\tilde{\mathcal{F}}_n = \{\tilde{f} = D^{-1}f : f \in \mathcal{F}_n\}$, it follows that

$$E[\|\mathbb{G}_n\|_{\mathcal{F}_n}] = D\cdot E[\|\mathbb{G}_n\|_{\tilde{\mathcal{F}}_n}] \lesssim DB_1K_1^{-1/2}v^{1/2}\left(1 + \frac{B_2K_1^{3/2}}{\sqrt{n}v^{3/2}}\right) \lesssim B_3K_0^{-2}K_1^{1/2}v^{1/2}\left(1 + \frac{B_2K_1^{3/2}}{\sqrt{n}v^{3/2}}\right),$$

for some $B_3 > 0$. The conclusion follows by observing that with $v = K_3(\log n)n^{-1/3}$, $K_1 = K_2 \log n$, and all sufficiently large n, we have

$$E[\|\mathbb{G}_n\|_{\mathcal{F}_n}] \lesssim C_3(\log n)n^{-1/6}(1+C_4) \lesssim \frac{A\nu}{2},$$

where $C_3 = B_3 K_0^{-2} K_2^{1/2} K_3^{1/2}$ and $C_4 = B_2(K_2/K_3)^{3/2}$.

Appendix B. Proof of lemma and theorem in Section 3

Notation. To avoid heavy notations, some of them are used in Appendix A but redefined here. Define $\tau_\eta(a) = E[\sigma^2(X'\eta_0)|X'\eta = a]$ and $\tau_{\eta_0}(a) = \tau_0(a)$ (note that $\tau_0(x'\eta_0) = \sigma^2(x'\eta_0)$). Let $\hat{\tau}_\eta = \hat{\tau}_\eta(x'\eta)$ be the isotonic estimator obtained by (3.3) for a given η, $\mathscr{W}$ be the support of $W := (1, X', Z')'$, $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i'\hat{\eta} \le t\}$, and $M_n(t) = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 \mathbb{I}\{X_i'\hat{\eta} \le t\}$.

B.1. **Proof of Lemma 2.** The main part of the proof is similar to that of Lemma 1. Recall that q_n^* is the $(n^{-1/3})$-th population quantile of $(X'\eta_0)$ and q_n is the $(n^{-1/3})$-th sample quantile of $\{X_i'\hat{\eta}\}_{i=1}^n$ with $\hat{\eta}$ estimated by (3.4). To proceed, we use the following lemma:

Lemma 4. *Under Assumptions M1-M6, it holds*

(i): $\hat{\eta} - \eta_0 = O_p(n^{-1/2})$,

(ii): $\tau_{\hat{\eta}}(a) - \tau_0(a) = O_p(n^{-1/2})$ *for each* a, *and* $\|\tau_{\hat{\eta}} - \tau_0\|_{2,P} = O_p(n^{-1/2})$.

The proof of this lemma is in Appendix B.3. Based on Lemma 4 (i), Assumptions M2-M3, and properties of the sample quantile, we obtain $q_n - q_n^* = O_p(n^{-1/2}) = o_p(n^{-1/3})$, which implies $c^* = \lim_{n\to\infty} n^{1/3}(q_n^* - x_L) = \text{plim}_{n\to\infty} n^{1/3}(q_n - x_L) < \infty$. By Assumption M2, Lemma 4 (ii), and similar arguments in Appendix A.1, we have

$$\begin{aligned}
& n^{1/3}\{\hat{\tau}_{\hat{\eta}}(q_n) - \tau_0(q_n)\} = n^{1/3}\{\hat{\tau}_{\hat{\eta}}(q_n) - \tau_{\hat{\eta}}(q_n)\} + o_p(1) \\
= \; & n^{1/3}[\{\hat{\tau}_{\hat{\eta}}(q_n) - \tau_{\hat{\eta}}(x_L)\} - \{\tau_{\hat{\eta}}(q_n) - \tau_0(x_L)\}] + o_p(1) \\
\xrightarrow{d} \; & D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2}\right)(1) - \underset{n\to\infty}{\text{plim}}\, n^{1/3}\{\tau_0(q_n) - \tau_0(x_L)\} \\
\overset{d}{\sim} \; & D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2}\right)(1) - \lim_{n\to\infty} n^{1/3}\{\tau_0(q_n^*) - \tau_0(x_L)\} \\
\overset{d}{\sim} \; & D^L_{[0,\infty)}\left(\sqrt{\frac{\sigma_\varepsilon^2(x_L)}{c^* f_X(x_L)}}\mathcal{W}_t + \tau_0'(x_L)\frac{t^2 c^*}{2} - \tau_0'(x_L)c^* t\right)(1),
\end{aligned}$$

