SUPPLEMENT TO "ERRORS IN THE DEPENDENT VARIABLE OF QUANTILE REGRESSION MODELS"

(*Econometrica*, Vol. 89, No. 2, March 2021, 849–873)

JERRY HAUSMAN
Department of Economics, MIT and NBER

HAOYANG LIU
Research and Statistics Group, Federal Reserve Bank of New York

YE LUO
HKU Business School, The University of Hong Kong

CHRISTOPHER PALMER
Sloan School of Management, MIT and NBER

SUPPLEMENTAL APPENDIX A: ESTIMATOR IMPLEMENTATION DETAILS

IN THIS APPENDIX, we present implementation details for our maximum likelihood estimator. Additional details and code to run the estimator can be found at https://github.com/palmercj/EIV-QR.

The main step is Step 5, the piecewise-linear sieve-ML estimator described in Section 3.1. Because this piecewise-linear estimator is computationally intensive, we use a series of preliminary steps to find start values in the neighborhood of the optimum.[16] These steps significantly reduce the time required for convergence of the piecewise-linear estimator.

(1) We estimate quantile regression on a grid of knots $[t_1, t_2, \ldots, t_J]$, where J is the number of knots, and denote the estimate as $\widehat{\beta}_{\mathrm{QR}}(\cdot)$.

(2) We run 40 weighted least squares (WLS) iterations using $\widehat{\beta}_{\mathrm{QR}}(\cdot)$ from Step 1 as the start value. Using WLS in some fashion is a common technique in quantile regression computational programs and in our case is motivated by the fact that under a normality assumption of the EIV term ε, the maximum likelihood estimator is equivalent to a weighted least squares one. Supplemental Material Appendix Section A.1 demonstrates this equivalence and also specifies the weights for the WLS iterations. We denote the weighted least squares estimate as $\widehat{\beta}_{\mathrm{WLS}}(\cdot)$.

(3) We estimate a piecewise-constant maximum likelihood estimator using $(\widehat{\beta}_{\mathrm{WLS}}(\cdot), \sigma_D)$ as the start value, where σ_D is a default start value for EIV parameters. In our simulations, where we estimate EIVs as mixtures of three normals, our start values for the EIV parameters specify three equally weighted mixtures with means -1,

Jerry Hausman: jhausman@mit.edu
Haoyang Liu: haoyang.liu@ny.frb.org
Ye Luo: kurtluo@hku.hk
Christopher Palmer: cjpalmer@mit.edu

[16] A key reason for added computational complexity for a piecewise-linear estimator relative to, for example, a piecewise-constant one is that we ensure $\widehat{\beta}(\cdot)$ is continuous at all knots $t_1, t_2, \ldots, t_J$ in the τ grid over $(0, 1)$. As such, if we change the estimated slope of $\widehat{\beta}(\cdot)$ between t_1 and t_2, it will affect $\widehat{\beta}(t_2)$ and also the level of $\widehat{\beta}(\cdot)$ for the entire range of $[t_2, 1]$. This makes $E_n[\log g(y|x, \beta, \sigma)]$ a highly nonlinear function of b_l, $l = 1, 2, \ldots, J+1$ (the coefficients for the spline functions) with many interaction terms of b_l.

 https://doi.org/10.3982/ECTA14667

0, and 1 and 1 as the standard deviation of each mixture. We search for the log-likelihood maximizer using an interior-point constrained optimizer, to which we feed analytical gradients and Hessian matrix. We constrain the standard deviations for the mixture components to be positive, the EIV distribution to be mean zero, the weights to sum to 1, and require that the sum of the weights for the first and second mixtures does not exceed 1 to ensure that the third weight is non-negative. We denote the resulting estimates as $(\widehat{\beta}_{\mathrm{PC}}(\cdot), \widehat{\sigma}_{\mathrm{PC}})$.

(4) We then calculate start values for the piecewise-linear estimator by sorting $\widehat{\beta}_{\mathrm{PC}}(\cdot)$ by $E_n[x]^T\widehat{\beta}_{\mathrm{PC}}(\cdot)$ (Chernozhukov, Fernandez-Val, and Galichon (2009)). Note that the log-likelihood of a piecewise-constant sieve in our setting is invariant to rearranging τ. For example, swapping the value of $\widehat{\beta}_{\mathrm{PC}}(\cdot)$ in $[t_1, t_2]$ with the one in $[t_2, t_3]$ does not change the empirical log-likelihood $L_n(\theta) = E_n[\log g(y|x, \theta)]$ as long as $t_3 - t_2 = t_2 - t_1$. We rearrange the elements of $\widehat{\beta}_{\mathrm{PC}}(\cdot)$ so that $x^T\widehat{\beta}_{PC}(\tau)$ is monotonically increasing in τ at the mean of the covariates. Then we connect $(\frac{t_j+t_{j+1}}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_j+t_{j+1}}{2}))$ with $(\frac{t_{j+1}+t_{j+2}}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_{j+1}+t_{j+2}}{2}))$, $j = 1, \ldots, J$.[17] We denote this constructed $\beta(\cdot)$ as $\tilde{\beta}_{\mathrm{PL}}(\cdot)$. We use $(\tilde{\beta}_{\mathrm{PL}}(\cdot), \widehat{\sigma}_{\mathrm{PC}})$ as the starting values for Step 5, piecewise-linear estimation.

(5) We run a piecewise-linear maximum likelihood estimator, and denote the final estimate as $(\widehat{\beta}_{\mathrm{PL}}(\cdot), \widehat{\sigma}_{\mathrm{PL}})$. Implementation details for the piecewise-linear estimator are similar to the ones in the piecewise-constant estimator, described in Step 4. We have also found genetic-algorithm-based optimizers to perform well in some applications.

A.1. *Weighted Least Squares*

Under a normality assumption of the EIV term ε, the maximization of $E_n[\log g(y|x, \theta)]$ reduces to the minimization of a simple weighted least squares problem. Suppose the disturbance ε is modeled as a normal random variable. Then the maximization problem (3.3) becomes the following, with the parameter vector $\theta = (\beta(\cdot), \sigma)$:

$$\max_{\theta} E_n\big[\log g(y|x, \theta)\big] := E_n\left[\int_\tau^1 \frac{f\big(y - x^T\beta(\tau)|\sigma\big)}{\int_0^1 f\big(y - x^T\beta(u)|\sigma\big)\,du}\left(-\frac{1}{2}\log\big(2\pi\sigma^2\big) - \frac{\big(y - x^T\beta(\tau)\big)^2}{2\sigma^2}\right)d\tau\right]. \tag{A.1}$$

Equation (A.1) demonstrates that the maximization problem of $\beta(\cdot)$ is to minimize the sum of weighted least squares. As in standard normal MLE, the FOC for $\beta(\cdot)$ does not depend on σ^2. Given an initial estimate of a weighting matrix W, the weighted least squares estimates of β and σ are

$$\widehat{\beta}(\tau_j) = \big(X^TW_jX\big)^{-1}X^TW_jy,$$

$$\widehat{\sigma} = \sqrt{\frac{1}{NJ}\sum_j\sum_i w_{ij}\widehat{\varepsilon}_{ij}^2},$$

[17] For the first interval we connect $(\frac{t_1}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_1}{2}))$ with $(\frac{t_1+t_2}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_1+t_2}{2}))$, and then extend it backward to $\tau = 0$. For the last interval, we connect $(\frac{t_{J-1}+t_J}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_{J-1}+t_J}{2}))$ with $(\frac{t_J+1}{2}, \widehat{\beta}_{\mathrm{PC}}(\frac{t_J+1}{2}))$, and extend it to $\tau = 1$

where W_j is the diagonal matrix formed from the jth column of W, which has elements w_{ij}. Given estimates $\widehat{\varepsilon}_j = y - X^T\widehat{\beta}(\tau_j)$ and $\widehat{\sigma}$, the weights w_{ij} for observation i in the estimation of $\beta(\tau_j)$ are

$$w_{ij} = \frac{\phi(\widehat{\varepsilon}_{ij}/\widehat{\sigma})}{\frac{1}{J}\sum_j \phi(\widehat{\varepsilon}_{ij}/\widehat{\sigma})},$$

where $\phi(\cdot)$ is the PDF of a standard normal distribution.

SUPPLEMENTAL APPENDIX B: ADDITIONAL SIMULATION RESULTS

In this appendix, we present Monte Carlo simulation results (mean bias and MSE) under alternative data-generating processes. For each design, quasi-ML estimation continues to treat the measurement error as a mixture of three normals. After simulating measurement error under alternative measurement error distributions (all normalized such that ε has equal variance across designs), Supplemental Material Appendix Table BIV presents results when a 99-knot sieve is used to approximate $\beta(\cdot)$. Finally, Supplemental Material Appendix Table BV reports bootstrapped confidence interval coverage probabilities for each parameter.

TABLE BI

MEAN BIAS AND MEAN SQUARED ERROR: $\varepsilon \sim 3\mathcal{N}$[a]

	β_1		β_2		β_3	
Quantile	QR	SMLE	QR	SMLE	QR	SMLE
			I. Mean Bias			
0.1	−2.926	−0.034	0.146	0.006	0.135	0.018
0.2	−2.488	0.000	0.224	0.005	0.144	−0.004
0.3	−2.074	0.007	0.266	0.005	0.130	0.006
0.4	−1.511	−0.007	0.249	0.007	0.088	0.005
0.5	−0.401	−0.041	0.101	0.016	−0.013	0.005
0.6	1.058	−0.024	−0.124	0.008	−0.121	0.000
0.7	1.940	−0.003	−0.237	0.008	−0.141	−0.005
0.8	2.602	0.033	−0.284	0.000	−0.125	−0.001
0.9	3.353	0.062	−0.283	−0.008	−0.097	0.005
$\overline{\lvert\text{Bias}\rvert}$	2.039	0.023	0.213	0.007	0.110	0.005
			II. Mean Squared Error			
0.1	8.565	0.040	0.021	0.005	0.018	0.006
0.2	6.189	0.020	0.051	0.003	0.021	0.003
0.3	4.301	0.038	0.071	0.008	0.017	0.007
0.4	2.284	0.019	0.062	0.005	0.008	0.004
0.5	0.162	0.033	0.010	0.011	0.000	0.006
0.6	1.119	0.019	0.016	0.004	0.015	0.003
0.7	3.763	0.032	0.056	0.011	0.020	0.004
0.8	6.770	0.018	0.081	0.003	0.016	0.001
0.9	11.242	0.032	0.081	0.004	0.010	0.002
$\overline{\text{MSE}}$	4.933	0.028	0.050	0.006	0.014	0.004

[a]Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and sieve MLE across 500 Monte Carlo simulations of $n = 100{,}000$ observations using data simulated from the data-generating process described in Section 4. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE BII

MEAN BIAS AND MEAN SQUARED ERROR: $\varepsilon \sim t$[a]

