Estimation in functional linear quantile regression

Kengo Kato
Hiroshima University
e-mail: kkato@hiroshima-u.ac.jp

Abstract: This paper studies estimation in functional linear quantile regression in which the dependent variable is scalar while the covariate is a function, and the conditional quantile for each fixed quantile index is modeled as a linear functional of the covariate. Here, we suppose that covariates are discretely observed and sampling points may differ across subjects, where the number of measurements per subject increases as the sample size. Also, we allow the quantile index to vary over a given subset of the open unit interval, so the slope function is a function of two variables: (typically) time and quantile index. Likewise, the conditional quantile function is a function of the quantile index and the covariate. We consider an estimator for the slope function based on the principal component basis. An estimator for the conditional quantile function is obtained by a plug-in method. Since the so-constructed plug-in estimator not necessarily satisfies the monotonicity constraint with respect to the quantile index, we also consider a class of monotonized estimators for the conditional quantile function. We establish rates of convergence for these estimators under suitable norms, showing that these rates are optimal in a minimax sense under some smoothness assumptions on the covariance kernel of the covariate and the slope function. Empirical choice of the cut-off level is studied by using simulations.

AMS 2000 subject classifications: 62G20.
Keywords and phrases: functional data, nonlinear ill-posed problem, principal component analysis, quantile regression.

1. Introduction

Quantile regression, initially developed by the seminal work of [29], is one of the most important statistical methods in measuring the impact of covariates on dependent variables. An attractive feature of quantile regression is that it allows us to make inference on the entire conditional distribution by estimating several different conditional quantiles. Some basic materials on quantile regression and its applications are summarized in [28].

This paper studies estimation in functional linear quantile regression in which the dependent variable is scalar while the covariate is a function, and the conditional quantile for a fixed quantile index is modeled as a linear functional of the covariate. The model that we consider is an extension of functional linear regression to the quantile regression case. Here, we suppose that covariates are discretely observed and sampling points may differ across subjects, where the...
number of measurements per subject increases as the sample size. Also, we allow the quantile index to vary over a given subset of the open unit interval, so the slope function is a function of two variables: (typically) time and quantile index. Likewise, the conditional quantile function is a function of the quantile index and the covariate. We consider the problem of estimating the slope function as well as the conditional quantile function itself. The estimator we consider for the slope function is based on the principal component analysis (PCA). Expanding the covariate and the slope function in terms of the PCA basis, the model is transformed into a quantile regression model with an infinite number of regressors. Truncating the infinite sum by the first \( m \) (say) terms, we may apply a standard quantile regression technique to estimating the first \( m \) coefficients in the basis expansion of the slope function at each quantile index, where \( m \) diverges as the sample size. In practice, the population PCA basis is unknown, so it is replaced by a suitable estimator for it. Once the estimator for the slope function is available, an estimator for the conditional quantile is obtained by a plug-in method. Since the so-constructed plug-in estimator not necessarily satisfies the monotonicity constraint with respect to the quantile index, we also consider a class of monotonized estimators for the conditional quantile function.

In summary, we have the following three types of estimators in mind:

(i) a PCA-based estimator for the slope function;
(ii) a plug-in estimator for the conditional quantile function;
(iii) monotonized estimators for the conditional quantile function.

We establish rates of convergence for these estimators under suitable norms, showing that these rates are optimal in a minimax sense under some smoothness assumptions on the covariance kernel of the covariate and the slope function.

In practice, we have to choose the cut-off level empirically. We suggest some criteria, namely (integrated-)AIC, BIC and GACV to choose the cut-off level. We study the performance of these criteria by using simulations. In our limited simulation experiments, although none of these criteria clearly dominated the others, (integrated-)BIC worked relatively stably.

Functional data have become increasingly important. We refer the reader to [34] for a comprehensive treatment on functional data analysis. Earlier theoretical studies in functional data analysis have focused mainly on functional linear mean regression models [see 9, 10, 37, 6, 22, 14, 27, 39, 15, and references cited in these papers]. Among them, [22] established fundamental results in functional linear mean regression, deriving sharp rates of convergence for a PCA-based estimator for the slope function under some smoothness assumptions. Note that they assumed that covariates are continuously observed. Other than functional linear mean regression, [33] developed estimation methods for generalized functional linear models using series expansions of covariates and slope functions.

While not many, there are some earlier papers on estimating conditional quantiles with function-valued covariates. [7] studied smoothing splines estimators for functional linear quantile regression models, while their established rates are not sharp. [18] considered nonparametric estimation of conditional quantiles.
when covariates are functions, which is a different topic than ours. [11] considered an “indirect” estimation of conditional quantiles. They modeled the conditional distribution as the composition of some (possibly unknown) link function and a linear functional of the covariate. They first estimated the conditional distribution function by adapting the method developed in [33] and then estimated the conditional quantile function by inverting the estimated conditional distribution function. In the quantile regression literature, there are two ways to estimate conditional quantiles. One is to directly model conditional quantiles and estimate unknown parameters by minimizing check functions. The other is to estimate conditional distribution functions and invert them to estimate conditional quantile functions. We refer to the former as a “direct” method while to the latter as an “indirect” method. The approach taken in this paper is classified into a direct method, while that of [11] is classified into an indirect method. Note that although their method is flexible, they only established consistency of the estimator.

Conditional quantile estimation offers a variety of fruitful applications for data containing function-valued covariates. A leading example in which conditional quantile estimation with function-valued covariates is useful appears in analysis of growth data [11]. Suppose that we have a growth data set of girls’ heights between age 1 and 18, say, where multiple measurements may occur at some ages. Use girl’s growth history between age 1 and 12 as a covariate, and her height at age 18 as a response. Then, conditional quantile estimation gives us an overall picture of the predictive distribution of girl’s height at age 18 given her growth history between age 1 and 12, which is more informative than just knowing the mean response. In addition to growth data, functional quantile regression has been applied in analysis of ozone pollution data [8] and EL Niño data [18]. We believe that functional linear quantile regression modeling is a benchmark modeling in conditional quantile estimation when covariates are functions, just as linear quantile regression modeling is so when covariates are vectors.

Our estimator for the slope function (at a fixed quantile index) can be understood as a regularized solution to an empirical version of a nonlinear ill-posed inverse problem that corresponds to the “normal equation” in the quantile regression case, where the regularization is controlled by the cut-off level. The paper is thus in part related to the literature on statistical nonlinear inverse problems, which is still an ongoing research area [see 4, 26, 31, 12, 19]. On the other hand, in the mean regression case, the normal equation becomes a linear ill-posed inverse problem. [22] considered two regularized estimators for the slope function based on the normal equation in the mean regression case. Conceptually, the problems handled in our and their papers are different in their nature: linearity and nonlinearity.

From a technical point of view, establishing sharp rates of convergence for our estimators is challenging. Our proof strategy builds upon the techniques developed in the asymptotic analysis for M-estimators with diverging numbers of parameters [see, for example, 23]. However, the additional complication arises essentially because “regressors” here are estimated ones and the estimation error
has to be taken into account, which requires some new techniques. Additionally, discretization errors bring a further technical complication.

Finally, the setting here is similar to Section 3 of [14]: covariates are densely but discretely observed, and the discretization error is taken into account in the analysis. However, the paper does not cover the case in which covariates are discretely observed with measurement errors because of the technical complication. A formal theoretical analysis in such a case is left in a future work.

The remainder of the paper is organized as follows. Section 2 presents the model and estimators. Section 3 gives the main results in which we derive rates of convergence for the estimators. Section 4 discusses empirical choice of the cut-off level. Proofs of the main results are given in Sections 5 and 6. Some technical results are provided in Appendix.

Notation: For $z \in \mathbb{R}^k$, let $\|z\|_2$ denote the Euclidean norm of $z$. For any integer $k \geq 2$, let $S^{k-1}$ denote the set of all unit vectors in $\mathbb{R}^k$: $S^{k-1} = \{ z \in \mathbb{R}^k : \|z\|_2 = 1 \}$. For any $y, z \in \mathbb{R}$, let $y \vee z = \max\{y, z\}$ and $y \wedge z = \min\{y, z\}$. Let $1(\cdot)$ denote the indicator function. For any given (random or non-random, scalar or vector) sequence $(z_i)_{i=1}^n$, $E_n[z_i] = n^{-1} \sum_{i=1}^n z_i$, which should be distinguished from the population expectation $E[\cdot]$. For any two sequences of positive constants $r_n$ and $s_n$, we write $r_n \asymp s_n$ if the ratio $r_n/s_n$ is bounded and bounded away from zero. Let $L_2[0, 1]$ denote the usual $L_2$ space with respect to the Lebesgue measure for functions defined on $[0, 1]$. Let $\| \cdot \|$ denote the $L_2$-norm: $\|f\|^2 = \int_0^1 f^2(t)dt$. For any finite set $I$, $\text{Card}(I)$ denotes the cardinality of $I$.

2. Methodology

2.1. Functional linear quantile regression modeling

Let $(Y, X)$ be a pair of a scalar random variable $Y$ and a random function $X = (X(t))_{t \in T}$ on a bounded closed interval $T$ in $\mathbb{R}$. Without loss of generality, we assume $T = [0, 1]$. By “random function”, we mean that $X(t)$ is a random variable for each $t \in [0, 1]$. We assume a mild regularity condition on the path property of $X$. Let $D[0, 1]$ denote the space of all càdlàg functions on $[0, 1]$, equipped with the Skorohod metric [see 3]. We assume that the map $t \mapsto X(t)$ is càdlàg almost surely. Equip $D[0, 1]$ with the Borel $\sigma$-field. Then, $X$ can be taken as a $D[0, 1]$-valued random variable. Since $D[0, 1]$ is a Polish space, and the product space $\mathbb{R} \times D[0, 1]$ with the product metric is also Polish, the regular conditional distribution of $Y$ given $X$ exists.

Let $Q_{Y|X}(\cdot \mid X)$ denote the conditional quantile function of $Y$ given $X$. Let $U$ be a given subset of $(0, 1)$ that is away from 0 and 1, i.e., for some small $\epsilon \in (0, 1/2)$, $U \subset [\epsilon, 1 - \epsilon]$. For each $u \in U$, we assume that $Q_{Y|X}(u \mid X)$ can be written as a linear functional of $X$, i.e., for each $u \in U$, there exist a scalar constant $a(u) \in \mathbb{R}$ and a scalar function $b(\cdot, u) \in L_2[0, 1]$ such that

$$Q_{Y|X}(u \mid X) = a(u) + \int_0^1 b(t, u)X(t)dt, \quad u \in U,$$

(1)
where \( X^c(t) = X(t) - \mathbb{E}[X(t)] \). Typical examples of \( \mathcal{U} \subset (0,1) \) are: (i) \( \mathcal{U} = \{u\} \) (singleton); (ii) \( \mathcal{U} = \{u_1, \ldots, u_K\} \) with \( 0 < u_1 < \cdots < u_K < 1 \) (finite set); (iii) \( \mathcal{U} = [u_L, u_U] \) with \( 0 < u_L < u_U < 1 \) (bounded closed interval). Formally, we allow for all these possibilities.

The model (1) is a natural extension of standard linear quantile regression models to function-valued covariates, and was first formulated in [7]. In what follows, we consider to estimate the slope function \((t,u) \mapsto b(t,u)\) and the conditional quantile function \((u,x) \mapsto Q_{Y\mid X}(u \mid x)\) as well.

### 2.2. Estimation strategy

We base our estimation strategy on the principal component analysis (PCA). Define the covariance kernel \( K(s,t) = \text{Cov}(X(s), X(t)) \). Then, by the Hilbert-Schmidt theorem, \( K(s,t) \) admits the spectral expansion

\[
K(s,t) = \sum_{j=1}^{\infty} \kappa_j \phi_j(s) \phi_j(t), \quad \kappa_1 \geq \kappa_2 \geq \cdots \geq 0,
\]

where \( \{\phi_j\}_{j=1}^{\infty} \) is an orthonormal basis for \( L_2[0,1] \). We will later assume that there are no ties in \( \kappa_j \), i.e., \( \kappa_1 > \kappa_2 > \cdots > 0 \). Since \( \{\phi_j\}_{j=1}^{\infty} \) is an orthonormal basis for \( L_2[0,1] \), we have the following expansions in \( L_2[0,1] \):

\[
X^c(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad b(t,u) = \sum_{j=1}^{\infty} b_j(u) \phi_j(t),
\]

where \( \xi_j \) and \( b_j(u) \) are defined by

\[
\xi_j = \int_0^1 X^c(t) \phi_j(t) dt, \quad b_j(u) = \int_0^1 b(t,u) \phi_j(t) dt.
\]

Here, \( \xi_j \) are called “principal scores” for \( X \). The expansion for \( X^c \) is called “Karhunen-Loeve expansion”. This leads to the expression \( \int_0^1 b(t,u) X^c(t) dt = \sum_{j=1}^{\infty} b_j(u) \xi_j \). Then, the model (1) is transformed into a quantile regression model with an infinite number of “regressors”:

\[
Q_{Y\mid X}(u \mid x) = a(u) + \sum_{j=1}^{\infty} b_j(u) \xi_j, \quad u \in \mathcal{U}.
\] (2)

Note that \( \mathbb{E}[\xi_j] = 0, \mathbb{E}[\xi_j^2] = \kappa_j \) and \( \mathbb{E}[\xi_j \xi_k] = 0 \) for all \( j \neq k \).

We first consider to estimate the slope function \((t,u) \mapsto b(t,u)\). To this end, we estimate the function \( b(\cdot,u) \) for each \( u \in \mathcal{U} \) and collect them to construct a final estimator for \((t,u) \mapsto b(t,u)\). To explain the basic idea, suppose for a while that (i) \( X \) were continuously observable; and (ii) the covariance kernel \( K(s,t) \) were known. The problem then reduces to finding suitable estimates of the coefficients \( b_j(u) \). Let \((Y_1, X_1), \ldots, (Y_n, X_n)\) be independent copies of \((Y, X)\).
For each \(i = 1, \ldots, n\), let \(\xi_{ij}\) be the principal scores for \(X_i\). Pick any \(u \in U\). Then, a plausible approach to estimating \(b(\cdot, u)\) is to truncate \(\sum_{j=1}^{\infty} b_j(u)\xi_j\) by \(\sum_{j=1}^{m} b_j(u)\xi_j\) for some large \(m\), and estimate only the first \(m\) coefficients \(b_1(u), \ldots, b_m(u)\) using a standard quantile regression technique. Let \(m = m_n\) be the “cut-off” level such that \(1 \leq m \leq n - 1\) and \(m \to \infty\) as \(n \to \infty\). Estimate \(a(u)\) and the first \(m\) coefficients \(b_1(u), \ldots, b_m(u)\) of \(b(\cdot, u)\) by

\[
(\hat{a}(u), \hat{b}(u), \ldots, \hat{b}_m(u)) = \arg \min_{a,b_1,\ldots,b_m} \mathbb{E}_n[\rho_a(Y_i - a - \sum_{j=1}^m b_j \xi_{ij})],
\]

where \(\rho_a(y) = \{u - 1(y \leq 0)\}y\) is the check function. Note that for \(u = 0.5\), \(\rho_{0.5}(\cdot)\) is equivalent to the absolute value function. Here, recall that \(\mathbb{E}_n[z_i] = n^{-1} \sum_{i=1}^{n} z_i\) for any sequence \(\{z_i\}_{i=1}^{n}\). The resulting estimator for the slope function \((t,u) \mapsto \hat{b}(t,u)\) is given by

\[
\hat{b} : (t,u) \mapsto \hat{b}(t,u), \quad \hat{b}(t,u) = \sum_{j=1}^{m} \hat{b}_j(u) \phi_j(t), \quad t \in [0,1], u \in U.
\]

However, this “estimator” is infeasible since (i) \(X\) is usually discretely observed; and (ii) \(K(s,t)\) is unknown. Because of (i), it is usually not possible to directly estimate the covariance kernel \(K(s,t)\) by the empirical one (since \(\mathbb{E}_n[(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))]\) with \(\bar{X}(t) = n^{-1} \sum_{i=1}^{n} X_i(t)\) is unavailable for some \((s,t)\)). Similarly to [14], we consider the following setting:

1. For each \(i = 1, \ldots, n\), \(X_i\) is only observed at \(L_i+1\) discrete points \(0 = t_{i1} < t_{i2} < \cdots < t_{iL_i+1} = 1\). Typically, \(\max_{1 \leq i \leq n} \max_{1 \leq l \leq L_i} (t_{i,l+1} - t_{il}) \to 0\) as \(n \to \infty\) is assumed.

