On High-Dimensional Misspecified Quantile Regression

by

Alexander Giessing

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2018

Doctoral Committee:

Professor Xuming He, Chair
Assistant Professor Andreas Hagemann
Associate Professor Long Nguyen
Assistant Professor Gongjun Xu

Alexander Giessing

giessing@umich.edu

ORCID iD: 0000-0002-6917-0652

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my advisor Xuming He. While I entered graduate school to independently explore and to freely pursue a topic of my own interest, Xuming's advice and guidance, his patience (which I surely abused too often!), his seemingly unlimited availability to meet and to discuss questions, and our weekly group meetings were in fact indispensable to me. I feel extremely fortunate to have him as my mentor and advisor.

I would like to thank my committee members Andreas Hagemann, Long Nguyen and Gongjun Xu. I appreciate that they took time to look at my work and responded with thoughtful, critical comments. I am also grateful for the time and effort which Moulinath Banerjee, Long Nguyen, and Ji Zhu invested in writing letters of recommendation on several occasions.

I would like to thank Jingshen Wang for joining me in a timely reading group on modern empirical process theory and our on-going collaboration. I look forward to seeing our ideas getting published.

Lastly, I would like to thank my parents for their constant encouragement and support.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In this dissertation we develop theory for inference and uncertainty quantification for potentially misspecified quantile regression processes when the number of predictor variables increases with or exceeds the sample size. Potential misspecification of the fitted model is a fundamental problem in statistics which is exacerbated by today's high-dimensional datasets, and quantile regression is often used in complex situations in which misspecifications are highly likely.

We make the following contributions: First, we establish a uniform-in-model strong Bahadur representation for misspecified quantile regression processes when the number of predictor variables increases and provide tight error bounds on its remainder term which hold uniformly over growing collections of quantile regression functions. Second, we derive an almost sure de-biased representation of the Lasso-penalized high-dimensional misspecified quantile regression process and analyze the theoretical properties of the misspecified post-Lasso quantile regression estimator. Third, to quantify the uncertainty associated with a misspecified quantile regression function we analyze its predictive risk and expected optimism. We propose uniformly consistent estimators for both quantities when the number of regression functions is growing moderately with the sample size. Empirical evidence shows that our estimators perform favorably against cross-validation estimates. Forth, we develop a set of new exponential and maximal inequalities which allow to control the fluctuations of a collection of suprema of empirical processes over classes of unbounded functions when both the collection of function classes and the complexity of each individual the function class grow with the sample size. These new inequalities are instrumental in deriving the theoretical results in this dissertation.

CHAPTER 1

Introduction

Misspecification is a universal phenomenon in statistical modeling and correct specification is rare in reality. Today's high-dimensional datasets exacerbate this problem as they often exhibit heteroscedasticity, asymmetries, and missingness. Quantile regression, which models the impact of predictor variables on the conditional distribution of a response variable, is routinely used in complex situations in which misspecification is highly likely. In this dissertation we therefore develop theory for inference and uncertainty quantification for potentially misspecified quantile regressions when the dimension of the regression functions increases with or exceeds the sample size.

To study the asymptotic properties of a statistical estimator it is often useful to approximate it by a sum of independent random variables with a higher-order remainder term. Such first-order approximations are called Bahadur representations after the pioneering work by Bahadur (1966). In Chapter 2, we consider collections of potentially misspecified quantile regression processes and derive a strong Bahadur representation which holds uniformly over a continuum of quantile levels and a growing number of regression functions. Such a uniform-in-model Bahadur representation has recently proved useful in developing theory for post-selection and selective inference in high-dimensional statistical models based on (penalized) least squares estimators (e.g. Berk et al., 2013; Lee et al., 2016; Tian and Taylor, 2017; Kuchibhotla et al., 2018). We do not pursue this strand of applications in our dissertation; however, in principle, our results allow to develop analogous (asymptotic) theory for post-selection and selective inference on the quantile regression process.

The derivation of the strong uniform-in-model Bahadur representation relies on new exponential and maximal inequalities which allow us to control the fluctuation of collections suprema of empirical processes over a classes of unbounded functions when the collection well as the complexity of the function classes depend on the sample size. Our new inequalities are based on Panchenko's (2003) concentration inequality and the conceptually simple (but technically challenging) idea of re-organizing the individual functions classes into equivalence classes of functions consisting of functions with roughly equal variance and

equal metric entropy. Our inequalities thus have a flavor of the inequalities for ratio-type and self-normalized empirical processes (e.g. Bercu et al., 2002; Giné and Koltchinskii, 2006; Massart, 2007). They significantly improve over similar exponential inequalites in Lemma 16 and 18 in Belloni and Chernozhukov (2011) and the maximal inequality in Theorem 3.1 in van der Vaart and Wellner (2011).

Sparse modeling of high-dimensional datasets has attracted a flurry of recent research. In the case of correctly specified high-dimensional sparse linear models a common strategy is to use an ℓ_1-penalized estimator to enforce sparsity on the vector of estimated coefficients. This approach was first proposed by Tibshirani (1996) in the context of least squares problems; its extension to the quantile regression problem was developed and analyzed by Belloni and Chernozhukov (2011). Recently, Bühlmann and van de Geer (2015) investigated ℓ_1-penalized estimators for misspecified high-dimensional sparse models of the conditional mean. In Chapter 3, we extend the Bühlmann and van de Geer's (2015) analysis to the problem of conducting inference on high-dimensional sparse quantile regression processes when the assumed linear regression function is misspecified. We establish strong consistency of the ℓ_1-penalized misspecified quantile regression process and derive a strong de-biased Bahadur-type representation for the misspecified quantile regression process. This generalizes and strengthens a similar representation of the quantile regression process by Bradic and Kolar (2017). We also provide a theoretical analysis of the post-Lasso estimator for the misspecified quantile regression process. In particular, we are able to characterize certain effects of model misspecification on the refitted quantile regression process. This complements results on the correctly specified post-Lasso estimator derived by Belloni and Chernozhukov (2011). The results in Chapter 3 are based on the strong uniform-in-model Bahadur representation and the new exponential and maximal inequalities derived in Chapter 2 and thus illustrate the potential of these technical developments.

Predictive modeling is at the core of many scientific disciplines, including business, engineering, finance, and public health. A natural way to gauge the predictive capability of a statistical model is to estimate its predictive risk. In recent years, the predictive risk from quantile models has gained significant interest in finance and risk management to assess the value-at risk and expected shortfall of investments (e.g. Xiao et al., 2015; Gaglianone et al., 2011; Engle and Manganelli, 2004) and to solve portfolio choice problems (e.g. Cahuich and Hernández-Hernández, 2013; He and Zhou, 2011; Bassett et al., 2004). In Chapter 4 we therefore analyze the predictive risk of possibly misspecified quantile regression models. We contribute to the theory of predictive risk evaluation of quantile regression models by deriving two (asymptotic) characterizations of the expected optimism of the in-sample risk and proposing a uniformly consistent, de-biased estimator of the predictive risk. These

two characterizations generalize and unify existing results on covariance penalties (Efron, 2004), covariance inflation criteria (Tibshirani and Knight, 1999), generalized degrees of freedom (e.g. Ye, 1998), and robust and generalized Akaike-type selection criteria for misspecified quantile regression models (e.g. Lv and Liu, 2014; Portnoy, 1997; Burman and Nolan, 1995). Both characterizations show that large part of the expected optimism can be attributed to a nonlinear function of the quantile level, the conditional density of the response variable given the predictors and the (weighted) covariance matrix of the predictors. We conclude that the commonly used notion of effective degree of freedom for a statistical model has a richer content for misspecified models.

These theoretical investigations lend themselves to a simple plug-in estimator for the expected optimism. We establish its uniform consistency over a class of candidate models and, based on this result, propose a uniformly consistent, de-biased estimate of the predictive risk. While these are large sample results, empirical evidence suggests that the de-biasing procedure provides a significant correction even in finite samples and when the model size is fixed and relatively small compared to the sample size. A comparison of our de-biased estimate against the popular method of cross-validation is favorable for our procedure.

CHAPTER 2

A Strong Uniform-In-Model Bahadur Representation for Misspecified Quantile Regression Processes

2.1 Introduction

To study the asymptotic properties of a statistical estimator it is often useful to approximate it by a sum of independent random variables with a higher-order remainder term. Such first-order approximations are called Bahadur representations after the pioneering work by Bahadur (1966). Bahadur representations may be used to determine the asymptotic distribution of a single estimator or to establish joint and process weak convergence of collections of estimators. Moreover, tight bounds on the higher-order remainder term may give insight into the impact of model misspecification on the asymptotic properties of the estimator.

The wide applicability of Bahadur-type representations has led to the development of a rich theory. For example, in the classical large sample framework, in which the sample size tends to infinity and the dimension of the parameter space is fixed, Carroll (1978) obtained strong representations for M-estimators characterized by their score functions; Niemiro (1992) and Arcones (1996) derived strong representations for M-estimators defined by minimizing convex functions; Koenker and Portnoy (1987), Babu (1989), and Arcones (1998) established strong representations for the quantile regression and the least absolute deviation estimator; He and Shao (1996) obtained strong representations for general M-estimators under non-stochastic design; and Wu (2007) considered weak and strong representations for M-estimators of linear regression models with dependent errors.

The quest for understanding the large sample behavior of estimates when the dimension of the parameter space diverges with the sample size has led to yet another body of literature. Initiated by Huber (1973) and further developed by Portnoy (1985), Mammen (1989)

and Welsh (1989) the definitive papers on Bahadur-type representations for estimators in increasing dimensions are Bai and Wu (1994) and He and Shao (2000).

The consolidation of the field of high-dimensional statistics over the last two decades has recently renewed the interest in first-order approximations of estimators by sums of independent random variables, e.g. Chernozhukov et al. (2013, 2014). It is now widely recognized that the major challenge in establishing asymptotic theory for inference, hypothesis testing and uncertainty quantification in high dimensions lies in properly accounting for the model selection procedure that is part of all high-dimensional estimation techniques, e.g. Leeb and Pötscher (2005, 2006); Zhang and Zhang (2013); van de Geer et al. (2014); Lee et al. (2016). To address this issue a number of researcher have obtained application-specific maximal inequalities and Bahadur-type representations that hold uniformly over collections of models. For example, Berk et al. (2013) obtained post-selection coverage guarantees for confidence intervals of least squares estimators that hold simultaneously over a range of models; Belloni and Chernozhukov (2011, 2013) assessed properties of quantile regression and least squares estimators after model selection via maximal inequalities that hold uniformly over collections of models; and Kuchibhotla et al. (2018) obtained a Bahadur representation for least squares estimators that holds uniformly over all subsets of possible models based on given number of predictors. We think that this development is a promising step towards a principled theory for high-dimensional inference. Therefore, in this paper, we derive a strong uniform-in-model Bahadur representation for the quantile regression processes in increasing dimensions. For illustrations and applications we refer to our companion work, in which the results established here are applied to three important statistical problems: the analysis of the post-selection quantile regression estimator under misspecification, the high-dimensional de-biased quantile regression process, and the predictive risk of misspecified quantile regression models.

The contributions of this paper are twofold. The first and main contribution is to establish tight almost sure bounds on the remainder term of the Bahadur representation for potentially misspecified quantile regression processes in increasing dimensions that hold uniformly over a growing collection of models. Our results hold for non-identically distributed and non-Gaussian data. We show that under mild regularity conditions the quantile regression process for estimators constructed from m out of d possible predictors and based on a sample of size n can be approximated by a sum of independent variables up to an error of order $O\left((m\log(ed/m)+\log\log n)^{3/4}n^{-3/4}\right)$ almost surely and uniformly over all possible models of size less or equal to m provided that $m\log(ed/m)=o(n)$. Thus, our bound matches the optimal stochastic order of the remainder term of Bahadur representations for quantiles in fixed dimension and under i.i.d data established in Kiefer (1967). As a side result

we obtain a better understanding of the nature of the restricted nonlinearity impact condition for the quantile regression process introduced in Belloni and Chernozhukov (2011). Our uniform-in-model Bahadur representation is applicable to parametric, nonparametric, and nonlinear quantile regression estimators. Easy corollaries include consistency and process weak convergence of the quantile regression process under misspecification and increasing number of parameters.

The second contribution is a new exponential inequality to control the fluctuation of a collection M of suprema of empirical processes over a class $\mathscr{F}$ of unbounded functions when the collection M as well as the complexity of the function class $\mathscr{F}$ depend on the sample size. The inequality is based on Panchenko's (2003) concentration inequality and the (conceptually) simple idea of splitting the class $\mathscr{F}$ and collection M into slices consisting of functions and models with roughly equal variance and equal metric entropy. The inequality thus has a flavor of the inequalities for ratio-type and self-normalized empirical processes developed by, among others, Bartlett et al. (2005), Giné and Koltchinskii (2006), Massart (2007) and Bercu et al. (2002). We use this exponential inequality to derive an almost sure bound on the suprema of empirical processes indexed by an unbounded function class. Establishing an almost sure bound requires proving a bounded law of iterated logarithm for the supremum of an empirical process over a class of unbounded functions, which is a major challenge in our proof. Our result does not follow from the existing theory on ratio-type statistics which hold for bounded functions only. It also does not follow from known results on self-normalized empirical processes. We complement this general exponential inequality with a maximal inequality for the mean of the supremum of an empirical process indexed by a function class relevant to quantile regression. This quantile specific result significantly improves upon the bounds derived in a more generic setting in van der Vaart and Wellner (2011).

Since the seminal work of Koenker and Bassett (1978) quantile regression is one of the main topics in statistics and econometrics. An appealing feature of quantile regression is that it allows the practitioner to conduct inference on the entire conditional distribution of the response variable by estimating a collection of different conditional quantiles. We refer to Koenker (2005) for a standard textbook on quantile regression. For studies of the statistical properties of misspecified quantile regression in the classical large sample setting with a fixed number of parameters we refer to Kim and White (2003), Angrist et al. (2006), and Noh et al. (2013). The results in Angrist et al. (2006) comprise a weak Bahadur representation. Potential misspecification of the quantile regression process when the number of parameters increases has been addressed partially in specific settings in two recent papers: Under the assumption that the misspecification vanishes asymptotically Belloni et al. (2017)

showed that the quantile regression process based on series estimators can be strongly approximated by a sequence of Gaussian processes. Under similar assumptions Chao et al. (2017) established a Bahadur representation (and its convergence to a limit Gaussian process) of the quantile processes of semi- and nonparametric regression with exponential tail bounds on the error terms. Neither of the two representations achieve the known optimal rate of the remainder term of the Bahadur representations, nor do they hold (nor can they be extended to hold) uniformly over a growing collection of models, almost surely and under persistent misspecification. In fact, our representation comprises the results on series and nonparametric quantile regression processes as special cases.

The remainder of the paper is organized as follows. In Section 2.2 we introduce the statistical framework and present the main results, i.e. the almost sure and finite sample uniform-in-model Bahadur representation for potentially misspecified quantile regression processes. Using this result, we briefly comment on the uniform-in-model consistency of the misspecified quantile regression process and discuss the (restricted) nonlinearity impact condition which has been introduced by Belloni and Chernozhukov (2011) in the context of ℓ_1-penalized high dimensional quantile regression. In Section 2.3 we provide the heuristics behind our proof strategy, the general exponential inequality for collections of suprema of empirical processes over classes of unbounded functions and the quantile regression specific maximum inequality. We conclude in Section 2.4 with a brief discussion about further applications and generalizations of the concepts developed in this paper. We defer all proofs and most technical details to Section 2.5.

We explain the notation used in the paper. In what follows, we implicitly index all parameters by the sample size n. Thus, when making asymptotic statements, we assume that $n \to \infty$ and $d = d_n \to \infty$ and $m = m_n \to \infty$. But we omit the index whenever this does not cause confusion. Constants $c, C, c_1, c_2, \ldots$ are understood to be independent of n and may change from line to line. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max a, b$, and $a \wedge b = \min\{a, b\}$. We denote the ℓ_2-norm by $\|\cdot\|_2$, the ℓ_1-norm by $\|\cdot\|_1$, the ℓ_∞-norm by $\|\cdot\|_\infty$ and the "ℓ_0-norm" by $\|\cdot\|_0$ (i.e. the number of nonzero components). For $r > 0$, $v \in \mathbb{R}^d$ we use $\mathscr{B}^d(v, r)$ to denote the ball in $\mathbb{R}^d$ with center at v and radius r with respect to the Euclidean norm. We shall abbreviate $\mathscr{B}^d(0, 1)$ to $\mathscr{B}^d$. Analogously, we denote spheres of radius $r > 0$ by $\mathscr{S}^d(v, r)$ and write $\mathscr{S}^d$ for $\mathscr{S}^d(0, 1)$. Given a vector $X \in \mathbb{R}^d$ and a set of indices $M \subseteq \{1, \ldots, d\}$ we let X_M denote the vector $\{X^{(j)}, j \in M\}$. The cardinality of M is denoted by $|M|$. We denote the quantile loss function for quantile level $\tau \in (0, 1)$ by $\rho_\tau(u) = u(1 - 1\{u \leq 0\})$ and its subgradient by $\varphi_\tau(u) = \tau - 1\{u \leq 0\}$. We use the terms subgradient and quantile regression score function interchangeably.

Throughout, we use the empirical process notation as defined in van der Vaart and

Wellner (1996); however, to accommodate the triangular array setting we introduce the following modifications: The symbol $\mathrm{E}[\cdot]$ denotes the expectation with respect to a generic probability measure P (which depends on the context). $\mathbb{P}_n$ denotes the empirical measure of the random vectors $\{Z_{ni}, 1 \leq i \leq n\}$ and $\mathbb{E}_{nN}[\cdot]$ denotes the empirical average over the first $N \leq n$ random vectors (ordered by their indices) distributed according to the empirical measure $\mathbb{P}_n$, i.e. $\mathbb{E}_{nN}[f] := \mathbb{E}_{nN}[f(Z_{ni})] = N^{-1}\sum_{i=1}^{N} f(Z_{ni})$. In addition, we define $\bar{\mathrm{E}}_{nN}[f] = \mathbb{E}_{nN}[\mathrm{E}[f]]$ and $\mathbb{G}_{nN}(f) = \sqrt{N}(\mathbb{E}_{nN}[f] - \bar{\mathrm{E}}_{nN}[f])$, and we denote the symmetrized process by $\mathbb{G}^{\circ}_{nN}(f)$. For $r \geq 1$ we denote the $L^r(\mathbb{P}_n)$-norm by $\|f\|_{\mathbb{P}_n,r} = (\mathbb{E}_{nn}[|f|^r])^{1/r}$. We write $\ell^{\infty}(\mathscr{T})$ for the set of all uniformly bounded real-valued functions on $\mathscr{T} \subset (0,1)$.

2.2 Main results

In this section we formulate the general framework within which we plan to analyze the misspecified quantile regression process problem in increasing dimensions, state the regularity conditions, and provide our main theoretical result. We include brief discussions of related results in the literature.

2.2.1 Setting

Let $\{(Y_n, X_n), (Y_{ni}, X_{ni}), 1 \leq i \leq n\}$ be a triangular array of row-wise independent random vectors, where $Y_n \in \mathbb{R}$ is a continuous response variable, $X_n \in \mathbb{R}^d$ a vector of predictor variables and the pair (Y_n, X_n) has joint distribution F_n. The joint distribution F_n may change with the sample size n, as we allow the number of predictor variables d to grow with n. We are interested in the conditional quantile function (CQF) of Y_n given $X_{n,M}$ for a set of quantile levels $\tau \in \mathscr{T} \subset (0,1)$ and all subset $M \subseteq \{1, \ldots, d\}$ of predictor variables,

$$Q_{Y_n}(\tau | X_{n,M}) = \inf\left\{ y : F_{Y_n|X_{n,M}}(y|X_{n,M}) \geq \tau \right\}, \tag{2.1}$$

where $F_{Y_n|X_{n,M}}(\cdot|X_{n,M})$ is the conditional distribution function of Y_n given $\left\{X_n^{(j)} : j \in M\right\}$. The purpose of linear quantile regression is to approximate the CQF by a linear regression function. To this end, we introduce the population quantile regression vector $\theta^*_{n,M}(\tau)$ at quantile level τ and based on model M as the solution to

$$\min_{\theta \in \mathbb{R}^{|M|}} \bar{\mathrm{E}}_{nn}[\rho_\tau(Y_{ni} - X'_{ni,M}\theta) - \rho_\tau(Y_{ni} - Q_{Y_n}(\tau|X_{n,M}))] \tag{2.2}$$

and its sample analogue, the sample (or estimated) quantile regression vector $\hat{\theta}_{n,M}(\tau)$ at quantile level τ and based on model M, as the solution to

$$\min_{\theta \in \mathbb{R}^{|M|}} \mathbb{E}_{nn}[\rho_\tau(Y_{ni} - X'_{ni,M}\theta)]. \tag{2.3}$$

In both displays, $\rho_\tau(u) = u(\tau - 1\{u \leq 0\})$ denotes the quantile loss function introduced by Koenker and Bassett (1978). An estimate of the CQF at quantile level τ based on model M and the predictor X_n is then formed as

$$\widehat{Q}_{Y_n}(\tau|X_{n,M}) = X'_{n,M}\hat{\theta}_{n,M}(\tau). \tag{2.4}$$

Since we are interested in estimating the CQF for a set of quantile levels $\mathscr{T} \subset (0,1)$ and a collection of models $M \subseteq \{1,\ldots,d\}$, we solve the problem (2.3) for all $\tau \in \mathscr{T}$ and all $M \subseteq \{1,\ldots,d\}$ to obtain a collection of quantile regression processes

$$\hat{\theta}_{n,M}(\cdot) = \{\hat{\theta}_{n,M}(\tau) : \tau \in \mathscr{T}\}, \quad M \subseteq \{1,\ldots,d\}. \tag{2.5}$$

We do not assume that the true CQF is indeed a linear function of the vector of predictor variables X_n and therefore the estimated CQF $\widehat{Q}_{Y_n}(\tau|X_{n,M})$ and the quantile regression process $\hat{\theta}_{n,M}(\cdot)$ may both be misspecified. For interpretations, justifications, and theoretical properties of quantile regressions under misspecification in the classical large sample framework, we refer to Angrist et al. (2006). In brief, they showed that under mild regularity conditions and for a given set of predictors M the quantile regression process $\hat{\theta}_{n,M}(\cdot)$ is uniformly consistent for $\theta^*_{n,M}(\cdot)$ and $\sqrt{n}(\hat{\theta}_{n,M}(\cdot) - \theta^*_{n,M}(\cdot))$ converges weakly to a zero mean Gaussian process which is fully characterized by its covariance function. We discuss related results in the context of our more general setting after deriving the strong uniform-in-model Bahadur representation.

Since the predictor variables $X_{n1},\ldots,X_{nn}$ can be evaluations of a d-dimensional dictionary of technical regressors with respect to random variables $Z_{n1},\ldots,Z_{nn}$, the here outlined framework is general enough to include many commonly used quantile regression estimators such as linear and nonlinear quantile regression (e.g. Koenker and Park, 1996), locally linear and polynomial quantile regression (e.g. Lee, 2003; Chaudhuri, 1991), series estimators (e.g. Belloni et al., 2017), and semi- and nonparametric quantile regression (e.g. Chao et al., 2017; Horowitz and Lee, 2005).

2.2.2 Conditions

To obtain a good uniform control over the error rate of the remainder term in the Bahadur representation for the collection $\left\{\hat{\theta}_{n,M}(\cdot), M \subseteq \{1,\ldots,d\}\right\}$ of quantile regression processes, we introduce the following technical assumptions.

(T1) *The data $\{(Y_{ni}, X_{ni}), 1 \leq i \leq n\}$, $(Y_{ni}, X_{ni}) \in \mathbb{R} \times \mathbb{R}^d$ are row-wise independent random vectors with distribution F_{ni}. Dimension d may change with the sample size and distribution F_{ni} may change with the sample size n and index $i \leq n$.*

(T2) *The conditional density $f_{Y_{ni}|X_{ni}}$ of Y_{ni} given predictor variables X_{ni} is uniformly bounded from above, i.e. there exits $f_+ < \infty$ such that for all $n \in \mathbb{N}$ and all $m \in \{1,\ldots,d\}$,*

$$\max_{M:|M|=m} \max_{i \leq n} \sup_{a \in \mathbb{R}} \sup_{x \in \mathbb{R}^d} \left| f_{Y_{ni}|X_{ni,M}}(a|x) \right| \leq f_+.$$

(T3) *The conditional density $f_{Y_{ni}|X_{ni,M}}$ of Y_{ni} given the predictor variable $X_{ni,M}$, is Hölder continuous with exponent $\alpha \in \left[\frac{1}{2}, 1\right]$, i.e. there exists a constant $f_H > 0$ such that all $n \in \mathbb{N}$, $a, b \in \mathbb{R}$ and $m \in \{1,\ldots,d\}$,*

$$\max_{M:|M|=m} \max_{i \leq n} \sup_{x \in \mathbb{R}^m} \left| f_{Y_{ni}|X_{ni,M}}(a|x) - f_{Y_{ni}|X_{ni,M}}(b|x) \right| \leq f_H |a-b|^{\alpha}.$$

(T4) *The predictors X_{ni} are vectors of random variables with finite $4+\delta$ moment for some $\delta > 0$, and there exists an absolute constant $\mu_4 < \infty$ such that for all $n \in \mathbb{N}$,*

$$\max_{\|u\|_1 \leq 1} \mathrm{E}\left[\max_{i \leq n} |X_{ni}' u|^{4+\delta} \right]^{1/(4+\delta)} \leq \mu_4.$$

(T5) *The maximum eigenvalue of the matrix of second moments of the predictor variables X_{ni} is uniformly bounded from above by a function of the dimension $|M|$, i.e. for all $n \in \mathbb{N}$ and all $m \in \{1,\ldots,d\}$,*

$$\left(\max_{M:|M|=m} \sup_{\|u\|_2=1} u' \bar{\mathrm{E}}_{nn}\left[X_{ni,M} X_{ni,M}' \right] u \right) \vee 1 \leq \bar{\varphi}_{\max}(m).$$

(T6) *The minimum eigenvalue of the matrix of weighted second moments of the predictor*

variables $X_{n,M}$,

$$D_{n,M}(\tau) = \bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni,M}}\left(X'_{ni,M}\theta^*_{ni,M}(\tau)|X_{ni,M}\right)X_{ni,M}X'_{ni,M}\right],$$

is uniformly bounded from below by a function of the dimension $|M|$*, i.e. for all* $n \in \mathbb{N}$ *and* $m \in \{1,\dots,d\}$,

$$\inf_{\tau\in\mathscr{T}}\left(\min_{M:|M|=m}\sup_{\|u\|_2=1} u'D_{n,M}(\tau)u\right) > \bar{\varphi}_{\min}(m).$$

Conditions (T1) – (T6) impose mild assumptions on the distribution of response and predictor variables. The uniformity in n is necessary as we consider triangular arrays; uniformity in the model size $|M|$ is necessary as we want to control the remainder term uniformly over all models $M \subseteq \{1,\dots,d\}$. The boundedness and smoothness assumptions (T2) and (T3) on the conditional density are fairly standard in the quantile regression literature. The assumption that the conditional density is (only) Hölder continuous adds additional flexibility to our framework; often the density is simply assumed to be continuous (e.g. Koenker and Portnoy, 1987; Belloni and Chernozhukov, 2011). Assumption (T4) is very mild; it is common practice in the literature on quantile regression in increasing dimensions to assume almost sure boundedness of the predictors, i.e. $\|X_{ni,M}\|_2 \leq \xi_m = O\left(n^b\right)$ almost surely for some $b > 0$ (e.g. Belloni et al., 2017; Chao et al., 2017). Boundedness of the predictor variables is a rather restrictive assumption. We remove this assumption and are still able to significantly improve the rate of the remainder compared to above authors. Assumptions (T5) and (T6) constitute another relaxation of assumptions commonly imposed in the literature. For example, Belloni et al. (2017), Chao et al. (2017) and Bradic and Kolar (2017) assume that the eigenvalues of the Gram matrix are almost surely bounded from below and above by a constant. Replacing this restrictive assumption with a more flexible condition relating the lower and upper bounds to the dimension of the Gram matrix allows for broader applications of our theory. Naturally, the functions $\bar{\varphi}_{\max}(m)$ and $\bar{\varphi}_{\min}(m)$ cannot behave arbitrary; the range of manageable behavior is content of the next two rate conditions.

(R1) Let $\alpha \in [1/2, 1]$ *be the Hölder exponent of the conditional density* $f_{Y_{ni}|X_{ni,M}}$,

$$\max_{1\leq m\leq d}\left\{\bar{\kappa}^{\alpha-1}(m)\bar{\varphi}^{1/2}_{\max}(m)\left(\frac{m\log\left(ed/m^{1/2}\right)+\log\log n}{n}\right)^{(2\alpha-1)/4}\right\} = O(1),$$

where $\bar{\kappa}(m) = \frac{\bar{\varphi}_{\max}(m)}{\bar{\varphi}_{\min}(m)}$.

(R2) Let $\alpha \in [1/2, 1]$ *be the Hölder exponent of the conditional density* $f_{Y_{ni}|X_{ni,M}}$,

$$\max_{1 \leq m \leq d} \left\{ \bar{\kappa}^{\alpha-1}(m) \bar{\varphi}^{1/2}_{\max}(m) \left(\frac{m \log(ed/m) + \log\log n}{n} \right)^{(2\alpha-1)/4} \right\} = O(1),$$

$$\max_{1 \leq m \leq d} \left\{ \bar{\kappa}(m) \bar{\varphi}^{\alpha/2}_{\max}(m) \left(\frac{m \log(ed/m) + \log\log n}{n} \right)^{\alpha/4} \right\} = O(1),$$

where $\bar{\kappa}(m) = \frac{\bar{\varphi}_{\max}(m)}{\bar{\varphi}_{\min}(m)}$.

In order to establish a uniform-in-model Bahadur representation only one of the two rate conditions needs to be satisfied. The implications of and differences between (R1) and (R2) are subtle and become the clearest when one considers the case of a single model of dimension m (i.e. $d = m$), with continuous conditional density (i.e. $\alpha = 1$), and conditioning number $\bar{\kappa}(m) = O(1)$ (independent of model size m). In this case, (R1) simplifies to $\bar{\varphi}_{\max}(m)(m + m\log m + \log\log n)^{1/2} n^{-1/2} = O(1)$. This is a weak assumption as it relates only to the maximum eigenvalue of the population Gram matrix, not the sample Gram matrix. It is easily satisfied by unbounded predictor variables. In contrast, the boundedness assumptions in Belloni et al. (2017), Chao et al. (2017) and Bradic and Kolar (2017) exclude such predictors.

Under the above scenario, the two conditions in (R2) simplify to the single condition $\bar{\varphi}_{\max}(m)(m + \log\log n)^{1/2} n^{-1/2} = O(1)$. Compared to (R1) we see that (R2) is the weaker assumption in this specific case. However, in the general case when the conditioning number does depend on m and multiple models are under consideration (R1) appears to be the weaker assumption. In fact, (R2) introduces an additional nonlinear constraint on the eigenvalues of the population Gram matrix which resembles a (restricted) nonlinearity impact (RNI) condition with explicit rates. RNI conditions were introduced to the quantile regression literature by Belloni and Chernozhukov (2011) in the context of high-dimensional ℓ_1-penalized estimation. The two rate conditions (R1) and (R2) reflect a fact already discussed in He and Shao (1996); i.e. under stronger design conditions the $\log m$-factor may be removed (see e.g. Bai and Wu, 1994), under more general design conditions it cannot.

Anticipating insights garnered from our proofs, we note that the condition (R1) constrains the growth rate of the logarithm of the expected value of the Lipschitz constant of the quantile regression objective function (i.e. $m \log\left(ed\bar{\mathrm{E}}[\|X_{ni,M}\|_2]/m\right) \asymp \log m\left(ed/m^{1/2}\right)$). The stronger rate conditions in Chao et al. (2017) (i.e. $m^3 \xi_m^2 (\log n)^3 = o(n)$ with $\|X_{ni,M}\|_2 \leq \xi_m = O(n^b)$ almost surely), are the result of a weaker control of the fluctuations of the

empirical process indexed by the quantile loss function.

2.2.3 Uniform-in-model Bahadur representations for the quantile regression process

In this section we present our main theoretical results. We begin with a uniform-in-model consistency result and then state two results regarding uniform-in-model Bahadur representations under different rate assumptions (R1) and (R2).

Theorem 2.1 (Uniform-in-Model Strong Consistency of the Quantile Regression Process). *Suppose that Assumptions (T1) – (T6). Let $\mathscr{T}$ be a compact subset of $(0,1)$. Then, there exist absolute constants $c_0, N_0 > 0$ such that for all $n > N_0$, all $\tau \in \mathscr{T}$ and all $M \subseteq \{1,\dots,d\}$,*

$$\|\hat{\theta}_{n,M}(\tau) - \theta^*_{n,M}(\tau)\|_2 \leq c_0 \frac{\bar{\varphi}^{1/2}_{\max}(|M|)}{\bar{\varphi}_{\min}(|M|)} \left(\frac{|M| \log\left(ed/|M|^{1/2}\right) + \log\log n}{n} \right)^{1/2} \quad a.s.$$

We emphasize that the statement holds without any rate assumption and uniformly for the entire collection $\left\{\hat{\theta}_{n,M}(\cdot),\, M \subseteq \{1,\dots,d\}\right\}$ of quantile regression processes. To the best of our knowledge this is first uniform-in-model consistency result for (potentially) misspecified quantile regression processes. Kuchibhotla et al. (2018) recently proposed uniform-in-model consistency results for the least square estimator and argued that their results extend to general M-estimators with twice continuously differentiable objective functions. Their theory does not seem to cover the quantile regression process based on the non-differentiable quantile loss.

Remark 2.1. *It is instructive to consider the special case of a single model M of dimension $|M| = m = m_n \to \infty$ as $n \to \infty$. Under the mild assumptions that $m \log m = o(n)$ and $\bar{\varphi}^{1/2}_{\max}(m)\bar{\varphi}^{-1}_{\min}(m) = O(1)$, Theorem 2.1 simplifies to*

$$\sup_{\tau \in \mathscr{T}} \|\hat{\theta}_{n,M}(\tau) - \theta^*_{n,M}(\tau)\|_2 = O\left(\left(\frac{m + \log\log n}{n} \right)^{1/2} \right) \quad a.s. \tag{2.6}$$

Thus, up to the $\log\log n$-factor this matches the rate of convergence for a single fixed quantile vector obtained in Theorem 2.1 He and Shao (1996). The proof of our theorem shows that the $\log\log n$-factor can be removed if one is satisfied with bounding the remainder term only in probability.

Remark 2.2. *Under the scenario in Remark 2.1, we improve the consistency result for*

quantile series estimators in Belloni et al. (2017) (Theorem 1). To obtain a comparable rate of consistency (up to the $\log\log n$*-factor) Belloni et al. (2017) assume* $m\xi_m^2(\log n)^2 = o(n)$*, where, translating their notation into ours,* $\xi_m = \sup_{\omega\in\Omega}\|X_M(\omega)\|_2$*. Thus, they need stronger moment assumptions (i.e. boundedness of the predictors) and more stringent growth conditions on the model size m.*

Using this uniform-in-model consistency result of Theorem 2.1 we obtain the following strong uniform-in-model Bahadur representation for the misspecified quantile regression process in increasing dimensions.

Theorem 2.2 (Uniform-in-Model Strong Bahadur Representation for the Quantile Regression Process). *Suppose that Assumptions (T1) – (T6) and (R1) hold. Let* $\mathscr{T}$ *be a compact subset of* $(0,1)$*. Then, there exist universal constants* $c_0, N_0 > 0$ *such that for all* $n > N_0$*, all* $\tau \in \mathscr{T}$ *and all* $M \subseteq \{1,\dots,d\}$*,*

$$\hat{\theta}_{n,M}(\tau) - \theta^*_{n,M}(\tau) = D^{-1}_{n,M}(\tau)\mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X_{ni,M}] + r_{n,M}(\tau),$$

and

$$\|r_{n,M}(\tau)\|_2 \leq c_0\bar{\kappa}^2(|M|)\left(\frac{|M|\log\left(ed/|M|^{1/2}\right) + \log\log n}{n}\right)^{3/4} \quad a.s.$$

In contrast to the consistency result of Theorem 2.1 we now need to impose the mild rate condition (R1). The bound on the remainder term has the exponent $3/4$ which is known to be optimal for the quantile regression (e.g. Kiefer, 1967; Koenker, 2005). However, the log-factors appearing in the bound may not be optimal and under different conditions on the design matrix those factors may be removed. We refer to the discussion of Assumptions (R1) and (R2) in Section 2.2.1.

We are not aware of comparable results on uniform-in-model Bahadur representations for the quantile regression process. Again, the results from Kuchibhotla et al. (2018) do not apply here since they do not hold for M-estimators based on non-differentiable objective functions and processes indexed by a continuous variable such as the quantile level $\tau \in \mathscr{T}$.

Remark 2.3. *The strength of our representation is best illustrated in comparison to the Bahadur representation of Theorem 5.1 in Chao et al. (2017). We consider again the scenario of Remark 2.1 with one fixed model M of dimension* $|M| = m$*, continuous conditional density, and upper bounds on the extreme eigenvalues* $\bar{\varphi}_{\max}(m)$ *and* $\bar{\varphi}_{\min}(m)$*. To establish their Bahadur representation Chao et al. (2017) impose the rate condition* $m\xi_m^2\log n = o(n)$*,*

where $\xi_m = \sup_{\omega\in\Omega} \|X_M(\omega)\|_2$. *Under this condition they derive exponential tail bounds on the remainder term* $r_{n,M}(\tau)$ *of the Bahadur representation of the form*

$$\mathrm{P}\left(\sup_{\tau\in\mathscr{T}} \|r_{n,M}(\tau)\|_2 \le \mathfrak{R}_{n,M}(t_n)\right) \ge 1 - 2e^{-t_n},$$

where $t_n \ll n/\xi_m^2$ *and* $\mathfrak{R}_{n,M}(t_n)$ *is a complicated expression depending on, among other things, the sample size n, the model size m, the Lipschitz constant* ξ_m, *and* t_n. *The leading term of* $\mathfrak{R}_{n,M}(t_n)$ *is of order* $O\left(c_n\left(\frac{m\log n}{n}\right)^{1/2} \vee c_n\left(\frac{t_n}{n}\right)^{1/2} \vee \frac{m\xi_m}{n}\right)$, *for a sequence* $c_n = o(1)$ *whose exact rate depends on the design matrix. Thus, without further assumptions Chao et al. (2017) do not achieve the optimal bound on the remainder term with exponent* $3/4$. *Moreover, the condition on* t_n *would require severe restrictions on the growth rate of the model size m if one wanted to use their exponential tail bound to derive an almost sure bound. Finally, their representation holds only for one model M, while ours applies to the entire collection of models* $M \subseteq \{1,\dots,d\}$.

The bound on the error rate in Theorem 2.2 is unsatisfactory if we consider only one model M of growing dimension $|M| = m$ in the sense that the bound contains a $\log m$-factor. The next result shows that this factor can be dispensed with if one imposes a more restrictive rate condition on the design matrix.

Theorem 2.3 (Uniform-in-Model Strong Bahadur Representation and Consistency under Restricted Nonlinearity Impact). *Suppose that Assumptions (T1) – (T6) and (R2) hold. Let* $\mathscr{T}$ *be a compact subset of* $(0,1)$. *Then, there exist universal constants* $c_0, N_0 > 0$ *such that for all* $n > N_0$, *all* $\tau \in \mathscr{T}$ *and all* $M \subseteq \{1,\dots,d\}$,

$$\hat\theta_{n,M}(\tau) - \theta^*_{n,M}(\tau) = D^{-1}_{n,M}(\tau)\mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X_{ni,M}] + r_{n,M}(\tau),$$

and

$$\|r_{n,M}(\tau)\|_2 \le c_0\bar\kappa^2(|M|)\left(\frac{|M|\log(ed/|M|) + \log\log n}{n}\right)^{3/4} \quad a.s.$$

and also

$$\|\hat\theta_{n,M}(\tau) - \theta^*_{n,M}(\tau)\|_2 \le c_0\frac{\bar\varphi^{1/2}_{\max}(|M|)}{\bar\varphi_{\min}(|M|)}\left(\frac{|M|\log(ed/|M|) + \log\log n}{n}\right)^{1/2} \quad a.s.$$

We conclude this section with an easy corollary derived from Theorem 2.2 regarding the weak convergence of the potentially misspecified quantile regression process for a single

model M with growing dimension $|M| = m$. We can safely omit its proof since the technical part, stochastic equicontinuity of the centered quantile regression score as a function of τ and θ, is part of our proof of the strong Bahadur representation. Moreover, total boundedness of the class of score functions in $L^2(\mathrm{P})$ follows from the moment condition (T4) and weak convergence of the marginals by the multivariate central limit theorem.

Corollary 2.1 (Weak Convergence of the Quantile Regression Process). *Suppose that Assumptions (T1) – (T6) and (R1) hold for model M. In addition, suppose that $F_{ni} = F$ for all $i, n \in \mathbb{N}$ and that $|M|^3(\log|M|)^3 = o(n)$. Let $\mathscr{T}$ be a compact subset of $(0,1)$ and $u_M \in \mathscr{S}^{|M|}$. Then,*

$$n^{1/2}\big(u_M'\hat{\theta}_{n,M}(\cdot) - u_M'\theta_M^*(\cdot)\big) \rightsquigarrow \mathbb{G}(\cdot) \quad \textit{in} \quad \ell^\infty(\mathscr{T}),$$

where $\mathbb{G}(\cdot)$ is a centered Gaussian process with covariance function

$$\Sigma(\tau_1, \tau_2; u_M) = \big(\tau_1 \wedge \tau_2 - \tau_1\tau_2\big)u_M'D_M^{-1}(\tau_1)\mathrm{E}[X_M X_M']D_M^{-1}(\tau_2)u_M.$$

A point-wise weak convergence result for the quantile regression vector in increasing dimensions follows from Corollary 2.1 in He and Shao (2000); the first weak convergence result for the entire quantile regression process in increasing dimension was established only recently by Chao et al. (2017) (Theorem 2.1). In a similar vein Belloni et al. (2017) showed that the quantile regression process in increasing dimensions can be strongly approximated by a sequence of Gaussian processes. Our weak convergence result in Corollary 2.1 is a relevant contribution to this body of theory as it significantly relaxes the constraints on the growth rate of the parameter space compared to Chao et al. (2017) and Belloni et al. (2017).

Remark 2.4. *We leave the question how to formulate the weak convergence of the collection $\{\hat{\theta}_{n,M}(\cdot),\ M \subseteq \{1,\dots,d\}\}$ of quantile regression processes when the number of models increases with the sample size for future work.*

2.3 Upper bounds on collections of suprema of empirical processes

In this section we explain the three major technical challenges in the proof of the uniform-in-model Bahadur representation for misspecified quantile regression processes and motivate our approach to solving them. The main results in this section are a new exponential inequality to control the fluctuation of a collection of suprema of empirical processes over a

class of unbounded functions and a new local maximal inequality for the empirical process indexed by the quantile regression score function which adapts to the variance of score function.

2.3.1 Three opportunities for improvements

The key step in proving Theorem 2.1 and Theorems 2.2 and 2.3 is to show almost sure asymptotic equicontinuity of the quantile regression loss function ρ_τ and the score function φ_τ uniformly over all possible models $M \subseteq \{1,\dots,d\}$. More precisely, for any decreasing sequence $r_n(|M|) \downarrow 0$ (which may depend on the model size $|M|$) we need to derive good (i.e. almost sure) error rates on the following suprema of empirical processes:

$$\begin{aligned}
&\max_{M\subseteq\{1,\dots,d\}} \sup_{\tau\in\mathscr{T}} \sup_{\|\theta_M-\theta^*_{n,M}(\tau)\|_2\le r_n(|M|)} \mathbb{G}_{nn}\Big(\rho_\tau\big(Y_{ni}-X'_{ni,M}\theta_M\big)-\rho_\tau\big(Y_{ni}-X'_{ni,M}\theta^*_{n,M}(\tau)\big)\Big),\\
&\max_{M\subseteq\{1,\dots,d\}} \sup_{\tau\in\mathscr{T}} \sup_{\|\theta_M-\theta^*_{n,M}(\tau)\|_2\le r_n(|M|)} \mathbb{G}_{nn}\Big(\varphi_\tau\big(Y_{ni}-X'_{ni,M}\theta_M\big)-\varphi_\tau\big(Y_{ni}-X'_{ni,M}\theta^*_{n,M}(\tau)\big)\Big).
\end{aligned} \tag{2.7}$$

With Talagrand's (1996b) (Theorem 1.4) concentration inequality for empirical processes, which relates the control of the supremum of an empirical process over a class of function $\mathscr{F}$ to the expectation of this supremum, the problem of finding good bounds on the suprema in eq. (2.7) reduces to the simpler one of finding bounds on their expected values. This can be done under uniform entropy conditions which measure the complexity of a class $\mathscr{F}$ by providing uniform (in probability measure Q) upper bounds on the number $N\big(\varepsilon,\mathscr{F},L^r(Q)\big)$ of $L^r(Q)$-balls of radius ε that are necessary to cover $\mathscr{F}$. A widely-used inequality in this context is given in Theorem 3.1 in Giné and Koltchinskii (2006). Informally, we can summarize the content of this theorem as follows: Suppose that $\mathscr{F}$ admits a bounded envelope function $F \le U$ and that $\log N\big(\varepsilon,\mathscr{F},L^r(Q)\big) \le H(\|F\|_{L^2(Q)}/\varepsilon)$ for some non-decreasing function H satisfying some mild regularity conditions. Then, given a collection $X_1,\dots,X_n$ of i.i.d random variables with law P,

$$\mathrm{E}\Big[\sup_{f\in\mathscr{F}}\Big|\frac{1}{\sqrt{n}}\mathbb{G}_n\big(f(X_i)\big)\Big|\Big] \le C\left[\frac{\sigma}{\sqrt{n}}\sqrt{H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)\bigvee\frac{U}{n}H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)}\right],$$

where $C>0$ is a positive number depending on H, and σ satisfies

$$\sup_{f\in\mathscr{F}}\mathrm{E}[f^2] \le \sigma^2 \le \|F\|^2_{L^2(P)}.$$

A naive application of this theorem to bound the expected values of the suprema in eq. (2.7) leads to a three-fold unsatisfactory result: First, the bound depends linearly on the largest model size d; it does not yield an adaptive bound of order $|M|\log(ed/|M|)$. Second, the theorem only applies to uniformly bounded function classes; while we want to work under much weaker moment conditions to allow for unbounded predictors. Third, in the case of the second supremum in eq. (2.7) the common choice for function H (i.e. Theorem 2.6.7 in van der Vaart and Wellner (1996) combined with Theorem 3 in Andrews (1994)) results in an upper bound with exponent $1/2$ not $3/4$ and thus fails to properly adapt to the variance of the empirical process. These three failures explain the worse rates and more stringent assumptions in Belloni et al. (2017) and Chao et al. (2017).

We resolve these issues as follows: First, to address the problem of obtaining an adaptive bound in the model size $|M|$ we split the function class $\mathscr{F}$ into slices of functions with roughly equal variances and equal metric entropy. We then separately compare the empirical processes on each slice with their expectations using an appropriate concentration inequality. When putting everything back together we weight the slices according to their relative importance in relation to the entire function class. To do so, we normalize the empirical process on each slice and consider self-normalized empirical processes. This new approach is related to but different from techniques developed for ratio-type empirical processes. In fact, it is straightforward to stratify the function class such that there exists a ratio-type process which dominates our self-normalized empirical processes. However, the techniques developed for ratio-type empirical processes yield tail bounds which depend on all strata; while our approach yields tail bounds which are independent of any stratum. Thus, our approach provides a tighter control of the tail probabilities. Our proof technique also differs from the peeling device used in the context of generic chaining. Chaining bounds depend on different scales and therefore yield deviation and concentration inequalities involving non-adaptive multi-scale terms. We refer the reader to Pollard (1995), Giné and Koltchinskii (2006) and Koltchinskii (2011) for a detailed study of ratio-type empirical processes and applications to statistical learning theory and nonparametric statistics. For a comprehensive introduction to majorizing measures and generic chaining we refer to Talagrand (1996a).

Second, to solve the problem of working with unbounded predictors, we resort to Panchenko's (2003) concentration inequality for self-normalized empirical processes indexed by an unbounded function class. In order to turn the exponential tail bounds on our self-normalized empirical processes into almost sure bounds we derive a maximal inequality for self-normalized empirical processes and a corresponding bounded law of iterated logarithm. This requires proving a version of Ottaviani's maximal inequality for self-normalized empirical processes; an auxiliary result that may be of independent interest. The proof of the

bounded law of iterated logarithm combines a traditional truncation argument with Marton's Coupling Inequality (Massart, 2007, Proposition 2.21). This allows us to consider the cases of function classes whose complexity grows with the sample size. A common theme in our proofs are symmetrization arguments via Rademacher random variables. We thus provide a more elementary approach to concentration and maximal inequalities for self-normalized empirical processes than Bercu et al. (2002) who invoke Herbst's argument to a suitably modified logarithmic Sobolev inequality.

Third, we obtain the optimal exponent of $3/4$ on the expected value of the supremum over the score function by revisiting the proofs of Theorem 3.1 in Giné and Koltchinskii (2006) and Theorem 2.1 in van der Vaart and Wellner (2011) and exploiting the special structure of the quantile regression score function. More precisely, we extensively use the fact that the score function is the the product of a linear and a bounded function which is differentiable in quadratic mean. Thus, our result holds (in principle) for a rich class of robust estimators. However, we do not strive for the most general result and leave this for future research.

2.3.2 General exponential inequalities for collections of suprema of empirical processes

The main theoretical contribution in this section is Lemma 2.3 which provides an almost sure upper bound and an exponential inequality for collections of suprema of empirical processes. However, the ost used results throughout the proofs of the Bahadur representation are the more specific Theorem 2.4 and Corollary 2.2.

We start with the following preliminary lemma which provides a bounded law of iterated logarithm for the supremum of an empirical process with explicit constants for a function class that may change with the sample size n.

Lemma 2.1. *Let $\{X_{ni}, i \leq n\}$ be a triangular array of row-wise independent random vectors on a measurable space $\mathscr{X}$ and let $\mathscr{F}_n = \{f : \mathscr{X} \to \mathbb{R}\}$ be classes of measurable functions with countable separants. In addition, suppose that for each $\mathscr{F}_n$ there exists an envelope function $F_n : \mathscr{X} \to (0, \infty]$ such that $\mathrm{E}[\max_{i \leq n} F_n^2(X_{ni})] < \infty$ for all $n \in \mathbb{N}$. Then, there exist absolute constants $N_0 > 0$ such that for all $n > N_0$,*

$$\sup_{f \in \mathscr{F}_n} |\mathbb{G}_{nn}(f)| \leq 4\mathrm{E}\Big[\sup_{f \in \mathscr{F}_n} |\mathbb{G}^{\circ}_{nn}(f)|\Big] + 12\sqrt{2\log\log n}\,\bar{\mathrm{E}}_{nn}[F_n^2]^{1/2} \quad a.s.$$

and

$$\sup_{f\in\mathscr{F}_n}|\mathbb{G}^{\circ}_{nn}(f)| \leq \mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}^{\circ}_{nn}(f)|\Big] + 3\sqrt{2\log\log n}\,\bar{\mathrm{E}}_{nn}[F_n^2]^{1/2} \quad a.s.$$

This bounded law of iterated logarithm differs from the one provided in Theorem 10 in Ledoux and Talagrand (1989) in several ways. Most importantly, our result holds under weak finite second moment conditions on the envelope function, while Ledoux and Talagrand (1989) impose a Kolmogorov-type almost sure boundedness condition on the functions $f \in \mathscr{F}_n$. The variance factor that we obtain may not be the best possible, but unless one imposes more stringent assumptions it does not seem to be improvable.

Remark 2.5. *As the proof of Theorem 10 in Ledoux and Talagrand (1989) our proof is based on a randomization argument. However, unlike them we do not use randomization via uniformly on $[-1,1]$ distributed random variables, but Rademacher variables distributed on $\{-1,1\}$. This allows us to combine a truncation argument with Marton's Coupling Inequality (e.g. Massart, 2007, Proposition 2.21). Using Marton's Coupling Inequality instead of directly exploiting the Sub-Gaussianity of the conditional Rademacher average helps us to avoid regularity conditions related to Dudley's entropy integral (such as uniform entropy conditions) and therefor yields explicit constants. However, we do not claim that those constants are optimal.*

The next lemma is a maximal version of Panchenko's (2003) concentration inequality for self-normalized empirical processes. The result appears to be new.

Lemma 2.2. *Let $\{X_{ni}, i \leq n\}$ be a triangular array of row-wise independent random vectors on a measurable space $\mathscr{X}$ and let $\mathscr{F}_n = \{f : \mathscr{X} \to \mathbb{R}\}$ be classes of measurable functions with countable separants. Let $\{\tilde{X}_{ni}, i \leq n\}$ be an independent copy of $\{X_{ni}, i \leq n\}$. Denote the mixed uniform variance of function class $\mathscr{F}_n$ by*

$$\mathbb{V}_{nn}(\mathscr{F}_n) = \mathrm{E}\left[\sup_{f\in\mathscr{F}_n}\frac{1}{n}\sum_{i=1}^{n}\Big(f(X_{ni}) - f(\widetilde{X}_{ni})\Big)^2 \mid (X_{n1},\ldots,X_{nn})\right],$$

Then, for $t > 0$,

$$\mathrm{P}\Bigg(\max_{1\leq k\leq n}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)| \geq 7\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big] + 14t^{1/2}\Big(\mathbb{V}_{nn}(\mathscr{F}_n) + \mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]\Big)^{1/2}\Bigg) \leq 12ee^{-t/2}.$$

Remark 2.6. *Note that the variance term on the right of the inequality sign in the probability operator,* $\left(\mathbb{V}_{nn}(\mathscr{F}_n) + \mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]\right)^{1/2}$*, is a random quantity. This sets this inequality apart from its classical counterparts with a deterministic control of the variance such as the Ottaviani's and Lévi's maximal inequalities (e.g. van der Vaart and Wellner, 1996, Propositions A.1.1 and A.1.2).*

We are now ready to give the main theoretical result of this section. This lemma develops a general framework for slicing a class of functions $\mathscr{F}$ into smaller sub-classes of equal variance or equal metric entropy by introducing the notion of equivalence classes and (sub)probability measures defined on these equivalence classes. While this idea seems very natural to us, we are not aware of a comparable approach of the same level of generality in the literature on empirical processes.

