Biometrika (2020), **103**, 1, *pp*. 1–14
Printed in Great Britain Advance Access publication on 31 July 2018

Adaptive Risk Bounds for Quantile Trend Filtering

BY OSCAR HERNAN MADRID PADILLA
Department of Statistics, University of California, Los Angeles,
520 Portola Plaza, Los Angeles, CA 90095.
oscar.madrid@stat.ucla.edu

AND SABYASACHI CHATTERJEE
Department of Statistics at University of Illinois at Urbana-Champaign,
725 S. Wright St. M/C 374 Champaign, IL 61820
sc1706@illinois.edu

SUMMARY

We study quantile trend filtering, a recently proposed method for one-dimensional nonparametric quantile regression. We show that the penalized version of quantile trend filtering attains minimax rates, off by a logarithmic factor, for estimating the vector of quantiles when its kth discrete derivative belongs to the class of bounded variation signals. Our results also show that the constrained version of trend filtering attains minimax rates in the same class of signals. Furthermore, we show that if the true vector of quantiles is piecewise polynomial, then the constrained estimator attains optimal rates up to a logarithmic factor. All of our results hold based on a robust metric and under minimal assumptions of the data generation mechanism. We also illustrate how our technical arguments can be used for analysing other shape constrained problems with quantile loss. Finally, we provide extensive experiments that show that quantile trend filtering can perform well, based on mean squared error criteria, under Gaussian, Cauchy, and t-distributed errors.

Some key words: Total variation, nonparametric quantile regession, local adaptivity, fused lasso.

1. INTRODUCTION

1.1. *Introduction*

In this paper we focus on the problem of nonparametric quantile regression for the sequence model. Specifically, given a random vector $y \in R^n$ and a quantile level $\tau \in (0,1)$, our goal is to estimate θ^*, the vector of τ-quantiles of y, given as

$$\theta_i^* = F_{y_i}^{-1}(\tau), \quad \text{for } i = 1, \ldots, n.$$

Here F_{y_i} is the cumulative distribution function of y_i, and we assume that $y_1, \ldots, y_n$ are independent.

Throughout this paper, our focus is on signals that have small rth total variation as in Tibshirani [2014]. The latter is defined as

$$\mathrm{TV}^{(r)}(\theta^*) := n^{r-1}\|D^{(r)}\theta^*\|_1,$$

with

$$D^{(1)} = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{pmatrix} \in R^{(n-1)\times n}$$

and for $r > 1$, we define $D^{(r)} = D^{(1)}D^{(r-1)}$, where $D^{(1)}$ is of the appropriate dimension. Since we assume that the rth total variation $\mathrm{TV}^{(r)}(\theta^*)$ is small, it is natural to consider the estimator

$$\hat{\theta}^{(r)} \;=\; \underset{\theta\in R^n}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \;+\; \lambda \|D^{(r)}\theta\|_1 \right\}, \tag{1}$$

for a tuning parameter $\lambda > 0$. The intuition behind (1) is that $r = 1$ produces piecewise constant estimates, $r = 2$ piecewise quadratic, and for general r the estimator is piecewise polynomial of degree at most $r - 1$.

We highlight that the case of $r = 1$ in (1) appeared in Li and Zhu [2007] within a context of array CGH data for cancer studies. When $r = 1$ we refer to the estimator as quantile fused lasso. More recently, with an application to air quality data, Brantley et al. [2019] proposed the general quantile trend filtering of order r. We use the convention that $D^{(0)} \in R^{n\times n}$ is the identity matrix.

A related estimator that we will consider is

$$\begin{aligned} \hat{\theta}^{(r)}_C \;=\; & \underset{\theta\in R^n}{\arg\min} \;\; \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \\ & \text{subjec to} \;\; \|D^{(r)}\theta\|_1 \;\le\; n^{1-r}V, \end{aligned} \tag{2}$$

where V is a tuning parameter. Thus, $\hat{\theta}^{(r)}_C$ is the constrained version of $\hat{\theta}^{(r)}$ defined in (1).

1.2. *Summary of results*

Our goal in this paper is to extend the regression trend filtering theory to the setting of quantile regression. The first result that we generalize comes from Mammen and van de Geer [1997], Tibshirani [2014]. These authors proved that trend filtering regression, the estimator

$$\hat{\theta} = \underset{\theta\in R^n}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \theta_i)^2 \;+\; \lambda \|D^{(r)}\theta\|_1 \right\}, \tag{3}$$

with an appropriate choice of λ, satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta^*_i\right)^2 \;=\; O_{\mathrm{pr}}\left\{ n^{-2r/(2r+1)} \right\}, \tag{4}$$

where the left hand size is known as the mean squared error (MSE). Such result holds under the assumption that the data are Sub-Gaussian and requires $\mathrm{TV}^{(r)}(\theta^*) = O(1)$. In this paper, we extend (4) to the quantile setting and show that $\hat{\theta}^{(r)}$ defined in (1) satisfies

$$\Delta_n^2\left\{\theta^* - \hat{\theta}^{(r)}\right\} \;=\; O_{\mathrm{pr}}\left\{ n^{-2r/(2r+1)} \left(\log n\right)^{1/(2r+1)} \right\}. \tag{5}$$

Throughout, we use the notation $\Delta_n^2 \,:\, R^n \to R$ for the function given as

$$\Delta_n^2(v) \;=\; \frac{1}{n}\sum_{i=1}^{n} \min\left\{|v_i|, v_i^2\right\}, \qquad \text{for } v \in R^n, \tag{6}$$

which, up to constants, is a Huber loss, see Huber [1964].

