Adaptive Quantile Trend Filtering

Oscar Hernan Madrid Padilla[1] and Sabyasachi Chatterjee[2]

[1]Department of Statistics, University of California, Los Angeles
[2]Department of Statistics, University of Illinois at Urbana-Champaign

July 15, 2020

Abstract

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.

Keywords: 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. 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. This problem arises in many areas of statistics with applications in finance (e.g. Benoit and Van den Poel (2009); Belloni et al. (2019)), marketing (e.g. Coad and Rao (2006); Perlich et al. (2007)), environmental sciences (e.g. Knight and Ackerly (2002); Cade and Noon (2003); Wasko and Sharma (2014)), astronomy (e.g. Sanderson et al. (2006); Pata and Schindler (2015)) among others.

In addition to the measurements $\{y_i\}_{i=1}^n$, we will assume that there is also side information in the form of structural constraints on the parameter θ^*. For instance, one possible structural constraint can be that for most i's the parameters θ_i^* and θ_{i+1}^* are similar to each other. This can be formalized as

$$\sum_{1=1}^{n} |\theta_i^* - \theta_{i+1}^*| \; = \; O(1). \tag{1}$$

Equation (1) holds if θ^* is piecewise constant with a small number of pieces. Moreover, we will also consider models where θ^* can be piecewise linear, or more generally piecewise polynomial. Additionally, our results have implications in high-dimensional regression, and for other shape constrained estimators.

However, for a better flow of this paper, we focus on estimators of the form

$$\hat{\theta}^{(k+1)} = \underset{\theta \in R^n}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\Delta^{(k+1)}\theta\|_1 \right\}, \tag{2}$$

for a tuning parameter $\lambda > 0$. Here, for $k = 0$, we define

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

and for $k > 1$, we define $\Delta^{(k+1)} = \Delta^{(1)}\Delta^{(k)}$, where $\Delta^{(1)}$ is of the appropriate dimension. The case of $k = 0$ in (2) appeared in Li and Zhu (2007) in the context of array CGH data for cancer studies. When $k = 0$ 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 k. We use the convention that $\Delta^{(0)} \in R^{n\times n}$ is the identity matrix.

A related estimator that we will consider is

$$\begin{aligned} \hat{\theta}_C^{(k+1)} = \underset{\theta \in R^n}{\arg\min} \quad & \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \\ \text{subjec to} \quad & \|\Delta^{(k+1)}\theta\|_1 \le \tfrac{V}{n^k}, \end{aligned} \tag{3}$$

where V is a tuning parameter. Thus, $\hat{\theta}_C^{(k+1)}$ is the constrained version of $\hat{\theta}^{(k+1)}$ defined in (2).

1.2 Summary of results

We now list the contributions that we make in this paper. Throughout the paper we denote by $D_n^2 : R^n \to R$ the function given as

$$D_n^2(v) = \frac{1}{n}\sum_{i=1}^{n} \min\left\{|v_i|, v_i^2\right\}, \quad \text{for } v \in R^n, \tag{4}$$

which, up to constants, is a Huber loss, see Huber (1964).

Supposing that the kth order difference of θ^* has total variation with canonical scaling (of order n^{-k}), we show that

$$D_n^2\left\{\theta^* - \hat{\theta}^{(k+1)}\right\} = O_{\text{pr}}\left\{n^{-(2k+2)/(2k+3)} (\log n)^{1/(2k+3)}\right\}. \tag{5}$$

This results holds under general assumptions on the distributions F_{y_i} which include cases such as the Gaussian, Cauchy, t-distribution, and most generic distributions. Notably, (5) implies that $\hat{\theta}^{(k+1)}$ attains minimax rates, up to logarithmic factors, for estimating signals in the class with kth order difference having total variation at most as that of $\Delta^{(k)}\theta^*$. Thus, our results generalize to the quantile regression setting, previously known theory for trend filtering in one dimension under sub-Gaussian data, see Mammen and van de Geer (1997); Tibshirani (2014); Wang et al. (2016).

Under general conditions, we show that the constrained estimator $\hat{\theta}_C^{(k+1)}$ attains the rate $n^{-(2k+2)/(2k+3)}$ in terms of the loss $D_n^2(\cdot)$. This is for an appropriate choice of the tuning parameter V. Thus, for an ideal choice of the tuning parameter, the constrained estimator attains minimax rates for estimating θ^*. This

result is in line with Guntuboyina et al. (2017b) which proved that the same rate under the mean squared error loss is attained by trend filtering under sub-Gaussian errors. However, our result takes one step further in imposing minor assumptions and prooves that the constrained quantile trend filtering estimator is robust to heavy-tailed distributions of the errors.

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

$$D_n^2\left\{\theta^* - \hat{\theta}_C^{(k+1)}\right\} = O_{\text{pr}}\left\{\frac{(s+1)}{n}\log\left(\frac{en}{s+1}\right)\right\}.$$

Thus, when the true quantiles vector is piecewise polynomial, with a minimal spacing condition between change points, the quantile trend filtering estimator attains minimax rate. We refer the reader to Guntuboyina et al. (2017b) for the corresponding result for trend filtering under the assumption of Gaussian errors.

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 $\Delta^{(k+1)}$ in (3) 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 $D_n^2(\cdot)$ 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).

A notable difference between our work and the current literature on trend filtering has to do with the loss function for which we provide upper bounds for different estimators. Specifically, we consider $D_n^2(\cdot)$ whereas previous work consisted of upper bounds on the mean squared error. This difference along with the quantile trend filtering nature allows us to provide results that hold under very general assumptions, which go beyond sub-Gaussian error models.

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

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

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.