Lemma 2.3. *Let* $\{X_{ni}, i \leq n\}$ *be a triangular array of row-wise independent random vectors on a measurable space* $\mathscr{X}$ *and let* $\mathscr{F}_n = \{f : \mathscr{X} \to \mathbb{R}\}$ *be classes of measurable functions with countable separants. Let* $\mathscr{R}_n \subseteq \mathscr{F}_n \times \mathscr{F}_n$ *be an equivalence relation on* $\mathscr{F}_n$ *such that the quotient set* $\mathscr{F}_n/\mathscr{R}_n$ *is countable. Let* ν_n *be a (sub)probability measure on* $\mathscr{F}_n/\mathscr{R}_n$*. For* $f \in \mathscr{F}_n$ *denote its corresponding equivalence class by* $[f]_{\mathscr{R}_n} = \{g \in \mathscr{F}_n : (f,g) \in \mathscr{R}_n\}$*. Let* $\delta \in (0,1)$*. Then, with probability at least* $1-\delta$*, for all* $f \in \mathscr{F}_n$*,*

$$|\mathbb{G}_{nn}(f)| \leq \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] + 2^{3/2}\mathbb{V}_{nn}^{1/2}([f]_{\mathscr{R}_n})\left(\log\frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\frac{1}{\delta}\right)^{1/2}.$$

Moreover, there exist $c_0, c_1, N_0 > 0$ *such that for all* $n > N_0$ *and for all* $f \in \mathscr{F}_n$*,*

$$\begin{aligned}|\mathbb{G}_{nn}(f)| \leq\ & c_0 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\ & + c_1\left(\mathbb{V}_{nn}([f]_{\mathscr{R}_n}) + \mathrm{E}[\mathbb{V}_{nn}([f]_{\mathscr{R}_n})]\right)^{1/2}\left(\log\frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} \quad a.s.\end{aligned}$$

Constants $c_0 = 7$ *and* $c_1 = 20$ *work.*

Note that the right hand side of above inequality depends only on the supremum over the equivalence classes $[f]_{\mathscr{R}_n}$ and the logarithm of the associated weights $\nu_n([f]_{\mathscr{R}_n})$. This inequality is therefore useful if it is possible to construct relatively small equivalence classes with small envelope functions, and if it is possible to define sensible probability measure on the quotient space. The equivalence relations that we have in mind, partition the large function class $\mathscr{F}_n$ into smaller sub-classes according to their metric entropy (and thus place models M of the same size $|M| = m$ in the same sub-class). In applications we usually

choose the probability measure ν_n such that it assigns mass to sub-classes in a way that is inverse proportional to the entropy of the sub-class.

Remark 2.7. *Constructing a probability measure on the quotient space is relatively easy. We only need to define a probability measure on the coarse quotient set $\mathscr{F}_n/\mathscr{R}_n$ and its corresponding σ-algebra $\sigma(\mathscr{F}_n/\mathscr{R}_n) = \left\{ \cup_{j\in J} S_j : S_j \in \mathscr{F}_n/\mathscr{R}_n, J \in \{1,\ldots,|\mathscr{F}_n/\mathscr{R}_n|\} \right\}$.*

Remark 2.8. *Compared with the non-asymptotic maximal inequality for a collection of empirical processes given in Belloni and Chernozhukov (2011) (Lemma 19) we note the following: Our non-asymptotic inequality is exponential (with the bound depending on $\log(1/\delta)$) while Belloni and Chernozhukov's (2011) inequality is polynomial (with the bound depending on $\sqrt{1/\delta}$). Thus, it is not possible to derive tight almost sure bounds from their inequality.*

Combining Lemma 2.3 with Lemma 2.1 readily yields the following Theorem which we use throughout the proofs of the strong uniform-in-model Bahadur representations.

Theorem 2.4. *Recall notation and assumptions of Lemma 2.3. Further, suppose that each $S_n \in \mathscr{F}_n/\mathscr{R}_n$ has an envelope function $F_{S_n} : \mathscr{X} \to (0,\infty]$ such that $\mathrm{E}[\max_{i\le n} F_{S_n}^4(X_{ni})] < \infty$. Then, there exist $c_0, c_1, c_2, N_0 > 0$ such that for all $n > N_0$ and for all $f \in \mathscr{F}_n$,*

$$
\begin{aligned}
|\mathbb{G}_{nn}(f)| \le\ & c_0 \mathrm{E}\Big[\sup_{f\in[f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)| \Big] \\
& + c_1 \left(\mathrm{E}\Big[\sup_{f\in[f]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}(f^2)| \Big] \right)^{1/2} \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n \right)^{1/2} \\
& + c_2 \Big(\sup_{f\in[f]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2] \Big)^{1/2} \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n \right)^{1/2} \\
& + c_3 \left(\bar{\mathrm{E}}_{nn}[F_{[f]_{\mathscr{R}_n}}^4] \right)^{1/4} \left(\frac{\log\log n}{n} \right)^{1/4} \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n \right)^{1/2} \quad a.s.
\end{aligned}
$$

Constants $c_0 = 7$, $c_1 = 69$, $c_2 = 149$, and $c_3 = 223$ work.

This result shows that the empirical process $\mathbb{G}_{nn}(f)$ can be bounded uniformly over all $f \in \mathscr{F}_n$ and almost surely in terms of the second moment of f, the fourth moment of envelop functions $F_{[f]_{\mathscr{R}_n}}$ and the mean of $\sup_{f\in[f]_{\mathscr{R}_n}} \mathbb{G}_{nn}(f)$ and $\sup_{f\in[f]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}} \mathbb{G}_{nn}(f^2)$. All suprema are taken over equivalent classes $[f]_{\mathscr{R}_n}$ only. Thus, in order to leverage this fact, the next step is to find equivalence classes that generate partitions on which the expected values of the suprema over $[f]_{\mathscr{R}_n}$ are much smaller than over $\mathscr{F}_n$, and on which the four terms in on the right hand side of the inequality are of comparable orders.

A different way of looking at Theorem 2.4 is by comparing it with Lemma 2.1: We see that we have essentially traded off higher moment conditions against smaller classes over which to compute (expected values of) suprema.

We conclude this section with a straightforward but useful corollary.

Corollary 2.2. *Let $\{X_{ni}, i \leq n\}$ be a triangular array of row-wise independent random vectors on a measurable space $\mathscr{X}$. Let $\mathscr{F}_n = \mathscr{G}_n \circ \mathscr{H}_n = \{g \circ h : g \in \mathscr{G}_n, h \in \mathscr{H}_n\}$, where $\mathscr{G}_n = \{g : \mathscr{D} \to \mathbb{R}\}$ and $\mathscr{H} = \{h : \mathscr{X} \to \mathscr{D}\}$. Let $\mathscr{G}_n$ be a class of measurable functions with countable separant with an envelope function $G_n : \mathscr{D} \to (0, \infty]$ such that $\mathrm{E}[\max_{i \leq n} G_n^4(X_{ni})] < \infty$. Let $\mathscr{R}_{\mathscr{H}} \subseteq \mathscr{F}_n \times \mathscr{F}_n$ be an equivalence relation on $\mathscr{H}_n$ such that the quotient set $\mathscr{H}_n / \mathscr{R}_{\mathscr{H}}$ is countable. Let ν_n be a (sub)probability measure on $\mathscr{H}_n / \mathscr{R}_{\mathscr{H}}$.*

Then, there exist $c_0, c_1, c_2, N_0 > 0$ such that for all $n > N_0$ and for all $f = g \circ h \in \mathscr{F}_n$,

$$
\begin{aligned}
|\mathbb{G}_{nn}(f)| \leq\ & c_0 \mathrm{E}\Big[\sup_{g \in \mathscr{G}_n} |\mathbb{G}_{nn}(g \circ h)| \Big] \\
& + c_1 \Big(\mathrm{E}\Big[\sup_{g \in \mathscr{G}_n} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}((g \circ h)^2)| \Big] \Big)^{1/2} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \\
& + c_2 \Big(\sup_{g \in \mathscr{G}_n} \bar{\mathrm{E}}_{nn}[(g \circ h)^2] \Big)^{1/2} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \\
& + c_3 \Big(\bar{\mathrm{E}}_{nn}[(G_n \circ h)^4] \Big)^{1/4} \left(\frac{\log\log n}{n} \right)^{1/4} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \quad a.s.
\end{aligned}
$$

Constants $c_0 = 7$, $c_1 = 69$, $c_2 = 149$, and $c_3 = 223$ work.

Remark 2.9. *This corollary allows the following important refinement in the case of convex functions parametrized by a possibly uncountable index set: Suppose that $\mathscr{G}_n = \{g_\theta : g_\theta \text{ is convex}, \theta \in \Theta_n\}$ and that there exists a countable set T_n such that $\Theta_n \subseteq \mathrm{conv}(T_n)$. Then,*

$$
\begin{aligned}
|\mathbb{G}_{nn}(f)| \leq\ & c_0 \mathrm{E}\Big[\sup_{\theta \in \Theta_n} |\mathbb{G}_{nn}(g_\theta \circ h)| \Big] \\
& + c_1 \Big(\mathrm{E}\Big[\sup_{t \in T_n} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}((g_t \circ h)^2)| \Big] \Big)^{1/2} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \\
& + c_2 \Big(\sup_{t \in T_n} \bar{\mathrm{E}}_{nn}[(g_t \circ h)^2] \Big)^{1/2} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \\
& + c_3 \Big(\bar{\mathrm{E}}_{nn}[(G_n \circ h)^4] \Big)^{1/4} \left(\frac{\log\log n}{n} \right)^{1/4} \left(\log \frac{1}{\nu_n([h]_{\mathscr{R}_{\mathscr{H}}})} + \log\log n \right)^{1/2} \quad a.s.
\end{aligned}
$$

2.3.3 Quantile regression specific maximal inequalities

In this section we present specific results on the quantile regression loss and score functions. The main theoretical result is Lemma 2.5 which provides the optimal exponent $3/4$ for the upper bound on the expected value of the supremum over the quantile score function. This result is central to deriving the optimal rate on the remainder term of the Bahadur representation because it is used to bound the two leading terms of the almost sure bound in Theorem 2.4. It can be easily generalized to other function classes; see comments below. In Lemata 2.7–2.9 we establish uniform-in-model almost sure asymptotic equicontinuity results for function classes specific to quantile regression.

We begin with a lemma to bound the moments of linear functions.

Lemma 2.4. *Let X be an $\mathbb{R}^d$-valued random vector and denote the largest eigenvalues of $\mathrm{E}[XX']$ by $\varphi_{\max}(d)$. Suppose that there exists $p \geq 1$ and an absolute constant $\mu_{2p} < \infty$ such that $\max_{1\leq k\leq d} \mathbb{E}\left[(X^{(k)})^{2p}\right] \leq \mu_{2p}$. Let $K = 1/(e-\sqrt{e})$. Then,*

$$\sup_{\|v\|_2=1} \mathrm{E}\left[|X'v|^{2p}\right]^{1/(2p)} \leq 2^{2+1/(2p)} K^{1/2} p^{1/2} \mu_{2p} + \varphi_{\max}^{1/2}(d),$$

and

$$\max_{1\leq j\leq d} \mathrm{E}\left[|X^{(j)}|^{2p}\right]^{1/(2p)} \leq 2^{2+1/(2p)} K^{1/2} p^{1/2} \mu_{2p} + \mu_1.$$

The next lemma provides a bound on the expected value of the supremum of empirical processes indexed by the quantile regression score function. While stated for a very specific class of functions, the result holds more generally for any function that can be decomposed into the product of a linear function and a bounded function which is differentiable in quadratic mean.

Lemma 2.5. *Suppose that Assumptions (T1), (T2), (T4), and (T5) hold for a fixed model of size d. For $r_n > 0$ and $\|v\|_2 = 1$, $v \in \mathbb{R}^d$ define*

$$\mathscr{F}_n = \left\{ (Y,X) \mapsto 1\left\{X'\theta_1 < Y \leq X'\theta_2\right\} X'v : \|\theta_1 - \theta_2\| \leq r_n,\, \theta_1, \theta_2 \in \mathbb{R}^d \right\}$$

Then, there exists an absolute constant $c_5 > 0$ such that

$$\frac{1}{\sqrt{n}} \mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} \mathbb{G}_n(f) \Big] \leq c_5 \bar{\varphi}_{\max}^{3/2}(d) \left(\frac{d}{n}\right)^{3/4} + c_5 f_+^{1/2} \bar{\varphi}_{\max}^{3/2}(d) r_n^{1/2} \left(\frac{d}{n}\right)^{1/2}.$$

Remark 2.10. *By choosing $r_n = (d/n)^{1/2}$ we obtain the desired exponent $3/4$.*

The following technical lemma allows us to break up the supremum of an empirical process into smaller chunks which lend themselves naturally to manageable equivalence classes. We usually use it together with Corollary 2.2. We use the first part to break up the supremum over the quantile regression loss function and the second part to break up the supremum over the score function.

Lemma 2.6. *Let $\{X_{ni}, i \leq n\}$ be a triangular array of row-wise independent random vectors on $\mathbb{R}^d$, $S \subseteq \{1,\dots,d\}$ be a set of indices, $m \in \{1,\dots,d\}$, and $F : \mathbb{R} \to \mathbb{R}$ a real-valued map.*

1) Let $\varepsilon \in (0, \frac{1}{2}]$. There exists a set $\mathscr{M}_m^d \subset \mathscr{B}^d$ such that $|\mathscr{M}_m^d| \leq \left(1+\frac{2}{\varepsilon}\right)^m \left(\frac{ed}{m}\right)^m$, $\|v\|_0 \leq m$ for all $v \in \mathscr{M}_m^d$, and

$$\sup_{\substack{\|u\|_2=1 \\ u\in\mathbb{R}^m}} \sup_{S:|S|=m} \mathbb{G}_{nn}\big(F(X'_{ni,S}u)\big) \leq \max_{v\in\mathscr{M}_m^d} \sup_{u\in\mathscr{B}^d_{\mathrm{supp}(v)}(v,\varepsilon)} \mathbb{G}_{nn}\big(F(X'_{ni}u)\big).$$

2) Suppose that F is linear. There exist a set $\mathscr{M}_m^d \subset \mathscr{B}^d$ and an absolute constant $C > 0$ such that $|\mathscr{M}_m^d| \leq C\left(\frac{5ed}{m}\right)^{4m}$, $\|v\|_0 \leq m$ for all $v \in \mathscr{M}_m^d$, and

$$\sup_{\substack{\|u\|_2=1 \\ u\in\mathbb{R}^m}} \sup_{S:|S|=m} \mathbb{G}_{nn}\big(F(X'_{ni,S}u)\big) \leq \max_{v\in\mathscr{M}_m^d} \mathbb{G}_{nn}\big(F(X'_{ni}v)\big).$$

In the next three lemmata we establish uniform-in-model almost sure asymptotic equicontinuity of the quantile regression score function and the quantile regression loss function, and derive an almost sure bound on the subgradient.

Lemma 2.7. *Suppose that Assumptions (T1), (T2), (T4), and (T5) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$. Define*

$$\mathscr{F}_n = \Big\{(Y,X) \mapsto \varphi_\tau(Y - X'_M\theta_{1,M})X_M - \varphi_\tau(Y - X'_M\theta_{2,M})X_M : \\ \|\theta_{1,M} - \theta_{2,M}\|_2 \leq r_n(|M|),\ \theta_1,\theta_2 \in \mathbb{R}^d,\ \tau \in \mathscr{T},\ M \subseteq \{1,\dots,d\}\Big\}.$$

Then, there exist absolute constants $c_6, N_6 > 0$ such that for all $n > N_6$ and all $f_{\tau,M,\theta_1,\theta_2} \in \mathscr{F}_n$,

$$\frac{1}{\sqrt{n}}\|\mathbb{G}_{nn}(f_{\tau,M,\theta_1,\theta_2})\|_2 \leq c_6\bar{\varphi}^{3/2}_{\max}(|M|)\left(\frac{|M|\log(ed/|M|)+\log\log n}{n}\right)^{3/4} \\ + c_6 f_+^{1/2}\bar{\varphi}^{3/2}_{\max}(|M|)\left(\frac{|M|\log(ed/|M|)+\log\log n}{n}\right)^{1/2} r_n^{1/2}(|M|) \quad a.s.$$

Lemma 2.8. *Suppose that Assumptions (T1), (T2), (T4), and (T5) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$. Define*

$$\mathscr{F}_n = \Big\{ (Y,X) \mapsto \varphi_\tau\big(Y - X_M'\theta^*_{n,M}(\tau)\big)X_M : \tau \in \mathscr{T}, M \subseteq \{1,\dots,d\} \Big\}.$$

Then, there exist absolute constants $c_7, N_7 > 0$ such that for all $n > N_7$, and all $f_{\tau,M} \in \mathscr{F}_n$,

$$\|\mathbb{E}_{nn}[f_{\tau,M}]\|_2 \leq c_7 \bar{\varphi}^{1/2}_{\max}(|M|) \left(\frac{|M| \log(ed/|M|) + \log\log n}{n} \right)^{1/2} a.s.$$

Lemma 2.9. *Suppose that Assumptions (T1), (T2), (T4), and (T5) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$ and $r_n \in (0,1]$. Define*

$$\begin{aligned}\mathscr{F}_n = \Big\{ &(Y,X) \mapsto \rho_\tau(Y - X_M'\theta_M) - \rho_\tau(Y - X_M'\theta^*_{n,M}(\tau)) : \\ &\|\theta_M - \theta^*_{n,M}(\tau)\|_2 \leq r_n(|M|),\ \theta \in \mathbb{R}^d,\ \tau \in \mathscr{T},\ M \subseteq \{1,\dots,d\} \Big\}.\end{aligned}$$

Then, there exist absolute constants $c_8, N_8 > 0$ such that for all $n > N_8$,

$$\frac{1}{\sqrt{n}} \sup_{f \in \mathscr{F}_n} |\mathbb{G}_{nn}(f)| \leq c_8 \bar{\varphi}^{1/2}_{\max}(|M|) \left(\frac{|M| \log\left(ed/|M|^{1/2}\right) + \log\log n}{n} \right)^{1/2} r_n(|M|) \quad a.s.$$

2.4 Discussion

In this paper we derive almost sure uniform-in-model Bahadur representations and uniform-in-model maximal inequalities for potentially misspecified quantile regression processes. Such Bahadur-type representations and inequalities are useful for developing a principled theory for inference in high dimensions (e.g. Berk et al., 2013; Belloni and Chernozhukov, 2011, 2013; Kuchibhotla et al., 2018; Giessing and He, 2018). We anticipate that they will prove particularly useful for developing an asymptotic theory of selective inference for the quantile regression process (e.g. Lee et al., 2016; Tian and Taylor, 2017). To obtain a tight bound on the error rate of the Bahadur representation which adapts well to different models we propose an intuitive approach based on slicing and weighting function classes by borrowing ideas from majorizing measures, ratio-type, and self-normalized empirical processes. These results are applicable beyond quantile regression and may even be easier to apply in high-dimensional problems other than quantile regression processes. For applications of the uniform-in-model Bahadur representation we refer to our companion paper on inference for the high-dimensional quantile regression process under misspecification.

2.5 Proofs

2.5.1 Proofs of Section 2.2

2.5.1.1 Proof of Theorem 2.1

Proof. For $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, and $R_n(|M|) > 0$ define

$$K_{n,M}(\tau) = \{\theta \in \mathbb{R}^{|M|} : \|\theta - \theta^*_{n,M}(\tau)\|_2 = R_n(|M|)\}$$

Suppose that we have shown that there exists an $N_0 > 0$ such that for all $n > N_0$ and for al $\tau \in \mathscr{T}$ the centered check loss evaluated at any point $\theta \in K_{n,M}(\tau)$ is positive. Since the centered check loss is convex and negative, when evaluated at the minimizer $\hat{\theta}_{n,\lambda_n}(\tau)$, we conclude that for all $\tau \in \mathscr{T}$ and $M \subseteq \{1,\dots,d\}$ the minimizer $\hat{\theta}_{n,M}(\tau)$ lies almost surely in $K_{n,M}(\tau)$.

Consider the map $\theta \mapsto \bar{\mathrm{E}}_{nn}\big[\rho_\tau\big(Y_{ni} - X'_{ni,M}\theta\big) - \rho_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)\big]$, where $\theta \in K_{n,M}(\tau)$. This map is convex and by optimality of $\theta^*_{n,M}(\tau)$ a second-order Taylor expansion around $\theta^*_{n,M}(\tau)$ gives the following uniform lower bound

$$\bar{\mathrm{E}}_{nn}\big[\rho_\tau\big(Y_{ni} - X'_{ni,M}\theta\big) - \rho_\tau\big(Y_{ni} - X'_{ni}\theta^*_{n,M}(\tau)\big)\big] \geq \bar{\varphi}_{\min}(|M|)\|\theta - \theta^*_{n,M}(\tau)\|_2^2. \tag{2.8}$$

By eq. (2.8) and Lemma 2.9 there exists an $N_0 > 0$ such that for all $n > N_0$, uniformly in $\tau \in \mathscr{T}$, $\theta \in \partial D_{n,M}(\tau)$, and $M \subseteq \{1,\dots,d\}$,

$$\begin{aligned}
&\mathbb{E}_{nn}\big[\rho_\tau\big(Y_{ni} - X'_{ni,M}\theta\big) - \rho_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)\big] \\
&\quad \geq \bar{\mathrm{E}}_{nn}\big[\rho_\tau\big(Y_{ni} - X'_{ni,M}\theta\big) - \rho_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)\big] \\
&\quad - \left|\frac{1}{\sqrt{n}}\mathbb{G}_{nn}\big(\rho_\tau\big(Y_{ni} - X'_{ni,M}\theta\big) - \rho_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)\big)\right| \\
&\quad \geq \bar{\varphi}_{\min}(|M|)\|\theta - \theta^*_{n,M}(\tau)\|_2^2 \\
&\quad - c_8\left(\frac{|M|\log\left(ed/|M|^{1/2}\right) + \log\log n}{n}\right)^{1/2} \bar{\varphi}^{1/2}_{\max}(|M|)\|\theta - \theta^*_{n,M}(\tau)\|_2 \\
&\quad = R_n(|M|) \\
&\quad \times \left(\bar{\varphi}_{\min}(|M|)R_n(|M|) - c_8\left(\frac{|M|\log\left(ed/|M|^{1/2}\right) + \log\log n}{n}\right)^{1/2} \bar{\varphi}^{1/2}_{\max}(M)\right) \quad a.s.
\end{aligned} \tag{2.9}$$

To bound (2.9) away from 0 define

$$R_n(|M|) = c_9 \frac{\bar{\varphi}_{\max}^{1/2}(|M|)}{\bar{\varphi}_{\min}(|M|)} \left(\frac{|M| \log\left(ed/|M|^{1/2}\right) + \log\log n}{n} \right)^{1/2}, \tag{2.10}$$

where $c_9 > c_8$ is a (large) absolute constant. □

2.5.1.2 Proof of Theorem 2.2

Proof. For $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, and $R_n(|M|) > 0$ define

$$K_{n,M}(\tau) = \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \theta - \theta^*_{n,M}(\tau) - B_{n,M}(\tau) \right\|_2 \leq R_n(|M|) \right\},$$

where

$$B_{n,M}(\tau) = D^{-1}_{n,M}(\tau) \mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big) X_{ni,M}].$$

Suppose that we have shown that there exists an $N_0 > 0$ such that for all $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$ and $n > N_0$ the directional derivative of the centered quantile regression loss function evaluated at any point $\theta \in \partial K_{n,M}(\tau)$ and pointing in the direction of the outward normal vector on $K_{n,M}(\tau)$ in θ is strictly positive. Since the quantile regression loss function is convex, this implies that for all $\tau \in \mathscr{T}$ and $M \subseteq \{1,\dots,d\}$ the minimizer $\hat{\theta}_{n,M}(\tau)$ of the quantile regression loss function is almost surely contained in $K_{n,M}(\tau)$.

Step 1. Lower bound on directional derivative. Denote the outward normal vector on $K_{n,M}(\tau)$ in point θ by $\eta_{n,M}(\theta,\tau)$, i.e.

$$\eta_{n,M}(\theta,\tau) = \big(\theta - \theta^*_{n,M}(\tau) - B_{n,M}(\tau)\big) R_n^{-1}(|M|).$$

Then, uniformly in $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, $\theta \in K_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$\begin{aligned}
&\eta'_{n,M}(\theta,\tau) \bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M}\big) X_{ni,M}X'_{ni,M}]\big(\theta - \theta^*_{n,M}(\tau)\big) \\
&\quad = \eta'_{n,M}(\theta,\tau) \bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_n(\tau)|X_{ni,M}\big) X_{ni,M}X'_{ni,M}]\eta_{n,M}(\theta,\tau) R_n(|M|) \\
&\qquad + \eta'_{n,M}(\theta,\tau) \bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M}\big) X_{ni,M}X'_{ni,M}] B_{n,M}(\tau) \\
&\quad \geq \bar{\varphi}_{\min}(|M|) R_n(|M|) + \mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big) X'_{ni,M}] \eta_{n,M}(\theta,\tau).
\end{aligned} \tag{2.11}$$

By Assumption (T3) and Lemma 2.4 there exists an absolute constant $c_8 > 0$ such that

uniformly in $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, $\theta \in K_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$
\begin{aligned}
&\Big| -\bar{\mathrm{E}}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta\big)X'_{ni,M}]\eta_{n,M}(\theta,\tau) \\
&\quad - \eta'_{n,M}(\theta,\tau)\bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M}\big)X_{ni,M}X'_{ni,M}]\big(\theta - \theta^*_{n,M}(\tau)\big)\Big| \\
&\quad = \left| \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \int_{X'_{ni,M}\theta^*_{n,M}(\tau)}^{X'_{ni,M}\theta} \Big(f_{Y_{ni}|X_{ni,M}}(t|X_{ni,M}) \right.\right. \\
&\qquad\qquad \left.\left. - f_{Y_{ni}|X_{ni,M}}(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M})\Big)X'_{ni,M}dt\right]\eta_{n,M}(\theta,\tau)\right| \\
&\quad \le \frac{f_H}{1+\alpha}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left|(\theta - \theta^*_{n,M}(\tau))'X_{ni,M}\right|^{1+\alpha}\left|X'_{ni,M}\eta_{n,M}(\theta,\tau)\right|\right] \\
&\quad \le \|\theta - \theta^*_{n,M}(\tau)\|_2^{1+\alpha}\frac{f_H}{1+\alpha}\sup_{\|u\|_2=1}\bar{\mathrm{E}}_{nn}[(X'_{ni,M}u)^{2+2\alpha}]^{1/2}\sup_{\|u\|_2=1}\bar{\mathrm{E}}_{nn}[(X'_{ni,M}u)^2]^{1/2} \\
&\quad \le c_8 f_H \bar{\varphi}_{\max}^{1+\alpha/2}(|M|)\|\theta - \theta^*_{n,M}(\tau)\|_2^{1+\alpha}. \qquad (2.12)
\end{aligned}
$$

Since the quantile loss is convex we can lower bound the directional derivative at point θ in direction $\eta_{n,M}(\theta,\tau)$ by the inner product of $\eta_{n,M}(\theta,\tau)$ and the subgradient $-\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X'_{ni,M}\theta)X_{ni,M}]$. By Lemma 2.7 there exists $N_6 > 0$ such that for all $n > N_6$, uniformly in $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, $\theta \in K_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$
\begin{aligned}
&-\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X'_{ni,M}\theta)X'_{ni,M}]\eta_{n,M}(\theta,\tau) \\
&\quad \ge -\mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X'_{ni,M}]\eta_{n,M}(\theta,\tau) \\
&\quad -\bar{\mathrm{E}}_{nn}[\varphi_\tau(Y_{ni} - X'_{ni,M}\theta)X'_{ni,M}]\eta_{n,M}(\theta,\tau) \\
&\quad -\Big\| -\frac{1}{\sqrt{n}}\mathbb{G}_{nn}\big(\varphi_\tau(Y_{ni} - X'_{ni}\theta)X_{ni,M} - \varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X_{ni,M}\big)\Big\|_2 \\
&\quad \ge -\mathbb{E}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X'_{ni,M}]\eta_n(\theta,\tau) \\
&\quad + \eta'_{n,M}(\theta,\tau)\bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M}\big)X_{ni,M}X'_{ni,M}]\big(\theta - \theta^*_{n,M}(\tau)\big) \\
&\quad -\Big\| -\frac{1}{\sqrt{n}}\mathbb{G}_{nn}\big(\varphi_\tau(Y_{ni} - X'_{ni}\theta)X_{ni,M} - \varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X_{ni,M}\big)\Big\|_2 \\
&\quad -\Big\|\bar{\mathrm{E}}_{nn}[\varphi_\tau\big(Y_{ni} - X'_{ni,M}\theta^*_{n,M}(\tau)\big)X'_{ni,M}] \\
&\qquad\qquad -\bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni,M}}\big(X'_{ni,M}\theta^*_{n,M}(\tau)|X_{ni,M}\big)X_{ni,M}X'_{ni,M}]\big(\theta - \theta^*_{n,M}(\tau)\big)\Big\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \bar{\varphi}_{\min}(|M|)R_n(|M|) - c_6\bar{\varphi}^{3/2}_{\max}(|M|)\left(\frac{|M|\log(ed/|M|)+\log\log n}{n}\right)^{3/4} \\
&- c_6 f_+^{1/2}\bar{\varphi}^{3/2}_{\max}(|M|)\left(\frac{|M|\log(ed/|M|)+\log\log n}{n}\right)^{1/2}\|\theta-\theta^*_{n,M}(\tau)\|_2^{1/2} \qquad (2.13)\\
&- c_8 f_H\bar{\varphi}^{1+\alpha/2}_{\max}(|M|)\|\theta-\theta^*_{n,M}(\tau)\|_2^{1+\alpha} \quad a.s.
\end{aligned}
$$

where the last inequality follows from lines (2.11) and (2.12).

Step 2. Choice of $R_n(|M|)$ and existence of suitable constants. Recall that by Theorem 2.1 there exists an $N_7 > 0$ such that for all $n > N_7$, uniformly in $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, $\theta \in D_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$
\|\theta-\theta^*_{n,M}(\tau)\|_2 \leq c_7\frac{\bar{\varphi}^{1/2}_{\max}(|M|)}{\bar{\varphi}_{\min}(|M|)}\left(\frac{|M|\log\left(ed/|M|^{1/2}\right)+\log\log n}{n}\right)^{1/2} \quad a.s. \qquad (2.14)
$$

By above display (2.14) and Assumption (R1) there exists a constant $c_9 > 0$ such that

$$
c_8 f_h\bar{\varphi}^{1+\alpha/2}_{\max}\|\theta-\theta^*_{n,M}(\tau)\|_2^{1+\alpha} \leq c_9\bar{\kappa}^2(|M|)\left(\frac{|M|\log\left(ed/|M|^{1/2}\right)+\log\log n}{n}\right)^{3/4}.
$$

For a sufficiently large constant $c_{10} \geq 1 + c_9 + c_6 + c_6 c_7^{1/2} f_+^{1/2}$, set

$$
R_n(|M|) = c_{10}\bar{\kappa}^2(|M|)\left(\frac{|M|\log\left(ed/|M|^{1/2}\right)+\log\log n}{n}\right)^{3/4}. \qquad (2.15)
$$

Step 3. Completion of the proof. Plug in eq. (2.15) and (2.14) into eq. (2.13) and conclude that there exists $N_8 > 0$ such that for all $n > N_8$ uniformly in $\tau \in \mathscr{T}$, $M \subseteq \{1,\dots,d\}$, $\theta \in D_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$
-\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X'_{ni,M}\theta)X'_{ni,M}]\eta_{n,M}(\theta,\tau) \geq \bar{\varphi}_{\min}(|M|)R_n(|M|) > 0 \quad a.s.
$$

□

2.5.1.3 Proof of Theorem 2.3

Proof. The proof follows from by the same arguments as the proof of Theorem 2.2; however, we need to modify the argument in Step 2. Instead of invoking the consistency result from Theorem 2.1 we proceed as follows:

By Lemma 2.8 there exists an $N_7 > 0$ such that for all $n > N_7$, uniformly in $\tau \in \mathscr{T}$,

$M \subseteq \{1,\dots,d\}$, $\theta \in D_{n,M}(\tau)$, and $\eta_{n,M}(\theta,\tau)$,

$$\begin{aligned}\|\theta - \theta^*_{n,M}(\tau)\|_2 &\leq \left\|\theta - \theta^*_{n,M}(\tau) - B_{n,M}(\tau)\right\|_2 + \|B_{n,M}(\tau)\|_2 \\ &\leq R_n(|M|) + c_7 \frac{\bar{\varphi}^{1/2}_{\max}(|M|)}{\bar{\varphi}_{\min}(|M|)} \left(\frac{|M|\log(ed/|M|) + \log\log n}{n}\right)^{1/2} \quad a.s.,\end{aligned}$$

where the second inequality follows from (R2). Using this upper bound and $R_n(|M|)$ from the statement of the theorem, straightforward but somewhat tedious calculations show that the directional derivative in eq. (2.13) is almost surely strictly bounded away from zero. This concludes the proof. □

2.5.2 Proofs of Section 2.3.2

2.5.2.1 Proof of Lemma 2.1

Proof. **Step 1.** For $f \in \mathscr{F}_n$ define

$$\mathbb{G}_{nN}(f) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big(f(X_{ni}) - \mathrm{E}[f(X_{ni})]\Big).$$

For notational convenience, whenever $N > n$ we interpret the last $N-n+1$ summands in $\mathbb{G}_{nN}$ as 0. We write Lt to denote the function $\max(1,\log t)$ and LLt for the composition $L(Lt)$, $t \geq 0$. For $c_0 \geq 1$ define

$$w_{nNN'}(c_0) = 4\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}^{\circ}_{nN}(f)|\Big] + 4c_0\sqrt{2LLN'}\,\bar{\mathrm{E}}_{nN}[F_n^2]^{1/2}.$$

We make the following two observations: First, for all $n \in \mathbb{N}$,

$$\min_{2^\ell < N \leq 2^{\ell+1}} w_{n,2N,N}(c_0) \geq w_{n,2^{\ell+1},2^\ell}(c_0). \tag{2.16}$$

Indeed, for all $n \in \mathbb{N}$ and $f \in \mathscr{F}_n$, $\big\{\sqrt{N}\mathbb{G}_{nN}(f)\big\}_{N\in\mathbb{N}}$ is a martingale with respect to its natural filtration. Therefore, by Jensen's inequality $\big\{\sup_{f\in\mathscr{F}_n}\sqrt{N}|\mathbb{G}_{nN}(f)|\big\}_{N\in\mathbb{N}}$ is a sub-martingale. Hence, for all $\ell \in \mathbb{N}$ and $n \geq 2^\ell$,

$$\min_{2^\ell < N \leq 2^{\ell+1}} \mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nN}(f)|\Big] \geq \frac{1}{\sqrt{2}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{n2^\ell}(f)|\Big].$$

Moreover, for all $\ell \in \mathbb{N}$ and $n \geq 2^\ell$,

$$\min_{2^\ell < N \leq 2^{\ell+1}} \bar{\mathrm{E}}_{nN}[F_n^2] \geq \frac{1}{2} \bar{\mathrm{E}}_{n2^\ell}[F_n^2].$$

Whence, inequality (2.16) follows.

Second, for all $\eta > 0$ and all $\ell, n \in \mathbb{N}$,

$$\begin{aligned}
&\max_{2^\ell < N \leq 2^{\ell+1}} \mathrm{P}\left(\sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{nN}(f)|}{4\sqrt{2LL2^\ell}\, \bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]^{1/2}} > 1+\eta \right) \\
&\qquad \leq \max_{2^\ell < N \leq 2^{\ell+1}} \frac{\bar{\mathrm{E}}_{nN}[F_n^2]}{(1+\eta)^2 4^2 (2LL2^\ell)\, \bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]} \\
&\qquad \leq \frac{2\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]}{(1+\eta)^2 32 (LL2^\ell)\, \bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]} \\
&\qquad < \frac{1}{2}.
\end{aligned} \tag{2.17}$$

Similarly, for all $\eta > 0$, and all $\ell, n \in \mathbb{N}$,

$$\sup_{f \in \mathscr{F}_n} \mathrm{P}\Big(|\mathbb{G}_{n2^{\ell+1}}(f)| > (1+\eta) w_{n,2^{\ell+1},2^\ell}(1) \Big) < \frac{1}{4}. \tag{2.18}$$

Step 2. For $\eta > 0$ define

$$A_m = \left\{ \max_{n \geq 2^m} \sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{nn}(f)|}{w_{nnn}(3)} > 1+\eta \right\}.$$

The lower bound from eq. (2.16), Ottaviani's Inequality, and the upper bound from eq. (2.17) give

$$\begin{aligned}
\mathrm{P}\big(A_m\big) &\leq \max_{n \in \mathbb{N}} \mathrm{P}\left(\max_{N \geq 2^m} \sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{nN}(f)|}{w_{n,2N,N}(3)} > 1+\eta \right) \\
&\leq \max_{n \in \mathbb{N}} \mathrm{P}\left(\max_{\ell \geq m} \max_{2^\ell < N \leq 2^{\ell+1}} \sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{nN}(f)|}{w_{n,2^{\ell+1},2^\ell}(3)} > 1+\eta \right) \\
&\leq \max_{n \in \mathbb{N}} \left\{ 1 - \max_{2^\ell < N \leq 2^{\ell+1}} \mathrm{P}\left(\sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{nN}(f)|}{4\sqrt{2LL2^\ell}\, \bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]^{1/2}} > 1+\eta \right) \right. \\
&\qquad \left. - \mathrm{P}\left(\sup_{f \in \mathscr{F}_n} \frac{|\mathbb{G}_{n2^{\ell+1}}(f)|}{4\sqrt{2LL2^\ell}\, \bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]^{1/2}} > 1+\eta \right) \right\}^{-1}
\end{aligned}$$

$$\times \mathrm{P}\left(\sup_{f\in\mathscr{F}_n} \frac{|\mathbb{G}_{n2^{\ell+1}}(f)|}{w_{n,2^{\ell+1},2^\ell}(1)} > 1+\eta \right)$$
$$\leq 4 \max_{n\in\mathbb{N}} \sum_{\ell=m}^{\infty} \mathrm{P}\left(\sup_{f\in\mathscr{F}_n} \frac{|\mathbb{G}_{n2^{\ell+1}}(f)|}{w_{n,2^{\ell+1},2^\ell}(1)} > 1+\eta \right).$$

Thus, eq. (2.18) and the Symmetrization Lemma 2.3.7 in van der Vaart and Wellner (1996),

$$\mathrm{P}(A_m) \leq 16 \max_{n\in\mathbb{N}} \sum_{\ell=m}^{\infty} \mathrm{P}\left(\sup_{f\in\mathscr{F}_n} \frac{|\mathbb{G}^\circ_{n2^{\ell+1}}(f)|}{w_{n,2^{\ell+1},2^\ell}(1)} > \frac{1}{4}(1+\eta) \right). \tag{2.19}$$

The tail probability on the right side of eq. (2.19) can be further upper bounded by

$$16 \max_{n\in\mathbb{N}} \sum_{\ell=m}^{\infty} \mathrm{P}\left(\sup_{f\in\mathscr{F}_n} \frac{|\mathbb{G}^\circ_{n2^{\ell+1}}(f)|}{w_{n,2^{\ell+1},2^\ell}(1)} > \frac{1}{4}(1+\eta),\, \mathbb{E}_{n2^{\ell+1}}[F_n^2] \leq 2\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2] \right)$$
$$+ 16 \max_{n\in\mathbb{N}} \sum_{\ell=m}^{\infty} \mathrm{P}\left(\mathbb{E}_{n2^{\ell+1}}[F_n^2] > 2\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2] \right)$$
$$= A + B. \tag{2.20}$$

Step 3. Bound on A. We now derive an exponential inequality for the tail probability of the conditional Rademacher process $\sup_{f\in\mathscr{F}_n} |\mathbb{G}^\circ_{n2^{\ell+1}}(f)| \mid (X_{n1},\ldots,X_{n2^{\ell+1}})$ via Marton's Coupling Inequality. An upper bound for term A follows by integrating out $(X_{n1},\ldots,X_{n2^{\ell+1}})$.

For $a \in \{-1,1\}^{2^{\ell+1}}$ define

$$Z_{n2^{\ell+1}}(a) = \sup_{f\in\mathscr{F}_n} \frac{1}{2^{(\ell+1)/2}} \left| \sum_{i=1}^{2^{\ell+1}} a_i f(X_{ni}) \right| \mid (X_{n1},\ldots,X_{n2^{\ell+1}}).$$

To simplify notation in the following we do not make the conditioning on $(X_{n1},\ldots,X_{n2^{\ell+1}})$ explicit. Clearly, for all $a,b \in \{-1,1\}^{2^{\ell+1}}$,

$$Z_{n2^{\ell+1}}(a) - Z_{n2^{\ell+1}}(b) \leq \sup_{f\in\mathscr{F}_n} \frac{1}{2^{(\ell+1)/2}} \left| \sum_{i=1}^{2^{\ell+1}} (a_i - b_i) f(X_{ni}) \right|$$
$$\leq \sup_{f\in\mathscr{F}_n} \frac{2}{2^{(\ell+1)/2}} \sum_{i=1}^{2^{\ell+1}} |f(X_{ni})| 1\{a_i \neq b_i\}$$
$$\leq \frac{2}{2^{(\ell+1)/2}} \sum_{i=1}^{2^{\ell+1}} F_n(X_{ni}) 1\{a_i \neq b_i\} \tag{2.21}$$

Let B be the probability distribution of the vector ε of independent Rademacher random

variables on $\{-1,1\}^{2^{\ell+1}}$ and let Q be some probability distribution which is absolutely with respect to B. Denote by $\mathscr{P}(\mathrm{B},\mathrm{Q})$ the coupling of B and Q. By eq. 2.21 we have for any $\mathbb{Q} \in \mathscr{P}(\mathrm{B},\mathrm{Q})$,

$$\begin{aligned}
&\mathrm{E}_{\mathrm{Q}}[Z_{n2^{\ell+1}}] - \mathrm{E}_{\mathrm{B}}[Z_{n2^{\ell+1}}] \\
&\quad \leq \frac{2}{2^{(\ell+1)/2}} \int_{\{-1,1\}^{2^{\ell+1}} \times \{-1,1\}^{2^{\ell+1}}} \left(\sum_{i=1}^{2^{\ell+1}} F_n(X_{ni}) 1\{a_i \neq b_i\} \right) d\mathbb{Q}(a,b) \\
&\quad = \frac{2}{2^{(\ell+1)/2}} \sum_{i=1}^{2^{\ell+1}} F_n(X_{ni})\, \mathbb{Q}\big\{(a,b) \in \{-1,1\}^{2^{\ell+1}} \times \{-1,1\}^{2^{\ell+1}}, a_i \neq b_i\big\}.
\end{aligned}$$

Whence, by the Cauchy-Schwarz Inequality followed by Marton's Coupling Inequality (e.g. Massart, 2007, Proposition 2.21),

$$\begin{aligned}
\mathrm{E}_{\mathrm{Q}}[Z_{n2^{\ell+1}}] - \mathrm{E}_{\mathrm{B}}[Z_{n2^{\ell+1}}] &\leq 2^{1/2} \left(\frac{2}{2^{\ell+1}} \sum_{i=1}^{2^{\ell+1}} F_n^2(X_{ni}) \right)^{1/2} \\
&\qquad \times \left(\sum_{i=1}^{2^{\ell+1}} \mathbb{Q}^2\big\{(a,b) \in \{-1,1\}^{2^{\ell+1}} \times \{-1,1\}^{2^{\ell+1}}, a_i \neq b_i\big\} \right)^{1/2} \\
&\leq \sqrt{2v\mathbf{K}(\mathrm{Q},\mathrm{B})},
\end{aligned}$$

where $v_{n2^{\ell+1}} = 2\mathbb{E}_{n2^{\ell+1}}[F_n^2]$ and $\mathbf{K}(\mathrm{Q},\mathrm{B})$ denotes the Kullbach-Leibler divergence between distributions Q and B. Thus, by Lemma 2.13 in Massart (2007), for any $t > 0$,

$$\log \mathrm{E}_{\mathrm{B}}\left[e^{t(Z_{n2^{\ell+1}} - \mathrm{E}_{\mathrm{B}}[Z_{n2^{\ell+1}}])} \right] \leq \frac{v_{n2^{\ell+1}}}{2} t^2.$$

Hence, Chernoff's Inequality implies that

$$\mathrm{P}\left(\sup_{f \in \mathscr{F}_n} |\mathbb{G}^{\circ}_{n2^{\ell+1}}(f)| \geq \mathrm{E}\Big[\sup_{f \in \mathscr{F}_n} |\mathbb{G}^{\circ}_{n2^{\ell+1}}(f)| \Big] + t \mid (X_{n1}, \ldots, X_{n2^{\ell+1}}) \right) \leq e^{-t^2/(2v_{n2^{\ell+1}})}. \tag{2.22}$$

The bound on A follows now by integrating out the tail bound over $(X_{n1}, \ldots, X_{n2^{\ell+1}})$, i.e. by eq. (2.22),

$$A \leq 16 \max_{n \in \mathbb{N}} \sum_{\ell=m}^{\infty} \int_{\left\{ \omega : \mathbb{E}_{n2^{\ell+1}}[F_n^2](\omega) \leq 2\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2] \right\}}$$

$$\exp\left\{-\frac{(1+\eta)^2 4^2\left(2LL2^{\ell}\right)\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]}{16\mathbb{E}_{n2^{\ell+1}}[F_n^2](\omega)}\right\}d\mathrm{P}(\omega)$$

$$\leq 16\sum_{\ell=m}^{\infty} e^{-(1+\eta)^2 4^2(2LL2^{\ell})/32}$$

$$<\infty. \tag{2.23}$$

Step 4. Bound on B. By assumption $\mathrm{E}[\max_{i\leq N} F_n^2(X_{Ni})]<\infty$ for all $n,N\in\mathbb{N}$. Note that without loss of generality we can assume that $\mathrm{E}[\max_{i\leq N} F_n^2(X_{Ni})]\geq 1$; otherwise, we rescale by the rate at which the second moment vanishes. Thus,

$$\begin{aligned}
B &\leq 16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}\mathrm{P}\Big(\mathbb{E}_{n2^{\ell+1}}[F_n^2]>2\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2],\, F_n^2(X_{ni})\leq 2^{\ell+1},\, 1\leq i\leq 2^{\ell+1}\Big)\\
&\quad+16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}\mathrm{P}\Big(\exists i\in\{1,\ldots,2^{\ell+1}\}: F_n^2(X_{ni})>2^{\ell+1}\Big)\\
&\leq 16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}\mathrm{P}\left(\frac{1}{2^{(\ell+1)/2}}\mathbb{G}_{n2^{\ell+1}}[F_n^2 1\{F_n^2\leq 2^{\ell+1}\}]>\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]\right)\\
&\quad+16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}2^{\ell+1}\mathrm{P}\Big(\max_{i\leq n}F_n^2(X_{ni})>2^{\ell+1}\Big)\\
&\leq 16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}\frac{2}{2^{\ell+1}}\frac{\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^4 1\{F_n^2\leq 2^{\ell+1}\}]}{\left(\bar{\mathrm{E}}_{n2^{\ell+1}}[F_n^2]\right)^2}\\
&\quad+16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}2^{\ell+1}\mathrm{P}\Big(\max_{i\leq n}F_n^2(X_{ni})>2^{\ell+1}\Big)\\
&\leq 16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}\frac{2}{2^{\ell+1}}\int_0^{2^{\ell+1}}2t\mathrm{P}\left(\max_{i\leq n}F_n^2(X_{ni})>t\right)dt\\
&\qquad+16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}2^{\ell+1}\mathrm{P}\Big(\max_{i\leq n}F_n^2(X_{ni})>2^{\ell+1}\Big)\\
&=64\max_{n\in\mathbb{N}}\int_0^1\left\{\sum_{\ell=m}^{\infty}2^{\ell+1}t\mathrm{P}\left(\max_{i\leq n}F_n^2(X_{ni})>2^{\ell+1}t\right)\right\}dt\\
&\qquad+16\max_{n\in\mathbb{N}}\sum_{\ell=m}^{\infty}2^{\ell+1}\mathrm{P}\Big(\max_{i\leq n}F_n^2(X_{ni})>2^{\ell+1}\Big)\\
&<\infty,
\end{aligned} \tag{2.24}$$

where in the second last line we use a change of variable and Beppo Levi's monotone convergence theorem guarantees that we can interchange integration and summation.

Step 4. Note that the sequence $(A_m)_{m\in\mathbb{N}}$ is decreasing. Thus, eq. (2.23) and (2.24) and

the continuity of the P-measure imply that

$$\mathrm{P}\left(\limsup_{n\to\infty}\sup_{f\in\mathscr{F}_n}\frac{|\mathbb{G}_{nn}(f)|}{w_{nnn}(3)}>1+\eta\right)=\mathrm{P}\left(\lim_{m\to\infty}A_m\right)=\lim_{m\to\infty}\mathrm{P}(A_m)=0. \tag{2.25}$$

Since $\eta>0$ arbitrary, we conclude that there exists an $N_0>0$ such that for all $n>N_0$,

$$\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\le w_{nnn}(3)\quad a.s. \tag{2.26}$$

This proves the first statement. The second statement follows from the fact that we do not need to apply the Symmetrization Lemma 2.3.7 (van der Vaart and Wellner, 1996) when considering the symmetrized process $\mathbb{G}^{\circ}_{nn}(f)$. □

2.5.2.2 Proof of Lemma 2.2

Proof. Define

$$\mathbb{V}_{nN}(\mathscr{F}_n)=\mathrm{E}\left[\sup_{f\in\mathscr{F}_n}\frac{1}{N}\sum_{i=1}^{N}\left(f(X_{ni})-f(\widetilde{X}_{ni})\right)^2\mid(X_{n1},\ldots,X_{nn})\right],$$
$$\mathbb{W}_{nNN'}(\mathscr{F}_n)=\mathrm{E}\left[\sup_{f\in\mathscr{F}_n}\frac{1}{N'}\sum_{i=N'+1}^{N}\left(f(X_{ni})-f(\widetilde{X}_{ni})\right)^2\mid(X_{n,N'+1},\ldots,X_{nn})\right],$$

and for $f\in\mathscr{F}_n$,

$$\mathbb{G}_{nN}(f)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(f(X_{ni})-\mathrm{E}[f(X_{ni})]\Big).$$

For notational convenience, whenever $N>n$ we interpret the last $N-n+1$ summands in $\mathbb{V}_{nN}(\mathscr{F}_n)$, $\mathbb{W}_{nNN'}(\mathscr{F}_n)$, and $\mathbb{G}_{nN}(f)$ as 0. We note the following: Since $\sqrt{N}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nN}(f)|$ is a sub-martingale with respect to its natural filtration we have, for all $n\ge k$,

$$\sqrt{n}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big]\ge\sqrt{k}\Big[\mathrm{E}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)|\Big]. \tag{2.27}$$

Also, since the summands of $\mathbb{V}_{nn}(\mathscr{F}_n)$ are non-negative we have, for all $n\ge k$,

$$n\mathbb{V}_{nn}(\mathscr{F}_n)\ge k\mathbb{V}_{nk}(\mathscr{F}_n), \tag{2.28}$$

and

$$k\mathbb{W}_{nnk}(\mathscr{F}_n) \geq n\mathbb{V}_{nn}(\mathscr{F}_n) - k\mathbb{V}_{nk}(\mathscr{F}_n). \tag{2.29}$$

To upper bound the tail probability of the maximum over the partial sums by the tail probability of the full sum (up to n), we proceed as in the proof of Ottaviani's inequality (e.g. Ledoux and Talagrand, 1996, Lemma 6.2). The introduction of the random quantity $\mathbb{V}_{nn}(\mathscr{F}_n)$ requires several careful modifications of the classical proof. Once this upper bound is established, the claim follows from Theorem 1 in Panchenko (2003). Let

$$\begin{aligned}\tau = \min\Bigg\{ k \leq n : \sqrt{k}\left(\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nk}(f)| - 2t^{1/2}\mathbb{V}_{nk}^{1/2}(\mathscr{F}_n)\right) \\ > 7\sqrt{n}\left(\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nn}(f)|\Big] + 2t^{1/2}\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]^{1/2}\right)\Bigg\}.\end{aligned}$$

Note that the events $\{\tau = k\}$ depend only on $(X_{n1},\ldots,X_{nk})$ and are disjoint. Define

$$\begin{aligned}A_k = \Bigg\{ \sup_{f\in\mathscr{F}_n} |\sqrt{n}\mathbb{G}_{nn}(f) - \sqrt{k}\mathbb{G}_{nk}(f)| + 2\sqrt{k}t^{1/2}\mathbb{W}_{nnk}^{1/2}(\mathscr{F}_n) \\ \leq 6\sqrt{n}\left(\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nn}(f)|\Big] + 2t^{1/2}\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]^{1/2}\right)\Bigg\}.\end{aligned}$$

The event A_k depends only on $(X_{n,k+1},\ldots,X_{nn})$; thus it is independent of $\{\tau = k\}$. If $\{\tau = k\}$ and A_k occur together, then also

$$B_n = \left\{ \sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nn}(f)| > \mathrm{E}\Big[\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nn}(f)|\Big] + 2t^{1/2}\mathbb{V}_{nn}^{1/2}(\mathscr{F}_n)\right\}.$$

Indeed, by eq. (2.29),

$$\begin{aligned}&\sqrt{n}\left(\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nn}(f)| - 2t^{1/2}\mathbb{V}_{nn}^{1/2}(\mathscr{F}_n)\right) \\ &\quad\geq \sqrt{k}\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nk}(f)| - \sup_{f\in\mathscr{F}_n} |\sqrt{n}\mathbb{G}_{nn}(f) - \sqrt{k}\mathbb{G}_{nk}(f)| - 2\sqrt{n}t^{1/2}\mathbb{V}_{nn}^{1/2}(\mathscr{F}_n) \\ &\quad\geq \sqrt{k}\left(\sup_{f\in\mathscr{F}_n} |\mathbb{G}_{nk}(f)| - 2t^{1/2}\mathbb{V}_{nk}^{1/2}(\mathscr{F}_n)\right) \\ &\quad- \left(\sup_{f\in\mathscr{F}_n} |\sqrt{n}\mathbb{G}_{nn}(f) - \sqrt{k}\mathbb{G}_{nk}(f)| + 2\sqrt{k}t^{1/2}\mathbb{W}_{nnk}^{1/2}(\mathscr{F}_n)\right)\end{aligned}$$

$$> \sqrt{n}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big].$$

Thus, it follows that

$$\mathrm{P}(B_n) = \sum_{k=1}^{n}\mathrm{P}(\tau=k,B_n) \geq \sum_{k=1}^{n}\mathrm{P}(\tau=k,A_k) \geq \min_{j\leq n}\mathrm{P}\left(A_j\right)\sum_{k=1}^{n}\mathrm{P}(\tau=k). \tag{2.30}$$

Note that by independence and eq. (2.28),

$$\begin{aligned}
&\sum_{k=1}^{n}\mathrm{P}(\tau=k)\\
&\quad=\mathrm{P}\left(\max_{n\leq k}\sqrt{k}\left(\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)|-2t^{1/2}\mathbb{V}_{nk}^{1/2}(\mathscr{F}_n)\right)\right.\\
&\qquad\left.>7\sqrt{n}\left(\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big]+2t^{1/2}\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]^{1/2}\right)\right)\\
&\quad\geq\mathrm{P}\left(\max_{n\leq k}\sqrt{k}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)|\geq 7\sqrt{n}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big]\right.\\
&\qquad\left.+\sqrt{n}2t^{1/2}\left(\mathbb{V}_{nn}^{1/2}(\mathscr{F}_n)+7\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]^{1/2}\right)\right)\\
&\quad\geq\mathrm{P}\left(\max_{n\leq k}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)|\geq 7\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big]\right. &(2.31)\\
&\qquad\left.+14t^{1/2}\Big(\mathbb{V}_{nn}(\mathscr{F}_n)+\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]\Big)^{1/2}\right). &(2.32)
\end{aligned}$$

Also, by Markov's inequality and eq. (2.27) and (2.28),

$$\begin{aligned}
\min_{k\leq n}\mathrm{P}(A_k) &= 1-\max_{k\leq n}\mathrm{P}(A_k^c)\\
&\geq 1-2\max_{k\leq n}\mathrm{P}\left(\sqrt{k}\left(\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)|+2t^{1/2}\mathbb{V}_{nk}^{1/2}(\mathscr{F}_n)\right)\right.\\
&\qquad\left.\geq 3\sqrt{n}\left(\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big]+2t^{1/2}\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]^{1/2}\right)\right)\\
&\geq\frac{1}{3}. &(2.33)
\end{aligned}$$