The upper bound in (5) holds under general assumptions on the distributions F_{y_i} which include heavy-tailed distributions such as the Cauchy distribution. Notably, (5) implies that $\hat{\theta}^{(r)}$ attains minimax rates, up to logarithmic factors, for estimating signals in the class of signals with bounded rth order total variation. Additionally, we show that in the same class, the constrained estimator (2) attains the rate $n^{-2r/(2r+1)}$ under the error metric $\Delta_n^2(\cdot)$.

Supposing that the vector $D^{(r-1)}\theta^*$ has s change points satisfying a minimal spacing condition, we prove that, with an ideal tuning parameter, the estimator $\hat{\theta}_C^{(r)}$ satisfies

$$\Delta_n^2\left\{\theta^* - \hat{\theta}_C^{(r)}\right\} = O_{\mathrm{pr}}\left\{\frac{(s+1)}{n}\log\left(\frac{en}{s+1}\right)\right\}. \tag{7}$$

Thus, when the true quantiles vector is piecewise polynomial, with a minimal spacing condition between change points, the quantile trend filtering estimator attains minimax rates. We refer the reader to Guntuboyina et al. [2017b] which showed that the constrained trend filtering estimator satisfies

$$\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i^*\right)^2 = O_{\mathrm{pr}}\left\{\frac{(s+1)}{n}\log\left(\frac{en}{s+1}\right)\right\}.$$

More recently, **?** showed a similar bound, with extra log factors, that holds for the penalized estimator (3) for $r \in \{1,2,3,4\}$. However, these results from Guntuboyina et al. [2017a] and **?** require sub-Gaussian assumptions on the errors whereas (7) holds for general distributions and does not require any moment conditions.

Our proof technique sheds light upon obtaining convergence for other quantile regression estimators. One important example is the two-dimensional quantile fused lasso obtained by replacing $D^{(r)}$ in (2) with ∇, the incidence matrix of a $n^{1/2} \times n^{1/2}$ grid in two dimensions, see for instance Hutter and Rigollet [2016]. We show that the resulting estimator, under the loss $\Delta_n^2(\cdot)$ and general assumptions, attains the rate $n^{-1/2}\log n$. This matches the theory for two-dimensional total variation under sub-Gaussian errors in Hutter and Rigollet [2016], Chatterjee and Goswami [2019]. Another application of our theory is in high-dimensional regression. In the weak sparsity setting, with a fixed design, we show that ℓ_1-constrained quantile regression can consistently estimate the vector of regression coefficients, but without requiring a restricted eigenvalue condition as in Belloni et al. [2011], Fan et al. [2014].

1.3. *Previous work*

Since its introduction by Koenker and Bassett Jr [1978], quantile regression has become a prominent tool in statistics. The attractiveness of quantile regression is due to its flexibility for modelling conditional distributions, construction of predictive models, and even outlier detection applications. The problem of one-dimensional nonparametric quantile regression goes back at least to Utreras [1981], Cox [1983], Eubank [1988] who focused on median regression. However, it was not until Koenker et al. [1994] that a more general treatment was provided with the introduction of quantile smoothing splines in one dimension. These are defined as the solution to problems of the form

$$\underset{g\in\mathcal{C}}{\text{minimize}}\left[\sum_{i=1}^{n}\rho_\tau\{y_i - g(x_i)\} + \lambda\left\{\int_0^1 |g''(x)|^p dx\right\}^{1/p}\right],$$

assuming that $0 < x_i < \ldots < x_n < 1$, where $\lambda > 0$ is a tuning parameter, $p \geq 1$, and $\mathcal{C}$ a suitable class of functions.

The theoretical properties of quantile smoothing splines were studied in He and Shi [1994]. Specifically, the authors in He and Shi [1994] demonstrated that quantile smoothing splines attain the rate $n^{-2r/(2r+1)}$, for estimating quantile functions in the class of Hölder functions of exponent r. To the best of our knowledge, He and Shi [1994] is the closer theoretical quantile regression work to ours given the connections between trend filtering and locally adaptive splines. We refer the reader to **?** for a thorough discussion highlighting that trend filtering is in fact a special case of discrete splines.

In the context of median regression in one dimension, the authors in Brown et al. [2008] showed that a wavelet-based quantile regression approach attains minimax rates for estimating the median function, when the latter belongs to Besov spaces. However, despite the optimality of wavelet methods, it is known that total variation based methods can outperform wavelet methods in practice, see Tibshirani [2014], Wang et al. [2016]. Thus, we focus on trend filtering based estimators as in (1).

A precursor of trend filtering can be traced back in the machine learning literature to Rudin et al. [1992] who proposed total variation methods for image denoising applications. In the statistics literature, trend filtering was introduced as locally adaptive regression splines in Mammen and van de Geer [1997]. In its version in (3) trend filtering was independently introduced by **?** and Kim et al. [2009] .

On the computational front, it is known that Problem (3) with $r = 1$ can be solved in $O(n)$, see for instance Johnson [2013]. More recently, Hochbaum and Lu [2017] showed that the corresponding quantile fused lasso estimator, Problem (1) with $r = 1$, can be found in $O(n \log n)$ operations. For $r > 1$, Brantley et al. [2019] proposed an alternating direction method of multipliers (ADMM) based algorithm for computing quantile trend filtering estimators.