where the first and second equalities follow from Lemma 4 (ii), the convergence follows from a similar argument to (A.15), the first distribution relation follows from Lemma 4 (ii), Assumption M2(iv), and $q_n^* - q_n = o_p(n^{-1/3})$, and the second distribution relation follows from the fact that the $D^L_{[0,\infty)}$ is a linear operator for a linear function of t.

B.2. **Proof of Theorem 2.** Similar to Theorem 1, it is sufficient for the conclusion to prove the following lemma.

Lemma 5. *Under Assumptions M1-M6, it holds*

(i): $\|\hat{\tau}_\eta\|_\infty = O_p(\log n)$ *uniformly over* $\eta \in \mathcal{B}(\eta_0, \delta_0)$,

(ii): $\|\hat{\tau}_{\hat{\eta}} - \tau_0\|_{2,P}^2 = O_p((\log n)^2 n^{-2/3})$,

(iii): $E\left[\|\mathbb{G}_n\|_{\mathcal{F}_n}\right] \leq \frac{A\nu}{2}$ *holds for any constants* $A > 0$ *and* $\nu > 0$, *and all sufficiently large* n, *where* $\mathcal{F}_n$ *is the function class defined as*

$$\mathcal{F}_n = \left\{ f_n(w,u) = \mathbb{I}\{x'\eta > q_n\} \left(\frac{1}{\tau(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)} \right) w_h u : \begin{array}{c} \tau \geq 0 \text{ is monotone increasing on } I_\eta, \\ \|\tau\|_\infty \leq C \log n, \ \ \|\tau - \tau_\eta\|_{2,P}^2 \leq C r_n, \\ \mathbb{I}(x'\eta > q_n)/\tau(x'\eta) \leq 1/K_0, \\ h \in \{1 : \dim(w)\} \end{array} \right\},$$

with C *and* K_0 *being some positive constants, and* $r_n = (\log n)^2 n^{-2/3}$.

B.2.1. *Proof of Lemma 5 (i).* The proof is adapted from BDJ (2019, eq. (7.11) on p.3297). For fixed η, let $\{\hat{U}_{\eta,i}^2\}_{i=1}^n$ be a permutation of $\{\hat{U}_j^2\}_{j=1}^n$, which is arranged according to the monotonically ordered series $\{X_i'\eta\}_{i=1}^n$. The min-max formula of the isotonic regression says

$$\min_{1 \leq k \leq n} \frac{\sum_{i=1}^k \hat{U}_{\eta,i}^2}{k} \leq \hat{\tau}_\eta(x'\eta) \leq \max_{1 \leq k \leq n} \frac{\sum_{i=k}^n \hat{U}_{\eta,i}^2}{n-k+1},$$

for each $x \in \mathscr{X}$ and $\eta \in \mathcal{B}(\eta_0, \delta_0)$, which implies $\min_{1 \leq j \leq n} \hat{U}_j^2 \leq \hat{\tau}_\eta(x'\eta) \leq \max_{1 \leq j \leq n} \hat{U}_j^2$ for each $x \in \mathscr{X}$. Thus, it is sufficient for the conclusion to show that

$$\max_{1 \leq j \leq n} \hat{U}_j^2 = O_p(\log n). \tag{B.1}$$

Observe that

$$\max_{1 \leq j \leq n} \hat{U}_j^2 \leq \max_{1 \leq j \leq n} U_j^2 + 2Rk||\hat{\theta}_{\text{OLS}} - \theta||_\infty \max_{1 \leq j \leq n} |U_j| + R^2 k^2 ||\hat{\theta}_{\text{OLS}} - \theta||_\infty^2,$$