2. Based on the discrete observations, for each \(i = 1, \ldots, n\), we construct an interpolated function \(\hat{X}_i = (\hat{X}_i(t))_{t \in [0,1]}\) for \(X_i = (X_i(t))_{t \in [0,1]}\).

Here, we shall use a simple interpolation rule (see also the later remark):

\[
\hat{X}_i(t) = \sum_{l=1}^{L_i} X(t_{il})1(t \in [t_{il}, t_{il+1})), \quad i = 1, \ldots, n.
\]

The observed time points \(t_{i1}, \ldots, t_{iL_i+1}\) (and \(L_i\)) should be indexed by the sample size \(n\); however it is suppressed for the notational convenience. Suppose now that the interpolated functions \(\hat{X}_1, \ldots, \hat{X}_n\) are obtained. Then, we may estimate the covariance kernel \(K(s,t)\) by

\[
\hat{K}(s,t) = \mathbb{E}_n[(\hat{X}_i(s) - \bar{X}(s))(\hat{X}_i(t) - \bar{X}(t))],
\]

where \(\bar{X}(t) = n^{-1} \sum_{i=1}^{n} \hat{X}_i(t)\). Let \(\hat{K}(s,t) = \sum_{j=1}^{\infty} \hat{k}_j \phi_j(s) \hat{\phi}_j(t)\) be the spectral expansion of \(\hat{K}(s,t)\) where \(\hat{k}_1 \geq \hat{k}_2 \geq \cdots \geq 0\) and \(\{\hat{\phi}_j\}_{j=1}^{\infty}\) is an orthonormal basis for \(L_2[0,1]\). Each principal score \(\xi_{ij}\) is estimated by

\[
\hat{\xi}_{ij} = \int_0^1 (\hat{X}_i(t) - \bar{X}(t)) \hat{\phi}_j(t) dt.
\]
Then, the coefficients \( a(u) \) and \( b_1(u), \ldots, b_m(u) \) are estimated by

\[
(\hat{a}(u), \hat{b}_1(u), \ldots, \hat{b}_m(u)) = \arg \min_{a,b_1,\ldots,b_m} \mathbb{E}_n[\rho_u(Y_i - a - \sum_{j=1}^m b_j \hat{\xi}_{ij})].
\]

(4)

The resulting estimator for the slope function \((t,u) \mapsto b(t,u)\) is given by

\[
\hat{b}: (t,u) \mapsto \hat{b}(t,u), \hat{b}(t,u) = \sum_{j=1}^m \hat{b}_j(u) \hat{\phi}_j(t), \quad t \in [0,1], u \in \mathcal{U}.
\]

The optimization problem (4) can be transformed into a linear programming problem and can be solved by using standard statistical softwares. Once the estimator for the slope function is obtained, the conditional \( u \)-quantile of \( Y \) given \( X = x \) for a given function \( x = (x(t))_{t \in [0,1]} \in L^2[0,1] \) is estimated by a plug-in method:

\[
\hat{Q}_{Y|X}(u \mid x) = \hat{a}(u) + \int_0^1 \hat{b}(t,u)(x(t) - \tilde{\bar{X}}(t))dt.
\]

Empirical choice of the cut-off level will be discussed in Section 4.

The basis \( \{\phi_j\}_{j=1}^\infty \) is called the (population) PCA basis. Alternatively, one may use other basis functions independent of the data, such as Fourier and Wavelet bases, in which case the analysis becomes more tractable. A potential drawback of using such basis functions is that, as discussed in \cite{15}, using the “first” \( m \) basis functions is less motivated. The PCA basis is a benchmark basis in functional data analysis, which is the reason why the PCA basis is used in this paper. Other estimation methods such as smoothing splines \cite{14} and a reproducing kernel Hilbert space approach \cite{39} could be adapted in the quantile regression case, which is left as a future topic.

The interpolation rule used here may be replaced by any other reasonable interpolation rule. For example, a plausible alternative is to use

\[
\tilde{X}_i^{mid}(t) = \sum_{i=1}^{L_i} X(t_i) + \frac{X(t_{i,L_i+1})}{2}1(t \in [t_i, t_{i,L_i+1})), \quad i = 1, \ldots, n.
\]

It is not hard to see that the theory below also applies to this interpolation rule. In practice, this interpolation rule may be more recommended since it uses all the discrete observations \( X_i(t_{i1}), \ldots, X_i(t_{i,L_i+1}) \).

2.3. Connection to nonlinear ill-posed inverse problems

For any fixed \( u \in \mathcal{U} \), our estimator \( \hat{b}(\cdot,u) \) can be understood as a regularized solution to an empirical version of a nonlinear inverse problem that corresponds to the “normal equation”:

\[
\mathcal{A}(u, b(\cdot,u)) = 0,
\]

(5)
where the map \( A : \mathcal{U} \times L_2[0,1] \rightarrow L_2[0,1] \) is defined by
\[
A(u, g)(\cdot) = \mathbb{E}[\{u - 1(Y \leq \int_0^1 g(t)X^c(t)dt)\}X^c(\cdot)],
\]
\[
eq \mathbb{E}[\{u - F_{Y|X}(\int_0^1 g(t)X^c(t)dt \mid X)\}X^c(\cdot)], \quad u \in \mathcal{U}, g \in L_2[0,1].
\]
Here, \( F_{Y|X}(y|X) \) denotes the conditional distribution function of \( Y \) given \( X \). For the sake of simplicity, we have ignored the constant term. Observe that for any fixed \( u \in \mathcal{U} \), the map \( A(u, \cdot) : L_2[0,1] \rightarrow L_2[0,1] \) is a nonlinear operator. In fact, using an approximation \( X^c_i \approx \sum_{j=1}^{\infty} \hat{\xi}_{ij} \phi_j =: \hat{X}_i \), our estimator \( \hat{b}(\cdot, u) \) is an approximate solution to an empirical version of (5) over the linear subspace spanned by \( \{\hat{\phi}_1, \ldots, \hat{\phi}_m\} \):
\[
\hat{A}(u, \hat{b}(\cdot, u)) \approx 0,
\]
where the map \( \hat{A} : \mathcal{U} \times L_2[0,1] \rightarrow L_2[0,1] \) is defined by
\[
\hat{A}(u, g)(\cdot) = \mathbb{E}_n[\{u - 1(Y_i \leq \int_0^1 g(t)\hat{X}^c_i(t)dt)\}\hat{X}^c_i(\cdot)], \quad u \in \mathcal{U}, g \in L_2[0,1].
\]
To see (6), observe that
\[
\hat{A}(u, \hat{b}(\cdot, u))(\cdot) = \sum_{j=1}^m \mathbb{E}_n[\{u - 1(Y_i \leq \sum_{k=1}^m \hat{\xi}_{ik}\hat{b}_k(u))\}\hat{\xi}_{ij}] \phi_j(\cdot).
\]
The first order condition to (4) implies that
\[
\mathbb{E}_n[\{u - 1(Y_i \leq \sum_{k=1}^m \hat{\xi}_{ik}\hat{b}_k(u))\}] \hat{\xi}_{ij} \approx 0, \quad 1 \leq j \leq m,
\]
which leads to (6) [the discussion here is informal to give an intuition behind our estimator]. Note that solving (4) is computationally more appealing than directly searching a solution to (6) as the former problem is convex while the latter is not.

Meanwhile, as long as the map \( y \mapsto F_{Y|X}(y|x) \) is continuous, for any fixed \( u \in \mathcal{U} \), the nonlinear inverse problem (5) is locally ill-posed at \( b(\cdot, u) \) in the sense of Hofmann and Scherzer [24, Definition 1.1], i.e., there exists a sequence of functions \( \{g_n\} \) in a neighborhood of \( b(\cdot, u) \) (in \( L_2[0,1] \)) such that \( A(u, g_n) \rightarrow A(u, b(\cdot, u)) \) but \( g_n \not\rightarrow b(\cdot, u) \) in the \( L_2 \)-norm. To see this, take a sequence of functions \( \{g_n\} \) in a neighborhood of \( b(\cdot, u) \) such that \( g_n \not\rightarrow b(\cdot, u) \) but \( g_n \xrightarrow{w} b(\cdot, u) \) in the \( L_2 \)-norm, where \( \xrightarrow{w} \) means the weak convergence in \( L_2[0,1] \). Then, by the weak convergence, we have
\[
\int_0^1 g_n(t)X^c(t)dt \rightarrow \int_0^1 b(t, u)X^c(t)dt.
\]
By the continuity of the map \( y \mapsto F_{Y|X}(y|X) \), (7) implies that \( A(u, g_n) \rightarrow A(u, b(\cdot, u)) \) despite \( g_n \not\rightarrow b(\cdot, u) \). This suggests that any sensible estimation procedure based on the normal equation (5) has to involve some regularizations. In our case, the regularization is done by restricting the parameter space for \( b(\cdot, u) \) to a sequence of finite dimensional subspaces, where the cut-off level \( m \) plays a role of regularization parameter.
2.4. Monotonization

Suppose in this section that \( \mathcal{U} \) is a bounded closed interval: \( \mathcal{U} = [u_L, u_U] \) with \( 0 < u_L < u_U < 1 \). The conditional quantile function \( Q_{Y|X}(u \mid x) \) is monotonically nondecreasing in \( u \). However, the plug-in estimator \( \hat{Q}_{Y|X}(u \mid x) \) constructed is not necessarily so. To circumvent this problem, we may monotonize \( u \) tonally nondecreasing in \( u \).\( u \), also measures the difficulty of estimating the slope function (Assumption (A3), \( \alpha > 0 \)).

Suppose in this section that \( \mathcal{U} \) is a bounded closed interval: \( \mathcal{U} = [u_L, u_U] \) with \( 0 < u_L < u_U < 1 \). The conditional quantile function \( Q_{Y|X}(u \mid x) \) is monotonically nondecreasing in \( u \). However, the plug-in estimator \( \hat{Q}_{Y|X}(u \mid x) \) constructed is not necessarily so. To circumvent this problem, we may monotonize \( u \) tonally nondecreasing in \( u \).\( u \), also measures the difficulty of estimating the slope function (Assumption (A3), \( \alpha > 0 \)).

Monotonization estimate is at least as good as the initial estimate \( \hat{Q}_{Y|X}(u \mid x) \) given above. Then, we have for all \( q \in [1, \infty] \),

\[
\left[ \int_{\mathcal{U}} \left| \hat{Q}_{Y|X}^q(u \mid x) - Q_{Y|X}(u \mid x) \right|^q \, du \right]^{1/q} \leq \left[ \int_{\mathcal{U}} \left| \hat{Q}_{Y|X}^q(u \mid x) - Q_{Y|X}(u \mid x) \right|^q \, du \right]^{1/q},
\]

where an obvious modification is made when \( q = \infty \).

3. Rates of convergence

In this section, we derive rates of convergence for the estimators defined in the previous section, and argue their optimality. We make the following assumptions. Let \( C_1 > 1 \) be some sufficiently large constant. First of all, we assume:

(A1) \( \{Y_i, X_i\}_{i=1}^\infty \) is i.i.d. with \( (Y, X) \).
(A2) \( \int_1^\infty \mathbb{E}[X^4] \, dt \leq C_1 \) and \( \mathbb{E}[\xi_j^4] \leq C_1 \kappa_j^2 \) for all \( j \geq 1 \).

The i.i.d. assumption is conventional. It is beyond the scope of the paper to extend the theory to dependent data. Note that [25] discussed weakly dependent functional data. Assumption (A2) is a mild moment restriction.

(A3) For some \( \alpha > 1 \), \( C_1^{-1} j^{-\alpha} \leq \kappa_j \leq C_1 j^{-\alpha} \) and \( \kappa_j - \kappa_{j+1} \geq C_1^{-1} j^{-\alpha-1} \) for all \( j \geq 1 \).
(A4) For some \( \beta > \alpha/2 + 1 \), \( \sup_{u \in \mathcal{U}} |b_j(u)| \leq C_1 j^{-\beta} \) for all \( j \geq 1 \).
(A5) Let \( F_{Y|X}(y \mid X) \) denote the conditional distribution function of \( Y \) given \( X \). Then, the map \( y \mapsto F_{Y|X}(y \mid X) \) is twice continuously differentiable with \( f_{y|X}(y \mid X) = \partial F_{y|X}(y \mid X) / \partial y \) and \( f_{y|X}'(y \mid X) = \partial f_{y|X}(y \mid X) / \partial y \).

Furthermore, \( f_{y|X}(y \mid X) \vee |f_{y|X}'(y \mid X)| \leq C_1 \).

(A6) \( \inf_{u \in \mathcal{U}} f_{y|X}(Q_{Y|X}(u \mid X) \mid X) \geq C_1^{-1} \).

Assumptions (A3) and (A4) are adapted from (3.2) and (3.3) of [22]. In assumption (A3), \( \alpha \) measures the smoothness of the covariance kernel \( K \), which also measures the difficulty of estimating the slope function \( (t, u) \mapsto b(t, u) \). The second part of assumption (A3) is to require the spacings among \( \kappa_j \) not to be too
small, which ensures identifiability of eigenfunctions \( \phi_j \) and thereby sufficient estimation accuracy of \( \phi_j \). Assumption (A4) determines the “smoothness” of the function \( t \mapsto b(t,u) \). The condition that \( \beta > \alpha/2 + 1 \) requires the function \( t \mapsto b(t,u) \) to be sufficiently smooth relative to \( K \) uniformly in \( u \in U \). See Hall and Horowitz [22, p.74] for some related discussions on these assumptions. Assumptions (A5) and (A6) are specific to the quantile regression case. Both assumptions are standard in the quantile regression literature when \( X \) is a vector. Assumption (A6) ensures sufficient identifiability of the conditional \( u \)-quantiles for \( u \in U \).

(A7) For each \( i = 1, \ldots, n \), \( X_i \) is observed only at discrete points \( 0 = t_{i1} < t_{i2} < \cdots < t_{iL_i+1} = 1 \). Define \( \Delta = \Delta_n = \max_{1 \leq i \leq n} \max_{1 \leq l \leq L_i} \{t_{i,l+1} - t_{il}\} \). Then, \( \Delta \to 0 \) as \( n \to \infty \).

(A8) There exists a constant \( \gamma \in (0,2] \) such that \( K(t,t) - 2K(s,t) + K(s,s) \leq C_1(t-s)\gamma \) for all \( s,t \in [0,1] \) with \( s < t \).