	β_1		β_2		β_3	
Quantile	QR	SMLE	QR	SMLE	QR	SMLE
			I. Mean Bias			
0.1	−1.965	0.103	0.142	0.018	0.103	−0.004
0.2	−1.091	0.098	0.112	−0.005	0.050	−0.011
0.3	−0.650	0.012	0.087	−0.017	0.021	−0.013
0.4	−0.338	−0.060	0.062	0.017	0.000	0.003
0.5	−0.063	−0.053	0.032	0.017	−0.015	0.010
0.6	0.226	0.009	−0.005	−0.003	−0.030	0.000
0.7	0.580	0.008	−0.052	−0.001	−0.045	−0.005
0.8	1.095	0.061	−0.114	−0.018	−0.061	−0.010
0.9	2.080	0.044	−0.200	−0.015	−0.083	−0.010
$\overline{\lvert\text{Bias}\rvert}$	0.899	0.050	0.090	0.012	0.045	0.007
			II. Mean Squared Error			
0.1	3.862	0.220	0.020	0.016	0.011	0.027
0.2	1.191	0.069	0.013	0.006	0.003	0.006
0.3	0.423	0.120	0.008	0.019	0.001	0.016
0.4	0.114	0.070	0.004	0.010	0.000	0.006
0.5	0.004	0.147	0.001	0.032	0.000	0.013
0.6	0.052	0.057	0.000	0.012	0.001	0.004
0.7	0.337	0.123	0.003	0.031	0.002	0.010
0.8	1.199	0.031	0.013	0.007	0.004	0.003
0.9	4.326	0.020	0.040	0.009	0.007	0.005
$\overline{\text{MSE}}$	1.279	0.095	0.011	0.016	0.003	0.010

[a]Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and sieve quasi-MLE modeling the error term as a mixture of three normals across 500 Monte Carlo simulations of $n = 100{,}000$ observations each. The data are simulated from the data-generating process described in Section 4 but measurement error generated as a Student's t random variable with three degrees of freedom, multiplied by $\sqrt{3.5}$ to ensure the variance of the measurement error is equal across simulation designs. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE BIII

MEAN BIAS AND MEAN SQUARED ERROR: $\varepsilon \sim$ Laplace[a]

	β_1		β_2		β_3	
Quantile	QR	SMLE	QR	SMLE	QR	SMLE
			I. Mean Bias			
0.1	−2.507	0.058	0.176	0.017	0.127	0.000
0.2	−1.332	0.030	0.127	−0.003	0.054	0.001
0.3	−0.776	−0.055	0.094	0.015	0.019	0.007
0.4	−0.394	−0.050	0.065	0.019	−0.001	0.002
0.5	−0.059	0.022	0.034	−0.005	−0.017	−0.007
0.6	0.285	0.035	−0.005	−0.009	−0.032	−0.011
0.7	0.697	0.043	−0.052	−0.022	−0.047	−0.003
0.8	1.330	0.028	−0.127	−0.010	−0.069	−0.007
0.9	2.631	0.016	−0.243	−0.009	−0.102	−0.011
$\overline{\lvert\text{Bias}\rvert}$	1.112	0.038	0.103	0.012	0.052	0.006
			II. Mean Squared Error			
0.1	6.284	0.160	0.031	0.015	0.016	0.026
0.2	1.775	0.055	0.016	0.006	0.003	0.008
0.3	0.603	0.156	0.009	0.028	0.000	0.020
0.4	0.156	0.069	0.004	0.013	0.000	0.007
0.5	0.004	0.198	0.001	0.041	0.000	0.018
0.6	0.081	0.063	0.000	0.013	0.001	0.005
0.7	0.486	0.116	0.003	0.033	0.002	0.011
0.8	1.770	0.026	0.016	0.007	0.005	0.003
0.9	6.923	0.024	0.060	0.010	0.011	0.006
$\overline{\text{MSE}}$	2.009	0.096	0.016	0.018	0.004	0.012

[a]Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and sieve quasi-MLE modeling the error term as a mixture of three normals across 500 Monte Carlo simulations of $n = 100{,}000$ observations each. The data are simulated from the data-generating process described in Section 4 but measurement error generated as a Laplace random variable with $\lambda = 2.29$ to ensure the variance of the measurement error is equal across simulation designs. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE BIV

MEAN BIAS AND MEAN SQUARED ERROR: 99 KNOTS[a]

	β_1		β_2		β_3	
Quantile	QR	SMLE	QR	SMLE	QR	SMLE
			I. Mean Bias			
0.1	−2.926	−0.231	0.146	0.009	0.135	0.007
0.2	−2.488	−0.208	0.224	−0.002	0.144	0.010
0.3	−2.074	−0.211	0.266	0.013	0.130	0.003
0.4	−1.511	−0.176	0.249	−0.007	0.088	−0.003
0.5	−0.401	−0.193	0.101	0.006	−0.013	0.001
0.6	1.058	−0.214	−0.124	0.018	−0.121	0.003
0.7	1.940	−0.223	−0.237	0.020	−0.141	−0.002
0.8	2.602	−0.173	−0.284	0.005	−0.125	0.004
0.9	3.353	−0.113	−0.283	−0.004	−0.097	0.000
$\overline{\lvert\text{Bias}\rvert}$	2.039	0.194	0.213	0.009	0.110	0.004
			II. Mean Squared Error			
0.1	8.565	0.330	0.021	0.008	0.018	0.014
0.2	6.189	0.283	0.051	0.012	0.021	0.019
0.3	4.301	0.298	0.071	0.021	0.017	0.015
0.4	2.284	0.273	0.062	0.023	0.008	0.016
0.5	0.162	0.274	0.010	0.029	0.000	0.013
0.6	1.119	0.287	0.016	0.034	0.015	0.012
0.7	3.763	0.283	0.056	0.025	0.020	0.008
0.8	6.770	0.237	0.081	0.021	0.016	0.006
0.9	11.242	0.091	0.081	0.013	0.010	0.004
$\overline{\text{MSE}}$	4.933	0.262	0.050	0.021	0.014	0.012

[a]Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and sieve MLE across 500 Monte Carlo simulations of $n = 100{,}000$ observations using data simulated from the data-generating process described in Section 4 and when a sieve of $J = 99$ knots is used in estimation. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE BV

COVERAGE PROBABILITIES FOR BOOTSTRAPPED CONFIDENCE INTERVALS[a]

	I. Coefficient Functions		
Quantile	β_1	β_2	β_3
0.0625	1.00	0.98	0.95
0.1250	1.00	0.99	0.95
0.1875	0.99	0.98	0.98
0.2500	0.98	0.98	0.99
0.3125	1.00	0.99	0.99
0.3750	0.96	0.96	1.00
0.4375	0.99	0.99	0.99
0.5000	0.99	0.99	0.99
0.5625	1.00	1.00	1.00
0.6250	1.00	0.99	1.00
0.6875	1.00	1.00	0.98
0.7500	0.99	1.00	0.99
0.8125	0.97	0.97	1.00
0.8750	0.95	0.98	0.96
0.9375	0.85	0.93	0.95

	II. Distributional Parameters	
	Parameter	Coverage
mixture weights	λ_1	0.99
	λ_2	0.96
	λ_3	0.95
mixture means	μ_1	0.97
	μ_2	1
	μ_2	0.97
mixture standard deviations	σ_1	0.96
	σ_2	0.98
	σ_3	0.96
Average coverage across all parameters		0.98

[a]Notes: Table reports coverage probabilities for the sieve MLE bootstrapped confidence intervals across 100 Monte Carlo simulations of 100 bootstrap draws each using data simulated from the data-generating process described in Section 4. For each Monte Carlo draw from the data-generating process, we estimate our sieve-ML estimator, bootstrap the simulated data, reestimate the model parameters, and then construct a confidence interval centered at the first sieve-ML estimate for that Monte Carlo draw ± 1.96 times the standard deviation of that parameter's estimates across that Monte Carlo simulation's bootstrap draws. The reported coverage probabilities above indicate the fraction of Monte Carlo simulations for which the true parameter is contained inside the bootstrapped confidence intervals.

SUPPLEMENTAL APPENDIX C: DATA APPENDIX

Following the sample selection criteria of Angrist, Chernozhukov, and Fernández-Val (2006), our data come from 1% samples of decennial census data available via IPUMS.org (Ruggles, Genadek, Goeken, Grover, and Sobek (2015)) from 1980 to 2010. Copies of the extracts we use are available at https://github.com/palmercj/EIV-QR. From each database, we select annual wage income, education, age, and race data for prime-age (age 40–49) white males who have at least five years of education, were born in the United States, had positive earnings and hours worked in the reference year, and whose responses for age, education, and earnings were not imputed (which would have been an additional source of measurement error). Our dependent variable is log weekly wage, obtained as annual wage income divided by weeks worked. For 1980, we take the number of years of education to be the highest grade completed and follow the methodology of Angrist, Chernozhukov, and Fernández-Val (2006) to convert the categorical education variable in 1990, 2000, and 2010 into a measure of the number of years of schooling. Experience is defined as age minus years of education minus five. For 1980, 1990, and 2000, we use the exact extract of Angrist, Chernozhukov, and Fernández-Val (2006), and draw our own data to extend the data to include the 2010 census. Table CI reports summary statistics for the variables used in the regressions in the text. Wages for 1980–2000 were expressed in 1989 dollars after deflating using the Personal Consumption Expenditures Index. As slope coefficients in a log-linear quantile regression specification are unaffected by scaling the dependent variable, we do not deflate our 2010 data.

Although quantile regression recovers effects on the conditional distribution of the outcome, it is worth noting that given the substantial variation in wages left unexplained by the Mincer model, the empirical difference between effects on the unconditional and conditional distributions of the dependent variable is likely small. See DiNardo, Fortin, and Lemieux (1996) and Powell (2013) for further discussion and methods that recover effects on the unconditional distribution. Because of the relatively low goodness of fit of equation (5.1) (as is the case in many cross-sectional applied microeconomics settings), over 63% of the observations in the top unconditional decile are also in the top conditional decile.

TABLE CI

EDUCATION AND WAGES SUMMARY STATISTICS[a]

Year	1980	1990	2000	2010
Log weekly wage	6.43	6.48	6.50	8.37
	(0.66)	(0.69)	(0.74)	(0.76)
Education	12.99	13.97	13.90	14.12
	(3.08)	(2.66)	(2.41)	(2.39)
Experience	25.38	24.45	24.45	24.55
	(4.32)	(4.01)	(3.60)	(3.83)
Number of Observations	60,051	80,115	90,201	98,292

[a]Notes: Table reports summary statistics for the census data used in the quantile wage regressions in the text. The 1980, 1990, and 2000 data sets come from Angrist, Chernozhukov, and Fernández-Val (2006). We extend the sample to include 2010 census microdata from IPUMS (Ruggles et al. (2015)).

SUPPLEMENTAL APPENDIX D: ADDITIONAL PROOFS OF LEMMAS AND THEOREMS

D.1. *Lemmas and Theorems in Section 3*

The following lemmas are used in the proofs of Lemmas 2 and 3.