In general graphs and with sub-Gaussian noise, Wang et al. [2016] proposed a generalization of trend filtering including theoretical and computational developments. In the particular case of the fused lasso on general graphs, Padilla et al. [2018] proved a general upper bound that only depends on the total variation along the graph and the sample size. Fan et al. [2018] studied an ℓ_0 estimator inspired by total variation regularization. Padilla et al. [2020] proved that the fused lasso in geometric graphs attains minimax results for piecewise Lipchitz classes. Ortelli and van de Geer [2019] studied connections between fused lasso on graphs and the lasso estimator from Tibshirani [1996].

2. MAIN RESULTS

2.1. *Constrained estimator on bounded variation class of signals*

We start studying the constrained quantile trend filtering estimator as defined in (2). Here, we start by stating the modelling assumptions needed to arrive at our first result concerning bounded variation classes of signals. Throughout the paper, we assume that $\tau \in (0, 1)$, and $r \in \{1, 2, \ldots\}$ are fixed. The quantities $\epsilon_i = y_i - \theta_i^*, i = 1, \ldots, n$ are referred as the errors. We also write $V^* = \mathrm{TV}^{(r)}(\theta^*)$. Clearly, $\theta^* \in K$, where

$$K = \left\{ \theta \in R^n \, : \, \mathrm{TV}^{(r)}(\theta) \leq V^* \right\}. \tag{8}$$

Notice that when $r = 1$ the set K becomes the class of bounded variation signals. The case of $r > 1$ corresponds to higher order bounded variation classes, see Tibshirani [2014] for an overview.

Our main assumption stated next requires that for each y_i, there exists a neighborhood around the quantile such that within such neighborhood the cumlative distribution function of y_i grows linearly away from θ_i^*.

Assumption 1. There exists a constant $L > 0$ such that for $\delta \in R^n$ satisfying $\|\delta\|_\infty \leq L$ we have that

$$\min_{i=1,\ldots,n} |F_{y_i}(\theta_i^* + \delta_i) - F_{y_i}(\theta_i^*)| \geq \underline{f}\,|\delta_i|,$$

for some cinstant $\underline{f} > 0$, and where F_{y_i} is the cumulative distribution function of y_i.

If the cumulative distribution functions F_{y_i} have probability density functions f_{y_i}, then Assumption 1 is a weaker condition than requiring that

$$\inf_{\|\delta\|_\infty \leq L} \min_{i=1,\ldots,n} f_{y_i}(\theta_i^* + \delta_i) \geq \underline{f},$$

which appeared as Condition 2 in He and Shi [1994], and it is related to condition D.1 in Belloni et al. [2011]. Furthermore, we hightlight that Assumption 1 will hold for most common distributions including the Cauchy distribution.

We are now ready to state our first result. This shows that quantile trend filtering attains optimal rates for estimating signals in K. The proof of this result is deferred to the Supplementary material.

THEOREM 1. *Under Assumption* 1*, and if* V *in* (2) *is chosen such that* $V \geq V^*$ *then*

$$\Delta_n^2\left\{\theta^* - \hat{\theta}_C^{(r)}\right\} = O_{\mathrm{pr}}\left[n^{-(2r)/(2r+1)} V^{r/(2r+1)} \max\left\{1, \left(\frac{V}{n^{r-1}}\right)^{2r/(2r+1)}\right\}\right].$$

Notably, under the canonical scaling $V^* = O(1)$, Theorem 1 shows that the constrained quantile trend filtering estimator attains minimax rates for estimating θ^* in the class of parameters K, see Mammen and van de Geer [1997], Tibshirani [2014], Guntuboyina et al. [2017b]. However, unlike previous results on trend filtering, our result holds without the strong assumption that the errors are sub-Gaussian. This explains why the upper depends on the loss $\Delta_n^2(\cdot)$ defined in (6).

On another note, the role of τ is not made explicit in Theorem 1. This is because τ is fixed. However, from Assumption 1 and the proof of Theorem 1, it can be seen that the closer τ is to $\{0, 1\}$, the larger the constants are in the upper bound in Theorem 1. For symmetric distributions, the closer τ is to 0.5 the less difficult it becomes to estimate the vector of τ-quantiles θ^*.

2.2. *Constrained estimator fast rates of convergence*

We now show that quantile trend filtering enjoys fast rates of convergence in the sense of Guntuboyina et al. [2017b]. Thus, quantile trend filtering, just like trend filtering, can adapt to potential discontinuities of the signal, and it attains optimal rates of convergence for estimating piecewise polynomial signals. This is stated next.

THEOREM 2. *Suppose that* $s = \|D^{(r)}\theta^*\|_0$*, let* $S = \{j : (D^{(r)}\theta^*)_j \neq 0\}$ *and suppose that*

$$\min_{\ell=1,\ldots,s+1} (j_{\ell+1} - j_\ell) \geq \frac{cn}{s+1},$$

for some constant c *satisfying* $0 \leq c \leq 1$*, and where* $j_0 = 1$*,* $j_{s+1} = n - r$*, with* $j_1, \ldots, j_s$ *are the elements of* S*. Under Assumption* 1*, and* V *in Problem* (2) *chosen as* $V = V^*$*, we have that*

$$\Delta_n^2\left\{\theta^* - \hat{\theta}_C^{(r)}\right\} = O_{\mathrm{pr}}\left[\max\left\{\frac{V^*}{n^{r-1}}, 1\right\} \frac{(s+1)}{n} \log\left(\frac{en}{s+1}\right)\right].$$

For the case of median regression with sub-Gaussian errors and with cannonical scaling $V^* = O(1)$, Theorem 2 shows that the constrained quantile trend filtering estimator attains, off by a logarithmic factor, the rate attained by an oracle estimator that knows the set S, see Guntuboyina et al. [2017b]. However, Theorem 2 holds for general distributions and quantiles going beyond sub-Gaussian distributions.