Assumptions (A7) and (A8) are a set of sampling assumptions on \( X_i \). A rate restriction will be imposed on \( \Delta \). Assumption (A7) in particular requires \( \min_{1 \leq i \leq n} L_i \to \infty \) as \( n \to \infty \), which means that each set of discrete points \( t_{i1}, \ldots, t_{iL_i+1} \) has to be dense in \([0,1]\) as the sample size grows. Assumption (A8) is an additional assumption on the smoothness of the covariance kernel. For example, \( \gamma = 1 \) if \( K(s,t) \) is Lipschitz continuous and \( \gamma = 2 \) if \( K(s,t) \) is twice continuously differentiable. The value of \( \gamma \) controls the discretization error. Note that possible values of \( \gamma \) depend on the value of \( \alpha \). Typically, if \( \alpha \) is sufficiently large, (A8) is satisfied with \( \gamma = 2 \). Assumption (A8) is similar in spirit to (A2) of [14], in which they directly assumed some smoothness of the random function \( t \mapsto X(t) \) to deal with the discretization error (roughly speaking, their \( 2\kappa \) corresponds to our \( \gamma \)).

Let \( F = \mathcal{F}(C_1, \alpha, \beta, \gamma) \) denote the set of all distributions of \((Y,X)\) compatible with assumptions (A2)-(A6) and (A8) for a given (admissible) values of \( C_1, \alpha, \beta \) and \( \gamma \) (such that \( \mathcal{F} \neq \emptyset \)). The following theorem, which will be proved in Section 5 below, establishes rates of convergence for the slope estimator \((t,u) \mapsto \hat{b}(t,u)\).

**Theorem 1.** Suppose that assumptions (A1)-(A8) are satisfied. Take \( m \asymp n^{1/(\alpha+2\beta)} \). Then, we have

\[
\lim_{M \to \infty} \limsup_{n \to \infty} \sup_{F \in \mathcal{F}} \mathbb{P}_F \left[ \sup_{u \in U} \int_0^1 \{\hat{b}(t,u) - b(t,u)\}^2 dt > M n^{-(2\beta-1)/(\alpha+2\beta)} \right] = 0, \tag{9}
\]

provided that \((n \vee (\log n)m^{3\alpha+3})\Delta^\gamma = O(1) \) as \( n \to \infty \).

Inspection of the proof of Theorem 1 shows that, if \( X \) were continuously observable, under assumptions (A1)-(A6), the rate of convergence of the estimator based on the direct empirical covariance kernel will be \( n^{-(2\beta-1)/(\alpha+2\beta)} \). The side condition that \((n \vee (\log n)m^{3\alpha+3})\Delta^\gamma = O(1) \) is assumed to make the discretization error negligible. This condition seems not quite restrictive. For example, if \( \beta \geq \alpha + 3/2 \) and \( \gamma = 2 \), it is satisfied as long as \( \Delta = O((n \log n)^{1/2}) \), which seems to be mild in view of some applications in functional data analysis.
The following theorem, which will be proved in Section 6, establishes rates of convergence for $Q_{Y|X}(u \mid x)$. For the notational convenience, define

$$\mathcal{E}(\hat{Q}_{Y|X}, u) = \int \{\hat{Q}_{Y|X}(u \mid x) - Q_{Y|X}(u \mid x)\}^2 dP_X(x),$$

where $P_X$ denotes the distribution of $X$ (defined on $D[0, 1]$).

**Theorem 2.** Suppose that assumptions (A1)-(A8) are satisfied. Take $m \asymp n^{1/(\alpha+2\beta)}$. Then, we have

$$\lim_{M \to \infty} \lim_{n \to \infty} \sup_{F \in \mathcal{F}} \sup_{u \in \mathcal{U}} \mathbb{P} \left( \sup_{u \in \mathcal{U}} \mathcal{E}(\hat{Q}_{Y|X}, u) > Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)} \right) = 0, \quad (10)$$

provided that $(n \vee (\log n)m^{3\alpha+3}) \Delta^\gamma = O(1)$ as $n \to \infty$.

For monotonized estimators, the following corollary directly follows in view of (8) and Theorem 2.

**Corollary 1.** Let $\mathcal{U}$ be a bounded closed interval in $(0, 1)$. Suppose that all the assumptions of Theorem 2 are satisfied. Let $\hat{Q}^\dagger_{Y|X}(u \mid x)$ be any monotonized estimator for $Q_{Y|X}(u \mid x)$ given in Section 2.4. Then, we have

$$\lim_{M \to \infty} \lim_{n \to \infty} \sup_{F \in \mathcal{F}} \sup_{u \in \mathcal{U}} \mathbb{P} \left( \int_{\mathcal{U}} \mathcal{E}(\hat{Q}^\dagger_{Y|X}, u) du > Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)} \right) = 0.$$

Here, note that the rate $n^{-(\alpha+2\beta-1)/(\alpha+2\beta)}$ attained in estimating $Q_{Y|X}(u \mid x)$ is faster than the rate $n^{-(2\beta-1)/(\alpha+2\beta)}$ attained in estimating $b(t, u)$.

In what follows, we discuss optimality of these rates.

**Proposition 1.** Suppose that assumptions (A1)-(A6) and (A8) are satisfied. Let $\gamma$ be such that

$$0 < \gamma < \begin{cases} \alpha - 1, & \text{if } \alpha \leq 3, \\ 2, & \text{if } \alpha > 3. \end{cases} \quad (11)$$

Then, there exists a constant $M > 0$ such that for $\mathcal{F} = \mathcal{F}(C_1, \alpha, \beta, \gamma),$

$$\lim_{n \to \infty} \inf_{b \in \mathcal{F}} \sup_{F \in \mathcal{F}} \mathbb{P} \left( \sup_{u \in \mathcal{U}} \int_0^1 \{\hat{b}(t, u) - b(t, u)\}^2 dt > Mn^{-(2\beta-1)/(\alpha+2\beta)} \right) > 0,$$

where $\inf_{b}$ is taken over all estimators for the slope function $(t, u) \mapsto b(t, u)$ based on $(Y_1, X_1), \ldots, (Y_n, X_n)$. Similarly, there exists a constant $M > 0$ such that for $\mathcal{F} = \mathcal{F}(\gamma, \alpha, \beta, \gamma),$

$$\lim_{n \to \infty} \inf_{Q_{Y|X}} \sup_{F \in \mathcal{F}} \mathbb{P} \left( \sup_{u \in \mathcal{U}} \mathcal{E}(\hat{Q}_{Y|X}, u) > Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)} \right) > 0,$$

where $\inf_{Q_{Y|X}}$ is taken over all estimators for the conditional quantile function $Q_{Y|X}(u \mid x)$ based on $(Y_1, X_1), \ldots, (Y_n, X_n)$. In case of $\mathcal{U}$ being
a bounded closed interval in $(0, 1)$, there exists a constant $M > 0$ such that for $\mathcal{F} = \mathcal{F}(C_1, \alpha, \beta, \gamma)$,

$$
\liminf_{n \to \infty} \inf_{\bar{Q}_Y \in \mathcal{F}} \sup_{F \in \mathcal{F}} P_F \left[ \int_{\mathcal{U}} E(\bar{Q}_Y | X, u) du > M n^{- (\alpha + 2 \beta - 1) / (\alpha + 2 \beta)} \right] > 0,
$$

where the previous convention applies.

The side condition (11) is a compatibility condition between assumptions (A3) and (A8). It is not addressed here whether this condition is tight. However, some restriction between $\alpha$ and $\gamma$ is required in establishing lower bounds of rates of convergence to guarantee that the class $\mathcal{F}(C_1, \alpha, \beta, \gamma)$ is at least nonempty. Proposition 1 shows that under this side condition the rates established in Theorems 1, 2 and Corollary 1 are indeed optimal in the minimax sense. A proof of Proposition 1 is given in Appendix A.

4. Empirical choice of the cut-off level

In this section, we suggest three criteria to choose $m$, and investigate their performance by simulations. We use a heuristic reasoning to derive selection criteria. Suppose that $\mathcal{U}$ is a singleton: $\mathcal{U} = \{ u \}$. Suppose that there is no truncation bias, i.e., $b(t, u) = \sum_{j=1}^m b_j(u) \phi_j(t)$ and $Q_{Y|X}(u|X) = a(u) + \sum_{j=1}^m b_j(u) \xi_j$. Then, the infeasible estimator $(\hat{a}(u), \hat{b}_1(u), \ldots, \hat{b}_m(u))'$ defined by (3) can be regarded as a (conditional) maximum likelihood estimator when the conditional distribution of $Y$ given $X$ has the asymmetric Laplace density of the form:

$$
f(y|X, u, \sigma) = \frac{u(1-u)}{\sigma} \exp \left\{ -\frac{1}{\sigma} \rho(u - a(u) - \sum_{j=1}^m \hat{b}_j(u) \xi_j) \right\},
$$

where $\sigma > 0$ is a scale parameter. This suggests the following analogues of AIC and BIC in the present context:

$$
\text{AIC}(u) = \log \left[ \frac{1}{n} \sum_{i=1}^n \rho_u(Y_i - \hat{a}(u) - \sum_{j=1}^m \hat{b}_j(u) \xi_{ij}) \right] + \frac{(m + 1)}{n},
$$

$$
\text{BIC}(u) = \log \left[ \frac{1}{n} \sum_{i=1}^n \rho_u(Y_i - \hat{a}(u) - \sum_{j=1}^m \hat{b}_j(u) \xi_{ij}) \right] + \frac{(m + 1) \log n}{n}.
$$

See also Koenker [28, Section 4.9.1] for some related discussion. According to [38], we may also consider an analogue of GACV as follows:

$$
\text{GACV}(u) = \frac{\sum_{i=1}^n \rho_u(Y_i - \hat{a}(u) - \sum_{j=1}^m \hat{b}_j(u) \xi_{ij})}{n - (m + 1)}.
$$

In case of $\mathcal{U}$ being a bounded closed interval, define the integrated-AIC, BIC, and GACV as follows:

$$
\text{IAIC} = \int_{\mathcal{U}} \text{AIC}(u) du, \quad \text{IBIC} = \int_{\mathcal{U}} \text{BIC}(u) du, \quad \text{IGACV} = \int_{\mathcal{U}} \text{GACV}(u) du.
$$
In case of \( \mathcal{U} \) being a set of finite grid points, each integral is replaced by the summation over the grid points.

We carried out a small Monte Carlo study to investigate the finite sample performance of these criteria. In all cases, the number of Monte Carlo repetitions was 1,000. The numerical results obtained in this section were carried out by using the matrix language Ox [16]. The Ox code for solving quantile regression problems supplied on Professor Koenker’s website was used. See also [30] for some computational aspects of quantile regression problems.

The simulation design is described as follows:

\[
Y = \int_0^1 \varphi(t) X(t) dt + \varepsilon, \\
\varphi(t) = \sum_{j=1}^{50} \varphi_j \varphi_j(t), \quad \varphi_1 = 0.3, \quad \varphi_j = 4(-1)^{j+1} j^{-2}, j \geq 2, \quad \varphi_j(t) = 2^{1/2} \cos(j \pi t), \\
X(t) = \sum_{j=1}^{50} \gamma_j Z_j \varphi_j(t), \quad \gamma_j = (-1)^{j+1} j^{-n/2}, \quad \alpha \in \{1.1, 2\}, \quad Z_j \sim U[-3^{1/2}, 3^{1/2}], \\
\varepsilon \sim N(0,1) \text{ or Cauchy}, \quad n \in \{100, 200, 500\}.
\]

Each \( X_i \) is observed at 201 equally spaced grid points on \([0,1]\). In this design, we have

\[
Q_{Y|X}(u \mid X) = F_{\varepsilon}^{-1}(u) + \int_0^1 \varphi(t) X(t) dt,
\]

where \( F_{\varepsilon}^{-1}(\cdot) \) is the quantile function of \( \varepsilon \). Thus, \( a(u) = F_{\varepsilon}^{-1}(u) \) and \( b(t,u) \equiv \varphi(t) \) (\( b(t,u) \) is independent of \( u \)). We considered two cases for \( \mathcal{U} \): (a) \( \mathcal{U} = \{0.5\} \) and (b) \( \mathcal{U} = \{0.15, 0.2, \ldots, 0.85\} \). In each case, the performance was measured by

\[
\text{QA-MISE} = \frac{1}{\text{Card}(\mathcal{U})} \sum_{u \in \mathcal{U}} \mathbb{E} \left[ \int_0^1 \{ \hat{b}(t,u) - b(t,u) \}^2 dt \right] ; \text{ or},
\]

\[
\text{QA-MISE} = \frac{1}{\text{Card}(\mathcal{U})} \sum_{u \in \mathcal{U}} \mathbb{E} \left[ \int \{ \hat{Q}_{Y|X}(u \mid x) - Q_{Y|X}(u \mid x) \}^2 dP_X(x) \right],
\]

where \( P_X \) denotes the distribution of \( X \) and QA-MISE is the abbreviation of “quantile-averaged mean integrated squared error”.

The simulation results for case (a) are summarized in Figures 1-4. Figures 1 and 2 show the performance of the selection criteria for the normal error case, while Figures 3 and 4 show that for the Cauchy error case. In each figure, “Fixed” refers to the performance of the estimator with fixed \( m \). In the normal error case, BIC worked better than other two criteria. On the other hand, in the Cauchy error case, AIC and GACV worked better than BIC. Looking closely at these figures, one finds that AIC and GACV performed quite badly in some cases (see the bottom half in Figure 1). Although none of these criteria clearly dominated the others, BIC worked relatively stably. These figures also show that
as $\alpha$ increases from 1.1 to 2, the performance of $\hat{b}(\cdot, 0.5)$ becomes worse, while that of $Q_{Y|X}(0.5 \mid x)$ becomes better. This is consistent with the theoretical results in the previous section. Essentially similar comments apply to case (b), Figures 5-8.

5. Proof of Theorem 1

We divide the proof into three subsections. Some technical results are proved in Appendix B. To avoid the notational complication, uniformity in $F \in \mathcal{F}$ will be suppressed. Let $C > 0$ denote a generic constant of which the value may change from line to line. In most cases, qualification “almost surely” will be suppressed. In some parts of the proofs, we use empirical process techniques. We follow the basic notation used in [36].

5.1. Reduction of the problem

Let $b_0(u) = a(u)$ and $\xi_{i0} = \tilde{c}_{i0} = 1$. For any $v_0, \ldots, v_m$, write $v_m = (v_0, \ldots, v_m)'$. For $v_m \in \mathbb{R}^{m+1}$ and $w_m \in \mathbb{R}^{m+1}$, write $v_m \cdot w_m = \sum_{j=0}^{m} v_j w_j$. Then,

$$\hat{b}^m(u) = (\hat{b}_0(u), \hat{b}_1(u), \ldots, \hat{b}_m(u))' = \arg \min_{\hat{b}_m \in \mathbb{R}^{m+1}} \mathbb{E}_n [\rho_u(Y_i - \tilde{c}_m \cdot \hat{b}^m)].$$

We use a further re-parameterization. Let $\eta_{ij} = \kappa_j^{-1/2} \xi_{ij}, \hat{\eta}_{ij} = \kappa_j^{-1/2} \hat{\xi}_{ij}$, $d_j(u) = \kappa_j^{1/2} b_j(u)$ and $\hat{d}_j(u) = \kappa_j^{1/2} \hat{b}_j(u)$. Note that $\mathbb{E}[\eta_{ij}] = 0, \mathbb{E}[\eta_{ij}^2] = 1$, and $\mathbb{E}[\eta_{ij} \eta_{ik}] = 0$ for all $j \neq k$. Then,

$$\hat{d}^m(u) = (\hat{d}_0(u), \hat{d}_1(u), \ldots, \hat{d}_m(u))' = \arg \min_{\hat{d}_m \in \mathbb{R}^{m+1}} \mathbb{E}_n [\rho_u(Y_i - \tilde{c}_m \cdot \hat{d}^m)].$$

We first consider to bound $\sup_{u \in \mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2}$.