LEMMA 6: *The space* $M[B_1 \times B_2 \times B_3 \times \cdots \times B_{d_x}]$ *is a compact and complete space under* L^p *for any* $p \geq 1$.

PROOF OF LEMMA 6: For bounded monotonic functions, pointwise convergence is equivalent to uniform convergence, making a space of bounded monotonic functions compact under any L^p norm for $p \geq 1$. Hence the product space $B_1 \times B_2 \times \cdots \times B_{d_x}$ is compact. It is complete since the L^p functional space is complete and the limit of monotonic functions is still monotonic. *Q.E.D.*

LEMMA 7—Donskerness of Θ: *The set of functions*

$$\mathcal{G} = \{h(y, x, \beta(\cdot), \sigma) := \log(g(y|x, \beta(\cdot), \sigma)) | (\beta(\cdot), \sigma) \in \Theta\}$$

is μ*-Donsker, where* μ *is the joint PDF of* (y, x).

PROOF: By Theorem 2.7.5 of Van Der Vaart and Wellner (1996), the bracketing number $N_{[]}$ of the space of uniformly bounded monotone functions $\mathcal{F}$ satisfies $\log N_{[]}(\varepsilon, \mathcal{F}, L_r(Q)) \leq K\frac{1}{\varepsilon}$ for every probability measure Q and every $r \geq 1$ and a constant K which depends only on r. Consider a collection of functions $\mathcal{F} := q(y, x, \theta) | \theta \in \Theta$ such that

$$|q(y, x, \theta_1) - q(y, x, \theta_2)| \leq \|\theta_1 - \theta_2\|_2 w(y, x), \tag{D.1}$$

$$E_Q[|w(y, x)|^2] < \infty, \tag{D.2}$$

where Q is some probability measure on (y, x). Since Θ is a product space of bounded monotone functions M and a finite-dimensional bounded compact set Σ, the bracketing number of $\mathcal{F}$ given measure Q is also bounded by $\log N_{[]}(\varepsilon, \mathcal{F}, L_2(Q)) \leq K d_x \frac{1}{\varepsilon}$, where K is a constant depending only on Θ and $w(y, x)$. Therefore, $\int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, Q)} < \infty$, that is, $\mathcal{F}$ is Donsker.

In particular, let $q = \log g$ and $Q = \mu$, where μ is the joint PDF of (x, y). By Assumption 5(6), equation (D.1) holds with $w(y|x) := \int_0^1 |y - x^T\beta(\tau)|^\gamma \, d\tau$. Equation (D.2) is satisfied by Assumption 5(3). Hence, $\mathcal{G}$ is μ-Donsker. *Q.E.D.*

PROOF OF LEMMA 2: To show the consistency of the ML estimator, it is sufficient to prove the satisfaction of the following regularity conditions (Newey and McFadden (1994)):

(1) The parameter space $\Theta = M \times \Sigma$ is compact.
(2) Global identification holds, that is, there exists no other $\theta' = (\beta', \sigma') \in \Theta$ such that $E[\log \int_0^1 f(y - x^T\beta'(\tau)|\sigma') \, d\tau] = E[\log \int_0^1 f(y - x^T\beta_0(\tau)|\sigma_0) \, d\tau]$.
(3) The objective function $E[\log \int_0^1 f(y - x^T\beta'(\tau)|\sigma') \, d\tau]$ is continuous for all $\theta' = (\beta', \sigma') \in \Theta$.
(4) Stochastic equicontinuity of $E_n[\log \int_0^1 f(y - x^T\beta(\tau)|\sigma)]$, with $\theta \in \Theta$.

Condition 1 is established by Lemma 6. Condition 2 is provided by Lemma 1. Condition 3 holds under Assumption 5. For the proof of point 4, see Lemma 7 above. Therefore, the ML estimator defined herein is consistent. *Q.E.D.*

The following lemma establishes mild ill-posedness (Assumption 6(1)).

LEMMA 8—Sufficient Condition for Mild Ill-posedness: *If the function f satisfies Assumptions* 1, 4, 5, *and* 7 *with degree $\lambda > 0$, then for any θ,*

$$\frac{1}{J^\lambda} \precsim \inf_{p\in\Theta_J-\theta,\, p\neq 0,\, \|\xi\|\le C\|p_\beta\|} \frac{\|p\|_d}{\|p\|},$$

where $\|p\|_d := |p^T\mathcal{I}p|^{\frac{1}{2}}$, p_β is the component of p related to β, ξ is the component of p related to σ, and C is some fixed constant.

Note that if we assume that there is sufficient nonlinearity in the function $\partial^{\lambda-1}g_\sigma$, then Lemma 8 holds even without the condition that $\|\xi\| \le C\|p_\beta\|$.

PROOF: Suppose f satisfies the discontinuity condition in Assumption 7 with degree of ill-posedness $\lambda > 0$, and without loss of generality, assume $c_\delta = 1$. For simplicity, we will prove the statement for a piecewise constant (0th-order) spline, but for any fixed order of spline, the proof is similar. We can then assume that the parameter vector can be written $p = (p_1^T, \ldots, p_J^T, \xi^T)^T$, where each p_j is a constant $d_x \times 1$-dimensional vector, and ξ is a $d_\sigma \times 1$-dimensional vector. Define $f_{\tau_i} := f_{\tau_i}(y|x, \sigma_0, \beta_0) = \int_{\frac{i-1}{J}}^{\frac{i}{J}} x^T f'(y - x^T\beta_0(\tau)|\sigma_0)\, d\tau$, and recall $g_\sigma(y|x, \sigma_0, \beta_0) := \int_0^1 f_\sigma(y - x^T\beta_0(\tau)|\sigma_0)\, d\tau$. Denote $l_J := (f_{\tau_1}^T, f_{\tau_2}^T, \ldots, f_{\tau_J}^T, g_\sigma^T)^T$. Then for any $p_J \in \Theta_J - \theta$, $p_J^T\mathcal{I}p_J = E[\int_{\mathbb{R}} \frac{(l_J^T p_J)^2}{g}\, dy] \ge CE[(l_J^T p_J)^2]$ for some constant $C > 0$ since g is bounded from above.

Define $c := \inf_{x\in\mathcal{X}_1,\tau\in[0,1]}(x^T\beta_0'(\tau)) > 0$, where $\mathcal{X}_1$ is a bounded open subset of $\mathcal{X}$ with non-trivial probability density of x. Let $S(\lambda) := \sum_{i=0}^{\lambda-1}\left(\binom{\lambda-1}{i}\right)^2$ be a constant that only depends on λ, where $\binom{b}{a}$ stands for the combinatorial number choosing a elements from a set with size b. Then

$$\begin{aligned}
&\lambda S(\lambda) E\left[\int_{\mathbb{R}} \left(l_J^T p_J\right)^2 dy\right] \\
&\quad = E\left[\left(\sum_{j=0}^{\lambda-1}\binom{\lambda-1}{j}^2\right)\left(\sum_{j=0}^{\lambda-1}\int_{\mathbb{R}}\left(l_J^T\left(y + \frac{jc}{2(\lambda-1)uJ}\right)p_J\right)^2 dy\right)\right],
\end{aligned}$$

where $u > 0$ is a constant that will be specified later. For convenience, we abbreviate β_0 and σ_0 as β and σ.

Define the interval $Q_J^i := [a + x^T\beta(\frac{i-1/2}{J}) - \frac{c}{2uJ}, a + x^T\beta(\frac{i-1/2}{J}) + \frac{c}{2uJ}]$. By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
&E\left[\left(\sum_{j=0}^{\lambda-1}\binom{\lambda-1}{j}^2\right)\left(\sum_{j=0}^{\lambda-1}\int_{\mathbb{R}}\left(l_J^T\left(y+\frac{jc}{2(\lambda-1)uJ}\right)p_J\right)^2 dy\right)\right]\\
&\quad\geq E_{x\in\mathcal{X}_1}\left[\int_{\mathbb{R}}\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}l_J^T\left(y+\frac{jc}{2(\lambda-1)uJ}\right)p_J\right)^2 dy\right]\\
&\quad\geq E_{x\in\mathcal{X}_1}\left[\int_{Q_J^i}\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}l_J^T\left(y+\frac{jc}{2(\lambda-1)uJ}\right)p_J\right)^2 dy\right].
\end{aligned}
$$

WLOG, we abbreviate $E_{x\in\mathcal{X}_1}$ as E since $x\in\mathcal{X}_1$ has positive probability density. By definition,

$$
\begin{aligned}
&\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}l_J^T\left(y+\frac{jc}{2(\lambda-1)uJ}\right)p_J\\
&\quad=\sum_{i=1}^{J}\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_i}\left(y+\frac{jc}{2(\lambda-1)uJ}\right)x^T p_i\right)\\
&\qquad+\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}g_\sigma\left(y+\frac{jc}{2(\lambda-1)uJ}\right)^T\xi.
\end{aligned}
$$

By the discontinuity assumption and finite differencing, for each $i=1,2,\ldots,J$,

$$
\begin{aligned}
&\frac{1}{2^{\lambda-1}}\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_i}\left(y+\frac{jc}{2(\lambda-1)uJ}\right)x^T p_i\right)\\
&\quad=\big(1+o(1)\big)f_{\tau_i}^{(\lambda-1)}\left(y+\frac{c}{4uJ}\right)\left(\frac{c}{2(\lambda-1)uJ}\right)^{\lambda-1}x^T p_i.
\end{aligned}
$$

For u being large enough and for any $i'=1,2,\ldots,J$ and $y\in Q_J^i=[a+x^T\beta(\frac{i+1/2}{J})-\frac{c}{2uJ}, a+x^T\beta(\frac{i+1/2}{J})+\frac{c}{2uJ}]$,

$$
\begin{aligned}
f_{\tau_{i'}}^{(\lambda-1)}\left(y+\frac{c}{4uJ}\right)&=\int_{(i'-1)/J}^{i'/J}f^{(\lambda)}\left(y+\frac{c}{4uJ}-x^T\beta(\tau)\right)d\tau\\
&=\int_{x^T\beta((i'-1)/J)}^{x^T\beta(i'/J)}f^{(\lambda)}\left(y+\frac{c}{4uJ}-z\right)\frac{1}{x^T\beta'\big(q(z|x)\big)}\,dz,
\end{aligned}
$$

where $q(z|x)=\inf(\tau: x^T\beta(\tau)\leq z)$ is well defined given $x^T\beta(\tau)$ is strictly monotonic.