2.3. *Penalized trend filtering estimator*

We now provide theoretical guarantees for the penalized quantile trend filtering estimator (3). From a computational point of view the penalize quantile trend filtering presents a more appealing method than its constrained counterpart. To elaborate on this point, both (1) and (2) are linear programs that can be solved using any linear programming software. However, for large sized problems linear programming can become burdensome. To address that, previous authors (e.g Hochbaum and Lu [2017], Brantley et al. [2019]) have studied different types of algorithms that can efficiently solve the penalized quantile trend filtering problem. This in contrast to the constrained quantile trend filtering problem that has not received attention from a computational perspective due to its inherit difficulty.

Next, we state our main result for penalized quantile trend filtering.

THEOREM 3. *Suppose that Assumption* 1 *holds and* $V^* = \Theta(1)$. *Then there exists a choice of* λ *for Problem* 1 *satisfying*

$$\lambda = \Theta\left\{ n^{(2r-1)/(2r+1)} \left(\log n\right)^{1/(2r+1)} \|D^{(r)}\theta^*\|_1^{-(2r-1)/(2r+1)} \right\},$$

such that

$$\Delta_n^2\left\{\theta^* - \hat{\theta}^{(r)}\right\} = O_{\rm pr}\left\{ n^{-(2r)/(2r+1)} \left(\log n\right)^{1/(2r+1)} \right\}.$$

Theorem 3 shows that, under the loss $\Delta_n^2(\cdot)$, penalized trend filtering attains minimax rates, up to a logarithmic factor, for estimating signals in the class of bounded variation and its higher order versions. The extra logarithmic factors are the main difference between Theorems 3 and 1. The proof of Theorem 3 uses tools discussed in Section 4 combined with a careful construction of a restricted set in the spirit of Belloni et al. [2011], and exploiting results from Wang et al. [2016] and Guntuboyina et al. [2017b]. Finally, if V^* is allowed to grow to infinity, then, both, the choice of λ and the upper bound in Theorem 3 would need to be inflated with polynomial functions of V^*.

3. OTHER APPLICATIONS

3.1. *Fused lasso in grid graphs*

Total variation denoising in multiple dimensions has attracted tremendous attention due to its application to image denoising problems Rudin et al. [1992]. In this subsection, we study the problem of quantile fused lasso in d dimensions. In particular, we will exploit ideas from Section 4 combined with results from Hutter and Rigollet [2016] to obtain an upper bound, under the loss $\Delta_n^2(\cdot)$.

More precisely, we consider the $n^{1/d} \times \ldots \times n^{1/d}$ d-dimensional grid $G_d = (\{1, \ldots, n\}, E_n)$. For a signal $\theta \in R^n$ we define its total variation along G_d as

$$\|\nabla\theta\|_1 := \sum_{\{i,j\}\in E_n} |\theta_i - \theta_j|,$$

where ∇ is the usual edge vertex incidence matrix of the graph grid G_d. With this notation, we consider the estimator

$$\hat{\theta} \;=\; \underset{\theta \in K}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \right\}, \tag{9}$$

where $K = \{\theta \in R^n \,:\, \|\nabla\theta\|_1 \leq V\}$ for some tuning parameter $V > 0$.

We are now ready to state our main result in this subsection.

THEOREM 4. *Suppose that Assumptions* 1 *holds with $\theta^* \in R^n$ the vector of τ-quantiles of y. If V in (*9*) is chosen to satisfy $V \geq \|\nabla\theta^*\|_1$ and $\|\nabla\theta^*\|_1 = O(n^{1-1/d})$, then*

$$\Delta_n^2(\hat{\theta} - \theta^*) \;=\; O_{\rm pr}\left\{ \frac{V(\log n)^2}{n} \right\},$$

for $d = 2$, and

$$\Delta_n^2(\hat{\theta} - \theta^*) \;=\; O_{\rm pr}\left(\frac{V \log n}{n} \right),$$

*for $d > 2$, where $\hat{\theta}$ is the estimator defined in (*9*).*

Notice that Theorem 4 requires that true signal has total variation along G_d which is of order $O(n^{1-1/d})$. This is a standard setting for denoising in grid graphs, in fact Sadhanala and Tibshirani [2017] refers to this scaling of the total variation as "canonical". Under this condition and Assumption 1, Theorem 4 shows that quantile fused lasso in d dimensions attains minimax rates under the loss $\Delta_n^2(\cdot)$ provided that $V \asymp V^*$. These rates match those in Chatterjee and Goswami [2019] for the constrained fused lasso in two dimensions, see also Hutter and Rigollet [2016] for the corresponding result for the penalized estimator in d dimensions. However, unlike previous results, Theorem 4 holds with a different metric than the mean squared error and it holds under more general settings than sub-Gaussian errors.

3.2. *High-dimensional quantile regression*

Next, we focus on high-dimensional quantile regression. Specifically we consider the constrained version of the ℓ_1-QR estimator defined in Knight and Fu [2000] and studied in Belloni et al. [2011]. ℓ_1-QR is commonly used as a robust tool for variable selection and prediction with high-dimensional covariates, consisting of the quantile version of the lasso estimator from Tibshirani [1996].

More specifically, suppose that we are given $\{(x_i, y_i)\}_{i=1}^n \subset R^p \times R$ with the $\{x_i\}_{i=1}^n$ fixed, and with $y_1, \ldots, y_n$ independent and satisfying the quantile relation

$$F_{y_i}^{-1}(\tau) = x_i^\top \theta^*, \tag{10}$$

where F_{y_i} is the cumulative distribution function of y_i, with $\theta^* \in R^p$ and $\|\theta^*\|_1 = s$. With this setting, we focus on the goal of estimating θ^*. Towards that end, we consider the estimator

$$\hat{\theta} \;=\; \underset{\theta \in K}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \theta) \right\}, \tag{11}$$

where $K = \{\theta \in R^p \,:\, \|\theta\|_1 \leq s\}$.