**Lemma 1.** Suppose that for all $\epsilon > 0$, there exist constants $c > 0$ and $M > 0$ possibly depending on $\epsilon$ such that

$$\lim_{n \to \infty} \inf \mathbb{P} \left\{ - \mathbb{E}_n \{ u - 1(Y_i \leq \tilde{c}_m \cdot (d^m(u) + M \sqrt{m/n} \hat{h}^m)) \} \left( \hat{h}^m \cdot \hat{c}_m \right) \right\} > c \sqrt{m/n}, \forall u \in \mathcal{U}, \forall \hat{h}^m \in \mathbb{S}^m \} > 1 - \epsilon.$$

Then, we have

$$\lim_{n \to \infty} \sup \mathbb{P} \left\{ \sup_{u \in \mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M \sqrt{m/n} \right\} \leq \epsilon.$$

**Proof.** The proof is divided into three steps.

Step 1: $\|\mathbb{E}_n \{ u - 1(Y_i \leq \tilde{c}_m \cdot \hat{d}^m(u)) \} \hat{h}^m \|_{\ell^2} \leq \left( (m+1)/n \right) \max_{1 \leq i \leq n} \|\hat{h}^m\|_{\ell^2}$. The proof is based on the next lemma.
Fig 1. Performance of selection criteria. Case (a). Estimation of $b(-, 0.5)$.

Fig 2. Performance of selection criteria. Case (a). Estimation of $Q_{Y|X}(0.5 \mid x)$.
Fig 3. Performance of selection criteria. Case (a). Estimation of $b(\cdot, 0.5)$.

Fig 4. Performance of selection criteria. Case (a). Estimation of $Q_{Y|X}(0.5 | x)$. 
Fig 5. Performance of selection criteria. Case (b). Estimation of \((t, u) \mapsto b(t, u)\).

Fig 6. Performance of selection criteria. Case (b). Estimation of \(Q_{Y|X}(u | x)\).
Fig 7. Performance of selection criteria. Case (b). Estimation of \((t, u) \rightarrow b(t, u)\).

Fig 8. Performance of selection criteria. Case (b). Estimation of \(Q_{Y|X}(u \mid x)\).
Lemma 2. Let \( \{ (y_i, z'_i) \}_{i=1}^n \) be a sequence of pairs of non-stochastic variables \((y_i, z'_i)\) where \(y_i \in \mathbb{R}\) and \(z_i \in \mathbb{R}^k\). Pick any \(u \in (0, 1)\). Let \( \hat{d}(u) \in \mathbb{R}^k \) be any solution to the minimization problem
\[
\min_{d \in \mathbb{R}^k} \mathbb{E}_n[\rho_u(y_i - z'_i d)].
\]
Then, we have
\[
\|\mathbb{E}_n[\{u - 1(y_i \leq z'_i \hat{d}(u))\}z_i]\|_2^2 
\leq n^{-1} \text{Card} \{\{i \in \{1, \ldots, n\} : y_i = z'_i \hat{d}(u)\} \cap \{1 \leq i \leq n\} \max \|z_i\|_2.
\]

Proof of Lemma 2. Follows from a small modification of El-Attar et al. [17, Lemma 2.1].

Recall that \( \hat{\eta}_i^m \) depends only on \( X'_1 := \{X_1, \ldots, X_n\} \) and not on \( Y_1, \ldots, Y_n \). Since the conditional distribution of \( Y_1, \ldots, Y_n \) given \( X'_1 \) is absolutely continuous, by Sard’s theorem, \( \sup_{u \in \mathcal{U}} \text{Card} \{(i \in \{1, \ldots, n\} : Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)) \} \leq m + 1 \) almost surely. To be more precise, pick any subset \( I \subset \{1, \ldots, n\} \) such that \( \text{Card}(I) \geq m + 2 \). Conditional on \( X'_1 \), consider the set
\[
S_I = \{(\hat{\eta}_i^m, \delta^m) : \delta^m \in \mathbb{R}^{m+1} \} \subset \mathbb{R}^{\text{Card}(I)}.
\]
Then, \( S_I \) is a linear subspace of dimension at most \( m + 1 \). Suppose that \( Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u) \) for all \( i \in I \) for some \( u \in \mathcal{U} \). Then, \( (Y_i)_{i \in I} \in S_I \), by which we have
\[
\mathbb{P}\{Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u), \forall i \in I, \exists u \in \mathcal{U} \mid X'_1 \} \leq \mathbb{P}\{(Y_i)_{i \in I} \in S_I \mid X'_1 \}. \tag{12}
\]
However, by Sard’s theorem [see 32], the Lebesgue measure of \( S_I \) in \( \mathbb{R}^{\text{Card}(I)} \) is zero, and by the absolute continuity of the conditional distribution of \( (Y_i)_{i \in I} \) given \( X'_1 \), the right side of (12) is zero. Thus, we conclude that
\[
\mathbb{P}\left\{ \sup_{u \in \mathcal{U}} \text{Card} \{(i \in \{1, \ldots, n\} : Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)) \} \geq m + 2 \mid X'_1 \right\} 
\leq \sum_{\text{Card}(I) \geq m + 2} \mathbb{P}\{Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u), \forall i \in I, \exists u \in \mathcal{U} \mid X'_1 \} = 0,
\]
by which we have \( \sup_{u \in \mathcal{U}} \text{Card} \{(i \in \{1, \ldots, n\} : Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)) \} \leq m + 1 \) almost surely. Then, the conclusion of Step 1 follows from an application of Lemma 2.

Step 2: We have
\[
\max_{1 \leq i \leq n} \|\hat{\eta}_i^m\|_2 = o_P\{\log n\}^{-1}\sqrt{n/m}. \tag{13}\]
We defer the proof of (13) to Appendix B.

Step 3: Proof of the lemma.

Define
\[
\hat{h}^m(u) = \begin{cases} 
\frac{\hat{d}^m(u) - d^m(u)}{\|\hat{d}^m(u) - d^m(u)\|_2}, & \text{if } \hat{d}^m(u) \neq d^m(u), \\
0, & \text{otherwise}.
\end{cases}
\]
Then, by Steps 1 and 2, we have
\[ \sup_{u \in \mathcal{U}} |\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot \hat{d}^m(u))\}(\hat{h}^m(u) \cdot \hat{\eta}_i^m)| = o_P(\sqrt{m/n}). \]
Define the event
\[ \mathcal{E}_n := \left\{ -\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + M \sqrt{m/nh^m}))\}(h^m \cdot \hat{\eta}_i^m) \right\} > c\sqrt{m/n}, \forall u \in \mathcal{U}, \forall h^m \in \mathcal{S}^m \}
Since the map
\[ l \mapsto -\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + l \sqrt{m/nh^m}))\}(h^m \cdot \hat{\eta}_i^m) \]
is non-decreasing for all \( u \in \mathcal{U} \) and \( h^m \in \mathcal{S}^m \), the event \( \mathcal{E}_n \) is also written as
\[ \mathcal{E}_n = \left\{ -\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + l \sqrt{m/nh^m}))\}(h^m \cdot \hat{\eta}_i^m) \right\} > c\sqrt{m/n}, \forall u \in \mathcal{U}, \forall l \geq M, \forall h^m \in \mathcal{S}^m \}
Thus, as \( n \to \infty \),
\[ \mathbb{P} \left\{ \sup_{u \in \mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M \sqrt{m/n} \right\} 
= \mathbb{P} \left\{ \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M \sqrt{m/n}, \exists u \in \mathcal{U} \right\} 
\leq \mathbb{P} \left\{ \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M \sqrt{m/n}, \exists u \in \mathcal{U} \right\} \cap \mathcal{E}_n + \mathbb{P}(\mathcal{E}_n^c) 
\leq \mathbb{P} \left\{ -\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot \hat{d}^m(u))\}(\hat{h}^m(u) \cdot \hat{\eta}_i^m) > c\sqrt{m/n}, \exists u \in \mathcal{U} \right\} + \mathbb{P}(\mathcal{E}_n^c) 
\leq o(1) + (1 + o(1))\epsilon. \]
This completes the proof of Lemma 1. \hfill \Box

5.2. Verification of the hypothesis of Lemma 1

Pick any \( h^m = (h_0, h_1, \ldots, h_m)' \in \mathcal{S}^m \). For a given \( M > 1 \), let \( \delta^m = (\delta_0, \delta_1, \ldots, \delta_m)' = M \sqrt{m/nh^m} \). Then,
\[ -\mathbb{E}_n\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m))\}(h^m \cdot \hat{\eta}_i^m) 
= -\mathbb{E}_n\{u - 1(Y_i \leq Q_{Y|X}(u | X_i))\}(h^m \cdot \hat{\eta}_i^m) 
+ \mathbb{E}_n\{F_{Y|X}(\hat{\eta}_i^m \cdot (d^m(u) + \delta^m) | X_i) - F_{Y|X}(Q_{Y|X}(u | X_i) | X_i)(h^m \cdot \hat{\eta}_i^m)\} 
+ n^{-1/2} \mathbb{E}_n\{1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) - 1(Y_i \leq Q_{Y|X}(u | X_i))(h^m \cdot \hat{\eta}_i^m)\} 
=: I + II + III. \]
where we have used the fact that $F_{Y|X}(Q_{Y|X}(u \mid X) \mid X) = u$ and

$$n^{-1/2}G_{n1X} \{ (1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) - 1(Y_i \leq Q_{Y|X}(u \mid X_i)) \} (\hat{h}_i^m \cdot \hat{\eta}_i^m)$$

$$:= \mathbb{E} \{ (1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) - 1(Y_i \leq Q_{Y|X}(u \mid X_i)) \\
- F_{Y|X}(\hat{\eta}_i^m \cdot (d^m(u) + \delta^m) \mid X_i) + F_{Y|X}(Q_{Y|X}(u \mid X_i) \mid X_i) \} (\hat{h}_i^m \cdot \hat{\eta}_i^m).$$

We separately bound the terms $I$, $II$ and $III$ uniformly in $u \in \mathcal{U}$ and $h^m \in \mathbb{S}^m$. In what follows, stochastic orders are interpreted independent of $M$. Note that

$$\hat{\eta}_i^m \cdot (d^m(u) + \delta^m) - Q_{Y|X}(u \mid X_i)$$

$$= \sum_{j=0}^m (d_j(u) + \delta_j)\hat{\eta}_{ij} - \sum_{j=0}^\infty d_j(u)\eta_{ij}$$

$$= \hat{\eta}_i^m \cdot \delta^m + (\hat{\eta}_i^m - \eta_i^m) \cdot d^m(u) + \sum_{j=m+1}^\infty d_j(u)\eta_{ij}$$

$$=: \hat{\eta}_i^m \cdot \delta^m + \hat{\delta}_i(u).$$

Bounding $I$: observe that

$$I \geq -\|\mathbb{E} \{(u - 1(Y_i \leq Q_{Y|X}(u \mid X_i))\}\hat{\eta}_i^m\|_{\ell^2}.$$ 

Using the relation

$$1(Y_i \leq Q_{Y|X}(u \mid X_i)) = 1(F_{Y|X}(Y_i | X_i) \leq u)$$

$$= 1(U_i \leq u), \text{ with } U_i = F_{Y|X}(Y_i | X_i),$$

we have

$$\sup_{u \in \mathcal{U}} \|\mathbb{E} \{(u - 1(U_i \leq u))\}\hat{\theta}_{ij}\|_{\ell^2}$$

$$\leq \sqrt{\sum_{j=0}^m \sup_{u \in \mathcal{U}} (\mathbb{E} \{(u - 1(U_i \leq u))\}\hat{\theta}_{ij})^2}.$$

Here, $U_1, \ldots, U_n$ are independent uniform random variables on $(0, 1)$ independent of $X_1^n := \{X_1, \ldots, X_n\}$. Pick any $0 \leq j \leq m$. Let $\sigma_1, \ldots, \sigma_m$ be independent Rademacher random variables independent of $(U_1, X_1), \ldots, (U_n, X_n)$. Since $U_1, \ldots, U_n$ are independent from $X_1^n$, applying the symmetrization inequality [see Lemma 2.3.1 of 36] conditional on $X_1^n$, we have

$$\mathbb{E} \left[ \sup_{u \in \mathcal{U}} (\mathbb{E} \{(u - 1(U_i \leq u))\}\hat{\theta}_{ij})^2 \mid X_1^n \right] \leq 4\mathbb{E} \left[ \sup_{u \in \mathcal{U}} (\mathbb{E} \{(u - 1(U_i \leq u))\}\hat{\theta}_{ij})^2 \mid X_1^n \right].$$

We make use of Proposition 3 in Appendix C to bound the right side. Consider the class of functions

$$\mathcal{G} = \{\mathbb{R} \times \mathbb{R} \ni (y, z) \mapsto 1(y \leq u)z : u \in \mathcal{U}\}.$$
Then, we have
\[
\mathbb{E} \left[ \sup_{u \in \mathcal{U}} (\mathbb{E}_n[\sigma_i(1(U_i \leq u)\hat{\eta}_{ij})]^2 \mid X_i^1) \right] = \mathbb{E} \left[ \sup_{g \in \mathcal{G}} (\mathbb{E}_n[\sigma_i g(U_i, \hat{\eta}_{ij})]^2 \mid X_i^1) \right].
\]

It is standard to see that \( \mathcal{G} \) is a VC subgraph class with VC index \( \leq 3 \). Thus, by Theorem 2.6.7 of [36], there exist universal constants \( A \geq c \) and \( W \geq 1 \) such that, for envelope function \( G(y, z) = |z| \),
\[
N(\epsilon | G\|_{L_2(P_n)}, \mathcal{G}, L_2(P_n)) \leq (A/\epsilon)^W, \quad 0 < \forall \epsilon \leq 1,
\]
where \( P_n \) denotes the empirical probability on \( \mathbb{R} \times \mathbb{R} \) that assigns probability \( n^{-1} \) to each \((U_i, \hat{\eta}_{ij}), i = 1, \ldots, n \). Therefore, by Proposition 3, we conclude that
\[
\mathbb{E} \left[ \sup_{u \in \mathcal{U}} (\mathbb{E}_n[\sigma_i 1(U_i \leq u)\hat{\eta}_{ij}]^2 \mid X_i^1) \right] \leq n^{-1}D\mathbb{E}_n[\hat{\eta}_{ij}^2],
\]
where \( D \) is a universal constant. Since \( 0 \leq j \leq m \) is arbitrary, we have
\[
\sup_{u \in \mathcal{U}} \mathbb{E}_n[\{u - 1(Y_i \leq Q_{Y|X}(u \mid X_i))\hat{\eta}_{i}^m\}]^{\|m\|_2} = O_P\{((\mathbb{E}_n[\hat{\eta}_{i}^m]\|m\|_2)^2)^{1/2}n^{-1/2}\}.
\]

We shall show in Appendix B that
\[
\mathbb{E}_n[\hat{\eta}_{i}^m]\|m\|_2 = O_P(m),
\]
by which we have
\[
\sup_{u \in \mathcal{U}} \mathbb{E}_n[\{u - 1(Y_i \leq Q_{Y|X}(u \mid X_i))\hat{\eta}_{i}^m\}]^{\|m\|_2} = O_P(\sqrt{m/n}).
\]