Other models for nonparametric quantile regression have also been studied in the literature. Kim et al. (2007) considered a linear model with random coefficients. Horowitz and Lee (2005) focused on additive models and provided convergence rates that match with those in He and Shi (1994), when the different

components are r times continuously differentiable. He et al. (1998) introduced quantile smoothing splines in two dimensions. Takeuchi et al. (2006); Li et al. (2007) and Zhang et al. (2016) developed reproducing kernel Hilbert spaces methods. Yu (1999); Liu and Wu (2009), and Racine and Li (2017) examined kernel based approaches. Pratesi et al. (2009) proposed a penalized splines procedure. Fan and Liu (2016) considered a semi parametric quantile model. Meinshausen (2006) proposed the quantile random forest. Zhao et al. (2017) focused on a version of the celebrated single index model for quantile regression.

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 (2). We recall that trend filtering regression is the solution to

$$\underset{\theta \in R^n}{\text{minimize}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 \, + \, \lambda \| \Delta^{(k+1)} \theta \|_1 \right\}, \tag{6}$$

where $\lambda > 0$ is a tuning parameter.

Hence $\hat{\theta}^{(k+1)}$ as defined in (2) is the quantile version of (6). The case of trend filtering with $k = 0$ was introduced in Rudin et al. (1992) for image denoising applications. In the statistics literature, trend filtering was introduced as locally adaptive regression splines in Mammen and van de Geer (1997). Tibshirani et al. (2005) then defined difference based estimators of the form (6) that have led to further development of methods for trend filtering such as in Kim et al. (2009); Tibshirani and Taylor (2011, 2012).

On the computational front, it is known that Problem (6) with $k = 0$ 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 (2) with $k = 0$, can be found in $O(n \log n)$ operations. For $k > 0$, Brantley et al. (2019) proposed an alternating direction method of multipliers (ADMM) based algorithm for computing quantile trend filtering estimators.

As for the theoretical understanding of trend filtering, Mammen and van de Geer (1997) and Tibshirani (2014) showed that if $\| \Delta^{(k+1)} \theta^* \|_1 = O(n^{-k})$ then

$$\frac{1}{n} \sum_{i=1}^{n} (\theta_i^* - \hat{\theta}_i)^2 = O_{\text{pr}} \left(n^{-(2k+2)/(2k+3)} \right),$$

with $\hat{\theta}$ the solution to (6) with an appropriate tuning parameter, and under the assumption that the errors are independent and sub-Gaussian. A similar result also holds for the constrained version of (6), see Guntuboyina et al. (2017b).

Furthermore, when $S = \{j \, : \, (\Delta^{(k+1)} \theta^*)_j \neq 0\}$, under the assumption of independent and identically distributed Gaussian errors, and a minimal spacing condition, Guntuboyina et al. (2017b) showed that

$$\frac{1}{n} \sum_{i=1}^{n} (\theta_i^* - \hat{\theta}_i)^2 = O_{\text{pr}} \left\{ \frac{(s+1)}{n} \log \left(\frac{en}{s+1} \right) \right\}, \tag{7}$$

where $\hat{\theta}$ is the kth order constrained trend filtering estimator, and $s = |S|$. A similar upper bound, with an extra logarithmic factor, was proven in Lin et al. (2017) for the penalized fused lasso estimator. It is an open question whether or not an upper bound as (7) holds for the penalized trend filtering estimator (6) with $k > 0$.

Aside from one-dimensional nonparametric regression, the statistical properties of trend filtering have been extensively studied in other contexts. In two-dimensional grid graphs, Sadhanala et al. (2016) proved

lower bounds for estimation under squared error loss. Hutter and Rigollet (2016) showed that penalized total variation denoising attains minimax rates under canonical scaling. A similar result was recently stated in Chatterjee and Goswami (2019) for the corresponding total variation constrained estimator.

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 Constrained estimator

2.1 Bounded variation class of signals

We now study the constrained quantile trend filtering estimator as defined in (3). 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 $k \in \{0,1,2,\dots\}$ are fixed.

Our first assumption stated next simply requires that θ^*, the vector of τ-quantiles, has kth discrete derivative which has bounded variation. We also require that the measurements y_i are independent.

Assumption 1. *We write $\theta_i^* = F_{y_i}^{-1}(\tau)$ for $i = 1,\dots,n$, and $V^* := n^k \|\Delta^{(k+1)}\theta^*\|_1$ satisfies $V^* = O(1)$. Here F_{y_i} is cumulative distribution function of y_i for $i = 1,\dots,n$. Also, $y_1,\dots,y_n$ are assumed to be independent.*

Throughout the paper, the quantities $\epsilon_i = y_i - \theta_i^*$, $i = 1,\dots,n$ are referred as the errors.

Clearly, $\theta^* \in K$, where

$$K = \left\{ \theta \in R^n \,:\, \|\Delta^{(k+1)}\theta\|_1 \leq \frac{V^*}{n^k} \right\}. \tag{8}$$

Notice that when $k = 0$ the set K becomes the class of bounded variation signals. The choice of $k > 0$ corresponds to higher order bounded variation classes, see Tibshirani (2014) for an overview.

Our next condition requires that for each y_i, there exists a neighborhood around the quantile such that within such neighborhood the probability density function of y_i is bounded by below. A related assumption appeared as D.1 in Belloni et al. (2011), and Condition 2 in He and Shi (1994).

Assumption 2. *There exists a constant $L > 0$ such that for $\delta \in R^n$ satisfying $\|\delta\|_\infty \leq L$ we have that*

$$\min_{i=1,\dots,n} f_{y_i}(\theta_i^* + \delta_i) \geq \underline{f},$$

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

Notice that Assumption 2 will hold for most common distributions such as Normal, Cauchy, and t-distribution.