By construction, for u large enough, $a + x^T\beta(\tau) \in Q_J^i$ implies that $\tau \in [(i-1)/J, i/J]$. Hence, for any $i' \neq i$, $\int_{(i'-1)/J}^{i'/J} f^{(\lambda)}(y + \frac{c}{4uJ} - x^T\beta(\tau))\,d\tau \leq C\frac{1}{uJ}$ for some generic constant $C > 0$, since $f^{(\lambda)}(y + \frac{c}{4uJ} - x^T\beta(\tau))$ is continuous and bounded from above. For $i' = i$,

$$\int_{(i'-1)/J}^{i'/J} f^{(\lambda)}\left(y + \frac{c}{4uJ} - x^T\beta(\tau)\right)d\tau = \int_{x^T\beta((i'-1)/J)}^{x^T\beta(i'/J)} f^{(\lambda)}\left(y + \frac{c}{4uJ} - z\right)\frac{1}{x^T\beta'\big(q(z|x)\big)}\,d,$$

when there exists a $z \in (x^T\beta((i'-1)/J), x^T\beta(i'/J))$ such that $y + \frac{c}{4uJ} - z = a$ for any $y \in Q_J^i$, which implies an integration over the Dirac delta function. Hence, for $i' = i$,

$$\left|\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_{i'}}\left(y + \frac{jc}{2(\lambda-1)uJ}\right)\right)\right| \geq \left(\frac{1}{c} - C\frac{1}{uJ}\right)\left(\frac{c}{uJ}\right)^{\lambda-1},$$

and if $i' \neq i$,

$$\left|\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_{i'}}\left(y + \frac{jc}{2(\lambda-1)uJ}\right)\right)\right| \precsim \left(\frac{c}{uJ}\right)^{\lambda}.$$

Similarly,

$$\left\|\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}g_{\sigma}\left(y + \frac{jc}{2(\lambda-1)uJ}\right)^T\right\| \precsim \left(\frac{c}{uJ}\right)^{\lambda}.$$

Accordingly, when u is chosen to be large enough, and $\|\xi\| \precsim \|p_\beta\|$, the sum

$$\begin{aligned}&\frac{1}{2^{\lambda-1}}\left(\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_i}\left(y + \frac{jc}{2(\lambda-1)uJ}\right)x^T p_i\right)\\ &\quad= \big(1 + o(1)\big)f_{\tau_i}^{(\lambda-1)}\left(y + \frac{c}{4uJ}\right)\left(\frac{c}{2(\lambda-1)uJ}\right)^{\lambda-1}x^T p_i\end{aligned}$$

is dominated by the term $\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_i}(y + \frac{jc}{2(\lambda-1)uJ})x^T p_i$ for $y \in Q_J^i$, where

$$\left|\sum_{j=0}^{\lambda-1}(-1)^j\binom{\lambda-1}{j}f_{\tau_i}\left(y + \frac{jc}{2(\lambda-1)uJ}\right)\right| \geq C\left(\frac{1}{uJ}\right)^{\lambda-1}$$

for some constant $C > 0$.

Noting that the intervals $\{Q_J^i\}$ do not intersect each other for any $J \to \infty$ and u being a large enough constant,

$$\begin{aligned}\lambda S(\lambda)E\left[\int_{\mathbb{R}}\big(l_J^T p_J\big)^2\,dy\right] &\geq S(\lambda)\sum_{i=1}^{J}E\left[\int_{Q_J^i}\sum_{j=1}^{\lambda}\left(l_J^T\left(y + \frac{jc}{2(\lambda-1)uJ}\right)p_J\right)^2 dy\right]\\ &\asymp E\left[\frac{c}{\lambda uJ}\sum_{i=1}^{J}\left(\frac{1}{(uJ)^{\lambda-1}}x_i^T p_i\right)^2\right] \asymp \frac{1}{J^{2\lambda-1}}\sum_{j=1}^{J}E\big[x_j^2 p_j^2\big] \asymp \frac{1}{J^{2\lambda}}\|p\|_2^2.\end{aligned}$$

Since $\lambda S(\lambda)$ is a constant that only depends on λ, then $p^T \mathcal{I} p \gtrsim \frac{1}{J^{2\lambda}} \|p\|_2^2$ for any $\|\xi\| \leq C\|p_\beta\|$. *Q.E.D.*

PROOF OF LEMMA 3: By Lemma 7, the set of log-likelihood functions indexed by $\widehat{\theta}_n \in \Theta$ is Donsker such that the sample-average log-likelihood converges uniformly to its population counterpart:

$$E\big[-\log g(y|x, \widehat{\theta}_n)\big] \leq E_n\big[-\log g(y|x, \widehat{\theta}_n)\big] + o_p(1).$$

By Chen (2007), there exists a $\theta_n^* \to \theta_0$ as $J_n \to \infty$ where $\theta_n^* \in \Theta_{J_n}^r$ given that $d_2(\theta_0, \Theta_{J_n}^r) = O(J_n^{-\min(p,r)})$, denoting the degree of smoothness of $\beta_0(\cdot)$ as p. Because $\widehat{\theta}_n$ is the minimizer of the negative log-likelihood, $E_n[-\log g(y|x, \widehat{\theta}_n)] \leq E_n[-\log g(y|x, \theta_n^*)]$. Again, by uniform convergence,

$$E_n\big[-\log g\big(y|x, \theta_n^*\big)\big] \leq E\big[-\log g\big(y|x, \theta_n^*\big)\big] + o_p(1) \leq E\big[-\log g(y|x, \theta_0)\big] + o_p(1),$$

where the last step used the continuity of the population log-likelihood function around θ_0. Since Θ is compact, by identification (i.e., Theorem 1), we have $\widehat{\theta}_n \xrightarrow{p} \theta_0$ as $J_n \to \infty$. *Q.E.D.*

PROOF OF LEMMA 4: By Lemma 3, we know that sieve-ML estimators for β_0 and σ_0 are consistent, that is, $\|\widehat{\sigma} - \sigma_0\| \xrightarrow{p} 0$ and $\|\widehat{\beta}_J - \beta_0\|_2 \xrightarrow{p} 0$. MLE by definition implies that $E_n[\log(g(y|x, \widehat{\beta}_J, \sigma))] \geq E_n[\log(g(y|x, \beta_J^*, \sigma_0))]$, where by construction of the sieve, there exists a β_J^* such that $\|\beta_J^* - \beta_0\|_2 \leq C J_n^{-r-1}$ for some generic constant $C > 0$. Therefore, $\|(\widehat{\beta}_J, \widehat{\sigma}) - (\beta_J^*, \sigma_0)\| \xrightarrow{p} 0$ as $J_n \to \infty$.

By Lemma 7, $\mathcal{G} = \{h(y, x, \beta(\cdot), \sigma) := \log(g(y|x, \beta(\cdot), \sigma)) | (\beta(\cdot), \sigma) \in \Theta\}$ is Donsker. Thus by stochastic equicontinuity,

$$\begin{aligned} &E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] \\ &\quad = E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] + o_p(1/\sqrt{n}), \end{aligned}$$

implying that

$$\begin{aligned} &E\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] \\ &\quad = E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - o_p(1/\sqrt{n}) \geq -o_p(1/\sqrt{n}). \end{aligned}$$

Define $G_n := \sqrt{n}(E_n - E)$. By the maximal inequality, for any $\delta > 0$,

$$\begin{aligned} &E\Big[\max_{\|\beta_J - \beta_J^*\|_2 < \delta} \big|G_n \log g(y|x, \widehat{\beta}_J, \widehat{\sigma}) - G_n \log g\big(y|x, \beta_J^*, \sigma_0\big)\big|\Big] \\ &\quad \leq K_1 \int_0^\delta \sqrt{\log N\big(r, M, \|\cdot\|_2\big)}\, dr, \end{aligned}$$

where $N(r, M, \|\cdot\|_2)$ is the covering number of r balls on M, the space of β, and K_1 is a generic constant. We want to show that $N(r, M, \|\cdot\|_2)$ is bounded by a polynomial of $1/r$. Define F as the space of univariate monotone functions mapping from $[0, 1]$ to bounded intervals depending on a matrix X and the bounds of $\beta(\cdot)$. Let $\alpha(\tau) := X^T \beta(\tau)$, where $X = (x_1, \ldots, x_{d_x})$ is a $d_x \times d_x$ invertible matrix consisting of d_x linearly independent

vectors $x \in \mathcal{X}$. Then each component of $\alpha(\tau)$ is strictly monotonic, $\alpha(\cdot)$ belongs to F, and $\beta(\tau) = (X^{-1})^T \alpha(\tau)$. Then the covering number of M will be the same as the covering number of F^{d_x} up to a multiplicative constant K_2 depending on X. Therefore, $N(r, M, \|\cdot\|_2) < K_2 N(r, F^{d_x}, \|\cdot\|_2) < K_3/r^{d_x}$ for some positive constant K_3 and

$$E\Big[\max_{\|\beta_J - \beta_J^*\|_2 < \delta} \big|G_n \log g(y|x, \widehat{\beta}_J, \widehat{\sigma}) - G_n \log g\big(y|x, \beta_J^*, \sigma_0\big)\big|\Big] \le K_4 \delta \sqrt{-\log \delta}$$

for a positive constant K_4 and δ small enough. Letting $\delta = \max(\|\widehat{\sigma} - \sigma\|, \|\widehat{\beta}_J - \beta_J^*\|)$,

$$\begin{aligned}
&E\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] \\
&\quad = E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - O_p\bigg(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\bigg) \\
&\quad \ge -O_p\bigg(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\bigg).
\end{aligned}$$

By consistency of $(\widehat{\beta}_J, \widehat{\sigma})$, $\delta \xrightarrow{p} 0$.

Since $E[\log g(y|x, \beta, \sigma)]$ is maximized at (β_0, σ_0), the Hadamard derivative of $E[\log g(y|x, \beta_0, \sigma_0)]$ with respect to $\beta \in \Theta$ is 0. By Assumption 5(2), the $\log g(\cdot|\cdot, \cdot, \cdot)$ function is twice differentiable with bounded derivatives up to the second order. Therefore, for some generic constant $C_1 > 0$,

$$\begin{aligned}
&E\big[\log g\big(y|x, \beta_J^*, \sigma_0\big)\big] - E\big[\log g(y|x, \beta_0, \sigma_0)\big] \\
&\quad \ge -C_1 \big\|\beta_J^* - \beta_0\big\|_2^2 \ge -C_1 C^2 J_n^{-2r-2} = O\bigg(\frac{1}{n}\bigg).
\end{aligned}$$

Then

$$\begin{aligned}
&E\big[\log g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big] - E\big[\log g(y|x, \beta_0, \sigma_0)\big] \\
&\quad = E\big[\log g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big] - E\big[\log g\big(y|x, \beta_J^*, \sigma_0\big)\big] \\
&\qquad + E\big[\log g\big(y|x, \beta_J^*, \sigma_0\big)\big] - E\big[\log g(y|x, \beta_0, \sigma_0)\big] \\
&\quad \ge -O_p\bigg(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\bigg) - C_1 C^2 J_n^{-2r-2} = -O_p\bigg(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\bigg),
\end{aligned}$$

where the last step used the assumption that $\frac{J_n^{2r+2}}{n} \to \infty$.

