Adaptive quantile regression

S.A. van de Geer[a]

[a]Mathematical Institute, University of Leiden,
P.O. Box 9512, 2300 RA Leiden, The Netherlands

In this paper, the estimation of a regression quantile function is studied. We consider a linear regression model with dimension m, where m may be as large as the number of observations n. As penalty on the quantile regression estimator, the sum of the absolute value if its coefficients is used. We show that the estimator adapts to the unknown smoothness of the underlying quantile regression function, as well as to unknown identifiablity properties.

1. INTRODUCTION

Consider $n \geq 2$ independent observations on a response variable $Y_i \in \mathbf{R}$ and a covariable $x_i \in \mathcal{X}$, $i = 1, \ldots, n$. We study the estimation of the β-quantile $g_0(x_i)$ of the distribution of Y_i given the covariable x_i, for $i = 1, \ldots, n$. Here, $0 < \beta < 1$ is a given number, chosen by the statistician. The quantile regression problem was introduced by [6]. There, $g_0(x)$ is modelled as a linear function of a fixed (small) number of parameters. In this paper, we propose a linear model with many parameters, say m, where m may be as large as the number of observations n. Moreover, we do not require strong assumptions on the distribution of the observations, in particular on the degree of identifiabililty of the unknown regression function g_0. Our aim is to construct an estimator that adapts to the amount of parameters needed, as well as to identifiability properties of g_0.

Let $\gamma = \gamma_\beta$ be the quantile regression loss function

$$\gamma(y) = \beta |y| \mathrm{l}\{y < 0\} + (1 - \beta)|y| \mathrm{l}\{y \geq 0\}. \tag{1}$$

Consider the empirical loss function

$$\Gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i - g(x_i)), \tag{2}$$

and let

$$\Gamma(g) = \mathbf{E}\Gamma_n(g) \tag{3}$$

be the theoretical loss function. Here, and throughout, we regard the covariables $x_1, \ldots, x_n$ as fixed (i.e., we work conditionally on the observed values of the covariables). It is then easily seen that

$$g_0 = \arg\min_g \Gamma(g), \tag{4}$$

where the minimum is taken over the class of all functions $g : \mathcal{X} \to \mathbf{R}$. More specifically, let for $i = 1, \ldots, n$,

$$F_i(y) = \mathbf{P}(Y_i \leq y), \ y \in \mathbf{R}, \tag{5}$$

be the distribution function of Y_i. Then, assuming the inverse of F_i at β exists, we have

$$g_0(x_i) = F_i^{-1}(\beta) = \arg\min_{c \in \mathbf{R}} \mathbf{E}\gamma(Y_i - c), \ i = 1, \ldots, n. \tag{6}$$

Now, the idea is to estimate g_0 by using the empirical counterpart, i.e., the minimizer of Γ_n. We then need to somehow control complexity, since the overall minimizer $\arg\min_g \Gamma_n(g)$ just reproduces the data $(Y_i)_{i=1}^n$. To avoid overfitting, one can add a penalty $\mathrm{pen}(g)$ to the empirical loss function. The penalized quantile regression estimator is

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \{\Gamma_n(g) + \mathrm{pen}(g)\}. \tag{7}$$

Here, $\mathcal{G}$ is an a priori model class, which may be the class of *all* functions $g : \mathcal{X} \to \mathbf{R}$.

We now come to the description of the linear model, and the penalty. Let for $j = 1, \ldots, m$, $\psi_j : \mathcal{X} \to \mathbf{R}$ be given functions. We assume that $\psi_1, \ldots, \psi_m$ are linearly independent in $L_2(Q_n)$, with $Q_n = \sum_{i=1}^n \delta_{x_i}/n$ the empirical measure of the covariables. (Thus in particular, we assume that $m \leq n$). Take $\mathcal{G}$ as (a subset of) the linear functions

$$g = g_\alpha = \sum_{j=1}^m \alpha_j \psi_j, \ \alpha \in \mathbf{R}^d. \tag{8}$$

We let

$$\Sigma_n = \int \psi \psi^T dQ_n, \tag{9}$$

where $\psi = (\psi_1, \ldots, \psi_m)^T : \mathcal{X} \to \mathbf{R}^m$. Denote the smallest eigenvalue of Σ_n as $\lambda^2_{\min}$, and its largest eigenvalue as $\lambda^2_{\max}$. Invoke the normalization

$$\lambda_{\max} = 1. \tag{10}$$

and choose the penalty

$$\mathrm{pen}(g_\alpha) = \lambda_n \sum_{j=1}^m |\alpha_j|, \tag{11}$$

with λ_n a regularization parameter, to be specified (see Theorem 1). We call (11) an L_1-penalty. It is closely related to soft-thresholding ([3]), and to the LASSO ([16,5]). Note that with this penalty, the computation of the penalized quantile estimator is quite feasible (using e.g. interior point methods, see also [14]).

In [7,8], total variation type penalties on the function g or on (first or higher order) derivates of g are used. Asymptotic theory for such estimators can be found in [13]. These estimators have local adaptive properties, but they may not be globally adaptive. Aiming at a globally adaptive procedure (see Section 2 for more details), one may propose to use model selection among a collection of linear (high-dimensional) models, and use a penalty proportional to model dimension (AIC ([1]), BIC ([15]), etc.). However, it is not at all

clear whether such an approach is robust against violations of regularity conditions on the distribution of Y_i, $i = 1, \ldots, n$ (for instance, the assumption of existence of Lebesgue densities which are strictly positive near the β-quantile (assumption (i) in Theorem 4.2 of [6])). Such regularity conditions ensure quadratic behavior of the loss function near its minimum. In our setup, the behavior near the minimum may be unknown.

Below, a general condition (13) on the identifiability of g_0 is presented. Here, and in the sequel, we use the notation

$$\|g\|^2 = \sum_{i=1}^{n} g^2(x_i), \ g: \ \mathcal{X} \to \mathbf{R}. \tag{12}$$

Thus $\|\cdot\|$ is the $L_2(Q_n)$-norm.