The statement follows by combining eq. (2.30)-(2.33) and Theorem 1 in Panchenko

(2003) (with $\alpha = 1$),

$$\begin{aligned}
&\mathrm{P}\left(\max_{n\le k}\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nk}(f)| \ge 7\mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big] + 14t^{1/2}\Big(\mathbb{V}_{nn}(\mathscr{F}_n)+\mathrm{E}[\mathbb{V}_{nn}(\mathscr{F}_n)]\Big)^{1/2}\right)\\
&\quad\le 3\mathrm{P}\left(\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)| > \mathrm{E}\Big[\sup_{f\in\mathscr{F}_n}|\mathbb{G}_{nn}(f)|\Big] + 2t^{1/2}\mathbb{V}_{nn}^{1/2}(\mathscr{F}_n)\right)\\
&\quad\le 12ee^{-t/2}.
\end{aligned}$$

□

2.5.2.3 Proof of Lemma 2.3

Proof. We only give the proof of the almost sure statement. The finite sample statement follows trivially from Theorem 1 in Panchenko (2003). For $S_n \in \mathscr{F}_n/\mathscr{R}_n$ define

$$\mathbb{V}_{nN}(S_n) = \mathrm{E}\left[\sup_{f\in S_n}\frac{1}{N}\sum_{i=1}^{N}\Big(f(X_{ni}) - f(\widetilde{X}_{ni})\Big)^2 \mid (X_{n1},\ldots,X_{nn})\right],$$

and for $f \in S_n$,

$$\mathbb{G}_{nN}(f) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(f(X_{ni}) - \mathrm{E}[f(X_{ni})]\Big).$$

For notational convenience, whenever $N > n$ we interpret the last $N-n+1$ summands in $\mathbb{V}_{nN}(S_n)$ and $\mathbb{G}_{nN}(f)$ as 0. We write Lt to denote the function $\max(1,\log t)$ and LLt for the composition $L(Lt)$, $t \ge 0$. Let

$$\begin{aligned}
\mathbb{U}_{nNN'}(S_n) = {}& 7\mathrm{E}\Big[\sup_{f\in S_n}|\mathbb{G}_{nN}(f)|\Big]\\
&+ 14\Big(\mathbb{V}_{nN}(S_n)+\mathrm{E}\left[\mathbb{V}_{nN}(S_n)\right]\Big)^{1/2}2^{1/2}\Big(LLN' + \log\frac{1}{\nu_n(S_n)}\Big)^{1/2},
\end{aligned}$$

and for arbitrary $\eta > 0$ and $m \in \mathbb{N}$,

$$A_m = \left\{\max_{n\ge 2^m}\sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n}\sup_{f\in S_n}\frac{|\mathbb{G}_{nn}(f)|}{\mathbb{U}_{nnn}(S_n)} > 1+\eta\right\}.$$

Then, by two applications of the union bound and Lemma 2.2,

$$\begin{aligned}
\mathrm{P}(A_m) &\leq \max_{n\in\mathbb{N}} \mathrm{P}\left(\max_{N\geq 2^m} \sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f)|}{\mathbb{U}_{nNN}(S_n)} > 1+\eta \right) \\
&\leq \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \mathrm{P}\left(\max_{\ell\geq m} \max_{2^\ell < N\leq 2^{\ell+1}} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f)|}{\mathbb{U}_{nNN}(S_n)} > 1+\eta \right) \\
&\leq \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sum_{\ell=m}^{\infty} \mathrm{P}\left(\max_{N\leq 2^{\ell+1}} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f)|}{\mathbb{U}_{nN2^\ell}(S_n)} > 1+\eta \right) \\
&\leq 12e \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sum_{\ell=m}^{\infty} e^{-(1+\eta)^2[-\log \nu_n(S_n)+(\log \ell)+(\log\log 2)]} \\
&\leq 12e \left(\max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \nu_n(S_n)^{(1+\eta)^2} \right) \sum_{\ell=m}^{\infty} e^{-(1+\eta)^2[\log \ell+\log\log 2)]} \\
&< \infty. \qquad (2.34)
\end{aligned}$$

The sequence of sets $(A_m)_{m\in\mathbb{N}}$ is decreasing. Thus, by continuity of the P-measure,

$$\mathrm{P}\left(\limsup_{n\to\infty} \sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sup_{f\in S_n} \frac{|\mathbb{G}_{nn}(f)|}{\mathbb{U}_{nnn}(S_n)} > 1+\eta \right) = \mathrm{P}\left(\lim_{m\to\infty} A_m \right) = \lim_{m\to\infty} \mathrm{P}(A_m) = 0, \quad (2.35)$$

where the last equality follows from eq. (2.34). Since eq. (2.35) holds for all $\eta > 0$ we conclude that there exists an $N_0 > 0$ such that for all $n > N_0$,

$$\max_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sup_{f\in S_n} \frac{|\mathbb{G}_{nn}(f)|}{\mathbb{U}_{nnn}(S_n)} \leq 1 \quad a.s.$$

This establishes the claim. □

2.5.2.4 Proof of Theorem 2.4

Proof. Note that

$$\begin{aligned}
\mathbb{V}_{nn}([f]_{\mathscr{R}_n}) &\leq 2 \sup_{f\in[f]_{\mathscr{R}_n}} \mathbb{E}_{nn}[f^2] + 2\mathrm{E}\Big[\sup_{f\in[f]_{\mathscr{R}_n}} \mathbb{E}_{nn}[f^2] \Big] \\
&\leq \frac{2}{\sqrt{n}} \sup_{f\in[f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f^2)| + \frac{2}{\sqrt{n}} \mathrm{E}\Big[\sup_{f\in[f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f^2)| \Big] + 4 \sup_{f\in[f]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2].
\end{aligned}$$

We can use non-adaptive bound from Lemma 2.1 to control the random quantity in above display. We then combine this bound with the adaptive bound from Lemma 2.2 and

conclude that there exist constants $c_0, c_1, c_2, N_0 > 0$ such that for all $n > N_0$ and all $f \in \mathscr{F}_n$,

$$
\begin{aligned}
&|\mathbb{G}_{nn}(f)| \\
&\leq c_0 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\
&+ 4c_1 \sup_{f \in [f]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2]^{1/2} \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} \\
&+ 24^{1/2} c_1 \bar{\mathrm{E}}_{nn}[F^4_{[f]_{\mathscr{R}_n}}]^{1/4} \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} \left(\frac{\log\log n}{n}\right)^{1/4} \\
&+ 2^{3/2} c_1 \bar{\mathrm{E}}_{nn}[F^2_{[f]_{\mathscr{R}_n}}]^{1/2} \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} n^{-1/4} \\
&+ (2 + 2^{1/2}) c_1 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f^2)|\Big]^{1/2} \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} n^{-1/4} \\
&\leq c_0 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\
&+ c_1 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}(f^2)|\Big]^{1/2} \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} \\
&+ c_2 \left(\sup_{f \in [f]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2]^{1/2} + \bar{\mathrm{E}}_{nn}[F^4_{[f]_{\mathscr{R}_n}}]^{1/4} \left(\frac{\log\log n}{n}\right)^{1/4}\right) \\
&\quad \times \left(\log \frac{1}{v_n([f]_{\mathscr{R}_n})} + \log\log n\right)^{1/2} \quad a.s.
\end{aligned}
$$

□

2.5.3 Proofs of Section 2.3.3

2.5.3.1 Proof of Lemma 2.4

Proof. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$ be a vector if i.i.d. Rademacher variables independent of $X = \left(X^{(1)}, \ldots, X^{(d)}\right)$. Let $K = 1/(e - \sqrt{e})$, $p \geq 1$, and $\|v\|_2 = 1$, $v \in \mathbb{R}^d$. The key idea is to apply Khinchine's inequality; this allows us to exploit the fact that v lies on the unit sphere in $\mathbb{R}^d$ and gives a bound independent of d. We have

$$
\mathrm{E}\left[|X'v|^{2p}\right]^{1/(2p)} \leq \left(\mathrm{E}\left[\left|\sum_{k=1}^{d} X^{(k)} v_k - \mathrm{E}\left[X^{(k)} v_k\right]\right|^{2p}\right]\right)^{1/(2p)} + \left|\mathrm{E}\left[X'v\right]\right|
$$

$$
\begin{aligned}
&\leq 2\left(\mathrm{E}\left[\left|\sum_{k=1}^{d}\varepsilon_k X^{(k)}v_k\right|^{2p}\right]\right)^{1/(2p)} + \left|\mathrm{E}\left[(X'v)^2\right]\right|^{1/2} \\
&\leq 2^{2+1/(2p)}K^{1/2}p^{1/2}\left(\mathrm{E}\left[\left(\sum_{k=1}^{d}\left(X^{(k)}v_k\right)^2\right)^{p}\right]\right)^{1/(2p)} + \bar{\varphi}_{\max}^{1/2}(d) \\
&\leq 2^{2+1/(2p)}K^{1/2}p^{1/2}\left(\sum_{k=1}^{d}\left(\mathrm{E}\left[\left(X^{(k)}v_k\right)^{2p}\right]\right)^{1/p}\right)^{1/2} + \bar{\varphi}_{\max}^{1/2}(d) \\
&= 2^{2+1/(2p)}K^{1/2}p^{1/2}\left(\sum_{k=1}^{d}\mu_{2p}^2 v_k^2\right)^{1/2} + \bar{\varphi}_{\max}^{1/2}(d) \\
&= 2^{2+1/(2p)}K^{1/2}p^{1/2}\mu_{2p} + \bar{\varphi}_{\max}^{1/2}(d), \qquad (2.36)
\end{aligned}
$$

where the second inequality follows from Jensen's inequality and the symmetrization inequality for conditional Rademacher averages, the third from Khinchine's inequality for conditional Rademacher averages, the fourth from Minkowski's integral inequality, and the remaining inequalities from elementary calculations. Note that he first term on the right side of eq. (2.36) is independent of d and v. This concludes the proof of the first statement. The second statement can be proved analogously. □

2.5.3.2 Proof of Lemma 2.5

Proof. We combine a combinatorial (i.e. sample distribution independent) bound on the entropy of a conditional Rademacher average with a probabilistic (i.e. sample distribution dependent) bound on the second and fourth moments of the corresponding unconditional process. This allows us to leverage the fact that the variance of the (unconditional) process is proportional to r_n.

Step 1. Let $p \in \{1,2\}$ and $r_n > 0$ and consider the following class of functions:

$$
\mathscr{F}_{p,r_n} = \left\{(Y,X) \mapsto 1\left\{X'\theta_1 < Y \leq X'\theta_2\right\}(X'v)^p : \|\theta_1 - \theta_2\| \leq r_n,\, \theta_1,\theta_2 \in \mathbb{R}^d\right\}.
$$

Note that $G^p(X) = |X'v|^p \vee 1$ is an envelop function for $\mathscr{F}_{p,r_n}$. Denote the normalized function class by

$$
\widetilde{\mathscr{F}}_{p,r_n} = \left\{f/G^p : f \in \mathscr{F}_{p,r_n}\right\},
$$

and the collection subgraphs of $f \in \widetilde{\mathscr{F}}_{p,\infty}$ by $\widetilde{\Gamma}_p$. Then,

$$\begin{aligned}
\widetilde{\Gamma}_1 &= \left\{(y,x,t) : f(y,x) \geq t,\, f \in \widetilde{\mathscr{F}}_{1,\infty}\right\} \\
&= \left\{x'v/G(x) \geq t\right\} \cap \left\{y > x'\theta_1,\, \theta_1 \in \mathbb{R}^d\right\} \cap \left\{y \leq x'\theta_2,\, \theta_2 \in \mathbb{R}^d\right\}, \\
\widetilde{\Gamma}_2 &= \left\{(y,x,t) : f(y,x) \geq t,\, f \in \widetilde{\mathscr{F}}_{2,\infty}\right\} \\
&= \left(\left\{x'v/G(x) \geq |t|^{1/2}\right\} \cup \left\{x'v/G(x) \leq -|t|^{1/2}\right\} \right) \\
&\qquad \cap \left\{y > x'\theta_1,\, \theta_1 \in \mathbb{R}^d\right\} \cap \left\{y \leq x'\theta_2,\, \theta_2 \in \mathbb{R}^d\right\}.
\end{aligned}$$

Hence, by Lemma 2.6.18 in van der Vaart and Wellner (1996) $\widetilde{\mathscr{F}}_{1,\infty}$ and $\widetilde{\mathscr{F}}_{2,\infty}$ are VC-subgraph with VC-indices at most $2d+6$. Therefore, by Haussler's VC bound (e.g. van der Vaart and Wellner, 1996, Theorem 2.6.7) for any probability measure Q and $r \geq 1$ the $L^r(Q)$ covering numbers of $\widetilde{\mathscr{F}}_{p,\infty}$ satisfy

$$N\left(\eta, \widetilde{\mathscr{F}}_{p,\infty}, L^r(Q)\right) \leq C_p(2d+6)(16e)^{2d+6}\left(\frac{1}{\eta}\right)^{r(2d+5)}, \tag{2.37}$$

where $0 < \eta < 1$ and $C_p > 0$ is a universal constant depending only on $p \in \{1,2\}$.

For $\delta > 0$, define

$$\sigma_{n,p}(G,\delta) = \sup_{f \in \mathscr{F}_p(r_n)} \|f\|_{\mathbb{P}_n,2}\, \|G^p\|_{\mathbb{P}_n,2+\delta}.$$

By construction $G^p \geq 1$ and therefore $\sigma_{n,p}(G,\delta) \geq \sup_{f \in \mathscr{F}_{p,r_n}} \|f\|_{\mathbb{P}_n,2}$. Moreover, for arbitrary $f_1, f_2 \in \mathscr{F}_{p,r_n}$, by the generalized Hölder inequality,

$$\begin{aligned}
&\|f_1 - f_2\|^2_{\mathbb{P}_n,2}\, \sigma^{-1}_{n,p}(G,\delta) \\
&\leq \sup_{f \in \mathscr{F}_{p,r_n}} \frac{2}{n}\sum_{i=1}^n \left| \frac{f_1(Y_{ni},X_{ni})}{G^p(X_{ni})} - \frac{f_2(Y_{ni},X_{ni})}{G^p(X_{ni})} \right| |f(Y_{ni},X_{ni})G^p(X_{ni})|\, \sigma^{-1}_{n,p}(G,\delta) \\
&\leq 2\,\|f_1/G^p - f_2/G^p\|_{\mathbb{P}_n,2+4/\delta} \sup_{f \in \mathscr{F}_{p,r_n}} \|f\|_{\mathbb{P}_n,2}\, \|G^p\|_{\mathbb{P}_n,2+\delta}\, \sigma^{-1}_{n,p}(G,\delta) \\
&= 2\,\|f_1/G^p - f_2/G^p\|_{\mathbb{P}_n,2+4/\delta}.
\end{aligned} \tag{2.38}$$

Therefore, by eq. (2.38) the covering numbers satisfy

$$\begin{aligned}
N\left(\eta\, \sigma_{n,p}(G,\delta), \mathscr{F}_{p,r_n}, L^2(\mathbb{P}_n)\right) &\leq N\left(\eta^2/2, \widetilde{\mathscr{F}}_{p,r_n}, L^{2+4/\delta}(\mathbb{P}_n)\right) \\
&\leq N\left(\eta^2/2, \widetilde{\mathscr{F}}_{p,\infty}, L^{2+4/\delta}(\mathbb{P}_n)\right)
\end{aligned} \tag{2.39}$$

and, applying eq. (2.37) for $Q = \mathbb{P}_n$ and $r = 2 + 4/\delta$ to the right side of eq. (2.39), we arrive at

$$N\left(\eta\, \sigma_{n,p}(G,\delta'), \mathscr{F}_{p,r_n}, L^2(\mathbb{P}_n)\right) \leq C_p(2d+6)(16e)^{2d+6}\left(\frac{2}{\eta}\right)^{(2+4/\delta)(2d+5)}. \tag{2.40}$$

Step 2. Combinatorial bound. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be a vector of i.i.d. Rademacher random variables independent of $(Y_{n1}, X_{n1}), \ldots, (Y_{nn}, X_{nn})$. By Dudley's entropy inequality (e.g. van der Vaart and Wellner, 1996, Corollary 2.2.8), inequality (2.40), a change of variables, two applications of Cauchy-Schwarz and one of Jensen's inequality,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}} \mathbb{G}_{nn}(f)\Big] \\
&\leq 2\mathrm{E}\left[\sup_{f\in\mathscr{F}_{p,r_n}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_{ni}, X_{ni})\right] \\
&\leq \frac{C_p}{\sqrt{n}}\mathrm{E}\left[\int_0^{\sigma_{n,p}(G,\delta)} \left(\log N\left(\eta, \mathscr{F}_{p,r_n}, L^2(\mathbb{P}_n)\right)\right)^{1/2} d\eta\right] \\
&= \frac{C_p}{\sqrt{n}}\mathrm{E}\left[\sigma_{n,p}(G,\delta)\int_0^1 \left(\log N\left(\eta\, \sigma_{n,p}(G,\delta), \mathscr{F}_{p,r_n}, L^2(\mathbb{P}_n)\right)\right)^{1/2} d\eta\right] \\
&\leq C_p\left(\frac{2d+6}{n}\right)^{1/2} \mathrm{E}\left[\sigma_{n,p}(G,\delta)\right]\int_0^1 \left(\log\left(1/\eta\right)\right)^{1/2} d\eta \\
&\leq C_p\left(\frac{2d+6}{n}\right)^{1/2} \mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}} \mathbb{E}_{nn}[f^2]\Big]^{1/2} \bar{\mathrm{E}}_{nn}[G^{p(2+\delta)}]^{1/(2+\delta)}
\end{aligned} \tag{2.41}$$

where $C_p > 0$ is a universal constant (that may change from line to line) depending only on $p \in \{1,2\}$.

Step 3. Moment bounds. On the one hand, by the boundedness of the conditional density of $Y_n|X_n$,

$$\begin{aligned}
&\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}} \mathbb{E}_{nn}[f^2]\Big] \\
&\leq \frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}} \mathbb{G}_{nn}(f^2)\Big] + \sup_{f\in\mathscr{F}_{p,r_n}} \bar{\mathrm{E}}_{nn}[f^2] \\
&\leq \frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}} \mathbb{G}_{nn}(f^2)\Big] \\
&\quad + \sup_{f\in\mathscr{F}_{p,r_n}} \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n (X_{ni}'v)^{2p}\left|F_{Y_n|X_n}(X_{ni}'\theta_2|X_{ni}) - F_{Y_n|X_n}(X_{ni}'\theta_1|X_{ni})\right|\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}}\mathbb{G}_{nn}(f^2)\Big] + \mathrm{E}\left[\frac{f_+}{n}\sum_{i=1}^{n}|X_{ni}'\theta_2 - X_{ni}'\theta_1|G^{2p}(X_{ni})\right] \\
&\leq \frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}}\mathbb{G}_{nn}(f^2)\Big] + r_n\,\varphi_{\max}(d)\,f_+\,\bar{\mathrm{E}}_{nn}[G^{4p}]^{1/2}
\end{aligned}
\tag{2.42}
$$

On the other hand, since $0 \leq f^2 \leq g^{2p} \leq G^{2p}$,

$$
\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{p,r_n}}\mathbb{E}_{nn}[f^2]\Big] \leq \bar{\mathrm{E}}_{nn}[G^{2p}]. \tag{2.43}
$$

Step 4. We now combine the combinatorial bound with the moment bounds. We explain each of the inequalities below.

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbb{E}\Big[\sup_{f\in\mathscr{F}_{1,r_n}}\mathbb{G}_n(f)\Big] \\
&\leq C_1\left(\frac{d}{n}\right)^{1/2}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{1,r_n}}\mathbb{E}_{nn}[f^2]\Big]^{1/2}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)} \\
&\leq C_1\left(\frac{d}{n}\right)^{1/2}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\left|\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{2,r_n}}\mathbb{G}_{nn}(f)\Big]\right|^{1/2} \\
&\quad + C_1 f_+^{1/2}\bar{\varphi}_{\max}^{1/2}(d)\,r_n^{1/2}\left(\frac{d}{n}\right)^{1/2}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\,\bar{\mathrm{E}}_{nn}[G^4]^{1/4} \\
&\leq C_1C_2^{1/2}\left(\frac{d}{n}\right)^{3/4}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\mathrm{E}\Big[\sup_{f\in\mathscr{F}_{2,r_n}}\mathbb{E}_{nn}[f^2]\Big]^{1/4}\bar{\mathrm{E}}_{nn}[G^{4+2\delta)}]^{1/(4+2\delta)} \\
&\quad + C_1 f_+^{1/2}\bar{\varphi}_{\max}^{1/2}(d)\,r_n^{1/2}\left(\frac{d}{n}\right)^{1/2}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\,\bar{\mathrm{E}}_{nn}[G^4]^{1/4} \\
&\leq c_0\left(\frac{d}{n}\right)^{3/4}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\,\bar{\mathrm{E}}_{nn}[G^4]^{1/4}\,\bar{\mathrm{E}}_{nn}[G^{4+2\delta)}]^{1/(4+2\delta)} \\
&\quad + c_0 f_+^{1/2}\bar{\varphi}_{\max}^{1/2}(d)\,r_n^{1/2}\left(\frac{d}{n}\right)^{1/2}\bar{\mathrm{E}}_{nn}[G^{2+\delta}]^{1/(2+\delta)}\,\bar{\mathrm{E}}_{nn}[G^4]^{1/4},
\end{aligned}
\tag{2.44}
$$

where $c_0 > 0$ is a universal constant. The first inequality follows from combinatorial bound (2.41) for $p = 1$, the second inequality follows from moment bound (2.42) for $p = 1$, the third inequality follows from combinatorial bound (2.41) for $p = 2$, and the last inequality follows from moment bound (2.43) for $p = 2$.

By Lemma 2.4 we can bound the moments of envelop function G by multiples of $\bar{\varphi}_{\max}(d)$. In particular, there exists a universal constant $c_1 > 0$ such that eq. (2.44) simplifies

to

$$\frac{1}{\sqrt{n}}\mathbb{E}\Big[\sup_{f\in\mathscr{F}_{1,r_n}}\mathbb{G}_{nn}(f)\Big]\leq c_1\bar{\varphi}_{\max}^{3/2}(d)\left(\frac{d}{n}\right)^{3/4}+c_1 f_+^{1/2}\bar{\varphi}_{\max}^{3/2}(d)r_n^{1/2}\left(\frac{d}{n}\right)^{1/2}.$$

□

2.5.3.3 Proof of Lemma 2.6

Proof. **Step 1. Reduction to supremum over m-sparse unit ball.** Let $S\subseteq\{1,\dots,d\}$ be a set of indices with cardinality $|S|=m$, $u\in\mathbb{R}^m$, and $v\in\mathbb{R}^d$, with support S and $v_S=u$. Then, for any $X\in\mathbb{R}^d$, we have

$$X_S'u=X'v.$$

Therefore,

$$\sup_{\substack{\|u\|_2=1\\ u\in\mathbb{R}^m}}\sup_{S:|S|=m}\mathbb{G}_{nn}\big(F(X_{ni,S}'u)\big)\leq\max_{v\in\mathscr{E}_m^d}\mathbb{G}_{nn}\big(F(X_{ni}'v)\big).$$

where $\mathscr{E}_m^d\subset\mathscr{B}^d$ denotes the m-sparse subset of the unit ball in $\mathbb{R}^d$, i.e.

$$\mathscr{E}_m^d=\Big\{v\in\mathbb{R}^d:\|v\|_2\leq 1,\|v\|_0\leq m\Big\}.$$

Thus, taking the supremum over models (of size m) is equivalent to taking the supremum over sparse unit vectors (with support sets of size at most m).

Step 2. Proof of the first statement. Decompose $\mathscr{E}_m^d$ in the following way:

$$\mathscr{E}_m^d=\bigcup_{k=1}^m\bigcup_{S:|S|=k}\mathscr{B}_S^d,$$

where $\mathscr{B}_S^d$ is the subset of the unit ball in $\mathbb{R}^d$ with support set S, i.e.

$$\mathscr{B}_S^d=\Big\{v\in\mathbb{R}^d:\|v\|_2\leq 1,\ j\notin S\implies v_j=0\Big\}.$$

Let $\mathscr{N}_S^d$ be an ε-net of $\mathscr{B}_S^d$ and define

$$\mathscr{M}_m^d=\bigcup_{k=1}^m\bigcup_{S:|S|=k}\mathscr{N}_S^d.$$

Since $|\mathscr{N}_S^d| \le \left(1+\frac{2}{\varepsilon}\right)^{|S|}$ for all S, we have

$$|\mathscr{M}_m^d| \le \sum_{k=1}^{m} \binom{d}{k} \left(1+\frac{2}{\varepsilon}\right)^k \le \left(1+\frac{2}{\varepsilon}\right)^m \left(\frac{ed}{m}\right)^m.$$

Thus, by the decomposition of $\mathscr{E}_m^d$ into the union of sparse unit balls $\mathscr{B}_S^d$,

$$\max_{v\in\mathscr{E}_m^d} \mathbb{G}_{nn}\big(F(X'_{ni}v)\big) \le \max_{v\in\mathscr{M}_m^d} \sup_{u\in\mathscr{B}^d_{\mathrm{supp}(v)}(v,\varepsilon)} \mathbb{G}_{nn}\big(F(X'_{ni}u)\big).$$

Step 3. Proof of the second statement. We now show that for $\varepsilon \in (0, \frac{1}{2}]$ there exist an absolute constant $C > 0$ (independent of d, m, n and ε) and a countable finite set $\widetilde{\mathscr{M}}_m^d \subset \mathscr{B}^d$ such that

$$\mathscr{E}_m^d \subset \mathrm{conv}(\widetilde{\mathscr{M}}_m^d) \quad \text{and} \quad |\widetilde{\mathscr{M}}_m^d| \le C\left(1+\frac{2}{\varepsilon}\right)^{4m}\left(\frac{ed}{m}\right)^{4m}.$$

We then argue that since F is linear it suffices to take the maximum over $\widetilde{\mathscr{M}}_m^d$ instead of the supremum over $\mathrm{conv}(\widetilde{\mathscr{M}}_m^d)$.

Let $\mathscr{M}_m^d$ be as above. By Lemma 7.1 in Koltchinskii (2011) with $b = 0$, we have

$$\mathscr{E}_m^d \subset 4\,\mathrm{conv}\left(\mathscr{M}_m^d\right),$$

Since taking the Minkowski summation and forming the convex hull commute, we have

$$4\,\mathrm{conv}(\mathscr{M}_m^d) = \mathrm{conv}(4\,\mathscr{M}_m^d).$$

By the classical sum set estimate (e.g. Tao and Vu, 2006, Lemma 2.1) we have

$$|4\,\mathscr{M}_m^d| \le \binom{|\mathscr{M}_m^d|+3}{4} \le C|\mathscr{M}_m^d|^4 \le C\left(1+\frac{2}{\varepsilon}\right)^{4m}\left(\frac{ed}{m}\right)^{4m},$$

where $C > 0$ is an absolute constant. Finally, set $\widetilde{\mathscr{M}}_m^d = 4\,\mathscr{M}_m^d$.

Since F is a linear function, so is the map $v \mapsto \mathbb{G}_{nn}\big(F(X'_{ni}v)\big)$. Hence, the Bauer Maximum Principle (e.g. Aliprantis and Border, 2006, Corollary 7.69, p. 298) guarantees that at least one of the maximizers $v \mapsto \mathbb{G}_{nn}\big(F(X'_{ni}v)\big)$ over the closed, convex set $\mathrm{conv}(\widetilde{\mathscr{M}}_m^d)$ is an extreme point of $\mathrm{conv}(\widetilde{\mathscr{M}}_m^d)$. But the set of extreme points of $\mathrm{conv}(\widetilde{\mathscr{M}}_m^d)$ is just $\widetilde{\mathscr{M}}_m^d$. The statement holds for any $\varepsilon \in (0, \frac{1}{2}]$; in particular, it holds for $\varepsilon = \frac{1}{2}$. This proves the second statement. □

2.5.3.4 Proof of Lemma 2.7

Proof. The result follows from Theorem 2.4 for a suitably chosen equivalence relation $\mathscr{R}_n$ and probability measure ν_n. The equivalence relation $\mathscr{R}_n$ should be constructed such that it induces a partition $\mathscr{F}_n/\mathscr{R}_n$ in which each sub-class of functions has a small (local) envelope function and finite (local) metric entropy, and the probability measure ν_n should assign the same mass to elements with the same metric entropy, thus forming a link between the local metric entropy and the metric entropy of the original function class $\mathscr{F}_n$. Often it helps embed the function class under consideration into a larger function class with a simpler structure.

Step 1. Reduction to decoupled function class. Note that functions under consideration can be decomposed into the difference of two simpler functions:

$$\begin{aligned}&\varphi_\tau(Y - X_M'\theta_{1,M})X'u - \varphi_\tau(Y - X_M'\theta_{2,M})X_M'u \\ &\quad = 1\left\{X_M'\theta_{1,M} < Y \le X_M'\theta_{2,M}\right\}X_M'u - 1\left\{X_M'\theta_{2,M} < Y \le X_M'\theta_{1,M}\right\}X_M'u.\end{aligned}$$

Thus, it suffices to find an upper bound for the supremum of the expected value over the function class

$$\begin{aligned}\mathscr{F}_1 = \Big\{&(Y,X) \mapsto 1\left\{X_M'\theta_{1,M} < Y \le X_M'\theta_{2,M}\right\}X_M'u : \\ &M \subseteq \{1,\dots,d\},\, u \in \mathscr{B}^{|M|},\, \|\theta_{1,M} - \theta_{2,M}\| \le r_n(|M|),\, \theta_1,\theta_2 \in \mathbb{R}^d\Big\},\end{aligned}$$

where $\mathscr{B}^{|M|}$ is the unit ball in $\mathbb{R}^{|M|}$. The dependence of the indicator function $1\{X_M'\theta_{1,M} < Y \le X_M'\theta_{2,M}\}$ and the inner product $X_M'u$ on the same support set M complicates the analysis. It turns out that it is easier to consider the larger, decoupled function class

$$\begin{aligned}\mathscr{F}_2 = \Big\{&(Y,X) \mapsto 1\left\{X_M'\theta_{1,M} < Y \le X_M'\theta_{2,M}\right\}X_S'u : \\ &M,S \subseteq \{1,\dots,d\},\, |S| = |M|,\, u \in \mathscr{B}^{|S|},\, \|\theta_{1,M} - \theta_{2,M}\| \le r_n(|M|),\, \theta_1,\theta_2 \in \mathbb{R}^d\Big\}.\end{aligned}$$

By Lemma 2.6 there exist finite sets $\mathscr{M}_m^d \subset \mathscr{B}^d$, $m \in \{1,\dots,d\}$ such that the supremum over function class $\mathscr{F}_2$ can be upper bounded by the supremum over the following function class

$$\begin{aligned}\mathscr{F}_3 = \Big\{&(Y,X) \mapsto 1\left\{X_M'\theta_{1,M} < Y \le X_M'\theta_{2,M}\right\}X'v : \\ &M \subseteq \{1,\dots,d\}, v \in \mathscr{M}_{|M|}^d,\, \|\theta_{1,M} - \theta_{2,M}\| \le r_n(|M|),\, \theta_1,\theta_2 \in \mathbb{R}^d\Big\}.\end{aligned}$$

This function class is easy to deal with as it is the product of a collection of bounded function $(Y,X) \mapsto 1\left\{X_M'\theta_{1,M} < Y \leq X_M'\theta_{2,M}\right\}$ and a collection of an unbounded linear function $(Y,X) \mapsto X'v$. In particular, while the collection of bounded functions is uncountable (note that the parameters (θ_1,θ_2) range in an uncountable subset of $\mathbb{R}^d$), the collection of the unbounded linear function is finite (the parameter v ranges in $\mathscr{M}_m^d$, $m \in \{1,\ldots,d\}$). This is structure is important in order to obtain tight uniform almost sure bounds via Theorem 2.4.

Step 2. Equivalence relation and probability measure. Define the equivalence relation $\mathscr{R}_n \subseteq \mathscr{F}_3 \times \mathscr{F}_3$ by

$$(f_{v^1,M^1,\theta_1^1,\theta_2^1}, f_{v^2,M^2,\theta_1^2,\theta_2^2}) \in \mathscr{R}_n \iff \left\{v^1 = v^2,\ |M^1| = |M^2|,\ v^1,v^2 \in \mathscr{M}_{|M^1|}^d\right\},$$

and the probability measure $\nu_n : \sigma(\mathscr{F}_3/\mathscr{R}_n) \to [0,1]$ by $\nu_n(\emptyset) = 0$ and

$$\nu_n(P_{v,M}) = c_\nu^{-1}\left(\frac{5ed}{|M|}\right)^{-3|M|},\ P_{v,M} \in \mathscr{F}_3/\mathscr{R}_n,$$

where $c_\nu > 0$ is such that that $1 = \sum_{P_{v,M} \in \mathscr{F}_{r_n}/\mathscr{R}_n} \nu_n(P_{v,M})$. Note that $0 < c_\nu < \frac{e}{2}C$, where $C > 0$ is the absolute constant from Lemma 2.6. Indeed,

$$\begin{aligned} c_\nu = \sum_{P_{v,M} \in \mathscr{F}_{r_n}/\mathscr{R}_n} c_\nu \nu_n(P_{v,M}) &= \sum_{k=1}^d |\mathscr{M}_k^d| \binom{d}{k} \left(\frac{5ed}{k}\right)^{-3k} \\ &\leq C\sum_{k=1}^d \left(\frac{5ed}{k}\right)^{-k} \leq C\sum_{k=1}^\infty (5e)^{-k} < \frac{C}{1-(5e)^{-1}}. \end{aligned}$$

Further, note that each subclass $P_{v,M} \in \mathscr{F}_3/\mathscr{R}_n$ is a VC-subgraph class of functions with VC-index at most $2|M|+6$ and envelop function $G_v(X) = |X'v| \vee 1$. Thus, the elements in the partition $\mathscr{F}_3/\mathscr{R}_n$ have comparable metric entropy and small (i.e. dimension independent) local envelops (note that $\|v\|_2 \leq 1$ is fixed!).

Step 3. By Theorem 2.4 and Lemma 2.5 there exist absolute constants $c_6, N_0 > 0$ such that for all $n > N_0$ and all $f_{v,M,\theta_1,\theta_2} \in \mathscr{F}_3$ we have

$$\begin{aligned} &\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{v,M,\theta_1,\theta_2})| \\ &\quad\leq \frac{7}{\sqrt{n}}\mathrm{E}\Big[\sup_{f \in [f_{v,M,\theta_1,\theta_2}]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\ &\quad+ 69\mathrm{E}\Big[\sup_{f \in [f_{v,M,\theta_1,\theta_2}]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]^{1/2} \end{aligned}$$

$$
\begin{aligned}
&\quad \times \left(\frac{\log(eC/2) + 2|M|\log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&+ 149 \left(\sup_{f \in [f_{v,M,\theta_1,\theta_2}]_{\mathscr{F}_n}} \bar{\mathrm{E}}_{nn}[f^2] \right)^{1/2} \\
&\quad \times \left(\frac{\log(eC/2) + 2|M|\log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&+ 223 \left(\bar{\mathrm{E}}_{nn}[G_{v,M}^4] \right)^{1/4} \\
&\quad \times \left(\frac{\log(eC/2) + 2|M|\log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \left(\frac{\log\log n}{n} \right)^{1/4} \\
&\leq c_6 \bar{\varphi}_{\max}^{3/2}(|M|) \left(\frac{|M|}{n} \right)^{3/4} + c_6 f_+^{1/2} \bar{\varphi}_{\max}^{3/2}(|M|) \left(\frac{|M|}{n} \right)^{1/2} r_n^{1/2}(|M|) \\
&\quad + c_6 \bar{\varphi}_{\max}(|M|) \left(\frac{|M|}{n} \right)^{1/4} \left(\frac{|M|\log(ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&\quad + c_6 f_+^{1/2} \bar{\varphi}_{\max}(|M|) \left(\frac{|M|\log(ed/|M|) + \log\log n}{n} \right)^{1/2} r_n^{1/2}(|M|) \\
&\quad + c_6 \bar{\varphi}_{\max}^{1/2}(|M|) \left(\frac{|M|\log(ed/|M|) + \log\log n}{n} \right)^{1/2} \left(\frac{\log\log n}{n} \right)^{1/4}.
\end{aligned}
\tag{2.45}
$$

Two times this upper bound (i.e. display (2.45)) yields an almost sure upper bound for any function f in the original function class $\mathscr{F}_n$. Since $\bar{\varphi}_{\max}(|M|) \geq 1$ this can be further upper bounded as in the statement of the lemma. □

2.5.3.5 Proof of Lemma 2.8

Proof. **Step 1. Equivalence relation and probability measure**. Note that for $f \in \mathscr{F}_n$,

$$
\begin{aligned}
\|\mathbb{E}_{nn}[f]\|_2 = \frac{1}{\sqrt{n}} \|\mathbb{G}_{nn}(f)\|_2 &\leq \left\| \frac{1}{n} \sum_{i=1}^n \left(X_{ni,M} - \mathrm{E}[X_{ni,M}] \right) \right\|_2 \\
&+ \left\| \frac{1}{n} \sum_{i=1}^n \left(1\{Y_{ni} \leq X_{ni,M}' \theta_{n,M}^*(\tau)\} X_{ni,M} - \mathrm{E}[1\{Y_{ni} \leq X_{ni,M}' \theta_{n,M}^*(\tau)\} X_{ni,M}] \right) \right\|_2.
\end{aligned}
$$

By Lemma 2.6 there exist finite sets $\mathscr{M}_m^d \subseteq \mathscr{B}^d$, $m \in \{1,\ldots,d\}$ such that it suffices to consider the following two function classes:

$$
\begin{aligned}
\mathscr{F}_1 &= \left\{ (Y,X) \mapsto 1\left\{ Y \leq X_M' \theta_{n,M}^*(\tau) \right\} X'v : \tau \in \mathscr{T}, v \in \mathscr{M}_{|M|}^d, M \subseteq \{1,\ldots,d\} \right\}, \\
\mathscr{F}_2 &= \left\{ (Y,X) \mapsto X'v : v \in \mathscr{M}_{|M|}^d, M \subseteq \{1,\ldots,d\} \right\}.
\end{aligned}
$$

We only derive bounds involving function class $\mathscr{F}_1$; bounds for function class $\mathscr{F}_2$ can be established analogously. We define the equivalence relation $\mathscr{R}_1 \subseteq \mathscr{F}_1 \times \mathscr{F}_1$ by

$$(f_{v^1,M^1,\tau^1}, f_{v^2,M^2,\tau^2}) \in \mathscr{R}_1 \iff \left\{ v^1 = v^2,\ |M^1| = |M^2|,\ v^1, v^2 \in \mathscr{M}^d_{|M^1|} \right\},$$

and the sub-probability measure $\nu_1 : \sigma(\mathscr{F}_1/\mathscr{R}_1) \to [0,1]$ by $\nu_n(\emptyset) = 0$ and

$$\nu_1(P_{v,M}) = \frac{2}{eC} \left(\frac{5ed}{|M|} \right)^{-3|M|}, \ P_{v,M} \in \mathscr{F}_1/\mathscr{R}_1,$$

where $C > 0$ is the absolute constant from Lemma 2.6. Note that each subclass $P_{v,M} \in \mathscr{F}_1/\mathscr{R}_1$ is a VC-subgraph class of functions with VC-index at most $|M|+2$ and envelop function $G_{v,M}(X) = |X'_M v| \vee 1$.

Step 2. By Theorem 2.4 there exist $c_7, N_0 > 0$ such that for all $n > N_0$ and all $f_{v,M,\tau} \in \mathscr{F}_1$ we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \mathbb{G}_{nn}(f_{v,M,\tau}) \\
&\quad \le \frac{7}{\sqrt{n}} \mathrm{E}\Big[\sup_{f \in [f_{v,M,\tau}]_{\mathscr{R}_1}} |\mathbb{G}_{nn}(f)| \Big] \\
&\qquad + 69 \mathrm{E}\Big[\sup_{f \in [f_{v,M,\tau}]_{\mathscr{R}_1}} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}(f^2)| \Big]^{1/2} \\
&\qquad\quad \times \left(\frac{\log(eC/2) + 2|M| \log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&\qquad + 149 \left(\sup_{f \in [f_{v,M,\tau}]_{\mathscr{R}_1}} \bar{\mathrm{E}}_{nn}[f^2] \right)^{1/2} \\
&\qquad\quad \times \left(\frac{\log(eC/2) + 2|M| \log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&\qquad + 223 \left(\bar{\mathrm{E}}_{nn}[G^4_{v,M}] \right)^{1/4} \\
&\qquad\quad \times \left(\frac{\log(eC/2) + 2|M| \log(5ed/|M|) + \log\log n}{n} \right)^{1/2} \left(\frac{\log\log n}{n} \right)^{1/4} \\
&\quad \le c_7 \bar{\varphi}_{\max}(|M|) \left(\frac{|M|}{n} \right)^{1/2} + c_7 \bar{\varphi}^{1/2}_{\max}(|M|) \left(\frac{|M| \log(ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&\qquad + c_7 \bar{\varphi}_{\max}(|M|) \left(\frac{|M|}{n} \right)^{1/4} \left(\frac{|M| \log(ed/|M|) + \log\log n}{n} \right)^{1/2} \\
&\qquad + c_7 \bar{\varphi}^{1/2}_{\max}(|M|) \left(\frac{|M| \log(ed/|M|) + \log\log n}{n} \right)^{1/2} \left(\frac{\log\log n}{n} \right)^{1/4},
\end{aligned} \tag{2.46}$$

where the bounds on $\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}_n}}|\mathbb{G}_{nn}(f)|\Big]$ and $\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}_n}}|\mathbb{G}_{nn}(f^2)|\Big]$ follow from eq. (2.41). Display (2.46) is also an upper bound for functions from function class $\mathscr{F}_2$. Therefore, four times this upper bound (eq. (2.46)) yields an almost sure upper bound for any function f in the original function class $\mathscr{F}_n$. □

2.5.3.6 Proof of Lemma 2.9

Proof. **Step 1. Reduction to decoupled and countable function class.** Recall the function class in the statement of the lemma:

$$\mathscr{F}_1 = \Big\{(Y,X)\mapsto \rho_\tau(Y-X_M'\theta_M)-\rho_\tau(Y-X_M'\theta^*_{n,M}(\tau)):\\ \|\theta_M-\theta^*_{n,M}(\tau)\|_2\le r_n(|M|),\,\theta\in\mathbb{R}^d,\,\tau\in\mathscr{T},\,M\subseteq\{1,\dots,d\}\Big\}.$$

Note that $\mathscr{F}_1$ is contained in the larger, decoupled function class

$$\mathscr{F}_2 = \Big\{(Y,X)\mapsto \rho_\tau(Y-X_M'\theta^*_{n,M}(\tau)-X_S'\delta_S)-\rho_\tau(Y-X_M'\theta^*_{n,M}(\tau)):\\ \|\delta_S\|_2\le r_n(|S|),\,\delta\in\mathbb{R}^d,\;\tau\in\mathscr{T},\;|M|=|S|,\;M,S\subseteq\{1,\dots,d\}\Big\}.$$

By Lemma 2.6 there exist finite sets $\mathscr{M}^d_m\subset\mathscr{B}^d\big(r_n(m)\big)$, $m\in\{1,\dots,d\}$, $\|v\|_0\le m$ for all $v\in\mathscr{M}^d_m$, and $\varepsilon_M=|M|^{-1/2}$ such that the supremum over function class $\mathscr{F}_2$ can be upper bounded by the maximum over

$$\mathscr{F}_3 = \Big\{(Y,X)\mapsto \rho_\tau(Y-X_M'\theta^*_{n,M}(\tau)-X'\delta)-\rho_\tau(Y-X_M'\theta^*_{n,M}(\tau)):\\ \delta\in\mathscr{B}^d_{\mathrm{supp}(v)}\big(v,\varepsilon_M\cdot r_n(|M|)\big),\;v\in\mathscr{M}^d_{|M|},\;\tau\in\mathscr{T},\;M\subseteq\{1,\dots,d\}\Big\}.$$

Step 2. Equivalence relation and probability measure. Define the equivalence relation $\mathscr{R}_n\subseteq\mathscr{F}_3\times\mathscr{F}_3$ by

$$(f_{\tau^1,\delta^1,v^1,M^1},f_{\tau^2,\delta^2,v^2,M^2})\in\mathscr{R}_n\\ \iff\Big\{|M^1|=|M^2|,\;v^1=v^2,\;v^1,v^2\in\mathscr{M}^d_{|M^1|},\;\delta^1,\delta^2\in\mathscr{B}^d_{\mathrm{supp}(v^1)}\big(v^1,\varepsilon_M\cdot r_n(|M|)\big)\Big\},$$

and the sub-probability measure $\nu_n:\sigma(\mathscr{F}_3/\mathscr{R}_n)\to[0,1]$ by $\nu_n(\emptyset)=0$ and

$$\nu_n(P_{v,M})=\frac{2}{e}\left(1+\frac{2}{\varepsilon_M}\right)^{-|M|}\left(\frac{ed}{|M|}\right)^{-3|M|},\quad P_{v,M}\in\mathscr{F}_3/\mathscr{R}_n,$$

By the Lipschitz continuity of the check loss ρ_τ, each subclass $P_{v,M}\in\mathscr{F}_3/\mathscr{R}_n$ has envelop

function

$$G_{v,M}(X) = \sup_{\delta \in \mathscr{B}^d_{\mathrm{supp}(v)}(v, \varepsilon \cdot r_n(|M|))} |X'\delta| \leq \varepsilon_M \cdot \|X_M\|_2 r_n(|M|) + |X'v|.$$

Step 3. In the previous two steps we have effectively upper bounded the quantile loss function by the composition of two functions; one depending on $v \in \mathscr{M}^d_m$ and another depending on $\delta \in \mathscr{B}^d_{\mathrm{supp}(v)}\big(v, \varepsilon_M \cdot r_n(|M|)\big)$. We are now in the position of applying Corollary 2.2. We conclude that there exist $c_9, N_9 > 0$ such that for all $n > N_9$ and all $f_{\tau,\delta,v,M} \in \mathscr{F}_3$

$$\begin{aligned}
\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\delta,v,M})| \leq{}& \frac{c_9}{\sqrt{n}}\mathrm{E}\Big[\sup_{f \in [f_{\tau,\delta,v,M}]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\
&+ c_9 \left(\mathrm{E}\Big[\sup_{f \in [f_{\tau,\delta,v,M}]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]\right)^{1/2} \left(\frac{|M| \log\big(ed/(|M|\varepsilon_M)\big) + \log\log n}{n}\right)^{1/2} \\
&+ c_9 \left(\sup_{f \in [f_{\tau,\delta,v,M}]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2} \left(\frac{|M| \log\big(ed/(|M|\varepsilon_M)\big) + \log\log n}{n}\right)^{1/2} \\
&+ c_9 \Big(\bar{\mathrm{E}}_{nn}[G^4_{v,M}]\Big)^{1/4} \left(\frac{|M| \log\big(ed/(|M|\varepsilon_M)\big) + \log\log n}{n}\right)^{1/2} \left(\frac{\log\log n}{n}\right)^{1/4}.
\end{aligned} \tag{2.47}$$

To bound the first term on the right side of eq. (2.47), symmetrize the centered process (e.g. van der Vaart and Wellner, 1996, Lemma 2.3.7), combine the Lipschitz-continuity of the check loss with the contraction principle for Rademacher averages (e.g. Ledoux and Talagrand, 1996, Theorem 4.12) and Proposition 3.2 in Koltchinskii (2011) for the expected value of suprema over classes of linear functions,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f \in [f_{\tau,\delta,v,M}]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\
&\quad\leq \frac{2}{\sqrt{n}}\mathrm{E}\Big[\sup_{f \in [f_{\tau,\delta,M}]_{\mathscr{R}_n}} |\mathbb{G}^{\circ}_{nn}(f)|\Big] \\
&\quad\leq 2\mathrm{E} \sup_{\delta \in \mathscr{B}^d_{\mathrm{supp}(v)}(v, \varepsilon_M \cdot r_n(|M|))} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X'_{ni}(\delta - v)\right| + 2\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X'_{ni} v\right| \\
&\quad\leq \varepsilon_M \cdot 2\bar{\varphi}^{1/2}_{\max}(|M|) \left(\frac{|M|}{n}\right)^{1/2} r_n(|M|) + 2\bar{\varphi}^{1/2}_{\max}(|M|) \frac{r_n(|M|)}{\sqrt{n}}.
\end{aligned} \tag{2.48}$$

The first part of the second term on the right side of eq. (2.47) can be upper bounded in

the same way and using the moment bounds from Lemma 2.4, i.e.

$$\left(\mathrm{E}\Big[\sup_{f\in[f_{\tau,\delta,v,M}]_{\mathscr{R}_n}}\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]\right)^{1/2}\leq \varepsilon_M^{1/2}\cdot c_{10}\bar{\varphi}_{\max}^{1/4}(|M|)\left(\frac{|M|}{n}\right)^{1/4} r_n^{1/2}(|M|) + c_{10}\bar{\varphi}_{\max}^{1/4}(|M|)\frac{r_n^{1/2}(|M|)}{n^{1/4}}, \tag{2.49}$$

where $c_{10}>0$ is an absolute constant.

The first part of the third term on the right side of eq. (2.47) can be upper bounded using the Lipschitz-continuity,

$$\left(\sup_{f\in[f_{\tau,\delta,v,M}]_{\mathscr{R}_n}}\bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2}\leq \varepsilon_M\cdot\bar{\varphi}_{\max}^{1/2}(|M|)r_n(|M|)+\bar{\varphi}_{\max}^{1/2}(|M|)r_n(|M|). \tag{2.50}$$

The first part of the fourth term on the right side of eq. (2.47) can be upper bounded using the Lipschitz-continuity and the moment bounds from Lemma 2.4,

$$\begin{aligned}\left(\bar{\mathrm{E}}_{nn}[G_{v,M}^4]\right)^{1/4} &\leq \varepsilon_M\cdot 15^{1/4}\left(\bar{\mathrm{E}}_{nn}[\|X_{ni,M}\|_2^4]\right)^{1/4} r_n(|M|)+15^{1/4}\left(\bar{\mathrm{E}}_{nn}[(X_{ni}'v)^4]\right)^{1/4}\\ &= c_{11}r_n(|M|),\end{aligned} \tag{2.51}$$

where $c_{10}>0$ is an absolute constant.

By combining (2.48)-(2.51) with eq. (2.47) we conclude that there exists $c_{12}>0$ such that for all $n>N_9$ and all $f_{\tau,\theta_1,\theta_2,M}\in\mathscr{F}_n$,

$$\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\theta_1,\theta_2})|\leq c_{12}\left(\frac{|M|\log\left(ed/|M|^{1/2}\right)+\log\log n}{n}\right)^{1/2}\bar{\varphi}_{\max}^{1/2}(|M|)r_n(|M|)\quad a.s.$$

□

CHAPTER 3

Inference for High-Dimensional Misspecified Quantile Regression Processes

3.1 Introduction

We consider the problem of conducting inference on high-dimensional sparse quantile regression processes when the assumed linear regression function is misspecified. That is, we consider a scenario in which the number of available predictor variables d may far exceed the sample size n, but only a small number of unknown predictor variables is indeed relevant for modeling the conditional quantile function of the response variable. Misspecification enters the picture in two ways: For one thing, we do not assume that the collection of available predictor variables contains all relevant predictors. Thus, any fitted model may suffer from an omitted variable bias. For another, we allow the true conditional quantile function to be nonlinear while we fit only linear regression functions. Thus, the fitted models may be considered linear approximations to the unknown true quantile function.

In the case of correctly specified high-dimensional sparse linear models a common strategy is to use an ℓ_1-penalized estimator to enforce sparsity on the vector of estimated coefficients. This approach was first proposed by Tibshirani (1996) in the context of least squares problems; its extension to the quantile regression problem was developed and analyzed by Belloni and Chernozhukov (2011). Recently, Bühlmann and van de Geer (2015) investigated ℓ_1-penalized estimators for misspecified high-dimensional sparse models of the conditional mean.

In this paper, we aim to extend the analysis of Bühlmann and van de Geer (2015) to the ℓ_1-penalized misspecified quantile regression estimator. To this end, we first derive a strong consistency result of the ℓ_1-penalized quantile regression process. We then use this result to construct a de-biased (or de-sparsified) quantile regression process which converges weakly to a Gaussian limit process. This generalizes and strengthens results obtained by Bradic and Kolar (2017) on correctly specified high-dimensional quantile regression processes. Second,

we analyze the theoretical properties of the refitted quantile regression process when the refit is based on a model selected by Lasso. Our post-selection consistency and sparsity results are comparable to those derived by Belloni and Chernozhukov (2011, 2013). Moreover, we are able to make explicit the impact of misspecification on the empirical sparsity of the refitted quantile regression process.

A key conceptual contribution of this paper is the notion of "quantile sublevel sets". This concept allows an intuitive formulation of sufficient conditions for the estimated quantile regression vector to be consistent for the best (sparse) approximation to the true CQF. Moreover, we find that the penalty parameter for which the restricted cone condition holds (or equivalently, the penalty parameter for which the subdifferential of the ℓ_1-penalized quantile regression problem contains zero) depends among other things on the degree of misspecification of the best approximation to the true CQF. Therefore, in general, the pivotal penalty parameter for correctly specified models proposed by Belloni and Chernozhukov (2011) does not provide the correct amount of penalization if the model is misspecified.

We organize our paper as follows. In Section 3.2 we introduce a general framework for quantile regression in high dimensions, state regularity conditions, and discuss the concept of τ-quantile sublevel sets in relation to the restricted cone property of ℓ_1-penalized estimators. In Section 3.3 we present our main results on debiased and post-selection quantile regression estimators. In Section 3.4 we state auxiliary results. We conclude in Section 3.5 with a brief discussion about future applications and generalizations of the presented results. Proofs to all theorems and auxiliary results can be found in Section 3.6.

We explain the notation used in the paper. In what follows, we implicitly index all parameters by the sample size n. Thus, when making asymptotic statements, we assume that $n \to \infty$ and $d = d_n \to \infty$ and $m = m_n \to \infty$. But we omit the index whenever this does not cause confusion. Constants $c, C, c_1, c_2, \ldots$ are understood to be independent of n and may change from line to line. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max a, b$, and $a \wedge b = \min\{a, b\}$. We denote the ℓ_2-norm by $\|\cdot\|_2$, the ℓ_1-norm by $\|\cdot\|_1$, the ℓ_∞-norm by $\|\cdot\|_\infty$, and the operator norm (which returns the largest singular value of a matrix) by $\|\cdot\|_{op}$. For $r > 0$, $v \in \mathbb{R}^d$ we use $\mathscr{B}^d(v, r)$ to denote the ball in $\mathbb{R}^d$ with center at v and radius r with respect to the Euclidean norm. We shall abbreviate $\mathscr{B}^d(0, 1)$ to $\mathscr{B}^d$. Analogously, we denote spheres of radius $r > 0$ by $\mathscr{S}^d(v, r)$ and write $\mathscr{S}^d$ for $\mathscr{S}^d(0, 1)$. Given a vector $X \in \mathbb{R}^d$ and a set of indices $M \subseteq \{1, \ldots, d\}$ we let X_M denote the vector $\{X^{(j)}, j \in M\}$. The cardinality of M is denoted by $|M|$. We denote the quantile loss function for quantile level $\tau \in (0, 1)$ by $\rho_\tau(u) = u(1 - 1\{u \leq 0\})$ and its subgradient by $\varphi_\tau(u) = \tau - 1\{u \leq 0\}$. We use the terms subgradient and quantile regression score function interchangeably.

Throughout, we use the empirical process notation of van der Vaart and Wellner (1996).