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

Theorem 1. *Under Assumptions 1–2, and V in (3) is chosen as $V = V^*$ then*

$$D_n^2\left\{\theta^* - \hat{\theta}_C^{(k+1)}\right\} = O_{\text{pr}}\left\{n^{-(2k+2)/(2k+3)}\right\}.$$

Notably, 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 $D_n^2(\cdot)$ defined in (4). Moreover, Theorem 1 shows that quantile trend filtering is a robust estimator.

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^*.

Finally, we conclude with a remark regarding estimation of multiple quantiles simultaneously.

Remark 1. *Let $\Lambda \subset (0, 1)$ be a finite set. Consider the estimator*

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

where $\{V(\tau)\}$ are tuning parameters. Define $\theta_i^*(\tau) = F_{y_i}^{-1}(\tau)$ for all $\tau \in \Lambda$. If Assumptions 1–2 hold for each $\theta^*(\tau)$ instead of θ^*, then the proof of Theorem 1 implies that

$$\sum_{\tau\in\Lambda} D_n^2\left\{\theta^*(\tau) - \hat{\theta}_C^{(k+1)}(\tau)\right\} = O_{\text{pr}}\left\{n^{-(2k+2)/(2k+3)}\right\},$$

provided that $V(\tau) = n^k\|\Delta^{(k+1)}\theta^*(\tau)\|_1$. This shows that the upper bound in Theorem 1 also holds for estimation of multiple quantiles simulatenously. A similar phenomenon also holds for all other constrained estimators results that we provide. Hence, for simplicity we focus on the analysis of single quantiles.

2.2 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 $s = \|\Delta^{(k+1)}\theta^*\|_0$, let $S = \{j : (\Delta^{(k+1)}\theta^*)_j \neq 0\}$ 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 $j_0 = 1$, $j_{s+1} = n - k - 1$, with $j_1, \ldots, j_s$ are the elements of S. Under Assumptions 1–2, and V in Problem (3) chosen as $V = V^*$, we have that

$$D_n^2\left\{\theta^* - \hat{\theta}_C^{(k+1)}\right\} = O_{\text{pr}}\left\{\frac{(s+1)}{n}\log\left(\frac{en}{s+1}\right)\right\}.$$

For the case of median regression with sub-Gaussian errors, 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.

3 Penalized trend filtering estimator

We now provide theoretical guarantees for the penalized quantile trend filtering estimator (6). 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 (2) and (3) 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 trend filtering.

Theorem 3. *Suppose that Assumptions 1–2. Then there exists a choice of λ for Problem 2 satisfying*

$$\lambda = \Theta\left\{n^{(2k+1)/(2k+3)} (\log n)^{1/(2k+3)} \|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}\right\},$$

such that

$$D_n^2\left\{\theta^* - \hat{\theta}^{(k+1)}\right\} = O_{\text{pr}}\left\{n^{-(2k+2)/(2k+3)} (\log n)^{1/(2k+3)}\right\}.$$

Theorem 3 shows that, under the loss $D_n^2(\cdot)$, 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 4.1 combined with a careful construction of restricted set in the spirit of Belloni et al. (2011), and exploiting results from Wang et al. (2016) and Guntuboyina et al. (2017b).

4 Other applications

4.1 Proof ideas

Before providing other applications of our proof techniques, 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 (2) and (3) 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 $D_n^2(\cdot)$ and the quantile population loss.

Lemma 4. *Suppose that $\theta_i^* = F_{y_i}^{-1}(\tau)$ and Assumptions 2 holds. Then there exists a constant $C > 0$ such that for all $\delta \in R^n$, we have*

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

for some positive constant C.

Lemma 7 does not depend on trend filtering and in fact can be used with other shape constrained estimators. Two different ways that we use Lemma 7 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 7 and the optimality of $\hat{\theta}$, it can be proven that

$$E\left\{D_n^2\left(\theta^* - \hat{\theta}\right)\right\} \leq E\left\{\frac{M_n(\hat{\theta})}{C\,n}\right\} \leq \frac{2}{n} E\left[\sup_{v \in K}\left\{M_n(v) - \hat{M}_n(v)\right\}\right], \tag{9}$$

provided that $\theta^* \in K$, see the Appendix for details. Hence, in order to give an upper bound for $E\{D_n^2(\theta^* - \hat{\theta})\}$, it is enough to provide an upper bound for the right most term in (9). We do that in the Appendix 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

$$E\left(\sup_{v \in K} \sum_{i=1}^{n} \xi_i (v_i - \theta_i^*)\right), \tag{10}$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables. The quantity (10) 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 Appendix reduce the problem of controlling $D_n^2(\theta^* - \hat{\theta})$ to that of deriving an upper bound on

$$E\left(\sup_{v \in K\,:\, D_n^2(v) \leq \eta^2} \sum_{i=1}^{n} \xi_i (v_i - \theta_i^*)\right), \tag{11}$$

for a carefully chosen $\eta > 0$. To give an upper bound for (11), we exploit results from Guntuboyina et al. (2017b) which controls a similar quantity obtained by replacing $D_n^2(\cdot)$ with the mean squared error.

4.2 2D fused lasso

Total variation denoising in two 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 two dimensions. In particular, we will exploit ideas from Section 4.1 combined with results from Chatterjee and Goswami (2019) to obtain an upper bound, under the loss $D_n^2(\cdot)$.