Let $z(y|x) = g(y|x, \widehat{\beta}_J, \widehat{\sigma}) - g(y|x, \beta_0, \sigma_0)$ and define $\|z(y|x)\|_1 := \int_{-\infty}^{\infty} |z(y|x)|\, dy$. Then by Pinsker's inequality conditional on each value of x,

$$\begin{aligned}
E_x\big[\big\|z(y|x)\big\|_1^2\big] &\le 2E_x\big[D\big(g(y|x, \beta_0, \sigma_0) \,\|\, g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] \\
&\le 2\big(E\big[\log\big(g(y|x, \beta_0, \sigma_0)\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big]\big) \\
&= O_p\bigg(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\bigg), \qquad \text{(D.3)}
\end{aligned}$$

where $D(P \,\|\, Q)$ is the K-L divergence between two probability distributions P and Q.

Now consider the characteristic functions of $x^T\widehat{\beta}_J(\tau)$ and $x^T\beta_0(\tau)$ conditional on x and given that $\tau \sim U[0,1]$,

$$\phi_{x\widehat{\beta}_J}(s|x) = \frac{\int_{-\infty}^{\infty} g(y|x,\widehat{\beta}_J,\widehat{\sigma})e^{isy}\,dy}{\phi_\varepsilon(s|\widehat{\sigma})} \quad \text{and} \quad \phi_{x\beta_0}(s|x) = \frac{\int_{-\infty}^{\infty} g(y|x,\beta_0,\sigma_0)e^{isy}\,dy}{\phi_\varepsilon(s|\sigma_0)}.$$

Then for any x and s, $|\phi_{x\widehat{\beta}_J}(s|x)\phi_\varepsilon(s|\widehat{\sigma}) - \phi_{x\beta_0}(s|x)\phi_\varepsilon(s|\sigma_0)| = |\int_{-\infty}^{\infty} z(y|x)e^{isy}\,dy| \le \|z(y|x)\|_1$. Defining $m(s) := \phi_\varepsilon(s|\sigma_0)/\phi_\varepsilon(s|\widehat{\sigma})$ and dividing both sides by $\phi_\varepsilon(s|\sigma)\phi_{x\beta_0}(s|x)$,

$$\left| m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right| \le \frac{\|z(y|x)\|_1}{|\phi_{x\beta_0}(s|x)\phi_\varepsilon(s|\sigma_0)|}. \tag{D.4}$$

Plugging (D.4) back into (D.3), we have

$$\begin{aligned} E_x\left[\left| m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right|^2\right] &\le E_x\left[\frac{\|z(y|x)\|_1^2}{|\phi_{x\beta_0}(s|x)\phi_\varepsilon(s|\sigma_0)|^2}\right] \\ &\le E_x\left[\|z(y|x)\|_1^2\right]\frac{s^2}{C^2\phi_\varepsilon(s|\sigma_0)^2} \\ &= o_p\left(E_x\left[\|z(y|x)\|_1^2\right]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right), \end{aligned} \tag{D.5}$$

where in the last step we require that $s \in [-l, l]$ for some $l > 0$ such that $|\phi_{x\beta_0}(s|x)|$ is bounded away from zero. Using the fact that for any random variable a and any number b, $\mathrm{Var}(a) \le E[(a-b)^2]$, we have that

$$E_x\left[\left| m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right|^2\right] \ge \mathrm{Var}_x\left(\frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}\right).$$

Inequality (D.5) then implies that

$$\mathrm{Var}_x\left(\frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}\right) \precsim_p E_x\left[\|z(y|x)\|_1^2\right]\frac{1}{\phi_\varepsilon(s|\sigma_0)}. \tag{D.6}$$

Applying Assumption 8, inequality (D.6) implies that

$$E_x\left[\left| \frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)} \right|^2\right] = O_p\left(E_x\left[\|z(y|x)\|_1^2\right]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right). \tag{D.7}$$

We can rewrite $m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}$ as $(m(s)-1) - \frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)}$. Using that $\frac{1}{2}a^2 - b^2 \leq (a-b)^2$ for any $a, b \in \mathbb{R}$, we can bound inequality (D.7) from below such that

$$\frac{1}{2}E_x\big[\big|m(s)-1\big|^2\big] - E\left[\left|\frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)}\right|^2\right]$$

$$\leq E_x\left[\left|m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}\right|^2\right] \tag{D.8}$$

$$= O_p\left(E_x\big[\big\|z(y|x)\big\|_1^2\big]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right). \tag{D.9}$$

Combining (D.8) with (D.7),

$$E_x\big[\big|m(s)-1\big|^2\big] = O_p\left(E_x\big[\big\|z(y|x)\big\|_1^2\big]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right),$$

or, equivalently, for any $s \in [-l, l]$ where l is some fixed constant,

$$\big|\phi_\varepsilon(s|\sigma_0) - \phi_\varepsilon(s|\widehat{\sigma})\big|^2 = O_p\big(E_x\big[\big\|z(y|x)\big\|_1^2\big]\big). \tag{D.10}$$

Applying Assumption 5(6) along with (D.10), it follows that $\|\widehat{\sigma} - \sigma_0\|^2 = O_p(E_x[\|z(y|x)\|_1^2]) = O_p(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}})$.

If $\|\widehat{\sigma} - \sigma\| > \|\widehat{\beta}_J - \beta_J^*\|$, then $\delta = \|\widehat{\sigma} - \sigma\|$, and it follows that $\|\widehat{\sigma} - \sigma_0\|^2 = O_p(\frac{\log n}{n})$.

If $\|\widehat{\sigma} - \sigma\| \leq \|\widehat{\beta}_J - \beta_J^*\|$, then $\delta = \|\widehat{\beta}_J - \beta_J^*\|$, and $\|\widehat{\sigma} - \sigma_0\|^2 = O_p(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}})$.

Therefore, $\|\widehat{\sigma} - \sigma_0\|^2 = O_p(\max(\frac{\log n}{n}, \frac{\|\widehat{\beta}_J - \beta_J^*\|\sqrt{-\log\|\widehat{\beta}_J - \beta_J^*\|}}{\sqrt{n}}))$. *Q.E.D.*

The following lemma will be instrumental in proving asymptotic normality.

LEMMA 9: *Under Assumptions* 1, 4, 5, 6(1) (*the mildly ill-posed case*), 8, *and* $\frac{J_n^{2r+2}}{n} \to \infty$, $\|\widehat{\beta}_J - \beta_J^*\| = O_p((J_n^{2\lambda}\frac{\log^{\frac{1}{2}} n}{n^{\frac{1}{2}}})^{\frac{1}{2\lambda+1}})$.

PROOF: Our argument follows the proof of Lemma 4. Let $z(y|x) := g(y|x, \widehat{\beta}_J, \sigma_0) - g(y|x, \beta_J^*, \sigma_0)$. By Pinsker's inequality conditional on each value of x,

$$E_x\big[\big\|z(y|x)\big\|_1^2\big] \leq 2E_x\big[D\big(g(y|x, \widehat{\beta}_J, \sigma_0) \parallel g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big]$$

$$\leq 2\big(E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big]\big),$$

where $D(P \parallel Q)$ is the K-L divergence between two probability distributions P and Q.

By the maximal inequality, for any $\delta > 0$,

$$E\Big[\max_{\|\beta_J - \beta_J^*\|_2 < \delta}\big|G_n \log g\big(y|x, \beta_J^*, \sigma_0\big) - G_n \log g(y|x, \beta_J, \sigma_0)\big|\Big]$$

$$\leq K\int_0^\delta \sqrt{\log N\big(r, M, \|\cdot\|_2\big)}\, dr,$$

where $N(r, M, \|\cdot\|_2)$ is the covering number of r balls on M (the space of β) and K is a generic constant. Since M is a bounded and co-monotone space ($x^T\beta(\tau)$ is monotone in τ for all $x \in \mathcal{X}$), $N(r, M, \|\cdot\|_2) < \delta^{d_x}$. Therefore,

$$E\Big[\max_{\|\beta_J - \beta_J^*\|_2 < \delta} \big|G_n \log g\big(y|x, \beta_J^*, \sigma_0\big) - G_n \log g(y|x, \beta_J, \sigma_0)\big|\Big] \le \delta\sqrt{-\log\delta}$$

and $|G_n \log g(y|x, \beta_J^*, \sigma_0) - G_n \log g(y|x, \widehat{\beta}_J, \sigma_0)| = O_p(\widehat{\delta}\sqrt{-\log\widehat{\delta}})$, where $\widehat{\delta} := \|\beta_J^* - \widehat{\beta}_J\|$. Using a similar argument as in the proof of Lemma 4, we can show that $\|\widehat{\sigma} - \sigma_0\|^2 = O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}})$. Thus,

$$E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] \tag{D.11}$$

$$= \frac{1}{\sqrt{n}} G_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - \frac{1}{\sqrt{n}} G_n\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] \tag{D.12}$$

$$+ E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big]. \tag{D.13}$$

The terms in (D.12) are $O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}})$. For the terms in (D.13), we have

$$\begin{aligned} &E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] \\ &\quad = E_n\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] \\ &\qquad + E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big]. \end{aligned}$$

We know that $E_n[\log(g(y|x, \beta_J^*, \sigma_0))] - E_n[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] \le 0$ by the first-order condition. We also have

$$E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \widehat{\sigma})\big)\big] - E_n\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] = O_p\big(\|\widehat{\sigma} - \sigma_0\|^2\big) = O_p\left(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}\right).$$

Combining the results on different terms in (D.11), we have

$$E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] \precsim_p \frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}.$$

It follows that

$$E_x\big[\big\|z(y|x)\big\|_1^2\big] \le 2E\big[\log\big(g\big(y|x, \beta_J^*, \sigma_0\big)\big)\big] - E\big[\log\big(g(y|x, \widehat{\beta}_J, \sigma_0)\big)\big] = O_p\left(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}\right).$$

Now consider the characteristic functions of $x^T\widehat{\beta}_J(\tau)$ and $x^T\beta_J^*(\tau)$ conditional on x, $\tau \sim U[0, 1]$:

$$\phi_{x\widehat{\beta}_J}(s|x) = \frac{\int_{-\infty}^{\infty} g(y|x, \widehat{\beta}_J, \sigma_0)e^{isy}\,dy}{\phi_\varepsilon(s|\sigma_0)} \quad \text{and} \quad \phi_{x\beta_J^*}(s|x) = \frac{\int_{-\infty}^{\infty} g\big(y|x, \beta_J^*, \sigma_0\big)e^{isy}\,dy}{\phi_\varepsilon(s|\sigma_0)}.$$

It follows that $|\phi_{x\widehat{\beta}_J}(s|x)\phi_\varepsilon(s|\sigma_0) - \phi_{x\beta_J^*}(s|x)\phi_\varepsilon(s|\sigma_0)| = |\int_{-\infty}^{\infty} z(y|x)e^{isy}\,dy| \le \|z(y|x)\|_1$. Then

$$\left|\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)\right| \le_p \frac{\left\|z(y|x)\right\|_1}{\phi_\varepsilon(s|\sigma_0)}.$$