However, since we will be working in a triangular array setting, we introduce the following modifications: The symbol $\mathrm{E}[\cdot]$ denotes the expectation with respect to a generic probability measure P (which depends on the context). $\mathbb{P}_n$ denotes the empirical measure of the random vectors $\{Z_{ni}, 1 \leq i \leq n\}$ and $\mathbb{E}_{nN}[\cdot]$ denotes the empirical average over the first $N \leq n$ random vectors (ordered by their indices) distributed according to the empirical measure $\mathbb{P}_n$, i.e. $\mathbb{E}_{nN}[f] := \mathbb{E}_{nN}[f(Z_{ni})] = N^{-1}\sum_{i=1}^{N} f(Z_{ni})$. In addition, we define $\bar{\mathrm{E}}_{nN}[f] = \mathbb{E}_{nN}[\mathrm{E}[f]]$ and $\mathbb{G}_{nN}(f) = \sqrt{N}(\mathbb{E}_{nN}[f] - \bar{\mathrm{E}}_{nN}[f])$, and we denote the symmetrized process by $\mathbb{G}^{\circ}_{nN}(f)$. For $r \geq 1$ we denote the $L^r(\mathbb{P}_n)$-norm by $\|f\|_{\mathbb{P}_n,r} = (\mathbb{E}_{nn}[|f|^r])^{1/r}$. We write $\ell^{\infty}(\mathscr{T})$ for the set of all uniformly bounded real-valued functions on $\mathscr{T} \subset (0,1)$.

3.2 Framework

In this section we lay out a general framework for misspecified quantile regression processes in high dimensions, provide regularity conditions, and introduce the notion of τ-quantile sublevel sets which provide intuitive sufficient conditions under which the ℓ_1-penalized misspecified quantile regression estimates are consistent for the best sparse approximation of the true CQF.

3.2.1 Setting

We consider a triangular array $\{(Y_n, X_n), (Y_{ni}, X_{ni}), 1 \leq i \leq n\}$ of row-wise independent random vectors, where $Y_n \in \mathbb{R}$ is a continuous response variable, $X_n \in \mathbb{R}^d$ is a vector of predictor variables and the pair (Y_n, X_n) has joint distribution F_n. The dimension d of the predictor variables may be much larger than the sample size n; e.g. $d = O(n^b)$ for some $b > 0$, and the joint distribution F_n may depend on the sample size n as well. Our object of interest is the conditional quantile function (CQF) of Y_n given X_n for a set of quantile levels $\tau \in \mathscr{T} \subset (0,1)$

$$Q_{Y_n}(\tau|X_n) = \inf\left\{y : F_{Y_n|X_n}(y|X_n) \geq \tau\right\}, \tag{3.1}$$

where $F_{Y_n|X_n}(\cdot|X_n)$ is the conditional distribution function of Y_n given X_n. We do not impose any specific structural relation on the predictors X_n and the CQF (such as linearity in X_n or else), but we do assume that a small number of predictor variables suffices to capture (most) of the behavior of the CQF across all quantile levels $\tau \in \mathscr{T}$. To make this more precise, let us introduce the population quantile regression vector $\theta^*_n(\tau)$ at quantile level τ as the

solution to

$$\min_{\theta\in\mathbb{R}^d} \bar{\mathrm{E}}_{nn}[\rho_\tau(Y_{ni} - X_{ni}'\theta) - \rho_\tau(Y_{ni} - Q_{Y_n}(\tau|X_n))], \tag{3.2}$$

and its support set as

$$S_n^*(\tau) = \Big\{ j \in \{1,\ldots,d\} : |\theta_n^{*(j)}(\tau)| > 0 \Big\}. \tag{3.3}$$

We define the maximal cardinality of the support set $|S_n^*| = \sup_{\tau\in\mathscr{T}} |S_n^*(\tau)|$ and call it the "(maximal) sparsity level" or the "(maximal) size of the best linear approximation to the truth". In what follows, we always assume that $|S_n^*|$ is much smaller than the sample size n. Under this assumption a natural estimator for $\theta_n^*(\tau)$ is the solution $\hat{\theta}_{n,\lambda_n}(\tau)$ to the ℓ_1-penalized quantile regression problem,

$$\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{nn}[\rho_\tau(Y_{ni} - X_{ni}'\theta)] + \lambda_n(\tau) \sum_{j=1}^{d} |\theta_j|, \tag{3.4}$$

where $\lambda_n(\tau) > 0$ is a quantile level specific penalty parameter. In above displays, $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$ denotes the quantile loss function introduced by Koenker and Bassett (1978). Since we are interested in estimating the CQF for a set of quantile levels $\mathscr{T} \subset (0,1)$ we solve the problem (3.4) for all $\tau \in \mathscr{T}$ and obtain the quantile regression process

$$\hat{\theta}_{n,\lambda_n}(\cdot) = \{\hat{\theta}_{n,\lambda_n}(\tau) : \tau \in \mathscr{T}\}. \tag{3.5}$$

For a comprehensive analysis of the asymptotic and finite sample properties of the ℓ_1-penalized quantile regression process when the true CQF is a linear function of the predictor variables X_n we refer to the seminal work by Belloni and Chernozhukov (2011). Among other things, they establish weak consistency of the quantile regression process $\{\hat{\theta}_{n,\lambda_n}(\tau), \tau \in \mathscr{T}\}$ in the ℓ_2- and $L_2(\mathbb{P}_n)$-norm and weak consistency of the quantile regression function $\widehat{Q}_{Y_n|X_n}(\tau|X) = X'\hat{\theta}_{n,\lambda_n}(\tau)$ in the $L_2(P)$-norm uniformly over all $\tau \in \mathscr{T}$. They also show weak model selection consistency under beta-min-type conditions, and analyze the properties of the refitted post-lasso estimator.

3.2.2 Restricted cone property and quantile sublevel sets

Let $S \subseteq \{1,\dots,d\}$ be a set of indices and $c \geq 0$ be a non-negative constant. Define the $C(S,c)$-cone by

$$C(S,c) := \left\{\delta \in \mathbb{R}^d : \|\delta_{S^c}\|_1 \leq c\|\delta_S\|_1,\ \|\delta_S\|_0 \leq n\right\}. \tag{3.6}$$

These cones are key quantities in the analysis of ℓ_1-penalized estimators since under suitable conditions the solution to an ℓ_1-penalized M-estimation problem lies in such a cone (e.g. Candes and Tao, 2007; Bickel et al., 2009; Belloni and Chernozhukov, 2011, 2013). Leveraging this fact, it is then possible to develop an asymptotic theory for the penalized estimator. The following lemma shows that a similar result is also true for the ℓ_1-penalized misspecified quantile regression estimate defined in eq. (3.4).

Lemma 3.1 (Restricted Cone Property). *Let $\mathscr{T}$ be a compact subset of $(0,1)$. Let $c_0 > 1$ and set $\bar{c} = \frac{c_0+1}{c_0-1}$. Suppose that for all $\tau \in \mathscr{T}$,*

$$\lambda_n(\tau) \geq c_0 \left\|\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}]\right\|_\infty \quad a.s.$$

Then, for all $n > N_0$ and all $\tau \in \mathscr{T}$,

$$\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau) \in C(S_n^*(\tau),\bar{c}) \quad a.s.$$

Remark 3.1. *The main theoretical results in the next section rely to a large extend on this lemma and the technical Lemma 3.2 which combines convexity arguments and the special geometry of the $C(S,c)$-cone in order to bound suprema of (quantile regression specific) empirical processes indexed by function classes defined on $C(S,c)$-cones.*

Above Lemma 3.1 reveals that the choice of penalty parameters $\{\lambda_n(\tau) : \tau \in \mathscr{T}\}$ is crucial. In the case of correctly specified quantile regression processes Belloni and Chernozhukov (2011) and Koenker (2011) propose a simulation-based approach to finding a penalty level $\{\lambda_n(\tau) : \tau \in \mathscr{T}\}$ that satisfies the condition in Lemma 3.1 with high probability. However, their approach relies on the pivotal properties of the gradient of the quantile regression loss function when evaluated at the true CQF. There simulation-based approach is infeasible if the regression model is misspecified. Kato (2011) explores the case of asymptotically vanishing misspecification and finds that under certain conditions Belloni and Chernozhukov's (2011) choice of the penalty parameter still leads to consistent estimates. However, since we work under the assumption of persistent (i.e. non-vanishing) misspecification his result is of little practical use to us.

In the following, we introduce a concept that allows us to gain at least some theoretical inside in the asymptotic behavior of "good" penalty parameters in the case of persistent misspecification. For each $\tau \in \mathscr{T}$ define the "τ-quantile sublevel set" as

$$\left\{(Y,X) \in \mathbb{R} \times \mathbb{R}^d : Y \leq X'\theta_n^*(\tau)\right\}. \tag{3.7}$$

The set in above display contains all pairs $(Y,X) \in \mathbb{R} \times \mathbb{R}^d$ which lie on or below the best linear approximation $X'\theta_n^*(\tau)$ to the true CQF at quantile level τ; hence the name. If the quantile regression process is correctly specified on $\mathscr{T}$, i.e. $X'\theta_n^*(\tau) = Q_Y(\tau|X)$ for all $\tau \in \mathscr{T}$, then, by monotonicity of the quantile function, the τ-quantile sublevel sets are linearly ordered by inclusion, i.e.

$$\left\{(Y,X) \in \mathbb{R} \times \mathbb{R}^d : Y \leq X'\theta_n^*(\tau_1)\right\} \subseteq \left\{(Y,X) \in \mathbb{R} \times \mathbb{R}^d : Y \leq X'\theta_n^*(\tau_2)\right\} \quad \text{whenever } \tau_1 \leq \tau_2,\ \tau_1, \tau_2 \in \mathscr{T}. \tag{3.8}$$

If the true CQF is not linear in the predictor variables, this ordering does not hold in general and the extent to which the ordering fails is a natural measure for the severeness of the misspecification of the linear approximation of the true quantile regression process on $\mathscr{T}$. Clearly, if the ordering in eq. (3.8) holds for a set $\mathscr{T}$ of quantiles, the collection of τ-quantile sublevel sets forms a VC-class of sets with VC-index 2, while its VC-index can range from 2 to d if the ordering is violated (e.g. van der Vaart and Wellner, 1996, Lemma 2.6.16). Given these considerations it is immediate that for correctly specified quantile regression processes with high probability $\|\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}]\|_\infty \asymp \left(\frac{\log d}{n}\right)^{1/2}$. However, if the quantile regression process is misspecified there exists a positive constants $C > 0$ such that

$$C^{-1}\left(\frac{\log d}{n}\right)^{1/2} \lesssim \left\|\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}]\right\|_\infty \lesssim C\left(\frac{d\log d}{n}\right)^{1/2} \tag{3.9}$$

with high probability. Since $d \gg n$ the bound on the right hand side diverges to infinity. To establish consistency results it is therefore necessary to impose conditions on the rate at which the VC-index $V_{\mathscr{L}}$ of the collection

$$\mathscr{L} := \left\{\left\{(Y,X) \in \mathbb{R} \times \mathbb{R}^d : Y \leq X'\theta_n^*(\tau)\right\} : \tau \in \mathscr{T}\right\} \tag{3.10}$$

can grow as d and n diverge. In other words, one has to impose conditions on how closely the collection of best linear approximations $\{X'\theta_n^*(\tau) : \tau \in \mathscr{T}\}$ to the true quantile regression process on $\mathscr{T}$ resembles a collection of true quantile functions $Q_{Y_n|X_n}(\tau|X)$ in terms of satisfying the monotonicity relation in eq. (3.8). Taking the collection of τ-quantile sublevel

sets as the starting point to characterize process misspecification enables us to describe the degree of misspecification for the entire process without having to describe the precise misspecification at each quantile level.

Remark 3.2. *Note that the dimensionality of the vector of regression coefficients $\theta_n^*(\tau)$ does not determine the VC-index $V_{\mathscr{L}}$ of the collection of quantile sublevel sets. It is easy to see that the VC-index $V_{\mathscr{L}}$ is small if this map is bounded variation, and $V_{\mathscr{L}} = 2$ if the map is monotonically increasing. Moreover, $2 \leq V_{\mathscr{L}} \leq |\cup_{\tau\in\mathscr{T}} S_n^*(\tau)|$.*

Remark 3.3. *Invoking Sudakov's minoration for Gaussian variables (e.g. Ledoux and Talagrand, 1996, Theorem 3.18) it is easy to see if the data are i.i.d. Gaussian, then $\lambda_n(\tau) \gtrsim \left(\frac{\log d}{n}\right)^{1/2}$ uniformly for all $\tau \in \mathscr{T}$. Thus, from above remark we conclude that if the sparsity levels $S_n^*(\tau)$ satisfy $|\cup_{\tau\in\mathscr{T}} S_n^*(\tau)| \lesssim \log d$, a "good" penalty parameter should be of the order $\lambda_n(\tau) \asymp \left(\frac{\log d}{n}\right)^{1/2}$. We conjecture that Sudakov's minoration for Rademacher variables (e.g. Ledoux and Talagrand, 1996, Theorem 4.15) can be used to establish such a result for more general random variables.*

Remark 3.4. *One may argue that the VC-index of the collection of sets $\mathscr{L}$ does not measure an approximation (or misspecification) bias but rather the variance of the quantile regression score function. However, this variance is an increasing function in the variability of the map $\tau \mapsto X'\theta_n^*(\tau)$, i.e. the VC-index $V_{\mathscr{L}}$.*

3.2.3 Conditions

We introduce several mild conditions on the (joint) distribution of response Y_n and predictor variables X_n. These conditions are the high-dimensional analogues to the assumptions necessary to establish uniform-in-model consistency and Bahadur representation in the previous chapter.

We begin with an adaptation of the concept of restricted eigenvalues to the quantile regression process setting: We define the (S,c)-restricted minimum eigenvalue of the matrix of weighted second moments of the predictor variables by

$$\bar{\varphi}_{\min}(S,c) = \inf_{\tau\in\mathscr{T}} \inf_{\delta\in C(S,c)\cap B(0,1)} \delta' \bar{\mathrm{E}}_{nn}[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau)|X_{ni})X_{ni}X_{ni}']\delta, \tag{3.11}$$

and the (S,c)-restricted maximum eigenvalue of the matrix of second moments of the predictor variables by

$$\bar{\varphi}_{\max}(S,c) = \max_{\delta\in C(S,c)\cap B(0,1)} \delta' \bar{\mathrm{E}}_{nn}[X_{ni}X_{ni}']\delta < \infty. \tag{3.12}$$

Equipped with these definitions we can now state the regularity conditions.

(H1) *The data $\{(Y_{ni}, X_{ni}), 1 \leq i \leq n\}$, $(Y_{ni}, X_{ni}) \in \mathbb{R} \times \mathbb{R}^d$ are row-wise independent random vectors with distribution F_{ni}. The dimension of the predictor variables satisfies $d = O(n^b)$, for some $b > 0$. The distribution F_{ni} may change with the sample size n and index $i \leq n$.*

(H2) *The conditional density $f_{Y_{ni}|X_{ni}}$ of Y_{ni} given predictor variables X_{ni} is uniformly bounded from above, i.e. there exits $f_+ < \infty$ such that for all $n \in \mathbb{N}$,*

$$\max_{M:|M|\leq n} \max_{i\leq n} \sup_{a\in\mathbb{R}} \sup_{x\in\mathbb{R}^d} \left| f_{Y_{ni}|X_{ni,M}}(a|x) \right| \leq f_+.$$

(H3) *The conditional density $f_{Y_{ni}|X_{ni,M}}$ of Y_{ni} given the predictor variable $X_{ni,M}$, is Hölder continuous with exponent $\alpha \in \left[\frac{1}{2}, 1\right]$, i.e. there exists a constant $f_H > 0$ such that all $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$,*

$$\max_{M:|M|\leq n} \max_{i\leq n} \sup_{x\in\mathbb{R}^{|M|}} \left| f_{Y_{ni}|X_{ni,M}}(a|x) - f_{Y_{ni}|X_{ni,M}}(b|x) \right| \leq f_H |a-b|^\alpha.$$

(H4) *The predictors X_{ni} are vectors of random variables with finite $4+\delta$ moment for some $\delta > 0$, and there exists an absolute constant $\mu_4 < \infty$ such that for all $n \in \mathbb{N}$,*

$$\max_{\|u\|_1\leq 1} \mathrm{E}\left[\max_{i\leq n} |X'_{ni}u|^{4+\delta} \right]^{1/(4+\delta)} \leq \mu_4.$$

(H5) *The maximum eigenvalue of the matrix of second moments of the predictor variables X_{ni} is bounded from above uniformly over all models $M \subseteq \{1, \ldots, d\}$ with dimension at most n, i.e.*

$$\left(\max_{M:|M|\leq n} \sup_{\|u\|_2=1} u' \bar{\mathrm{E}}_{nn}\left[X_{ni,M} X'_{ni,M} \right] u \right) \vee 1 \leq \bar{\varphi}_{\max}(n).$$

(H6) *The minimum eigenvalue of the matrix of weighted second moments of the predictor variables $X_{n,M}$, is bounded from below uniformly over all models $M \subseteq \{1, \ldots, d\}$ with dimension at most n, i.e.*

$$\inf_{\tau\in\mathscr{T}} \left(\min_{M:|M|\leq n} \sup_{\|u\|_2=1} u' D_{n,M}(\tau) u \right) > \bar{\varphi}_{\min}(n).$$

Conditions (H1) – (H6) impose mild assumptions on the distribution of response and predictor variables. They are the almost identical to the ones of the low dimensional setting within which we derived the uniform-in-model Bahadur representation. The difference is that we only need to put constraints on models up to size n. We therefore refrain from discussing these conditions. We impose the following rate conditions on the restricted eigenvalues defined in eq. (3.11) and eq. (3.12).

(S1) *The collection $\mathscr{L}$ of quantile sublevel sets has finite VC-index $V_{\mathscr{L}} = O(\log d)$.*

(S2) *There exists an absolute constant $c > 0$ such that VC-index $V_{\mathscr{L}}$ and model size $|S_n^*|$ satisfy*

$$\bar{\kappa}^{\alpha-1}(S_n^*, c)\bar{\varphi}_{\max}^{1/2}(S_n^*, c)|S_n^*|^{(2\alpha-1)/4}\left(\frac{\log d + \log\log n}{n}\right)^{(2\alpha-1)/4} = O(1),$$

where $\bar{\kappa}(S,c) = \frac{\bar{\varphi}_{\max}(S,c)}{\bar{\varphi}_{\min}(S,c)}$ and $\alpha \in [\frac{1}{2}, 1]$ is the Hölder exponent of the conditional density $f_{Y_{ni}|X_{ni}}$.

Condition (S1) constrains the growth rate of the VC-index $V_{\mathscr{L}}$ which captures the process misspecification of the collection of approximate quantile regression functions $\{X'\theta_n^*(\tau) : \tau \in \mathscr{T}\}$. We have argued in Section 3.2.2 why this is necessary. Note that in the low-dimensional setting the VC-index $V_{\mathscr{L}}$ does show up as it enters only through the restricted cone property which is only relevant for ℓ_1-penalized estimates (see Lemma 3.1). Condition (S2) is the analogue to rate condition (R1) in the low-dimensional setting.

3.3 Main results

In this section we present an almost sure de-biased representation of the high-dimensional quantile regression process and provide an analysis of the theoretical properties of the misspecified quantile regression process after Lasso-based model selection.

3.3.1 A de-biased representation for the high-dimensional quantile regression process

Establishing asymptotic properties of estimates based on ℓ_1-penalized regression problems as the one in eq. (3.4) when the number of predictors d far exceeds the sample size n is difficult due to a persistent bias (e.g. Fan and Lv, 2011). One of the techniques developed to adjust for this bias, is the so-called de-biased (or de-sparsified) Lasso estimator proposed by Zhang

and Zhang (2013) and van de Geer et al. (2014). This estimator includes a correction term in the candidate estimator $\hat{\theta}_{n,\lambda_n}(\tau)$ which removes the penalization bias. Recently, Bradic and Kolar (2017) proposed such a de-biased estimator for the ℓ_1-penalized quantile regression process assuming that the true CQF is linear and data are i.i.d Sub-Gaussian.

Using our uniform-in-model Bahadur representation we obtain a similar but stronger result under much weaker assumptions. In fact, we do not just obtain a de-biased estimator, but a proper de-biased representation of the ℓ_1-penalized quantile regression process. The foundation of this result is the following strong consistency result of the ℓ_1-penalized quantile regression vector.

Theorem 3.1 (Strong Consistency of the ℓ_1-Penalized Quantile Regression Vector). *Suppose that Assumptions (H1) – (H6) and (S1) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$ and $\bar{c} > 0$. Then, there exist $c_0, C_\lambda, N_0 > 0$ such that for all $n > N_0$ and all $\lambda_n(\tau) = \lambda_n \geq C_\lambda \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n}\right)^{1/2}$,*

$$\sup_{\tau\in\mathscr{T}} \|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\|_2 \leq c_0 \left(\frac{\bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}(S_n^*, \bar{c})} \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n} \right)^{1/2} \bigvee \frac{|S_n^*|^{1/2} \lambda_n}{\bar{\varphi}_{\min}(S_n^*, \bar{c})} \right) \quad a.s.$$

Remark 3.5. *Observe that the penalty level λ_n has to be chosen such that it dominates the model degree of misspecification measured by the VC-index $V_{\mathscr{L}}$ of the collection of quantile sublevel sets. For a discussion how and under what conditions this might be achievable we refer to Kato (2011), who discussed these questions in the context of Group Lasso-penalized quantile regression processes.*

Using above consistency result we can establish the following de-biased representation of potentially misspecified quantile regression processes in high dimensions.

Theorem 3.2 (Strong De-biased Representation of the ℓ_1-Penalized Quantile Regression Vector). *Suppose that Assumptions (H1) – (H6) and (S1) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$ and $\bar{c} > 0$. There exist $c_0, C_\lambda, N_0 > 0$ such that for all $n > N_0$, $\tau \in \mathscr{T}$, and $\lambda_n(\tau) = \lambda_n \geq C_\lambda \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n}\right)^{1/2}$,*

$$\begin{aligned} &\hat{\theta}_{n,\lambda_n}(\tau) + \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)^{-1} \mathbb{E}_{nn}\left[\varphi_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))X_{ni}\right] \\ &= \theta_n^*(\tau) + \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)^{-1} \mathbb{E}_{nn}\left[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}\right] + r_n(\tau), \end{aligned}$$

and

$$\sup_{\tau\in\mathscr{T}} \|r_n(\tau)\|_\infty = O\left(\bar{\kappa}^{1/2}(S_n^*,\bar{c})\left(\frac{\log d + |S_n^*|\log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2}\left(|S_n^*|^{1/2}\lambda_n\right)^{1/2} \bigvee \bar{\kappa}^{1+\alpha}(S_n^*,\bar{c})\left(|S_n^*|^{1/2}\lambda_n\right)^{1+\alpha}\right) \quad a.s.,$$

and for any fixed $u\in\mathbb{R}^d$, $\|u\|_2 = 1$,

$$\sup_{\tau\in\mathscr{T}} |r_n(\tau)'u| = O\left(\bar{\kappa}^{1/2}(S_n^*,\bar{c})\bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})\left(\frac{|S_n^*|\log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2}\left(|S_n^*|^{1/2}\lambda_n\right)^{1/2} \bigvee \bar{\kappa}^{1+\alpha}(S_n^*,\bar{c})\bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})\left(|S_n^*|^{1/2}\lambda_n\right)^{1+\alpha}\right) \quad a.s.$$

If also Assumption (S2) holds and $\lambda_n \asymp \left(\frac{V_{\mathscr{L}}+\log d+\log\log n}{n}\right)^{1/2}$, *then*

$$\sup_{\tau\in\mathscr{T}} \|r_n(\tau)\|_\infty \vee \frac{\sup_{\tau\in\mathscr{T}}|r_n(\tau)'u|}{\bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})} = O\left(\bar{\kappa}^2(S_n^*,\bar{c})|S_n^*|^{3/4}\left(\frac{\log d+\log\log n}{n}\right)^{3/4}\right) \quad a.s.$$

This result improves on the one obtained by Bradic and Kolar (2017) in several ways. First, under Assumption (S2) our representation achieves the known optimal rate (exponent $3/4$) on the remainder term even if the data are non-identically distributed and the predictors follow a heavy-tailed distribution. Since Bradic and Kolar (2017) derived their result from the weak consistency result in Belloni and Chernozhukov (2011) they impose the same (or even more stringent) boundedness and distributional assumptions. Second, our representation does not only hold component-wise in the ℓ_∞-norm, but also for the projection onto the unit-sphere. In the context of hypothesis testing this allows for more interesting contrasts. Lastly, our representation holds almost surely and not just with high probability which opens the door for applications in predictive risk estimation as in Giessing and He (2018).

Remark 3.6. *This representation can be turned into a de-biased estimator by plugging in a consistent estimate of* $\left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)^{-1}$. *We refer to Bradic and Kolar (2017) for one possible estimator.*

We conclude this section with an easy corollary on the weak convergence of the de-biased estimator to a Gaussian limit process.

Corollary 3.1. *Suppose that the conditions of Theorem 3.2 are met. In addition, suppose that $F_{ni} = F$ for all $i, n \in \mathbb{N}$ and that $|S_n^*|^3(\log|S_n^*|)^3 = o(n)$. Let $\mathscr{T}$ be a compact subset of $(0,1)$. Let $u \in \mathscr{S}^d$ and*

$$\sup_{\tau\in\mathscr{T}} \left\| \widehat{Q}_n(\tau) - \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau)) X_{ni}X_{ni}' \right] \right)^{-1} \right\|_{op} \to 0.$$

Then,

$$n^{1/2}u'\left(\hat{\theta}_{n,\lambda_n}(\cdot) - \theta_n^*(\cdot) + \widehat{Q}_n(\cdot)\mathbb{E}_{nn}\left[\varphi_{(\cdot)}(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\cdot))X_{ni} \right] \right) \rightsquigarrow \mathbb{G}(\cdot) \quad in \quad \ell^\infty(\mathscr{T}),$$

where $\mathbb{G}(\cdot)$ is a centered Gaussian process with covariance function

$$\begin{aligned}\Sigma(\tau_1,\tau_2;u) &= \left(\tau_1 \wedge \tau_2 - \tau_1\tau_2\right) \\ &\times u'\left(\bar{\mathrm{E}}_n\left[f_{Y_i|X_i}(X_i'\theta_n^*(\tau_1))X_iX_i' \right] \right)^{-1} \mathrm{E}[XX']\left(\bar{\mathrm{E}}_n\left[f_{Y_i|X_i}(X_i'\theta_n^*(\tau_2))X_iX_i' \right] \right)^{-1} u.\end{aligned}$$

Remark 3.7. *An analogous result holds for the process*

$$n^{1/2}u'\widehat{D}_n(\tau)\left(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\right) + \mathbb{E}_{nn}\left[\varphi_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))X_{ni}\right], \quad \tau \in \mathscr{T},$$

where $\widehat{D}_n(\tau)$ is consistent estimate of $\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}' \right]$. Note that $\widehat{D}_n(\tau)$ is often easier to construct than the estimate of the inverse, $\widehat{Q}_n(\tau)$. This is yet another improvement of our representation over Bradic and Kolar's (2017) results.

Remark 3.8. *We leave a more comprehensive exploration of these results, including applications to hypothesis testing, to future research.*

3.3.2 Theoretical properties of the misspecified quantile regression estimate post-Lasso selection

In this section we provide theoretical results pertaining to the post-Lasso quantile regression estimator under model misspecification. In particular, we show how misspecification of the quantile regression process, measured by the VC-index $V_{\mathscr{L}}$ of the collection τ-quantile sublevel sets, impacts empirical sparsity and consistency of the post-Lasso estimate.

The theoretical properties of post-Lasso estimators have already been studied by several authors. For example, Belloni and Chernozhukov (2011) provided a comprehensive analysis

of the post-Lasso estimator for correctly specified quantile regression models, and Belloni and Chernozhukov (2013) developed a more general theory for post-Lasso least squares estimators covering misspecified models. In general, our results match well with their theoretical findings and there are not many new insights to gain. However, since we base our analysis on the strong uniform-in-model Bahadur representation for misspecified quantile regression processes our proofs differ from theirs. We obtain slightly stronger results under weaker regularity conditions.

We denote the model selected by solving the Lasso problem in eq. (3.4) by

$$\widehat{S}_{n,\lambda_n}(\tau) = \left\{ j \in \{1,\dots,d\} : |\hat{\theta}^{(j)}_{n,\lambda_n}(\tau)| > 0 \right\}, \ \tau \in \mathscr{T}.$$

Based on $\widehat{S}_{n,\lambda_n}(\tau)$ we define the post-Lasso quantile regression estimator $\tilde{\theta}_n(\tau)$ as the solution to

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{nn}[\rho_\tau(Y_{ni} - X_{ni}'\theta)] \ \text{ s.t. } \ \theta^{(j)} = 0 \ \text{ for each } j \notin \widehat{S}_{n,\lambda_n}(\tau), \ \ \tau \in \mathscr{T}. \tag{3.13}$$

The first result in this section shows that as in the case of a correctly specified quantile regression model (e.g. Belloni and Chernozhukov, 2011, Theorem 3) the sparsity level of the ℓ_1-penalized quantile regression estimate under misspecification can be controlled by choosing the penalty level high enough. Notably, the sparsity control is always affected by the degree of misspecification measured in terms of $V_{\mathscr{L}}$.

Theorem 3.3 (Empirical Sparsity). *Suppose that Assumptions (H1) – (H6) and (S1) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$, $\bar{c} > 0$, and $C_\lambda > 0$ as in Theorem 3.1. Let*

$$\lambda_n \geq C_\lambda \, \bar{\varphi}^{1/2}_{\max}(|S^*_n|, \bar{c}) \, \bar{\varphi}^{3/2}_{\max}(n, \bar{c}) \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n} \right)^{1/2},$$

and suppose that

$$\frac{\bar{\varphi}^{1/2}_{\max}(|S^*_n|, \bar{c})}{\bar{\varphi}^{3/2}_{\max}(n, \bar{c})} \gtrsim |S^*_n|^{1/2} \left(\frac{\log d + \log\log n}{n} \right)^{1/2}.$$

Then, there exist absolute constants $c_0, N_0 > 0$ such that for all $n > N_0$ and all $\tau \in \mathscr{T}$,

$$|\widehat{S}_n(\tau)| \leq |S^*_n| \times c_0 \bar{\varphi}^2_{\max}(n, \bar{c}) + c_0 \bar{\varphi}^2_{\max}(n, \bar{c}) \frac{V_{\mathscr{L}}}{n\lambda_n^2} \quad a.s.$$

Theorem 3.3 shows that it is possible to control the empirical sparsity at (or around) the true sparsity level $|S^*_n|$ by choosing a large penalty level. Therefore, invoking the uniform-in-

model consistency in low dimensions we obtain the following post selection consistency result for the post-Lasso quantile regression estimator.

Theorem 3.4 (Post-Selection Consistency). *Suppose that Assumptions (H1) – (H6), (R1), and (S1) hold. Let $\mathscr{T}$ be a compact subset of $(0,1)$, $\bar{c} > 0$, and $C_\lambda > 0$ as in Theorem 3.1. Let $\lambda_n \geq C_\lambda \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n}\right)^{1/2}$. Then, there exist constants $c_0, c_1, N_0 > 0$ such that for all $n > N_0$ and all $\tau \in \mathscr{T}$,*

$$
\begin{aligned}
&\|\tilde{\theta}_n(\tau) - \theta_n^*(\tau)\|_2 \\
&\leq c_0 1\{S_n^*(\tau) \subseteq \widehat{S}_n(\tau)\} \\
&\times \frac{\bar{\varphi}_{\max}^{1/2}(\widehat{S}_n(\tau))}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau))} \left(\frac{|\widehat{m}_n(\tau)| \log(ed/|\widehat{m}_n(\tau)|^{1/2}) + |S_n^*(\tau)| + \log\log n}{n} \right)^{1/2} \\
&+ c_0 1\{S_n^*(\tau) \not\subseteq \widehat{S}_n(\tau)\} \times \frac{\bar{\varphi}_{\max}^{1/2}(\widehat{S}_n(\tau))}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau))} \left(\frac{|\widehat{S}_n(\tau)| \log(ed/|\widehat{S}_n(\tau)|^{1/2}) + \log\log n}{n} \right)^{1/2} \\
&+ c_1 1\{S_n^*(\tau) \not\subseteq \widehat{S}_n(\tau)\} \times \|\theta_n^*(\tau) - \hat{\theta}_{n,\lambda_n}(\tau)\|_2^{1/2} \\
&\times \left(\frac{|S_n^*|^{1/2} \lambda_n}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau) \cup S_n^*(\tau))} \bigvee \frac{\bar{\varphi}_{\max}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau) \cup S_n^*(\tau))} \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n} \right)^{1/2} \right)^{1/2},
\end{aligned}
$$

where $\widehat{m}_n(\tau) = \widehat{S}_n(\tau) \setminus S_n^*(\tau)$.

If also Assumption (S2) holds and $\frac{\lambda_n}{\bar{\varphi}_{\max}^{1/2}(|S_n^*|, \bar{c})\, \bar{\varphi}_{\max}^{3/2}(n, \bar{c})} \asymp \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n}\right)^{1/2}$, then, almost surely,

$$
\begin{aligned}
&\sup_{\tau \in \mathscr{T}} \|\tilde{\theta}_n(\tau) - \theta_n^*(\tau)\|_2 \\
&= O\left(\left(\frac{\bar{\varphi}_{\max}(n, \bar{c}) \bar{\varphi}_{\max}^{1/2}(n)}{\bar{\varphi}_{\min}(n)} + \frac{\bar{\varphi}_{\max}^{1/2}(|S_n^*|, \bar{c}) \bar{\varphi}_{\max}^{3/4}(n, \bar{c})}{\bar{\varphi}_{\min}^{1/2}(|S^*|, \bar{c}) \bar{\varphi}_{\min}^{1/2}(n)} \right) \left(\frac{|S_n^*| \log d + |S_n^*| \log\log n}{n} \right)^{1/2} \right).
\end{aligned}
$$

We conclude this section with the a corollary on the post-selection representation of the post-selection estimator which holds whenever the true support set $S_n^*(\tau)$ is contained in the estimated support set $\widehat{S}_n(\tau)$.

For $S \subset \{1, \ldots, d\}$ let $\iota_S : \mathbb{R}^{|S|} \to \mathbb{R}^d$ be the map that embeds a lower dimensional vector $v \in \mathbb{R}^{|S|}$ into the higher dimensional ambient space $\mathbb{R}^d$, i.e.

$$
\left(\iota_S\{v\}\right)_j = \begin{cases} v_j & \text{for } j \in S \\ 0 & \text{otherwise.} \end{cases}
$$

Corollary 3.2 (Post-Selection Representation). *Suppose that Assumptions (H1) – (H6), (R1), and (S1) hold. Let $\bar{c} > 0$, $C_\lambda > 0$ as in Theorem 3.1, and $\lambda_n \geq C_\lambda \left(\frac{V_{\mathscr{L}} + \log d + \log\log n}{n}\right)^{1/2}$. If $S_n^*(\tau) \subseteq \widehat{S}_n(\tau)$ for all $\tau \in \mathscr{T} \subset (0,1)$, then there exist constants $c_0, c_1, N_0 > 0$ such that for all $n > N_0$ and all $\tau \in \mathscr{T}$,*

$$\begin{aligned}
&\tilde{\theta}_n(\tau) - \theta_n^*(\tau) \\
&\quad = \iota_{\widehat{S}_n(\tau)} \left\{ \left(\bar{\mathbb{E}}_{nn} \left[f_{Y_{ni}|X_{ni,\widehat{S}_n(\tau)}} \left(X'_{ni,\widehat{S}_n(\tau)} \theta^*_{n,\widehat{S}_n(\tau)}(\tau) | X_{ni,\widehat{S}_n(\tau)} \right) X_{ni,\widehat{S}_n(\tau)} X'_{ni,\widehat{S}_n(\tau)} \right] \right)^{-1} \right. \\
&\qquad\qquad \left. \mathbb{E}_{nn}[\varphi_\tau\left(Y_{ni} - X'_{ni,\widehat{S}_n(\tau)} \theta^*_{n,\widehat{S}_n(\tau)}(\tau)\right) X_{ni,\widehat{S}_n(\tau)}] \right\} \\
&\quad + r_n(\tau),
\end{aligned}$$

and

$$\|r_n(\tau)\|_2 \leq c_0 \bar{\kappa}^2(|\widehat{S}_n|) \left(\frac{|\widehat{m}_n| \log(ed/|\widehat{m}_n|^{1/2}) + |\widehat{S}_n| + \log\log n}{n} \right)^{3/4} \quad a.s.,$$

where $|\widehat{m}_n| = \max_{\tau \in \mathscr{T}} |\widehat{m}_n(\tau)|$, and $\widehat{m}_n(\tau) = \widehat{S}_n(\tau) \setminus S_n^*(\tau)$.

This result complements the de-biased representation from the previous section. It shows that if $S_n^*(\tau) \subseteq \widehat{S}_n(\tau)$, then the non-zero components of the estimated regression vector satisfy a (classical) strong Bahadur representation.

3.4 Technical results

In this section we collect technical auxiliary results needed to establish the main results. These auxiliary results are mostly generalizations of auxiliary lemmatas first established in low dimensions.

Lemma 3.2. *Let $\{X_{ni}, i \leq n\}$ be a triangular array of row-wise independent random vectors on $\mathbb{R}^d$, $S \subseteq \{1, \ldots, d\}$ be a set of indices, $m \in \{1, \ldots, d\}$, and $F : \mathbb{R} \to \mathbb{R}$ a real-valued map.*

1) Suppose that F is convex. Let $\varepsilon \in (0, \frac{1}{2}]$. There exist a set $\mathscr{M}_m^d \subset \mathscr{B}^d$ and an absolute constant $C > 0$ such that

$$|\mathscr{M}_m^d| \leq C \left(1 + \frac{2}{\varepsilon}\right)^{2(2+c)m} \left(\frac{ed}{m}\right)^{2(2+c)m},$$

$\|v\|_0 \le m$ *for all* $v \in \mathscr{M}_m^d$, *and*

$$\max_{S:|S|=m} \sup_{u \in C(S,c)} \mathbb{E}_{nn}\big(F(X'_{ni,S}u)\big) \le \max_{v \in \mathscr{M}_m^d} \sup_{u \in \mathscr{B}^d_{\mathrm{supp}(v)}(v,\varepsilon)} \mathbb{E}_{nn}\big(F(X'_{ni}u)\big).$$

2) *Suppose that* F *is linear. There exist a set* $\mathscr{M}_m^d \subset \mathscr{B}^d$ *and an absolute constant* $C > 0$ *such that* $|\mathscr{M}_m^d| \le C\left(\frac{5ed}{m}\right)^{2(2+c)m}$, $\|v\|_0 \le m$ *for all* $v \in \mathscr{M}_m^d$, *and*

$$\max_{S:|S|=m} \sup_{u \in C(S,c)} \mathbb{G}_{nn}\big(F(X'_{ni,S}u)\big) \le \max_{v \in \mathscr{M}_m^d} \mathbb{G}_{nn}\big(F(X'_{ni}v)\big).$$

Lemma 3.3. *Suppose that Assumptions (H1) – (H6) hold. Let* $\mathscr{T}$ *be a compact subset of* $(0,1)$, $\bar{c} = \frac{c_0+1}{c_0-1}$, *and* $r_n \in (0,1]$. *Define*

$$\mathscr{F}_n = \Big\{(Y,X) \mapsto \rho_\tau(Y - X'\theta) - \rho_\tau(Y - X'\theta_n^*(\tau)) : \\ \theta \in \mathbb{R}^d,\, \theta - \theta_n^*(\tau) \in C(S_n^*(\tau), \bar{c}) \cap \mathscr{B}^d(r_n),\, \tau \in \mathscr{T}\Big\}.$$

Then, there exist absolute constants $c_{12}, N_{12} > 0$ *such that for all* $n > N_{12}$,

$$\frac{1}{\sqrt{n}} \sup_{f \in \mathscr{F}_n} |\mathbb{G}_{nn}(f)| \le c_{12} \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})\, r_n \quad a.s.$$

For the next lemma we introduce the following new notation: For $n \in \mathbb{N}$ let $[n] = \{1, \dots, n\}$ and for $I \subseteq \mathbb{N}$, $|I| < \infty$, and $S_n \in \mathscr{F}_n/\mathscr{R}_n$ define

$$\mathbb{V}_{nN}(S_n; I) = \mathrm{E}\left[\sup_{f \in S_n} \frac{1}{N} \sum_{i \in I \cap [N \wedge n]} \Big(f(X_{ni}) - f(\widetilde{X}_{ni})\Big)^2 \mid (X_{n1}, \dots, X_{nn})\right],$$

and for $f \in S_n$,

$$\mathbb{G}_{nN}(f; I) = \frac{1}{\sqrt{N}} \sum_{i \in I \cap [N \wedge n]} \Big(f(X_{ni}) - \mathrm{E}[f(X_{ni})]\Big).$$

Lemma 3.4. *Let* $\{X_{ni}, i \le n\}$ *be a triangular array of row-wise independent random vectors on a measurable space* $\mathscr{X}$ *and let* $\mathscr{F}_n = \{f : \mathscr{X} \to \mathbb{R}\}$ *be classes of measurable functions with countable separants. Let* $\mathscr{R}_n \subseteq \mathscr{F}_n \times \mathscr{F}_n$ *be an equivalence relation on* $\mathscr{F}_n$ *such that the quotient set* $\mathscr{F}_n/\mathscr{R}_n$ *is countable. Let* ν_n *be a (sub)probability measure on* $\mathscr{F}_n/\mathscr{R}_n$. *For* $f \in \mathscr{F}_n$ *denote its corresponding equivalence class by* $[f]_{\mathscr{R}_n} = \{g \in \mathscr{F}_n : (f,g) \in \mathscr{R}_n\}$. *Let*

$\delta \in (0,1)$. *There exist* $c_0, c_1, N_0 > 0$ *such that for all* $n > N_0$, $f \in \mathscr{F}_n$ *and* $I \subseteq \{1,\dots,n\}$,

$$
\begin{aligned}
|\mathbb{G}_{nn}(f;I)| \leq\ & c_0 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f;I)|\Big] \\
& + c_1 \Big(\mathbb{V}_{nn}([f]_{\mathscr{R}_n};I) + \mathrm{E}[\mathbb{V}_{nn}([f]_{\mathscr{R}_n};I)]\Big)^{1/2} \\
& \times \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n + |I| \log(2en/|I|)\right)^{1/2} \quad a.s.
\end{aligned}
$$

Moreover, if each $S_n \in \mathscr{F}_n/\mathscr{R}_n$ *has an envelope function* $F_{S_n} : \mathscr{X} \to [1,\infty]$ *such that* $\mathrm{E}[\max_{i \leq n} F_{S_n}^4(X_{ni})] < \infty$. *Then, there exist* $c_0, c_1, c_2, N_0 > 0$ *such that for all* $n > N_0$ *and for all* $f \in \mathscr{F}_n$ *and* $I \subseteq \{1,\dots,n\}$, *almost surely,*

$$
\begin{aligned}
|\mathbb{G}_{nn}(f;I)| \leq\ & c_0 \mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f;I)|\Big] \\
& + c_1 \left(\mathrm{E}\Big[\sup_{f \in [f]_{\mathscr{R}_n}} \frac{1}{\sqrt{n}} |\mathbb{G}_{nn}(f^2;I)|\Big]\right)^{1/2} \\
& \times \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n + |I| \log(2en/|I|)\right)^{1/2} \\
& + c_2 \Big(\sup_{f \in [f]_{\mathscr{R}_n}} \bar{\mathrm{E}}_{nn}[f^2;I]\Big)^{1/2} \\
& \times \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n + |I| \log(2en/|I|)\right)^{1/2} \\
& + c_3 \left(\bar{\mathrm{E}}_{nn}[F_{[f]_{\mathscr{R}_n}}^4;I]\right)^{1/4} \\
& \times \left(\frac{\log\log n}{n}\right)^{1/4} \left(\log \frac{1}{\nu_n([f]_{\mathscr{R}_n})} + \log\log n + |I| \log(2en/|I|)\right)^{1/2}.
\end{aligned}
$$

3.5 Discussion

In this paper we provide theoretical results for inference on high-dimensional misspecified quantile regression processes. We establish almost sure consistency of the ℓ_1-penalized estimated quantile regression process, derive an almost sure de-biased representation for the high-dimensional quantile regression process, weak convergence of the de-biased estimator, and analyze theoretical properties of misspecified quantile regression estimates post-Lasso selection.

We have emphasized the theoretical results and have only given rough sketches of possible applications. The almost sure de-biased representation arguably provides the most

interesting opportunities for applications, including applications to quantile treatment effect estimation with high-dimensional covariates and goodness-of-fit testing for locally misspecified quantile regression index models. Moreover, the proof of the de-biased representation of the ℓ_1-penalized quantile regression process seems to be generalizable to other (robust) M-estimators.

We leave unanswered several important questions regrading practical implementations of our results. These questions include the problem of choosing a sensible penalty level under persistent misspecification; the estimation of the inverse of the high-dimensional weighted covariance matrix featuring in the almost sure de-biased representation. We leave these questions to future research.

3.6 Proofs

3.6.1 Proofs of Section 3.3

3.6.1.1 Proof of Lemma 3.1

Proof. The proof strategy is standard (e.g. Belloni and Chernozhukov, 2013, which establishes a similar result but for the least squares estimator). By optimality of $\hat{\theta}_{n,\lambda_n}(\tau)$ and assumption on $\lambda_n(\tau)$,

$$\begin{aligned}
0 &\geq \mathbb{E}_{nn}[\rho_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))] - \mathbb{E}_{nn}[\rho_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))] \\
&\quad + \lambda_n(\tau)\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1 - \lambda_n(\tau)\|\theta_n^*(\tau)\|_1 \\
&\geq -\mathbb{E}_{nn}[\varphi(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}']\big(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\big) \\
&\quad + \lambda_n(\tau)\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1 - \lambda_n(\tau)\|\theta_n^*(\tau)\|_1 \\
&\geq -\big\|\mathbb{E}_{nn}[\varphi(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}]\big\|_\infty \big\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\big\|_1 \\
&\quad + \lambda_n(\tau)\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1 - \lambda_n(\tau)\|\theta_n^*(\tau)\|_1 \\
&\geq -\frac{\lambda_n(\tau)}{c_0}\big\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\big\|_1 + \lambda_n(\tau)\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1 - \lambda_n(\tau)\|\theta_n^*(\tau)\|_1 \quad a.s.
\end{aligned}$$

Thus,

$$(1 + c_0^{-1})\big\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\big\|_1 \geq \big\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\big\|_1 + \|\hat{\theta}_{n,\lambda_n}(\tau)\|_1 - \|\theta_n^*(\tau)\|_1 \quad a.s. \tag{3.14}$$

Elementary calculations

$$\left\|\hat{\theta}_{n,\lambda_n}(\tau)-\theta_n^*(\tau)\right\|_1+\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1-\|\theta_n^*(\tau)\|_1\geq 2\|\hat{\theta}_{n,\lambda_n,S^*(\tau)^c}(\tau)\|_1 \quad a.s. \tag{3.15}$$

Combining inequalities (3.14) and (3.15),

$$\frac{c_0+1}{c_0-1}\left\|\hat{\theta}_{n,\lambda_n,S^*(\tau)}(\tau)-\theta_n^*(\tau)\right\|_1\geq\|\hat{\theta}_{n,\lambda_n,S^*(\tau)^c}(\tau)\|_1 \quad a.s.$$

Moreover, the regularized quantile regression problem can be recast as a linear programming problem and it is known that its solution satisfies $\|\hat{\theta}_{n,\lambda_n,S^*(\tau)}(\tau)\|_0\leq n$ (e.g. Koenker, 2005). □

3.6.1.2 Proof of Theorem 3.1

Proof. **Step 1. Verification of cone condition.** We show that there exist constants $C_\lambda, N_0 > 0$ such for all $n > N_0$,

$$\left\|\mathbb{E}_{nn}[\varphi_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))X_{ni}]\right\|_\infty\leq C_\lambda\left(\frac{V_{\mathscr{L}}+\log d+\log\log n}{n}\right)^{1/2} \quad a.s. \tag{3.16}$$

Thus, by Lemma 3.1 $\hat{\theta}_{n,\lambda_n}(\tau)$ satisfies the almost sure cone condition whenever λ_n is at least as large as right side of above display.

Consider the following function class:

$$\mathscr{F}=\left\{(Y,X)\mapsto\left(\tau-1\left\{Y\leq X'\theta_n^*(\tau)\right\}\right)X^{(j)}:\ \tau\in\mathscr{T},\ j\in\{1,\ldots,d\}\right\}.$$

Define the equivalence relation $\mathscr{R}\subseteq\mathscr{F}\times\mathscr{F}$ by

$$(f_{j,\tau_1}\,f_{k,\tau_2})\in\mathscr{R}\iff\{j=k\},$$

and the probability measure $\nu:\sigma(\mathscr{F}/\mathscr{R})\to[0,1]$ by $\nu(\emptyset)=0$ and

$$\nu(P_j)=d^{-1},\ \ P_j\in\mathscr{F}/\mathscr{R}.$$

By assumption (D1) each subclass $P_j\in\mathscr{F}/\mathscr{R}$ is a VC-subgraph class of functions with VC-index at most $V_{\mathscr{L}}+4$ and envelop function $G_j(X)=|X^{(j)}|\vee 1$.

By Theorem 2.4 there exists $N_0 > 0$ such that for all $n > N_0$ and all $f_{j,\tau} \in \mathscr{F}$ we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathbb{G}_{nn}(f_{j,\tau}) \leq{}& \frac{7}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{j,\tau}]_{\mathscr{R}}}|\mathbb{G}_{nn}(f)|\Big] \\
&+ 69\left(\mathrm{E}\Big[\sup_{f\in[f_{j,\tau}]_{\mathscr{R}}}\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]\right)^{1/2}\left(\frac{\log d+\log\log n}{n}\right)^{1/2} \\
&+ 149\left(\sup_{f\in[f_{j,\tau}]_{\mathscr{R}}}\bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2}\left(\frac{\log d+\log\log n}{n}\right)^{1/2} \\
&+ 223\left(\bar{\mathrm{E}}_{nn}[G_j^4]\right)^{1/4}\left(\frac{\log d+\log\log n}{n}\right)^{1/2}\left(\frac{\log\log n}{n}\right)^{1/4} \\
&\leq c_{12}\left(\frac{V_{\mathscr{L}}}{n}\right)^{1/2} + c_{12}\left(\frac{\log d+\log\log n}{n}\right)^{1/2} \\
&+ c_{12}\left(\frac{\log d+\log\log n}{n}\right)^{1/2}\left(\frac{\log\log n}{n}\right)^{1/4} \quad a.s.,
\end{aligned} \tag{3.17}$$

where the bounds on $\frac{7}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{j,\tau}]_{\mathscr{R}_1}}|\mathbb{G}_{nn}(f)|\Big]$ and $69\left(\mathrm{E}\Big[\sup_{f\in[f_{j,\tau}]_{\mathscr{R}}}\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]\right)^{1/2}$ follow from eq. (2.41) and $c_{12} > 0$ depends on $\mathrm{E}_{nn}[|X_{ni}^{(j)}|^2]$ and $\mathrm{E}_{nn}[|X_{ni}^{(j)}|^4]$. Thus, there exist $C_\lambda, N_0 > 0$ such that for all $n > N_0$, eq. (3.16) holds.

Step 2. Lower bound on centered quantile loss function. We proceed as in Step 2 of the proof of Theorem 2.1. For $\tau \in \mathscr{T}$ and $R_n(\lambda_n, S_n^*) > 0$ define

$$D_{n,\lambda_n}(\tau) = \{\theta \in \mathbb{R}^d : \|\theta - \theta_n^*(\tau)\|_2 = R_n(\lambda_n, S_n^*)\}$$

The general idea is similar to the proof idea of Theorem 2.2: Suppose that we have shown that there exists an $N_0 > 0$ such that for all $n > N_0$ and for all $\tau \in \mathscr{T}$ the centered regularized check loss evaluated at any point $\theta \in D_{n,\lambda_n}(\tau) \cap C(S_n^*(\tau), \bar{c})$ is positive. Since for all $\tau \in \mathscr{T}$ the centered regularized check loss is convex and negative, when evaluated at the minimizer $\hat{\theta}_{n,\lambda_n}(\tau)$, and by Step 1 we conclude that $\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\|_2 \leq R_n(\lambda_n, S_n^*)$ for all $\tau \in \mathscr{T}$.

Consider the map $\theta \mapsto \bar{\mathrm{E}}_{nn}\big[\rho_\tau\big(Y_{ni} - X_{ni}'\theta\big) - \rho_\tau\big(Y_{ni} - X_{ni}'\theta_n^*(\tau)\big)\big]$, $\theta \in D_{n,\lambda_n}(\tau) \cap C(S_n^*(\tau), \bar{c})$. Thus, by optimality of $\theta_n^*(\tau)$ and convexity a second-order Taylor expansion around $\theta_n^*(\tau)$ gives the following uniform lower bound

$$\bar{\mathrm{E}}_{nn}\big[\rho_\tau\big(Y_{ni} - X_{ni}'\theta\big) - \rho_\tau\big(Y_{ni} - X_{ni}'\theta_n^*(\tau)\big)\big] \geq \bar{\varphi}_{\min}(S_n^*, \bar{c})\|\theta - \theta_n^*(\tau)\|_2^2. \tag{3.18}$$

By eq. (3.18), Step 1, and Lemma 3.3 there exists an $N_1 \geq N_0$ such that for all $n > N_1$,

uniformly in $\tau \in \mathscr{T}$ and $\theta \in \partial D_n(\tau) \cap C(S_n^*(\tau), \bar{c})$,

$$
\begin{aligned}
&\mathbb{E}_{nn}\left[\rho_\tau\left(Y_{ni} - X_{ni}'\theta\right) - \rho_\tau\left(Y_{ni} - X_{ni}'\theta_n^*(\tau)\right)\right] + \lambda_n(\tau)\left(\|\theta\|_1 - \|\theta_n^*(\tau)\|_1\right) \\
&\geq \bar{\mathrm{E}}_{nn}\left[\rho_\tau\left(Y_{ni} - X_{ni}'\theta\right) - \rho_\tau\left(Y_{ni} - X_{ni}'\theta_n^*(\tau)\right)\right] \\
&\quad - \left|\frac{1}{\sqrt{n}}\mathbb{G}_{nn}\left(\rho_\tau\left(Y_{ni} - X_{ni}'\theta\right) - \rho_\tau\left(Y_{ni} - X_{ni}'\theta_n^*(\tau)\right)\right)\right| \\
&\quad - \lambda_n(\tau)(1+\bar{c}) \sum_{j \in S_n^*(\tau)} |\theta^{(j)} - \theta_n^{*(j)}| \\
&\geq \bar{\varphi}_{\min}(S_n^*, \bar{c})\|\theta - \theta_n^*(\tau)\|_2^2 \\
&\quad - c_9 \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})\|\theta - \theta_n^*(\tau)\|_2 \\
&\quad - \lambda_n(\tau)(1+\bar{c})|S_n^*|^{1/2}\|\theta - \theta_n^*(\tau)\|_2 \\
&= R_n(\lambda_n, S_n^*)\left(\bar{\varphi}_{\min}(S_n^*, \bar{c}) R_n(\lambda_n, S_n^*)\right. \\
&\quad \left. - c_9 \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c}) - \lambda_n(\tau)(1+\bar{c})|S_n^*|^{1/2}\right) \quad a.s.
\end{aligned}
\tag{3.19}
$$

To bound (3.19) away from 0 set for some large constant $c_{11} > 0$,

$$
R_n(\lambda_n, S_n^*) = c_{11}\left(\frac{\bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}(S_n^*, \bar{c})}\left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \vee \frac{|S_n^*|^{1/2}\lambda_n}{\bar{\varphi}_{\min}(S_n^*, \bar{c})}\right).
$$

This concludes the proof. □

3.6.1.3 Proof of Theorem 3.2

Proof. The subgradient of the objective function of the ℓ_1-penalized quantile regression problem cannot be minorized (asymptotically) by a quadratic function of the centered quantile regression vector. Therefore, we need to proceed differently than in the proof of Theorem 2.2: We use a simple Taylor approximation argument and combine it with the consistency result of Theorem 3.1.