More precisely, we consider the $n^{1/2} \times n^{1/2}$ two-dimensional grid $G_{2D} = (\{1, \ldots, n\}, E_n)$. For a signal $\theta \in R^n$ we define its total variation along G_{2D} as

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

where ∇ is the usual edge vertex incidence matrix of size $2n^{1/2}(n^{1/2} - 1)\,\mathrm{x}\,n$ of the graph G_{2D}. With this notation, we consider the estimator

$$\hat{\theta} = \underset{\theta \in K}{\arg\min} \left\{\sum_{i=1}^{n} \rho_\tau(y_i - \theta_i)\right\}, \tag{12}$$

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

Next we state an assumption which is the analogous of Assumption 1 for the context of two-dimensional total variation denoising.

Assumption 3. *Writing $\theta_i^* = F_{y_i}^{-1}(\tau)$ for $i = 1, \ldots, n$ and $V^* := \|\nabla\theta^*\|_1$, we require that $V^* = O(n^{1/2})$. Here, F_{y_i} is cumulative distribution function of y_i for $i = 1, \ldots, n$. We assume $y_1, \ldots, y_n$ to be independent.*

Notice that Assumption 3 requires that true signal has total variation along G_{2D} which is of order $O(n^{1/2})$. This is a standard setting for 2D denoising, in fact Sadhanala and Tibshirani (2017) refers to this scaling of the total variation as "canonical".

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

Theorem 5. *Suppose that Assumptions 2–3 hold with $\theta^* \in R^n$ the vector of τ-quantiles of y. If V in (12) is chosen as $V = \|\nabla\theta^*\|_1$ then*

$$D_n^2(\hat{\theta} - \theta^*) \;=\; O_{\mathrm{pr}}\left(\frac{\log n}{n^{1/2}}\right),$$

where $\hat{\theta}$ is the estimator defined in (12).

Theorem 5 shows that quantile fused lasso in two dimensions attains minimax rates under the loss $D_n^2(\cdot)$ and the canonical scaling. 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. However, unlike previous results, Theorem 5 holds with a different metric than the mean squared error and it holds under more general settings than sub-Gaussian errors.

4.3 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

$$F_{y_i}^{-1}(\tau) = x_i^\top \theta^*, \tag{13}$$

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

$$\hat{\theta} \;=\; \operatorname*{arg\,min}_{\theta \in K} \left\{ \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \theta) \right\}, \tag{14}$$

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 4. *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*

$$\min_{i=1,\ldots,n} f_{y_i}(x_i^\top \theta^* + u) \geq \underline{f},$$

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 5. *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* $\max_{j=1,\dots,p} \|X_{\cdot,j}\| \leq n^{1/2}$.

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

Theorem 6. *Suppose that Assumptions 4–5 hold. Then there exists a constant $C > 0$ such that*

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

where $\hat{\theta}$ is the estimator defined in (14).

We emphasise that Theorem 6 holds without conditions on the eigenvalues of the design matrix. This is a crucial difference from previous work in the literature that relies on the 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 $D_n^2(\cdot)$ 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 6.

5 Expriments

We now proceed to illustrate with simulations the empirical performance of quantile trend filtering. As benchmark methods, we consider the fused lasso (6) with $k = 0$ (FL), trend filtering of order 1 (TF1), and quantile splines (QS) using the R package "fields". For quantile trend filtering we consider the penalized estimator (2) with choices $k = 0$ and $k = 1$ which we denote as QFL and QTF1 respectively. These are implemented in R via ADMM, similarly to Brantley et al. (2019). We 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

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

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, $i = 1,\dots,n$. 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 $i \in \{1,\dots,n\} \cup \{n - 2\lfloor n/3\rfloor + 1,\dots,n\}$ 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 $\text{Cauchy}(0,1)$.

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	QFL	QTF1	QS	FL	TF1
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	*	*

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 $t(3)$.