where k is the dimension of θ. From Lemma 7.1 of BDJ, Assumption M2 guarantees $\max_{1 \leq j \leq n} U_j^2 = O_p(\log n)$. By the same reasoning for the proof of Lemma 3, we have $\max_{1 \leq j \leq n} |U_j| = O_p(\log n)$ and $||\hat{\theta}_{\text{OLS}} - \theta||_\infty = O_p(n^{-1/2})$. Thus, we have $\|\hat{\tau}_\eta\|_\infty = O_p(\log n)$. Since different η only changes the permutation $\{\hat{U}_{\eta,i}^2\}_{i=1}^n$ but not $\max_{1 \leq j \leq n} \hat{U}_j^2$, we have $\|\hat{\tau}_\eta\|_\infty = O_p(\log n)$ uniformly over $\eta \in \mathcal{B}(\eta_0, \delta_0)$.

B.2.2. *Proof of Lemma 5 (ii).* The main part of the proof is similar to those of Lemma 3 (ii) and Proposition 4 of BGH-supp. Define

$$\begin{aligned} g_{\eta,\tau}(u,x) &= \{2u^2 \tau(x'\eta) - \tau(x'\eta)^2\} - \{2u^2 \tau_\eta(x'\eta) - \tau_\eta(x'\eta)^2\}, \\ R_{n,\eta} &= \frac{2}{n} \sum_{j=1}^n \left(-2U_j W_j(\hat{\theta}_{\text{OLS}} - \theta) + \{W_j(\hat{\theta}_{\text{OLS}} - \theta)\}^2 \right) \{\hat{\tau}_\eta(X_j'\eta) - \tau_\eta(X_j'\eta)\}, \\ d_2^2(\tau_1, \tau_2) &= -E[2\tau_1\tau_2 - \tau_1^2 - \tau_2^2], \end{aligned}$$

Following reasoning similar to that presented for (A.21)-(A.26), we have for some C and K,

$$\begin{aligned}
& P(\sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)} d_2^2(\hat{\tau}_\eta,\tau_\eta)\geq \epsilon_n^2) \\
\leq\ & P\left(\begin{array}{c} \sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)} d_2(\hat{\tau}_\eta,\tau_\eta)\geq\epsilon_n,\ \ \sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)}\|\hat{\tau}_\eta\|_\infty\leq K\log n, \\ \sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)}\int g_{\eta,\hat{\tau}}(u,x)d(\mathbb{P}_n-P)(u,x)+R_{n,\eta}-d_2^2(\hat{\tau}_\eta,\tau_\eta)\geq 0, \\ \sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)}|R_{n,\eta}|\leq Cn^{-1}(\log n)^2 \end{array}\right) \\
& +P(|R_{n,\eta}|>Cn^{-1}(\log n)^2)+P\left(\sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)}\|\hat{\tau}_\eta\|_\infty>K\log n\right) \\
=:\ & P_1+P_2+P_3.
\end{aligned}$$

Lemma 5 (i) implies $P_3\to 0$, and $P_2\to 0$ follows from similar arguments for (A.26). For P_1, we define

$$\begin{aligned}
\mathcal{T} &= \{\tau:\tau \text{ is positive and monotone increasing function on } I_\eta,\ \ \|\tau\|_\infty\leq K\log n\}, \\
\mathcal{G} &= \{g(x,u)=\{2u^2\tau(x'\eta)-\tau(x'\eta)^2\}-\{2u^2\tau_\eta(x'\eta)-\tau_\eta(x'\eta)^2\}:\ \tau\in\mathcal{T}\}, \\
\mathcal{G}_v &= \{g\in\mathcal{G}: d_2(\tau,\tau_\eta)\leq v\},
\end{aligned}$$

for each $\eta\in\mathcal{B}(\eta_0,\delta_0)$. By similar arguments for Lemma 3 (ii) and Proposition 4 of BGH-supp, we can obtain

$$P_1 \ \leq\ \sum_{s=0}^{\infty} E\left[\|\mathbb{G}_n g\|_{\mathcal{G}_{2^{s+1}\epsilon_n}}\right]/\{\sqrt{n}2^{2s}\epsilon_n^2-Cn^{-1/2}(\log n)^2\},$$

and

$$\sup_{\eta\in\mathcal{B}(\eta_0,\delta_0)}\int\{\hat{\tau}_\eta(x'\eta)-\tau_\eta(x'\eta)\}^2dF(x)=O_p((\log n)^2n^{-2/3}). \tag{B.2}$$

By combining (B.2), Lemma 4, and the triangle inequality, we obtain $\|\hat{\tau}_{\hat{\eta}}-\tau_0\|_{2,P}^2=O_p((\log n)^2n^{-2/3})$.