This approach can be applied to any consistent estimator; however, the resulting representation does not always provide a linearization of the estimator. In the case of quantile regressions the non-linear term of the representation is negligible under additional assumptions on the growth rate of predictors versus sample size (e.g. Belloni et al., 2017, Lemma 34, p. 106) and would therefore lead to the same result as our Theorem 2.2 (the virtue of our

proof of Theorem 2.2 is that we do not need this assumption).

We only provide proof of the uniform bound on the remainder $r_n(\tau)$ in the ℓ_∞-norm. The proof of the bound on the projected remainder $|r_n(\tau)'u|$ is very similar.

Step 1. Expansion. Let $\{(\widetilde{Y}_{ni}, \widetilde{X}_{ni}), i \le n\}$ be an independent copy of the triangular array $\{(Y_{ni}, X_{ni}), i \le n\}$. Then,

$$
\begin{aligned}
&- \mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))X_{ni}] \\
&= -\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))X_{ni}] + \mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] \\
&\quad + \bar{\mathrm{E}}_{nn}[\varphi_\tau(\widetilde{Y}_{ni} - \widetilde{X}_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))\widetilde{X}_{ni}] - \bar{\mathrm{E}}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] \\
&\quad - \mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] + \bar{\mathrm{E}}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] \\
&\quad - \bar{\mathrm{E}}_{nn}[\varphi_\tau(\widetilde{Y}_{ni} - \widetilde{X}_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))\widetilde{X}_{ni}] \\
&\quad - \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)\left(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\right) \\
&\quad + \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)\left(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\right) \\
&= \mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] \\
&\quad + \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)\left(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\right) + r_{n,1}(\tau) + r_{n,2}(\tau),
\end{aligned}
$$

where

$$
\begin{aligned}
r_{n,1}(\tau) = &-\mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))X_{ni}] + \mathbb{E}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}] \\
&+ \bar{\mathrm{E}}_{nn}[\varphi_\tau(\widetilde{Y}_{ni} - \widetilde{X}_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))\widetilde{X}_{ni}] - \bar{\mathrm{E}}_{nn}[\varphi_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))X_{ni}],
\end{aligned}
$$

and

$$
\begin{aligned}
r_{n,2}(\tau) = &-\bar{\mathrm{E}}_{nn}[\varphi_\tau(\widetilde{Y}_{ni} - \widetilde{X}_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))\widetilde{X}_{ni}] \\
&- \left(\bar{\mathrm{E}}_{nn}\left[f_{Y_{ni}|X_{ni}}(X_{ni}'\theta_n^*(\tau))X_{ni}X_{ni}'\right]\right)\left(\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\right).
\end{aligned}
$$

Step 2. Bound on $r_{n,1}(\tau)$ in the ℓ_∞-norm. Set

$$
\begin{aligned}
R_n &\equiv R_n(\lambda_n, S_n^*) \\
&= c_0 \left(\frac{\bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}(S_n^*, \bar{c})}\left(\frac{|S_n^*|\log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \vee \frac{|S_n^*|^{1/2}\lambda_n}{\bar{\varphi}_{\min}(S_n^*, \bar{c})}\right),
\end{aligned}
$$

where $c_0 > 0$ is the constant from Theorem 3.1. Then, there exists $N_0 > 0$ such that for all

$n > N_0$ and all $\tau \in \mathscr{T}$,

$$\|\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau)\|_2 \leq R_n \quad a.s.$$

Let $\mathscr{M}_n^d\big(S_n^*(\tau)\big) \subset \mathscr{B}^d(R_n)$ be the countable finite set introduced in the proof of Lemma 3.3 and note for all $n > N_0$ and all $\tau \in \mathscr{T}$,

$$\hat{\theta}_{n,\lambda_n}(\tau) - \theta_n^*(\tau) \in C(S_n^*(\tau), \bar{c}) \cap \mathscr{B}^d(R_n) \quad a.s.$$

Further, observe that as in the proof of Lemma 2.7 the functions under consideration can be decomposed into the difference of two simple functions:

$$\begin{aligned}
&\varphi_\tau(Y - X'\hat{\theta}_{n,\lambda_n}(\tau))X - \varphi_\tau(Y - X'\theta_n^*(\tau))X \\
&\quad = 1\left\{X'\theta_n^*(\tau) < Y \leq X'\hat{\theta}_{n,\lambda_n}(\tau)\right\}X - 1\left\{X'\hat{\theta}_{n,\lambda_n}(\tau) < Y \leq X\theta_n^*(\tau)\right\}X
\end{aligned}$$

Thus, in order to bound $\|r_{n,1}(\tau)\|_\infty$ uniformly over $\tau \in \mathscr{T}$ it suffices to consider the following class of functions:

$$\begin{aligned}
\mathscr{F} = \Big\{&(Y,X) \mapsto 1\left\{X'\theta_1 < Y \leq X'\theta_2\right\}X^{(j)} : \\
&\theta_1, \theta_2 \in \mathbb{R}^d, (\theta_1 - \theta_2) \in \mathscr{M}_n^d\big(S_n^*(\tau)\big), \tau \in \mathscr{T}, j \in \{1, \ldots, d\}\Big\}.
\end{aligned}$$

Analogous to Step 2 in the proof of Lemma 3.3 define the equivalence relation $\mathscr{R}_n \subseteq \mathscr{F} \times \mathscr{F}$ by

$$\begin{aligned}
(f_{\tau^1,\theta_1^1,\theta_2^1,j}, f_{\tau^2,\theta_1^2,\theta_2^2,k}) \in \mathscr{R}_n \iff \Big\{&j = k,\ \theta_1^1 - \theta_2^1 = \theta_1^2 - \theta_2^2,\ |S_n^*(\tau^1)| = |S_n^*(\tau^2)|, \\
&(\theta_1^1 - \theta_2^1), (\theta_1^2 - \theta_2^2) \in \mathscr{M}^d\big(S_n^*(\tau^1)\big)\Big\},
\end{aligned}$$

and the probability measure $\nu_n : \sigma(\mathscr{F}/\mathscr{R}_n) \to [0,1]$ by $\nu_n(\emptyset) = 0$ and

$$\nu_n(P_{\tau,\theta_1,\theta_2,j}) = c_\nu^{-1} d^{-1} \left(\frac{5ed}{|S_n^*(\tau)|}\right)^{-2(3+\bar{c})|S_n^*(\tau)|}, \quad P_{\tau,\theta_1,\theta_2,j} \in \mathscr{F}/\mathscr{R}_n,$$

where $c_\nu > 0$ is such that that $1 = \sum_{P_{\tau,j} \in \mathscr{F}_\mathscr{M}/\mathscr{R}_n} \nu_n(P_{\tau,j})$. Observe the additional factor d^{-1} in the probability measure, which takes care of the ℓ_∞-norm. Also, note that $0 < c_\nu < \frac{eC}{2}$ (with $C > 0$ the constant from Lemma 2.6) and that each subclass $P_{\tau,\theta_1,\theta_2,j} \in \mathscr{F}/\mathscr{R}_n$ has envelop function $G_j(X) = X^{(j)} \vee 1$. Thus, by Theorem2.4 there exists $N_1 \geq N_0$ such that for

all $n > N_1$ and all $f_{\tau,\theta_1,\theta_2,j} \in \mathscr{F}$,

$$\begin{aligned}
\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\theta_1,\theta_2,j})| &\leq \frac{7}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{\tau,\theta_1,\theta_2,j}]_{\mathscr{R}_n}}|\mathbb{G}_{nn}(f)|\Big] \\
&+ 149\left(\sup_{f\in[f_{\tau,\theta_1,\theta_2,j}]_{\mathscr{R}_n}}\bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2} \\
&\times\left(\frac{\log(e/2)+\log d+3|S_n^*|\log(5ed/|S_n^*|)+\log\log n}{n}\right)^{1/2} \\
&+ 223\left(\bar{\mathrm{E}}_{nn}[G_j^4]\right)^{1/4} \\
&\times\left(\frac{\log(e/2)+\log d+3|S_n^*|\log(5ed/|M|)+\log\log n}{n}\right)^{1/2}\left(\frac{\log\log n}{n}\right)^{1/4}.
\end{aligned} \tag{3.20}$$

Upper bound the first term on the right side of eq. (3.20) by

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{\tau,\theta_1,\theta_2,j}]_{\mathscr{R}_n}}|\mathbb{G}_{nn}(f)|\Big] &\leq \frac{2}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{\tau,\theta_1,\theta_2,j}]_{\mathscr{R}_n}}|\mathbb{G}^{\circ}_{nn}(f)|\Big] \\
&= 2\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i 1\left\{X_{ni}'\theta_1 < Y_{ni} \leq X_{ni}'\theta_2\right\}X_{ni}^{(j)}\right| \\
&= 2\mathrm{E}\Big[\max_{u\in\{-1,1\}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i 1\left\{X_{ni}'\theta_1 < Y_{ni} \leq X_{ni}'\theta_2\right\}X_{ni}^{(j)}u\Big] \\
&\leq 2\left(\frac{\log 2}{n}\right)^{1/2} f_+^{1/2}\sup_{\|v\|_2=1}\left(\mathrm{E}\Big[\frac{1}{n}\sum_{i=1}^n |X_{ni}'v||X_{ni}^{(j)}|^2\Big]\right)^{1/2}\|\theta_1-\theta_2\|_2^{1/2} \\
&\leq 2(\log 2)^{1/2} f_+^{1/2}\mu_4\bar{\varphi}_{\max}^{1/4}(S_n^*,\bar{c})\left(\frac{R_n}{n}\right)^{1/2}.
\end{aligned} \tag{3.21}$$

The first part of the second term can be bounded similarly,

$$\left(\sup_{f\in[f_{\tau,\theta_1,\theta_2,j}]_{\mathscr{R}_n}}\bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2} \leq f_+^{1/2}\mu_4\bar{\varphi}_{\max}^{1/4}(S_n^*,\bar{c})R_n^{1/2}, \tag{3.22}$$

and the first term of the third term is bounded by

$$\left(\bar{\mathrm{E}}_{nn}[G_j^4]\right)^{1/4} \leq \mu_4. \tag{3.23}$$

Combining eq. (3.21)-(3.23) with eq. (3.20) we conclude that there exists an absolute

constant $c_1 > 0$ such that for all $n \geq N_1$ and for all $f_{\tau,\theta_1,\theta_2,j} \in \mathscr{F}$,

$$\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\theta_1,\theta_2,j})| \leq c_1 \bar{\varphi}_{\max}^{1/4}(S_n^*, \bar{c}) \left(\frac{R_n}{n}\right)^{1/2}$$
$$+ c_1 \bar{\varphi}_{\max}^{1/4}(S_n^*, \bar{c}) \left(\frac{\log d + |S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} R_n^{1/2} \quad a.s.$$

Thus, for all $n \geq N_1$,

$$\sup_{\tau \in \mathscr{T}} \|r_{1,n}(\tau)\|_\infty$$
$$= O\left(\frac{\bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}^{1/2}(S_n^*, \bar{c})} \left(|S_n^*|^{1/2}\lambda_n\right)^{1/2} \left(\frac{\log d + |S_n^*| \log(ed/|S_n^*|) + \log\log n}{n}\right)^{1/2}\right) \quad a.s. \tag{3.24}$$

Step 3. Bound on $r_{n,2}(\tau)$ in the ℓ_∞-norm. By the quadratic approximation of eq. (2.12), there exist $c_2 > 0$ such that for all $n \geq N_1$,

$$\sup_{\tau \in \mathscr{T}} \|r_{n,2}(\tau)\|_\infty \leq c_2 f_H \bar{\varphi}_{\max}^{1/2+\alpha/2}(S_n^*, \bar{c}) \mu_2 R_n^{1+\alpha} = O\left(\frac{\bar{\varphi}_{\max}^{1+\alpha}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}^{1+\alpha}(S_n^*, \bar{c})} \left(|S_n^*|^{1/2}\lambda_n\right)^{1+\alpha}\right) \quad a.s. \tag{3.25}$$

□

3.6.1.4 Proof of Theorem 3.3

Proof. **Step 1. Empirical sparsity via complementary slackness condition.** Recall that the ℓ_1-penalized quantile regression problem can be written as a linear programming problem:

$$\min_{(v^+,v^-,\theta^+,\theta^-) \in \mathbb{R}^{2n+2d}} \Big\{ \mathbb{E}_{nn}[\tau v_i^+ + (1-\tau) v_i^-]$$
$$+ \lambda_n \sum_{j=1}^d (\theta_j^+ + \theta_j^-) \ \ s.t. \ \ v_i^+ - v_i^- = Y_{ni} - X_{ni}'(\theta_j^+ - \theta_j^-),\ i \leq n \Big\}.$$

The dual of this problem is

$$\max_{a \in \mathbb{R}^n} \left\{ \mathbb{E}_{nn}[Y_{ni} a_i] \ \ s.t. \ \ \left|\mathbb{E}_{nn}[X_{ni}^{(j)} a_i]\right| \leq \lambda_n,\ j \leq d, \ \ and \ \ \tau - 1 \leq a_i \leq \tau,\ i \leq n \right\}.$$

Denote by $\hat{a}_n(\tau)$ the solution of the dual problem. By the complementary slackness

conditions associated with the dual problem (i.e. Belloni and Chernozhukov, 2011, Lemma 9), we have the following characterization of the empirical sparsity

$$\lambda_n |\widehat{S}_n(\tau)|^{1/2} = \left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} \hat{a}_{ni}(\tau)] \right\|_2. \tag{3.26}$$

Since $|\widehat{S}_n(\tau)| \leq n$ for all $\tau \in \mathscr{T}$, we have

$$\sup_{\|u\|_2=1} \mathrm{E}_{nn}\left[\frac{1}{n} \sum_{i=1}^{n} (u' X_{ni,\widehat{S}_n(\tau)})^2 \right] \leq \bar{\varphi}_{\max}(n). \tag{3.27}$$

Expand

$$\begin{aligned}
&\left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} \hat{a}_{ni}(\tau)] \right\|_2 \\
&\leq \left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} (\hat{a}_{ni}(\tau) - \varphi_\tau(Y_{ni} - X_{ni}' \hat{\theta}_n(\tau))] \right\|_2 \\
&\quad + \left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} (\varphi_\tau(Y_{ni} \leq X_{ni}' \hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}' \theta_n^*(\tau)))] \right\|_2 \\
&\quad + \left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} \varphi_\tau(Y_{ni} - X_{ni}' \theta_n^*(\tau))] \right\|_2 \\
&= A + B + C.
\end{aligned} \tag{3.28}$$

Step 2. Upper bound on C. To bound the third term on the right of eq. (3.28) note that by Lemma 2.8 there exist absolute constants $N_1, c_1 > 0$ such that for all $n > N_1$ and all $\tau \in \mathscr{T}$,

$$\begin{aligned}
&\left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} \varphi_\tau(Y_{ni} - X_{ni}' \theta_n^*(\tau))] \right\|_2 \\
&\leq c_1 \bar{\varphi}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{V_{\mathscr{L}}}{n} \right)^{1/2} + c_1 \bar{\varphi}_{\max}^{1/2}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n| \log(ed/|\widehat{S}_n|) + \log\log n}{n} \right)^{1/2} \quad a.s.
\end{aligned} \tag{3.29}$$

Step 3. Upper bound on B. Let

$$R_n(\lambda_n, S_n^*) = c_0 \left(\frac{\bar{\varphi}_{\max}^{1/2}(S_n^*, \bar{c})}{\bar{\varphi}_{\min}(S_n^*, \bar{c})} \left(\frac{|S_n^*| \log(ed/|S_n^*|^{1/2}) + \log\log n}{n} \right)^{1/2} \bigvee \frac{|S_n^*|^{1/2} \lambda_n}{\bar{\varphi}_{\min}(S_n^*, \bar{c})} \right),$$

where $c_0 > 0$ is the constant from Theorem 3.1. To bound the second term on the right of eq. (3.28) consider

$$\left\| \mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)} (\varphi_\tau(Y_{ni} \leq X_{ni}' \hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}' \theta_n^*(\tau)))] \right\|_2$$

$$
\begin{aligned}
&\leq \Big\| \frac{1}{\sqrt{n}} \mathbb{G}_{nn}[X_{ni,\widehat{S}_n(\tau)}(\varphi_\tau(Y_{ni} \leq X_{ni}'\hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}'\theta_n^*(\tau)))] \Big\|_2 \\
&+ \sup_{M:|M|=|\widehat{S}_n(\tau)|} \Big\| \mathrm{E}_{nn}[X_{ni,M}(\varphi_\tau(Y_{ni} \leq X_{ni}'\hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}'\theta_n^*(\tau)))] \Big\|_2 .
\end{aligned} \tag{3.30}
$$

By Lemma 2.7 and Theorem 3.1 there exist absolute constants $N_2, c_2 > 0$ such that for all $n > N_2$ and all $\tau \in \mathscr{T}$,

$$
\begin{aligned}
&\Big\| \frac{1}{\sqrt{n}} \mathbb{G}_{nn}[X_{ni,\widehat{S}_n(\tau)}(\varphi_\tau(Y_{ni} \leq X_{ni}'\hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}'\theta_n^*(\tau)))] \Big\|_2 \\
&\leq c_2 \bar{\varphi}_{\max}^{3/2}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n|}{n} \right)^{3/4} \\
&+ c_2 f_+^{1/2} \bar{\varphi}_{\max}^{3/2}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n| \log(ed/|\widehat{S}_n|) + \log\log n}{n} \right)^{1/2} R_n^{1/2}(\lambda_n, S_n^*) \quad a.s.
\end{aligned} \tag{3.31}
$$

By Theorem 3.1 and two applications of Cauchy-Schwarz,

$$
\begin{aligned}
&\sup_{M:|M|=|\widehat{S}_n(\tau)|} \Big\| \mathrm{E}_{nn}[X_{ni,M}(\varphi_\tau(Y_{ni} \leq X_{ni}'\hat{\theta}_n(\tau)) - \varphi_\tau(Y_{ni} \leq X_{ni}'\theta_n^*(\tau)))] \Big\|_2 \\
&\leq \bar{\varphi}_{\max}(\widehat{S}_n, \bar{c}) f_+ R_n(\lambda_n, S_n^*).
\end{aligned} \tag{3.32}
$$

Step 4. Upper bound on A. To bound the first term in eq. (3.28) note that $\hat{a}_{ni}(\tau) \neq \varphi(Y_{ni} - X_{ni}'\hat{\theta}_n(\tau)$ only if $Y_{ni} = X_{ni}'\hat{\theta}_n(\tau)$ and that the penalized quantile regression fit can interpolate at most $|\widehat{S}_n(\tau)| \leq |\widehat{S}_n| \leq n$ points with probability one (e.g. Koenker, 2005). Further, note that $|\hat{a}_{ni}(\tau) - \varphi(Y_{ni} - X_{ni}'\hat{\theta}_n(\tau)| \leq 1$ for all $i \leq n$, $n \in \mathbb{N}$. Thus,

$$
\begin{aligned}
&\Big\| \mathbb{E}_{nn}[(\hat{a}_{ni}(\tau) - \varphi(Y_{ni} - X_{ni}'\hat{\theta}_n(\tau)) X_{ni,\widehat{S}_n(\tau)}] \Big\|_2 \\
&\leq \max_{\substack{I:|I|\leq|\widehat{S}_n| \\ I\in\{1,\ldots,n\}}} \sup_{u\in\mathscr{E}(\widehat{S}_n(\tau))} \frac{1}{n} \sum_{i\in I} |u' X_{ni}| \\
&\leq \max_{\substack{I:|I|\leq|\widehat{S}_n| \\ I\in\{1,\ldots,n\}}} \sup_{u\in\mathscr{E}(\widehat{S}_n(\tau))} \frac{1}{\sqrt{n}} \mathbb{G}_{nn}(|u' X_{ni}|; I) + \max_{\substack{I:|I|\leq\widehat{S}_n \\ I\in\{1,\ldots,n\}}} \sup_{u\in\mathscr{E}(\widehat{S}_n(\tau))} \bar{\mathrm{E}}_{nn}\big[|u' X_{ni}|; I\big],
\end{aligned} \tag{3.33}
$$

where

$$
\mathscr{E}(S) = \Big\{ u \in \mathbb{R}^d : \|u\|_0 = |S| \Big\}, \quad S \subseteq \{1, \ldots, d\}.
$$

The second term on the right of display (3.33) can be upper bounded using Lemma 2.4,

$$\max_{\substack{I:|I|\le\widehat{S}_n\\ I\in\{1,\dots,n\}}}\sup_{u\in\mathscr{E}(\widehat{S}_n(\tau))}\bar{\mathrm{E}}_{nn}\big[|u'X_{ni};I\big]\le\bar{\varphi}^{1/2}_{\max}(\widehat{S}_n,\bar{c})\left(\frac{|\widehat{S}_n|}{n}\right)^{1/2} \tag{3.34}$$

To bound the first term on the right of display (3.33) we now establish an adaptive upper bound uniformly over all index sets I of cardinality $|I|\le n$ (this suffices since we know that $|\widehat{S}_n(\tau)|\le|\widehat{S}_n|\le n$ uniformly in $\tau\in\mathscr{T}$). To this end, consider the following class of functions:

$$\mathscr{F}=\Big\{(Y,X)\mapsto|X_M'v|:M\subseteq\{1,\dots,d\},\ |M|\le n,\ v\in\mathscr{M}^d_{|M|}\Big\},$$

where $\mathscr{M}^d_m\subset\mathscr{B}^d$, $m\in\{1,\dots,n\}$, is the finite set defined in Lemma 2.6.

Observe that $\mathscr{F}$ is countably finite and define the equivalence relation $\mathscr{R}\subseteq\mathscr{F}\times\mathscr{F}$ by

$$(f_{v^1,M^1},f_{v^2,M^2})\in\mathscr{R}\iff\Big\{v^1=v^2,\ |M^1|=|M^2|\Big\},$$

and the probability measure $\nu_n:\sigma(\mathscr{F}/\mathscr{R})\to[0,1]$ by $\nu_n(\emptyset)=0$ and

$$\nu_n(P_{v,M})=c_\nu^{-1}\left(\frac{5ed}{|M|}\right)^{-2|M|}\left(\frac{en}{|M|}\right)^{-|M|},\ P_{v,M}\in\mathscr{F}/\mathscr{R},$$

where $c_\nu>0$ is such that $1=\sum_{P_{v,M}\in\mathscr{F}/\mathscr{R}}\nu_n(P_{v,M})$. Note that $0<c_\nu<\frac{eC}{2}$, where $C>0$ is the constant from Lemma 2.6. Indeed,

$$\begin{aligned}c_\nu&=\sum_{P_{v,M}\in\mathscr{F}/\mathscr{R}}c_\nu\nu_n(P_{v,M})=\sum_{k=1}^d\sum_{j=1}^n|\mathscr{M}^d_k|\binom{d}{k}\binom{n}{j}\left(\frac{5ed}{k}\right)^{-2k}\left(\frac{en}{j}\right)^{-j}\\&\le C\sum_{k=1}^d\left(\frac{5ed}{k}\right)^{-k}<\frac{C}{1-(5e)^{-1}}.\end{aligned}$$

Further, note that each subclass $P_{v,M}\in\mathscr{F}/\mathscr{R}$ has envelop function $G_{v,M}(X)=|X_M'v|\vee1$ (where $\|v\|_2\le1$ is fixed).

Now, instead of applying Theorem 2.4 as usual, invoke Lemma 3.4 to handle the maximum over the index set $I\subset\{1,\dots,n\}$: We conclude that there exist absolute constants

$N_3, c_3 > 0$ such that for all $n > N_3$, $f_{v,M} \in \mathscr{F}$ and $I \subseteq \{1,\dots,n\}$,

$$\begin{aligned}
\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{v,M};I)| &\leq \frac{c_3}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}}}|\mathbb{G}_{nn}(f;I)|\Big] \\
&+ c_3\left(\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}_n}}\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2;I)|\Big]\right)^{1/2} \\
&\times\left(\frac{|M|\log(ed/|M|)+\log\log n+|I|\log(2en/|I|)}{n}\right)^{1/2} \\
&+ c_3\left(\sup_{f\in[f_{v,M}]_{\mathscr{R}}}\bar{\mathrm{E}}_{nn}[f^2;I]\right)^{1/2}\left(\frac{|M|\log(ed/|M|)+\log\log n+|I|\log(2en/|I|)}{n}\right)^{1/2} \\
&+ c_3\left(\bar{\mathrm{E}}_{nn}[G_{v,M}^4;I]\right)^{1/4} \\
&\times\left(\frac{|M|\log(ed/|M|)+\log\log n+|I|\log(2en/|I|)}{n}\right)^{1/2}\left(\frac{\log\log n}{n}\right)^{1/4}.
\end{aligned} \tag{3.35}$$

To bound the first term on the right side of eq. (3.35) symmetrize the centered process and exploit the Sub-Gaussianity of (conditional) Rademacher averages,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}}}|\mathbb{G}_{nn}(f;I)|\Big] &\leq 2\mathrm{E}\left|\frac{1}{n}\sum_{i\in I}\varepsilon_i|X'_{ni,M}v|\right| \\
&\leq c_4\bar{\varphi}_{\max}^{1/2}(|M|,\bar{c})\frac{|I|^{1/2}}{n} \\
&\leq c_4\frac{\bar{\varphi}_{\max}^{1/2}(|M|,\bar{c})}{n^{1/2}},
\end{aligned}$$

where $c_4 > 0$ is an absolute constant. A bound of the first part of the second term the right side of eq. (3.35) follows from similar arguments and the moment bound in Lemma 2.4,

$$\left(\mathrm{E}\Big[\sup_{f\in[f_{v,M}]_{\mathscr{R}_n}}\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2;I)|\Big]\right)^{1/2} \leq c_5\frac{\bar{\varphi}_{\max}^{1/2}(|M|,\bar{c})}{n^{1/2}},$$

where $c_5 > 0$ is an absolute constant.

To bound the first part of the third term on the right side of eq. (3.35) use Lemma 2.4 and compute

$$\left(\sup_{f\in[f_{v,M}]_{\mathscr{R}}}\bar{\mathrm{E}}_{nn}[f^2;I]\right)^{1/2} = \bar{\varphi}_{\max}^{1/2}(|M|,\bar{c})\left(\frac{|I|}{n}\right)^{1/2} \leq \bar{\varphi}_{\max}^{1/2}(|M|,\bar{c}).$$

Similarly, the first part of the forth term on the right of eq. (3.35) can be bounded using Lemma 2.4, i.e. there exist an absolute constant $c_6 > 0$ such that

$$\left(\bar{\mathrm{E}}_{nn}[G^4_{u,M};I]\right)^{1/4} \leq c_6 \bar{\varphi}^{1/2}_{\max}(|M|,\bar{c}) \left(\frac{|I|}{n}\right)^{1/2} \leq c_6 \bar{\varphi}^{1/2}_{\max}(|M|,\bar{c}).$$

Thus, there exist absolute constants $c_6, N_6 > 0$ such that for all $n > N_6 \geq N_3$,

$$\begin{aligned} &\max_{\substack{I:|I|\leq|\widehat{S}_n(\tau)| \\ I\in\{1,\ldots,n\}}} \sup_{u\in\mathscr{C}(\widehat{S}_n(\tau))} \frac{1}{\sqrt{n}}\mathbb{G}_{nn}(|u'X_{ni}|;I) \\ &\qquad \leq c_6 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|\log(ed/|\widehat{S}_n|)+\log\log n}{n}\right)^{1/2} \quad a.s. \end{aligned} \tag{3.36}$$

Thus, combining bounds eq. (3.34) and (3.36) we conclude that there exist absolute constants $c_6, N_6 > 0$ such that for all $n > N_6$ and all $\tau \in \mathscr{T}$,

$$\begin{aligned} &\left\|\mathbb{E}_{nn}[X_{ni,\widehat{S}_n(\tau)}(\hat{a}_{ni}(\tau) - \varphi(Y_{ni} - X'_{ni}\hat{\theta}_n(\tau))]\right\|_2 \\ &\qquad \leq c_6 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|\log(ed/|\widehat{S}_n|)+\log\log n}{n}\right)^{1/2} \quad a.s. \end{aligned} \tag{3.37}$$

Step 5. Completing the bound. The bounds on A, B, and C imply that there exist $c_7, N_7 > 0$ such that for all $n > N_7$ and all $\tau \in \mathscr{T}$,

$$\begin{aligned} \lambda_n|\widehat{S}_n(\tau)|^{1/2} \leq\ & c_7 \bar{\varphi}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{V_{\mathscr{L}}}{n}\right)^{1/2} \\ &+ c_7 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|\log(ed/|\widehat{S}_n|)+\log\log n}{n}\right)^{1/2} \\ &+ c_7 \bar{\varphi}^{3/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|}{n}\right)^{3/4} \\ &+ c_7 \bar{\varphi}^{3/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|\log(ed/|\widehat{S}_n|)+\log\log n}{n}\right)^{1/2} R^{1/2}_n(\lambda_n,S^*_n) \\ &+ c_7 \bar{\varphi}_{\max}(\widehat{S}_n,\bar{c}) R_n(\lambda_n,S^*_n) \\ &+ c_7 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n,\bar{c}) \left(\frac{|\widehat{S}_n|\log(ed/|\widehat{S}_n|)+\log\log n}{n}\right)^{1/2} \end{aligned}$$

$$
\begin{aligned}
&\leq c_7 \bar{\varphi}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{V_{\mathscr{L}}}{n} \right)^{1/2} + c_7 \bar{\varphi}^{3/2}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n|}{n} \right)^{3/4} \\
&+ c_7 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n| \log(ed/|\widehat{S}_n|^{1/2}) + \log\log n}{n} \right)^{1/2} \\
&+ c_7 \bar{\varphi}^{3/2}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{|\widehat{S}_n| \log(ed/|\widehat{S}_n|^{1/2}) + \log\log n}{n} \right)^{1/2} |S_n^*|^{1/4} \lambda_n^{1/2} \\
&+ c_7 \bar{\varphi}_{\max}(\widehat{S}_n, \bar{c}) |S_n^*|^{1/2} \lambda_n
\end{aligned}
\tag{3.38}
$$

Rearrange eq. (3.38) to obtain,

$$
\begin{aligned}
|\widehat{S}_n(\tau)|^{1/2} &\left(\lambda_n - c_7 \frac{\bar{\varphi}^{3/2}_{\max}(\widehat{S}_n, \bar{c})}{n^{1/2}} \left(\frac{|\widehat{S}_n|}{n} \right)^{1/4} - c_7 \bar{\varphi}^{1/2}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{\log d + \log\log n}{n} \right)^{1/2} \right. \\
&\left. - c_7 \bar{\varphi}^{3/2}_{\max}(\widehat{S}_n, \bar{c}) \left(\frac{\log d + \log\log n}{n} \right)^{1/2} |S_n^*|^{1/4} \lambda_n^{1/2} \right) \\
&\leq c_7 \bar{\varphi}_{\max}(|\widehat{S}_n|, \bar{c}) |S_n^*|^{1/2} \lambda_n + c_7 \bar{\varphi}_{\max}(|\widehat{S}_n|, \bar{c}) \left(\frac{V_{\mathscr{L}}}{n} \right)^{1/2}.
\end{aligned}
$$

Exploiting the conditions on λ_n and the restricted eigenvalues and the fact that $\widehat{S}_n(\tau)| \leq n$, we conclude that there exist constants $c_8, N_8 > 0$ such that for all $n > N_8$ and all $\tau \in \mathscr{T}$,

$$
|\widehat{S}_n(\tau)| \leq |S_n^*| \times c_8 \bar{\varphi}^2_{\max}(n, \bar{c}) + c_8 \bar{\varphi}^2_{\max}(n, \bar{c}) \frac{V_{\mathscr{L}}}{n \lambda_n^2} \quad a.s.
$$

□

3.6.1.5 Proof of Theorem 3.4

Proof. We combine the bound on the empirical sparsity form Theorem 3.3 with the uniform-in-model consistency result from Theorem 2.1 and the consistency result for ℓ_1-penalized quantile regression from Theorem 3.1. For this proof only, we introduce the following notation: For a d-dimensional vector $\theta(\tau) \in \mathbb{R}^d$ with support S we write $\theta(\tau; S)$ when we want to emphasize the support set S. This not to be confused with $\theta_S(\tau) \in \mathbb{R}^{|S|}$ which denotes the projection of $\theta(\tau)$ onto the set of coordinates S. However, observe that $\|\theta(\tau; S)\|_2 =$

$\|\theta_S(\tau)\|_2$. Note that with this notation we have the following identity:

$$\tilde{\theta}_n(\tau) = \hat{\theta}_n(\tau; \widehat{S}_n(\tau)).$$

Case 1. $S_n^*(\tau) \subseteq \widehat{S}_n(\tau)$**.** Observe that in this case we have $\theta_n^*(\tau; \widehat{S}_n(\tau)) = \theta_n^*(\tau)$. Moreover, the logarithm of the VC-index associated with the $|\widehat{S}_n(\tau)|$-dimensional quantile regression vector (of which $|S_n^*|$ coordinates are known to be included) is, up to additive and multiplicative absolute constants,

$$\log \binom{d - |S_n^*(\tau)|}{\widehat{m}_n(\tau)} + |S_n^*(\tau)| \leq |\widehat{m}_n(\tau)| \log\left(ed/|\widehat{m}_n(\tau)|\right) + |S_n^*(\tau)|,$$

where $\widehat{m}_n(\tau) = \widehat{S}_n(\tau) \setminus S_n^*(\tau)$ denotes the number of incorrectly included predictors. Hence, by Theorem 2.1 there exist $c_0, N_0 > 0$ such that for all $n > N_0$ and all $\tau \in \mathscr{T}$,

$$\begin{aligned}
\|\tilde{\theta}_n(\tau) - \theta_n^*(\tau)\|_2 &\leq \|\hat{\theta}_n(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau; \widehat{S}_n(\tau))\|_2 + \|\theta_n^*(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau)\|_2 \\
&= \|\hat{\theta}_n(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau; \widehat{S}_n(\tau))\|_2 \\
&\leq c_0 \frac{\bar{\varphi}_{\max}^{1/2}(\widehat{S}_n(\tau))}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau))} \left(\frac{|\widehat{m}_n(\tau)| \log(ed/|\widehat{m}_n(\tau)|^{1/2}) + |S_n^*(\tau)| + \log\log n}{n} \right)^{1/2} \quad a.s.
\end{aligned}$$

Case 2. $S_n^*(\tau) \not\subseteq \widehat{S}_n(\tau)$**.** In this case, $\|\theta_n^*(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau)\|_2 \neq 0$. By Theorem 2.1, there exist $c_0, N_0 > 0$ such that for all $n > N_0$ and all $\tau \in \mathscr{T}$,

$$\begin{aligned}
\|\tilde{\theta}_n(\tau) - \theta_n^*(\tau)\|_2 &\leq \|\hat{\theta}_n(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau; \widehat{S}_n(\tau))\|_2 + \|\theta_n^*(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau)\|_2 \\
&\leq c_0 \frac{\bar{\varphi}_{\max}^{1/2}(\widehat{S}_n(\tau))}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau))} \left(\frac{|\widehat{S}_n(\tau)| \log(ed/|\widehat{S}_n(\tau)|^{1/2}) + \log\log n}{n} \right)^{1/2} \\
&\quad + \|\theta_n^*(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau)\|_2 \quad a.s.
\end{aligned}$$

In the remainder of the proof we establish an upper bound on the second term on the right in above display. Observe that by optimality of $\theta_n^*(\tau)$ and convexity of the loss function,

$$\begin{aligned}
&\bar{\mathrm{E}}_{nn}\left[\rho_\tau\left(Y_{ni} - X_{ni}'\theta_n^*(\tau; \widehat{S}_n(\tau))\right) - \rho_\tau(Y_{ni} - X_{ni}'\theta_n^*(\tau))\right] \\
&\quad \geq \bar{\varphi}_{\min}(\widehat{S}_n(\tau) \cup S_n^*(\tau)) \|\theta_n^*(\tau; \widehat{S}_n(\tau)) - \theta_n^*(\tau)\|_2^2. \qquad (3.39)
\end{aligned}$$

Moreover, since $\theta_n^*(\tau; \widehat{S}_n(\tau))$ and $\hat{\theta}_{n,\lambda_n}(\tau)$ have the same support set and by optimality of

$\theta_n^*(\tau;\widehat{S}_n(\tau))$,

$$\begin{aligned}
&\bar{\mathrm{E}}_{nn}\left[\rho_\tau\left(Y_{ni}-X_{ni}'\theta_n^*(\tau;\widehat{S}_n(\tau))\right)-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right]\\
&\quad\le \bar{\mathrm{E}}_{nn}\left[\rho_\tau\left(Y_{ni}-X_{ni}'\theta_n^*(\tau;\widehat{S}_n(\tau))\right)-\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))\right]\\
&\qquad\quad+\bar{\mathrm{E}}_{nn}\left[\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right]\\
&\quad\le \bar{\mathrm{E}}_{nn}\left[\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right]\\
&\quad= \frac{1}{\sqrt{n}}\mathbb{G}_{nn}\left(\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right)\\
&\qquad\quad-\mathbb{E}\left[\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right] \qquad (3.40)
\end{aligned}$$

The second term on the right side of eq. (3.40) can be upper bounded exploiting the cone property of the lasso estimate $\hat{\theta}_{n,\lambda_n}(\tau)$, i.e. uniformly in $\tau\in\mathscr{T}$,

$$\begin{aligned}
&\mathbb{E}\left[\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right] \qquad (3.41)\\
&\quad\le \lambda_n\left(\|\theta_n^*(\tau)\|_1-\|\hat{\theta}_{n,\lambda_n}(\tau)\|_1\right)\\
&\quad\le (1+\bar{c})|S_n^*|^{1/2}\lambda_n\|\theta_n^*(\tau)-\hat{\theta}_{n,\lambda_n}(\tau)\|_2 \quad a.s. \qquad (3.42)
\end{aligned}$$

The first term on the right side of eq. (3.40) can be upper bounded using the result on the centered quantile loss function from Lemma 3.3, i.e. uniformly in $\tau\in\mathscr{T}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbb{G}_{nn}\left(\rho_\tau(Y_{ni}-X_{ni}'\hat{\theta}_{n,\lambda_n}(\tau))-\rho_\tau(Y_{ni}-X_{ni}'\theta_n^*(\tau))\right)\\
&\le c_1(2+\bar{c})\bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})\left(\frac{|S_n^*|\log(ed/|S_n^*|^{1/2})+\log\log n}{n}\right)^{1/2}\|\theta_n^*(\tau)-\hat{\theta}_{n,\lambda_n}(\tau)\|_2 \ \ a.s., \qquad (3.43)
\end{aligned}$$

where $c_1>0$ is an absolute constant.

Combining eq. (3.39)-(3.43) we conclude that there exist absolute constants $c_2, N_2>0$ such that for all $n>N_2$ and all $\tau\in\mathscr{T}$,

$$\begin{aligned}
&\|\theta_n^*(\tau;\widehat{S}_n(\tau))-\theta_n^*(\tau)\|_2\\
&\quad\le c_2\|\theta_n^*(\tau)-\hat{\theta}_{n,\lambda_n}(\tau)\|_2^{1/2}\left(\frac{|S_n^*|^{1/2}\lambda_n}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau)\cup S_n^*(\tau))}\right.\\
&\qquad\left.\bigvee\frac{\bar{\varphi}_{\max}(S_n^*,\bar{c})}{\bar{\varphi}_{\min}(\widehat{S}_n(\tau)\cup S_n^*(\tau))}\left(\frac{|S_n^*|\log(ed/|S_n^*|^{1/2})+\log\log n}{n}\right)^{1/2}\right)^{1/2}
\end{aligned}$$

Completion of the proof. The statement of the theorem follows by combining Cases 1 and 2. □

3.6.1.6 Proof of Corollary 3.2

Proof. For this proof only, we introduce the following notation: For a d-dimensional vector $\theta(\tau) \in \mathbb{R}^d$ with support S we write $\theta(\tau;S)$ when we want to emphasize the support set S. This not to be confused with $\theta_S(\tau) \in \mathbb{R}^{|S|}$ which denotes the projection of $\theta(\tau)$ onto the set of coordinates S. However, observe that $\|\theta(\tau;S)\|_2 = \|\theta_S(\tau)\|_2$. Note that with this notation we have the following identity: $\tilde{\theta}_n(\tau) = \hat{\theta}_n(\tau;\widehat{S}_n(\tau))$. Since $S_n^*(\tau) \subseteq \widehat{S}_n(\tau)$, we also have $\theta_n^*(\tau;\widehat{S}_n(\tau)) = \theta_n^*(\tau)$. The claim now follows as Step 1 in the proof of Theorem 3.4 but using the uniform-in-model Bahadur representation from Theorem 2.2 instead of the uniform-in-model consistency result from Theorem 2.1. □

3.6.2 Proofs of Section 3.4

3.6.2.1 Proof of Lemma 3.2

Proof. **Proof of the first statement.** The proof is similar to the proof of Lemma 2.6. By Lemma 7.1 in Koltchinskii (2011), we have for $\varepsilon \in (0, \frac{1}{2}]$ and $S \subset \{1,\dots,d\}$, $|S| = m$,

$$C(S,c) \subset \operatorname{conv}\left(2(2+c)\,\mathscr{M}_m^d\right),$$

where

$$\mathscr{M}_m^d = \bigcup_{k=1}^{m} \bigcup_{S:|S|=k} \mathscr{N}_S^d,$$

and $\mathscr{N}_S^d$ is an ε-net of the sparse unit ball $\mathscr{B}_S^d$, i.e.

$$\mathscr{B}_S^d = \left\{v \in \mathbb{R}^d : \|v\|_2 \le 1,\ j \notin S \implies v_j = 0\right\}.$$

Without loss of generality we can assume that $2(2+c)$ is an integer. Hence, as in the proof of Lemma 2.6,

$$|2(2+c)\,\mathscr{M}_m^d| \le \binom{|\mathscr{M}_m^d| + 2c + 3}{2c+4} \le C|\mathscr{M}_m^d|^{2(2+c)} \le C\left(1+\frac{2}{\varepsilon}\right)^{2(2+c)m}\left(\frac{ed}{m}\right)^{2(2+c)m},$$

where $C > 0$ is an absolute constant. Note that

$$\operatorname{conv}\left(2(2+c)\,\mathscr{M}_m^d\right) \subseteq \operatorname{conv}\left(2(2+c)\,\widetilde{\mathscr{M}}_m^d\right)$$

for

$$\widetilde{\mathscr{M}}_m^d = \bigcup_{k=1}^{m} \bigcup_{S:|S|=k} \bigcup_{v \in \mathscr{N}_S^d} \mathscr{B}^d_{\operatorname{supp}(v)}(v, \varepsilon).$$

Since F is a convex function, so is the map $v \mapsto \mathbb{E}_{nn}\big(F(X_{ni}'v)\big)$. Hence, the Bauer Maximum Principle (e.g. Aliprantis and Border, 2006, Corollary 7.69, p. 298) applies and we conclude as in Lemma 2.6 that it suffices to consider the maximum over $2(2+c)\mathscr{M}_m^d$. Thus,

$$\max_{S:|S|=m} \sup_{u \in C(S,c)} \mathbb{E}_{nn}\big(F(X_{ni}'u)\big) \leq \max_{v \in 2(2+c)\mathscr{M}_m^d} \sup_{u \in \mathscr{B}^d_{\operatorname{supp}(v)}(v,\varepsilon)} \mathbb{E}_{nn}\big(F(X_{ni}'u)\big).$$

Proof of the second statement. Follows as the proof of the second statement of Lemma 2.6. Therefore, we do not reproduce it. □

3.6.2.2 Proof of Lemma 3.3

Proof. We would like to proceed as in the proof of Lemma 2.9, which is the low-dimensional analogue to this lemma. However, in high-dimensions breaking up the supremum (over the cone $C(S_n^*(\tau), \bar{c})$) into smaller chunks is more difficult. Obviously, the statement of Lemma 3.2 is much weaker than the corresponding statement of its low-dimensional analogue, Lemma 2.6. We address this problem by modifying Theorem 2.4 to exploit the Lipschitz continuity of the quantile loss function.

Step 1. Enlarged and reduced function classes. Recall the function class in the statement of the lemma:

$$\mathscr{F}_1 = \Big\{ (Y,X) \mapsto \rho_\tau(Y - X'\theta) - \rho_\tau(Y - X'\theta_n^*(\tau)) : \\ \theta \in \mathbb{R}^d,\, \theta - \theta_n^*(\tau) \in C(S_n^*(\tau), \bar{c}) \cap \mathscr{B}^d\big(r_n(|S_n^*(\tau)|)\big),\, \tau \in \mathscr{T} \Big\}.$$

By Lemma 3.2 (and its proof) for any $\varepsilon_\tau = |S_n^*(\tau)|^{-1/2}$ there exists a finite set $\mathscr{M}_\tau^d = \mathscr{M}^d_{|S_n^*(\tau)|} \subset \mathscr{B}^d\big(r_n(|S_n^*(\tau)|)\big)$ with cardinality

$$|\mathscr{M}_\tau^d| \leq C\left(1 + \frac{2}{\varepsilon_\tau}\right)^{2(2+\bar{c})|S_n^*(\tau)|} \left(\frac{ed}{|S_n^*(\tau)|}\right)^{2(2+\bar{c})|S_n^*(\tau)|},$$

such that

$$C(S_n^*(\tau),\bar{c}) \subseteq \operatorname{conv}\left(\mathscr{M}_\tau^d\right),$$

where $C>0$ is an absolute constant, and

$$\max_{S:|S|=m} \sup_{u\in C(S,c)} \mathbb{E}_{nn}\big(F(X'_{ni,S}u)\big) \le \max_{v\in\mathscr{M}_m^d} \sup_{u\in\mathscr{B}^d_{\mathrm{supp}(v)}(v,\varepsilon)} \mathbb{E}_{nn}\big(F(X'_{ni}u)\big).$$

Define the enlarged function class

$$\mathscr{F}_2 = \Big\{(Y,X)\mapsto \rho_\tau(Y-X'\theta)-\rho_\tau(Y-X'\theta_n^*(\tau)) : \\ \theta\in\mathbb{R}^d,\, \theta-\theta_n^*(\tau)\in conv\left(\mathscr{M}_\tau^d\right),\, \tau\in\mathscr{T}\Big\},$$

and the reduced function class

$$\mathscr{F}_3 = \Big\{(Y,X)\mapsto \rho_\tau(Y-X'\theta_n^*(\tau)-X'\delta)-\rho_\tau(Y-X'\theta_n^*(\tau)) : \\ \delta\in\mathscr{B}^d_{\mathrm{supp}(v)}\big(v,\varepsilon_\tau\cdot r_n(|S_n^*(\tau)|)\big),\, v\in\mathscr{M}_\tau^d,\, \tau\in\mathscr{T}\Big\}.$$

Note that the supremum over class $\mathscr{F}_1$ is dominated by the supremum over function class $\mathscr{F}_2$. Hence, we proceed with function class $\mathscr{F}_2$.

Step 2. Equivalence relation and probability measure. Define the equivalence relation $\mathscr{R}_n\subseteq\mathscr{F}_2\times\mathscr{F}_2$ by

$$(f_{\tau^1,\delta^1,v^1},\, f_{\tau^2,\delta^2,v^2})\in\mathscr{R}_n \iff \Big\{|S_n^*(\tau^1)|=|S_n^*(\tau^2)|,\, v^1=v^2,\, v^1,v^2\in\mathscr{M}^d_{\tau^1}, \\ \delta^1,\delta^2\in\mathscr{B}^d_{\mathrm{supp}(v)}\big(v,\varepsilon_{\tau^1}\cdot r_n(|S_n^*(\tau^1)|)\big)\Big\},$$

and the probability measure $\nu_n:\sigma(\mathscr{F}_2/\mathscr{R}_n)\to[0,1]$ by $\nu_n(\emptyset)=0$ and

$$\nu_n(P_{\tau,v}) = c_\nu^{-1}\left(1+\frac{2}{\varepsilon_\tau}\right)^{-2(2+\bar{c})|S_n^*(\tau)|}\left(\frac{ed}{|S_n^*(\tau)|}\right)^{-2(3+\bar{c})|S_n^*(\tau)|},\ P_{\tau,v}\in\mathscr{F}_2/\mathscr{R}_n,$$

where $c_\nu>0$ is such that that $1=\sum_{P_{\tau,v}\in\mathscr{F}_2/\mathscr{R}_n}\nu_n(P_{\tau,v})$. Note that $0<c_\nu<eC$. Indeed,

$$c_\nu = \sum_{P_{\tau,\delta}\in\mathscr{F}_2/\mathscr{R}_n} c_\nu\nu_n(P_{\tau,v}) = \sum_{S\in\{S_n^*(\tau):\tau\in\mathscr{T}\}} |\mathscr{M}^d(S)|\left(1+\frac{2}{\varepsilon_{|S|}}\right)^{-2(2+\bar{c})|S|}\left(\frac{ed}{|S|}\right)^{-2(3+\bar{c})|S|}$$

$$
\begin{aligned}
&\leq C \sum_{S \in \{S_n^*(\tau):\tau\in\mathscr{T}\}} \left(\frac{ed}{|S|}\right)^{-2|S|} \\
&\leq C \sum_{k=1}^{\infty} \binom{d}{k} \left(\frac{ed}{k}\right)^{-2k} \\
&< \frac{C}{1-e^{-1}}.
\end{aligned}
$$

Further, note that each sub-class $P_{\tau,v} \in \mathscr{F}_2/\mathscr{R}_n$ has envelop function

$$
G_{\tau,v}(X) = \sup_{\delta \in \mathscr{B}^d_{\mathrm{supp}(v)}(v, \varepsilon_\tau \cdot r_n(|S_n^*(\tau)|))} |X'\delta| \leq \varepsilon_\tau \cdot \|X_{|S_n^*(\tau)|}\|_2 r_n(|S_n^*(\tau)|) + |X'v|.
$$

Also, note that $\mathscr{F}_3$ has the same envelope function.

Step 3. Refinement of Theorem 2.4 for classes of Lipschitz continuous and convex functions. Since the quantile loss function ρ_τ is convex, upon revisiting the proof of Theorem 2.4 and invoking Lemma 3.2 we observe that the following result is true (see also Corollary 2.2): For the above defined equivalence relation $\mathscr{R}_n$ on $\mathscr{F}_2$ there exist $c_0, c_1, c_2, N_0 > 0$ such that for all $n > N_0$ and for all $f_{\tau,\delta,v} \in \mathscr{F}_2$,

$$
\begin{aligned}
\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\delta,v})| \leq{}& \frac{c_{11}}{\sqrt{n}} \mathrm{E}\Big[\sup_{f\in[f_{\tau,\delta,v}]_{\mathscr{R}_n}} |\mathbb{G}_{nn}(f)|\Big] \\
&+ c_{11}\left(\mathrm{E}\Big[\sup_{f\in[f_{\tau,\delta,v}]_{\mathscr{R}_n}\cap\mathscr{F}_3} \frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f^2)|\Big]\right)^{1/2} \\
&\times \left(\frac{|S_n^*(\tau)|\log\left(ed/(|S_n^*(\tau)|\varepsilon_\tau)\right) + \log\log n}{n}\right)^{1/2} \\
&+ c_{11}\left(\sup_{f\in[f_{\tau,\delta,v}]_{\mathscr{R}_n}\cap\mathscr{F}_3} \bar{\mathrm{E}}_{nn}[f^2]\right)^{1/2} \\
&\times \left(\frac{|S_n^*(\tau)|\log\left(ed/(|S_n^*(\tau)|\varepsilon_\tau)\right) + \log\log n}{n}\right)^{1/2} \\
&+ c_{11}\left(\bar{\mathrm{E}}_{nn}[G_{\tau,v}^4]\right)^{1/4} \\
&\times \left(\frac{|S_n^*(\tau)|\log\left(ed/(|S_n^*(\tau)|\varepsilon_\tau)\right) + \log\log n}{n}\right)^{1/2} \left(\frac{\log\log n}{n}\right)^{1/4}.
\end{aligned} \tag{3.44}
$$

The remainder of the proof follows analogous to Step 3 of the proof of Lemma 2.9 by

replacing $|M|$ with $|S_n^*| = \sup_{\tau\in\mathscr{T}} |S_n^*(\tau)|$ and the maximum eigenvalue $\bar{\varphi}(|M|)$ of a model of size $|M|$ by the maximum $(|S^*|,\bar{c})$-restricted maximum eigenvalue. Conclude that there exists $c_{10} > 0$ such that for all $n > N_{12}$ and all $f_{\tau,\delta,v} \in \mathscr{F}_2$,

$$\frac{1}{\sqrt{n}}|\mathbb{G}_{nn}(f_{\tau,\theta_1,\theta_2})| \le c_{12}\bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})\frac{r_n(|S_n^*|)}{\sqrt{n}} + c_{12}\left(\frac{|S_n^*|\log(ed/|S_n^*|^{1/2}) + \log\log n}{n}\right)^{1/2} \bar{\varphi}_{\max}^{1/2}(S_n^*,\bar{c})r_n(|S_n^*|) \quad a.s.$$

□

3.6.2.3 Proof of Lemma 3.4

Proof. The proof of the first statement is almost identical to the proof of Lemma 2.3. The proof of the second statement is exactly the same as the one of Theorem 2.4 and therefore we do not repeat it here.

Recall the definition given right before the statement of the lemma. For $n \in \mathbb{N}$ let $[n] = \{1,\dots,n\}$ and for $I \subseteq \mathbb{N}$, $|I| < \infty$, and $S_n \in \mathscr{F}_n/\mathscr{R}_n$ define

$$\mathbb{V}_{nN}(S_n;I) = \mathrm{E}\left[\sup_{f\in S_n}\frac{1}{N}\sum_{i\in I\cap[N\wedge n]}\Big(f(X_{ni}) - f(\widetilde{X}_{ni})\Big)^2 \mid (X_{n1},\dots,X_{nn})\right],$$

and for $f \in S_n$,

$$\mathbb{G}_{nN}(f;I) = \frac{1}{\sqrt{N}}\sum_{i\in I\cap[N\wedge n]}\Big(f(X_{ni}) - \mathrm{E}[f(X_{ni})]\Big).$$

Note that we do not divide the sums by the number of summands $|I\cap[N\wedge n]|$ but by the (larger) number N. Thus, we do not form averages and "waste" a factor $|I\cap[N\wedge n]|/|N|$. It turns out that wasting this factor allows us to apply Lemma 2.2 without changes and this is what makes the proof simple. Since the resulting bound is good enough for our purposes, we do not aim at deriving a sharper bound which would require the development of a concentration inequality for averages of top order statistics.