Bounding I: by Taylor's theorem, we have
\[
F_{Y|X}(Q_{Y|X}(u \mid X) + y \mid X) - F_{Y|X}(Q_{Y|X}(u \mid X) \mid X) = f_{Y|X}(Q_{Y|X}(u \mid X) \mid X)y + \frac{y^2}{2} \int_0^1 f_{Y|X}(Q_{Y|X}(u \mid X) + \theta y \mid X)(1 - \theta)d\theta
\]
\[
=: f_{Y|X}(Q_{Y|X}(u \mid X) \mid X)y + \frac{y^2}{2}R(u, y, X),
\]
by which we have, using (14),
\[
II = \mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u \mid X) \mid X_i)(\hat{\eta}_{i}^m \cdot \delta^m + \hat{r}_i(u))(h^m \cdot \hat{\eta}_{i}^m)]
\]
\[
+ \frac{1}{2}\mathbb{E}_n[(\hat{\eta}_{i}^m \cdot \delta^m + \hat{r}_i(u))^2R(u, \hat{\eta}_{i}^m \cdot \delta^m + \hat{r}_i(u), X_i)(h^m \cdot \hat{\eta}_{i}^m)]
\]
\[
\geq M\sqrt{m/n}\mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u \mid X) \mid X_i)(h^m \cdot \hat{\eta}_{i}^m)^2]
\]
\[
- C\mathbb{E}_n[(\hat{r}_i(u)(h^m \cdot \hat{\eta}_{i}^m))] - C\mathbb{E}_n[(\hat{\eta}_{i}^m \cdot \delta^m + \hat{r}_i(u))^2|h^m \cdot \hat{\eta}_{i}^m|]
\]
\[
\geq M\sqrt{m/n}\mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u \mid X) \mid X_i)(h^m \cdot \hat{\eta}_{i}^m)^2]
\]
\[
- C\mathbb{E}_n[\hat{r}_i^2(u)]^{1/2}\mathbb{E}_n[(h^m \cdot \hat{\eta}_{i}^m)^2]^{1/2}
\]
\[
- CM^2(m/n)(\max_{1 \leq i \leq n} \hat{\eta}_{i}^m \|m\|_2^2)\mathbb{E}_n[(h^m \cdot \hat{\eta}_{i}^m)^2]
\]
\[
- C(\max_{1 \leq i \leq n} \hat{\eta}_{i}^m \|m\|_2^2)\mathbb{E}_n[\hat{r}_i^2(u)],
\]
(16)
where we have used the fact that \( f_{Y|X}(y|X) \vee f_{Y'|X}(y'|X) \leq C \). By assumption (A5), there exists a small constant \( c_1 > 0 \) such that

\[
C \mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u \mid X_i)(h^m \cdot \hat{\eta}_i^m)^2) \geq c_1 \mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2], \forall u \in \mathcal{U}, \forall h^m \in \mathbb{S}^m.
\]

We shall show in Appendix B that

\[
\sup_{h^m \in \mathbb{S}^m} \mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2] - 1 = o_P(1), \tag{17}
\]

\[
\sup_{u \in \mathcal{U}} \mathbb{E}_n[\hat{c}_i^2(u)] = O_P(mn^{-1}). \tag{18}
\]

Thus, by (13), (15), (17) and (18), we have

\[
(16) \geq c_1 M \sqrt{m/n(1 - o_P(1))} - O_P(\sqrt{m/n}) - M^2 o_P(\sqrt{m/n}),
\]

where the stochastic orders are evaluated uniformly in \( u \in \mathcal{U} \) and \( h^m \in \mathbb{S}^m \).

Bounding III: Let \( \sigma_1, \ldots, \sigma_n \) be independent Rademacher random variables independent of the data \((Y_1, X_1), \ldots, (Y_n, X_n)\). Applying the symmetrization inequality conditional on \( X_1^n := \{X_1, \ldots, X_n\} \), we have

\[
\mathbb{E} \left[ \sup_{u \in \mathcal{U}, h^m \in \mathbb{S}^m} |n^{-1/2} \mathbb{G}_n|X| \{1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) - 1(Y_i \leq Q_{Y|X}(u \mid X_i))(h^m \cdot \hat{\eta}_i^m)\mid \mid X_1^n \} \right]
\]

\[
\leq 2\mathbb{E} \left[ \sup_{u \in \mathcal{U}, h^m \in \mathbb{S}^m} \mathbb{E}_n[\sigma_i(1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) - 1(Y_i \leq Q_{Y|X}(u \mid X_i))(h^m \cdot \hat{\eta}_i^m)\mid \mid X_1^n \right], \tag{19}
\]

where \( \delta^m \) is taken as \( \delta^m = M \sqrt{m/nh^m} \) in the suprema. Note that the symmetrization inequality is applicable since the regular conditional distribution of \((Y_1, \ldots, Y_n)\) given \( X_1^n \) exists and conditional on \( X_1^n \), \( Y_1, \ldots, Y_n \) are independent.

Consider the class of functions

\[
\mathcal{G} = \left\{ \mathbb{R} \times D[0,1] \times \mathbb{R}^{m+1} \ni (y, x, \eta^m) \mapsto \{1(y \leq \eta^m \cdot (d^m(u) + \delta^m)) - 1(y \leq Q_{Y|X}(u \mid x))(h^m \cdot \eta^m) : u \in \mathcal{U}, h^m \in \mathbb{S}^m, \delta^m = M \sqrt{m/nh^m} \right\}.\]

Then, we have

\[
(19) = 2\mathbb{E} \left[ \sup_{g \in \mathcal{G}} \mathbb{E}_n[\sigma_i g(Y_i, X_i, \hat{\eta}_i^m)\mid \mid X_1^n \right].
\]
We apply Proposition 2 in Appendix C to bound the right side. Note that $(X_i, \hat{\eta}_i^m)$ are measurable with respect to the $\sigma$-field generated by $X_i^n$, the regular conditional distribution of $(Y_1, \ldots, Y_n)'$ given $X_i^n$ exists, and conditional on $X_i^n$, $Y_1, \ldots, Y_n$ are independent. Observe that

$$\sup_{g \in \mathcal{G}} |g(Y_i, X_i, \hat{\eta}_i^m)| \leq \max_{1 \leq i \leq n} \|\hat{\eta}_i^m\|_{\ell^2} =: \hat{B},$$

and, by (14),

$$\sup_{g \in \mathcal{G}} \mathbb{E}_n \left[ \mathbb{E} \left[ g^2(Y_i, X_i, \hat{\eta}_i^m) \mid X_i^n \right] \right] = \sup_{u \in U, h^m \in S^m} \mathbb{E}_n \left[ |F_Y |X (\hat{\eta}_i^m \cdot (d^m(u) + \delta^m)) \mid X_i \right]
-F_Y |X (Q_Y |X (u \mid X_i) \mid X_i)((h^m \cdot \hat{\eta}_i^m)^2] \leq \sup_{u \in U, h^m \in S^m} \left\{ C \sqrt{m/n} \mathbb{E}_n \left[ \|h^m \cdot \hat{\eta}_i^m\| \right] + C \mathbb{E}_n \left[ |i(u)| (h^m \cdot \hat{\eta}_i^m)^2 \right] \right\} \leq CM \sqrt{m/n} \max_{1 \leq i \leq n} \|\hat{\eta}_i^m\|_{\ell^2} \sup_{h^m \in S^m} \mathbb{E}_n \left[ (h^m \cdot \hat{\eta}_i^m)^2 \right] + C \max_{1 \leq i \leq n} \|\hat{\eta}_i^m\|_{\ell^2} (\sup_{u \in U} \mathbb{E}_n \left[ |i(u)| \right])^{1/2} \left( \sup_{h^m \in S^m} \mathbb{E}_n \left[ (h^m \cdot \hat{\eta}_i^m)^2 \right] \right)^{1/2} =: \hat{\tau}^2.$$

We shall show in Appendix B that there exist some constants $c_2 \geq 1$ and $A' \geq 3\sqrt{\tau}$ such that

$$N(\hat{B}, \mathcal{G}, L_2(P_n')) \leq (A'/\epsilon)^{c_2m}, \ 0 < \forall \epsilon \leq 1, \quad (20)$$

where $P_n'$ denotes the empirical distribution on $\mathbb{R} \times D[0, 1] \times \mathbb{R}^{m+1}$ that assigns probability $n^{-1}$ to each $(Y_i, X_i, \hat{\eta}_i^m), \ i = 1, \ldots, n$. Therefore, by Proposition 2, we conclude that

$$\mathbb{E} \left[ \sup_{g \in \mathcal{G}} \left| \mathbb{E}_n \left[ \sigma_i g(Y_i, X_i, \hat{\eta}_i^m) \mid X_i^n \right] \right| \right] \leq 1(\hat{\tau} > 0)D' \left[ \sqrt{c_2m\hat{\tau}^2 \over n} \log {A' \hat{B} \over \hat{\tau}} + {c_2m \hat{B} \over n} \log {A' \hat{B} \over \hat{\tau}} \right], \quad (21)$$

provided that $\hat{\tau} \leq \hat{B}$, where $D'$ is a universal constant.

By (13), (15) and (17), and the fact that $M > 1$, we have

$$\hat{B} = o_P\{ (\log n)^{-1} \sqrt{n/m} \}, \ \hat{\tau}^2 = M o_P\{ (\log n)^{-1} \},$$

and there exists a small constant $c_3 > 0$ such that with probability approaching one

$$\hat{\tau}^2 \geq c_3 \hat{B} \sqrt{m/n}.$$
Thus, replacing $\hat{\tau}$ by $\hat{\tau} \wedge \hat{B}$ if necessary, (21) = $M^{1/2}o_P(\sqrt{m/n})$, by which we conclude that

$$III \geq -M^{1/2}o_P(\sqrt{m/n}),$$

where the stochastic order is evaluated uniformly in $u \in U$ and $h^m \in S^m$.

Taking these together, we now conclude that

$$I + II + III \geq c_1 M \sqrt{m/n} (1 - o_P(1)) - o_P(\sqrt{m/n}) - M^2 o_P(\sqrt{m/n}),$$

where the stochastic orders are evaluated uniformly in $u \in U$ and $h^m \in S^m$.

This immediately implies the hypothesis of Lemma 1.

5.3. Completion of the proof

We have shown that

$$\sup_{u \in U} \|d^m(u) - d^m(u)\|_{L^2} = O_P(\sqrt{m/n}),$$

which immediately implies that

$$\sup_{u \in U} \|\hat{b}^m(u) - b^m(u)\|_{L^2}^2 = O_P(\kappa^{-1} m n^{-1}) = O_P(m^{\alpha+1} n^{-1})$$

$$= O_P(n^{-(2\beta-1)/(\alpha+2\beta)}).$$

Observe that

$$\hat{b}(t,u) - b(t,u)$$

$$= \sum_{j=1}^{m} \hat{b}_j(u) \hat{\phi}_j(t) - \sum_{j=1}^{m} b_j(u) \hat{\phi}_j(t) + \sum_{j=1}^{m} b_j(u) \phi_j(t)$$

$$- \sum_{j=1}^{m} b_j(u) \phi_j(t) + \sum_{j=m+1}^{\infty} b_j(u) \phi_j(t)$$

$$= \sum_{j=1}^{m} (\hat{b}_j - b_j)(u) \hat{\phi}_j(t) + \sum_{j=1}^{m} b_j(u) (\hat{\phi}_j(t) - \phi_j(t)) + \sum_{j=m+1}^{\infty} b_j(u) \phi_j(t),$$

so that, uniformly in $u \in U$,

$$\int_0^1 (\hat{b}(t,u) - b(t,u))^2 dt$$

$$\leq 3\|\hat{b}^m(u) - b^m(u)\|_{L^2}^2 + \sum_{j=1}^{m} b_j^2(u) \|\hat{\phi}_j - \phi_j\|^2 + 3 \sum_{j=m+1}^{\infty} b_j^2(u)$$

$$= O_P(n^{-(2\beta-1)/(\alpha+2\beta)}) + 3 \sum_{j=1}^{m} b_j^2(u) \|\hat{\phi}_j - \phi_j\|^2 + O(m^{-2\beta+1})$$

$$= O_P(n^{-(2\beta-1)/(\alpha+2\beta)}) + 3 \sum_{j=1}^{m} b_j^2(u) \|\hat{\phi}_j - \phi_j\|^2.$$
By the proof of (18) in Appendix B, we see that
\[
m \sum_{j=1}^{m} b_j^2(u) \| \hat{\varphi}_j - \varphi_j \|^2 \leq C m \sum_{j=1}^{m} j^{-2\beta} \| \hat{\varphi}_j - \varphi_j \|^2
\]
\[
= O_p(m(n^{-1} + \Delta^\gamma + n^{-1}(\log n)m^{\alpha+3}\Delta^\gamma)) = O(mn^{-1})
\]
\[
= o_p((2\beta - 1)/(\alpha + 2\beta)).
\]
This completes the proof of Theorem 1. \hfill \Box

6. Proof of Theorem 2

Without loss of generality, we may assume that \( \mathbb{E}[X(t)] = 0 \) for all \( t \in [0,1] \). Let \( X_{n+1} \) be a copy of \( X \) independent of the data \( \mathcal{D}_n := \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \). Then,
\[
\mathcal{E}(\hat{Q}_Y|X, u) = \mathbb{E}[\{\hat{Q}_Y|X(u | X_{n+1}) - Q_Y|X(u | X_{n+1})\}^2 | \mathcal{D}_n].
\]
Let \( X_{n+1} = \sum_{j=1}^{\infty} \xi_{n+1,j} \varphi_j \). Observe that
\[
\hat{Q}_Y|X(u | X_{n+1}) = \hat{a}(u) + \sum_{j=1}^{m} \hat{b}_j(u) \xi_{n+1,j} + \int_{0}^{1} X_{n+1}(t) \sum_{j=1}^{m} \hat{b}_j(u)(\hat{\varphi}_j - \varphi_j)(t) dt
\]
\[
- \int_{0}^{1} \tilde{X}(t) \tilde{b}(t, u) dt, \quad \text{and}
\]
\[
Q_Y|X(u | X_{n+1}) = a(u) + \sum_{j=1}^{m} b_j(u) \xi_{n+1,j} + \sum_{j=m+1}^{\infty} b_j(u) \xi_{n+1,j}.
\]
Letting \( \eta_{n+1,j} = \kappa_j^{-1/2} \xi_{n+1,j} \), we have
\[
\{\hat{Q}_Y|X(u | X_{n+1}) - Q_Y|X(u | X_{n+1})\}^2
\]
\[
\leq C \left[ (\hat{a} - a)^2(u) + \left\{ \sum_{j=1}^{m} (\hat{d}_j - d_j)(u) \eta_{m+1,j} \right\}^2 + \left\{ \sum_{j=m+1}^{\infty} b_j(u) \xi_{n+1,j} \right\}^2 \right]
\]
\[
+ \int_{0}^{1} X_{n+1}^2(t) dt \int_{0}^{1} \left\{ \sum_{j=1}^{m} \hat{b}_j(u)(\hat{\varphi}_j - \varphi_j)(t) \right\}^2 dt + \left\{ \int_{0}^{1} \tilde{X}(t) \tilde{b}(t, u) dt \right\}^2.
\]
Taking expectation with respect to \( X_{n+1} \), we have
\[
\mathbb{E}[\{\hat{Q}_Y|X(u | X_{n+1}) - Q_Y|X(u | X_{n+1})\}^2 | \mathcal{D}_n] \leq C \left[ \| \hat{d}^m(u) - d^m(u) \|_2^2
\]
\[
+ \sum_{j=m+1}^{\infty} \kappa_j b_j^2(u) + \int_{0}^{1} \left\{ \sum_{j=1}^{m} \hat{b}_j(u)(\hat{\varphi}_j - \varphi_j)(t) \right\}^2 dt + \left\{ \int_{0}^{1} \tilde{X}(t) \tilde{b}(t, u) dt \right\}^2 \right].
\]
By the previous proof, we have
\[ \sup_{u \in U} \| \hat{d}^m(u) - d^m(u) \|_{L^2}^2 = O_P(m/n) = O_P(n^{-(\alpha + 2\beta - 1)/(\alpha + 2\beta)}), \]
and
\[ \sum_{j=m+1}^{m} \kappa_j b_j^2(u) \leq C \sum_{j=m+1}^{\infty} j^{-\alpha - 2\beta} = O(m^{-\alpha - 2\beta + 1}) = O(n^{-(\alpha + 2\beta - 1)/(\alpha + 2\beta)}). \]