Proof of Lemma 5 (iii). To avoid heavy notation, we use the same notation as in the proof of Lemma 3 (iii), but some notation is redefined here. Let

$$\mathcal{T}_{\mathcal{I},K_1}=\{\tau \text{ monotone non-decreasing on some interval } \mathcal{I} \text{ and } 0<\tau<K_1\}.$$

Assumption M2 guarantees $0<\underline{C}<\tau_0<\overline{C}<\infty$. Similar to the proof of Lemma 3 (iii), we calculate $H_B(\epsilon,\tilde{\mathcal{F}},\|\cdot\|_{B,P})$ and $\|\tilde{f}\|_{B,P}^2$, with $\tilde{\mathcal{F}}=\{\tilde{f}=D^{-1}f: f\in\mathcal{F}\}$, where the constant $D>0$ is determined later. Define $I_\eta^*=(a^L,a^U)$ with $a^L=\inf_{x\in\mathscr{X},\eta\in\mathcal{B}(\eta_0,\delta_0)}x'\eta$ and $a^U=\sup_{x\in\mathscr{X},\eta\in\mathcal{B}(\eta_0,\delta_0)}x'\eta$. Define

$$\mathcal{F}_n=\left\{f_n(w,u)=\mathbb{I}\{x'\eta>q_n\}\left(\frac{1}{\tau(x'\eta)}-\frac{1}{\tau_\eta(x'\eta)}\right)w_hu:\ \begin{array}{c}\tau\in\mathcal{T}_{I_\eta^*,K_1},\eta\in\mathcal{B}(\eta_0,\delta_0),\\ \|\tau-\tau_0\|_{2,P}^2\leq v^2,h\in\{1:\dim(w)\},\\ \mathbb{I}(x'\eta>q_n)/\tau(x)\leq 1/K_0\end{array}\right\},$$

where w_h is the h-th component of w. We set $2K_0=\underline{C}$, $K_1=K_2\log n$, and $v=K_3(\log n)n^{-1/3}$ for some positive constants K_2 and K_3.

By van der Vaart and Wellner (1996, Theorem 2.7.5), it holds for each $\epsilon \in (0, K_1)$,

$$H_B(\epsilon, \mathcal{T}_{I_\eta^*,K_1}, \|\cdot\|_P) \leq \frac{C'K_1}{\epsilon}.$$

Similarly to the univariate case, we can choose those bracket functions (τ^L, τ^U), which satisfy $\mathbb{I}\{x'\eta > q_n\}/\tau^L(x'\eta) \leq 1/K_0$. Then, we define

$$f^L(w,u) = \begin{cases} \mathbb{I}\{x'\eta > q_n\}\left(\frac{1}{\tau^U(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)}\right) w_h u & \text{if } w_h u \geq 0, \\ \mathbb{I}\{x'\eta > q_n\}\left(\frac{1}{\tau^L(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)}\right) w_h u & \text{if } w_h u < 0, \end{cases}$$

$$f^U(w,u) = \begin{cases} \mathbb{I}\{x'\eta > q_n\}\left(\frac{1}{\tau^L(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)}\right) w_h u & \text{if } w_h u \geq 0, \\ \mathbb{I}\{x'\eta > q_n\}\left(\frac{1}{\tau^U(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)}\right) w_h u & \text{if } w_h u < 0. \end{cases}$$