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

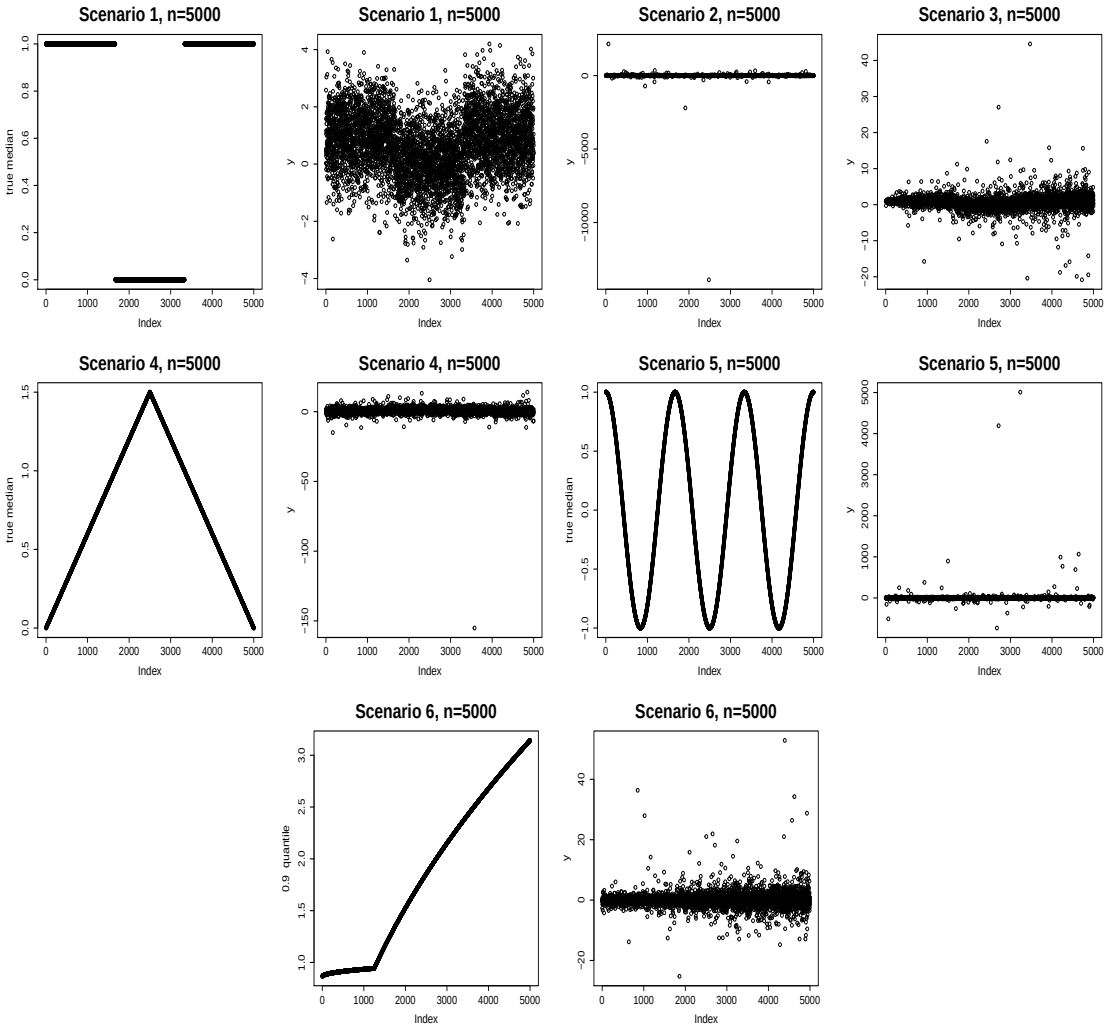

Figure 1: The top left panel shows θ^*, the true median, for Scenarios 1,2, and 3. The next three panels in the top row correspond to data generated according to Scenarios 1, 2 and 3. Similarly, the middle panels show the true median and an instance of data for Scenarios 4 and 5. Finally, from left to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

Acknowledgement

The authors thank Ryan Tibshirani for helpful and stimulating conversations.

A General lemmas

Before presenting our lemmas, we recall a well known fact that will be used in our proofs. For a set $K \subset R^n$, the Rademacher complexity of K is defined as

$$w_R(K) = E\left(\sup_{v \in K} \sum_{i=1}^{n} \xi_i v_i\right),$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables. Similarly, the Gaussian width of K is defined as

$$w_G(K) = E\left(\sup_{v \in K} \sum_{i=1}^{n} z_i v_i\right),$$

where $z_1, \ldots, z_n$ are independent standard normal random variables.

With this notation, it is known that (see Page 151 in Wainwright (2019) and also Tomczak-Jaegermann (1989); Bartlett and Mendelson (2002))

$$w_R(K) \leq \left(\frac{\pi}{2}\right)^{1/2} w_G(K). \tag{16}$$

We now prove lemmas that hold for general shape constrained problems with quantile regression. Throughout this section K is a subset of R^n.

Definition 1. 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 $M_{n,i}(\theta_i) = \mathbb{E}(\rho_\tau(z_i - \theta_i) - \rho_\tau(z_i - \theta_i^*))$ where $z \in R^n$ is an independent copy of y, the population loss becomes

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

With the notation from Definition 1, we consider the estimator

$$\hat{\theta} = \begin{array}{ll} \underset{\theta \in R^n}{\arg\min} & \hat{M}_n(\theta) \\ \text{subject to} & \theta \in K, \end{array} \tag{17}$$

and $\theta^* \in \arg\min_{\theta \in R^n} M_n(\theta)$.

Lemma 7. *With the notation from before,*

$$M_n(\hat{\theta}) \leq 2 \sup_{v \in K} \left\{ M_n(v) - \hat{M}_n(v) \right\}.$$

Proof.

$$\begin{aligned} M_n(\hat{\theta}) &= M_n(\hat{\theta}) - \hat{M}_n(\hat{\theta}) + \hat{M}_n(\hat{\theta}) - \hat{M}_n(\theta^*) + \hat{M}_n(\theta^*) - M_n(\theta^*) \\ &\leq M_n(\hat{\theta}) - \hat{M}_n(\hat{\theta}) + \hat{M}_n(\theta^*) - M_n(\theta^*) \\ &\leq 2 \sup_{v \in K} \left\{ M_n(v) - \hat{M}_n(v) \right\}, \end{aligned}$$

where the first inequality follows since $\hat{M}_n(\hat{\theta}) \leq \hat{M}_n(\theta^*)$. □

Lemma 8. *Suppose that Assumption 2 holds. Then there exists a constant c_τ such that for all $\delta \in R^n$, we have*

$$M_n(\theta^* + \delta) - M_n(\theta^*) \geq D^2(\delta) := \sum_{i=1}^{n} d(\delta_i)$$

where

$$d(x) = \begin{cases} c_\tau x^2 & \text{if } |x| \le L, \\ c_\tau |x| & \text{if } |x| > L, \end{cases} \tag{18}$$

for some constant $c_\tau > 0$ that scales like $\min\{\underline{f}L, \underline{f}\}$.

Proof. Suppose that $|\delta_i| \le L$, then as in Equation B.3 of Belloni et al. (2011),

$$\begin{aligned} M_{n,i}(\theta_i^* + \delta_i) - M_{n,i}(\theta_i^*) &= \int_0^{\delta_i} [F_{y_i}(\theta_i^* + z) - F_{y_i}(\theta_i^*)]\, dz \\ &= \int_0^{\delta_i} f_{y_i}(u(\theta_i^*, z)) z dz \\ &\ge \frac{\delta_i^2 \underline{f}}{2}, \end{aligned}$$

where $u(\theta_i^*, z)$ is a point between $\theta_i^* + z$ and θ_i^*, and the inequality follows from Assumption 2.