Observe that
\[ \left\{ \sum_{j=1}^{m} \hat{b}_j(u) (\hat{\phi}_j - \phi_j) \right\}^2 \]
\[ \leq 2 \left\{ \sum_{j=1}^{m} b_j(u) (\hat{\phi}_j - \phi_j) \right\}^2 + 2 \left\{ \sum_{j=1}^{m} (\hat{b}_j - b_j) (u) (\hat{\phi}_j - \phi_j) \right\}^2 \]
\[ = 2 \left\{ \sum_{j=1}^{m} b_j(u) (\hat{\phi}_j - \phi_j) \right\}^2 + 2 \left\{ \sum_{j=1}^{m} \kappa_j^{-1/2} (\hat{d}_j - d_j) (u) (\hat{\phi}_j - \phi_j) \right\}^2 \]
\[ \leq 2m \sum_{j=1}^{m} b_j^2(u) (\hat{\phi}_j - \phi_j)^2 + 2\| \hat{d}^m(u) - d^m(u) \|_{L^2}^2 \sum_{j=1}^{m} \kappa_j^{-1} (\hat{\phi}_j - \phi_j)^2, \]
by which we have
\[ \int_0^1 \left\{ \sum_{j=1}^{m} \hat{b}_j(u) (\hat{\phi}_j - \phi_j)(t) \right\}^2 dt \]
\[ \leq 2m \sum_{j=1}^{m} b_j^2(u) (\hat{\phi}_j - \phi_j)^2 + 2\| \hat{d}^m(u) - d^m(u) \|_{L^2}^2 \sum_{j=1}^{m} \kappa_j^{-1} (\hat{\phi}_j - \phi_j)^2. \]

By the previous proof, we see that
\[ m \sum_{j=1}^{m} b_j^2(u) (\hat{\phi}_j - \phi_j)^2 \leq C m \sum_{j=1}^{m} j^{-2\beta} (\hat{\phi}_j - \phi_j)^2 = O_P(mn^{-1}) \]
\[ = O_P(n^{-(\alpha + 2\beta - 1)/(\alpha + 2\beta)}), \]
while by the proof of (15), we have \( \sum_{j=1}^{m} \kappa_j^{-1} (\hat{\phi}_j - \phi_j)^2 = o_P(1) \). Thus, we conclude that
\[ \sup_{u \in U} \int_0^1 \left\{ \sum_{j=1}^{m} \hat{b}_j(u) (\hat{\phi}_j - \phi_j)(t) \right\}^2 dt = O_P(n^{-(\alpha + 2\beta - 1)/(\alpha + 2\beta)}). \]

Finally, we have
\[ \left\{ \int_0^1 \tilde{X}(t) \tilde{b}(t,u) dt \right\}^2 \leq \int_0^1 \tilde{X}^2(t) dt \int_0^1 \tilde{b}^2(t,u) dt = O_P(n^{-1}) \times O_P(1) = O_P(n^{-1}), \]
uniformly in $u \in U$. Taking these together, we conclude that
\[
\sup_{u \in U} E \left[ \left( \hat{Q}_{Y|X}(u | X_{n+1}) - Q_{Y|X}(u | X_{n+1}) \right)^2 \right] = O_P \left( n^{-(\alpha+2\beta-1)/(\alpha+2\beta)} \right).
\]
This completes the proof.

Acknowledgments

Most of this work was done while the author was visiting Department of Economics, MIT. He would like to thank Victor Chernozhukov and Hidehiko Ichimura for their suggestions and encouragements.

References

[1] Barlow, R., Bartholomew, J., Bremer, J., and Brunk, H. *Statistical Inference Under Order Restrictions*. Jonh Wiley & Sons.
[2] Belloni, A., Chernozhukov, V. and Fernández-Val, I. (2011). Conditional quantile processes based on series or many regressors. arXiv:1105.6154.
[3] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley.
[4] Bissantz, N., Hohage, T. and Monk, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems* 20 1773-1789.
[5] Bosq, D. (2000). *Linear Processes in Function Spaces*. Lecture Notes in Statistics, Springer.
[6] Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* 34 2159-2179.
[7] Cardot, H., Crambes, C. and Sarda, P. (2005). Quantile regression when the covariates are functions. *J. Nonparametric Statist.* 17 841-856.
[8] Cardot, H., Crambes, C. and Sarda, P. (2007). Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In: *Statistical Methods for Biostatistics and Related Fields* (eds. W. Hardle, Y. Mori and P. Vieu). pp.221-244. Springer.
[9] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.* 45 11-22.
[10] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear models. *Statist. Sinica* 13 571-591.
[11] Chen, K. and Müller, H.-G. (2011). Conditional quantile analysis when covariates are functions, with application to growth data. *J. Roy. Stat. Soc. Ser. B*, to appear.
[12] Chen, X. and Pouzo, D. (2011). Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica*, to appear.
[13] Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2009). Improving point and interval estimators of monotone rearrangement. *Biometrika* 96 559-575.
(14) Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* 37 35-72.

(15) Delaigle, A. and Hall, P. (2011). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* to appear.

(16) Doornik, J.A. (2002). *Object-Oriented Matrix Programming Using Ox*. 3rd edition. Timberlake Consultants Press.

(17) El-Attar, R.A., Vidyasagar, M. and Dutta, S.P.K. (1979). An algorithm for \( l_1 \)-norm minimization with application to nonlinear \( l_1 \)-approximation. *SIAM J. Numer. anal.* 16 70-86.

(18) Ferraty, F., Rabhi, A. and Vieu, P. (2005). Conditional quantiles for dependent functional data with application to the climatic EL Niño phenomenon. *Sankhya* 67 378-398.

(19) Gagliardini, P. and Scaillet, O. (2011). Nonparametric instrumental variables estimation of structural quantile effects. *Econometrica*, to appear.

(20) Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* 37 503-522.

(21) Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* 34 1143-1216.

(22) Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* 35 70-91.

(23) He, X. and Shao, Q.-M. (2000). On parameters on increasing dimensions. *J. Multivariate Anal.* 73 125-135.

(24) Hofmann, B. and Scherzer, O. (1998). Local ill-posedness and source conditions of operator equations in Hilbert space. *Inverse Problems* 14 1189-1206.

(25) Hormann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *Ann. Statist.* 38 1845-1884.

(26) Horowitz, J.L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* 74 1191-1208.

(27) James, G.M., Wang, J. and Zhu, J. (2009). Functional linear regression that’s interpretable. *Ann. Statist.* 37 2083-2108.

(28) Koenker, R. (2005). *Quantile Regression*. Oxford University Press.

(29) Koenker, R. and Bassett G.W. (1978). Regression quantiles. *Econometrica* 46 33-50.

(30) Koenker, R. and Portnoy, S. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statist. Science* 12 279-300.

(31) Loubes, J.-M. and Carenne, L. (2010). Penalized estimators for non linear inverse problems. *ESAIM: Probability and Statistics* 14 173-191.

(32) Milnor, J.W. (1965). *Topology from the Differentiable Viewpoint*. Princeton Landmarks in Mathematics. Princeton University Press.

(33) Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* 33 774-805.

(34) Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. Springer.
[35] Rudelson, M. (1999). Random vectors in the isotropic position. *J. Fun-
c tional Anal.* **164** 60-72.

[36] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and
Empirical Processes: With Applications to Statistics*. Springer.

[37] Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression
analysis for longitudinal data. *Ann. Statist.* **33** 2873-2903.

[38] Yuan, M. (2006). GACV for quantile smoothing splines. *Computational
Statistics and Data Analysis* **50** 813-829.

[39] Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach
to functional linear regression. *Ann. Statist.* **38** 3412-3444.
Appendix A: Proof of Proposition 1

Consider the same construction as in [22]. Let \( \phi_1(t) \equiv 1 \) and \( \phi_j(t) = 2^{1/2} \cos(j \pi t) \) for \( j \geq 1 \). Put \( \varphi_j = \theta_j j^{-\beta} \) for \( n^{1/(\alpha+2\beta)} + 1 \leq j \leq 2 n^{1/(\alpha+2\beta)} \) and \( \varphi_j = 0 \) otherwise where \([y]\) denotes the integer part of \( y \in \mathbb{R} \) and each \( \theta_j \) is either 0 or 1. Let \( Z_1, Z_2, \ldots \sim U[-3^{1/2}, 3^{1/2}] \) i.i.d. Take \( X(t) = \sum_{j=1}^{\infty} j^{-\alpha/2} Z_j \phi_j(t) \) and \( \varrho(t) = \sum_{j=[n^{1/(\alpha+2\beta)}]+1}^{2 n^{1/(\alpha+2\beta)}} \theta_j \phi_j(t) \). Consider a sequence of data generating processes

\[
Y = \int_0^1 \varrho(t) X(t) \, dt + \epsilon = \sum_{j=[n^{1/(\alpha+2\beta)}]+1}^{2 n^{1/(\alpha+2\beta)}} \theta_j j^{-(\alpha+2\beta)/2} Z_j + \epsilon, \quad \epsilon \sim N(0, 1), \ \epsilon \perp X.
\]

Then, we have

\[
Q_{Y|X}(u | X) = a(u) + \int b(t, u) X(t) \, dt,
\]

with

\[
a(u) = \Phi^{-1}(u), \quad b(t, u) \equiv \varrho(t), \quad f_{Y|X}(y | X) = \phi(y - \int_0^1 \varrho(t) X(t) \, dt),
\]

by which one sees that assumptions (A4)-(A6) are satisfied. Here, \( \phi(\cdot) \) and \( \Phi(\cdot) \) are the density and the distribution function of the standard normal distribution, respectively. Suppose that \( \alpha \leq 3 \). Then, since for any \( 0 < \gamma < \alpha - 1 \), \( t \mapsto \cos(t) \) is \( \gamma/2 \)-Hölder continuous (by the periodicity of the cosine function), we have

\[
\mathbb{E}[(X(s) - X(t))^2] \leq C|s - t|^\gamma \sum_{j=1}^{\infty} j^{-\alpha + \gamma} \leq C'|s - t|^\gamma, \quad \forall s, t \in [0, 1],
\]

where \( C \) and \( C' \) are some constants. This shows that Assumption (A8) is satisfied with \( 0 < \gamma < \alpha - 1 \) when \( \alpha \leq 3 \). For \( \alpha > 3 \), \( K(s, t) \) is twice continuously differentiable, so that Assumption (A8) is satisfied with \( 0 < \gamma \leq 2 \). Finally, by [22], for any estimator \( (t, u) \mapsto \hat{b}(t, u) \),

\[
\sup_u \sup_{t \in U} \int_0^1 \mathbb{E}[(\hat{b}(t, u) - b(t, u))^2] \, dt \\
\geq \sup_u \int_0^1 \mathbb{E}[(\hat{b}(t, u_0) - \varrho(t))^2] \, dt \quad (u_0 \text{ is any point in } U) \\
\geq D n^{-2(\beta - 1)/(\alpha + 2\beta)},
\]

where \( \sup_u \) denotes the supremum over all \( 2 n^{1/(\alpha + 2\beta)} \) different distributions of \((Y, X)\) obtained by taking different choices of \( \theta[n^{1/(\alpha + 2\beta)}] + 1, \ldots, \theta[2 n^{1/(\alpha + 2\beta)}] \), and \( D > 0 \) is a constant. The other assertions follow similarly. This completes the proof. \( \square \)

Appendix B: Proofs of (13), (15), (17), (18) and (20)

In this section, we provide proofs of (13), (15), (17), (18) and (20) omitted in Section 5. Throughout the section, we assume all the conditions of Theorem 1.
Observe that
\[ n \text{Furthermore, as Lemma 3. We have }\]
\[ \left( X_i(t) - X_i(s) \right) = \sum_{l=1}^{L_i} \left( X_i(t_{il}) - X_i(t) \right) 1(t \in [t_{il}, t_{il+1}]), \ t \in [0, 1), \]
\[ \text{Recall that } \]
\[ \text{Without loss of generality, we may assume that } \]
\[ \text{Define the (infeasible) empirical covariance kernel }\]
\[ \hat{K}^*(s, t) = \mathbb{E}_n[(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))], \]
\[ \text{where } \bar{X}(t) = n^{-1} \sum_{i=1}^n X_i(t). \text{ Let } \hat{K}^*(s, t) = \sum_{j=1}^{\infty} \hat{\kappa}_j^* \hat{\phi}_j^*(s)\hat{\phi}_j^*(t) \text{ be the spectral expansion of } \hat{K}^*(s, t) \text{ where } \hat{\kappa}_1^* \geq \hat{\kappa}_2^* \geq \cdots \geq 0 \text{ and } \{\hat{\phi}_j^*\}_{j=1}^{\infty} \text{ is an orthonormal basis for } L_2[0, 1]. \text{ Without loss of generality, we may assume that } \]
\[ \int \hat{\phi}_j^* \hat{\phi}_j^* \geq 0, \int \hat{\phi}_j^* \hat{\phi}_j^* \geq 0, \forall j \geq 1. \]

Here, to ease the notation, \( \int_0^1 f(t)dt \) is abbreviated as \( \int f \) for any function \( f : [0, 1] \to \mathbb{R} \). Define
\[ \tilde{\xi}_{ij}^* = \int (X_i - \bar{X}) \hat{\phi}_j^*, \quad \tilde{\eta}_{ij}^* = \kappa_j^{-1/2} \tilde{\xi}_{ij}^*. \]
Recall that \( \eta_{ij} = \kappa_j^{-1/2} \xi_{ij} = \kappa_j^{-1/2} \int X_i \phi_j \) and \( \tilde{\eta}_{ij} = \kappa_j^{-1/2} \tilde{\xi}_{ij} = \kappa_j^{-1/2} \int (\hat{X}_i - \bar{X}) \hat{\phi}_j \). We will frequently use the following decomposition: for \( j \geq 1, \)
\[ \hat{\eta}_{ij} - \eta_{ij} = \hat{\eta}_{ij} - \tilde{\eta}_{ij} + \tilde{\eta}_{ij} - \eta_{ij}, \]
\[ \hat{\eta}_{ij} - \tilde{\eta}_{ij} = \kappa_j^{-1/2} \int (\hat{X}_i - X_i) \hat{\phi}_j + \kappa_j^{-1/2} \int X_i (\hat{\phi}_j - \phi_j), \]
\[ - \kappa_j^{-1/2} \int (\bar{X} - \hat{X}) \hat{\phi}_j - \kappa_j^{-1/2} \int \hat{X} (\hat{\phi}_j - \phi_j) \]
\[ =: \Delta_{ij1} + \Delta_{ij2} - \Delta_{ij3} - \Delta_{ij4}, \]
\[ \tilde{\eta}_{ij} - \eta_{ij} = \kappa_j^{-1/2} \int X_i (\phi_j^* - \phi_j) - \kappa_j^{-1/2} \int \bar{X} \phi_j - \kappa_j^{-1/2} \int \hat{X} (\hat{\phi}_j - \phi_j), \]
\[ =: \Delta_{ij5} - \Delta_{ij6} - \Delta_{ij7}. \]

We prepare some lemmas. For any function \( R : [0, 1]^2 \to \mathbb{R} \), define \( |||R||| = (\int \int R^2(s, t)dsdt)^{1/2} \). Recall that \( \|f\|^2 = \int_0^1 f^2(t)dt \) for any \( f : [0, 1] \to \mathbb{R} \).