We write Lt to denote the function $\max(1,\log t)$ and LLt for the composition $L(Lt)$, $t \ge 0$. Let

$$\mathbb{U}_{nNN'}(S_n;I) = 7\mathrm{E}\Big[\sup_{f\in S_n}|\mathbb{G}_{nN}(f;I)|\Big] + 14\Big(\mathbb{V}_{nN}(S_n;I) + \mathrm{E}\left[\mathbb{V}_{nN}(S_n;I)\right]\Big)^{1/2}$$

$$\times 2^{1/2}\left(LLN' + 2|I|\log(2eN'/|I|) + \log\frac{1}{\nu_n(S_n)}\right)^{1/2},$$

and for arbitrary $\eta > 0$ and $m \in \mathbb{N}$,

$$A_m = \left\{ \max_{n\geq 2^m} \sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \max_{I\subseteq[n]} \sup_{f\in S_n} \frac{|\mathbb{G}_{nn}(f;I)|}{\mathbb{U}_{nnn}(S_n;I)} > 1+\eta \right\}.$$

Then, by three applications of the union bound and by Lemma 2.2,

$$\begin{aligned}
\mathrm{P}(A_m) &\leq \max_{n\in\mathbb{N}} \mathrm{P}\left(\max_{N\geq 2^m} \sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \max_{I\subseteq[N]} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f;I)|}{\mathbb{U}_{nNN}(S_n;I)} > 1+\eta \right) \\
&\leq \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \mathrm{P}\left(\max_{\ell\geq m} \max_{2^\ell < N \leq 2^{\ell+1}} \max_{I\subseteq[N]} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f;I)|}{\mathbb{U}_{nNN}(S_n;I)} > 1+\eta \right) \\
&\leq \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sum_{\ell=m}^{\infty} \sum_{I\subseteq[2^{\ell+1}]} \mathrm{P}\left(\max_{N\leq 2^{\ell+1}} \sup_{f\in S_n} \frac{|\mathbb{G}_{nN}(f;I)|}{\mathbb{U}_{nN2^\ell}(S_n;I)} > 1+\eta \right) \\
&\leq 12e \max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \sum_{\ell=m}^{\infty} \sum_{k=1}^{2^{\ell+1}} \binom{2^{\ell+1}}{k} e^{-(1+\eta)^2[-\log\nu_n(S_n)+(\log\ell)+(\log\log 2)+2k\log(2e2^\ell/k)]} \\
&\leq 12e \left(\max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \nu_n(S_n)^{(1+\eta)^2} \right) \sum_{\ell=m}^{\infty} \sum_{k=1}^{2^{\ell+1}} \left(\frac{e2^{\ell+1}}{k}\right)^{-k} e^{-(1+\eta)^2[\log\ell+\log\log 2)]} \\
&\leq 12e \left(\max_{n\in\mathbb{N}} \sum_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \nu_n(S_n)^{(1+\eta)^2} \right) \left(\sum_{k=1}^{\infty} e^{-k} \right) \sum_{\ell=m}^{\infty} e^{-(1+\eta)^2[\log\ell+\log\log 2)]} \\
&< \infty. \qquad (3.45)
\end{aligned}$$

The sequence of sets $(A_m)_{m\in\mathbb{N}}$ is decreasing. Thus, by continuity of the P-measure,

$$\begin{aligned}
\mathrm{P}\left(\limsup_{n\to\infty} \sup_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \max_{I\subseteq[n]} \sup_{f\in S_n} \frac{|\mathbb{G}_{nn}(f;I)|}{\mathbb{U}_{nnn}(S_n;I)} > 1+\eta \right) &= \mathrm{P}\left(\lim_{m\to\infty} A_m \right) \\
&= \lim_{m\to\infty} \mathrm{P}(A_m) = 0, \qquad (3.46)
\end{aligned}$$

where the last equality follows from eq. (2.34). Since eq. (3.46) holds for all $\eta > 0$ we conclude that there exists an $N_0 > 0$ such that for all $n > N_0$,

$$\max_{S_n\in\mathscr{F}_n/\mathscr{R}_n} \max_{I\subseteq[n]} \sup_{f\in S_n} \frac{|\mathbb{G}_{nn}(f;I)|}{\mathbb{U}_{nnn}(S_n;I)} \leq 1 \quad a.s.$$

This establishes the claim. □

CHAPTER 4

On the Predictive Risk in Misspecified Quantile Regression

4.1 Introduction

Predictive modeling is at the core of many scientific disciplines, including business, engineering, finance, and public health. A natural way to gauge the predictive capability of a statistical model is to estimate its predictive risk. The systematic study of the risk of a statistical procedure traces back to at least Stein (1981). Since then, the concept of risk has become an integral part of applied statistical modeling: predictive risk is routinely used to assess the complexity of statistical modeling procedures (e.g. Akaike, 1992; Mallows, 1973; Foster and George, 1994) to compare statistical models across different fitting techniques (e.g. Hastie and Tibshirani, 1990; Ye, 1998), and to choose tuning parameters that control bias-variance trade-offs (e.g. Donoho and Johnstone, 1995; Kou and Efron, 2002). In several special cases, Stein's (1981) theory of unbiased risk estimation provides simple estimates for the risk of a statistical model. However, in general, there does not exist a unified approach to estimating the predictive risk of a statistical model or procedure.

In this paper, we focus on the predictive risk of possibly misspecified quantile regression models. In addition to its role in applied statistical modeling as outlined above, in recent years the predictive risk from quantile models has also garnered significant interest in finance and risk management to assess the value-at risk and expected shortfall of investments (e.g. Xiao et al., 2015; Gaglianone et al., 2011; Engle and Manganelli, 2004; Chernozhukov and Umantsev, 2001) and to solve portfolio choice problems in the framework of Kahneman and Tversky's (1979) prospect theory (e.g. Cahuich and Hernández-Hernández, 2013; He and Zhou, 2011; Bassett et al., 2004)

We contribute to the theory of predictive risk evaluation of quantile regression models by deriving two (asymptotic) characterizations of the expected optimism of the in-sample risk and proposing a uniformly consistent, de-biased estimator of the predictive risk. Our first

characterization of the expected optimism provides a characterization comparable to Efron's (2004) covariance penalty and Tibshirani and Knight's (1999) covariance inflation criterion. The second characterization is related to robust and generalized Akaike-type selection criteria for misspecified quantile regression models (e.g. Lv and Liu, 2014; Portnoy, 1997; Burman and Nolan, 1995) and helps to assess the impact of under- and over-fitting on the predictive risk. Both characterizations show that large part of the expected optimism can be attributed to a nonlinear function of the quantile level, the conditional density of the response variable given the predictors and the (weighted) covariance matrix of the predictors. Specializing to location models, we glean additional insight into the expected optimism and its functional dependence on the conditional density and the number of predictors. As a consequence, the commonly used notion of effective degree of freedom for a statistical model has a richer content for misspecified models.

The second characterization of the expected optimism lends itself to a simple plug-in estimator. We establish its uniform consistency over a class of candidate models and, based on this result, propose a uniformly consistent, de-biased estimate of the predictive risk. Our theoretical analysis indicates that the de-biased estimator is particularly relevant in the case in which the dimension of candidate models grows at least in the order of the square root of the sample size. Empirical evidence suggests that the de-biasing procedure is practically relevant even when the model size is fixed and relatively small compared to the sample size. A comparison of our de-biased estimate against the popular method of cross-validation is favorable for our procedure.

To allow broad applicability of our the theoretical results, we develop our theory in a triangular array framework in which the number of predictor variables may grow with the sample size. We only require minimal assumptions on the joint distribution of the response and predictor variables. Notably, the response and the predictor variables can both be unbounded, their marginal distributions can be non-Gaussian, and their relationship (i.e. the conditional quantile functions) can be linear, nonlinear or nonparametric. Thus, our framework for quantile regression generalizes the frameworks of Lee (2016); Noh et al. (2013); Angrist et al. (2006); Kim and White (2003) who consider misspecified quantile regression models with a fixed number of parameters. Unlike the recent literature on quantile regression based on series, semi- and nonparametric estimators we do not assume that the misspecification error vanishes as more predictors are included in the regression function (Belloni et al., 2017; Chao et al., 2017). Naturally, our results continue to hold if the model is (asymptotically) correctly specified.

We organize this article as follows: In Section 4.2 we lay out a general framework for misspecified quantile regression models. We introduce necessary terminology and discuss

how to define the predictive risk of potentially misspecified quantile regression models. In Section 4.3 we derive two asymptotic characterizations of the expected optimism of the in-sample risk and discuss insights that we gain from these characterizations. In Section 4.4 we propose a nonparametric plug-in estimator for one of the asymptotic characterizations of the expected optimism and use it to construct a de-biased estimate of the predictive risk. We establish uniform consistency of both estimators. In Section 4.5 we report numerical evidence that our estimates of the expected optimism and the predictive risk are on target, and that the predictive risk estimate can be better than the commonly-used cross-validation approach. We conclude in Section 4.6 with additional remarks, and present all proofs in 4.7.The Supplementary Materials in 4.8 contain additional simulation results.

4.2 Misspecified quantile regression and predictive risk

4.2.1 Notation and framework

The setting of interest is a high-dimensional triangular array $\mathscr{D}_n = \{(Y_{ni}, X_{ni})\}_{i=1}^n$, where $(Y_{ni}, X_{ni}) \in \mathbb{R} \times \mathscr{X}$ are row-wise independent random vectors with distribution F_n which may change with the sample size n. As per convention the scalar variable Y_{ni} denotes the response variable and the vector $X_{ni} \in \mathscr{X}$ denotes a vector of covariates. We denote by $F_{Y_n|X_n}$ the conditional distribution of Y_{ni} given X_{ni}. We use subscripts on the expectation operator $\mathbb{E}$ to specify to which random variable the operator is applied to, i.e. $\mathbb{E}_{(Y_{n1},X_{n1})}$ means that expectation is only taken over (Y_{n1}, X_{n1}) whereas $\mathbb{E}_{\mathscr{D}_n}$ means that expectation is taken over the entire triangular array $\mathscr{D}_n$. Let

$$x \mapsto Z(x) = \big(Z_1(x), \ldots, Z_d(x)\big) \tag{4.1}$$

be a mapping from $\mathscr{X}$ into $\mathbb{R}^d$ and call the transformed covariates $Z(X_{n1}), \ldots, Z(X_{nn})$ predictor variables. We consider the case where the dimension d of the predictor variables growths with the sample size n and may be much larger than n. We call a subset $S \subseteq \{1, \ldots, d\}$ of predictors $Z(X_{ni})$ a model and write

$$Z_S(X_{ni}) = \big(Z_j(X_{ni})\big)_{j \in S}. \tag{4.2}$$

Denote the collection of models under consideration by M. We allow M to be as large as the power set of $\{1, \ldots, d\}$ and to grow with the sample size n. We write $|S|$ for the cardinality of a model S and denote the largest cardinality of models in M by m. Clearly, we have $m \leq d$.

The purpose of linear quantile regression is to approximate the true conditional quantile function (CQF) of Y_{ni} given X_{ni},

$$Q_{Y_n}(\tau|X_{ni}) = \inf\left\{y : F_{Y_n|X_n}(y|X_{ni}) \geq \tau\right\}, \tag{4.3}$$

by a linear function of the predictor variables $Z(X_{ni})$. To this end, we assume that the vectors of predictor variables $Z(X_{ni})$ consist of series functions with good approximation properties such as indicators, B-splines, regression splines, polynomials, Fourier series, or wavelets (e.g. Belloni et al., 2017; Chao et al., 2017). However, unlike them we do not require that the approximation error vanishes as the number of predictors m increases, i.e. we allow for persistent misspecification. We define the vector of regression coefficients $\theta^{\tau}_{n,S} = (\theta^{\tau}_{n1}, \ldots, \theta^{\tau}_{n|S|})'$ associated with model S as the solution to the quantile regression problem

$$\min_{\theta \in \mathbb{R}^{|S|}} \mathbb{E}_{\mathscr{D}_n}\left[\rho_\tau\big(Y_{n1} - Z_S(X_{n1})'\theta\big) - \rho_\tau\big(Y_n - Q_{Y_n}(\tau|X_{n1})\big)\right], \tag{4.4}$$

and the vector of estimated regression coefficients $\hat{\theta}^{\tau}_{n,S} = (\hat{\theta}^{\tau}_{n1}, \ldots, \hat{\theta}^{\tau}_{n|S|})'$ as the solution to the sample quantile regression problem

$$\min_{\theta \in \mathbb{R}^{|S|}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(Y_{ni} - Z_S(X_{ni})'\theta\big), \tag{4.5}$$

where $\rho_\tau(u) = (\tau - 1\{u \leq 0\})$ is the check loss (Koenker, 2005). The estimate of the true CQF of Y_n given X_n based on model S is given as

$$\widehat{Q}_{Y_n}(\tau|X_n, S) = Z_S(X_n)'\hat{\theta}^{\tau}_{n,S}. \tag{4.6}$$

Koenker and Bassett (1978) show that under mild conditions $\widehat{Q}_{Y_n}(\tau|X_n, S)$ is a consistent estimate of $Q_{Y_n}(\tau|X_n)$ if the true CQF is indeed linear in $Z_S(X_{ni})$ and the dimension of the predictor variables is fixed. The results on general M-estimators in He and Shao (2000) (Theorem 1), semi-parametric quantile regression Chao et al. (2017) and quantile series estimators Belloni et al. (2017) extend this consistency result to cases in which the dimension of the predictors m increases with the sample size n and the model S is (asymptotically) correctly specified.

Our setting of persistent misspecification is closely related to the framework of Angrist et al. (2006) and extends it to the case of growing numbers of predictors. Angrist et al. (2006) show (Theorem 1) that the solution to the quantile regression problem (4.4) under persistent misspecification can be interpreted as the best linear approximation to the true

CQF under a weighted square loss with (random) weights that down weight regions of the parameter space in which the conditional density of Y_{n1} given X_{n1}, $f_{Y_n|X_n}$, is low, i.e. $\theta^{\tau}_{n,S}$ equivalently solves

$$\min_{\theta\in\mathbb{R}^{|S|}} \mathbb{E}_{\mathscr{D}_n}\left[\omega^{\tau}\cdot\left(Q_{Y_n}(\tau|X_{n1})-Z_S(X_{n1})'\theta\right)^2\right], \tag{4.7}$$

where

$$\omega^{\tau}\equiv\omega^{\tau}(X_{n1},S)=\int_0^1(1-u)f_{Y_n|X_n}\Big(u\cdot Z_S(X_{n1})'\theta+(1-u)\cdot Q_{Y_n}(\tau|X_{n1})|X_{n1}\Big)du.$$

In this sense, even under persistent misspecification the vector of regression coefficients $\theta^{\tau}_{n,S}$ and its estimate $\hat{\theta}^{\tau}_{n,S}$ capture important features of the true CQF. Within this framework we can analyze the predictive risk of a model regardless of whether it is correctly specified or not. This allows us to derive theoretical results that hold for a wide range of modeling approaches, including linear, nonlinear, and nonparametric models.

4.2.2 Predictive risk and expected optimism

To introduce some terminology and to rationalize our approach to the predictive risk of potentially misspecified quantile regression models we briefly review the literature on predictive risk estimation.

Suppose that a model f is fitted to some data $\mathscr{Z}_n=\{Z_1,\ldots,Z_n\}$ producing an estimate $\hat{\mu}_n=f(\mathscr{Z}_n)$ for target μ. Predictive risk evaluation tries to assess how well $\hat{\mu}_n$ predicts μ at a future data point Z^0 independently generated from the same mechanism that produced $\mathscr{Z}_n$. To measure the error between $\hat{\mu}_n$ and μ one chooses a loss function L and defines the predictive risk as the average loss over current and future data, i.e.

$$\mathbb{E}_{\mathscr{Z}_n,Z^0}\left[L\Big(\mu(Z^0),\hat{\mu}_n(Z^0)\Big)\right]. \tag{4.8}$$

Two statistical theories have been developed to estimate this quantity, cross-validation (e.g. Stone, 1974, 1977; Allen, 1974; Golub et al., 1979; Wahba, 1990; Efron, 1983, 1986, 2004; Efron and Tibshirani, 1997) and covariance penalties, which include techniques such as Mallows's (1973) Cp, Akaike's (1998) information criterion (AIC) and final prediction error (FPE), Takeuchi's (1976) information criterion (TIC), and Stein's (1981) unbiased risk estimate (SURE). Neither of the two theories is strictly superior over the other: On the one hand, cross-validation techniques tend to produce estimates of the predictive risk that have a higher variance than estimates based on covariance penalties, since they split the

sample into test and training sets and thereby reduce the number of samples from which $\hat{\mu}_n$ is estimated (e.g. Efron, 2004). On the other hand, covariance penalties have only been derived for a limited number of loss functions, namely the square loss and the "q class of error measures" (Efron, 1986), whereas cross-validation techniques can be applied to any loss function L. Lastly, cross-validation techniques target the predictive risk directly, while covariance penalties provide as an intermediate result an estimate of the bias of the in-sample risk when used as estimate of the predictive risk. Following the terminology introduced by Efron (2004) we call the negative bias the "expected optimism" of the in-sample risk,

$$b_n(L,\mu) = \mathbb{E}_{\mathscr{Z}_n,Z^0}\left[L\Big(\mu(Z^0),\hat{\mu}_n(Z^0)\Big)\right] - \mathbb{E}_{\mathscr{Z}_n}\left[\frac{1}{n}\sum_{i=1}^{n} L\Big(\mu(Z_i),\hat{\mu}_n(Z_i)\Big)\right]. \tag{4.9}$$

The expected optimism is typically non-negative since the in-sample risk is usually downward biased as an estimate of the predictive risk. Given a consistent estimate $\hat{b}_n(L,\mu)$ of $b_n(L,\mu)$ one obtains a consistent and de-biased estimate of the predictive risk via

$$\frac{1}{n}\sum_{i=1}^{n} L\Big(\mu(Z_i),\hat{\mu}_n(Z_i)\Big) + \hat{b}_n(L,\mu). \tag{4.10}$$

The expected optimism is relevant beyond predictive risk estimation in model selection (e.g. Akaike, 1992; Foster and George, 1994), model comparison (e.g. Hastie and Tibshirani, 1990; Tibshirani and Knight, 1999; Kou and Efron, 2002) and computation of generalized degrees of freedom (e.g. Ye, 1998). Because of these potential applications we develop our predictive risk estimator for quantile regression models along the line of covariance penalties.

4.2.3 Predictive risk and expected optimism in quantile regression

We discuss the choice of the loss function to measure the predictive risk of a potentially misspecified quantile regression model S and define the associated expected optimism.

Let (Y_n^0, X_n^0) be a pair of data points drawn from F_n and independent of sample $\mathscr{D}_n = \{(Y_{ni}, X_{ni})\}_{i=1}^n$. Fix a model $S \subseteq \{1,\ldots,d\}$ and consider the estimate of the CQF of Y_n^0 given X_n^0 based on model S and sample $\mathscr{D}_n$, i.e.

$$\widehat{Q}_{Y_n^0}(\tau|X_n^0,S) = Z_S(X_n^0)'\hat{\theta}_{n,S}^{\tau}. \tag{4.11}$$

Since the true CQF of Y_n^0 given X_n^0, $Q_{Y_n^0}(\tau|X_n^0)$, is not an observable statistic given the data $\mathscr{D}_n$ and (Y_n^0, X_n^0), risk measures which assess directly the difference between estimate

$\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)$ and target $Q_{Y_n^0}(\tau|X_n^0)$, such as the mean squared prediction error or the mean absolute prediction error, do not have (simple) sample analogues. We therefore propose the following risk measure which depends only on observables.

Definition 4.1 (Predictive risk). *The predictive risk of quantile regression model S is*

$$\mathrm{PR}_n^\tau(S) = \mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\rho_\tau\big(Y_n^0 - \widehat{Q}_{Y_n^0}(\tau|X_n^0,S)\big) - \rho_\tau(Y_n^0)\right],$$

where (Y_n^0,X_n^0) is a pair of data points drawn from F_n and independent of sample $\mathscr{D}_n$.

The associated expected optimism of using the in-sample risk $\frac{1}{n}\sum_{i=1}^n \Big(\rho_\tau\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S) - \rho_\tau(Y_{ni})\Big)$ as an estimate of the predictive risk is defined as follows.

Definition 4.2 (Expected Optimism). *The expected optimism of quantile regression model S is*

$$b_n^\tau(S) = \mathrm{PR}_n^\tau(S) - \mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n\Big(\rho_\tau\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S)\big) - \rho_\tau(Y_{ni})\Big)\right].$$

Several comments are in order with regard to these two definitions: First, the reason for subtracting $\rho_\tau(Y_n^0)$ in Definition 4.1 (and $\rho_\tau(Y_{ni})$ in Definition 4.2) is purely technical: It allows us to dispense with moment conditions on the response variable Y_n^0. To see this, note that the check loss ρ_τ is Lipschitz continuous and hence the predictive risk $\mathrm{PR}_n^\tau(S)$ is upper bounded by $\mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\big|Z_S(X_n^0)'\hat{\theta}_{n,S}^\tau\big|$. For this expected value to be finite it suffices that the CQF of Y_n^0 given X_n^0 has finite second moments (e.g. Angrist et al., 2006).

Second, the expected optimism associated with the check loss ρ_τ can be related to generalized degrees of freedoms (e.g. Ye, 1998) and to model selection criteria in quantile regression (Portnoy, 1997), in robust regression (Burman and Nolan, 1995), in misspecified (linear) regression models (Bozdogan, 2000), and in misspecified generalized linear models (Lv and Liu, 2014). In particular, we note that Burman and Nolan's (1995) criterion reduces to a special case of our estimate of the predictive risk.

Third, the predictive risk based on the check loss ρ_τ has garnered significant interest in finance and risk management. For example, it is used in the context of value-at-risk (e.g. Xiao et al., 2015; Gaglianone et al., 2011), conditional value-at-risk and expected shortfall (e.g. Engle and Manganelli, 2004; Chernozhukov and Umantsev, 2001) and portfolio choice problems with Choquet expectation (e.g. Cahuich and Hernández-Hernández, 2013; He and Zhou, 2011; Bassett et al., 2004; Tversky and Kahneman, 1992).

Fourth, as an immediate consequence of Theorem 1 in Angrist et al. (2006) we have the following relation between the predictive risk based on the check loss ρ_τ and the (weighted)

mean squared prediction error.

Proposition 4.1. *Suppose that* $\mathbb{E}_{X_n^0}\left[Q_{Y_n^0}(\tau|X_n^0)\right]$ *is finite and* $\theta_{n,S}^\tau$ *uniquely solves* (3.2). *Then, the predictive risk satisfies*

$$\begin{aligned}\mathrm{PR}_n^\tau(S) = \mathbb{E}_{\mathscr{D}_n,X_n^0}\left[\omega^\tau(\mathscr{D}_n,X_n^0,S)\cdot\left(Q_{Y_n^0}(\tau|X_n^0)-\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)\right)^2\right] \\ +\left(\mathbb{E}_{X_n^0}\left[F_{Y_n^0|X_n^0}(0|X_n^0)\right]-\tau\right).\end{aligned}$$

where

$$\omega^\tau(\mathscr{D}_n,X_n^0,S)=\int_0^1(1-u)f_{Y_n^0|X_n^0}\left(u\cdot\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)+(1-u)\cdot Q_{Y_n^0}(\tau|X_n^0)|X_n^0\right)du.$$

Note that the second term in the quadratic expansion of the $\mathrm{PR}_n^\tau(S)$ is a constant in the interval $(-\tau,1-\tau)$ and independent of sample $\mathscr{D}_n$, quantile model S, and estimate $\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)$; only the first term depends on the data and the model. This first term is a weighted version of the mean squared prediction error with weights that down-weight regions on the real line where the conditional density of Y_n^0 given X_n^0 is low. Angrist et al. (2006) demonstrate that for most practical purposes the weight ω^τ tends to be constant across X_n^0. Thus, we can think of the predictive risk $\mathrm{PR}_n^\tau(S)$ as a nearly affine transformation of the mean squared prediction error $\mathbb{E}_{\mathscr{D}_n,X_n^0}\left[\left(Q_{Y_n^0}(\tau|X_n^0)-\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)\right)^2\right]$.

Lastly, the expected optimism controls the mean absolute prediction error between the estimate $\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)$ and the best linear approximation to the true CQF of Y_n^0 given X_n^0 based on model S, i.e. $Z_S(X_n^0)'\theta_{n,S}^\tau$.

Proposition 4.2. *Under the assumptions of Proposition 4.1,*

$$\mathbb{E}_{\mathscr{D}_n,X_n^0}\left[\left|Z_S(X_n^0)'\theta_{n,S}^\tau-\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)\right|\right]\leq b_n^\tau(S).$$

The proof follows from the lower bound $b_n^\tau(S)\geq\mathbb{E}_{\mathscr{D}_n,X_n^0}\Big[\rho_\tau\big(Y_n^0-\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)\big)$ $-\rho_\tau\big(Y_n^0-Z_S(X_n^0)'\theta_{n,S}^\tau\big)\Big]$ and Theorem 1 in Angrist et al. (2006).

In summary, the predictive risk as defined above has applications in finance and risk management and quantifies the uncertainty of using the estimate $\widehat{Q}_{Y_n^0}(\tau|X_n^0,S)$ to predict the true CQF of Y_n^0 given X_n^0. The associated expected optimism is of independent interest as it relates to commonly used model selection criteria.

4.2.4 Technical assumptions

For the theoretical investigations of the predictive risk and the expected optimism of potentially misspecified quantile regression models we require several assumptions which we discuss in this section. Since the quantile level τ is always pre-specified, we suppress the dependence on τ in some notation. Recall that $S \subseteq \{1,\dots,d\}$, $|S| \leq m$, and that M is a subset of the power set of $\{1,\dots,d\}$. Throughout, we assume that M contains at least two models, i.e. $|M| \geq 2$, and that $n \geq 16$, i.e. $\log\log n \geq 1$.

(A1) *The data $(Y_{ni}, X_{ni}) \in \mathbb{R} \times \mathscr{X}$ are row-wise independent random vectors with distribution F_n, where F_n may change with the sample size n.*

(A2) *The conditional density $f_{Y_n|X_n}$ of Y_n given X_n is uniformly bounded from above, i.e. there exits $\nu_+ < \infty$ such that*

$$\limsup_{n\to\infty} \sup_{a\in\mathbb{R}} \sup_{x\in\mathbb{R}^d} \left| f_{Y_n|X_n}(a|x) \right| \leq \nu_+.$$

(A3) *The conditional density $f_{Y_n|X_n}$ of Y_n given X_n is α-Hölder continuous for $\alpha \in \left[\frac{1}{2}, 1\right]$, i.e. there exists a constant $\nu_H > 0$ such that for any $a, b \in \mathbb{R}$,*

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \left| f_{Y_n|X_n}(a|x) - f_{Y_n|X_n}(b|x) \right| \leq \nu_H |a-b|^\alpha.$$

(A4) *The maximum eigenvalue of the matrix of second moments is uniformly bounded from above, i.e. there exists $\lambda_+ < \infty$ such that*

$$\limsup_{n\to\infty} \max_{S\in M} \lambda_{\max}\left(\mathbb{E}_{X_n}\left[Z_S(X_n) Z_S(X_n)' \right] \right) \leq \lambda_+,$$

and the minimum eigenvalue of the weighted second moment matrix is bounded from below by $\lambda_n > 0$,

$$\min_{S\in M} \lambda_{\min}\left(\mathbb{E}_{X_n}\left[f_{Y_n|X_n}\left(Z_S(X_n)' \theta^\tau_{n,S} | X_n \right) Z_S(X_n) Z_S(X_n)' \right] \right) > \lambda_n.$$

In the above assumptions the uniformity in n is necessary since we consider triangular arrays. Assumptions (A1), (A2), and (A3) with $\alpha = 1$ are fairly standard in the quantile regression literature (e.g. Angrist et al., 2006; Belloni et al., 2017; Chao et al., 2017). It is possible to relax the (implicit) assumption that the random variables are identically distributed within reach row; in fact independence suffices for our results. However, we do

not pursue these refinements in the present paper. The stringentness of Assumption (A4) depends on how fast λ_n is allowed to go to zero. We require the following technical rate condition on λ_n:

(A5) The minimum eigenvalue of the matrix of second moments, λ_n, is bounded below asymptotically in the following way:

$$\lambda_n \gtrsim \left(\frac{m \log |M| \log\log n}{n} \right)^{1/2 - 1/(4\alpha)}.$$

This rate condition is purely technical and difficult to motivate. Clearly, the condition is less stringent the larger α, i.e. the smoother the conditional density $F_{Y_n|X_n}$ of Y_n given X_n. In particular, if $\alpha = 1/2$, we require $\lambda_n = O(1)$; whereas in the case of a continuous conditional density, we allow $\lambda_n = O\left((m \log |M| \log\log n)^{1/4} n^{-1/4} \right)$. The rate condition relaxes the stronger boundedness assumptions on the largest and smallest eigenvalue of the weighted second moment matrix that prevail in the literature on quantile regression (Koenker, 2005). Together with the upper bound on the largest eigenvalue of the expected value of the Gram-matrix the rate condition implies that $m \lesssim n$. This is a much weak condition on the growth rate of the number of predictors than has been proposed in recent work on (misspecified) quantile regression with increasing number of predictors. E.g. Belloni et al. (2017) and Chao et al. (2017) require that $\zeta_m = \sup_{x \in \mathscr{X}} \|Z(x)\|_2 < \infty$ satisfies $m \zeta_m^2 (\log n)^2 = o(n)$. If the predictors are element-wise bounded, this amounts to the condition $m^2 (\log n)^2 = o(n)$. We shall see that our relaxed assumption on the growth rate is important in the theoretical analysis of the proposed estimate for the predictive risk in Section 4.4.

Lastly, we introduce the following moment condition on the predictors:

(A6) The vector $Z(X_n) = \big(Z_1(X_n), \ldots, Z_d(X_n)\big)$ is a vector of random variables with finite $8 + \delta$ moment, for some $\delta > 0$. In particular, for $1 \leq k \leq 8$, there exist constants $\mu_k > 0$ such that

$$\limsup_{n \to \infty} \max_{j = 1, \ldots, d} \left(\mathbb{E}_{X_n} \left[\left| Z_j(X_n) \right|^{k+\delta} \right] \right)^{1/(k+\delta)} \leq \mu_k.$$

This condition is significantly weaker than the uniform boundedness assumption on the map Z imposed in Belloni et al. (2017) and Chao et al. (2017) (i.e. $\zeta_m = \sup_{x \in \mathscr{X}} \|Z(x)\|_2 < \infty$). Again, uniformity in n is necessary since we consider triangular arrays.

4.3 Two asymptotic characterizations of the expected optimism

4.3.1 The covariance form of the expected optimism

In the case of ordinary least squares the expected optimism can be evaluated via Mallows's (1973) Cp, in the case of nonlinear least squares with Gaussian errors the expected optimism can be estimated via Stein's (1981) divergence formula, and for loss functions that belong to Efron's (2004) "q class of error measures" the expected optimism can be expressed as a function of the covariance of two observable quantities.

Since the expected optimism $b_n^\tau(S)$ from Definition 4.2 is based on the check loss ρ_τ none of above three results applies. However, the expected optimism $b_n^\tau(S)$ satisfies an approximate version of Efron's (2004) covariance representation.

Theorem 4.1 (Covariance Form of the Expected Optimism). *Suppose that Assumptions (A1) – (A6) from Section 4.2.4 hold. Then,*

$$b_n^\tau(S) = tr\left(Cov\left(\frac{1}{n}\sum_{i=1}^{n} Z_S(X_{ni})\varphi_\tau\left(Y_{ni} - Z_S(X_{ni})'\theta_{n,S}^\tau\right),\, \hat{\theta}_{n,S}^\tau - \theta_{n,S}^\tau\right)\right) \;+\; r_{n,1}(S),$$

where $\varphi_\tau(u) = \tau - 1\{u < 0\}$ *and*

$$\sup_{S\in M}\left|r_{n,1}(S)\right| = O\left(\frac{1}{\lambda_n^{3/2}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right)\;\; a.s.$$

We postpone a discussion of the rate of the remainder term to the next section. Focusing instead on the leading term of above approximation we observe the following: If the true CQF is indeed linear in $Z_S(X_n)$, i.e. $Q_{Y_n}(\tau|X_n) = Z_S(X_n)'\theta_{n,S}^\tau$, then the leading term of the optimism $b_n^\tau(S)$ can be re-formulated as

$$\frac{1}{n}\sum_{i=1}^{n} Cov\left(\varphi_\tau\left(Y_{ni} - Q_{Y_n}(\tau|X_{ni})\right),\, \widehat{Q}_{Y_n}(\tau|X_{ni},S)\right). \tag{4.12}$$

Thus, in this case the expected optimism is essentially a simple function of the covariances between estimates $\widehat{Q}_{Y_n}(\tau|X_{ni},S)$ and targets $Q_{Y_n}(\tau|X_{ni})$, $i = 1,\ldots,n$. This is reminiscent of Efron's (2004) results for the "q class of error measures".

Re-writing the leading term of the optimism $b_n^\tau(S)$ as the expected value of the gradient

of the check loss and the centered regression vector,

$$\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau\big(Y_{ni}-Z_S(X_{ni})'\theta^\tau_{n,S}\big)Z_S(X_{ni})'(\hat{\theta}^\tau_{n,S}-\theta^\tau_{n,S})\right], \tag{4.13}$$

we gain two more insights:

First, the covariance form of the expected optimism can be viewed as a first order linearization of the check loss. In particular, the covariance form is the (expected value) of the directional derivative of the check loss in direction $\hat{\theta}^\tau_{n,S}-\theta^\tau_{n,S}$ and evaluated at the vector of regression coefficients $\theta^\tau_{n,S}$. Since the check loss is convex this directional derivative is always non-negative, i.e. the leading term of the expected optimism non-negative. This confirms our statistical intuition that the bias of the in-sample risk as estimate of the predictive risk is negative.

Second, using the naive sample analogue $\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau\big(Y_{ni}-Z_S(X_{ni})'\hat{\theta}^\tau_{n,S}\big)X'_{ni,S}\hat{\theta}^\tau_{n,S}$ to estimate the expected optimism will inevitably result in a poor estimate because the gradient evaluated at its sample minimizer $\hat{\theta}^\tau_{n,S}$ is close to zero. Thus, even though the approximate covariance form does not dependent on the future (unattainable) data point (Y^0_n,X^0_n), it does not allow us to entirely bypass the computation of the expected value with respect to the unknown distribution F_n. A similar observation was first made by Efron (1986) about his covariance penalties. To overcome this difficulty he proposes a parametric bootstrap approach; below we show a different approach which does not rely on re-sampling.

4.3.2 The trace form of the expected optimism

As noted in Section 4.3.1 the predictive risk under check loss ρ_τ is almost an affine transformation of the mean squared prediction error. We might therefore expect that the expected optimism can be approximated by an expression similar to the penalty term in Mallows's (1973) Cp or Takeuchi's (1976) TIC. The following theorem shows that this intuition is correct.

Theorem 4.2 (Trace Form of the Expected Optimism). *Suppose that Assumptions (A1) – (A6) from Section 4.2.4. Then,*

$$b^\tau_n(S)=\frac{1}{n}tr\left(D^\tau_{n,0}(S)^{-1}D^\tau_{n,1}(S)\right)+r_{n,2}(S),$$

where

$$D^\tau_{n,0}(S)=\mathbb{E}_{X_{n1}}\left[f_{Y_n|X_n}\big(Z_S(X_{n1})'\theta^\tau_{n,S}|X_{n1}\big)Z_S(X_{n1})Z_S(X_{n1})'\right],$$

$$D^{\tau}_{n,1}(S) = \mathbb{E}_{X_{n1}}\left[\varphi^2_{\tau}\left(Y_{n1} - Z_S(X_{n1})'\theta^{\tau}_{n,S}\right)Z_S(X_{n1})Z_S(X_{n1})'\right],$$

with $\varphi_{\tau}(u) = \tau - 1\{u < 0\}$ *and*

$$\sup_{s\in M}\left|r_{n,2}(S)\right| = O\left(\frac{1}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right) \quad a.s.$$

We observe the following: First, under Assumptions (A1) – (A6) the trace from is roughly of order $O\left(\lambda_n^{-1}n^{-1}|S|\right)$ and hence dominates the remainder term $r_{n,2}(S)$. Therefore the trace form is a meaningful approximation of the expected optimism.

Second, in the literature on robust estimation the trace form is also known as "expected self-influence", i.e. the average influence that an observation has on its own fitted value (e.g. Hampel et al., 2005, p. 317). While at hindsight the connection between expected optimism and "expected self-influence" appears intuitive, it has not been made in the past, to the best of our knowledge.

Third, the trace form clearly resembles the complexity penalties of AIC-type model selection criteria for misspecified (linear) regression models (e.g. Takeuchi, 1976; Bozdogan, 2000) and misspecified generalized linear models (e.g. Lv and Liu, 2014). This similarity is expected since complexity penalties of AIC-type model selection criteria aim at estimating the expected optimism of the in-sample risk based on a loss function equal to the negative (pseudo) log-likelihood.

Lastly, by Theorem 4.2 the expected optimism is a nonlinear function of the conditional density $f_{Y_n|X_n}$, the quantile level τ and the (weighted) covariance of the predictors $Z_S(X_n)$. This property becomes more salient in the following special case of a simple location model.

Corollary 4.1 (Location Model). *Let* $Y_{ni} = X_{ni}'\theta_{S_0} + \varepsilon_{ni}$*, with i.i.d. covariates* X_{ni} *and i.i.d. errors* $\varepsilon_{ni} \sim F_{\varepsilon}$ *and density* f_{ε}*. Suppose that the* X_{ni} *and* ε_{ni} *are mutually independent for* $i = 1,\ldots,n$*. Let the map Z be the identity map so that the* $Z(X_{ni}) = X_{ni}$*. Suppose that the conditions of Theorem 4.2 hold and that the fitted model S contains the true model* S_0*, i.e.* $S_0 \subseteq S$*. Then,*

$$\frac{1}{n}tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right) = \frac{\tau(1-\tau)}{f_{\varepsilon}\left(F_{\varepsilon}^{-1}(\tau)\right)}\frac{|S|}{n}.$$

Corollary 4.1 is a simple consequence of the characterization of misspecified quantile regression as a weighted least squares problem (4.7). The statement of Corollary 4.1 implies that the expected optimism of an over-fitted model scales linearly in the size of the (over-)fitted model. Hence, if the true model is linear, this result provides a justification

to commonly used model selection criteria where the penalty term depends linearly on the size of the fitted model. In particular, the result also suggests that a good penalty for quantile regression models should depend on the quantile level τ and the density of the error distribution in the regression model. This fact has already been recognized earlier by Portnoy (1997) in the context of model selection for smoothing spline quantile regression. We also note that Burman and Nolan's (1995) generalized AIC for quantile regression models coincides with our trace form if either the predictors are fixed and orthogonal or the model is over-fitted as in Corollary 4.1.

Corollary 4.2 (Nested Quantile Regression Location Models). *Suppose that the data generating process is a (potentially nonlinear) location model. Let S_1 and S_2 be two models such that $S_1 \subseteq S_2$. The trace form of the larger model S_2 can be written in terms of the conditional density of Y_n given the predictors $Z_{S_1}(X_n)$ of the smaller model, i.e.*

$$\frac{1}{n} tr\left(D_0^\tau(S_2)^{-1} D_1^\tau(S_2)\right) = \frac{\tau(1-\tau)}{n} tr\left(D_0(S_1,S_2)^{-1} D_1(S_2)\right),$$

where

$$\begin{aligned} D_0(S_1,S_2) &= \mathbb{E}_{X_n}\left[f_{Y_n|Z_{S_1}(X_n)}\left(Z_{S_2}(X_n)'\theta_{S_2}^\tau | Z_{S_1}(X_n)\right) Z_{S_2}(X_n) Z_{S_2}(X_n)'\right], \\ D_1(S_2) &= \mathbb{E}_{X_n}\left[Z_{S_2}(X_n) Z_{S_2}(X_n)'\right]. \end{aligned}$$

Corollary 4.2 follows immediately from Theorem 1 in Angrist et al. (2006). The corollary is especially useful when comparing nested models and when $Z_{S_1}(X_n) = X_{S_1}$. It can also be used to quantify the effect that nuisance variables or under-fitting have on the expected optimism.

4.4 Consistent estimators for expected optimism and predictive risk

4.4.1 A plug-in estimator for the expected optimism

The trace form of Theorem 4.2 lends itself to a simple plug-in estimator for the expected optimism since the two matrices $D_{n,0}^\tau(S)$ and $D_{n,1}^\tau(S)$ are well-studied in the context of the (asymptotic) covariance matrix of the quantile regression vector (e.g. Koenker, 2005). In the

case of misspecification the following estimates for $D^{\tau}_{n,0}(S)$ and $D^{\tau}_{n,1}(S)$ have been proposed

$$\widehat{D}^{\tau}_{0,h}(S) = \frac{1}{2nh}\sum_{i=1}^{n} 1\{\left|Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S)\right| \leq h\} Z_S(X_{ni})Z_S(X_{ni})', \tag{4.14}$$

$$\widehat{D}^{\tau}_{n,1}(s) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau}\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S)\big) Z_S(X_{ni})Z_S(X_{ni})', \tag{4.15}$$

where h is a bandwidth parameter and $\varphi_{\tau}(u) = \tau - 1\{u < 0\}$ (e.g. Angrist et al., 2006; Belloni et al., 2017). We therefore propose the following plug-in estimate for the expected optimism $b^{\tau}_n(S)$,

$$\hat{b}^{\tau}_{n,h}(S) = \frac{1}{n} tr\left(\widehat{D}^{\tau^{-1}}_{0,h}(S)\widehat{D}^{\tau}_{n,1}(S)\right). \tag{4.16}$$

Since our regularity conditions are slightly more general than those in Belloni et al. (2017), the following consistency theorem does not follow from their Lemma 30. In particular, our Assumption (A5) on the growth rate of the number of predictors is less stringent than theirs. We shall see that this relaxation is important in the context of predictive risk estimation in Section 4.4.2.

Proposition 4.3 (Uniform Consistency of the Estimated Trace Form). *Suppose that Assumptions (A1) – (A6) from Section 4.2.4 hold, let $h > 0$ be the bandwidth parameter, and $r_n = \frac{1}{\lambda_n}\left(\frac{m\log|M|\log\log n}{n}\right)^{1/2}$. Then,*

$$\sup_{S\in M}\left|n\cdot\hat{b}^{\tau}_{n,h}(S) - tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right| = O_p\left(\frac{m\,h^{\alpha}}{\lambda_n^2} + \frac{m\,r_n}{h\lambda_n} + \frac{m\,r_n^{\alpha}}{\lambda_n^2}\right).$$

The first and second terms on the right hand side capture the variance and bias of the estimator with bandwidth h. They are standard in nonparametric smoothing. The third terms controls the bias induced by $\left\{\left(Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S)\right)\right\}_{i=1}^{n}$ at model S which serve as proxies for $\left\{\left(Y_n - Z_S(X_{ni})'\theta^{\tau}_{n,S}\right)\right\}_{i=1}^{n}$.

Specializing to the common case of a continuous conditional density $f_{Y_n|X_n}$, i.e. $\alpha = 1$, we observe the following: The optimal, mean-variance-balancing, bandwidth $h^* = (c_1/c_0)^{1/2}(\lambda_n r_n)^{1/2}$ with constants $c_0, c_1 > 0$ given in eq. (4.33) and (4.34), respectively. In principle these constants can be estimated from the data. But in practice, we find that the specific choice of the bandwidth has no significant effect. With bandwidth h^* the estimate $\hat{b}^{\tau}_{n,h}(S)$ is consistent at rate $O_p\left(m r_n^{1/2}\lambda_n^{-3/2} + m r_n \lambda_n^{-2}\right) = O_p\left(m r_n^{1/2}\lambda_n^{-3/2}\right)$. That is, $\hat{b}^{\tau}_{n,h}(S)$ is consistent at a rate that is the same as if the true errors $\left\{\left(Y_{ni} - \widehat{Q}_{Y_n}(\tau|X_{ni},S)\right)\right\}_{i=1}^{n}$ at model S were known.

Combining Theorem 4.2 and Proposition 4.3 we obtain the following consistency result.

Theorem 4.3 (Uniform Consistency of the Estimated Expected Optimism). *Let* $r_n = \frac{1}{\lambda_n}\left(\frac{m \log |M| \log\log n}{n}\right)^{1/2}$. *Under the conditions of Proposition 4.3,*

$$\sup_{S \in M} \left| \frac{\hat{b}^{\tau}_{n,h}(S)}{b^{\tau}_{n}(S)} - 1 \right| = O_p\left(n\lambda_n^{3/2} r_n^{5/2} + \frac{m\, h^{\alpha}}{\lambda_n^2} + \frac{m\, r_n}{h\lambda_n} + \frac{m\, r_n^{\alpha}}{\lambda_n^2} \right).$$

Since $\hat{b}^{\tau}_{n,h}(S)$ is the plug-in estimator for the trace form approximation, it is a biased estimate of the actual expected optimism $b^{\tau}_{n}(S)$. This deterministic bias is captured in the first term, the remaining three terms are already familiar from Proposition 4.3. Specializing once again to the case of a continuous conditional density we have under the optimal bandwidth h^* a rate of $O_p\big(n\lambda_n^{3/2} r_n^{5/2} + m\, r_n^{1/2}\lambda_n^{-3/2}\big) = O_p\big(n\lambda_n^{3/2} r_n^{5/2}\big)$. Thus, the deterministic error of using the trace form $tr\big(D^{\tau}_{n,0}(S)^{-1} D^{\tau}_{n,1}(S)\big)$ to approximate the expected optimism $b^{\tau}_{n}(S)$ dominates the stochastic estimation error. In other words, as point estimate $\hat{b}^{\tau}_{n}(S)$ is as good in estimating the expected optimism $b^{\tau}_{n}(S)$ as the unattainable trace form $tr\big(D^{\tau}_{n,0}(S)^{-1} D^{\tau}_{n,1}(S)\big)$.

4.4.2 A de-biased estimator of the predictive risk

As outlined in Section 4.2.2 given the consistent estimate of the expected optimism (4.16) we propose the following de-biased estimate of the predictive risk,

$$\widehat{PR}^{\tau}_{n,h}(S) = \frac{1}{n}\sum_{i=1}^{n} \Big(\rho_{\tau}\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau | X_{ni}, S)\big) - \rho_{\tau}(Y_{ni}) \Big) + \hat{b}^{\tau}_{n,h}(S). \tag{4.17}$$

We call this estimate "de-biased" because the in-sample risk $\frac{1}{n}\sum_{i=1}^{n} \big(\rho_{\tau}\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau | X_{ni}, S) - \rho_{\tau}(Y_{ni})\big)$ is itself already a consistent estimate for $\mathrm{PR}^{\tau}_{n,h}(S)$ in the sense that for any $S \in M$ with fixed model size $|S|$,

$$\left| \mathrm{PR}^{\tau}_{n,h}(S) - \frac{1}{n}\sum_{i=1}^{n} \Big(\rho_{\tau}\big(Y_{ni} - \widehat{Q}_{Y_n}(\tau | X_{ni}, S)\big) - \rho_{\tau}(Y_{ni}) \Big) \right| = O_p\left(n^{-1/2}\right). \tag{4.18}$$

We strengthen this fact in several ways: First, we show that under appropriate conditions our proposed estimator $\widehat{PR}^{\tau}_{n,h}(S)$ is consistent uniformly over all $S \in M$ and for models whose size $|S|$ grows with the sample size n. Second, we will see that for large models with size $|S| \gtrsim n^{1/2}$ the in-sample risk is no longer $n^{1/2}$-consistent for the predictive risk and that under certain conditions de-biasing the in-sample risk with $\hat{b}^{\tau}_{n,h}(S)$ restores the

$n^{1/2}$-consistency. We deduce these claims from the following general result.

Theorem 4.4 (Uniform Consistency of the De-biased Predictive Risk Estimate). *Suppose that Assumptions (A1) – (A6) from Section 4.2.4 hold. In addition, assume that $f_{Y_n|X_n}$ is uniformly bounded away from 0 for all n and that $\limsup_{n\to\infty} \mathbb{E}_{X_n}\left[Q^2_{Y_n}(\tau|X_n)\right] < \infty$. Let $h > 0$ be a bandwidth and $r_n = \frac{1}{\lambda_n}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}$. Then,*

$$\sup_{S\in M}\left|\widehat{PR}^{\tau}_{n,h}(S) - \mathrm{PR}^{\tau}_n(S)\right| = O_p\left(\left(\frac{\log|M|}{n}\right)^{1/2} + \frac{r_n}{n^{1/2}} + \lambda_n^{3/2} r_n^{5/2} + \frac{m\,h^{\alpha}}{\lambda_n^2 n} + \frac{m\,r_n}{h\lambda_n n} + \frac{m\,r_n^{\alpha}}{\lambda_n^2 n}\right),$$

The last four terms on the right hand side are familiar from the uniform consistency result of the trace form estimate for the expected optimism (i.e. Theorem 4.3), while the first two terms are related to the in-sample risk. Clearly, if $m = o(\lambda_n^{-2} n \log|M|\,\log\log n)$ and bandwidth h satisfies

$$\frac{1}{\lambda_n}\frac{m}{n}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2} \lesssim h \lesssim \frac{1}{\lambda_n^{2/\alpha}}\left(\frac{n}{m}\right)^{1/\alpha}, \tag{4.19}$$

then $\widehat{PR}^{\tau}_{n,h}(S)$ is consistent for $\mathrm{PR}^{\tau}_n(S)$ uniformly for all $S \in M$. However, we can learn more by considering special cases. To simplify this discussion, we consider the case in which the conditional density $f_{Y_n|X_n}$ is continuous and the bandwidth is chosen to balance the nonparametric estimation bias and variance (see discussion in Section 4.4.1). Then, Theorem 4.4 implies the following.

Corollary 4.3. *Suppose that the conditions of Theorem 4.4 hold, that the conditional density $f_{Y_n|X_n}$ is continuous, $\lambda_n^2 m = o\left(n\log|M|\,\log\log n\right)$ and $n^{1/4}h \asymp (m\log|M|\log\log n)^{1/4}$. Then,*

$$\sup_{S\in M}\left|\widehat{PR}^{\tau}_{n,h}(S) - \mathrm{PR}^{\tau}_n(S)\right| = O_p\left(\left(\frac{\log|M|}{n}\right)^{1/2} + \frac{1}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right).$$

These rates have an intuitive explanation: The first term $O\big(n^{-1/2}(\log|M|)^{1/2}\big)$ is related to the stochastic variability of the in-sample risk, the second term $O(\lambda_n^{-2}n^{-5/4}(m\log|M|\,\log\log n)^{5/4})$ is known from Theorem 4.2 to be the deterministic error of using the trace form $tr(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S))$ to approximate the expected optimism $b^{\tau}_n(S)$. Thus, unlike one might have suspected, it is not the nonparametric estimate of the expected optimism but the deterministic approximation of the expected optimism and the stochastic variability of the in-sample risk which limit the accuracy of our predictive risk estimate. It is easy to

verify that under the stated assumptions $\widehat{PR}^{\tau}_{n,h}(S)$ is consistent for $\mathrm{PR}^{\tau}_{n}(S)$ uniformly over all $S \in M$.

It is instructive to consider the implication of Corollary 4.3 under different growth regimes of the number of predictor variables. To this end, recall that the estimated trace form, $\hat{b}^{\tau}_{n,h}(S)$, is of order $O(\lambda_n^{-1} n^{-1} |S|)$. Hence, if $n^{1/2} \lesssim |S| \lesssim n$ the estimated trace form, $\hat{b}^{\tau}_{n,h}(S)$, dominates (rate-wise) the stochastic error and also the deterministic error (provided that we sharpen condition on m and n to $m = o(n\,\lambda_n^4 (\log |M|\, \log\log n)^{-5})$). Thus, in this regime the in-sample risk alone is not $n^{1/2}$-consistent for the predictive risk; de-biasing the in-sample risk is necessary to retain $n^{1/2}$-consistency.

However, if $|S| \lesssim n^{1/2}$ the stochastic error of he in-sample risk dominates (rate-wise) the estimate of the trace form. Thus, from the perspective of first order asymptotics the correction provided by the $\hat{b}^{\tau}_{n,h}(S)$ is not necessary in this regime. However, in Section 4.5 we report numerical evidence showing that even in this regime the de-biasing effect of $\hat{b}^{\tau}_{n,h}(S)$ is practically relevant.

As an aside, this discussion provides another explanation for the well-known fact that AIC-type model selection criteria are not model selection consistent: AIC-type penalties (based on estimates of the expected optimism) are too small to effectively discriminate between models of size $|S| \lesssim n^{1/2}$ since the stochastic variability of the in-sample risk is relatively large. For correctly specified (linear least squares regression) models with a fixed number of parameters this has already been recognized by e.g. Shao (1997) and Yang (2005).

4.5 Empirical Evidence

4.5.1 Set-up of the simulation study

We conduct Monte Carlo experiments to evaluate empirically the trace form approximation of the expected optimism and to corroborate the theoretical results from the previous two sections. We also compare the empirical performance of the trace form approximation to the commonly used cross-validated estimate of the predictive bias. Our Monte Carlo study uses four designs as the data generating processes (DGP), but only the results from DGP1 are given in the paper. The results from the other DGPs are qualitatively similar and details are given in the Supplementary Materials.

(DGP1) *Independent Gaussian design:* $y_i = x_{i1} + x_{i2} + x_{i3} + x_{i4} + \varepsilon_i$, *with* $x_i \sim_{iid} N(0, I_p)$ *independent of the errors* $\varepsilon_i \sim_{i.i.d.} N(0,4)$. *We use this process to illustrate the elementary properties of the predictive risk and the expected optimism from Corollaries 4.1 and 4.2. The joint Gaussianity of predictors and errors allows us to compute the exact value of the*

trace form with which we can assess the accuracy of our estimates. The variance of the error distribution is chosen such that signal-to-noise-ratio equals one.

(DGP2) *Correlated Gaussian design:* $y_i = x_{i1} + x_{i2} + x_{i3} + x_{i4} + \varepsilon_i$, *with* $\varepsilon_i \sim_{i.i.d.} N(0, 12.384)$ *independent of* $x_i \sim_{iid} N(0, \Sigma)$ *and* $\Sigma_{ij} = 0.8^{|i-j|}$ *for all* $i, j = 1, \ldots, p$. *The variance of the error distribution is chosen such that the signal-to-noise ratio equals one.*

(DGP3) *Heteroscedastic noise:* $y_i = x_{i1} + x_{i2} + x_{i3} + (1 + 1.5x_{i4})\varepsilon_i$, *where* $x_{ij} \sim_{i.i.d.} U([0,2])$ *for* $j = 1, \ldots, 4$ *independent of the errors* $\varepsilon_i \sim_{iid} N(0,1)$. *In this DGP the covariate* x_4 *is active for the conditional quantile functions except at the median.*

(DGP4) *Single interaction term with heavy-tailed noise:* $y_i = x_{i1} + x_{i2} + x_{i3} + 4x_{i3}x_{i4} + \varepsilon_i$, *where* ε_i *follow the t-distribution with 2 degrees of freedom independent of the predictors* $x_i \sim_{iid} N(0, I_p)$. *In this DGP all quantiles are non-linear functions of the covariates.*

We set the dimension of the space of covariates $\mathscr{X}$ equal to 50, and let Z be the identity map, so that the predictors are simply the covariates $X_1, \ldots, X_{50}$. We consider a collection of 176 candidate models with model seizes ranging between 0 to 50. This implies that we the size of the largest model under consideration is $m = 50$. We explain the choice of those candidate models in Section 4.5.2. Throughout the numerical experiments we keep the sample size fixed at $n = 500$. All reported estimates are averages over 10,000 independent realizations of the corresponding DGPs. To estimate the matrix $D_0(S)$ at quantile τ we use Powell's (1986) nonparametric estimator with uniform kernel function and bandwidth

$$c_{n,S} = \kappa_{n,S}\left(\Phi^{-1}(\tau + h_n) - \Phi^{-1}(\tau - h_n)\right),$$

where Φ denotes c.d.f. of the standard normal distribution, $\kappa_{n,S}$ is the minimum of the standard error and the inter-quartile-range of the estimated quantile regression residuals of model S, and

$$h_n = \frac{1}{n^{1/5}}\left(\frac{4.5\phi\left(\Phi^{-1}(\tau)\right)^4}{(2\Phi^{-1}(\tau)^2 + 1)^2}\right)^{1/5},$$

where ϕ denotes the p.d.f. of the standard normal distribution. Thus, $c_{n,S}$ satisfies the conditions of Theorems 4.3 and 4.4 which guarantee (uniform) consistency of the estimates of the expected optimism and the predictive risk; see Koenker (2005) for a detailed discussion of this choice of bandwidth.