Before arriving at our main result from this subsection, we first state some assumptions.

Assumption 2. The vector of quantiles θ^* belongs to K. Moreover, there exists a positive constant L such that for $u \in R$ satisfying $|u| \leq L$ we have that

$$\min_{i=1,\ldots,n} |F_{y_i}(x_i^\top \theta_i^* + u) - F_{y_i}(x_i^\top \theta_i^*)| \geq \underline{f}\,|u|,$$

for some $\underline{f} > 0$, where f_{y_i} is the probability density function of y_i.

The previous assumption is the version of Assumption 1 for the setting of high-dimensional regression. A related condition appeared in Belloni et al. [2011].

Our next assumption states that the columns of the design matrix are normalized. This is a standard condition in high-dimensional regression, see Rigollet and Hütter [2015] for a review.

Assumption 3. Let $X \in R^{n\times p}$ be the matrix whose ith row is the vector $x_i^\top$. Denote by $X_{\cdot,j}$ the jth column of X. We assume that $\max_{j=1,\ldots,p} \|X_{\cdot,j}\| \leq n^{1/2}$.

With the conditions from above, we now present our next result.

THEOREM 5. *Suppose that Assumptions* 2–3 *hold. Then there exists a constant* $C > 0$ *such that*

$$E\left[\Delta_n^2\left\{X(\hat{\theta} - \theta^*)\right\}\right] \leq Cs\left(\frac{\log p}{n}\right)^{1/2},$$

where $\hat{\theta}$ *is the estimator defined in* (11).

We emphasise that Theorem 5 holds without conditions on the eigenvalues of the design matrix. This is a crucial difference from previous work in the literature that relies on restricted eigenvalue conditions, see for instance Belloni et al. [2011], Fan et al. [2014], Sun et al. [2019]. However, the price we pay is that our upper bound is stated in terms of the function $\Delta_n^2(\cdot)$ rather than the mean squared error. Furthermore, our rate has an extra $s^{1/2}$ factor as compared to that of Theorem 2 in Belloni et al. [2011], which holds under stronger assumptions than the minimal assumptions in Theorem 5.

4. PROOF IDEAS

Next we present a proof sketch of our results. To that end, we define the empirical loss function

$$\hat{M}_n(\theta) = \sum_{i=1}^n \hat{M}_{n,i}(\theta_i),$$

where

$$\hat{M}_{n,i}(\theta_i) = \rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*).$$

Setting $M_{n,i}(\theta_i) = E\{\rho_\tau(z_i - \theta_i) - \rho_\tau(z_i - \theta_i^*)\}$ where $z \in R^n$ is an independent copy of y, the population loss becomes

$$M_n(\theta) = \sum_{i=1}^n M_{n,i}(\theta_i).$$

Hence, both (1) and (2) are based on a penalized and constrained version of the $\hat{M}_n$ respectively.

We are now ready to state the first step in the proof of all our theorems. This connects the function $\Delta_n^2(\cdot)$ and the quantile population loss.

LEMMA 1. *Suppose that* $\theta_i^* = F_{y_i}^{-1}(\tau)$ *and Assumptions* 1 *holds. Then there exists a constant* $C > 0$ *such that for all* $\delta \in R^n$*, we have*

$$\frac{M_n(\theta^* + \delta)}{n} \geq C\Delta_n^2(\delta),$$

for some positive constant C.

Lemma 1 does not depend on trend filtering and in fact can be used with other shape constrained estimators. Two different ways that we use Lemma 1 in this paper are the following. First, suppose that we are interested in a shape constrained estimator

$$\hat{\theta} \;=\; \underset{\theta \in K}{\arg\min}\; \hat{M}_n(\theta),$$

for a set $K \subset R^n$. Then by Lemma 1 and the optimality of $\hat{\theta}$, it can be proven that

$$E\left\{\Delta_n^2\left(\theta^* - \hat{\theta}\right)\right\} \leq E\left\{\frac{M_n(\hat{\theta})}{C\,n}\right\} \leq \frac{2}{n} E\left[\sup_{v \in K}\left\{M_n(v) - \hat{M}_n(v)\right\}\right], \tag{12}$$

provided that $\theta^* \in K$, see the Supplementary Material for details. Hence, in order to give an upper bound for $E\{\Delta_n^2(\theta^* - \hat{\theta})\}$, it is enough to provide an upper bound for the right most term in (12). We do that in the Supplementary material by using a symmetrization argument and Talagrand's contraction inequality, see for instance Van Der Vaart and Wellner [1996] and Ledoux and Talagrand [2013]. We reduce the problem to controlling

$$E\left(\sup_{v \in K} \sum_{i=1}^{n} \xi_i (v_i - \theta_i^*)\right), \tag{13}$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables. The quantity (13) is commonly known as the Rademacher complexity of the set K. It is well known that, up to a constant, the Rademacher complexity of a set is upper bounded by the Gaussian width or complexity of the same set, see Tomczak-Jaegermann [1989], Bartlett and Mendelson [2002], Wainwright [2019].