Using the relationship between the CDF and characteristic function of a random variable x (i.e., $F_x(w) = \frac{1}{2} - \int_{-\infty}^{\infty} \frac{\exp(iws)\phi_x(s)}{2\pi is}\,ds$), we have that

$$F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w) = \lim_{q\to\infty}\int_{-q}^{q} \frac{\exp(iws)}{2\pi is}\left(\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)\right)ds.$$

Then since in our sieve setting $\widehat{\beta}_J$ and β_J^* are rth-order spline functions with grid interval size of order $O(1/J_n)$, we know that $\max(|\phi_{x\widehat{\beta}_J}(s|x)|, |\phi_{x\beta_J^*}(s|x)|) \le J_n\frac{c}{s}$ for some constant $c > 0$. Therefore,

$$\begin{aligned} &E_x\left[\left|F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w)\right|\right] \\ &\quad\le E_x\left[\int_{-q}^{q}\left|\frac{\exp(iws)}{2\pi is}\right|\frac{\left\|z(y|x)\right\|_1}{\left|\phi_\varepsilon(s|\sigma_0)\right|}\,ds\right] + E_x\left[2\int_q^{\infty}\frac{1}{2\pi s}\left|\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)\right|ds\right]. \end{aligned} \tag{D.14}$$

The first term of the right-hand side of equation (D.14) is weakly bounded from above by

$$\frac{1}{2\pi}\int_{-q}^{q}\frac{1}{s\phi_\varepsilon(s|\sigma_0)}ds E_x\left[\left\|z(y|x)\right\|_1\right] = o_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}}\right),$$

where λ is the degree of mild ill-posedness. The second term of (D.14) is weakly bounded by $J_n\frac{4c}{q} \precsim J_n/q$. Putting these together, the right-hand side of (D.14) has an upper bound of $O_p(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}} + \frac{J_n}{q})$ for an arbitrary q.

Since $\widehat{\delta} = \|\beta_J^* - \widehat{\beta}_J\|_2 = O(E_x[|F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w)|])$, we have

$$\widehat{\delta} = O_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}} + \frac{J_n}{q}\right).$$

If $\widehat{\delta} = O_p(\frac{1}{n})$, then the conclusion holds. If $\widehat{\delta}$ converges to 0 slower than $\frac{1}{n}$, we have $\widehat{\delta} = O_p(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}\log^{\frac{1}{4}} n + \frac{J_n}{q})$, which implies $\widehat{\delta} = O_p(\frac{q^{2\lambda}}{n^{1/2}}\log^{\frac{1}{2}} n + \frac{J_n}{q})$. The optimal q is $(\frac{J_n n^{\frac{1}{2}}}{\log^{\frac{1}{2}} n})^{\frac{1}{2\lambda+1}}$. Then we have $\widehat{\delta} = \|\widehat{\beta}_J - \beta_J^*\| = O_p((J_n^{2\lambda}\frac{\log^{\frac{1}{2}} n}{n^{\frac{1}{2}}})^{\frac{1}{2\lambda+1}})$. *Q.E.D.*

PROOF OF THEOREM 2: Suppose $\widehat{\theta}_J = (\widehat{\beta}_J(\cdot), \widehat{\sigma}) \in \Theta_J^r$ is the rth-order sieve estimator. By the consistency of the sieve estimator established by Lemma 3, $\|\widehat{\theta}_J - \theta_0\|_2 \xrightarrow{p} 0$. It is easy to see that $\Theta_J^r \subset \Theta$. By Lemma 4, $\widehat{\sigma}$ will always converge to σ_0 at rate of at least $n^{-\frac{1}{4}}$. By construction of the sieve, there exists a set of parameters (β_J^*, σ_0) in Θ_J^r such that $\|\beta_J^* - \beta_0\|_2 = O(\frac{1}{J_n^{r+1}})$.

Let G_n denote the operator $\sqrt{n}(E_n - E)$. Then

$$\begin{aligned}\frac{1}{\sqrt{n}}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} &= E_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} \\ &= -E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})}, \end{aligned} \tag{D.15}$$

where we used the first-order condition $E_n(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})|_{(\widehat{\beta}_J,\widehat{\sigma})} = 0$. For the left-hand side of (D.15), by Donskerness of $\{(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})|_{(\tilde{\beta},\tilde{\sigma})} | (\tilde{\beta}, \tilde{\sigma}) \in \Theta\}$, we have

$$\frac{1}{\sqrt{n}}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} = \frac{1}{\sqrt{n}}\big(1 + o_p(1)\big)G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)},$$

which is asymptotically Gaussian. Next, we work on the right-hand side of (D.15). It can be expanded as

$$\begin{aligned}&-E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} \\ &\quad= -\left[E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)}\right] \\ &\qquad- E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)}, \end{aligned} \tag{D.16}$$

and then a Taylor expansion of the term inside brackets of (D.16) gives

$$\mathcal{I}_{\beta_J,\sigma_0}\big(\widehat{b}_J - b_J^*, \widehat{\sigma} - \sigma_0\big)^T + O_p\big(\|\widehat{b}_J - b_J^*\|^2 + \|\widehat{\sigma} - \sigma_0\|^2\big),$$

where $\widehat{b}_J$ and b_J^* denote the coefficient vectors for the spline functions in $\widehat{\beta}_J$ and β_J^*. The second term on the right-hand side of (D.16), $-E(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})|_{(\beta_J^*,\sigma_0)}$, equals

$$E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)},$$

because (β_0, σ_0) is the truth and therefore $E(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})|_{(\beta_0,\sigma_0)} = 0$.

Since $\|\beta_J^* - \beta_0\| = O(\frac{1}{J_n^{r+1}})$, by continuity

$$E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)} = O\big(\|\beta_J^* - \beta_0\|\big) = O\left(\frac{1}{J_n^{r+1}}\right).$$

Combining both sides of (D.15), we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\big(1 + o_p(1)\big)G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{(\beta_0,\sigma_0)} \\ &\quad= -\mathcal{I}_{\beta_J^*,\sigma_0}\big(\widehat{b}_J - b_J^*, \widehat{\sigma} - \sigma_0\big)^T + O_p\big(\|\widehat{b}_J - b_J^*\|^2 + \|\widehat{\sigma} - \sigma_0\|^2\big) + O_p\left(\frac{1}{J_n^{r+1}}\right). \end{aligned} \tag{D.17}$$

Since $\|\widehat{\sigma}-\sigma_0\|^2 = o_p(\frac{1}{\sqrt{n}})$, it is dominated by the Gaussian term on the left-hand side of (D.17). By Lemma 9 and the condition that $\frac{J_n^{4\lambda^2+6\lambda}\log(n)}{n}\to 0$, we know that $\|\widehat{\beta}_J-\beta_J^*\|^2 = J_n^{-\lambda}o_p(\|\widehat{\beta}_J-\beta_J^*\|)$ and $\|\widehat{b}_J-b_J^*\|^2 = J_n^{-\lambda}o_p(\|\widehat{b}_J-b_J^*\|)$, for J_n satisfying the growth rate conditions stated in the theorem. Therefore, (D.17) becomes

$$
\begin{aligned}
&-\big(1+o_p(1)\big)\mathcal{I}_{\beta_J^*,\sigma_0}\big(\widehat{b}_J-b_J^*,\widehat{\sigma}-\sigma_0\big)^T\\
&\quad=\frac{1}{\sqrt{n}}\big(1+o_p(1)\big)G_n\bigg(\frac{\partial\log g}{\partial\beta},\frac{\partial\log g}{\partial\sigma}\bigg)^T\bigg|_{(\beta_0,\sigma_0)}+O_p\bigg(\frac{1}{J_n^{r+1}}\bigg).
\end{aligned}
\tag{D.18}
$$

By continuity of the information matrix as a function of β, we know that the smallest eigenvalue of $\mathcal{I}_{\beta_J^*,\sigma_0}$ is on the same order as the smallest eigenvalue of $\mathcal{I}_{\beta_0,\sigma_0}$, that is, both are bounded by $\frac{c}{J_n^\lambda}$ from below with c as a constant. Hence (D.18) implies $\|\widehat{b}_J-b_J^*,\widehat{\sigma}-\sigma_0\| = J_n^\lambda O_p(\frac{1}{J_n^{r+1}},\frac{1}{\sqrt{n}})$, or

$$
\|\widehat{\beta}_J-\beta_0,\widehat{\sigma}-\sigma_0\| = O_p\bigg(\frac{1}{J_n^{r+1}}\bigg)+\big\|\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0\big\| = J_n^\lambda O_p\bigg(\frac{1}{J_n^{r+1}},\frac{1}{\sqrt{n}}\bigg),
$$

establishing the convergence rate of the sieve estimator. For asymptotic normality, note that if $J_n^{r+1}/\sqrt{n}\to\infty$, then the first term on the right-hand side of (D.18) dominates the second term on the right-hand side of (D.18), so we have

$$
\mathcal{I}_{\beta_J,\sigma_0}\big(\widehat{b}_J-b_J^*,\widehat{\sigma}-\sigma_0\big)^T=\frac{1}{\sqrt{n}}\big(1+o_p(1)\big)G_n\bigg(\frac{\partial\log g}{\partial\beta},\frac{\partial\log g}{\partial\sigma}\bigg)^T\bigg|_{(\beta_0,\sigma_0)}.
$$

Therefore, $\sqrt{n\kappa_J}(\widehat{b}_J-b_J^*,\widehat{\sigma}-\sigma_0)=\sqrt{\kappa_J}(1+o_p(1))\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}G_n(\frac{\partial\log g}{\partial\beta},\frac{\partial\log g}{\partial\sigma})^T|_{\beta_0,\sigma_0}$. By definition, we know that

$$
\mathcal{I}_{\beta_0,\sigma_0}=n\operatorname{Var}\bigg(G_n\bigg(\frac{\partial\log g}{\partial\beta},\frac{\partial\log g}{\partial\sigma}\bigg)\bigg|_{\beta_0,\sigma_0}\bigg).
\tag{D.19}
$$

By the growth condition $r+1>\lambda$, $\|\mathcal{I}_{\beta_0,\sigma_0}-\mathcal{I}_{\beta_J^*,\sigma_0}\|_2 = O(\frac{1}{J^r}) = o(\kappa_J)$. By the definition of κ_J as the smallest eigenvalue of $\mathcal{I}_{\beta_0,\sigma_0}$, we have

$$
\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0}-I_{\widetilde{J}}\big\|_2=\big\|\big(I_{\widetilde{J}}+(\mathcal{I}_{\beta_0,\sigma_0}-\mathcal{I}_{\beta_J^*,\sigma_0})\mathcal{I}^{-1}_{\beta_0,\sigma_0}\big)^{-1}-I_{\widetilde{J}}\big\|_2\to 0,
\tag{D.20}
$$

where $I_{\widetilde{J}}$ is the identity with dimension $\widetilde{J}\times\widetilde{J}$, and $\widetilde{J}=\dim(b_J)+d_\sigma$.