Identifiability condition. There exists a strictly increasing function $J: \ [0, \infty) \to [0, \infty)$ such that for all $g \in \mathcal{G}$,

$$\Gamma(g) - \Gamma(g_0) \geq \|g - g_0\| J(\|g - g_0\|). \tag{13}$$

In the regular case, J is linear, so that Γ is quadratic. If e.g., $J(\xi)$ decreases faster than linear as $\xi \to 0$, the regression function g_0 is harder to identify. We call J the *margin* function: it describes the sensitivity of $\Gamma(\cdot)$ to deviations from its minimum. Condition (13) will be referred to as the *margin condition*.

An important point is that the function J will generally be unknown, and that it will also be hard to estimate it with sufficient accuracy. Thus, we need a procedure which adapts to all possible J. Note also that the amount of identifiability may depend on possible a priori model assumptions, i.e., on the model class $\mathcal{G}$ for the regression function. We moreover make the rather trivial observation that one may replace $\mathcal{G}$ by an appropriate smaller class, but that one then has to take into account the (hopefully negligible) probability that $\hat{g}_n$ is not in the smaller class.

Example 1. Write $c_i = g_0(x_i)$ (so $F_i(c_i) = \beta$), and suppose that there exists constants, $0 < \epsilon < 1$, $\sigma > 0$ and $\rho > 0$ such that

$$|F_i(c) - F_i(c_i)| \geq |c - c_i|^\rho / \sigma^\rho, \tag{14}$$

for all $|c - c_i| \leq \epsilon$, $i = 1, \ldots, n$. Then, for $|c - c_i| \leq \epsilon$

$$\begin{aligned} \mathbf{E}\gamma(Y_i - c) - \mathbf{E}\gamma(Y_i - c_i) &= (c - c_i)(F_i(c) - F_i(c_i)) \\ &\geq |c - c_i|^{1+\rho} / \sigma^\rho, \ i = 1, \ldots, n. \end{aligned} \tag{15}$$

So then, when $|g(x_i) - g_0(x_i)| \leq \epsilon$, for all $i = 1, \ldots, n$, we have

$$\Gamma(g) - \Gamma(g_0) \geq \|g - g_0\|^{1+\tilde{\rho}} / \sigma^\rho, \tag{16}$$

where $\tilde{\rho} = \max(\rho, 1)$. Thus, then the margin condition (13) is met on the set

$$\mathcal{G}_0 = \{g: \ \max_{i=1,\ldots,n} |g(x_i) - g_0(x_i)| \leq \epsilon\}, \tag{17}$$

with J the function $J(\xi) = \xi^{1+\tilde{\rho}} / \sigma^\rho$, $\xi \geq 0$.

We will show in Theorem 1 of Section 3 that the quantile estimator with L_1-penalty adapts to the smoothness of g_0 as well as to the margin function J. In Section 2, we will explain in some detail what we mean by adaptation. Theorem 2 of Section 3 shows moreover that results can be extended to hold uniformly in β. In Section 4, we take a brief look at the concept of smoothness, considered in general terms of approximation theory. Section 5 presents the proof of Theorem 1 and Theorem 2.

2. THE ORACLE

Let the "dimension" of $g = g_\alpha$ be

$$d_g = \#\{\alpha_j \neq 0\}. \tag{18}$$

An "oracle" g_c^* is

$$g_c^* = \arg\min_{g \in \mathcal{G}} \tau_c^2(g|g_0). \tag{19}$$

where

$$\tau_c^2(g|g_0) = \left\{\Gamma(g) - \Gamma(g_0) + 4c(\lambda_n/\lambda_{\min})\sqrt{d_g}J^{-1}(4c(\lambda_n/\lambda_{\min})\sqrt{d_g})\right\}. \tag{20}$$

Here λ_n is the regularization parameter, and $c > 0$ is a constant. Furthermore, we let $\tau^2(g|g_0) = \tau_1^2(g|g_0)$, and we let $g^* = g_1^*$ be the oracle with constant $c = 1$. Its dimension is denoted by $d^* = d_{g^*}$. Note that g_c^* depends on the regularization parameter λ_n. We regard this parameter as to be chosen by the statistician, and hence it may not depend on unknown quantities. In fact we will choose it as

$$\lambda_n = u\sqrt{\frac{\log n}{n}}, \tag{21}$$

where u is equal, or larger than, some universal constant (see Theorem 1). The constant c however is allowed to depend on unknown quantities, i.e., on the distribution of Y_i, $i = 1, \ldots, n$. The constant 4 in our definition of τ_c^2 is only there because it comes out of our rough calculations in the proof of Theorem 1. It has no intrinsic meaning.

Now, we come to the question why we consider g^* (more generally g_c^*) as an oracle. The reason is that it represents the best trade-off, over all linear submodels, between "bias" and "variance", where we take the terminology in a very loose sense ("bias" corresponding to "approximation error" and "variance" to "estimation error"). Of course, g^* is non-random, and in fact, the "estimation error" comes rather from the estimator one could use if the set $\mathcal{J}^*$ (with cardinality d^*) of the non-zero coefficients of g^* were known. Let us give the heuristics here, without going into details.