When the set K is not compact, as it the case with trend filtering, exploiting the convexity of the quantile loss, our arguments in the Supplementary Material reduce the problem of controlling $\Delta_n^2(\theta^* - \hat{\theta})$ to that of deriving an upper bound on

$$E\left(\sup_{v \in K\,:\,\Delta_n^2(v) \leq \eta^2} \sum_{i=1}^{n} \xi_i (v_i - \theta_i^*)\right), \tag{14}$$

for a carefully chosen $\eta > 0$. To give an upper bound for (14), we exploit results from Guntuboyina et al. [2017b] which controls a similar quantity obtained by replacing $\Delta_n^2(\cdot)$ with the mean squared error.

5. EXPERIMENTS

We now proceed to illustrate with simulations the empirical performance of quantile trend filtering. As benchmark methods, we consider trend filtering (3) with $r = 1$ and $r = 2$ denoted as TF1 and TF2 respectively, and quantile splines (QS) using the R package "fields. Notice that TF1 and TF2 only provide estimates for $\tau = 0.5$. As for quantile trend filtering, we consider the penalized estimator (1) with choices $r = 1$ and $r = 2$ which we denote as QTF1 and QTF2 respectively. These are implemented in R via ADMM, similarly to Brantley et al. [2019]. We

Table 1. *Average mean squared error times 10, $\frac{10}{n}\sum_{i=1}^{n}(\theta_i^* - \hat{\theta}_I)^2$, averaging over 100 Monte carlo simulations for the different methods considered. Captions are described in the text.*

n	Scenario	τ	QTF1	QTF2	QS	TF1	TF2
10000	1	0.5	0.023	0.08	0.21	**0.016**	0.4
5000	1	0.5	0.046	0.12	0.23	**0.034**	0.65
1000	1	0.5	0.18	0.29	0.32	**0.12**	0.94
10000	2	0.5	**0.037**	0.11	0.13	4917385.2	5743.119
5000	2	0.5	**0.066**	0.15	0.17	25215.87	286.45
1000	2	0.5	**0.29**	0.43	0.45	354693.6	11522.6
10000	3	0.5	**0.015**	0.063	0.17	2.26	0.95
5000	3	0.5	**0.029**	0.092	0.18	0.14	0.65
1000	3	0.5	**0.13**	0.24	0.26	2.23	1.04
10000	4	0.5	0.045	**0.009**	0.015	0.065	0.016
5000	4	0.5	0.075	**0.019**	0.027	0.24	0.031
1000	4	0.5	0.30	**0.082**	0.098	0.29	0.31
10000	5	0.5	0.13	0.056	**0.041**	61625.82	134.80
5000	5	0.5	0.24	0.099	**0.086**	1063110.0	877.85
1000	5	0.5	1.92	**0.35**	**0.35**	1443060.0	11531.79
10000	6	0.9	0.18	**0.070**	0.075	*	*
5000	6	0.9	0.29	**0.13**	0.14	*	*
1000	6	0.9	1.19	**0.39**	0.41	*	*
10000	6	0.1	0.16	**0.065**	0.070	*	*
5000	6	0.1	0.31	**0.13**	0.14	*	*
1000	6	0.1	1.27	**0.46**	0.47	*	*

also compared against quantile random forest using the R package "quantregForest" but we omit the results due to poor performance.

For the different competing methods, we choose their corresponding penalty parameter to be the value that minimizes the average mean squared error over 100 Monte Carlo replicates. Here, for each instance of an estimator $\hat{\theta}$ the mean squared error is

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2,$$

with θ^* the vector of quantiles.

Next we describe the generative models or scenarios. For each scenario we generate 100 data sets for different values of n in the set $\{1000, 5000, 10000\}$. We then report the average mean squared error, based on optimal tuning, of the different competing methods. In each scenario the data are generated as

$$y_i = \theta_i^* + \epsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where $\theta^* \in R^n$, and the errors $\{\epsilon_i\}_{i=1}^n$ are independent with $\epsilon_i \sim F_i$ for some distributions F_i, $i = 1, \ldots, n$. We now discuss the different choices of θ^* and F_i's that we consider.

Scenario 1. In this case we take θ^* to satisfy $\theta_i^* = 1$ for $i \in \{1, \ldots, n\} \cup \{n - 2\lfloor n/3 \rfloor + 1, \ldots, n\}$ and $\theta_i^* = 0$ otherwise. As for the F_i's we use the distribution $N(0, 1)$.

Scenario 2. This is the same as Scenario 1, replacing $N(0, 1)$ with $\mathrm{Cauchy}(0, 1)$.

Scenario 3. Once again, we take θ^* as in Scenario 1. With regards to the F_i's, we generate $\epsilon_i \sim i^{1/2}/n^{1/2}v_i$, where $v_i \sim t(2)$. Here $t(2)$ denotes the t-dsitribution with 2 degrees of freedom.

Scenario 4. We set $\theta_i^* = 3(i/n)$, for $i \in \{1, \ldots, \lfloor n/2 \rfloor\}$, and $\theta_i^* = 3(1 - i/n)$ for $\{\lfloor n/2 \rfloor + 1, \ldots, n\}$. The errors are then independent draws from $t(3)$.