Recall Definitions 4.1 and 4.2 that the predictive risk and the expected optimism require the evaluation of a double expectation. Since the quantile regression vector is only implicitly

defined, this double expectation cannot be evaluated analytically. Instead, we use Monte Carlo estimates based on 50,000 samples to obtain values for the predictive risk and the expected optimism.

4.5.2 Estimation of the expected optimism

In Theorem 4.3 we establish uniform consistency of the estimated trace form for the expected optimism. In Figure 4.1 under DPG1, we plot the bias of 176 models (subsets of the 50 predictors) against their model sizes. We only consider 176 models because it is computationally expensive to evaluate the predictive risk and the expected optimism on all possible subsets of the 50 predictors. However, the special structure of the DGP together with Corollary 4.2 guarantee that this collection constitutes a representative subset of all possible models: The true DGP contains only four relevant predictors 1, 2, 3, and 4; those predictors are independent and identically distributed and contribute equally to the model (i.e. have the same regression coefficients). We can therefore stratify the collection of all possible subsets of the 50 predictors according to how many relevant predictors are included in a specific subset. This results in five collections of nested models indexed by 0 (relevant predictors), 1 (relevant predictor), ..., 4 (relevant predictors). By Corollary 4.2 the expected optimism of all nested models with j relevant predictors lie (approximately) on a ray emanating from the in-sample bias of the smallest model with j relevant predictors. Moreover, the slope of the ray is given by $\frac{\tau(1-\tau)}{500\phi_j}$, where ϕ_j denotes the value of the density of a centered normal random variable with variance j^2+1 evaluated at 0. The 176 models comprise the model that contains only the intercept and 35 models of each of the five stratified collections.

In Figure 4.1 the top gray line corresponds to the theoretical values of the trace form of models that have four relevant predictors and additional, irrelevant, predictors. The second line from the top corresponds to the theoretical values of the trace form of models that contain three relevant predictors and additional, irrelevant, predictors, and so forth. The last line (fifth from above) corresponds to models that do not contain any relevant predictors.

We observe that the estimates of the trace form (in red) lie on (or are very close) to theoretical values of the trace form uniformly for all 176 models. This confirms the fast uniform convergence rates obtained in Theorem 4.3. Note that the plot shows only 50 red dots and not as one might expect 176 dots. This is due to the fact that for DGP1 the value of the estimated trace form does not depend on the specific subset of predictors (i.e. S) but only on the size of the model (i.e. $|S|$), e.g. the two models with predictors $\{1,2,5\}$ and $\{3,4,10\}$ have the same trace form which is fully determined by the fact that they contain

two relevant and one irrelevant predictors. The expected optimism (in blue) does not follow the dashed gray lines of the theoretical values of the trace form as closely as the estimates do. This reflects the fact that the trace form is only an approximation to the expected optimism (see Theorem 4.2). The difference between the values of the trace form and the expected optimism appears to be negligible for models of size up to $20 \approx \sqrt{n}$ (recall that $n = 500$).

The vertical red lines indicate the standard deviations of the estimated trace forms. The standard deviation increases with the model size and, holding the number of nuisance predictor variables fixed, decreases with the number of relevant predictor variables that are included in the model. The latter effect is rather weak and can be best observed in the plot for the 80% quantile.

4.5.3 Comparison with cross-validated expected optimism

Cross-validation is a commonly-used method for estimating the predictive risk and the expected optimism. In this subsection we compare the trace form estimate with a 10-fold cross-validation estimate of the expected optimism.

Figure 4.2 shows the results of 10-fold cross-validation and the trace form for DGP1 at the median. We consider four representative models: Model I is the correct model (with predictors 1 to 4), Model II is an over-fitted model (with predictors 1 to 10), Model III is an under-fitted model (with predictors 1 and 2) and Model IV is the model that comprises the relevant predictors 1 and 2 and the irrelevant predictors 5 to 15. The vertical red line indicates the expected optimism. The white histograms show the empirical distribution of 10,000 cross-validation estimates of the expected optimism and the dark gray histograms show the empirical distribution of 10,000 trace form estimates of the expected optimism.

Both histograms are centered around the expected optimism; however, the estimate of the trace form concentrates significantly more around the target. As mentioned in Section 4.2.2 the reason for this is that the cross-validation estimate is based on a smaller sample size both for model estimation and for risk estimation.

4.6 Conclusion

In the present paper, we have derived two asymptotic approximations of the expected optimism, or the bias of the in-sample risk when used as an estimate of the predictive risk, and have proposed consistent estimates of the expected optimism and the predictive risk of potentially misspecified quantile regression models. The asymptotic approximations based on two explicit forms help us understand how the expected optimism depends on several

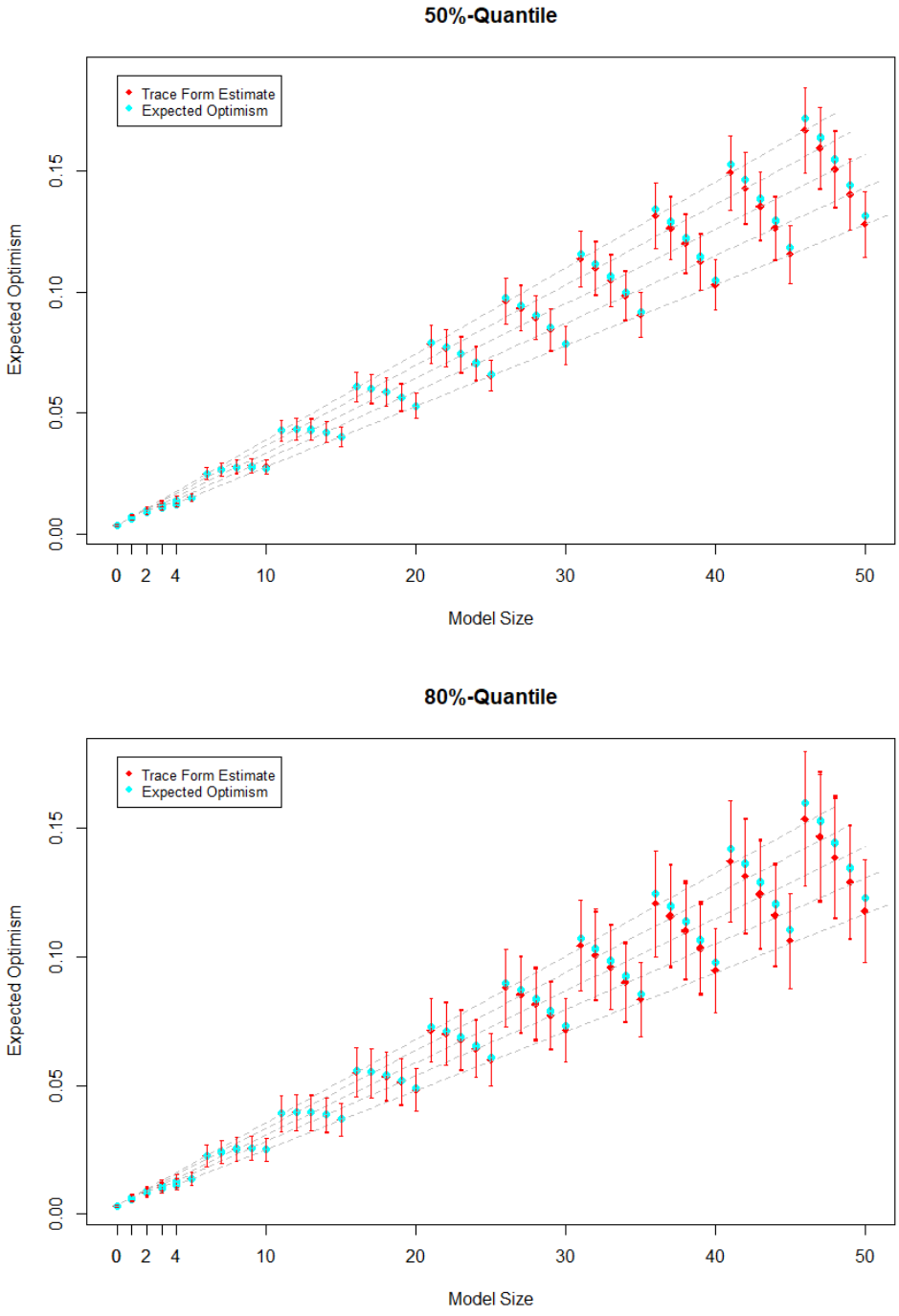

Figure 4.1: DGP1 trace form versus model size. Red: estimates of the trace form and standard errors. Blue: expected optimism. Dashed gray lines: exact evaluation of the trace form. Top: DGP1 with $\tau = 0.5$. Bottom: DGP1 with $\tau = 0.8$.

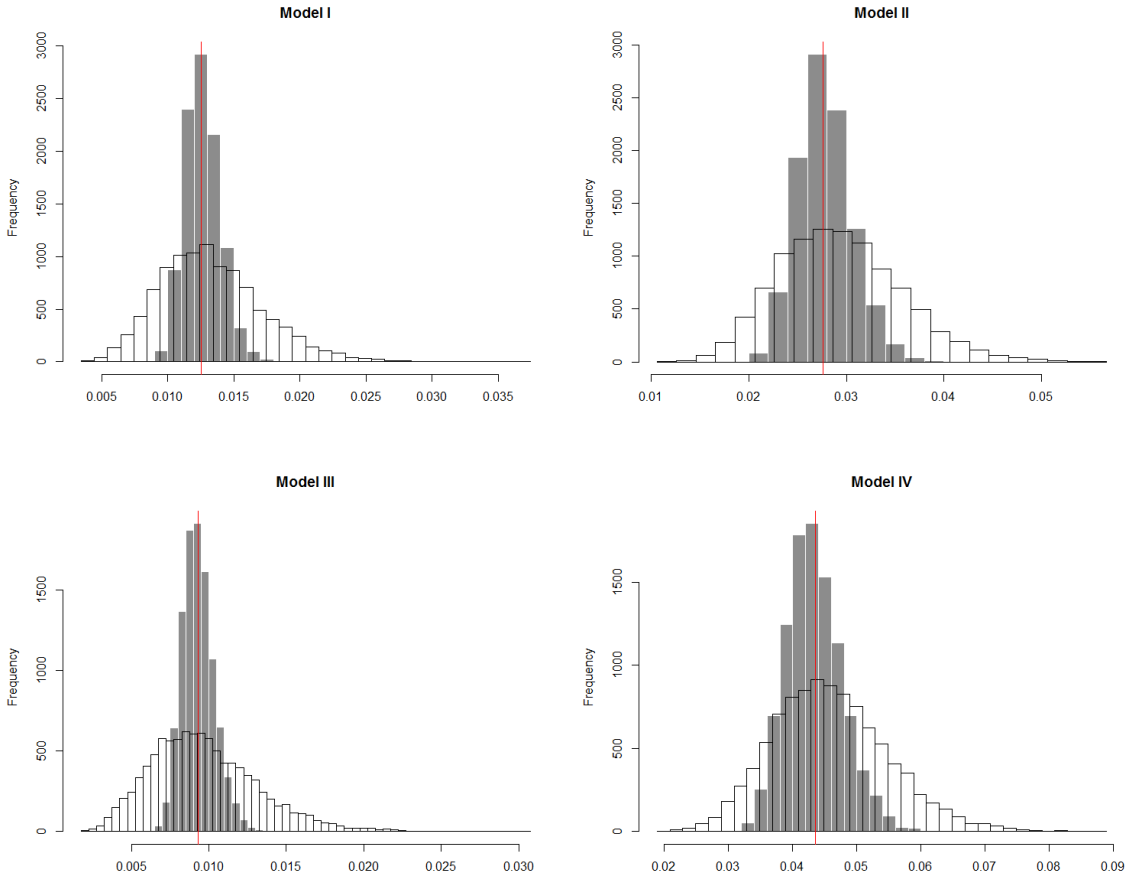

Figure 4.2: Expected optimism and trace form estimate (DGP1). Histograms of the 10-fold CV estimate of the expected optimism and the trace form estimate for DGP1 and $\tau = 0.5$. Red line: expected optimism. White histogram: 10-fold CV. Gray histogram: trace form estimate. Model I: correct model (with predictors 1 to 4), Model II: an over-fitted model (with predictors 1 to 10), Model III: an under-fitted model (with predictors 1 to 2) and Model IV that comprises the relevant predictors 1 and 2 and the irrelevant predictors 5 to 15.

factors, including the quantile level, the model misspecification bias, the model size, and sampling variability. In some simpler cases, the expected optimism is asymptotically linear in the model size, but for under-fitted or misspecified models in general, the relationship is far more complicated. The results show that commonly used AIC-type model selection criteria for quantile regression are not really good proxies of the predictive risk. The consistency results indicate that de-biasing the in-sample risk with an estimate of the expected optimism is necessary when considering models whose dimension grow with at least $n^{1/2}$. Empirical evidence suggests that even in the case of models with fixed dimension of the simple in-sample risk can be significantly improved via de-biasing.

The asymptotic approximations derived in the present paper are uniform in a class of candidate models, but those models are not data-dependent. An interesting question that relates more to model selection criteria is how well the bias and thus the predictive risk estimation hold up for data-dependent models. Clearly, additional research is needed to address this question.

4.7 Proofs

4.7.1 Additional notation and lemmata

We denote the check loss of the τth quantile by $\rho_\tau(u) = u(\tau - 1\{u < 0\})$ and the corresponding score function by $\varphi_\tau(u) = \tau - 1\{u < 0\}$. We define $Z_{ni,S} = Z_S(X_{ni})$, $Z^0_{ni,S} = Z_S(X^0_{ni})$, $\hat{\delta}^\tau_{n,S} = \hat{\theta}^\tau_{n,S} - \theta^\tau_{n,S}$, $e^\tau_{ni,S} = Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S}$, $\hat{e}^\tau_{ni,S} = Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}$, and $\hat{e}^{0\tau}_{n,S} = Y^0_n - Z^{0'}_{n,S}\hat{\theta}^\tau_{n,S}$. We use $C, c, c_0, c_1, \ldots$ to denote absolute constant that may change from line to line. Let $r_n = \frac{c_3}{\lambda_n}\left(\frac{m \log |M| \log\log n}{n}\right)^{1/2}$, where $c_3 > 0$ is the absolute constant from Lemma 4.6. Throughout we assume that $|M| \geq 2$ and $\log\log n > 1$, i.e. $n > 15$.

The proofs of Theorems 4.1 – 4.4 make use of the following lemmata. Their proofs can be found in Section 2.3.3.

Lemma 4.1 (Panchenko (2003)). *Let $X_1, \ldots, X_n, X^0_1, \ldots, X^0_n$ be i.i.d. random vectors on a measurable space $\mathscr{X}$ and let $\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R}\}$ be a countable class of measurable functions. Define the mixed uniform variance as*

$$V = \mathbb{E}_{X^0}\left[\sup_{f \in \mathscr{F}} \sum_{i=1}^n \left(f(X_i) - f(X^0_i)\right)^2\right].$$

Then for any $\alpha > 0$ and $t > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}}\sum_{i=1}^{n} f(X_i) \geq \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^{n} f(X_i)\right] + 2\sqrt{Vt}\right) \leq 2^{\alpha+1}\exp\left\{1 - \frac{\alpha}{\alpha+1}t\right\}$$

and

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}}\sum_{i=1}^{n} f(X_i) \leq \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^{n} f(X_i)\right] - 2\sqrt{Vt}\right) \leq 2^{\alpha+1}\exp\left\{1 - \frac{\alpha}{\alpha+1}t\right\}.$$

The next two technical lemmata are needed in the consistency proofs involving the matrices $D^{\tau}_{n,0}(S)$ and $D^{\tau}_{n,1}(S)$.

Lemma 4.2. *Let $\{(W_n, X_n), (W_{ni}, X_{ni}), i = 1, \ldots, n\}$ be a triangular array of row-wise i.i.d. random vectors in $\mathbb{R}\times\mathbb{R}^d$. Suppose that $W_n|X_n$ has a continuous distribution function and density bounded by $F_+ > 0$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function with envelope G such that $\mathbb{E}\left[G^{4+2\delta}\right] < \infty$ for some $\delta > 0$. Then, for all $r_n > 0$ there exists an absolute constant $c_0 > 0$ such that*

$$\begin{aligned}
&\mathbb{E}_{W,X}\left[\sup_{\|\theta\|_2=r_n}\frac{1}{n}\sum_{i=1}^{n}\left(1\left\{0 < W_{ni} \leq X'_{ni}\theta\right\}g(X_{ni}) - \mathbb{E}_{W,X}\left[1\left\{0 < W_{ni} \leq X'_{ni}\theta\right\}g(X_{ni})\right]\right)\right] \\
&\leq c_0\left(\frac{d}{n}\right)^{3/4}\mathbb{E}\left[G^4(X_{n1})\right]^{1/4}\mathbb{E}\left[G^{4+2\delta}(X_{n1})\right]^{1/(4+2\delta)}\mathbb{E}\left[G^{2+\delta}(X_{n1})\right]^{1/(2+\delta)} \\
&\quad + c_0 r_n^{1/2}\left(\frac{d}{n}\right)^{1/2}F_+^{1/2}\sup_{\|u\|_2=1}\mathbb{E}\left[(X'_{n1}u)G^2(X_{n1})\right]^{1/2}\mathbb{E}\left[G^{2+\delta}(X_{n1})\right]^{1/(2+\delta)}.
\end{aligned}$$

Lemma 4.3. *Let $\{X_{ni}, i = 1, \ldots, n\}$ be triangular array of row-wise i.i.d. $\mathbb{R}^d$-valued random vectors. Let $\max_{\|u\|_2=1}\mathbb{E}\left[(X'_{n1}u)^2\right] \leq \lambda_+$. Suppose that there exists $p \geq 1$ and an absolute constant $\mu_{4p} < \infty$, independent of d, such that $\max_{1\leq k\leq d}\mathbb{E}\left[(X^{(k)}_{n1})^{4p}\right] \leq \mu_{4p}$. Then,*

$$\sup_{\|v\|_2=1}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|X'_{ni}v|^{2p}\right]\right)^{1/(2p)} \leq 2^{2+1/(2p)}K^{1/2}p^{1/2}\mu_{4p}^{1/(2p)} + \lambda_+^{1/2},$$

where $K = 1/(e - \sqrt{e})$.

The next lemma provides an almost sure bound on the centered quantile regression score.

Lemma 4.4. *Suppose that Assumptions (A1) – (A6) hold. There exists an absolute constant*

$c_0 > 0$ *such that for any any* $r_n > 0$,

$$\sup_{S\in M}\sup_{\|\delta\|_2=r_n}\Big\|\frac{1}{n}\sum_{i=1}^{n} -\varphi_\tau(e^\tau_{ni,S}-Z'_{ni,S}\delta)Z_{ni,S}+\varphi_\tau(e^\tau_{ni,S})Z_{ni,S} \\ +\mathbb{E}_{\mathscr{D}_n}\left[\varphi_\tau(e^\tau_{ni,S}-Z'_{ni,S}\delta)Z_{ni,S}-\varphi_\tau(e^\tau_{ni,S})Z_{ni,S}\right]\Big\|_2 \\ \leq c_0 r_n^{1/2}\left(\frac{m\log|M|}{n}\right)^{1/2}+c_0\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4} \quad a.s.$$

The following lemma provides an almost sure bound on the (un-centered) quantile regression score.

Lemma 4.5. *Suppose that Assumptions (A1) – (A6) hold. There exists an absolute constant* $c_1 > 0$ *such that*

$$\sup_{S\in M}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z_{ni,S}\right\|_2 \leq c_1\left(\frac{m+\log|M|+\log\log n}{n}\right)^{1/2} \quad a.s.$$

The final lemma is a strengthened version of Theorems 2.1 and 2.2 in He and Shao (2000) in the special case of quantile regressions.

Lemma 4.6. *Suppose that Assumptions (A1) – (A6) hold. Then, there exists a universal constants* $c_2, c_3 > 0$ *such that*

$$\hat{\theta}^\tau_{n,S} = \theta^\tau_{n,S} + \left(\mathrm{E}\left[f_{Y_n|X_n}\left(Z'_{n1,S}\theta^\tau_{n,S}|X_{n1}\right)Z_{n1,S}Z'_{n1,S}\right]\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau\left(e^\tau_{n,S}\right)Z_{ni,S} + r^\tau_{n,S},$$

and

$$\sup_{S\in M}\left\|r^\tau_{n,S}\right\|_2 \leq \frac{c_2}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/4}\left(\frac{m\log|M|+\log\log n}{n}\right)^{1/2} \quad a.s.,$$

and

$$\sup_{S\in M}\|\hat{\theta}^\tau_{n,S}-\theta^\tau_{n,S}\|_2 \leq \frac{c_3}{\lambda_n}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2} \quad a.s.$$

4.7.2 Proof of Theorem 4.1

Proof of Theorem 4.1. **Step 1:** By Knight's identity,

$$\rho_\tau(u-v)-\rho_\tau(u) = -v\varphi_\tau(u)+\int_0^v\left(1\{u\leq s\}-1\{u\leq 0\}\right)ds,$$

for arbitrary $S \in M$, we can write the optimism as

$$
\begin{aligned}
&\mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,s}) - \rho(Y_{ni}) - \rho_\tau(Y_n^0 - Z^{0'}_{n,S}\hat{\theta}^\tau_{n,S}) + \rho_\tau(Y_n^0)\right] \\
&= \mathbb{E}_{\mathscr{D}_n}\left[-\frac{1}{n}\sum_{i=1}^n Z'_{ni,S}\hat{\delta}^\tau_{n,S}\varphi_\tau(e^\tau_{ni,S}) + \frac{1}{n}\sum_{i=1}^n \int_0^{Z'_{ni,S}\hat{\delta}^\tau_{n,S}} \left(1\{e^\tau_{ni,S} \le t\} - 1\{e^\tau_{ni,S} \le 0\}\right) dt\right] \\
&\quad - \mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[-Z^{0'}_{n,S}\hat{\delta}^\tau_{n,S}\varphi_\tau(e^{0\tau}_{n,S}) + \int_0^{Z^{0'}_{n,S}\hat{\delta}^\tau_{n,S}} \left(1\{\hat{e}^{0\tau}_{n,S} \le t\} - 1\{\hat{e}^{0\tau}_{n,S} \le 0\}\right) dt\right] \\
&= \left(-\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n Z'_{ni,S}\hat{\delta}^\tau_{n,S}\varphi_\tau(e^\tau_{ni,S})\right] + \mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[Z^{0'}_{n,S}\hat{\delta}^\tau_{n,S}\varphi_\tau(\hat{e}^{0\tau}_{n,S})\right]\right) \\
&\quad + \left(\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \int_0^{Z'_{ni,S}\hat{\delta}^\tau_{n,S}} \left(1\{e^\tau_{ni,S} \le t\} - 1\{e^\tau_{ni,S} \le 0\}\right) dt\right]\right. \\
&\quad \left. - \mathbb{E}_{\mathscr{D}_n,(Y^0,X^0)}\left[\int_0^{Z^{0'}_{n,S}\hat{\delta}^\tau_{n,S}} \left(1\{\hat{e}^{0\tau}_{n,S} \le t\} - 1\{\hat{e}^{0\tau}_{n,S} \le 0\}\right) dt\right]\right) \\
&= A_n(S) + B_n(S). \qquad (4.20)
\end{aligned}
$$

Step 2: Uniform upper bound on $B_n(S)$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Then,

$$
\begin{aligned}
\sup_{S\in M} B_n(S) &\le \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{1}{n}\sum_{i=1}^n \int_0^{Z'_{ni,S}\delta_S} \left(1\{e^\tau_{ni,S} \le t\} - 1\{e^\tau_{ni,S} \le 0\}\right) dt \right. \\
&\qquad \left. - \int_0^{Z^{0'}_{n,s}\delta_S} \left(1\{e^{0\tau}_{n,S} \le t\} - 1\{e^{0\tau}_{n,S} \le 0\}\right) dt\right] \\
&\le \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(Z'_{ni,S}\delta_S\right) 1\{0 \le e^\tau_{ni,S} \le Z'_{ni,S}\delta_S\}\right] \\
&\quad + \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{1}{n}\sum_{i=1}^n \varepsilon_i e^\tau_{ni,S} 1\{0 \le e^\tau_{ni,S} \le Z'_{ni,S}\delta_S\}\right] \\
&\le \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2 \le r_n} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i Z_{ni,S} 1\{0 \le e^\tau_{ni,S} \le Z'_{ni,S}\delta_S\}\right\|_2\right] r_n \qquad (4.21) \\
&\quad + \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{1}{n}\sum_{i=1}^n \varepsilon_i e^\tau_{ni,S} 1\{0 \le e^\tau_{ni,S} \le Z'_{ni,S}\delta_S\}\right]. \qquad (4.22)
\end{aligned}
$$

Bound on eq. (4.21). Note that after de-symmetrizing eq. (4.21) is upper bounded by

the centered quantile regression score. Thus, by the almost sure upper bound of Lemma 4.5,

$$\sup_{S\in M}\mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2\le r_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Z_{ni,S}1\{0\le e^{\tau}_{ni,S}\le Z'_{ni,S}\delta_S\}\right\|_2\right] r_n$$
$$= O\left(\frac{1}{\lambda_n^{3/2}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right).$$

Bound on eq. (4.22). Similarly to the bound on eq (4.21) we conclude that

$$\sup_{s\in M}\mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\sup_{\|\delta_S\|_2\le r_n}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i e^{\tau}_{ni,S}1\{0\le e^{\tau}_{ni,S}\le Z'_{ni,S}\delta_S\}\right]$$
$$= O\left(\frac{1}{\lambda_n^{3/2}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right).$$

Step 3: Uniform expansion of $A_n(S)$.

$$\begin{aligned}
&\sup_{S\in M}A_n(S)\\
&=\sup_{S\in M}\left|-\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}Z'_{ni,S}\hat{\delta}_{n,S}\varphi_\tau(e^{\tau}_{ni,S})\right]+\mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\frac{1}{n}\sum_{i=1}^{n}Z^{0\prime}_{ni}\hat{\delta}_{n,S}\varphi_\tau(\tilde{e}^{\tau}_{ni,S})\right]\right.\\
&\qquad\left.+tr\left(\frac{1}{n}\sum_{i=1}^{n}Cov\left(Z_{ni,S}\varphi_\tau(e^{\tau}_{ni,S}),\,\hat{\delta}_{n,S}\right)\right)\right|\\
&=\sup_{S\in M}\left|-\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}Z'_{ni,S}\hat{\delta}_{n,S}\varphi_\tau(e^{\tau}_{ni,S})\right]+tr\left(\frac{1}{n}\sum_{i=1}^{n}Cov\left(Z_{ni,S}\varphi_\tau(e^{\tau}_{ni,S}),\,\hat{\delta}_{n,S}\right)\right)\right|\\
&=0.
\end{aligned}\tag{4.23}$$

Step 4: Conclusion. The claim follows by combining the upper bounds in equations (4.20)– (4.23). □

4.7.3 Proof of Theorem 4.2

Proof of Theorem 4.2. We only need to approximate the covariance form of Theorem 4.1.

$$\sup_{S\in M}\left|tr\left(\frac{1}{n}\sum_{i=1}^{n}Cov\left(Z_{ni,S}\varphi_\tau(e^{\tau}_{ni,S}),\,\hat{\delta}_{n,S}\right)\right)-\frac{1}{n}tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right|$$
$$=\sup_{S\in M}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}Z'_{ni,S}\hat{\delta}_{n,S}\varphi_\tau(e^{\tau}_{ni,S})\right]-\frac{1}{n}tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right|$$

$$
\begin{aligned}
&= \sup_{S\in M}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z'_{ni,S}\left(D^\tau_{n,0}(S)^{-1}\frac{1}{n}\sum_{j=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z'_{ni,S}\right)\right]\right.\\
&-\frac{1}{n}tr\left(D^\tau_{n,0}(S)^{-1}D^\tau_{n,1}(S)\right)\\
&\left.\quad+\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z'_{ni,S}\left(\hat{\delta}^\tau_{ni,S}-D^\tau_{n,0}(S)^{-1}\frac{1}{n}\sum_{j=1}^{n}\varphi_\tau(e^\tau_{nj,S})Z_{nj,S}\right)\right]\right|\\
&= \sup_{S\in M}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z'_{ni,S}\left(\hat{\delta}^\tau_{ni,S}-D^\tau_{n,0}(S)^{-1}\frac{1}{n}\sum_{j=1}^{n}\varphi_\tau(e^\tau_{nj,S})Z_{nj,S}\right)\right]\right|\\
&\leq \sup_{S\in M}\mathbb{E}_{\mathscr{D}_n}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi_\tau(e^\tau_{ni,S})Z'_{ni,S}\right\|_2\left\|\hat{\delta}^\tau_{ni,S}-D^\tau_{n,0}(S)^{-1}\frac{1}{n}\sum_{j=1}^{n}\varphi_\tau(e^\tau_{nj,S})Z_{nj,S}\right\|_2\right]\\
&= O\left(c_3\left(\frac{m\log\log n}{n}\right)^{1/2}\frac{c_2}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/4}\left(\frac{m\log|M|+\log\log n}{n}\right)^{1/2}\right)\\
&= O\left(\frac{1}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4}\right),
\end{aligned}
$$

where the second to last equality follows from Lemmata 4.5 and 4.6. To conclude, combine this remainder term with the one of Theorem 4.1. □

4.7.4 Proof of Proposition 4.3

We split the proof of Proposition 4.3 in three parts.

Lemma 4.7. *Suppose that Assumptions (A1) – (A6) hold. Then,*

$$
\begin{aligned}
&\sup_{S\in M}\left|tr\left(D^\tau_{n,0}(S)^{-1}\left(\widehat{D}^\tau_{n,1}(S)-D^\tau_{n,1}(S)\right)\right)\right|\\
&\qquad= O_p\left(\frac{m}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}+\frac{m}{\lambda_n}\left(\frac{\log|M|}{n}\right)^{1/2}\right).
\end{aligned}
$$

Remark 4.1. *Since we use the quantile regression errors $\hat{e}^\tau_{ni,S}$ as proxies for the true errors $e^\tau_{ni,S}$ the process $tr\left(D^\tau_{n,0}(S)^{-1}\left(\widehat{D}^\tau_{n,1}(S)-D^\tau_{n,1}(S)\right)\right)$ is not centered. Therefore, we need to control not only the variance (standard deviation) of the process but also its deterministic drift. The deterministic drift is reflected in the first term, the variance in the second term. Note that the rate of deterministic drift can be written a $\frac{m}{\lambda_n}\times r_n$, where r_n is the rate at which the estimated quantile regression vector $\hat{\theta}^\tau_S$converges to θ^τ_S in probability, i.e. the rate at which the estimation bias of the residuals vanishes. As one expects, the rate of the term controlling the variance is proportional to the size of the maximal standard deviation (i.e.*

m) of the n summands and proportional to $(\log|M|)^{1/2}$*, where* $|M|$ *is the size of the finite set over which we take the supremum.*

Remark 4.2. *Clearly, under the stated assumptions, the first rate (controlling the bias) dominates the second rate (controlling the variance).*

Proof of Lemma 4.7. The goal is to apply Markov's inequality. Therefore, in the following we obtain upper bounds on the expected values of certain stochastic processes.

Step 1: Decomposition into deterministic bias and stochastic error terms.

$$
\begin{aligned}
&\sup_{S\in M}\left|tr\left(D_{n,0}^{\tau}(S)^{-1}\left(\widehat{D}_{n,1}^{\tau}(S)-D_{n,1}^{\tau}(S)\right)\right)\right| \\
&=\sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_{\tau}^{2}(\hat{e}_{ni,S}^{\tau})\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2-\mathbb{E}_{\mathscr{D}_n}\left[\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right]\right| \\
&\leq \sup_{S\in M}\sup_{\|\delta_S\|_2\leq r_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\varphi_{\tau}^{2}(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S)-\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\right)\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right. \\
&\qquad\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left(\varphi_{\tau}^{2}(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S)-\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\right)\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right]\right| \\
&\quad+\sup_{S\in M}\sup_{\|\delta_S\|_2\leq r_n}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\varphi_{\tau}^{2}(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S)-\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\right)\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right]\right| \\
&\quad+\sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2-\mathbb{E}_{\mathscr{D}_n}\left[\varphi_{\tau}^{2}(e_{ni,S}^{\tau})\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right]\right| \\
&=\sup_{S\in M}A_n(S)+\sup_{S\in M}B_n(S)+\sup_{S\in M}C_n(S). \qquad (4.24)
\end{aligned}
$$

Step 2: Upper bound on $\mathbb{E}_{\mathscr{D}_n}[\sup_{S\in M}A_n(S)]$**.** Let $\mathscr{D}_n^0$ be an independent copy of $\mathscr{D}_n$ and define

$$
\begin{aligned}
E_n(S)&=\mathbb{E}_{\mathscr{D}_n}\left[\sup_{\|\delta_S\|\leq r_n}\frac{1}{n}\sum_{i=1}^{n}\left((1-2\tau)1\{0<e_{ni,S}^{\tau}\leq Z_{ni,S}'\delta_S\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right.\right. \\
&\qquad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[(1-2\tau)1\{0<e_{ni,S}^{\tau}\leq Z_{ni,S}'\delta_S\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right]\right)\right], \\
W_n(S)&=\mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|\leq r_n}\frac{1}{n^2}\sum_{i=1}^{n}\left((1-2\tau)1\{0<e_{ni,S}^{\tau}\leq Z_{ni,S}'\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2\right.\right. \\
&\qquad\left.\left.-(1-2\tau)1\{0<e_{ni,S}^{0\tau}\leq Z_{ni,S}^{0\prime}\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^{0}\right\|_2^2\right)^2|\mathscr{D}_n\right].
\end{aligned}
$$

Note that for $1 \le p \le 4$,

$$\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \|Z_{ni}\|_2^{2p}\right]^{1/p} \le \mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \|Z_{ni}\|_{2p}^{2p}\right]^{1/p} m^{1/2-1/(2p)} \le \mu_{2p}^2 m^{1/2+1/(2p)}. \quad (4.25)$$

By Lemma 4.2 applied to $g(Z) = (1-2\tau)\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2$ and eq. (4.25) applied to the envelope function $G(Z) = \lambda_n^{-1}\|Z\|_2^2$,

$$E_n(S) \le \frac{c_0 m}{\lambda_n}\left(\frac{m}{n}\right)^{3/4} \bigvee \frac{c_0 m}{\lambda_n} r_n^{1/2}\left(\frac{m}{n}\right)^{1/2} \le \frac{c_0 m}{\lambda_n^{3/2}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4}, \quad (4.26)$$

where $c_0 > 0$ is an absolute constant independent of $S \in M$.

By the Hartman-Wintner law of iterated logarithm, Lemma 4.2 and eq. (4.25),

$$\begin{aligned}
&W_n(S)\\
&\le \sup_{\|\delta_S\|_2 \le r_n} \frac{2}{n^2}\sum_{i=1}^n (1-2\tau)^2 1\{0 < e_{ni,S}^{\tau} \le Z_{ni,S}'\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^4\\
&\quad + \mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{2}{n^2}\sum_{i=1}^n (1-2\tau)^2 1\{0 < e_{ni,S}^{0\tau} \le Z_{ni,S}^{0\prime}\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right\|_2^4\right]\\
&\le \frac{2}{n^2}\sum_{i=1}^n \left(\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^4 - \mathbb{E}_{\mathscr{D}_n}\left[\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^4\right]\right)\\
&\quad + \frac{4\nu_+}{n^2}\sum_{i=1}^n \mathbb{E}_{\mathscr{D}_n}\left[\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^4\right]\\
&\quad + \mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|_2 \le r_n} \frac{2}{n^2}\sum_{i=1}^n \left((1-2\tau)^2 1\{0 < e_{ni,S}^{0\tau} \le Z_{ni,S}^{0\prime}\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right\|_2^4\right.\right.\\
&\qquad\qquad \left.\left. - \mathbb{E}_{\mathscr{D}_n}\left[(1-2\tau)^2 1\{0 < e_{ni,S}^{0\tau} \le Z_{ni,S}^{0\prime}\delta\}\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right\|_2^4\right]\right)\right]\\
&\le \frac{c_1^2 m^2 (\log\log n)^{1/2}}{\lambda_n^2 h^2 n^{3/2}} + \frac{c_1^2 m^2 r_n}{\lambda_n^2 n^{3/2}} + \frac{c_1^2 m^2}{\lambda_n^2 n} \quad a.s.\\
&\le \frac{c_1^2 m^2}{\lambda_n^2 n} \quad a.s., \qquad (4.27)
\end{aligned}$$

where $c_1 > 0$ is an absolute constant independent of $S \in M$.

Note that

$$\left(\varphi_\tau^2(e_{ni,S}^{\tau} - Z_{ni,S}'\delta) - \varphi_\tau^2(e_{ni,S}^{\tau})\right)\left\|D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right\|_2^2$$

$$= (1-2\tau)1\{0 < e^{\tau}_{ni,S} \le Z'_{ni,S}\delta\} \left\| D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right\|_2^2.$$

Thus, for fixed $S \in M$ Lemma 4.1 and eq. (4.26)–(4.27) yield for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\|\delta_S\| \le r_n} \frac{1}{n} \sum_{i=1}^{n} \left(\left(\varphi^2_\tau(e^{\tau}_{ni,S} - Z'_{ni,S}\delta) - \varphi^2_\tau(e^{\tau}_{ni,S})\right) \left\| D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right\|_2^2 \right.\right. \\
&\qquad \left.\left. - \mathbb{E}_{\mathscr{D}_n}\left[\left(\varphi^2_\tau(e^{\tau}_{ni,S} - Z'_{ni,S}\delta) - \varphi^2_\tau(e^{\tau}_{ni,S})\right) \left\| D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right\|_2^2 \right] \right) \\
&\quad \ge \frac{c_0 m}{\lambda_n^{3/2}} \left(\frac{m \log|M| \log\log n}{n} \right)^{3/4} + 2\frac{c_1 m t^{1/2}}{\lambda_n n^{1/2}} \Bigg) \\
&\le 4e e^{-t/2}. \qquad (4.28)
\end{aligned}$$

Now, set t to $\log|M| + t^2$ and integrate out the tail bound,

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S \in M} A_n(S) \right] \\
&\le \frac{c_0 m}{\lambda_n^{3/2}} \left(\frac{m \log|M| \log\log n}{n} \right)^{3/4} + \frac{2c_1 m}{\lambda_n} \left(\frac{\log|M|}{n} \right)^{1/2} \\
&\quad + \frac{2c_1 m}{\lambda_n n^{1/2}} \int_0^{\infty} \mathbb{P}\left(\sup_{S \in M} A_n(S) \ge \frac{c_0 m}{\lambda_n^{3/2}} \left(\frac{m \log|M| \log\log n}{n} \right)^{3/4} \right. \qquad (4.29) \\
&\qquad \left. + \frac{2c_1 m}{\lambda_n n^{1/2}} \left(t + (\log|M|)^{1/2} \right) \right) dt \\
&\le \frac{c_0 m}{\lambda_n^{3/2}} \left(\frac{m \log|M| \log\log n}{n} \right)^{3/4} + \frac{2c_1 m}{\lambda_n} \left(\frac{\log|M|}{n} \right)^{1/2} + \frac{2c_1 m}{\lambda_n n^{1/2}} \int_0^{\infty} e^{-t^2/2} dt \\
&\le \frac{c_2 m}{\lambda_n} \left(\frac{m \log|M| \log\log n}{n} \right)^{1/2}, \qquad (4.30)
\end{aligned}$$

where $c_2 > 0$ is an absolute constant and the last inequality follows from the rate condition (A5).

Step 3: Upper bound on $\sup_{S \in M} B_n(S)$.

$$\begin{aligned}
&\sup_{S \in M} B_n(S) \\
&\le 2 \sup_{S \in M} \sup_{\|\delta_S\|_2 \le r_n} \mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n} \sum_{i=1}^{n} \left| F_{e^{\tau}_{n,S} | X_{n,S}}(Z'_{ni,S}\delta_S) - F_{e^{\tau}_{n,S} | X_{n,S}}(0) \right| \left\| D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right\|_2^2 \right] \\
&\le 2\nu_+ \sup_{S \in M} \sup_{\|\delta_S\|_2 \le r_n} \mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n} \sum_{i=1}^{n} \left| Z'_{ni,S}\delta_S \right| \left\| D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right\|_2^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\nu_+ \sup_{S\in M}\sup_{\|\delta_S\|_2\leq r_n} \mathbb{E}_{\mathscr{D}_n}\left[\delta_S' Z_{n,S} Z_{n,S}' \delta_S\right]^{1/2} \mathbb{E}_{\mathscr{D}_n}\left[\left\|D_{n,0}^{\tau}(S)^{-1/2} X_{ni,S}\right\|_2^4\right]^{1/2} \\
&\leq \frac{2\lambda_+^{1/2}\nu_+}{\lambda_n} r_n \sup_{S\in M} \mathbb{E}_{\mathscr{D}_n}\left[\left\|Z_{ni,S}\right\|_4^4\right]^{1/2} m^{(4-2)/4} \\
&\leq \frac{2\lambda_+^{1/2}\nu_+\mu_4^2}{\lambda_n} r_n m \\
&= 2c_3\lambda_+^{1/2}\nu_+\mu_4^2 \frac{m}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}. \qquad (4.31)
\end{aligned}
$$

Step 4: Upper bound on $\mathbb{E}_{\mathscr{D}_n}[\sup_{S\in M} C_n(S)]$. Let $\mathscr{D}_n^0$ be an independent copy of $\mathscr{D}_n$ and define

$$
\begin{aligned}
E_n(S) = \mathbb{E}_{\mathscr{D}_n}\Bigg[\sup_{u\in\{-1,1\}} \frac{1}{n}\sum_{i=1}^n \Bigg(&\varphi_\tau^2(e_{ni,S}^\tau)\left\|D_{n,0}^\tau(S)^{-1/2} Z_{ni,S}\right\|_2^2 u - \\
&\mathbb{E}_{\mathscr{D}_n}\left[\varphi_\tau^2(e_{ni,S}^\tau)\left\|D_{n,0}^\tau(S)^{-1/2} Z_{ni,S}\right\|_2^2 u\right]\Bigg)\Bigg], \\
W_n(S) = \mathbb{E}_{\mathscr{D}_n^0}\Bigg[\sup_{u\in\{-1,1\}} \frac{1}{n^2}\sum_{i=1}^n \Bigg(&\varphi_\tau^2(e_{ni,S}^\tau)\left\|D_{n,0}^\tau(S)^{-1/2} Z_{ni,S}\right\|_2^2 u - \\
&\varphi_\tau^2(e_{ni,S}^{0\tau})\left\|D_{n,0}^\tau(S)^{-1/2} Z_{ni,S}^0\right\|_2^2 u\Bigg)^2 \,|\, \mathscr{D}_n\Bigg].
\end{aligned}
$$

Now, proceed as in Step 2. We obtain,

$$
\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M} C_n(S)\right] \leq \frac{c_5 m}{\lambda_n}\left(\frac{\log|M|}{n}\right)^{1/2}, \qquad (4.32)
$$

where $c_5 > 0$ is an absolute constant independent of n, m, and M.

Step 5: Conclusion: The claim follows from Markov's inequality and the bounds (4.26)–(4.32). Note that the bound (4.29) is dominated by the bound (4.31). □

Lemma 4.8. *Suppose that Assumptions (A1) – (A6) hold. Then, for any $h > 0$,*

$$
\begin{aligned}
&\sup_{S\in M}\left\|D_{n,0}^\tau(S)^{-1}\left(\widehat{D}_{0,h}^\tau(S) - D_{n,0}^\tau(S)\right)\right\| \\
&= O_p\left(\frac{h^\alpha}{\lambda_n} \bigvee \frac{1}{\lambda_n^{1+\alpha}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{\alpha/2} \bigvee \frac{1}{\lambda_n h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}\right).
\end{aligned}
$$

Remark 4.3. *The process $\left\|D_{n,0}^\tau(S)^{-1}\left(\widehat{D}_{0,h}^\tau(S) - D_{n,0}^\tau(S)\right)\right\|$ is not centered; as in Lemma 4.7 we need to control variance and a deterministic drift term: the first term captures the*

bias of the non-parametric estimation technique, the second term captures the bias of using the quantile regression errors $\hat{e}^{\tau}_{ni,S}$ as proxies for the true errors $e^{\tau}_{ni,S}$, and the third term captures the variance of the non-parametric estimate. Note that the rate of the second drift term can be written as $\frac{1}{\lambda_n} \times {r_n}^{\alpha}$, where α is the Hölder-continuity coefficient and r_n is the rate at which the estimated quantile regression vector $\hat{\theta}^{\tau}_S$ converges to θ^{τ}_S in probability, i.e. the rate at which the estimation bias of the residuals vanishes. The $\log\log n$*-factor in the third term, is an artifact of our proof (for details, see comment at the beginning of the proof). However, apart from* $\log\log n$*-factor, the rate of the third term matches the rates of comparable results (e.g. Vershynin, 2012a, Theorem 5.45).*

Proof of Lemma 4.8. The operator norm requires a different approach than the proof of Lemma 4.7. Since we take the supremum over all $S \in M$ a natural idea is to use a uniform version of Rudelson's inequality (e.g. Rudelson and Vershynin, 2008, Lemma 3.6). However, Rudelson's uniform inequality requires bounded predictors Z_{ni} and is not easy to modify to also handle either dependent matrices or the supremum over $\delta_S \in \mathbb{R}^{|S|}$ with $\|\delta_S\|_2 \leq r_n$. Thus, instead of bounding the expected value and applying Makrov's inequality (as we did in the proof of Lemma 4.7), we use Lemma 4.1 to bound the tail probability, apply the union bound, and then integrate the tail probability to upper bound the expected value.

Let $K_h(u) = \frac{1}{2} 1\{|u| \leq h\}$.

Step 1: Decomposition into deterministic bias and stochastic error terms.

$$
\begin{aligned}
&\sup_{S \in M} \left\| D^{\tau}_{n,0}(S)^{-1} \left(\widehat{D}^{\tau}_{0,h}(S) - D^{\tau}_{n,0}(S) \right) \right\| \\
&= \sup_{S \in M} \sup_{\|v\|_2 = 1} \left| \frac{1}{nh} \sum_{i=1}^{n} \left(\left| v' D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right|^2 K_h \left(e^{\tau}_{ni,S} - Z'_{ni,S} \hat{\delta}_{n,s} \right) \right. \right. \\
&\quad \left. \left. - h \, \mathbb{E}_{\mathscr{D}_n} \left[f_{e^{\tau}_{n,S} | X_n}(0 | X_{ni}) \left| v' D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right|^2 \right] \right) \right| \\
&\leq \sup_{S \in M} \sup_{\|v\|_2 = 1} \sup_{\|\delta_S\|_2 \leq r_n} \left| \frac{1}{nh} \sum_{i=1}^{n} \left(\left| v' D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right|^2 \right. \right. \\
&\qquad \times \left[K_h \left(e^{\tau}_{ni,S} - Z'_{ni,S} \delta_S \right) - K_h \left(e^{\tau}_{ni,S} \right) \right] \\
&\quad \left. \left. - \mathbb{E}_{\mathscr{D}_n} \left[\left| v' D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right|^2 \left[K_h \left(e^{\tau}_{ni,S} - Z'_{ni,S} \delta_S \right) - K_h \left(e^{\tau}_{ni,S} \right) \right] \right] \right) \right| \\
&\quad + \sup_{S \in M} \sup_{\|v\|_2 = 1} \sup_{\|\delta_S\|_2 \leq r_n} \left| \mathbb{E}_{\mathscr{D}_n} \left[\frac{1}{nh} \sum_{i=1}^{n} \left| v' D^{\tau}_{n,0}(S)^{-1/2} Z_{ni,S} \right|^2 \left(K_h \left(e^{\tau}_{ni,S} - Z'_{ni,S} \delta_S \right) \right. \right. \right. \\
&\qquad \left. \left. \left. - h f_{e^{\tau}_{n,S} | X_n}(0 | X_{ni}) \right) \right] \right|
\end{aligned}
$$

$$
\begin{aligned}
&+\sup_{S\in M}\sup_{\|v\|_2=1}\left|\frac{1}{nh}\sum_{i=1}^{n}\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 K_h\left(e_{ni,S}^{\tau}\right)\right.\right.\\
&\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 K_h\left(e_{ni,S}^{\tau}\right)\right]\right)\right|\\
&=\sup_{S\in M}\sup_{\|v\|_2=1}A_n(S,v)+\sup_{S\in M}\sup_{\|v\|_2=1}B_n(S,v)+\sup_{S\in M}\sup_{\|v\|_2=1}C_n(S,v).
\end{aligned}
$$

Step 2: Upper bound on $\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\sup_{\|v\|_2=1}A_n(S,v)\right]$. Let $\mathscr{D}_n^0$ be an independent copy of $\mathscr{D}_n$ and define

$$
\begin{aligned}
E_n(S,v)&=\mathbb{E}_{\mathscr{D}_n}\left[\sup_{\|\delta_S\|\le r_n}\frac{1}{nh}\sum_{i=1}^{n}\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right.\right.\\
&\qquad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right]\right)\right],\\
W_n(S,v)&=\mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|\le r_n}\frac{1}{(nh)^2}\sum_{i=1}^{n}\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right.\right.\\
&\qquad\left.\left.-\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right|^2 1\{h<e_{ni,S}^{0\tau}\le Z_{ni,S}^{0\prime}\delta_S+h\}\right)^2|\mathscr{D}_n\right].
\end{aligned}
$$

By Lemma 4.2 applied to $g(Z)=\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2$ and Lemma 4.3 applied to the envelope $G(Z)\equiv g(Z)$,

$$
E_n(S,v)\le\frac{c_0}{\lambda_n h}\left(\frac{m}{n}\right)^{3/4}\bigvee\frac{c_0}{\lambda_n h}r_n^{1/2}\left(\frac{m}{n}\right)^{1/2}\le\frac{c_0}{\lambda_n^{3/2}h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4},\tag{4.33}
$$

where $c_0>0$ is an absolute constant independent of $S\in M$ and $v\in\mathbb{R}^{|S|}$.

By the Hartman-Wintner law of iterated logarithm, Lemma 4.2 and Lemma 4.3,

$$
\begin{aligned}
&W_n(S,v)\\
&\le\sup_{\|\delta_S\|_2\le r_n}\frac{2}{(nh)^2}\sum_{i=1}^{n}\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^4 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\\
&+\mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|_2\le r_n}\frac{2}{(nh)^2}\sum_{i=1}^{n}\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right|^4 1\{h<e_{ni,S}^{0\tau}\le Z_{ni,S}^{0\prime}\delta_S+h\}\right]\\
&\le\frac{2}{(nh)^2}\sum_{i=1}^{n}\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^4-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^4\right]\right)\\
&+\frac{4\nu_+}{(nh)^2}\sum_{i=1}^{n}\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^4\right]
\end{aligned}
\tag{4.34}
$$

$$
\begin{aligned}
&+\mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{\|\delta_S\|_2\le r_n}\frac{2}{(nh)^2}\sum_{i=1}^n\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right|^4 1\{h<e_{ni,S}^{0\tau}\le Z_{ni,S}^{0\prime}\delta_S+h\}\right.\right.\\
&\qquad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}^0\right|^4 1\{h<e_{ni,S}^{0\tau}\le Z_{ni,S}^{0\prime}\delta_S+h\}\right]\right)\right]\\
&\le\frac{c_1^2(\log\log n)^{1/2}}{\lambda_n^2h^2n^{3/2}}+\frac{c_1^2r_n}{\lambda_n^2h^2n^{3/2}}+\frac{c_1^2}{\lambda_n^2h^2n}\quad a.s.\\
&\le\frac{c_1^2}{\lambda_n^2h^2n}\quad a.s.,
\end{aligned}
\tag{4.35}
$$

where $c_1>0$ is an absolute constant independent of $S\in M$ and $v\in\mathbb{R}^{|S|}$.

By definition of $K_h(u)$,

$$
\begin{aligned}
&\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2\left[K_h\left(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S\right)-K_h\left(e_{ni,S}^{\tau}\right)\right]\\
&\quad=\frac{1}{2}\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2\left|1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right.\\
&\qquad\left.-1\{-h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S-h\}\right|.
\end{aligned}
$$

Thus, for fixed $S\in M$ and $v\in\mathbb{R}^{|S|}$ Lemma 4.1 and eq. (4.33)– (4.34) yield for any $t>0$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\|\delta_S\|\le r_n}\frac{1}{nh}\sum_{i=1}^n\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2\left[K_h\left(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S\right)-K_h\left(e_{ni,S}^{\tau}\right)\right]\right.\right.\\
&\qquad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2\left[K_h\left(e_{ni,S}^{\tau}-Z_{ni,S}'\delta_S\right)-K_h\left(e_{ni,S}^{\tau}\right)\right]\right]\right)\\
&\quad\ge\frac{c_0}{\lambda_n^{3/2}h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4}+2\frac{c_1t^{1/2}}{\lambda_nhn^{1/2}}\Bigg)\\
&\le 2\mathbb{P}\left(\sup_{\|\delta_S\|\le r_n}\frac{1}{nh}\sum_{i=1}^n\left(\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right.\right.\\
&\qquad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D_{n,0}^{\tau}(S)^{-1/2}Z_{ni,S}\right|^2 1\{h<e_{ni,S}^{\tau}\le Z_{ni,S}'\delta_S+h\}\right]\right)\right.\\
&\quad\left.\ge E_n(S,v)+2W_n^{1/2}(S,v)t^{1/2}\right)\\
&\le 8ee^{-t/2}.
\end{aligned}
\tag{4.36}
$$

Let $\mathscr{N}_S$ be an $\frac{1}{3}$-net of the $|S|$-dimensional unit sphere. Then $|\mathscr{N}_S|\le 7^{|S|}$ and for any symmetric $|S|\times|S|$-dimensional matrix A we have $\sup_{\|v\|=1}|v'Av|\le 3\sup_{v\in\mathscr{N}_S}|v'Av|$ (e.g.

Vershynin, 2012b, Lemma 5.4). Thus,

$$
\begin{aligned}
&\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\sup_{\|v\|_2=1} A_n(S,v)\right] \\
&\leq 3\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\sup_{v\in\mathscr{N}_S} A_n(S,v)\right] \\
&\leq \frac{3c_0}{\lambda_n^{3/2}h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4}+\frac{6c_1}{\lambda_n h}\left(\frac{\log|M|+\log\log n}{n}\right)^{1/2} \\
&\quad+\frac{6c_1}{\lambda_n h n^{1/2}}\int_0^\infty \mathbb{P}\left(\sup_{S\in M}\sup_{\|v\|_2=1} A_n(S,v)\geq \frac{c_0}{\lambda_n^{3/2}h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4}\right. \\
&\qquad\left.+\frac{2c_1}{\lambda_n h n^{1/2}}\left(t+(\log|M|+m\log 7)^{1/2}\right)\right)dt \\
&\leq \frac{3c_0}{\lambda_n^{3/2}h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{3/4}+\frac{6c_1}{\lambda_n h}\left(\frac{\log|M|+\log\log n}{n}\right)^{1/2} \\
&\qquad+\frac{6c_1}{\lambda_n h n^{1/2}}\int_0^\infty e^{-t^2/2}dt \\
&\leq \frac{c_2}{\lambda_n h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}, \qquad (4.37)
\end{aligned}
$$

where $c_2>0$ is an absolute constant and the last inequality follows from the rate condition (A5).