Denote $\Omega_J := \kappa_J\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0}\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}$. By (D.20), the largest eigenvalue of Ω_J is bounded by a constant, and

$$
\begin{aligned}
\big\|\Omega_J-\kappa_J\mathcal{I}^{-1}_{\beta_0,\sigma_0}\big\|_2&=\kappa_J\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0}\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}-\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\big\|_2\\
&\le\kappa_J\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0}\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}-\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\big\|_2+\kappa_J\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}-\mathcal{I}^{-1}_{\beta_0,\sigma_0}\big\|_2\\
&\le\kappa_J\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}\big\|_2\big\|\mathcal{I}_{\beta_0,\sigma_0}\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}-I_{\widetilde{J}}\big\|_2+\kappa_J\big\|\mathcal{I}^{-1}_{\beta_0,\sigma_0}\big\|_2\big\|\mathcal{I}^{-1}_{\beta_J^*,\sigma_0}-I_{\widetilde{J}}\big\|_2.
\end{aligned}
$$

By (D.19), $\|\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0} - I_{\tilde{J}}\|_2 \to 0$. By definition, $\kappa_J\|\mathcal{I}^{-1}_{\beta_0,\sigma_0}\|_2 = 1$. It is also straightforward to see that $\kappa_J\|\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}\|_2 = O(1)$. Therefore, $\|\Omega_J - \kappa_J\mathcal{I}^{-1}_{\beta_0,\sigma_0}\|_2 \to 0$. Then

$$\operatorname{Var}\left(\sqrt{\kappa_J}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}\right) = \kappa_J\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}\mathcal{I}_{\beta_0,\sigma_0}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0} = \Omega_J,$$

which has bounded eigenvalues. We next denote the submatrix of Ω_J for σ as $\Omega_{J,\sigma}$. Also, let $\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\sigma_0}$ denote the last d_σ rows of $\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}$. Let $\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\beta^*_J}$ denote the first $\dim(b_J)$ rows of $\mathcal{I}^{-1}_{\beta^*_J,\sigma_0}$. For $\widehat{\sigma} - \sigma_0$, we have

$$\sqrt{n\kappa_J}(\widehat{\sigma} - \sigma_0) = \sqrt{\kappa_J}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\sigma_0}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}.$$

Since the largest eigenvalue of Ω_J is bounded from above uniformly over J, we have that the matrix

$$\kappa_J\Omega_{J,\sigma} = \operatorname{Var}\left(\sqrt{\kappa_J}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\sigma_0}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}\right)$$

has eigenvalues bounded from above by constant and bounded away from 0 by κ_J.

For any $v \in \mathbb{R}^{d_\sigma}$, $\|v\| = 1$, we can define

$$z_n = v^T\sqrt{\kappa_J}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\sigma_0}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}$$

as a scalar random variable. Since $\operatorname{Var}(\sqrt{\kappa_J}\mathcal{I}^{-1}_{\beta^*_J,\sigma_0:\sigma_0}G_n(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})^T|_{\beta_0,\sigma_0}) = \kappa_J\Omega_{J,\sigma}$ has bounded eigenvalues, there exist positive constants $C_1, C_2 > 0$ such that $\sigma^2_{z_n} := \operatorname{Var}(z_n) \in [C_1\kappa_J, C_2]$.

For any fixed constant $\eta > 0$, we have

$$\begin{aligned}
&\frac{1}{\sigma^2_{z_n}}E\big[z_n^2 1(z_n > \eta\sqrt{n}\sigma_{z_n})\big] \\
&\quad\le \frac{1}{\sigma^2_{z_n}}E\left[\left\|\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{\beta_0,\sigma_0}\right\|^2 1\left(\left\|\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{\beta_0,\sigma_0}\right\|^2 > C_2^2\eta^2 n\sigma^2_{z_n}\right)\right],
\end{aligned}$$

where $1(\cdot)$ is the indicator function.

By Assumption 5(7) and the Markov inequality, $E[\|(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})|_{\beta_0,\sigma_0}\|^4] \le C(Jd_x + d_\sigma)^2$ for some constant $C > 0$, and

$$\begin{aligned}
&E\left[\left\|\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{\beta_0,\sigma_0}\right\|^2 1\left(\left\|\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{\beta_0,\sigma_0}\right\|^2 > C_2^2\eta^2 n\sigma^2_{z_n}\right)\right] \\
&\quad\le \frac{E\left[\left\|\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Big|_{\beta_0,\sigma_0}\right\|^4\right]}{C_2^2\eta^2 n\sigma^2_{z_n}} \le \frac{C(Jd_x + d_\sigma)^2}{C_2^2\eta^2 n\sigma^2_{z_n}}.
\end{aligned}$$

Therefore, for any fixed $\eta > 0$, $\frac{1}{\sigma^2_{z_n}}E[z_n^2 1(z_n > \eta\sqrt{n}\sigma_{z_n})] \precsim \frac{J^2}{C_2^2\eta^2 n\sigma^4_{z_n}} \precsim \frac{J^{2\lambda+2}}{n} \precsim \frac{J^{2r+2}}{n} \to 0$.

By the Lindeberg–Feller Triangular Central Limit Theorem,

$$\Omega_{J,\sigma}^{-1/2}\sqrt{\kappa_J}\mathcal{I}_{\beta_J^*,\sigma_0:\sigma_0}^{-1}G_n\left(\frac{\partial \log g}{\partial \beta},\frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}\xrightarrow{d}N(0,I_{d_\sigma}),$$

or equivalently,

$$\Omega_{J,\sigma}^{-1/2}\sqrt{n\kappa_J}(\widehat{\sigma}-\sigma_0)\xrightarrow{d}N(0,I_{d_\sigma}).$$

For convergence of $\widehat{\beta}$, we have

$$\sqrt{n\kappa_J}\big(\widehat{b}_J-b_J^*\big)=\sqrt{\kappa_J}\big(1+o_p(1)\big)\mathcal{I}_{\beta_J^*,\sigma_0:\beta_J^*}^{-1}G_n\left(\frac{\partial \log g}{\partial \beta},\frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}.$$

For any fixed τ,

$$\begin{aligned}\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_J^*(\tau)\big)&=\sqrt{n\kappa_J}\big(\widehat{b}_J-b_J^*\big)^TS^{(J)}(\tau)\\&=\sqrt{n\kappa_J}\big(1+o_p(1)\big)S^{(J)}(\tau)'\mathcal{I}_{\beta_J^*,\sigma_0:\beta_J^*}^{-1}G_n\left(\frac{\partial \log g}{\partial \beta},\frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{\beta_0,\sigma_0},\end{aligned}$$

where $S^{(J)}(\tau)=(S_1(\tau),\dots,S_{d_x\times(J+r)}(\tau))$ is the set of e base functions in the sieve space normalized such that $\|S^{(J)}(\tau)\|^2=O(1)$ as $J\to\infty$ and $\|S^{(J)}(\tau)\|^2$ denotes $\sum_{l=1}^{d_x\times(J+r)}|S_l(\tau)|^2$. Thus we have

$$\operatorname{Var}\left(\sqrt{\kappa_J}S^{(J)}(\tau)^T\mathcal{I}_{\beta_J^*,\sigma_0:\beta_J^*}^{-1}G_n\left(\frac{\partial \log g}{\partial \beta},\frac{\partial \log g}{\partial \sigma}\right)^T\Bigg|_{\beta_0,\sigma_0}\right)\le\big\|S^{(J)}\big\|^2\|\kappa_J\Omega_J\|_2,$$

which is uniformly bounded from the above for all J. We can now apply the Lindeberg–Feller Triangular Central Limit Theorem again and have

$$\Omega_{J,\tau}^{-1/2}\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_J^*(\tau)\big)\to_d N(0,I_{d_x}),$$

where $\Omega_{J,\tau}:=(S^{(J)}(\tau)\otimes I_{d_x})^T\Omega_{J,\beta}(S^{(J)}(\tau)\otimes I_{d_x})$, where $\Omega_{J,\beta}$ is the submatrix of Ω_J for β_J. Because $\|\beta_J^*-\beta_0\|=O(\frac{1}{J^{r+1}})=o(\frac{1}{\sqrt{n\kappa_J}})$, the bias term $\beta_J^*-\beta_0$ is dominated by $\widehat{\beta}_J(\tau)-\beta_J^*(\tau)$. Therefore,

$$\Omega_{J,\tau}^{-1/2}\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_0(\tau)\big)\to_d N(0,I_{d_x}).\qquad\textit{Q.E.D.}$$

PROOF OF THEOREM 3: By the same argument as in the proof of Theorem 2, we have

$$\mathcal{I}_{\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0}+O_p\big(\big\|\big(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0\big)\big\|_2^2\big)=\frac{-1}{\sqrt{n}}\mathbb{G}_{n,J_n}.\tag{D.21}$$

By setting J_n such that $\frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}}=\frac{1}{J_n}$, we have $(\frac{1}{2\lambda}-\eta)\log(n)<J_n^\zeta<\frac{1}{2\lambda}\log(n)$ for any small $\eta>0$ and n large enough. By Assumption 6(2), the minimum eigenvalue of $\mathcal{I}$ is bounded by $C\exp(-\lambda J_n^\zeta)$ for some $\lambda>0$, $\zeta>0$, and $C>0$. It follows that $\|\mathcal{I}_{\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0}\|\ge C\exp(-\lambda J_n^\zeta)\|(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)\|_2$.

(a) If $\|(\widehat{\beta}_J - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 \geq C_1/C\exp(-\lambda J_n^\zeta)$, for some constant C_1 large enough, then with probability approaching 1, we have $\|\mathcal{I}_{\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0}\|_2 > 2O_p(\|(\widehat{\beta}_J - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2^2)$, where $O_p(\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2^2)$ is the higher order residual term in the equation (D.21). It follows that $\|(\widehat{\beta}_J - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 \precsim_p \frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}}$.

(b) Else we have $\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 \leq C_1/C\exp(-\lambda J_n^\zeta) \leq C_1/C_n(1/2 - \eta\lambda) = o(\frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}})$.

Combining the two situations, we have $\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 = O_p(\frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}})$.