**Lemma 3.** We have
\[ \mathbb{E}_n[|||\hat{X}_i - X_i|||^2] = O_P(\Delta^\gamma), \quad |||\hat{K} - \bar{K}^*|||^2 = O_P(\Delta^\gamma). \]

Furthermore, as \( n \to \infty \), with probability approaching one,
\[ ||\hat{\phi}_j - \phi_j^*|| \leq Cj^{\alpha+1}|||\hat{K} - \bar{K}^*|||, \ 1 \leq \forall j \leq m. \]

**Proof.** Observe that
\[ \hat{X}_i(t) - X_i(t) = \sum_{i=1}^{L_i} (X_i(t_{il}) - X_i(t)) 1(t \in [t_{il}, t_{il+1}]), \ t \in [0, 1), \]
by which we have
\[ \| \hat{X}_i - X_i \|^2 = \sum_{l=1}^{L_i} \int_{t_{il}}^{t_{i,l+1}} (X_i(t_{il}) - X_i(t))^2 dt. \]

Taking expectation, we have
\[ \mathbb{E}[\| \hat{X}_i - X_i \|^2] = \sum_{l=1}^{L_i} \int_{t_{il}}^{t_{i,l+1}} \left\{ K(t, t) - 2K(t, t_{il}) + K(t_{il}, t) \right\} dt \]
\[ \leq C \sum_{l=1}^{L_i} (t_{i,l+1} - t_{il})^{\gamma + 1} \leq C \Delta \gamma, \]

where we have used assumption (A8). This leads to the first assertion. The second assertion follows from the Schwarz inequality and the first assertion. The third assertion needs some effort. By Bosq [5, Lemmas 4.2 and 4.3; see also the remark below], we have
\[ \sup_{j \geq 1} |\hat{k}_j - \kappa_j| \leq \| \hat{K}^* - K \|, \quad \sup_{j \geq 1} \hat{\chi}_j \| \hat{\phi}_j - \tilde{\phi}_j^* \| \leq 8^{1/2} \| \hat{K} - \hat{K}^* \|, \] (22)

where \( \hat{\chi}_j = \min\{\hat{k}_j - \hat{k}_j, \hat{k}_j - \hat{k}_{j+1}\} \) for \( j \geq 2 \) and \( \hat{\chi}_1 = \hat{k}_1 - \hat{k}_2 \). For some small constant \( c > 0 \), define the event
\[ \mathcal{E}_n = \{ \hat{\chi}_j \geq cj^{-\alpha - 1}, 1 \leq \forall j \leq m \}. \]

It suffices to show that \( \mathbb{P}(\mathcal{E}_n) \to 1 \). By the first inequality in (22),
\[ \hat{k}_k - \hat{k}_{k+1} \geq \kappa_k - \kappa_{k+1} - 2 \| \hat{K}^* - K \| \geq C^{-1} k^{-\alpha - 1} - 2 \| \hat{K}^* - K \|. \]

Since \( k^{-\alpha - 1} \geq m^{-\alpha - 1} \propto n^{-(\alpha + 1)/(\alpha + 2)\beta} \), \( \| \hat{K}^* - K \| = O_P(n^{-1/2}) \) (which follows by a simple calculation), and \( n^{-1/2} = o(n^{-(\alpha + 1)/(\alpha + 2)\beta}) \) (which follows by \( \beta > \alpha/2 + 1 \)), we have uniformly in \( 1 \leq k \leq m \),
\[ \hat{k}_k - \hat{k}_{k+1} \geq C^{-1} (1 - o_P(1)) k^{-\alpha - 1}, \]

which leads to that \( \mathbb{P}(\mathcal{E}_n) \to 1 \) by taking \( c \) sufficiently small. \( \square \)

**Remark 1.** Lemma 4.3 of [5] reads as follows: for functions \( Q, R : [0, 1]^2 \to \mathbb{R} \) having the spectral expansions in \( L_2[0,1]^2 \) of the form \( Q(s, t) = \sum_{j=1}^{\infty} \lambda_j \psi_j(s) \phi_j(t) \) and \( R(s, t) = \sum_{j=1}^{\infty} \nu_j \phi_j(s) \psi_j(t) \), where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \nu_1 \geq \nu_2 \geq \cdots \geq 0 \), and \( \{ \psi_j \}_{j=1}^{\infty} \) and \( \{ \phi_j \}_{j=1}^{\infty} \) are orthonormal bases for \( L_2[0,1] \), we have:
\[ \chi_j \| \phi_j - \psi_j \| \leq 8^{1/2} \| R - Q \| \]
for all \( j \geq 1 \) such that \( \chi_j > 0 \), where \( \chi_j = \min\{\lambda_j - \lambda_j, \lambda_j - \lambda_{j+1}\} \) for \( j \geq 2 \) and \( \chi_1 = \lambda_1 - \lambda_2 \). Here, we have assumed that \( \int \phi_j \psi_j \geq 0 \) for all \( j \geq 1 \). This lemma actually holds with \( \sup_{j \geq 1} \chi_j \| \phi_j - \psi_j \| \leq 8^{1/2} \| R - Q \| \) since the inequality trivially holds in case of \( \chi_j = 0 \).

The following useful result was established in [22].
Lemma 4. As \( n \to \infty \), with probability approaching one,
\[
\| \hat{\phi}_j^* - \phi_j \|^2 \leq 10 \sum_{k: k \neq j} (\kappa_j - \kappa_k)^{-2} \left\{ \int (\hat{K}^* - K)(s, t)\phi_j(s)\phi_k(t)dsdt \right\}^2, 1 \leq j \leq m,
\]
Furthermore, we have
\[
\sum_{k: k \neq j} (\kappa_j - \kappa_k)^{-2} \mathbb{E} \left\{ \left\{ \int (\hat{K}^* - K)(s, t)\phi_j(s)\phi_k(t)dsdt \right\}^2 \right\} = O(j^2 n^{-1}),
\]
uniformly in \( 1 \leq j \leq m \).

Proof. See Hall and Horowitz [22, p.83-84].

B.1. Proofs of (13) and (15)

We first prove (15). By Lemmas 3 and 4, we have
\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{\Delta}_{ij1}^2 \leq \mathbb{E}_n[\|X_i\|^2] \left\{ \sum_{j=1}^{m} \kappa_j^{-1} \right\} = O_P(m^{\alpha + 1} \Delta^\gamma) = o_P(1),
\]
\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{\Delta}_{ij2}^2 \leq \mathbb{E}_n[\|X_i\|^2] \left\{ \sum_{j=1}^{m} \kappa_j^{-1} \|\hat{\phi}_j^* - \hat{\phi}_j\|^2 \right\}
= O_P(1) \times O_P(\Delta^\gamma \sum_{j=1}^{m} j^{3\alpha + 2}) = O_P(m^{3\alpha + 3} \Delta^\gamma) = o_P(1),
\]
\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{\Delta}_{ij3}^2 \leq \mathbb{E}_n[\|X_i\|^2] \left\{ \sum_{j=1}^{m} \kappa_j^{-1} \|\hat{\phi}_j^* - \phi_j\|^2 \right\}
= O_P(1) \times O_P(n^{-1} \sum_{j=1}^{m} j^{\alpha + 2}) = O_P(m^{\alpha + 3} n^{-1}) = o_P(1).
\]
Similarly, we have
\[
\sum_{j=1}^{m} \hat{\Delta}_{j6}^2 = O_P(n^{-1} m^{\alpha + 1} \Delta^\gamma) = o_P(1),
\]
\[
\sum_{j=1}^{m} \hat{\Delta}_{j4}^2 = O_P(n^{-1} m^{3\alpha + 3} \Delta^\gamma) = o_P(1),
\]
\[
\sum_{j=1}^{m} \hat{\Delta}_{j7}^2 = O_P(m^{\alpha + 3} n^{-2}) = o_P(1).
\]
Using the decomposition \( X_i(t) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t) \), we have
\[
\hat{\Delta}_{j6} = \kappa_j^{-1/2} \int \hat{X} \phi_j = \kappa_j^{-1/2} \hat{\xi}_j = \hat{\eta}_j,
\]
by which we have
\[ \sum_{j=0}^{m} \hat{\Delta}_{j6}^2 = O_P(\sum_{j=1}^{m} \hat{\Delta}_{j6}^2) = O_P(\sum_{j=1}^{m} \mathbb{E}[\hat{\eta}_{j}^2]) = O_P(mn^{-1}) = o_P(1). \]

Finally, by a direct calculation, we have
\[ \mathbb{E}_n[\|\hat{\eta}_m\|_2^2] = O_P(m). \]

Taking these together, we obtain (15).

We now turn to prove (13). Observe that
\[ \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij1}^2 \leq \max_{1 \leq i \leq n} \|\hat{X}_i - X_i\|^2 \times O_P(m^{\alpha+1}), \]
\[ \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij2}^2 \leq \max_{1 \leq i \leq n} \|X_i\|^2 \times O_P(m^{3\alpha+3}\Delta^\gamma), \]
\[ \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij5}^2 \leq \max_{1 \leq i \leq n} \|X_i\|^2 \times O_P(m^{\alpha+3}n^{-1}). \]

Since \( \int \mathbb{E}[X^4] \leq C \), we have
\[ \max_{1 \leq i \leq n} \|X_i\|^2 = O_P(n^{1/2}), \]
which leads to that
\[ \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij2}^2 = O_P(n^{1/2}m^{3\alpha+3}\Delta^\gamma), \quad \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij5}^2 = O_P(m^{\alpha+3}n^{-1/2}). \]

Using the trivial bound \( \max_{1 \leq i \leq n} \|\hat{X}_i - X_i\|^2 \leq \sum_{i=1}^{n} \|\hat{X}_i - X_i\|^2 \), we also have
\[ \max_{1 \leq i \leq n} \sum_{j=1}^{m} \hat{\Delta}_{ij1}^2 = O_P(nm^{\alpha+1}\Delta^\gamma). \]

Similarly, since \( \mathbb{E}[\eta_{ij}^4] = \kappa_j^{-2}\mathbb{E}[\xi_{ij}^4] \leq C \) by Assumption (A2), we have
\[ \max_{1 \leq i \leq n} \sum_{j=0}^{m} \eta_{ij}^2 = O_P(mn^{1/2}). \]

Taking these together, we have
\[ \max_{1 \leq i \leq n} \|\hat{\eta}_m\|_2^2 = O_P(nm^{\alpha+1}\Delta^\gamma + n^{1/2}m^{3\alpha+3}\Delta^\gamma + m^{\alpha+3}n^{-1/2} + mn^{1/2}). \]

Since \( \alpha > 1, \beta > \alpha/2 + 1 \) and \( m^{3\alpha+3}\Delta^\gamma \to 0 \), there exists a small constant \( c > 0 \) (depending on \( \alpha \) and \( \beta \)) such that the right side is \( O_P(n^{-c}(n/m)) \). This implies (13).

\[ \Box \]
B.2. Proofs of (17) and (18)

We first prove (17). Observe that
\[(h^m \cdot \hat{\eta}^m_i)^2 - (h^m \cdot \eta_i^m)^2 = \{h^m \cdot (\hat{\eta}^m_i - \eta_i^m)\}^2 + 2(h^m \cdot \eta_i^m)\{h^m \cdot (\hat{\eta}^m_i - \eta_i^m)\},\]
by which we have for all \(h^m \in \mathbb{S}^m\),
\[
\left|\mathbb{E}_n[(h^m \cdot \hat{\eta}^m_i)^2] - \mathbb{E}_n[(h^m \cdot \eta_i^m)^2]\right| \leq \mathbb{E}_n[||\hat{\eta}^m_i - \eta_i^m||_2^2] + 2(\mathbb{E}_n[(h^m \cdot \eta_i^m)^2])^{1/2}(\mathbb{E}_n[||\hat{\eta}^m_i - \eta_i^m||_2^2])^{1/2}.
\]

By the proof of (15), we have
\[
\mathbb{E}_n[||\hat{\eta}^m_i - \eta_i^m||_2^2] = o_P(1).
\]

While by Rudelson’s inequality (Theorem 3 in Appendix C), we have
\[
\mathbb{E}\left[\sup_{h^m \in \mathbb{S}^m} |\mathbb{E}_n[(h^m \cdot \eta_i^m)^2] - 1|\right] \leq C \frac{\log n}{n} \mathbb{E}[\max_{1 \leq i \leq n} ||\eta_i^m||_2^2],
\]
provided that the right side is smaller than 1. Since \(\mathbb{E}[\eta_i^m] = \kappa_j^{-2}\mathbb{E}[\xi_j^2] \leq C\) for all \(j \geq 1\) by Assumption (A2), by Lemma 5 in Appendix C, we have
\[
\mathbb{E}[\max_{1 \leq i \leq n} ||\eta_i^m||_2^2] = O(mn^{1/2}).
\]

Therefore, we conclude that
\[
\sup_{h^m \in \mathbb{S}^m} |\mathbb{E}_n[(h^m \cdot \eta_i^m)^2] - 1| = O_P(n^{-1/4}m^{1/2}(\log n)^{1/2}) = o_P(1),
\]
so that uniformly in \(h^m \in \mathbb{S}^m\),
\[
\mathbb{E}_n[(h^m \cdot \hat{\eta}^m_i)^2] = \mathbb{E}_n[(h^m \cdot \eta_i^m)^2] + o_P(1) + O_P(1) \times o_P(1)
\]
\[= 1 + o_P(1).
\]

This completes the proof of (17).

We now turn to prove (18). Observe that
\[
\hat{\nu}_i^2(u) \leq 2\{(\hat{\eta}_i^m - \eta_i^m) \cdot d^m(u)\}^2 + 2\left\{\sum_{j=m+1}^{\infty} d_j(u)\eta_{ij}\right\}^2.
\]

Since \(\mathbb{E}[\eta_{ij}] = 0, \mathbb{E}[\eta_{ij}^2] = 1\) and \(\mathbb{E}[\eta_{ij}\eta_{ik}] = 0\) for all \(j \neq k\), we have
\[
\mathbb{E}\left[\left\{\sum_{j=m+1}^{\infty} d_j(u)\eta_{ij}\right\}^2\right] = \sum_{j=m+1}^{\infty} d_j^2(u) = \sum_{j=m+1}^{\infty} \kappa_j b_j^2(u)
\]
\[\leq C \sum_{j=m+1}^{\infty} j^{-\alpha - 2\beta} = O(mn^{-1}).
\]
By the proof of (15), we also have
\[
\sup_{u \in \mathcal{U}} \mathbb{E}_n[(\hat{\eta}_m - \eta^m) \cdot d^m(u)]^2 \leq C \sum_{j=1}^m j^{-\alpha - 2\beta} \mathbb{E}_n[(\hat{\eta}_j - \eta_j)^2] = O_P(\sum_{j=1}^m j^{-\alpha - 2\beta}(n^{-1}j^{\alpha+2} + \Delta^j j^{3\alpha+2})).
\]
Here, we have
\[
\sum_{j=1}^m j^{-2\beta + 2} = O(1), \quad \sum_{j=1}^m j^{-2\beta + 2\alpha + 2} = \begin{cases} O(1), & \text{if } -2\beta + 2\alpha + 2 < -1, \\ O(\log n), & \text{if } -2\beta + 2\alpha + 2 = -1, \\ O(m^{-2\beta + 2\alpha + 3}), & \text{if } -2\beta + 2\alpha + 2 > -1. \end{cases}
\]
Since \(m^{-2\beta + 2\alpha + 3} \lesssim n^{-1}m^{3\alpha + 3}\) and \(m^{3\alpha + 3} \asymp n\) when \(-2\beta + 2\alpha + 2 = -1\), we have
\[
\sum_{j=1}^m j^{-\alpha - 2\beta}(n^{-1}j^{\alpha+2} + \Delta^j j^{3\alpha+2}) = O(n^{-1} + \Delta^j n^{-1}(\log n)m^{3\alpha + 3}\Delta^\gamma) = O(n^{-1}).
\]
Taking these together, we obtain (18). This completes the proof.