Consider a d-dimensional model, with only the d coefficients in the index set $\mathcal{J}$, with cardinality d, possibly non-zero. Write

$$\mathcal{G}_d = \{g = \sum_{j \in \mathcal{J}} \alpha_j \psi_j\} \cap \mathcal{G} \tag{22}$$

for this model class and let $\hat{g}_{n,d}$ be the quantile estimator over this class, that is

$$\hat{g}_{n,d} = \arg\min_{g \in \mathcal{G}_d} \Gamma_n(g). \tag{23}$$

Let

$$g_{*,d} = \arg\min_{g \in \mathcal{G}_d} \Gamma(g) \tag{24}$$

be the best approximation of g_0 within the class $\mathcal{G}_d$. Then rewriting

$$\Gamma_n(\hat{g}_{n,d}) \leq \Gamma_n(g_{*,d}) \tag{25}$$

gives

$$\Gamma(\hat{g}_{n,d}) - \Gamma(g_0) \leq -\hat{\nu}_{n,d} + \Gamma(g_{*,d}) - \Gamma(g_0), \tag{26}$$

where

$$\hat{\nu}_{n,d} = [\Gamma_n(\hat{g}_{n,d}) - \Gamma_n(g_{*,d})] - [\Gamma(\hat{g}_{n,d}) - \Gamma(g_{*,d})] \tag{27}$$

is the "random part" of the problem. Now, empirical process theory (see e.g., [19,17], and their references) gives that $\hat{\nu}_{n,d}$ behaves like $\sqrt{d/n}\|\hat{g}_{n,d} - g_{*,d}\|$, i.e. except on a set $\bar{\mathbf{A}}$, with small probability $\mathbf{P}(\bar{\mathbf{A}})$,

$$|\hat{\nu}_{n,d}| \leq C\sqrt{d/n}\|\hat{g}_{n,d} - g_{*,d}\|, \tag{28}$$

where C is a constant depending on the distribution of the observations. After some straightforward calculations, we see that except on the set $\bar{\mathbf{A}}$,

$$\Gamma(\hat{g}_{n,d}) - \Gamma(g_0) \leq 2\tilde{\tau}^2(g_{*,d}|g_0) \tag{29}$$

where

$$\tilde{\tau}^2(g|g_0) = \Gamma(g) - \Gamma(g_0) + C\sqrt{\frac{d}{n}}J^{-1}(2C\sqrt{\frac{d}{n}}) \tag{30}$$

(Similar arguments are used in the proof of Theorem 1, where more details are presented.) So the best result (up to constants) are obtained for the class $\mathcal{J}_*$ with cardinality d_* corresponding to the non-zero coefficients of

$$g_* = \arg\min \tilde{\tau}^2(g|g_0). \tag{31}$$

Thus we see that apart from constants and the $\sqrt{\log n}$ term, the function g^* represents the ideal approximation of g_0 by a linear model with smallest possible dimension, taking into account the "bias" term $\Gamma(g) - \Gamma(g_0)$ as well as the "variance" term $C\sqrt{d_g/n}J^{-1}(2C\sqrt{d_g/n})$.

3. ADAPTATION

This section provides our main results. We give explicit constants, so that the dependencies on n, g_0 and other quantities, is evident. We have however not attempted to optimize these constants.

Theorem 1 below formulates an oracle inequality for the L_1-penalized quantile estimator

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \{\Gamma_n(g) + \text{pen}(g_\alpha)\}. \tag{32}$$

We took the terminology *oracle inequality* from [2], where a Gaussian sequence space model is considered (and in that context stronger results are obtained). We recall that

$$\mathcal{G} \subseteq \{g_\alpha = \sum_{j=1}^m \alpha_j \psi_j : \ \alpha \in \mathbf{R}^m\}, \tag{33}$$

and that

$$\Gamma_n(g) = \frac{1}{n}\sum_{i=1}^n \gamma(Y_i - g(x_i)) \tag{34}$$

is the quantile loss function, and

$$\text{pen}(g_\alpha) = \lambda_n \sum_{j=1}^m |\alpha_j| \tag{35}$$

is the L_1-penalty.

Theorem 1. *Assume that $\mathcal{G}$ is a convex set and that the margin condition (13) holds for each $g \in \mathcal{G}$, and some strictly increasing function J. Take $\lambda_n = (12+u)2\sqrt{\log n/n}$, with $u \geq 6$. Then for $n \geq N$, where N is such that $\lambda_n/n + \tau^2(g^*|g_0) \leq 1/8$, and for any $0 < \delta \leq 1$,*

$$\mathbf{P}\left(\Gamma(\hat{g}_n) - \Gamma(g_0) \geq (1+\delta)^2\left[\frac{\lambda_n}{n} + \tau_n^*(\delta)^2\right]\right) \leq 64R^2 \exp[-\frac{u^2 \log n}{1024R^2}], \tag{36}$$

where

$$\tau_n^*(\delta)^2 = \left[\Gamma(g^*) - \Gamma(g_0) + \frac{2}{\delta^2}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}J^{-1}(\frac{4}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*})\right], \tag{37}$$

and where $R = 1 + \|g^* - g_0\|$.

The estimator can be computed for a whole path of values of λ_n and β simultaneously (see [8] for a related problem). Note that the theorem allows one to immediatly obtain simultaneous inequalities for all β in (say) the finite grid $\beta \in \{k/n, (k+1)/n, \ldots, (n-k)/n\}$, $k = n\lfloor t \rfloor$, $t \in (0,1)$. It is in fact possible to extend Theorem 1 to hold uniformly for all β in some domain of interest, say $\beta \in \mathcal{B} \subset (0,1)$. This will be shown in Theorem 2.

Note that almost everything we have defined so far depends on β. For example, $g_0 = g_{0\beta}$, $\Gamma_n = \Gamma_{n,\beta}$ and $\Gamma = \Gamma_\beta$. In the margin condition (13), $J = J_\beta$ will generally depend on β as well. We will assume however that the *same* functions $\psi_1, \ldots, \psi_m$ are used in the linear model, and for simplicity also the *same* model class $\mathcal{G}$. The oracle g_β^* which minimizes

$$\tau_\beta^2(g|g_{0,\beta}) = \{\Gamma_\beta(g) - \Gamma_\beta(g_{0,\beta}) + 4(\lambda_n/\lambda_{\min})\sqrt{d_g}J_\beta^{-1}(4c(\lambda_n/\lambda_{\min})\sqrt{d_g})\}, \tag{38}$$

clearly also depends on β, as well as its dimension d_β^*. (Here, we have a ambiguity in notation as τ_c^2 and g_c^* were used in Section 2 with a different meaning, but we believe the risk of confusion is small.) The L_1-penalized β-quantile estimator will be denoted by $\hat{g}_{n,\beta}$.