By construction of the sieve, $\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 = O(\frac{1}{J_n})$. Hence, $\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 = O(\max(\frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}}, \frac{1}{J_n}))$. By assumption, we set J_n such that $\frac{\exp(\lambda J_n^\zeta)}{\sqrt{n}} = \frac{1}{J_n} = O(\frac{1}{\log^{1/\zeta}(n)})$. Therefore, the sieve estimator satisfies: $\|(\beta - \beta_J^*, \widehat{\sigma} - \sigma_0)\|_2 = O_p(\frac{1}{\log^{1/\zeta}(n)})$. *Q.E.D.*

PROOF OF LEMMA 5: A bootstrap process can be considered as putting non-negative weights $w_{i,n}$ on the ith observation. We require $E[w_{i,n}] = 1$, and $E[w_{i,n}]^2 = \sigma_{w,n}^2 < \infty$. One example is to let $(w_{i,1}, \ldots, w_{i,n}) \sim \text{Multinomial}(n, \frac{1}{n}, \ldots, \frac{1}{n})$, which is the nonparametric pairs bootstrap recommended in the text and used in the simulation and empirical results. The bootstrapped estimator $(\widehat{\beta}_J^b, \widehat{\sigma}^b)$ should satisfy the first-order condition

$$E_n^b\left[\left.\frac{\partial \log g^b}{\partial \beta}\right|_{\widehat{\beta}_J^b,\widehat{\sigma}^b}, \left.\frac{\partial \log g^b}{\partial \sigma}\right|_{\widehat{\beta}_J^b,\widehat{\sigma}^b}\right] = 0.$$

By Assumption 5(3), $E[\sup_{\beta,\sigma}|w_{i,n}\log g(y_i|x_i,\beta,\sigma)|] < \infty$. Moreover, by Assumption 5(2), $w_{i,n}\log g(y_i|x_i,\beta,\sigma)$ satisfies

$$E\big[\big|w\log g(y_i|x_i,\beta,\sigma) - w'\log g\big(y_i|x_i,\beta',\sigma'\big)\big|\big] \leq C_1\big(\big|w-w'\big| + \big\|(\beta,\sigma) - \big(\beta',\sigma'\big)\big\|\big)$$

for some generic constant $C_1 > 0$. By the ULLN for any $(\beta,\sigma) \in M \times \Sigma$, $E_n^b[\log g(\beta, \sigma)] \xrightarrow{p} E_n[\log g(\beta,\sigma)]$, which converges to $E[\log g(\beta,\sigma)]$ with probability approaching 1. Since $M \times \Sigma$ is compact and identification holds, it must be that $(\widehat{\beta}_J^b, \widehat{\sigma}^b) \xrightarrow{p} (\beta_0, \sigma_0)$. Therefore, $\|(\widehat{\beta}_J^b, \widehat{\sigma}^b) - (\widehat{\beta}_J, \widehat{\sigma})\| \xrightarrow{p} 0$.

Denote $G(\beta,\sigma) = (\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})$. By stochastic equicontinuity,

$$E_n^b\big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\big] - E_n\big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\big] = E_n^b\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] - E_n\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] + o_p\left(\frac{1}{\sqrt{n}}\right).$$

In the above equation, $E_n^b[G(\widehat{\beta}_J^b, \widehat{\sigma}^b)] = E_n[G(\widehat{\beta}_J, \widehat{\sigma})] = 0$ by the first-order condition. Thus we have

$$E_n^b\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] - E_n\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] + o_p\left(\frac{1}{\sqrt{n}}\right) = -\big(E_n\big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\big] - E_n\big[G(\widehat{\beta}_J, \widehat{\sigma})\big]\big). \quad \text{(D.22)}$$

Next we will show that the left-hand side of (D.22) is asymptotically normal and the right-hand side of (D.22) can be written as a matrix multiplied by $(\widehat{b}_J^b, \widehat{\sigma}^b) - (\widehat{b}_J, \widehat{\sigma})$, where $\widehat{b}_J$ and $\widehat{b}_J^b$ are the coefficients for the sieve functions in $\widehat{\beta}_J$ and $\widehat{\beta}_J^b$.

For the left-hand side of (D.22),

$$E_n^b\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] - E_n\big[G(\widehat{\beta}_J, \widehat{\sigma})\big] = E_n^b\big[(w_{i,n} - 1)G(\widehat{\beta}_J, \widehat{\sigma})\big].$$

Because $\operatorname{Var}(\sqrt{n}(w_{i,n}-1)G(\widehat{\beta}_J,\widehat{\sigma}))=E[G(\widehat{\beta}_J,\widehat{\sigma})G(\widehat{\beta}_J,\widehat{\sigma})^T]$, we have $\sqrt{n}\frac{n}{n-1}E_n^b[(w_{i,n}-1)G(\widehat{\beta}_J,\widehat{\sigma})]=O_p(\frac{1}{\sqrt{n}})$ under the $\|\cdot\|_2$ norm. By Theorem 2, $\|(\widehat{\beta}_J,\widehat{\sigma}_J)-(\beta_0,\sigma_0)\|=O_p(\frac{J_n^\lambda}{\sqrt{n}})$. Therefore,

$$E\big[G(\widehat{\beta}_J,\widehat{\sigma})G(\widehat{\beta}_J,\widehat{\sigma})^T\big]=\mathcal{I}_{\beta_0,\sigma_0}+O_p\bigg(\frac{J_n^\lambda}{\sqrt{n}}\bigg),$$

where $\mathcal{I}_{\beta_0,\sigma_0}:=E[G(\beta_0,\sigma_0)G(\beta_0,\sigma_0)^T]$. By assumption, $\frac{J_n^\lambda}{\sqrt{n}}/\operatorname{mineigen}(\mathcal{I}_{\beta_0,\sigma_0})\to 0$ as $n\to\infty$, thus $\mathcal{I}_{\beta_0,\sigma_0}$ dominates $O_p(\frac{J_n^\lambda}{\sqrt{n}})$.

For the right-hand side of (D.22), by Donskerness of $M\times\Sigma$ and stochastic equicontinuity, we have

$$E_n\big[G\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]-E_n\big[G(\widehat{\beta}_J,\widehat{\sigma})\big]=E\big[G\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]-E\big[G(\widehat{\beta}_J,\widehat{\sigma})\big]+o_p\bigg(\frac{1}{\sqrt{n}}\bigg),$$

where the remainder term $o_p(\frac{1}{\sqrt{n}})$ does not affect the derivation further and is dropped.

By Taylor expansion,

$$\begin{aligned}
&E\big[G\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]-E\big[G(\widehat{\beta}_J,\widehat{\sigma})\big]\\
&\quad=\mathcal{I}_{\widehat{\beta}_J,\widehat{\sigma}}\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)+O\big(||\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)||^2\big)\\
&\quad=\big(\mathcal{I}_{\beta_0,\sigma_0}+O\big(\|(\widehat{b}_J-b_0,\widehat{\sigma}-\sigma_0)\|\big)\big)\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)\\
&\qquad+O\big(\|\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)\|^2\big)\\
&\quad=\big(\mathcal{I}_{\beta_0,\sigma_0}+o_p(1)\big)\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)+O\big(\|\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)\|^2\big).
\end{aligned}$$

Combining different terms in (D.22), we have

$$\begin{aligned}
&\big(1+o_p(1)\big)\mathcal{I}_{\beta_0,\sigma_0}\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)+O\big(\|\big(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}\big)\|^2\big)\\
&\quad=\frac{1}{\sqrt{n}}\frac{n-1}{n}\bigg(\sqrt{n}\frac{n}{n-1}E_n^b\big[(w_{i,n}-1)G(\widehat{\beta}_J,\widehat{\sigma})\big]\bigg)=O_p\bigg(\frac{1}{\sqrt{n}}\bigg).
\end{aligned}\tag{D.23}$$

Similarly to Theorem 2, we need to show that $\|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\|^2$ is dominated by $(1+o_p(1))\mathcal{I}_{\beta_0,\sigma_0}(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})$. Stochastic equicontinuity implies that

$$E_n^b\big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]-E_n\big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]=E_n^b\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]-E_n\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]+o_p\bigg(\frac{1}{\sqrt{n}}\bigg)$$

or

$$E_n\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]-E_n\big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]=E_n^b\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]-E_n^b\big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]+o_p\bigg(\frac{1}{\sqrt{n}}\bigg),$$

where $E_n^b[\log g(\widehat{\beta}_J,\widehat{\sigma})]-E_n^b[\log g(\widehat{\beta}_J^b,\widehat{\sigma}^b)]\le 0$ by the optimality of $(\widehat{\beta}_J^b,\widehat{\sigma}^b)$. Thus we have

$$E_n\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]-E_n\big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\big]\le o_p\bigg(\frac{1}{\sqrt{n}}\bigg).$$

We also know that $E_n[\log g(\widehat{\beta}_J, \widehat{\sigma})] - E_n[\log g(\widehat{\beta}_J^b, \widehat{\sigma}^b)] \geq 0$ by the optimality of $(\widehat{\beta}_J, \widehat{\sigma})$. Hence,

$$\left|E_n\left[\log g\left(\widehat{\beta}_J^b, \widehat{\sigma}^b\right)\right] - E_n\left[\log g(\widehat{\beta}_J, \widehat{\sigma})\right]\right| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

With this, we can apply similar arguments as in Lemma 4 and Lemma 9 to show that

$$\left\|\widehat{\sigma}^b - \widehat{\sigma}\right\| = o_p\left(n^{-\frac{1}{4}}\right),$$

$$\left\|\widehat{\beta}_J^b - \widehat{\beta}_J\right\| = O_p\left(\left(J_n^{2\lambda}\frac{\log^{\frac{1}{2}} n}{n^{\frac{1}{2}}}\right)^{\frac{1}{2\lambda+1}}\right).$$

By an argument similar to the one in Theorem 2, (D.23) implies that $\sqrt{n\kappa_J}(\widehat{\beta}_J^b - \widehat{\beta}_J)$ and $\sqrt{n\kappa_J}(\widehat{\sigma}^b - \widehat{\sigma})$ have the same distributions as $\sqrt{n\kappa_J}(\widehat{\beta}_J - \beta_0)$ and $\sqrt{n\kappa_J}(\widehat{\sigma} - \sigma_0)$. *Q.E.D.*

REFERENCES

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile Regression Under Misspecification, With an Application to the US Wage Structure," *Econometrica*, 74 (2), 539–563. [8]

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Vol. 6, 5549–5632. [13]

CHERNOZHUKOV, V., I. FERNANDEZ-VAL, AND A. GALICHON (2009): "Improving Point and Interval Estimators of Monotone Functions by Rearrangement," *Biometrika*, 96 (3), 559–575. [2]

DINARDO, J., N. M. FORTIN, AND T. LEMIEUX (1996): "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach," *Econometrica*, 64 (5), 1001–1044. [8]

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, 2111–2245. [9]

POWELL, D. (2013): "A New Framework for Estimation of Quantile Treatment Effects: Nonseparable Disturbance in the Presence of Covariates," RAND Working Paper Series WR-824-1. [8]

RUGGLES, S., K. GENADEK, R. GOEKEN, J. GROVER, AND M. SOBEK (2015): "Integrated Public Use Microdata Series Version 6.0 [Machine-Readable Database]," University of Minnesota, Minneapolis. [8]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): "Weak Convergence," in *Weak Convergence and Empirical Processes*. New York: Springer, 16–28. [9]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 2 September, 2016; final version accepted 15 October, 2020; available online 22 October, 2020.