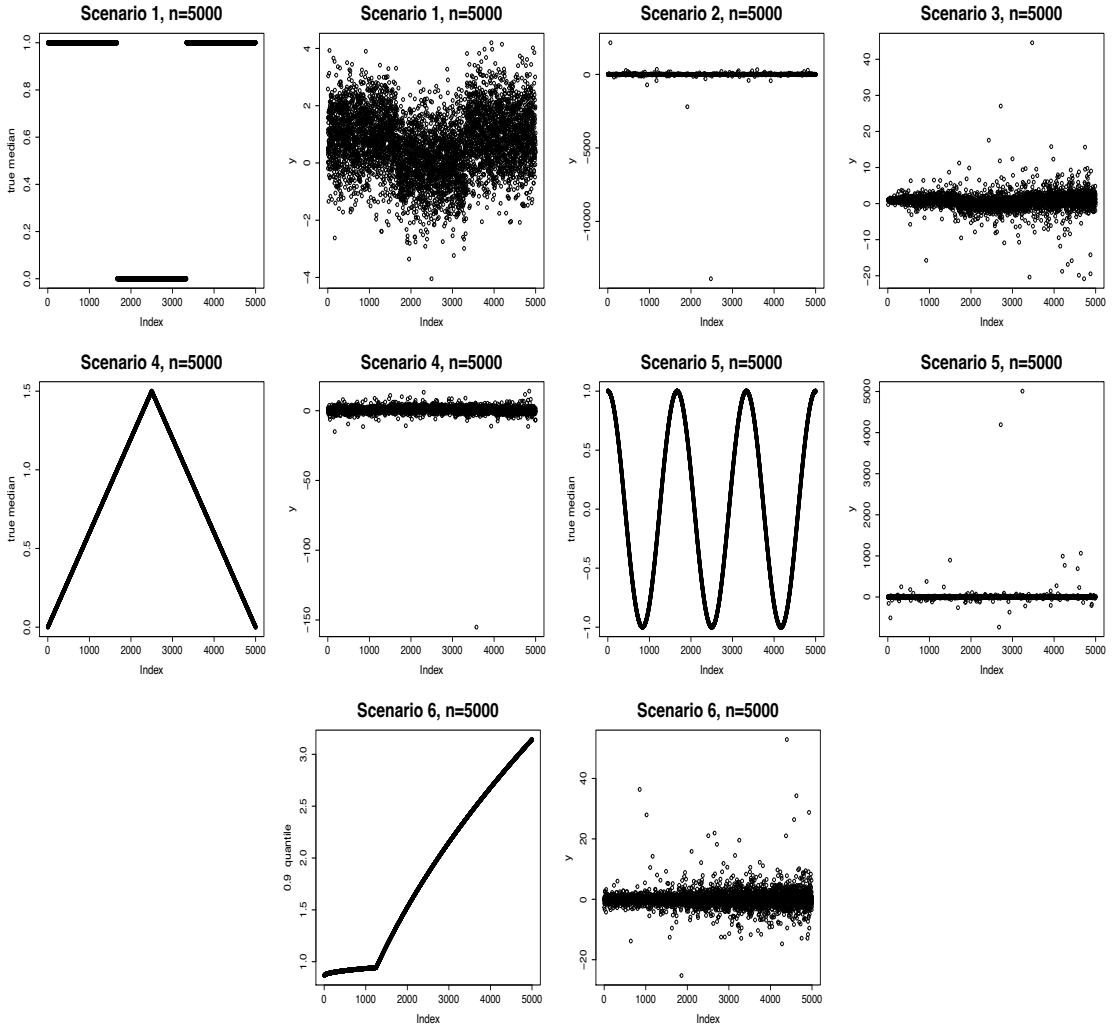

Fig. 1. The top left panel shows θ^*, the true median, for Scenarios 1,2, and 3. The next three panels in the top row correspond to data generated according to Scenarios 1, 2 and 3. Similarly, the middle panels show the true median and an instance of data for Scenarios 4 and 5. Finally, from left to right, the bottom row shows θ^* for Scenario 6 associated with $\tau = 0.9$, and an instance of data generated according to Scenario 6.

Scenario 5. The signal is taken as $\theta_i^* = \cos(6\pi i/n)$ for $i \in \{1, \ldots, n\}$. We then generate $\epsilon_i \sim^{ind} \text{Cauchy}(0, 1)$.

Scenario 6. For our last scenario we generate data as y as

$$y_i = \begin{cases} \frac{v_i(0.25\sqrt{(i/n)}+1.375)}{3} & \text{if } i \in \{1, \ldots, \lfloor n/2 \rfloor\} \\ \frac{v_i(7\sqrt{(i/n)}-2)}{3} & \text{if } i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}, \end{cases}$$

where $v_i \sim^{ind} t(2)$ for $i = 1, \ldots, n$.

Figure 5 illustrates the different scenarios that we consider, There, we can see that some of these scenarios have very heavy tail errors.

The results in Table 1 show that, overall, QTF1 and QTF2 outperforms the competitors. For Scenario 1 which consists of a piecewise constant signal with Gaussian errors, as expected, we can see that TF1 is the best method. For Scenarios 2–3. which have a piecewise constant median but heavy tail errors, the best method is QTF1. For Scenario 5, a model with a smooth median, the best method is quantile splines. Finally, for Scenarios 4 and 6 QTF2 outperforms the competitors, which is reasonable since in these scenarios θ^* is or can be well approximated by a piecewise linear signal.

6. DISCUSSION

We have studied quantile trend filtering in one dimension. Our risk adaptive bounds generalize previous work to quantile setting. The main advantage of our results is that they hold under very general conditions without requiring moment conditions and allowing for heavy-tailed distributions. However, unlike trend filtering with sub-Gaussian errors, our risk bounds are based on $\Delta_n^2(\cdot)$ instead of the mean squared error. While this two metrics are different, when the set over which the minimization is taken is uniformly bounded in $\|\cdot\|_\infty$ and such bound does not grow with n, then convergence rates with $\Delta_n^2(\cdot)$ also hold under the mean squared error.