Theorem 2. *Assume that $\mathcal{G}$ is a convex set and that the margin condition (13) holds for strictly increasing functions J_β, and for each $g \in \mathcal{G}$ and $\beta \in \mathcal{B}$. Take $\lambda_n = (12+u)2\sqrt{\log n/n}$, with $u \geq 6$. Then for $n \geq N$, where N is such that $\lambda_n/n + \sup_{\beta\in\mathcal{B}} \tau^2_\beta(g^*_\beta|g_{0,\beta}) \leq 1/8$, and for any $0 < \delta \leq 1$,*

$$\mathbf{P}\left(\Gamma_\beta(\hat{g}_{n,\beta}) - \Gamma_\beta(g_{0,\beta}) \geq (1+\delta)^2\left[\frac{\lambda_n}{n} + \tau^*_{n,\beta}(\delta)^2\right]\right) \leq 64R^2\exp[-\frac{u^2\log n}{1024R^2}], \tag{39}$$

where

$$\tau^*_{n,\beta}(\delta)^2 = \left[\Gamma_\beta(g^*_\beta) - \Gamma_\beta(g_{0,\beta}) + \frac{2}{\delta^2}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*_\beta}J^{-1}_\beta(\frac{4}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*_\beta})\right] \tag{40}$$

and where $R = 1 + \sup_{\beta\in\mathcal{B}} \|g^*_\beta - g_{0,\beta}\|$.

4. A TYPICAL EXAMPLE

In this section, we address the question what Theorem 1 has to say on adaptation to smoothness. We again fix $\beta \in (0,1)$ and drop subscripts β in the notation. Let us introduce a smoothness parameter s. For example, if g_0 is a function of one variable x in a bounded interval, say $x \in [0,1]$, one could describe the amount of smoothness by the value of the squared Sobolev pseudo-norm,

$$\int_0^1 |g_0^{(s)}(x)|^2 dx, \tag{41}$$

where $g_0^{(s)}$ is the s-th derivative of g_0. More generally, in higher dimensional space $\mathcal{X}$. the smoothness s of a function $g_0 : \mathcal{X} \to \mathbf{R}$ can be the *effictive* smoothness (such as (roughly) the number of derivatives divided by the dimension of $\mathcal{X}$). The parameter s may also be the smoothness parameter appearing in the definition of the Besov pseudo-norm, etc. Here, we will not provide any approximation theory nor define the concept of smoothness in any precise way. We simply *assume* that for some s, and c, and all $d \leq m$,

$$\inf_{\mathcal{G}_d}\inf_{g\in\mathcal{G}_d} \|g - g_0\| \leq cd^{-s}. \tag{42}$$

Here, the infimum is taken over the $\binom{m}{d}$ d-dimensional spaces

$$\mathcal{G}_d = \{\sum_{j\in\mathcal{J}} \alpha_j\psi_j\} \cap \mathcal{G},\ |\mathcal{J}| = d. \tag{43}$$

For smoothness classes where (42) is met, we refer to general work in approximation theory, for example [12], or [4].

Let us now assume that the margin condition (13) is met with $J(\xi) = \xi^r/\sigma^r$, for $\xi \leq \epsilon$, where $r > 0$, $\sigma > 0$ and $\epsilon > 0$ (see also Example 1). Thus

$$\Gamma(g) - \Gamma(g_0) \geq \|g - g_0\|^{r+1}/\sigma^r,\ g \in \mathcal{G},\ \|g - g_0\| \leq \epsilon. \tag{44}$$

Assume also the reverse holds for an appropriate constant η, i.e.,

$$\Gamma(g) - \Gamma(g_0) \leq \|g - g_0\|^{r+1}/\eta^r,\ g \in \mathcal{G},\ \|g - g_0\| \leq \epsilon. \tag{45}$$

We may assume that eventually $\|\hat{g}_n - g_0\| \leq \epsilon$ by convexity arguments, provided also $\|g^* - g_0\|$ is small enough. The convexity argument is explained in [18] (it is also used in the proof of Theorem 1).

The relevant expression appearing in Theorem 1 is the value $\tau_n^*(1)^2$ at g^* of the quantity

$$\Gamma(g) - \Gamma(g_0) + 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d_g}J^{-1}(4\frac{\lambda_n}{\lambda_{\min}}\sqrt{d_g}) \tag{46}$$

which can be bounded by

$$\frac{c^{r+1}}{\eta^r}\left(\|g - g_0\|^{-s(r+1)} + (\frac{4\lambda_n}{\lambda_{\min}})^{\frac{r+1}{r}}\frac{\sigma\eta^r}{c^{r+1}}d_g^{\frac{r+1}{2r}}\right). \tag{47}$$

Taking (47) as starting point, The optimal trade-off has solution

$$d_0^* = \frac{c^{r+1}}{\eta^r}\arg\min\{d^{-s(r+1)} + \delta_n d^{\frac{r+1}{2r}}\}, \tag{48}$$

where

$$\delta_n = (\frac{4\lambda_n}{\lambda_{\min}})^{\frac{r+1}{r}}\frac{\sigma\eta^r}{c^{r+1}}. \tag{49}$$

This optimal value is

$$d_0^* = (\frac{2rs}{\delta_n})^{\frac{2r}{(2rs+1)(r+1)}}, \tag{50}$$

where we tacitly assume that this is an integer. Inserting that value in (47) gives

$$\begin{aligned}\Gamma(g^*) - \Gamma(g_0) + 2\lambda_n\sqrt{d^*}J^{-1}(\frac{4\lambda_n}{\lambda_{\min}}\sqrt{d^*}) &\leq \frac{c^{r+1}}{\eta^r}c_{r,s}\delta_n^{\frac{2rs}{2rs+1}} \\ &= \frac{c^{r+1}}{\eta^r}c_{r,s}(\frac{4\lambda_n}{\lambda_{\min}})^{\frac{2s(r+1)}{2rs+1}}(\frac{\sigma\eta^r}{c^{r+1}})^{\frac{2rs}{2rs+1}}.\end{aligned} \tag{51}$$

Here,

$$c_{r,s} = (2rs)^{-\frac{2rs}{2rs+1}} + (2rs)^{\frac{1}{2rs+1}}. \tag{52}$$

Thus, taking λ_n of order $\sqrt{\frac{\log n}{n}}$ gives a rate of convergence of order

$$(\frac{\log n}{n})^{\frac{s(r+1)}{2rs+1}}. \tag{53}$$

For $r = 1$ this corresponds, up to the factor $(\log n)^{\frac{2s}{2s+1}}$, to the usual rate $n^{-2s/(2s+1)}$ for a model of smoothness s.