**B.3. Proof of (20)**

Consider the classes of functions
\[
\mathcal{G}_1 = \{\mathbb{R} \times D[0, 1] \times \mathbb{R}^{m+1} \ni (y, x, \eta^m) \mapsto 1(y \leq \eta^m \cdot (d^m(u) + \delta^m))(h^m \cdot \eta^m) : u \in \mathcal{U}, h^m \in S^m, \delta^m = M\sqrt{m/nh^m}\},
\]
and
\[
\mathcal{G}_2 = \{\mathbb{R} \times D[0, 1] \times \mathbb{R}^{m+1} \ni (y, x, \eta^m) \mapsto 1(y \leq Q_{Y|X}(u \mid x))(h^m \cdot \eta^m) : u \in \mathcal{U}, h^m \in S^m\}. \]

It is relatively standard to see that \(\mathcal{G}_1\) is a VC subgraph class with VC index bounded by \(cn\) for some constant \(c \geq 1\) [see 2, Lemma 18]. For \(\mathcal{G}_2\), observe first that \(1(y \leq Q_{Y|X}(u \mid x)) = 1(F_{Y|X}(y \mid x) \leq u)\). Since \(F_{Y|X}(y \mid x)\) is a fixed function, it is also shown that \(\mathcal{G}_2\) is a VC subgraph class with VC index bounded by \(c'n\) for some constant \(c' \geq 1\). The conclusion now follows from an application of Theorem 2.6.7 of [36] and a simple covering number calculation.

**Appendix C: Useful inequalities**

We introduce some useful inequalities.
Theorem 3 (Rudelson’s (1999) inequality). Let $Z_1, \ldots, Z_n$ be i.i.d. random vectors in $\mathbb{R}^k$ with $\Sigma := E[Z_1Z_1']$. Then, for all $k \geq c^2$,

$$
\mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n Z_iZ_i' - \Sigma \right\|_{\text{op}} \right] \leq \max\{\|\Sigma\|_{\text{op}}^{1/2} \delta, \delta^2\}, \quad \delta = C \sqrt{\frac{\log k}{n} \mathbb{E} \max_{1 \leq i \leq n} \|Z_i\|_{\ell_2}},
$$

where $\| \cdot \|_{\text{op}}$ is the operator norm and $C$ is a universal constant.

The expression of Theorem 3 is slightly different from Rudelson’s original form, but is directly deduced from his proof. Theorem 3 gives moment bounds on the difference between empirical and population Gram matrices in the operator norm. Recall that for any $k \times k$ symmetric matrix $A$, $\|A\|_{\text{op}} = \max_{v \in \mathbb{S}^{k-1}} |v'Av|$. To apply Rudelson’s inequality, we have to bound $\mathbb{E} \max_{1 \leq i \leq n} \|Z_i\|_{\ell_2}$, which is typically implemented by using the following lemma.

Lemma 5. Let $X_1, \ldots, X_n$ be arbitrary scalar random variables such that $\max_{1 \leq i \leq n} \mathbb{E}[|X_i|^r] < \infty$ for some $r \geq 1$. Then, we have

$$
\mathbb{E} \max_{1 \leq i \leq n} |X_i| \leq C_r n^{1/r},
$$

where $C_r$ is a constant depending only on $r$ and $\max_{1 \leq i \leq n} \mathbb{E}[|X_i|^r]$.

For the proof, see van der Vaart and Wellner [36, Lemma 2.2.2].

In what follows, we introduce “conditional” maximal inequalities. Below we assume the class of functions to be a “pointwise measurable class” to avoid a measurability complication. A class of measurable functions $\mathcal{G}$ on a measurable space $S$ is said to be pointwise measurable if there exists a countable class of measurable functions $\mathcal{H}$ on $S$ such that for any $g \in \mathcal{G}$, there exists a sequence \( \{h_m\} \subset \mathcal{H} \) with \( h_m(x) \to g(x) \) for all \( x \in S \). See Chapter 2.3 of [36]. This condition is satisfied in our application.

Proposition 2. Let $\Omega, \mathcal{A}, \mathbb{P}$ denote the underlying probability space. Let $\mathcal{D}$ be a sub $\sigma$-field of $\mathcal{A}$. Let $\{(u_i, v_i)\}_{i=1}^n$ be a sequence of random variables taking values in some measurable space $S$ such that $v_1, \ldots, v_n$ are $\mathcal{D}$-measurable, the regular conditional distribution of $(u_1, \ldots, u_n)$ given $\mathcal{D}$ exists, and conditional on $\mathcal{D}$, $u_1, \ldots, u_n$ are independent. Let $\mathcal{G}$ be a pointwise measurable class of functions on $S$ such that for some $\mathcal{D}$-measurable random variables $\hat{B}$ and $\hat{\tau}$,

$$
(i) \sup_{g \in \mathcal{G}} |g(u_i, v_i)| \leq \hat{B}, \quad 1 \leq \forall i \leq n,
$$

$$
(ii) \sup_{g \in \mathcal{G}} \mathbb{E}_n[\mathbb{E}[g^2(u_i, v_i) \mid \mathcal{D}]] \leq \hat{\tau}^2,
$$

$$
(iii) \hat{\tau} \leq \hat{B},
$$

almost surely. Suppose that there exist constants $A \geq 3\sqrt{e}$ and $W \geq 1$ such that

$$
N(\hat{B} \epsilon, \mathcal{G}, L_2(P_n)) \leq (A/\epsilon)^W, \quad 0 < \forall \epsilon \leq 1,
$$

(23)
where \( P_n \) denotes the empirical distribution on \( S \) that assigns probability \( n^{-1} \) to each \((u_i, v_i), i = 1, \ldots, n\). Let \( \sigma_1, \ldots, \sigma_n \) be independent Rademacher random variables defined on another probability space. Extend the underlying probability space by the product probability space. Then, we have

\[
E \left[ \sup_{g \in G} |E_n[\sigma_i g(u_i, v_i)]| \mid D \right] \leq 1(\hat{\tau} > 0)D \sqrt{\frac{2W}{n}} \log \frac{A\hat{B}}{\hat{\tau}} + \frac{W}{n} \log \frac{A\hat{B}}{\hat{\tau}}, \ a.s.,
\]

where \( D \) is a universal constant.

Proposition 2 is a conditional version of Proposition 2.1 of [20]. See also Theorem 3.1 of [21]. Here, \( \{(u_i, v_i)\}_{i=1}^n \) are not necessarily independent. However, conditional on \( D \), \( \{u_i\}_{i=1}^n \) are independent.

Proof of Proposition 2. The proof is a modification of that of Giné and Guillou [20, Proposition 2.1]. For the sake of completeness, we provide the full proof. Suppose that \( E[\sup_{g \in G} E_n[g^2(u_i, v_i)] \mid D] \wedge \hat{\tau} > 0 \). Otherwise the conclusion follows trivially. By Dudley’s inequality [see 36, Corollary 2.2.8], we have

\[
E \left[ \sup_{g \in G} \sqrt{n}E_n[\sigma_i g(u_i, v_i)] \mid \{u_i, v_i\}_{i=1}^n \right] \leq D \int_0^\theta \sqrt{\log N(\epsilon, G, L_2(P_n))} d\epsilon,
\]

where \( \theta := (\sup_{g \in G} E_n[g^2(u_i, v_i)]^{1/2} \mid D) \) and \( D \) is a universal constant. Suppose that \( \theta > 0 \). Using changes of variables, we have

\[
\int_0^\theta \sqrt{\log N(\epsilon, G, L_2(P_n))} d\epsilon = \hat{B} \int_0^{\theta/\hat{B}} \sqrt{\log N(\hat{B}\epsilon, G, L_2(P_n))} d\epsilon
\]

\[
\leq \hat{B} \sqrt{W} \int_0^{\theta/\hat{B}} \sqrt{\log(1/\epsilon)} d\epsilon
\]

\[
\leq \hat{B} \sqrt{W} \int_0^{\infty} \frac{\sqrt{\log(1/\epsilon)}}{\epsilon^2} d\epsilon.
\]

Integration by parts gives

\[
\int_0^{\infty} \frac{\sqrt{\log(1/\epsilon)}}{\epsilon^2} d\epsilon = \left[ -\frac{\sqrt{\log(1/\epsilon)^2}}{\epsilon} \right]_e + \frac{1}{2} \int_e^{\infty} \frac{1}{\epsilon^2 \sqrt{\log(1/\epsilon)}} d\epsilon
\]

\[
\leq \frac{\sqrt{\log(1/e)}}{e} + \frac{1}{2} \int_e^{\infty} \frac{\sqrt{\log(1/\epsilon)}}{\epsilon^2} d\epsilon,
\]

provided that \( c \geq e \), by which we have

\[
\int_e^{\infty} \frac{\sqrt{\log(1/\epsilon)}}{\epsilon^2} d\epsilon \leq \frac{2 \sqrt{\log c}}{c}, \text{ if } c \geq e.
\]

Since \( A\hat{B}/\theta \geq A \geq 3\sqrt{e} > e \), we have

\[
(24) \leq 2\sqrt{W} \theta \sqrt{\log(A\hat{B}/\theta)},
\]
by which we have, using Hölder’s inequality,
\[
E[(24) \mid D] \leq \sqrt{2W} \sqrt{E \left[ 1(\theta > 0)\theta \log \frac{A^2 \hat{B}^2}{\theta^2} \mid D \right]}.
\]
For any fixed \( c > 0 \), define \( f(u) = u \log (c/u) \) if \( u > 0 \) and \( f(0) = 0 \). Then, \( f(u) \) is concave on \([0, \infty)\). Thus, by Jensen’s inequality, the last expression is bounded by
\[
\sqrt{2W} \sqrt{E \left[ \sup_{g \in G} \mathbb{E}_n [g^2(u_i, v_i)] \mid D \right] \times \log \frac{A^2 \hat{B}^2}{\mathbb{E}_n \left[ \sup_{g \in G} \mathbb{E}_n [g^2(u_i, v_i)] \mid D \right]}}.
\]
Using the decomposition
\[
g^2(u_i, v_i) = \mathbb{E}[g^2(u_i, v_i) \mid D] + \{g^2(u_i, v_i) - \mathbb{E}[g^2(u_i, v_i) \mid D]\},
\]
and the symmetrization inequality conditional on \( D \), we have
\[
\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [g^2(u_i, v_i)] \mid D \right] \leq \sup_{g \in G} \mathbb{E}_n [\mathbb{E}[g^2(u_i, v_i) \mid D]] + 2\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [\sigma, g^2(u_i, v_i)] \mid D \right] \\
\leq \hat{\tau}^2 + 2\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [\sigma, g^2(u_i, v_i)] \mid D \right] \\
\leq \hat{\tau}^2 + 2\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [\sigma, g(u_i, v_i)] \mid D \right].
\]
Using now the contraction principle [see 36, Proposition A.3.2], we have
\[
\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [\sigma, g^2(u_i, v_i)] \mid D \right] \leq \hat{\tau}^2 + 8\hat{B} \mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [\sigma, g(u_i, v_i)] \mid D \right].
\]
Note that the right side is at most \( 9\hat{B}^2 \). Since for any given \( c > 0 \), the map \( u \mapsto u \log (c/u) \) is non-decreasing for \( 0 < u \leq c/e \), and \( A \geq 3\sqrt{\epsilon} \), we have
\[
\mathbb{E} \left[ \sup_{g \in G} \mathbb{E}_n [g^2(u_i, v_i)] \mid D \right] \times \log \frac{A^2 \hat{B}^2}{\mathbb{E}_n \left[ \sup_{g \in G} \mathbb{E}_n [g^2(u_i, v_i)] \mid D \right]} \\
\leq (\hat{\tau}^2 + 8\hat{B}Z) \log \frac{A^2 \hat{B}^2}{\hat{\tau}^2 + 8\hat{B}Z} \\
\leq (\hat{\tau}^2 + 8\hat{B}Z) \log \frac{A^2 \hat{B}^2}{\hat{\tau}^2} \\
= 2(\hat{\tau}^2 + 8\hat{B}Z) \log \frac{A\hat{B}}{\hat{\tau}},
\]
where
\[ Z := \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \mathbb{E}_n [\sigma_i g(u_i, v_i)] \mid D \right]. \]
Taking these together, we have
\[ \sqrt{n} Z \leq 2D \sqrt{W} \sqrt{\hat{\tau}^2 + 8 \hat{B} Z} \log \frac{A \hat{B}}{\hat{\tau}}. \]
Solving this inequality with respect to \( Z \) gives the desired bound. \( \square \)

**Proposition 3.** Consider the same setting as in Proposition 2. Instead of (i)-(iii) and \((23)\), suppose that there is an envelope function \( G \) for \( \mathcal{G} \) such that for some constants \( A \geq e \) and \( W \geq 1 \),
\[ N(\epsilon \|G\|_{L_2(P_n)}, \mathcal{G}, L_2(P_n)) \leq \frac{A}{\epsilon} W, \quad 0 < \forall \epsilon \leq 1. \]
Then, we have for all \( q \in [1, \infty) \),
\[ \left( \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \mathbb{E}_n [\sigma_i g(u_i, v_i)]^q \mid D \right] \right)^{1/q} \leq Dn^{-1/2} \left( \mathbb{E}_n [\mathbb{E}[G^q(u_i, v_i) \mid D]] \right)^{1/q} \sqrt{W \log A} \text{ a.s.,} \]
where \( \hat{q} = q \vee 2 \) and \( D \) is a universal constant.

**Proof.** By Dudley’s inequality,
\[ \left( \mathbb{E} \left[ \sup_{g \in \mathcal{G}} \mathbb{E}_n [\sigma_i g(u_i, v_i)]^q \mid \{(u_i, v_i)\}_{i=1}^n \right] \right)^{1/q} \leq D \int_0^\theta \sqrt{\log N(\epsilon, \mathcal{G}, L_2(P_n))} d\epsilon, \]
where \( \theta := (\sup_{g \in \mathcal{G}} \mathbb{E}_n [g^2(u_i, v_i)])^{1/2} \leq (\mathbb{E}_n [G^2(u_i, v_i)])^{1/2} \) and \( D \) is a universal constant. Using changes of variables implies that the right side is bounded by
\[ D (\mathbb{E}_n [G^2(u_i, v_i)])^{1/2} \sqrt{W} \int_0^1 \sqrt{\log(A/\epsilon)} d\epsilon. \]
If \( q \geq 2 \), then by Hölder’s inequality,
\[ \mathbb{E}(\mathbb{E}_n [G^2(u_i, v_i)]^{q/2} \mid D) \leq \mathbb{E} \mathbb{E}_n [G^q(u_i, v_i) \mid D] = \mathbb{E}_n [\mathbb{E}[G^q(u_i, v_i) \mid D]]. \]
On the other hand, if \( q \in [1, 2) \),
\[ \mathbb{E}(\mathbb{E}_n [G^2(u_i, v_i)]^{q/2} \mid D) \leq (\mathbb{E}_n [\mathbb{E}[G^2(u_i, v_i) \mid D]])^{q/2}. \]
This leads to the desired inequality. \( \square \)