One natural extension of our work is to consider estimation of multiple quantiles with trend filtering. This can be formulated as follows. Let $\Lambda \subset (0,1)$ be a finite set. Consider the estimator

$$\{\hat{\theta}_C^{(r)}(\tau)\} = \begin{array}{ll} \underset{\{\theta(\tau)\}_{\tau\in\Lambda}\subset R^n}{\arg\min} & \displaystyle\sum_{\tau\in\Lambda}\sum_{i=1}^{n}\rho_\tau(y_i-\theta_i(\tau)), \\ \text{subjec to} & \|D^{(r)}\theta(\tau)\|_1 \le n^{1-r}V(\tau), \ \ \forall\tau\in\Lambda \\ & \theta(\tau)\le\theta(\tau'), \quad \forall\tau<\tau', \quad \tau,\tau'\in\Lambda. \end{array}$$

where $\{V(\tau)\}$ are tuning parameters. Define $\theta_i^*(\tau) = F_{y_i}^{-1}(\tau)$ for all $\tau\in\Lambda$, and $i=1,\dots,n$. If Assumptions 1 holds for each $\theta^*(\tau)$ instead of θ^*, then the proof of Theorem 1 implies that

$$\sum_{\tau\in\Lambda}\Delta_n^2\left\{\theta^*(\tau)-\hat{\theta}_C^{(r)}(\tau)\right\} = O_{\rm pr}\left\{n^{-(2r)/(2r+1)}\right\},$$

provided that $V(\tau) = n^{r-1}\|D^{(r)}\theta^*(\tau)\|_1 = O(1)$. This shows that the upper bound in Theorem 1 also holds for estimation of multiple quantiles simulatenously. Similarly, we can also obtain a version of Theorem 2 for the case of multiple quantiles. However, we are not aware of how to handle the case where the set Λ can have growing size or infinetily many elements. This is left for future work.

As discussed in Sections 3–4, aside from quantile trend filtering in one dimension, our proof technique has implications to other quantile related problems. However, there are some limitations to our work. For instance, one framework where our machinery falls short is in isotonic regression. In such setting, we are not able to extend to quantile regressions the results from **?**.

Finally, we emphasize that when it comes to fast rates of convergence, estimation in the class of piecewise polynomial signals, we have only presented nearly minimax guarantees for the constrained version of quantile trend filtering. One potential way to prove the same for the penalized estimator could be to exploit some of the results from **?**. However, our proof technique would have to be significantly modified and it goes beyond the scope of this paper.

ACKNOWLEDGEMENT

The authors thank Ryan Tibshirani for helpful and stimulating conversations.

REFERENCES

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Alexandre Belloni, Victor Chernozhukov, et al. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Halley L Brantley, Joseph Guinness, and Eric C Chi. Baseline drift estimation for air quality data using quantile trend filtering. *arXiv preprint arXiv:1904.10582*, 2019.

Lawrence D Brown, T Tony Cai, Harrison H Zhou, et al. Robust nonparametric estimation via wavelet median regression. *The Annals of Statistics*, 36(5):2055–2084, 2008.

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2d total variation denoising. *arXiv preprint arXiv:1902.01215*, 2019.

Dennis D Cox. Asymptotics for m-type smoothing splines. *The Annals of Statistics*, pages 530–551, 1983.

Randall L Eubank. *Spline smoothing and nonparametric regression*, volume 90. M. Dekker New York, 1988.

Jianqing Fan, Yingying Fan, and Emre Barut. Adaptive robust variable selection. *Annals of statistics*, 42(1):324, 2014.

Zhou Fan, Leying Guan, et al. Approximate ℓ_0 -penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217–3245, 2018.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *arXiv preprint arXiv:1702.05113*, 2017a.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Spatial adaptation in trend filtering. *arXiv preprint arXiv:1702.05113*, 8, 2017b.

Xuming He and Peide Shi. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journaltitle of Nonparametric Statistics*, 3(3-4):299–308, 1994.

Dorit S Hochbaum and Cheng Lu. A faster algorithm solving a generalization of isotonic median regression and a class of fused lasso problems. *SIAM Journal on Optimization*, 27(4):2563–2596, 2017.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Statistics.*, page 73101, 1964.

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. *Annual Conference on Learning Theory*, 29:1115–1146, 2016.

Nicholas Johnson. A dynamic programming algorithm for the fused lasso and l_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51 (2):339–360, 2009.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Youjuan Li and Ji Zhu. Analysis of array cgh data for cancer studies using fused quantile regression. *Bioinformatics*, 23(18):2470–2476, 2007.

Enno Mammen and Sara van de Geer. Locally apadtive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

Francesco Ortelli and Sara van de Geer. Synthesis and analysis in total variation regularization. *arXiv preprint arXiv:1901.06418*, 2019.

Oscar Hernan Madrid Padilla, James Sharpnack, James G Scott, and Ryan J Tibshirani. The dfs fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:176–1, 2018.

Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M Witten. Adaptive nonparametric regression with the k-nearest neighbour fused lasso. *Biometrika*, 107(2):293–310, 2020.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

Leonid Rudin, Stanley Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *arXiv preprint arXiv:1702.05037*, 2017.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, pages 1–24, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1): 285–323, 2014.

Nicole Tomczak-Jaegermann. Banach-mazur distances and finite-dimensional operator ideals. pitman monographs and surveys in pure and applied mathematics, 38. *Pure and Applied Mathematics*, 38:395, 1989.

Florencio I Utreras. On computing robust splines and applications. *SIAM Journal on Scientific and Statistical Computing*, 2(2):153–163, 1981.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

[*Received on* 2 $January$ 2017. *Editorial decision on* 1 $April$ 2017]