5. PROOFS

The proof of Theorem 1 is a modification of arguments used in [11]. Throughout the proofs, we use the notation

$$\nu_n(\alpha) = [\Gamma_n(g_\alpha) - \Gamma_n(g^*)] - [\Gamma(g_\alpha) - \Gamma(g^*)],\ \alpha \in \mathbf{R}^m, \tag{54}$$

for the empirical process, and

$$I(\alpha) = \sum_{j=1}^{m} |\alpha_j|,\ \alpha \in \mathbf{R}^m, \tag{55}$$

for the L_1-norm. We furthermore write, for $M > 0$, $R > 0$,

$$\mathcal{A}_{M,R} = \{\alpha \in \mathbf{R}^m :\ I(\alpha - \alpha^*) \leq M,\ \|g_\alpha - g^*\| \leq R\}. \tag{56}$$

Lemma 3 and Lemma 4 study the "random part" of the problem: they provide a probability inequality for the empirical process. Theorem 1 uses these lemmas to arrive at its result. The proof Theorem 2 is established in a similar fashion, at the end of this section.

Lemma 3. *For all $M > 0$, $R > 0$, and $u > 0$, the following upper bound holds*

$$\mathbf{P}\left(\sup_{\alpha \in \mathcal{A}_{M,R}} |\nu_n(\alpha)| \geq (12+u)M\sqrt{\frac{\log n}{n}}\right) \leq \exp[-\frac{u^2((M^2/R^2) \vee 1)\log n}{32}]. \tag{57}$$

Here $a \vee b = \max(a, b)$.

Proof. Define the random variable

$$Z = \sup_{\alpha \in \mathcal{A}_{M,R}} |\nu_n(\alpha)|. \tag{58}$$

Set

$$U_i(\alpha) = \gamma(Y_i - g_\alpha(x_i)) - \gamma(Y_i - g^*(x_i)),\ i = 1, \ldots, n. \tag{59}$$

Then (see [9]),

$$\mathbf{P}(Z \geq \mathbf{E}(Z) + u) \leq \exp[-\frac{n^2u^2}{8b_n^2}], \tag{60}$$

where b_n^2 is assumed to satisfy

$$b_n^2 \geq \sup_{\alpha \in \mathcal{A}_{M,R}} \sum_{i=1}^{n} |U_i(\alpha) - \mathbf{E}U_i(\alpha)|^2. \tag{61}$$

But since γ is 1-Lipschitz,

$$|U_i(\alpha)| \leq |g_\alpha(x_i) - g^*(x_i)|, \tag{62}$$

so we may take

$$b_n^2 \leq \sup_{\alpha \in \mathcal{A}_{M,R}} 4n\|g_\alpha - g^*\|^2 \leq 4n(M^2 \wedge R^2), \tag{63}$$

where $a \wedge b = \min(a, b)$, and where we used the normalization $\lambda_{\max} = 1$, so that

$$\|g_\alpha - g^*\|^2 \leq \sum_{j=1}^m (\alpha_j - \alpha_j^*)^2, \tag{64}$$

and moreover,

$$\sum_{j=1}^m (\alpha_j - \alpha_j^*)^2 \leq I^2(\alpha - \alpha^*). \tag{65}$$

A symmetrization procedure (see e.g., [10] or [19]), yields that

$$\mathbf{E}(Z) \leq \mathbf{E}\left(\sup_{\alpha \in \mathcal{A}_{M,R}} |\frac{1}{n} \sum_{i=1}^n \epsilon_i U_i(\alpha)| \right), \tag{66}$$

where $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables. Because γ is 1-Lipschitz, we can apply the contraction principle ([10], Theorem 4.12)), which gives

$$\begin{aligned} \mathbf{E}\left(\sup_{\alpha \in \mathcal{A}_{M,R}} |\frac{1}{n} \sum_{i=1}^n \epsilon_i U_i(\alpha)| \right) &\leq 2\mathbf{E}\left(\sup_{\alpha \in \mathcal{A}_{M,R}} |\frac{1}{n} \sum_{i=1}^n \epsilon_i (g_\alpha(x_i) - g^*(x_i))| \right) \\ &\leq 2M\mathbf{E}\left(\max_{j=1,\ldots,m} |\frac{1}{n} \sum_{j=1}^m \epsilon_j \psi_j(x_i)| \right). \end{aligned} \tag{67}$$

Now, $\|\psi_j\| \leq 1$ for all $j = 1, \ldots, m$, because $\lambda_{\max} = 1$. Applying results in [19] (Chapter 2.2), we arrive at the bound

$$\mathbf{E}\left(\max_{j=1,\ldots,m} |\frac{1}{n} \sum_{j=1}^m \epsilon_j \psi_j(x_i)| \right) \leq 6\sqrt{\frac{\log n}{n}}. \tag{68}$$

As a consequence,

$$\mathbf{P}(Z \geq 12M\sqrt{\frac{\log n}{n}} + u) \leq \exp[-\frac{nu^2}{32(M^2 \wedge R^2)}]. \tag{69}$$