Suppose now that $\delta_i > L$. Then

$$\begin{aligned} M_{n,i}(\theta_i^* + \delta_i) - M_{n,i}(\theta_i^*) &= \int_0^{\delta_i} \{F_{y_i}(\theta_i^* + z) - F_{y_i}(\theta_i^*)\}\, dz \\ &\ge \int_{L/2}^{\delta_i} \{F_{y_i}(\theta_i^* + z) - F_{y_i}(\theta_i^*)\}\, dz \\ &\ge \int_{L/2}^{\delta_i} \{F_{y_i}(\theta_i^* + L/2) - F_{y_i}(\theta_i^*)\}\, dz \\ &= \left(\delta_i - \frac{L}{2}\right) \{F_{y_i}(\theta_i^* + L/2) - F_{y_i}(\theta_i^*)\} \\ &\ge \tfrac{\delta_i}{2} \tfrac{L\underline{f}}{2} \\ &=: |\delta_i| c_\tau, \end{aligned}$$

where the first two inequalities follow because F_{y_i} is monotone, and the third inequality by the mean value theorem and Assumption 2.

The case $\delta_i < -L$ can be handled similarly. The conclusion follows combining the three different cases. □

Next we state to conditions that generalize Assumptions 1–2 in the paper.

Assumption 6. *We write $\theta_i^* = F_{y_i}^{-1}(\tau)$, and assume that $\theta^* \in K$. Here F_{y_i} is cumulative distribution function of y_i for $i = 1, \ldots, n$. We require $y_1, \ldots, y_n$ to be independent.*

Assumption 7. *There exists a constant L such that for $\delta \in R^n$ satisfying $\|\delta\|_\infty \le L$ we have that*

$$\min_{i=1,\ldots,n} f_{y_i}(\theta_i^* + \delta_i) \ge \underline{f},$$

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

Corollary 9. *Under Assumptions 6–7,*

$$E\left\{D^2(\hat{\theta} - \theta^*)\right\} \le 2E\left[\sup_{v \in K} \left\{M_n(v) - \hat{M}_n(v)\right\}\right]. \tag{19}$$

Next, we proceed to bound the right hand side of Equation 19.

Lemma 10. *Under Assumption 6, we have that*

$$E\left\{\sup_{v\in K}\,(M_n-\hat{M}_n)(v)\right\}\leq 2\,E\left(\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i\hat{M}_{n,i}(v_i)\right),$$

where $\xi_1,\ldots,\xi_n$ are independent Rademacher variables independent of $\{y_i\}_{i=1}^n$.

Proof. Let $\tilde{y}_1,\ldots,\tilde{y}_n$ be an independent and identically distributed copy of $y_1,\ldots,y_n$, and let $\tilde{M}_{n,i}$ the version of $\hat{M}_{n,i}$ corresponding to $\tilde{y}_1,\ldots,\tilde{y}_n$. Then,

$$E\left(\sup_{v\in K}\,\sum_{i=1}^{n}[E\{\hat{M}_{n,i}(v_i)\}-\hat{M}_{n,i}(v_i)]\right) \;=\; E\left(\sup_{v\in K}\,\sum_{i=1}^{n}[E\{\tilde{M}_{n,i}(v_i)\}-\hat{M}_{n,i}(v_i)]\right).$$

Condition on $y_1,\ldots,y_n$ and let

$$X_v=\sum_{i=1}^{n}\left\{\tilde{M}_{n,i}(v_i)-\hat{M}_{n,i}(v_i))\right\}.$$

Then

$$\sup_{v\in K}\,E_{\tilde{y}_1,\ldots,\tilde{y}_n|y_1,\ldots,y_n}X_v\;\leq\;E_{\tilde{y}_1,\ldots,\tilde{y}_n|y_1,\ldots,y_n}\sup_{v\in K}\,X_v.$$

We can take the expected value with respect to $y_1,\ldots,y_n$ to get

$$\begin{aligned}
E\left[\sup_{v\in K}\,\sum_{i=1}^{n}\left\{M_{n,i}(v_i)-\hat{M}_{n,i}(v_i)\right\}\right] \;&\leq\; E\left[\sup_{v\in K}\,\sum_{i=1}^{n}\left\{\tilde{M}_{n,i}(v_i)-\hat{M}_{n,i}(v_i)\right\}\right]\\
&=\; E\left[\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i\left\{\tilde{M}_{n,i}(v_i)-\hat{M}_{n,i}(v_i)\right\}\right]\\
&\leq\; E\left\{\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i\tilde{M}_{n,i}(v_i)\right\}+\\
&\quad\; E\left\{\sup_{v\in K}\,\sum_{i=1}^{n}-\xi_i\hat{M}_{n,i}(v_i)\right\}\\
&=\; 2E\left\{\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i\hat{M}_{n,i}(v_i)\right\},
\end{aligned}$$

where the first equality follows because $(\xi_1(\tilde{M}_{n,1}(v_1)-\hat{M}_{n,1}(v_1)),\ldots,\xi_n(\tilde{M}_{n,n}(v_n)-\hat{M}_{n,n}(v_n)))$ and $(\tilde{M}_{n,1}(v_1)-\hat{M}_{n,1}(v_1),\ldots,\tilde{M}_{n,n}(v_n)-\hat{M}_{n,n}(v_n))$ have the same distribution. The second equality follows because

$-\xi_1,\ldots,-\xi_n$ are also independent Rademacher variables.