Step 3: Upper bound on $\sup_{S\in M}\sup_{\|v\|_2=1} B_n(S,v)$**.** The Hölder-continuity of $f_{e^\tau_{n,S}|X_n}$ yields

$$
\begin{aligned}
&\sup_{S\in M}\sup_{\|v\|_2=1} B_n(S,v) \\
&\leq \sup_{S\in M}\sup_{\|v\|_2=1}\sup_{\|\delta_S\|_2\leq r_n}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n\int K_1(u)\left[f_{e^\tau_{n,S}|X_n}(Z'_{ni,S}\delta_S+uh|X_{ni})\right.\right.\right. \\
&\qquad\left.\left.\left.-f_{e^\tau_{n,S}|X_n}(Z'_{ni,S}\delta_S|X_{ni})\right]du\times\left|v'D^\tau_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2\right]\right| \\
&\quad+\sup_{S\in M}\sup_{\|v\|_2=1}\sup_{\|\delta_S\|_2\leq r_n}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n\int K_1(u)\left[f_{e^\tau_{n,S}|X_n}(Z'_{ni,S}\delta|X_{ni})\right.\right.\right. \\
&\qquad\left.\left.\left.-f_{e^\tau_{n,S}|X_n}(0|X_{ni})\right]du\times\left|v'D^\tau_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2\right]\right| \\
&\leq \sup_{S\in M}\sup_{\|v\|_2=1}\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n\int K_1(u)\nu_H|uh|^\alpha du\left|v'D^\tau_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\sup_{S\in M}\sup_{\|v\|_2=1}\sup_{\|\delta_S\|_2\le r_n}\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\int K_1(u)\nu_H\left|Z'_{ni,S}\delta_S\right|^{\alpha}du\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2\right]\\
&\le\frac{\nu_H}{\lambda_n}\left(\int K_1(u)\left|u\right|^{\alpha}du\right)h^{\alpha}\\
&+\nu_H r_n^{\alpha}\sup_{S\in M}\sup_{\|v\|_2=1}\sup_{\|u\|_2=1}\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Z'_{ni,S}u\right|^{\alpha}\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2\right]\\
&\le\frac{\nu_H}{\lambda_n}\left(\int K_1(u)\left|u\right|^{\alpha}du\right)h^{\alpha}\\
&+\nu_H r_n^{\alpha}\sup_{S\in M}\sup_{\|u\|_2=1}\left(\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Z'_{ni,S}u\right|^2\right]\right)^{\alpha/2}\\
&\times\sup_{\|v\|_2=1}\left(\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^4\right]\right)^{1/2}\\
&\le\frac{c_5}{\lambda_n}h^{\alpha}+c_{\alpha}r_n^{\alpha}\lambda_n^{-1}\\
&\le\frac{c_5}{\lambda_n}h^{\alpha}+\frac{c_{\alpha}c_3^{\alpha}}{\lambda_n^{1+\alpha}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{\alpha/2},
\end{aligned}
\tag{4.38}
$$

where $c_5, c_3, c_\alpha > 0$ are absolute constants independent of $S \in M$ and $v \in \mathbb{R}^{|S|}$ (see Lemma 4.3)

.

Step 4: Upper bound on $\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\sup_{\|v\|_2=1}C_n(S,v)\right]$. Let $\mathscr{D}_n^0$ be an independent copy of $\mathscr{D}_n$ and define

$$
\begin{aligned}
E_n(S,v)&=\mathbb{E}_{\mathscr{D}_n}\left[\sup_{u\in\{-1,1\}}\frac{1}{nh}\sum_{i=1}^{n}\left(\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2K_h\left(e^{\tau}_{ni,S}\right)u\right.\right.\\
&\quad\left.\left.-\mathbb{E}_{\mathscr{D}_n}\left[\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2K_h\left(e^{\tau}_{ni,S}\right)u\right]\right)\right],\\
W_n(S,v)&=\mathbb{E}_{\mathscr{D}_n^0}\left[\sup_{u\in\{-1,1\}}\frac{1}{(nh)^2}\sum_{i=1}^{n}\left(\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z_{ni,S}\right|^2K_h\left(e^{\tau}_{ni,S}\right)u\right.\right.\\
&\quad\left.\left.-\left|v'D^{\tau}_{n,0}(S)^{-1/2}Z^0_{ni,S}\right|^2K_h\left(e^{0\tau}_{ni,S}\right)u\right)^2|\mathscr{D}_n\right].
\end{aligned}
$$

Now, proceed as in Step 2. We obtain,

$$
\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\sup_{\|v\|_2=1}C_n(S,v)\right]\le\frac{c_6}{\lambda_n h}\left(\frac{m\log|M|}{n}\right)^{1/2},
\tag{4.39}
$$

where $c_6 > 0$ is an absolute constant.

Step 5: Conclusion: The claim follows from Markov's inequality and the bounds (4.37)–(4.39). □

Proof of Proposition 4.3. We factor the stochastic process in two processes involving the processes in Lemmas 4.7 and 4.8. Then, convergence in probability at the given rate follows immediately.

Factorization.

$$
\begin{aligned}
&\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{0,h}(S)^{-1}\widehat{D}^{\tau}_{n,1}(S)\right)-tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right| \\
&=\sup_{S\in M}\left|tr\left(\left(\widehat{D}^{\tau}_{0,h}(S)^{-1}-D^{\tau}_{n,0}(S)^{-1}\right)\widehat{D}^{\tau}_{n,1}(S)\right)\right. \\
&\quad\left.+tr\left(D^{\tau}_{n,0}(S)^{-1}\left(\widehat{D}^{\tau}_{n,1}(S)-D^{\tau}_{n,1}(S)\right)\right)\right| \\
&\leq\left(\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|\right)\left(\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{n,1}(S)\widehat{D}^{\tau}_{0,h}(S)^{-1}\right)\right|\right) \\
&\quad+\sup_{S\in M}\left|tr\left(D^{\tau}_{n,0}(S)^{-1}\left(\widehat{D}^{\tau}_{n,1}(S)-D^{\tau}_{n,1}(S)\right)\right)\right| \\
&\leq\left(\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|\right)\left(\sup_{S\in M}\left|tr\left(D^{\tau}_{n,1}(S)D^{\tau}_{0,h}(S)^{-1}\right)\right|\right) \\
&\quad+\left(\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|\right) \\
&\quad\times\left(\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{n,1}(S)\widehat{D}^{\tau}_{0,h}(S)^{-1}-D^{\tau}_{n,1}(S)D^{\tau}_{0,h}(S)^{-1}\right)\right|\right) \\
&\quad+\sup_{S\in M}\left|tr\left(D^{\tau}_{n,0}(S)^{-1}\left(\widehat{D}^{\tau}_{n,1}(S)-D^{\tau}_{n,1}(S)\right)\right)\right|.
\end{aligned}
$$

Re-arranging and solving for $\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{0,h}(S)^{-1}\widehat{D}^{\tau}_{n,1}(S)\right)-tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right|$ gives

$$
\begin{aligned}
&\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{0,h}(S)^{-1}\widehat{D}^{\tau}_{n,1}(S)\right)-tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right| \\
&=\frac{\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|}{1-\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|}\sup_{S\in M}\left|tr\left(D^{\tau}_{n,1}(S)D^{\tau}_{0,h}(S)^{-1}\right)\right| \\
&\quad+\frac{\sup_{S\in M}\left|tr\left(D^{\tau}_{n,0}(S)^{-1}\left(\widehat{D}^{\tau}_{n,1}(S)-D^{\tau}_{n,1}(S)\right)\right)\right|}{1-\sup_{S\in M}\left\|\left(D_{0,S}-\widehat{D}^{h}_{0,S}\right)D^{\tau}_{n,0}(S)^{-1}\right\|}.
\end{aligned}
\tag{4.40}
$$

Thus, by Lemmata 4.7 and 4.8,

$$
\sup_{S\in M}\left|tr\left(\widehat{D}^{\tau}_{0,h}(S)^{-1}\widehat{D}^{\tau}_{n,1}(S)\right)-tr\left(D^{\tau}_{n,0}(S)^{-1}D^{\tau}_{n,1}(S)\right)\right|
$$

$$
\begin{aligned}
&= O_p\left(\frac{mh^\alpha}{\lambda_n^2} \bigvee \frac{m}{\lambda_n^{2+\alpha}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{\alpha/2} \bigvee \frac{m}{\lambda_n^2 h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2}\right) \\
&+ O_p\left(\frac{m}{\lambda_n^2}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2} + \frac{m}{\lambda_n}\left(\frac{\log|M|}{n}\right)^{1/2}\right) \\
&= O_p\left(\frac{mh^\alpha}{\lambda_n^2} + \frac{m}{\lambda_n^2 h}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{1/2} + \frac{m}{\lambda_n^{2+\alpha}}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{\alpha/2}\right),
\end{aligned}
\tag{4.41}
$$

by Assumption (A5) on the lower bound on λ_n.

□

4.7.5 Proof of Theorem 4.3

Proof of Theorem 4.3. By Theorem 4.2 we have $\inf_{S\in M} b_n^\tau(S) > 0$ and

$$
\inf_{S\in M} b_n^\tau(S) \gtrsim O\left(n^{-1}\right). \tag{4.42}
$$

Let $r_{n,2}$ be as defined in Theorem 4.2 and fix $T > 0$. By Theorem 4.2 there exists $T > 0$ and $N > 0$ such that for all $n \geq N$,

$$
\sup_{S\in M}\left|b_n^\tau(S) - \frac{1}{n}tr\left(D_{n,0}^\tau(S)^{-1}D_{n,1}^\tau(S)\right)\right| \leq Tr_{n,2}.
$$

Thus, by Proposition 4.3 for any $\varepsilon > 0$ there exist $T' > T$ and $N' \geq N$ such that for all $n > N'$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{S\in M}\left|\frac{\hat{b}_n^\tau(S)}{b_n^\tau(S)} - 1\right| > \frac{T}{\inf_{S\in M} b_n^\tau(S)}\left(\frac{m\,h^\alpha}{\lambda_n n} + \frac{m\,r_n}{nh} + \frac{m\,r_n^\alpha}{\lambda_n n} + r_{n,2}\right)\right) \\
&\leq \mathbb{P}\left(\sup_{S\in M}\left|\hat{b}_n^\tau(S) - \frac{1}{n}tr\left(D_{n,0}^\tau(S)^{-1}D_{n,1}^\tau(S)\right)\right|\right. \\
&\qquad \left. + \sup_{S\in M}\left|b_n^\tau(S) - \frac{1}{n}tr\left(D_{n,0}^\tau(S)^{-1}D_{n,1}^\tau(S)\right)\right| > \frac{T}{n}\left(\frac{h^\alpha}{\lambda_n} + \frac{r_n}{h} + \frac{m\,r_n^\alpha}{\lambda_n}\right) + Tr_{n,2}\right) \\
&\leq \mathbb{P}\left(\sup_{S\in M}\left|n\hat{b}_n^\tau(S) - tr\left(D_{n,0}^\tau(S)^{-1}D_{n,1}^\tau(S)\right)\right| > T\left(\frac{m\,h^\alpha}{\lambda_n} + \frac{m\,r_n}{h} + \frac{m\,r_n^\alpha}{\lambda_n}\right)\right) \\
&< \varepsilon.
\end{aligned}
$$

To conclude, note that by eq. (4.42)

$$\begin{aligned}\frac{r_{n,4}}{\inf_{S\in M} b_n^\tau(S)} &= O\left(\frac{n}{\lambda_n}\left(\frac{m\log|M|\,\log\log n}{n}\right)^{5/4} + \left(\frac{m\,h^\alpha}{\lambda_n} + \frac{m\,r_n}{h} + \frac{m\,r_n^\alpha}{\lambda_n}\right)\right)\\ &= O\left(n\lambda_n^{3/2} r_n^{5/2} + \frac{m\,h^\alpha}{\lambda_n} + \frac{m\,r_n}{h} + \frac{m\,r_n^\alpha}{\lambda_n}\right).\end{aligned}$$

□

4.7.6 Proof of Theorem 4.4

Step 1: Decomposition.

$$\begin{aligned}&\sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y_{ni}) + tr\left(\widehat D^\tau_{h,0}(S)^{-1}\widehat D^\tau_{n,1}(S)\right)\right.\\ &\qquad\left. -\mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y^0_{ni} - Z^{0'}_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y^0_{ni})\right]\right|\\ &\le \sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right.\\ &\qquad\left. -\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right]\right|\\ &\quad + \sup_{S\in M}\left|\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y_{ni})\right]\right.\\ &\qquad\left. -\mathbb{E}_{\mathscr{D}_n,(Y_n^0,X_n^0)}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y^0_{ni} - Z^{0'}_{ni,S}\hat\theta^\tau_{n,S}) - \rho_\tau(Y^0_{ni})\right] + \frac{1}{n}tr\left(D^\tau_{n,0}(S)^{-1}D^\tau_{n,1}(S)\right)\right|\\ &\quad + \sup_{S\in M}\frac{1}{n}\left|tr\left(\widehat D^\tau_{h,0}(S)^{-1}\widehat D^\tau_{n,1}(S)\right) - tr\left(D^\tau_{n,0}(S)^{-1}D^\tau_{n,1}(S)\right)\right|\\ &\quad + \sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S}) - \rho_\tau(Y_{ni})\right.\\ &\qquad\left. -\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S}) - \rho_\tau(Y_{ni})\right]\right|\\ &= \sup_{S\in M} A_n(S) + \sup_{S\in M} B_n(S) + \sup_{S\in M} C_n(S) + \sup_{S\in M} D_n(S). \qquad (4.43)\end{aligned}$$

Step 2: Bounds on $\sup_{S\in M} B_n(S)$ and $\sup_{S\in M} C_n(S)$. By Theorem 4.2,

$$\sup_{S\in M} B_n(S) = O\left(\lambda_n^{3/2} r_n^{5/2}\right) \quad a.s.,$$

and by Theorem 4.3,

$$\sup_{S\in M} C_n(S) = O_p\left(\frac{m\,h^{\alpha}}{\lambda_n^2 n} + \frac{m\,r_n}{h\lambda_n n} + \frac{m\,r_n^{\alpha}}{\lambda_n^2 n}\right).$$

Step 3: Bound on $\sup_{S\in M} D_n(S)$**.** Recall that $\theta^{\tau}_{n,S}$ solves the population quantile regression minimization problem under the constraint that the minimizer is a linear function, while $Q_{Y_n}(\tau|X_n)$ solves the unconstrained population quantile regression minimization problem. Therefore, we have for all $i = 1,\ldots,n$,

$$\begin{aligned}
0 &\leq \mathbb{E}_{\mathscr{D}_n}\left[\min_{S\in M}\rho_\tau(Y_{ni}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^{\tau}_{n,S}\right] \\
&\leq \mathbb{E}_{\mathscr{D}_n}\left[\min_{S\in M}\rho_\tau(Y_{ni}) - \rho_\tau(Y_{ni} - Q_{Y_n}(\tau|X_{ni}))\right] \\
&\leq \mathbb{E}_{\mathscr{D}_n}\left[\left|Q_{Y_n}(\tau|X_{ni}))\right|\right] < \infty. \qquad (4.44)
\end{aligned}$$

The chain of inequalities in 4.44, the convexity of the maximum operator together with Jensen's Inequality, and eq. 4.7 imply

$$\begin{aligned}
0 &\leq \mathbb{E}_{\mathscr{D}_n}\left[\min_{S\in M}\left\{\rho_\tau(Y_{ni}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^{\tau}_{n,S})\right\}\right] \\
&\leq \mathbb{E}_{\mathscr{D}_n}\left[\left|Q_{Y_n}(\tau|X_{ni}))\right|\right] + \mathbb{E}_{\mathscr{D}_n}\left[\min_{S\in M}\left\{-\rho_\tau(Y_{ni} - Z'_{ni,S}\theta^{\tau}_{n,S}) + \rho_\tau(Y_{ni} - Q_{Y_n}(\tau|X_{ni}))\right\}\right] \\
&\leq \mathbb{E}_{\mathscr{D}_n}\left[\left|Q_{Y_n}(\tau|X_{ni}))\right|\right] - \mathbb{E}_{\mathscr{D}_n}\left[\max_{S\in M}\left\{\rho_\tau(Y_{ni} - Z'_{ni,S}\theta^{\tau}_{n,S}) - \rho_\tau(Y_{ni} - Q_{Y_n}(\tau|X_{ni}))\right\}\right] \\
&\leq \mathbb{E}_{\mathscr{D}_n}\left[\left|Q_{Y_n}(\tau|X_{ni}))\right|\right] - \frac{\nu_-}{2}\mathbb{E}_{\mathscr{D}_n}\left[\max_{S\in M}\left(Z'_{ni,S}\theta^{\tau}_{n,S} - Q_{Y_n}(\tau|X_n)\right)^2\right].
\end{aligned}$$

For $i = 1,\ldots,n$ above inequality gives

$$\mathbb{E}_{\mathscr{D}_n}\left[\max_{S\in M}\left(Z'_{ni,S}\theta^{\tau}_{n,S} - Q_{Y_n}(\tau|X_{ni})\right)^2\right] \leq \frac{2}{\nu_-}\mathbb{E}_{\mathscr{D}_n}\left[\left|Q_{Y_n}(\tau|X_{ni}))\right|\right] < \infty. \qquad (4.45)$$

Let $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)$ be a vector of i.i.d. Rademacher variables independent of $\mathscr{D}_n$. By the Sub-Gaussianity of the conditional Rademacher average, the Lipschitz continuity of the quantile loss function, and eq. (4.45),

$$\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_{ni} - Z_{ni,S}\theta^{\tau}_{n,S}) - \rho_\tau(Y_{ni}) - \mathbb{E}_{\mathscr{D}_n}\left[\rho_\tau(Y_{ni} - Z_{ni,S}\theta^{\tau}_{n,S}) - \rho_\tau(Y_{ni})\right]\right|\right]$$

$$
\begin{aligned}
&\leq \left(\mathbb{E}_{\mathscr{D}_n} \left[\sup_{S \in M} \frac{1}{n} \sum_{i=1}^{n} \left(Z'_{ni,S} \theta^{\tau}_{n,S} \right)^2 \right] \right)^{1/2} \left(\frac{\log |M|}{n} \right)^{1/2} \\
&\leq \left[\left(\mathbb{E}_{\mathscr{D}_n} \left[\sup_{S \in M} \frac{1}{n} \sum_{i=1}^{n} \left(Z'_{ni,S} \theta^{\tau}_{n,S} - Q_{Y_n}(\tau | X_{ni}) \right)^2 \right] \right)^{1/2} \right. \\
&\qquad \left. + \left(\mathbb{E}_{\mathscr{D}_n} \left[\frac{1}{n} \sum_{i=1}^{n} Q^2_{Y_n}(\tau | X_{ni}) \right] \right)^{1/2} \right] \left(\frac{\log |M|}{n} \right)^{1/2} \\
&\leq c_1 \left(\frac{\log |M|}{n} \right)^{1/2},
\end{aligned}
$$

where $c_1 > 0$ depends on the constants in eq. (4.45). Thus,

$$
\sup_{S \in M} D_n(S) = O_p \left(\left(\frac{\log |M|}{n} \right)^{1/2} \right).
$$

Step 4: Bound on $\sup_{S \in M} A_n(S)$. Note that

$$
\begin{aligned}
2\Big(\rho_\tau(Y - Z'\theta_1) - \rho_\tau(Y - Z'\theta_2) \Big) &= Z'(\theta_1 - \theta_2) 1\{Y \geq Z'\theta_1\} 1\{Y \geq Z'\theta_2\} \\
&\quad - Z'(\theta_1 - \theta_2) 1\{Y < Z'\theta_1\} 1\{Y < Z'\theta_2\} \\
&\quad + (2Y - Z'\theta_1 - Z'\theta_2) 1\{Y \geq Z'\theta_1\} 1\{Y < Z'\theta_2\} \\
&\quad - (2Y - Z'\theta_1 - Z'\theta_2) 1\{Y < Z'\theta_1\} 1\{Y \geq Z'\theta_2\} \\
&\quad + (2\tau - 1) Z'(\theta_1 - \theta_2).
\end{aligned} \tag{4.46}
$$

Let $\mathscr{D}^0_n$ be an independent copy of $\mathscr{D}_n$ and define

$$
\begin{aligned}
W_n = \mathrm{E}_{\mathscr{D}^0_n} \Bigg[\max_{S \in M} \sup_{\|\delta_S\|_2 \leq r_n} \frac{1}{n^2} \sum_{i=1}^{n} \Big(&\rho_\tau(e^{\tau}_{ni,S} - Z'_{ni,S}\delta_S) - \rho_\tau(e^{\tau}_{ni,S}) \\
&- \rho_\tau(e^{0\tau}_{ni,S} - Z^{0'}_{ni,S}\delta_S) - \rho_\tau(e^{0\tau}_{ni,S}) \Big)^2 | \mathscr{D}_n \Bigg].
\end{aligned}
$$

By expansion (4.46),

$$
\begin{aligned}
W_n \leq &\max_{S \in M} \sup_{\|\delta_S\|_2 \leq r_n} \frac{2}{n^2} \sum_{i=1}^{n} \Big(\rho_\tau(e^{\tau}_{ni,S} - Z'_{ni,S}\delta_S) - \rho_\tau(e^{\tau}_{ni,S}) \Big)^2 \\
&+ \mathrm{E} \left[\max_{S \in M} \sup_{\|\delta_S\|_2 \leq r_n} \frac{2}{n^2} \sum_{i=1}^{n} \Big(\rho_\tau(e^{0\tau}_{ni,S} - Z^{0'}_{ni,S}\delta_S) - \rho_\tau(e^{0\tau}_{ni,S}) \Big)^2 \right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \max_{S\in M} \sup_{\|\delta_S\|_2\leq r_n} \frac{9}{n^2}\sum_{i=1}^{n}\left(\left(Z'_{ni,S}\delta_S\right)^2 - \mathbb{E}_{\mathscr{D}_n}\left[\left(Z'_{ni,S}\delta_S\right)^2\right]\right) \\
&\quad + \mathbb{E}_{\mathscr{D}_n^0}\left[\max_{S\in M} \sup_{\|\delta_S\|_2\leq r_n} \frac{9}{n^2}\sum_{i=1}^{n}\left(\left(Z^{0'}_{ni,S}\delta_S\right)^2 - \mathbb{E}_{\mathscr{D}_n}\left[\left(Z^{0'}_{ni,S}\delta_S\right)^2\right]\right)\right] \\
&\quad + \max_{S\in M} \sup_{\|\delta_S\|_2\leq r_n} \mathbb{E}_{\mathscr{D}_n}\left[\frac{18}{n^2}\sum_{i=1}^{n}\left(Z'_{ni,S}\delta_S\right)^2\right] \quad a.s. \qquad (4.47)
\end{aligned}
$$

As in Step 2 of the proof of Lemma 4.5, we conclude that there exists an absolute constant $c_2 > 0$ such that

$$
\begin{aligned}
&\max_{S\in M} \sup_{\|\delta_S\|_2\leq r_n} \frac{9}{n^2}\sum_{i=1}^{n}\left(\left(Z'_{ni,S}\delta_S\right)^2 - \mathbb{E}_{\mathscr{D}_n}\left[\left(Z'_{ni,S}\delta_S\right)^2\right]\right) \\
&\qquad \leq c_2\left(\frac{m\log|M|\log\log n}{n}\right)^{1/2}\frac{r_n^2}{n} \quad a.s.
\end{aligned}
$$

Thus, there exists an absolute constant $c_3 > 0$ such that

$$
W_n \leq c_3\frac{r_n^2}{n} \quad a.s. \qquad (4.48)
$$

By the second statement of Lemma 4.1 for any $t > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{S\in M}\frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right)\right. \\
&\qquad \left. -\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right] \geq 2W_n^{1/2}t\right) \\
&\leq \sum_{S\in M}\mathbb{P}\left(-\sup_{\|\delta_S\|_2\leq r_n} -\frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right)\right. \\
&\qquad \left. + \mathbb{E}_{\mathscr{D}_n}\left[\sup_{\|\delta_S\|_2\leq r_n} -\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right] \geq 2W_n^{1/2}t\right) \\
&\leq \sum_{S\in M}\mathbb{P}\left(\sup_{\|\delta_S\|_2\leq r_n} -\frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right)\right. \\
&\qquad \left. \leq \mathbb{E}_{\mathscr{D}_n}\left[\sup_{\|\delta_S\|_2\leq r_n} -\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_{ni} - Z'_{ni,S}\hat{\theta}^\tau_{n,S}) - \rho_\tau(Y_{ni} - Z'_{ni,S}\theta^\tau_{n,S})\right] - 2W_n^{1/2}t\right) \\
&\leq 4|M|ee^{-t^2/2}. \qquad (4.49)
\end{aligned}
$$

Analogously, we derive a bound on the probability for the lower tail via the first statement

of Lemma 4.1,

$$\mathbb{P}\left(\sup_{S\in M}\frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau(Y_{ni}-Z'_{ni,S}\hat{\theta}^\tau_{n,S})-\rho_\tau(Y_{ni}-Z'_{ni,S}\theta^\tau_{n,S})\right)\right.$$
$$\left.-\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_{ni}-Z'_{ni,S}\hat{\theta}^\tau_{n,S})-\rho_\tau(Y_{ni}-Z'_{ni,S}\theta^\tau_{n,S})\right]\leq -2W_n^{1/2}t\right)$$
$$\leq 4|M|ee^{-t^2/2}. \tag{4.50}$$

Thus, combining eq. (4.48)- (4.50) and setting $t=t'(\log|M|)^{1/2}$, there exists $N_0>0$ such that for all $n>N_0$,

$$\mathrm{P}\left(\sup_{S\in M}A_n(S)\geq c_4 r_n\frac{t^{1/2}}{n^{1/2}}\right)\leq 4ee^{-t'/2},$$

where $c_4>0$ is an absolute constant. Hence, as in Step 2 of the proof of Lemma 4.8 integrating this tail bound out yields for all $n>N_0$, $\mathbb{E}_{\mathscr{D}_n}[\sup_{S\in M}A_n(S)]\leq c_5\frac{r_n}{n^{1/2}}$, where $c_5>0$ is an absolute constant, and

$$\mathbb{E}_{\mathscr{D}_n}\left[\sup_{S\in M}A_n(S)\right]=O_p\left(\frac{r_n}{n^{1/2}}\right).$$

Step 5: Conclusion. Combining above bounds on $\sup_{S\in M}A_n(S)$ through $\sup_{S\in M}D_n(S)$ we have

$$\sup_{S\in M}\left|\widehat{PR}^\tau_{n,h}(S)-\mathrm{PR}^\tau_n(S)\right|$$
$$=O_p\left(\left(\frac{\log|M|}{n}\right)^{1/2}+\frac{r_n}{n^{1/2}}+\lambda_n^{3/2}r_{n,3}^{5/2}+\frac{m\,h^\alpha}{\lambda_n^2 n}+\frac{m\,r_{n,3}}{h\lambda_n n}+\frac{m\,r_{n,3}^\alpha}{\lambda_n^2 n}\right).$$

4.8 Supplementary Materials

In this section we provide further numerical evidence for the DGPs 2–4 from Section 4.5. The interpretations given in Section 4.5 apply to below plots as well. Qualitatively, the conclusion for DGPs 2–4 are the same for DGP1. However, the variations are sometimes higher in the results because DGP 2–4 involve more complex settings.

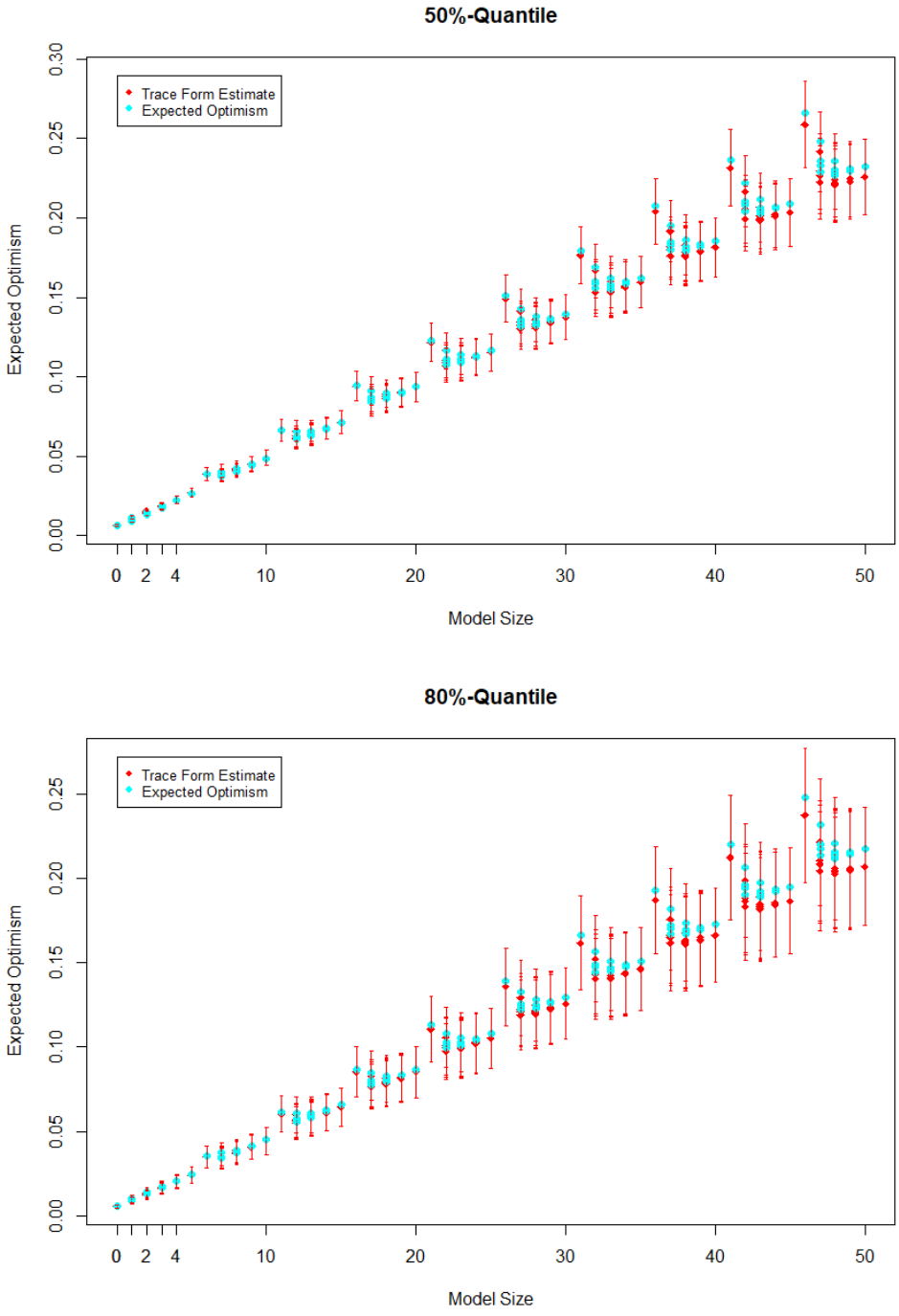

Figure 4.3: DGP2 trace form versus model size. Red: estimates of the trace form and standard errors. Blue: expected optimism. Top: DGP2 with $\tau = 0.5$. Bottom: DGP2 with $\tau = 0.8$.

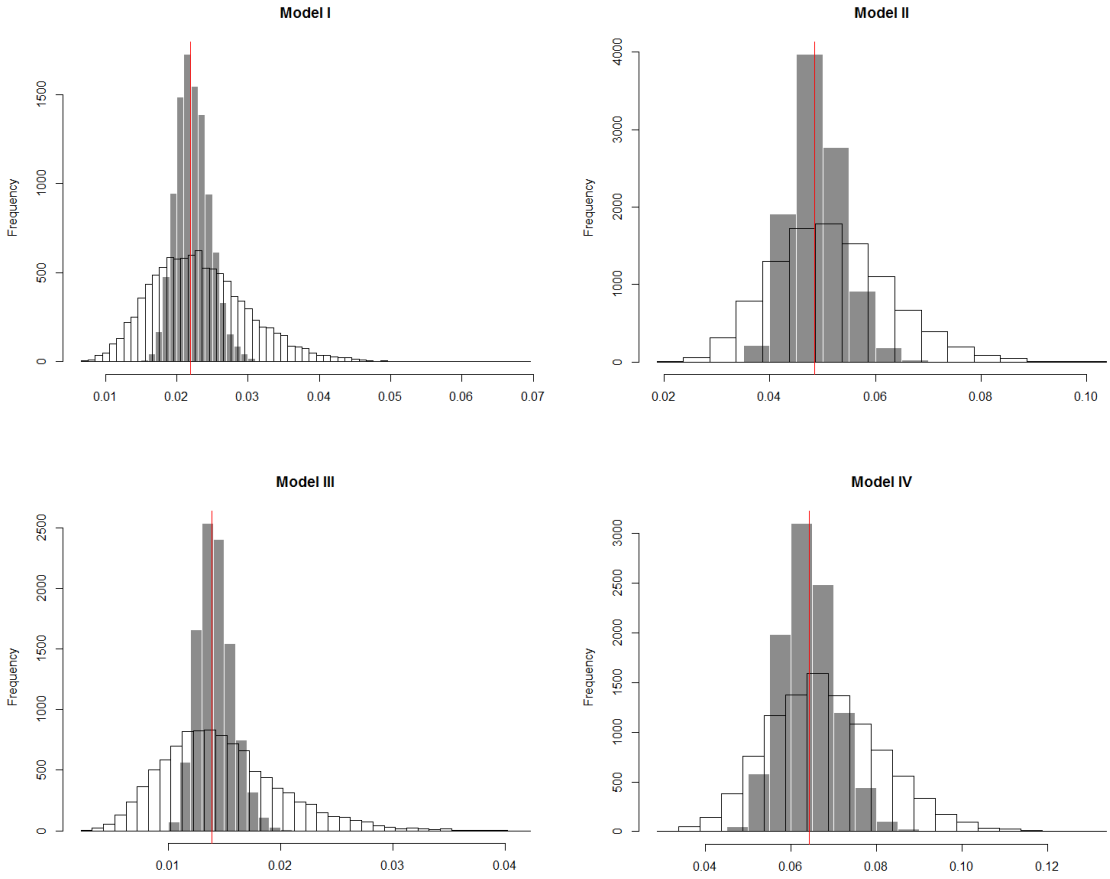

Figure 4.4: Expected optimism and trace form estimate (DGP2). Histograms of the 10-fold CV estimate of the expected optimism and the trace form estimate for DGP2 and $\tau = 0.5$. Red line: expected optimism. White histogram: 10-fold CV. Gray histogram: trace form estimate. Model I: correct model (with predictors 1 to 4), Model II: an over-fitted model (with predictors 1 to 10), Model III: an under-fitted model (with predictors 1 to 2) and Model IV that comprises the relevant predictors 1 and 2 and the irrelevant predictors 5 to 15.

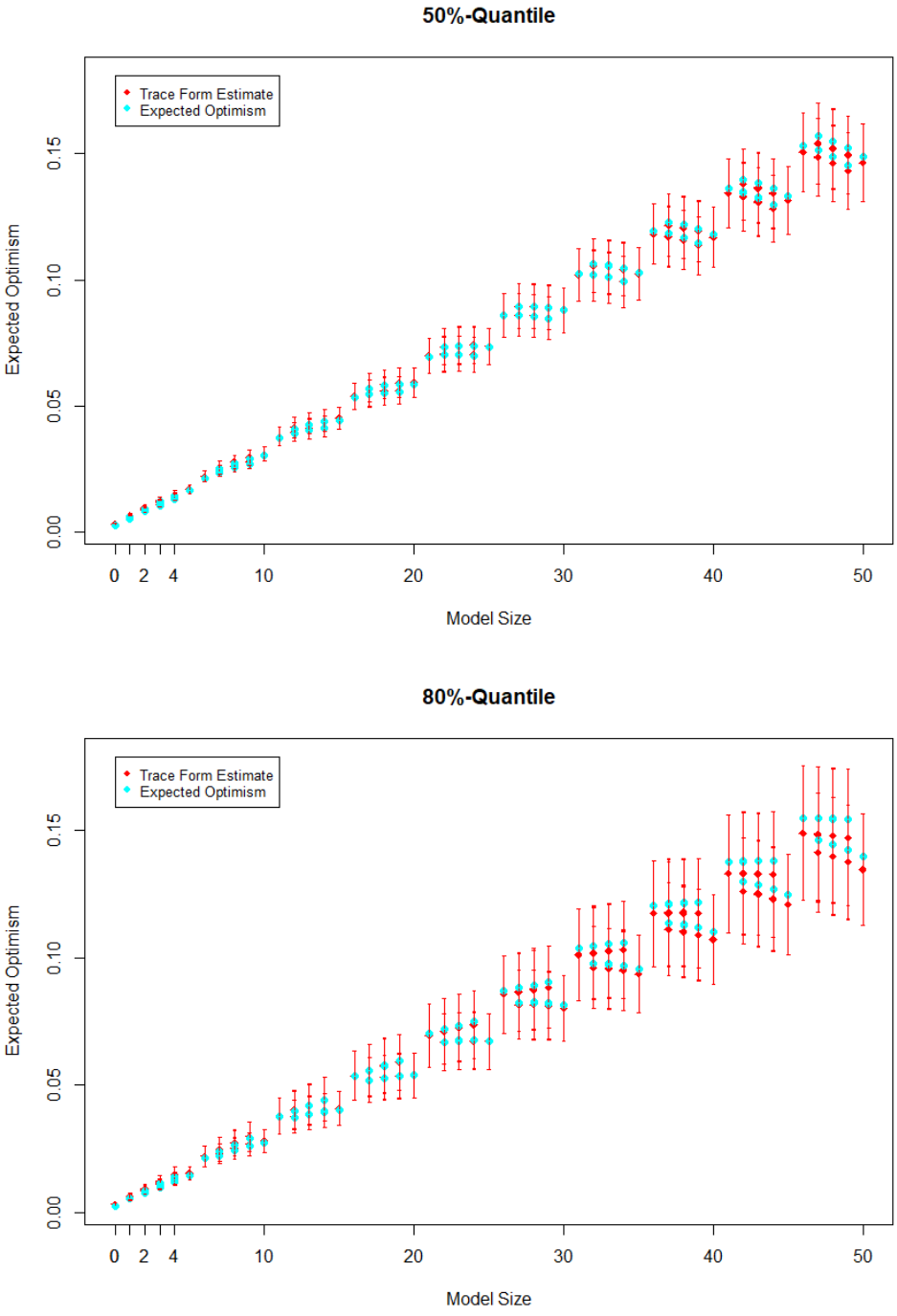

Figure 4.5: DGP3 trace form versus model size. Red: estimates of the trace form and standard errors. Blue: expected optimism. Top: DGP3 with $\tau = 0.5$. Bottom: DGP3 with $\tau = 0.8$.

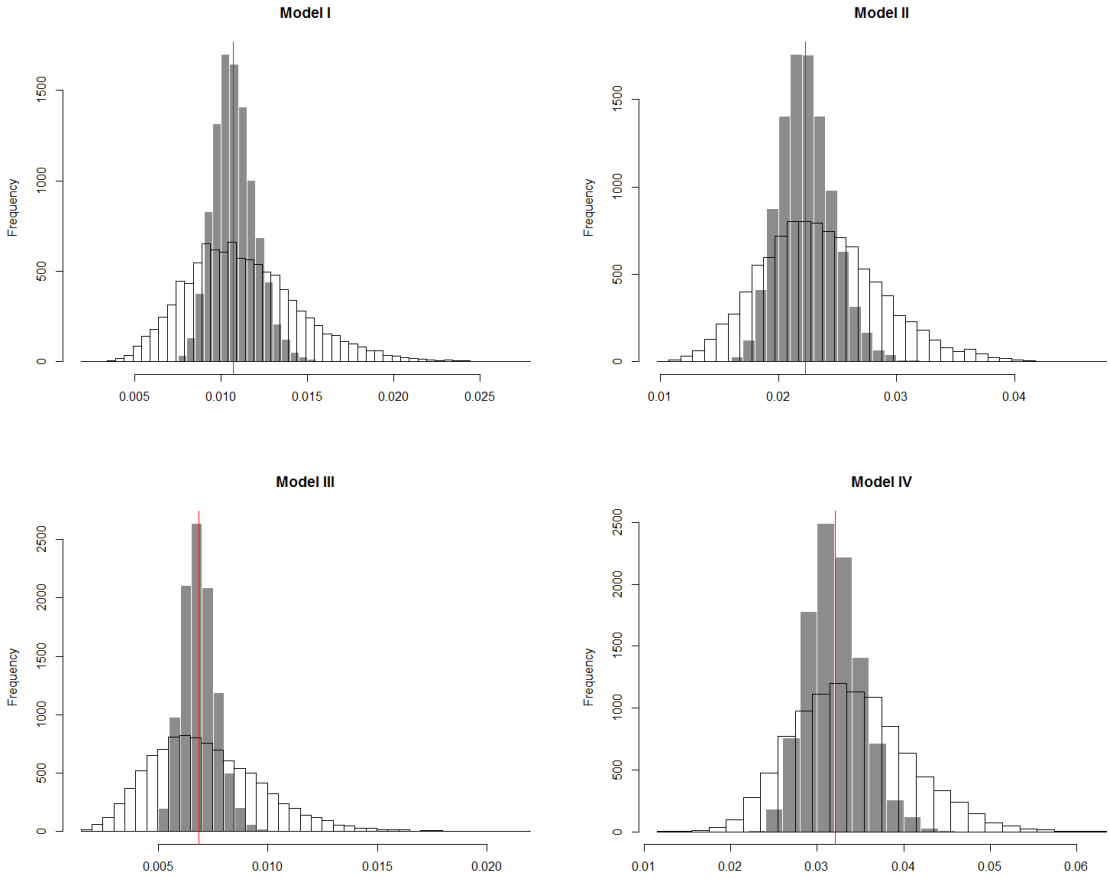

Figure 4.6: Expected optimism and trace form estimate (DGP3). Histograms of the 10-fold CV estimate of the expected optimism and the trace form estimate for DGP3 and $\tau = 0.5$. Red line: expected optimism. White histogram: 10-fold CV. Gray histogram: trace form estimate. Model I: correct model (with predictors 1 to 4), Model II: an over-fitted model (with predictors 1 to 10), Model III: an under-fitted model (with predictors 1 to 2) and Model IV that comprises the relevant predictors 1 and 2 and the irrelevant predictors 5 to 15.

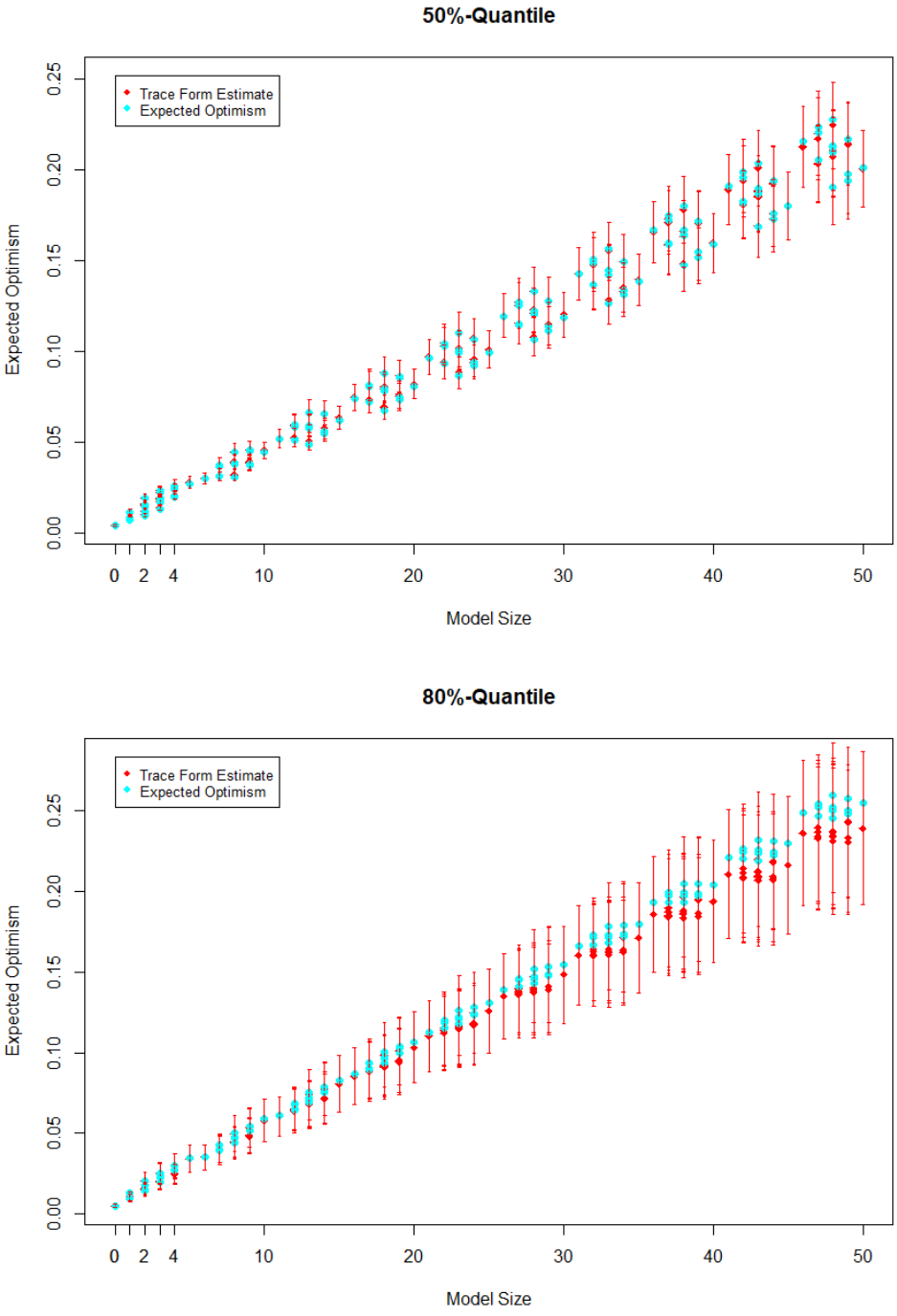

Figure 4.7: DGP4 trace form versus model size. Red: estimates of the trace form and standard errors. Blue: expected optimism. Top: DGP4 with $\tau = 0.5$. Bottom: DGP4 with $\tau = 0.8$.

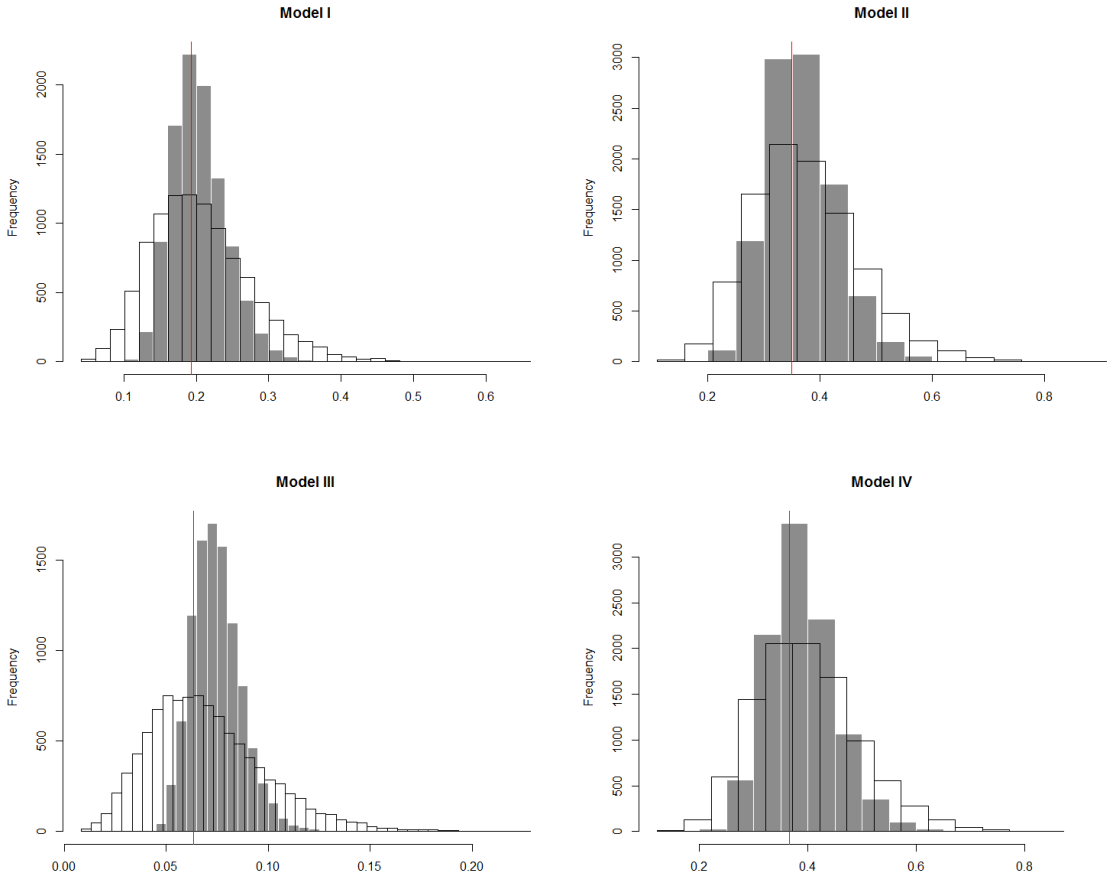

Figure 4.8: Expected optimism and trace form estimate (DGP4). Histograms of the 10-fold CV estimate of the expected optimism and the trace form estimate for DGP4 and $\tau = 0.5$. Red line: expected optimism. White histogram: 10-fold CV. Gray histogram: trace form estimate. Model I: correct model (with predictors 1 to 4), Model II: an over-fitted model (with predictors 1 to 10), Model III: an under-fitted model (with predictors 1 to 2) and Model IV that comprises the relevant predictors 1 and 2 and the irrelevant predictors 5 to 15.

BIBLIOGRAPHY

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 610–624, New York. Springer-Verlag.

Akaike, H. (1998). Statistical predictor identification. In Akaike, H., Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, Perspectives in Statistics, pages 137–151, New York. Springer-Verlag.

Aliprantis, C. and Border, K. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Studies in Economic Theory. Springer-Verlag, Heidelberg.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

Andrews, D. W. K. (1994). Empirical process methods in econometrics. In Z., Engle, R., Intriligator, M., and McFadden, D., editors, *Handbook of Econometrics, Volume IV*, pages 2247 – 2294, New York. North-Holland.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563.

Arcones, M. A. (1996). The Bahadur-Kiefer Representation of L_p Regression Estimators. *Econometric Theory*, 12(2):257–283.

Arcones, M. A. (1998). Second order representations of the least absolute deviation regression estimator. *Annals of the Institute of Statistical Mathematics*, 50(1):87–117.

Babu, G. J. (1989). Strong representations for lad estimators in linear models. *Probability Theory and Related Fields*, 83(4):547–558.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580.

Bai, Z. and Wu, Y. (1994). Limiting behavior of m-estimators of regression coefficients in high dimensional linear models i. scale dependent case. *Journal of Multivariate Analysis*, 51(2):211 – 239.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bassett, Jr., G. W., Koenker, R., and Kordas, G. (2004). Pessimistic portfolio allocation and choquet expected utility. *Journal of Financial Econometrics*, 2(4):477–492.

Belloni, A. and Chernozhukov, V. (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2017). Conditional quantile processes based on series or many regressors. *arxiv preprint,* https://arxiv.org/pdf/1105.6154.pdf.

Bercu, B., Gassiat, E., and Rio, E. (2002). Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *Ann. Probab.*, 30(4):1576–1604.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1):62 – 91.

Bradic, J. and Kolar, M. (2017). Uniform Inference for High-dimensional Quantile Regression: Linear Functionals and Regression Rank Scores. *ArXiv e-prints*.

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

Burman, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika*, 82(4):877–886.

Cahuich, L. D. and Hernández-Hernández, D. (2013). Quantile portfolio optimization under risk measure constraints. *Applied Mathematics & Optimization*, 68(2):157–179.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Carroll, R. J. (1978). On almost sure expansions for m-estimates. *The Annals of Statistics*, 6(2):314–318.

Chao, S.-K., Volgushev, S., and Cheng, G. (2017). Quantile processes for semi and nonparametric regression. *Electronic Journal of Statistics*, 11(2):3272–3331.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics*, 19(2):760–777.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, 26(1):271–292.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation [with comments, rejoinder]. *Journal of the American Statistical Association*, 99(467):619–642.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Engle, R. o. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating value-at-risk models via quantile regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Giessing, A. and He, X. (2018). On the predictive risk in misspecified quantile regression. *arxiv preprint,* https://arxiv.org/pdf/1802.00555.pdf.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2005). *Robust Statistics: The Approach Based on Influence Functions*. Probability and Statistics Series. John Wiley & Sons, New York.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, New York.

He, X. and Shao, Q.-M. (1996). A general bahadur representation of m -estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608–2630.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135.

He, X. D. and Zhou, X. Y. (2011). Portfolio choice via quantiles. *Mathematical Finance*, 21(2):203–231.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kato, K. (2011). Group Lass for high dimensional sparse quantile regression models. *ArXiv e-prints*.

Kiefer, J. (1967). On bahadur's representation of sample quantiles. *The Annals of Mathematical Statistics*, 38(5):1323–1342.

Kim, T.-H. and White, H. (2003). Estimation, inference, and specification testing for possibly misspecified quantile regression. In Fomby, T. B., Hill, R. C., Jeliazkov, I., Escanciano, J. C., and Hillebrand, E., editors, *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, volume 17 of *Advances in Econometrics*, pages 107–132. Emerald Group Publishing Limited.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, Cambridge.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265 – 283.

Koenker, R. and Portnoy, S. (1987). L-estimation for linear models. *Journal of the American Statistical Association*, 82(399):851–857.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Lecture Notes in Mathematics. Springer Verlag, New York.

Kou, S. C. and Efron, B. (2002). Smoothers and the C_p, generalized maximum likelihood, and extended exponential criteria. *Journal of the American Statistical Association*, 97(459):766–782.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. *arxiv preprint,* https://arxiv.org/pdf/1802.05801.pdf.

Ledoux, M. and Talagrand, M. (1989). Comparison theorems, random geometry and some limit theorems for empirical processes. *The Annals of Probability*, 17(2):596–631.

Ledoux, M. and Talagrand, M. (1996). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory*, 19(1):1–31.

Lee, Y.-Y. (2016). Interpretation and semiparametric efficiency in quantile regression under misspecification. *Econometrics*, 4(1).

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, 17(1):382–400.

Massart, P. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer-Verlag, Berlin.

Niemiro, W. (1992). Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, 20(3):1514–1533.

Noh, H., El Ghouch, A., and Van Keilegom, I. (2013). Assessing model adequacy in possibly misspecified quantile regression. *Computational Statistics & Data Analysis*, 57(1):558 – 569.

Panchenko, D. (2003). Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability*, 31(4):2068–2081.

Pollard, D. (1995). Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics*, 13(4):1403–1417.

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. *The Annals of Statistics*, 25(1):414–434.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, 32(1):143–155.

Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science*, 153:12–18.

Talagrand, M. (1996a). Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103.

Talagrand, M. (1996b). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563.

Tao, T. and Vu, V. (2006). *Additive Combinatorics*. Cambridge studies in advanced mathematics. Cambridge University Press.

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):529–546.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. and Wellner, J. A. (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Vershynin, R. (2012a). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

Vershynin, R. (2012b). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyinok, G., editors, *Compressed Sensing, Theory and Applications*, pages 210–268, Cambridge. Cambridge University Press.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.

Welsh, A. H. (1989). On m-processes and m-estimation. *The Annals of Statistics*, 17(1):337–361.

Wu, W. B. (2007). M -estimation of linear models with dependent errors. *The Annals of Statistics*, 35(2):495–521.

Xiao, Z., Guo, H., and Lam, M. S. (2015). Quantile regression and value at risk. In Lee, C.-F. and Lee, J. C., editors, *Handbook of Financial Econometrics and Statistics*, pages 1143–1167, New York. Springer-Verlag.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131.

Zhang, C.-H. and Zhang, S. S. (2013). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.