Replacing u by $uM\sqrt{\log n/n}$ completes the proof. □

Lemma 4. *We have for all $R \geq 1$ and $u \geq 6$*

$$\begin{aligned} \mathbf{P}\left(\sup_{\|g_\alpha - g^*\| \leq R} \frac{|\nu_n(\alpha)|}{I(\alpha - \alpha^*) + \frac{1}{n}} > (12 + u)2\sqrt{\frac{\log n}{n}} \right) \\ \leq 4R^2 \exp[-\frac{u^2 \log n}{64R^2}]. \end{aligned} \tag{70}$$

Proof. First we consider the set where $I(\alpha-\alpha^*)\le\frac{1}{n}$. By Lemma 3,

$$\mathbf{P}\left(\sup_{\alpha\in\mathcal{A}_{\frac{1}{n},R}}|\nu_n(\alpha)|>(12+u)\frac{2}{n}\sqrt{\frac{\log n}{n}}\right)\le\exp[-\frac{u^2\log n}{32}]. \tag{71}$$

Next, we consider $(1/n)<I(\alpha-\alpha^*)\le 1$. Take j_0 as the smallest integer such that $j_0+1>\log_2 n$. We find from Lemma 3,

$$\begin{aligned}\mathbf{P}&\left(\sup_{1/n<I(\alpha-\alpha^*)\le 1,\ \|g_\alpha-g^*\|\le R}\frac{|\nu_n(\alpha)|}{I(\alpha-\alpha^*)}>(12+u)2\sqrt{\frac{\log n}{n}}\right)\\&\le\sum_{j=0}^{j_0}\mathbf{P}\left(\sup_{\alpha\in\mathcal{A}_{2^{-j},R}}|\nu_n(\alpha)|>(12+u)2^{-j}\sqrt{\frac{\log n}{n}}\right)\\&\le(n+1)\exp[-\frac{u^2\log n}{32}]\le\exp[-\frac{u^2\log n}{64}].\end{aligned} \tag{72}$$

Next, we consider the set where $I(\alpha-\alpha^*)>1$. There

$$\begin{aligned}\mathbf{P}&\left(\sup_{I(\alpha-\alpha^*)>1,\ \|g_\alpha-g^*\|\le R}\frac{|\nu_n(\alpha)|}{I(\alpha-\alpha^*)}>(12+u)2\sqrt{\frac{\log n}{n}}\right)\\&\le\sum_{j=0}^{\infty}\mathbf{P}\left(\sup_{\alpha\in\mathcal{A}_{2^{j+1},R}}|\nu_n(\alpha)|>(12+u)2^{j+1}\sqrt{\frac{\log n}{n}}\right)\\&\le\sum_{j=0}^{\infty}\exp[-\frac{u^2 2^{2(j+1)}\log n}{32R^2}]\\&\le\sum_{j=0}^{\infty}\exp[-\frac{u^2(j+2)\log n}{32R^2}]\le 2R^2\exp[-\frac{u^2\log n}{32R^2}].\end{aligned} \tag{73}$$

□

Proof of Theorem 1. First, suppose we already know that $\|\hat g_n-g^*\|\le\bar R$, where we take $\bar R=2+4\|g^*-g_0\|$. Let $\mathbf{A}$ be the set

$$\mathbf{A}=\{|\nu_n(\alpha)|\le\lambda_n I(\alpha-\alpha^*)+\lambda_n/n,\ \text{for all}\ \|g_\alpha-g^*\|\le\bar R\}. \tag{74}$$

Then by Lemma 4,

$$\mathbf{P}(\mathbf{A})\ge 1-4\bar R^2\exp[-\frac{u^2\log n}{64\bar R^2}]. \tag{75}$$

So let us consider what happens on the set $\mathbf{A}$.

Clearly, the inequality

$$\Gamma_n(\hat g_n)+\lambda_n I(\hat\alpha_n)\le\Gamma_n(g^*)+\lambda_n I(\alpha^*) \tag{76}$$

may be rewritten in the form

$$\Gamma(\hat g_n)-\Gamma(g_0)\le-\nu_n(\hat\alpha_n)-\lambda_n[I(\hat\alpha_n)-I(\alpha_*)]+\Gamma(g^*)-\Gamma(g_0). \tag{77}$$

So on $\mathbf{A}$,

$$\Gamma(\hat{g}_n) - \Gamma(g_0) \leq \lambda_n I(\hat{\alpha}_n - \alpha^*) + \frac{\lambda_n}{n} - \lambda_n[I(\hat{\alpha}_n) - I(\alpha_*)] + \Gamma(g^*) - \Gamma(g_0). \tag{78}$$

Let $\mathcal{J}^* = \{j : \alpha_j^* \neq 0\}$, and let for any α, $I_1(\alpha) = \sum_{j\in\mathcal{J}^*} |\alpha_j|$ and $I_2(\alpha) = \sum_{j\notin\mathcal{J}^*} |\alpha_j|$. Since $I_2(\alpha - \alpha^*) = I_2(\alpha)$, we now find

$$\begin{aligned} \Gamma(\hat{g}_n) - \Gamma(g_0) &\leq \lambda_n I_1(\hat{\alpha}_n - \alpha^*) + \lambda_n I_2(\hat{\alpha}_n) \\ -\lambda_n[I_1(\hat{\alpha}_n) - I_1(\alpha^*)] - \lambda_n I_2(\hat{\alpha}_n) &+ \Gamma(g^*) - \Gamma(g_0) \\ = \lambda_n I_1(\hat{\alpha}_n - \alpha^*) - \lambda_n[I_1(\hat{\alpha}_n) - I_1(\alpha^*)] &+ \Gamma(g^*) - \Gamma(g_0). \end{aligned} \tag{79}$$

Since for any $a, b \in \mathbf{R}$, $|a| - |b| \leq |a - b|$, we arrive at

$$\Gamma(\hat{\alpha}_n) - \Gamma(\alpha_0) \leq 2\lambda_n I_1(\hat{\alpha}_n - \alpha^*) + \tfrac{\lambda_n}{n} + \Gamma(\alpha^*) - \Gamma(\alpha_0). \tag{80}$$