Note that (f^L, f^U) is a bracket for $f \in \mathcal{F}_n$. The bracket size is

$$\begin{aligned} &\|\tilde{f}^U - \tilde{f}^L\|_{B,P}^2 = \|D^{-1}f^U - D^{-1}f^L\|_{B,P}^2 \\ = \;& 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \int_{\mathscr{W}\times\mathbb{R}} \mathbb{I}\{x'\eta > q_n\} \left| \left(\frac{1}{\tau^L(x'\eta)} - \frac{1}{\tau^U(x'\eta)}\right) w_h u \right|^k dP(w,u) \\ \leq \;& 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \left\{ \frac{R^k k! M_0^{k-2} a_0 (2K_1)^{k-2}}{K_0^{2k}} \|\tau^U - \tau^L\|_P^2 \right\} \\ \leq \;& 2a_0 \left(\frac{R}{DK_0^2}\right)^2 \sum_{k=0}^{\infty} \left(\frac{2RM_0K_1}{DK_0^2}\right)^k \epsilon^2, \end{aligned}$$

where the first inequality follows from Assumption M2 (where we can choose $a_0, M_0 > 1$) and $\frac{\mathbb{I}\{x'\eta > q_n\}}{\tau^L(x'\eta)} \leq \frac{1}{K_0}$. Setting $D = 4M_0RK_1/K_0^2$ yields $\|\tilde{f}^U - \tilde{f}^L\|_{B,P} \leq \tilde{K}\epsilon$ for $\tilde{K} = \frac{a_0^{1/2}}{2M_0K_1}$, and thus

$$H_B(\epsilon, \tilde{\mathcal{F}}, \|\cdot\|_{B,P}) \leq \frac{\tilde{B}}{\epsilon}, \quad \text{for } \tilde{B} = \frac{C_2 a_0^{1/2}}{2M_0}. \tag{B.3}$$

Now we compute the Bernstein norm of $\tilde{f}$:

$$\begin{aligned} \|\tilde{f}\|_{B,P}^2 &= 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \int_{\mathscr{W}\times\mathbb{R}} \mathbb{I}\{x'\eta > q_n\} \left| \left(\frac{1}{\tau(x'\eta)} - \frac{1}{\tau_\eta(x'\eta)}\right) w_h u \right|^k dP(w,u) \\ &\leq 2\sum_{k=2}^{\infty} \frac{1}{k!D^k} \left\{ \frac{R^k k! M_0^{k-2} a_0 (2K_1)^{k-2}}{K_0^{2k}} \|\tau - \tau_0\|_P^2 \right\} \\ &\leq 2a_0 \left(\frac{R}{DK_0^2}\right)^2 \sum_{k=0}^{\infty} \left(\frac{2RM_0K_1}{DK_0^2}\right)^k v^2 \leq \frac{a_0}{4M_0^2} \frac{1}{K_1^2} v^2, \end{aligned}$$

where the first inequality follows from $\frac{\mathbb{I}\{x'\eta > q_n\}}{\tau(x'\eta)} \leq \frac{1}{K_0}$. This implies

$$\|\tilde{f}\|_{B,P} \leq B\frac{v}{K_1}, \quad \text{for } B = \frac{a_0^{1/2}}{2M_0}. \tag{B.4}$$

Combining (B.3) and (B.4), the remaining steps are the same as those in the proof of Lemma 3 (iii).

B.3. **Proof of Lemma 4.** Recall for fixed η, we first obtain $\hat{\tau}_\eta = \arg\min_{\tau \in \mathcal{M}} \frac{1}{n}\sum_{i=1}^n \{\hat{U}_i^2 - \tau(X_i'\eta)\}^2$ and then obtain $\hat{\eta}$ by $\hat{\eta} = \underset{\eta}{\mathrm{argmin}} ||\frac{1}{n}\sum_{i=1}^n X_i'\{\hat{U}_i^2 - \hat{\tau}_\eta(X_i'\eta)\}||^2$. We denote $E[X|X'\eta = x'\eta]$ by $E[X|x'\eta]$. The proof is similar to the ones in BGH and Balabdaoui and Groeneboom (2021) except that we need to handle the influence of the estimated dependent variables $\hat{U}_i^2$.