□

Lemma 11. *Suppose that Assumption 6 holds. Then*

$$E\left\{\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i\hat{M}_{n,i}(v_i)\right\}\;\leq\;E\left\{\sup_{v\in K}\,\sum_{i=1}^{n}\xi_i(v_i-\theta_i^*)\right\}.$$

Proof. Conditioning on $y_1,\ldots,y_n$, we notice that the functions $x\to M_{n,i}(\theta_i^*+x)$ are 1-Lipschitz continuous. The claim follows from Theorem 4.12 in Ledoux and Talagrand (2013). □

The following theorem can be used for proving upper bounds for general constraint estimators as in (17) when the set K is compact.

Theorem 12. *Under Assumption 6, we have that*

$$E\left\{M_n(\hat{\theta})\right\} \leq 4\,E\left\{\sup_{v\in K}\sum_{i=1}^{n}\xi_i(v_i-\theta_i^*)\right\}.$$

If in addition (2) holds, then

$$E\left\{D^2(\hat{\theta}-\theta^*)\right\} \leq 4\,E\left\{\sup_{v\in K}\sum_{i=1}^{n}\xi_i(v_i-\theta_i^*)\right\}.$$

Proof. This follows from Lemmas 7–11. □

B Theorem 1

B.1 Auxiliary lemmas for Proof of Theorem 1

To obtain the slow rates for trend filtering, we first bound

$$E\left\{\sup_{v\in K\,:\,D^2(v-\theta^*)\leq t^2}(M_n-\hat{M}_n)(v)\right\}$$

for all t. But as in the proof Lemmas 10 and 11, we have

$$E\left\{\sup_{v\in K\,:\,D^2(v-\theta^*)\leq t^2}(M_n-\hat{M}_n)(v)\right\} \leq \phi(t) := 2E\left\{\sup_{v\in K\,:\,D^2(v-\theta^*)\leq t^2}\sum_{i=1}^{n}\xi_i(v_i-\theta_i^*)\right\}. \tag{20}$$

Therefore, we proceed to bound $\phi(t)$ when $K=\{\theta\,:\,\|\Delta^{(k+1)}\theta\|_1\leq V^*n^{-k}\}$, where $V^*=n^k\|\Delta^{(k+1)}\theta^*\|_1$. Furthermore, we denote by $\mathcal{R}=\text{row}\{\Delta^{(k+1)}\}$ and by $\mathcal{R}^\perp$ the orthogonal complement of $\mathcal{R}$. We denote by $P_{\mathcal{R}}$ and $P_{\mathcal{R}^\perp}$ the orthogonal projections onto $\mathcal{R}$ and $\mathcal{R}^\perp$ respectively.

Lemma 13. *Let $\delta\in R^n$ with $D^2(\delta)\leq t^2$. Then*

$$\|P_{\mathcal{R}^\perp}\delta\|_\infty \leq \gamma(t,n) := 2(k+1)\left(\frac{t}{n^{1/2}}+\frac{t^2}{n}\right).$$

Proof. Let $v_1,\ldots,v_{k+1}$, an orthonormal basis of $\mathcal{R}^\perp$, which as in the proof Corollary 7 from Wang et al. (2016) can be taken to satsify the incoherence condition

$$\|v_j\|_\infty \leq \left(\frac{2}{n}\right)^{1/2}.$$

Then

$$P_{\mathcal{R}^\perp}\delta=\sum_{j=1}^{k+1}\delta^\top v_j v_j.$$

Hence,

$$|(P_{\mathcal{R}^\perp}\delta)_i| \leq (k+1)\left(\max_{j=1,\ldots,k+1}\|v_j\|_\infty\right)\left(\max_{j=1,\ldots,k+1}|\delta^\top v_j|\right) \leq (k+1)\left(\frac{2}{n}\right)^{1/2}\left(\max_{j=1,\ldots,k+1}|\delta^\top v_j|\right). \tag{21}$$

Now, for $j \in \{1, \ldots, k+1\}$, we have,

$$\begin{aligned} |\delta^\top v_j| &\leq \sum_{i=1}^n |\delta_i||v_{j,i}| \\ &= \sum_{i=1}^n |\delta_i||v_{j,i}|1_{\{|\delta_i|>L\}} + \sum_{i=1}^n |\delta_i||v_{j,i}|1_{\{|\delta_i|\leq L\}} \\ &\leq \|v_j\|_\infty \sum_{i=1}^n |\delta_i|1_{\{|\delta_i|>L\}} + \|v_j\| \left(\sum_{i=1}^n \delta_i^2 1_{\{|\delta_i|\leq L\}}\right)^{1/2} \\ &\leq \left(\tfrac{2}{n}\right)^{1/2} t^2 + t, \end{aligned} \tag{22}$$

where the first inequality follows from the triangle inequality, the second from Hölder and Cauchy–Schwarz inequalities, and the last by the definition of $D^2(\cdot)$. The claim follows combining (21) with (22). □

Lemma 14. *Let $\delta \in \mathbb{R}^n$ with $D^2(\delta) \leq t^2$. Then,*

$$D^2(P_{\mathcal{R}}\delta) \leq h(t,n) := 2\max\left\{L, \frac{1}{L}\right\}\left\{t^2 + 2t^2\gamma(t,n) + 16n(k+1)^2\left(\frac{t^2}{n} + \frac{t^4}{n^2}\right)\right\}.$$