Application of first the Cauchy-Schwarz inequality and the inequality $\sum_{j=1}^m (\alpha_j - \alpha_j^*)^2 \leq \|g_\alpha - g^*\|^2/\lambda_{\min}^2$, and then the triangle inequality yields

$$\begin{aligned} \Gamma(\hat{g}_n) - \Gamma(g_0) &\leq 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|\hat{g}_n - g^*\| + \frac{\lambda_n}{n} + \Gamma(g^*) - \Gamma(g_0) \\ \leq 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d_*}\|\hat{g}_n - g_0\| &+ 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|g^* - g_0\| + \frac{\lambda_n}{n} + \Gamma(g^*) - \Gamma(g_0) \\ &= I + II + III, \end{aligned} \tag{81}$$

where

$$I = 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|\hat{g}_n - g_0\|, \tag{82}$$

$$II = 2\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|g^* - g_0\|, \tag{83}$$

and

$$III = \frac{\lambda_n}{n} + \Gamma(g^*) - \Gamma(g_0). \tag{84}$$

If $I \geq \delta(II + III)$ and invoking margin condition (13), we now obtain

$$\|\hat{g}_n - g_0\| J(\|\hat{g}_n - g_0\|) \leq \Gamma(\hat{g}_n) - \Gamma(g_0) \leq 2(1 + \frac{1}{\delta})\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|\hat{g}_n - g_0\|. \tag{85}$$

This implies

$$\|\hat{g}_n - g_0\| \leq J^{-1}(2(\frac{1+\delta}{\delta})\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}) \tag{86}$$

and hence

$$\Gamma(\hat{\alpha}_n) - \Gamma(\alpha_0) \leq 2(\frac{1+\delta}{\delta})\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}J^{-1}(2(\frac{1+\delta}{\delta})\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}). \tag{87}$$

If $II \geq \delta(III)$, we get (by condition (13))

$$\|g^* - g_0\| J(\|g^* - g_0\|) \leq \Gamma(g^*) - \Gamma(g_0) \leq \frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}\|g^* - g_0\|, \tag{88}$$

which gives

$$\|g^* - g_0\| \leq J^{-1}(\frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}). \tag{89}$$

So then

$$II \leq \frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}J^{-1}(\frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d_*}). \tag{90}$$

So if $I \leq \delta(II + III)$ and $II \geq \delta(III)$,

$$\begin{aligned}\Gamma(\hat{g}_n) - \Gamma(g_0) &\leq (1+\delta)(II + III) \\ &\leq (1+\delta)(1+\frac{1}{\delta})\frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^**}J^{-1}(\frac{2}{\delta}\frac{\lambda_n}{\lambda_{\min}}\sqrt{d^*}).\end{aligned} \tag{91}$$

Finally, if $I \leq \delta(II + III)$ and $II \leq \delta(III)$, we clearly have

$$\Gamma(\hat{g}_n) - \Gamma(g_0) \leq (1+\delta)(II + III) \leq (1+\delta)^2(III). \tag{92}$$

We now come back to our starting point, namely the assumption $\|\hat{g}_n - g^*\| \leq \bar{R}$, with $\bar{R} = 2 + 4\|g^* - g_0\|$. We may replace in the proof we have so far, the function $\hat{g}_n$ by the convex combination $\hat{g}_{n,t} = t\hat{g}_n + (1-t)g^*$, where $t = 1/(1 + \|\hat{g}_n - g^*\|/\bar{R})$. This is because clearly,

$$\|\hat{g}_{n,t} - g^*\| \leq \bar{R}. \tag{93}$$

and because (77) is also true when $\hat{g}_n$ is replaced by $\hat{g}_{n,t}$. So, so far, we have shown that (36) holds for $\hat{g}_{n,t}$ instead of $\hat{g}_n$.

But now, for $\hat{g}_{n.t}$ implies that (take $\delta = 1$), on $\mathbf{A}$,

$$\|\hat{g}_{n,t} - g_0\| \leq 4(\frac{\lambda_n}{n} + \tau^2(g^*|g_0)), \tag{94}$$

which is less than $1/2$ for $n \geq N$. So then $\|\hat{g}_{n,t} - g^*\| \leq \|\hat{g}_{n,t} - g_0\| + \|g^* - g_0\| \leq \frac{1}{2} + \|g^* - g_0\|$. But then,

$$\begin{aligned}\|\hat{g}_n - g^*\| &\leq (1 + \|\hat{g}_n - g^*\|/\bar{R})(\frac{1}{2} + \|g^* - g_0\|) \\ &\leq \frac{1}{2} + \frac{3}{4}\|\hat{g}_n - g^*\| + \|g^* - g_0\|,\end{aligned} \tag{95}$$

by our choice $\bar{R} = 2 + 4\|g^* - g_0\| \geq 1 + 4\|g^* - g_0\|$. And now we found that

$$\|\hat{g}_n - g^*\| \leq 2 + 4\|g^* - g_0\| = \bar{R}. \tag{96}$$

Thus, our starting point is true on the set $\mathbf{A}$. This completes the proof. □

Proof of Theorem 2. We go back to Lemma 3, and show it holds uniformly in β. To this end, let us define

$$\bar{\mathcal{A}}_{M,R} = \{(\alpha, \bar{\alpha}) : \ I(\alpha - \bar{\alpha}) \leq M, \|g_\alpha - g_{\bar{\alpha}}\| \leq R\}. \tag{97}$$

Let

$$w_{n,\beta}(\alpha) = \Gamma_{n,\beta}(g_\alpha) - \Gamma_\beta(g_\alpha), \ \alpha \in \mathbf{R}^m. \tag{98}$$