The proof of consistency of $\hat{\eta}$ is similar to pp.16-17 of BGH-supp. By a similar argument in Balabdaoui and Groeneboom (2021, Lemma 3.2), under Assumptions M1-M3, we have

$$\frac{1}{n}\sum_{i=1}^n X_i'\{\hat{U}_i^2 - \hat{\tau}_\eta(X_i'\eta)\} = \frac{1}{n}\sum_{i=1}^n (X_i - E[X|X_i'\eta])\{\hat{U}_i^2 - \tau_\eta(X_i'\eta)\} + o_p(n^{-1/2}),$$

for each η, where we also use (B.2). Thus, it holds

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n X_i\{\hat{U}_i^2 - \hat{\tau}_{\hat{\eta}}(X_i'\hat{\eta})\}\right\| &= \min_\eta \left\|\frac{1}{n}\sum_{i=1}^n X_i\{\hat{U}_i^2 - \hat{\tau}_\eta(X_i'\eta)\}\right\| \\
&\leq \min_\eta \left\|\frac{1}{n}\sum_{i=1}^n (X_i - E[X|X_i'\eta])\{\hat{U}_i^2 - \tau_\eta(X_i'\eta)\} + o_p(n^{-1/2})\right\|.
\end{aligned}$$

The leading term inside the norm $\|\cdot\|$ of the last expression does not depend on the potentially non-smooth $\hat{\tau}_\eta$; it is a smooth function of η. Thus, under standard conditions for the method of moments, we have $\min_\eta \left\|\frac{1}{n}\sum_{i=1}^n (X_i - E[X_i|X_i'\eta])\{\hat{U}_i^2 - \tau_\eta(X_i'\eta)\}\right\| = 0$, and

$$\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^n X_i\{\hat{U}_i^2 - \hat{\tau}_{\hat{\eta}}(X_i'\hat{\eta})\} \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - E[X|X_i'\hat{\eta}])\{\hat{U}_i^2 - \hat{\tau}_{\hat{\eta}}(X_i'\hat{\eta})\} + o_p(n^{-1/2} + (\hat{\eta} - \eta)) \\
&= \int (x - E[X|x'\hat{\eta}])\{\hat{u}^2 - \tau_0(x'\eta_0)\} d(\mathbb{P}_n - P)(x, \hat{u}) \\
&\quad + \int (x - E[X|x'\hat{\eta}])\{\hat{u}^2 - \tau_{\hat{\eta}}(x'\hat{\eta})\} dP(x, \hat{u}) + o_p(n^{-1/2} + (\hat{\eta} - \eta)) \\
&=: I + II + o_p(n^{-1/2} + (\hat{\eta} - \eta)), \qquad \text{(B.5)}
\end{aligned}$$

where the second equality follows from similar arguments to pp.18-20 of BGH-supp and (B.2), and the third equality follows from a similar argument in pp.21-23 of BGH-supp.

Let $\hat{U}(w, u) = u - w'(\hat{\theta}_{\mathrm{OLS}} - \theta)$ and

$$\hat{e}(w, u) := \hat{U}(w, u)^2 - u^2 = -2w'(\hat{\theta}_{\mathrm{OLS}} - \theta)u + \{w'(\hat{\theta}_{\mathrm{OLS}} - \theta)\}^2. \qquad \text{(B.6)}$$

For I, we have

$$\begin{aligned}
I &= \int (x - E[X|x'\hat{\eta}])\{u^2 + \hat{e}(w, u) - \tau_0(x'\eta_0)\} d(\mathbb{P}_n - P)(w, u) \\
&= \int (x - E[X|x'\eta_0])\{u^2 - \tau_0(x'\eta_0)\} d(\mathbb{P}_n - P)(x, u) \\
&\quad + \int (x - E[X|x'\hat{\eta}])\hat{e}(w, u) d(\mathbb{P}_n - P)(w, u) + o_p(n^{-1/2}) \\
&= \int (x - E[X|x'\eta_0])\{u^2 - \tau_0(x'\eta_0)\} d(\mathbb{P}_n - P)(x, u) + o_p(n^{-1/2}), \qquad \text{(B.7)}
\end{aligned}$$

where the second equality follows from p.21 of BGH-supp, and the third equality follows from the facts that (a) $\hat{\theta}_{\text{OLS}} - \theta = O_p(n^{-1/2})$, (b) $\hat{e}(w,u)$ is a parametric function of w and u in a changing class indexed by $\hat{\theta}_{\text{OLS}}$ (see (B.6)), so its ϵ-entropy is of order $\log(1/\epsilon) \leq 1/\epsilon$ (see, e.g., Example 19.7 of van der Vaart and Wellner, 2000), and (c) similar arguments in pp.22-23 of BGH-supp. By Lemma 17 of BGH-supp we have

$$\tau_\eta(x'\eta) = \tau_0(x'\eta_0) + (\eta - \eta_0)(x - E[X|X'\eta_0 = x'\eta_0])\tau_0'(x'\eta_0) + o_p(\eta - \eta_0). \tag{B.8}$$

For II, observe that

$$\begin{aligned}
II &= \int (x - E[X|x'\hat{\eta}])\{u^2 - \tau_{\hat{\eta}}(x'\hat{\eta})\}dP(x,u) + \int (x - E[X|x'\hat{\eta}])\hat{e}(w,u)dP(w,u) \\
&= \left\{ \int (x - E[X|x'\eta_0])(x - E[X|X'\eta_0 = x'\eta_0])\tau_0'(x'\eta_0)dP(x) \right\} (\hat{\eta} - \eta_0) \\
&\quad + \int (x - E[X|x'\hat{\eta}])\hat{e}(w,u)dP(w,u) + o_p(\hat{\eta} - \eta_0) \\
&= \left\{ \int (x - E[X|x'\eta_0])(x - E[X|x'\eta_0])\tau_0'(x'\eta_0)dP(x) \right\} (\hat{\eta} - \eta_0) + O_p(n^{-1/2}) + o_p(\hat{\eta} - \eta_0) \\
&= B(\hat{\eta} - \eta_0) + O_p(n^{-1/2}) + o_p(\hat{\eta} - \eta_0), \qquad \text{(B.9)}
\end{aligned}$$

where the third equality follows from (B.8) and $(E[X|x'\hat{\eta}] - E[X|x'\eta_0])(\hat{\eta} - \eta_0) = o_p(\hat{\eta} - \eta_0)$, the fourth equality follows from $\hat{\theta}_{\text{OLS}} - \theta = O_p(n^{-1/2})$ and the definition of B in Assumption M6.

Combining (B.5), (B.7), and (B.9), we have

$$\hat{\eta} - \eta_0 = B^- \int (x - E[X|x'\eta_0])\{u^2 - \tau_0(x'\eta_0)\}d(\mathbb{P}_n - P)(x,u) + O_p(n^{-1/2}) + o_p(n^{-1/2} + (\hat{\eta} - \eta)),$$

where B^- is the Moore-Penrose inverse of B (see p.17 of BGH for more details). Therefore, we have $\hat{\eta} - \eta_0 = O_p(n^{-1/2})$. This result, combined with (B.8) and Assumptions M1 and M2, implies $\tau_{\hat{\eta}}(a) - \tau_0(a) = O_p(n^{-1/2})$ and $\|\tau_{\hat{\eta}} - \tau_0\|_{2,P} = O_p(n^{-1/2})$.

References

[1] Acemoglu, D. and P. Restrepo (2017) Secular stagnation? The effect of aging on economic growth in the age of automation, *American Economic Review*, 107, 174-179.

[2] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and E. Silverman (1955) An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics*, 26, 641-647.

[3] Babii, A. and R. Kumar (2023) Isotonic regression discontinuity designs. *Journal of Econometrics*, 234(2), 371-393.

[4] Balabdaoui, F., Durot, C. and H. Jankowski (2019) Least squares estimation in the monotone single index model, *Bernoulli*, 25, 3276-3310.

[5] Balabdaoui, F., Groeneboom, P. and K. Hendrickx (2019) Score estimation in the monotone single-index model, *Scandinavian Journal of Statistics*, 46, 517-544.

[6] Balabdaoui, F. and P. Groeneboom (2021) Profile least squares estimators in the monotone single index model, in *Advances in Contemporary Statistics and Econometrics*, pp. 3-22, Springer.

[7] Barlow, R. and H. Brunk (1972) The isotonic regression problem and its dual, *Journal of the American Statistical Association*, 67, 140-147.

[8] Bickel, P. J. (1978) Using residuals robustly I: Tests for heteroscedasticity, nonlinearity, *Annals of Statistics*, 6, 266-291.