It is clear that

$$w_{n,\beta}(\alpha) = \beta v_{1,n}(\alpha) + (1-\beta) v_{2,n}(\alpha), \tag{99}$$

where

$$v_{1,n}(\alpha) = l_{1,n}(\alpha) - l_1(\alpha), \tag{100}$$

$$l_{1,n}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - g_\alpha(x_i)| \mathrm{l}\{Y_i - g_\alpha(x_i) < 0\}, \tag{101}$$

and

$$l_1(\alpha) = \mathbf{E} l_{1,n}(\alpha). \tag{102}$$

Moreover,

$$v_{2,n}(\alpha) = l_{2,n}(\alpha) - l_2(\alpha), \tag{103}$$

$$l_{2,n}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - g_\alpha(x_i)| \mathrm{l}\{Y_i - g_\alpha(x_i) \geq 0\}, \tag{104}$$

and

$$l_2(\alpha) = \mathbf{E} l_{2,n}(\alpha). \tag{105}$$

Now, define

$$Z_1 = \sup_{(\alpha,\bar{\alpha}) \in \bar{\mathcal{A}}_{M,R}} |v_{1,n}(\alpha) - v_{1,n}(\bar{\alpha})|, \tag{106}$$

and

$$Z_2 = \sup_{(\alpha,\bar{\alpha}) \in \bar{\mathcal{A}}_{M,R}} |v_{2,n}(\alpha) - v_{2,n}(\bar{\alpha})|. \tag{107}$$

Also, let

$$Z = \sup_{\beta \in \mathcal{B}} \sup_{(\alpha,\bar{\alpha}) \in \bar{\mathcal{A}}_{M,R}} |w_{n,\beta}(\alpha) - w_{n,\beta}(\bar{\alpha})|. \tag{108}$$

As in the proof of Lemma 3, one can show that $\mathbf{E}Z_1 \leq 12M\sqrt{\log n/n}$ as well as $\mathbf{E}Z_2 \leq 12M\sqrt{\log n/n}$. Thus also

$$\mathbf{E}\mathbf{Z} \leq 12M\sqrt{\log n/n}. \tag{109}$$

Arguing as in the proof of Lemma 3 and Lemma 4, we now arrive at a uniform in β version of Lemma 4:

$$\mathbf{P}\left(\sup_{\beta\in\mathcal{B}}\sup_{\|g_\alpha-g_{\bar\alpha}\|\leq R}\frac{|v_{n,\beta}(\alpha)-v_{n,\beta}(\bar\alpha)|}{I(\alpha-\bar\alpha)+\frac{1}{n}}>(12+u)2\sqrt{\frac{\log n}{n}}\right)\leq 4R^2\exp[-\frac{u^2\log n}{64R^2}], \qquad (110)$$

for all $R\geq 1$ and $u\geq 6$.

Now, replace in the proof of Theorem 1, the set $\mathbf{A}$ by

$$\{|v_{n,\beta}(\alpha)-v_{n,\beta}(\alpha^*_\beta)|\leq\lambda_n I(\alpha-\alpha^*_\beta)+\lambda/n,\ \text{for all } \|g_\alpha-g^*_\beta\|\leq\bar R \text{ and all } \beta\in\mathcal{B}\}, \qquad (111)$$

and proceed as there. □

REFERENCES

1. Akaike, H. Information theory and an extension of the maximum likelihood principle. Proceedings 2nd International Symposium on Information Theory, P.N. Petrov and F. Csaki (eds.), Akademia Kiado, Budapest (1973) 267-281.
2. Donoho, D.L., and Johnstone, I.M. Ideal spatial adaptation via wavelet shrinkage. Biometrika No. 81 (1994) 425-455.
3. Donoho, D.L. Denoising via soft-thresholding. IEEE Transactions in Information Theory No. 41 (1995) 613-627.
4. Edmunds, E., and Triebel, H. Entropy numbers and approximation numbers in function spaces. II. Proceedings of the London Mathematical Society (3) No. 64 (1992) 153-169.
5. Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, New York, 2001.
6. Koenker, R., and Bassett Jr. G. Regression quantiles. Econometrica No. 46 (1978) 33-50.
7. Koenker, R., Ng, P.T., and Portnoy, S.L. Nonparametric estimation of conditional quantile functions. L_1 Statistical Analysis and Related Methods, Y. Dodge (ed.), Elsevier, Amsterdam (1992) 217-229.
8. Koenker, R., Ng, P.T., and Portnoy, S.L. Quantile smoothing splines. Biometrika No. 81 (1994) 673-680.
9. Massart, P. Some applications of concentration inequalities to statistics. Ann. Fac. Sci. Toulouse No. 9 (2000) 245-303.
10. Ledoux, M., and Talagrand, M. Probability in Banach Spaces, Isoperimetry and Processes. Springer, Berlin, 1991.
11. Loubes, J.-M., and van de Geer S. Adaptive estimation, using soft thresholding type penalties. Statistica Neerlandica No. 56 (2002) 453-478.
12. Pinkus, A. n-widths in Approxiamation Theory. Springer, New York, 1985.
13. Portnoy, S. Local asymptotics for quantile smoothing splines. Ann. Statist. No. 25 (1997) 414-434.
14. Portnoy, S., and Koenker, R. The Guassian hare and the Laplacian tortoise: computability of squared error versus absolute-error estimators, with discussion. Stat. Science No. 12 (1997) 279-300.
15. Schwarz, G. Estimating the dimension of a model. Ann. Statist. No. 6 (1978) 461-464.

16. Tibshirani, R. Regression analysis and selection via the LASSO. Journal Royal Statist. Soc. B No. 58 (1996) 267-288.
17. van de Geer, S. Empirical Processes in M-Estimation. Cambridge University Press, 2000.
18. van de Geer, S. M-estimation using penalties or sieves. J. Statist. Planning Inf. No. 108 (2002) 55-69.
19. van der Vaart, A.W., and Wellner, J.A. Weak Convergence and Empirical Processes, with Applications to Statistics. Springer, New York, 1996.