[9] Box, G. E. and W. J. Hill (1974) Correcting inhomogeneity of variance with power transformation weighting, *Technometrics*, 16, 385-389.

[10] Carroll, R. J. (1982) Adapting for heteroscedasticity in linear models, *Annals of Statistics*, 10, 1224-1233.

[11] Chetverikov, D. (2019) Testing regression monotonicity in econometric models, *Econometric Theory*, 35, 729-776.

[12] Cragg, J. G. (1983) More efficient estimation in the presence of heteroscedasticity of unknown form, *Econometrica*, 51, 751-764.

[13] Cragg, J. G. (1992) Quasi-Aitken estimation for heteroskedasticity of unknown form, *Journal of Econometrics*, 54, 179-201.

[14] Dümbgen, L. and V. G. Spokoiny (2001) Multiscale testing of qualitative hypotheses, *Annals of Statistics*, 29, 124-152.

[15] Fang, B., Guntuboyina, A., and Sen, B. (2021) Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.* 49(2): 769-792.

[16] Ghosal, S., Sen, A. and A. van der Vaart (2000) Testing monotonicity of regression, *Annals of Statistics*, 28, 1054-1082.

[17] Grenander, U. (1956) On the theory of mortality measurement. II., *Skand. Aktuarietidskr*, 39, 125-153.

[18] Groeneboom, P. and G. Jongbloed (2014) *Nonparametric Estimation under Shape Constraints*, Cambridge University Press.

[19] Hall, P. and N. E. Heckman (2000). Testing for monotonicity of a regression mean by calibrating for linear functions, *Annals of Statistics*, 28, 20-39.

[20] Hansen, A. H. (1939) Economic progress and declining population growth, *American Economic Review*, 29, 1-15.

[21] Hsu, Y. C., Liu, C. A. and X. Shi (2019) Testing generalized regression monotonicity, *Econometric Theory*, 35, 1146-1200.

[22] Jobson, J. D. and W. A. Fuller (1980) Least squares estimation when the covariance matrix and parameter vector are functionally related, *Journal of the American Statistical Association*, 75, 176-181.

[23] Kim, J. and D. Pollard (1990) Cube root asymptotics, *Annals of Statistics*, 18, 191-219.

[24] Kulikov, V. N. and Lopuhaä, H. P. (2006) The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Annals of Statistics*, 34(2), 742-768.

[25] Matzkin, R. L. (1994) Restrictions of economic theory in nonparametric methods, in Engle R. F. and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, pp. 2523-2558, Elsevier.

[26] Mincer, J. (1974) Schooling, Experience, and Earnings. *Human Behavior & Social Institutions,* No. 2.

[27] Newey, W. K. (1993) Efficient estimation of models with conditional moment restrictions, in Maddala, G. S., Rao, C. R. and H. D. Vinod (eds.), *Handbook of Statistics*, vol. 11, pp. 419-454, Elsevier.

[28] Rao, B. P. (1969) Estimation of a unimodal density, *Sankhyā*, A 31, 23-36.

[29] Rao, B. P. (1970) Estimation for distributions with monotone failure rate, *Annals of Mathematical Statistics*, 41, 507-519.

[30] Robinson, P. M. (1987) Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form, *Econometrica*, 55, 875-891.

[31] Ruud, P. A. (2000) *An introduction to classical econometric theory.* OUP Catalogue.

[32] van de Geer, S. (2000) Empirical Processes in M−estimation. *Cambridge University Press.*

[33] van der Vaart, A. W. (2000) *Asymptotic Statistics*, Cambridge University press.

[34] van der Vaart, A. W. and J. A. Wellner (1996) *Weak Convergence and Empirical Processes*, Springer.

[35] Woodroofe, M. and Sun, J. (1993) A penalized maximum likelihood estimate of f (0+) when f is non-increasing. *Statistica Sinica*, 501-515.

[36] Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., MIT Press.

[37] Wooldridge, J. M. (2013) *Introductory econometrics: A modern approach, 5th Ed.* Cengage learning.

School of Social Sciences, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan.

Email address: yarai@waseda.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: t.otsu@lse.ac.uk

Department of Economics, University of Mannheim, L7 3-5, 68161, Mannheim, Germany.

Email address: mengshan.xu@uni-mannheim.de