Proof. Set $\tilde{\delta} = P_{\mathcal{R}}\delta$. By Lemma 13 we have that $\|\tilde{\delta} - \delta\|_\infty \leq \gamma(t,n)$. Then

$$\begin{aligned} D^2(\tilde{\delta}) &= \sum_{i=1}^n |\tilde{\delta}_i|1_{\{|\tilde{\delta}_i|>L\}} + \sum_{i=1}^n \tilde{\delta}_i^2 1_{\{|\tilde{\delta}_i|\leq L\}} \\ &\leq \sum_{i=1}^n |\tilde{\delta}_i|1_{\{|\delta_i|>L+2\gamma(t,n)\}} + \sum_{i=1}^n \tilde{\delta}_i^2 1_{\{|\delta_i|\leq L-2\gamma(t,n)\}} \\ &\quad + \max\{L, L^2\}\sum_{i=1}^n \min\left\{\frac{|\tilde{\delta}_i|}{L}, \frac{\tilde{\delta}_i^2}{L^2}\right\} 1_{\{|\delta_i|\in(1-2\gamma(t,n),1+2\gamma(t,n))\}} \\ &\leq \max\{L, L^2\}\sum_{i=1}^n \frac{|\tilde{\delta}_i|}{L}1_{\{|\delta_i|>L\}} + \max\{L, L^2\}\sum_{i=1}^n \frac{\tilde{\delta}_i^2}{L^2}1_{\{|\delta_i|\leq L\}} \end{aligned}$$

and so

$$\begin{aligned} D^2(\tilde{\delta}) &\leq \max\{L, L^2\}\sum_{i=1}^n \frac{|\delta_i| + \gamma(t,n)}{L}1_{\{|\delta_i|>L\}} + \max\{L, L^2\}\sum_{i=1}^n \frac{2\delta_i^2 + 2\{\gamma(t,n)\}^2}{L^2}1_{\{|\delta_i|\leq L\}} \\ &\leq \tfrac{2\max\{L,L^2\}}{\min\{L,L^2\}}\left[t^2 + 2t^2\gamma(t,n) + 2n\{\gamma(t,n)\}^2\right] \\ &\leq 2\max\{L, \tfrac{1}{L}\}\left\{t^2 + 2t^2\gamma(t,n) + 16n(k+1)^2\left(\tfrac{t^2}{n} + \tfrac{t^4}{n^2}\right)\right\}. \end{aligned}$$

□

Lemma 15. *Let $\delta \in R^n$ with $\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k}$. Then there exists a constant $\tilde{C}_k$ depending on k such that $\|P_{\mathcal{R}}\delta\|_\infty \leq \tilde{C}_k V^*$.*

Proof. This follows since

$$\|P_{\mathcal{R}}\delta\|_\infty = \|\left\{\Delta^{(k+1)}\right\}^+ \Delta^{(k+1)}\delta\|_\infty \leq \|\left\{\Delta^{(k+1)}\right\}^+\|_\infty \|\Delta^{(k+1)}\delta\|_1 \leq O\{n^k\|\Delta^{(k+1)}\delta\|_1\},$$

where the last equality follows as in the proof of Corollary 7 from Wang et al. (2016) . □

Lemma 16. *Let $\delta \in R^n$, $\tilde{\delta} = P_{\mathcal{R}}\delta$ and suppose that $D^2(\delta) \leq t^2$, and $\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k}$. Then*

$$\|\tilde{\delta}\| \leq m(n,t) := \max\left\{\left(\tilde{C}_k V^*\right)^{1/2}, 1\right\} 2^{1/2} \max\left\{L^{1/2}, \tfrac{1}{L^{1/2}}\right\}\left(\left[1 + \{2\gamma(t,n)\}^{1/2} + 4(k+1)\right] t + \tfrac{4(k+1)t^2}{n^{1/2}}\right).$$

Proof. Notice that

$$\begin{aligned}\|\tilde{\delta}\|^2 &= \sum_{i=1}^n \tilde{\delta}_i^2 \\ &\leq \sum_{i=1}^n \tilde{\delta}_i^2 1_{\{|\tilde{\delta}_i| \leq L\}} + \sum_{i=1}^n \tilde{\delta}_i^2 1_{\{|\tilde{\delta}_i| > L\}} \\ &\leq \max\{\|\tilde{\delta}\|_\infty, 1\} D^2(\tilde{\delta}),\end{aligned}$$

and the claim follows from Lemmas 14–15. □

Lemma 17. *Under Assumptions 1–2, we have that*

$$\begin{aligned}E\left\{\sup_{\delta\,:\,\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k}\ D^2(\delta)\leq t^2} \sum_{i=1}^n \xi_i\delta_i\right\} \leq\ & C(k+1)^2\left\{\left(\tfrac{2}{n}\right)^{1/2} t^2 + t\right\} + C_k m(t,n)\left\{\tfrac{n^{1/2}V^*}{m(t,n)}\right\}^{1/(2+2k)} \\ & + C_k m(t,n)\{\log(en)\}^{1/2},\end{aligned}$$

for some positive constants C and C_k, and where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables.

Proof. Notice that

$$\begin{aligned}E\left\{\sup_{\delta\,:\,\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k},\ D^2(\delta)\leq t^2} \sum_{i=1}^n \xi_i\delta_i\right\} \leq\ & E\left\{\sup_{\delta\,:\,\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k},\ D^2(\delta)\leq t^2} \xi^\top P_{\mathcal{R}^\perp}\delta\right\} \\ & + E\left\{\sup_{\delta\,:\,\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k}\ D^2(\delta)\leq t^2} \xi^\top P_{\mathcal{R}}\delta\right\} \\ =:\ & T_1 + T_2.\end{aligned}$$

We now proceed to bound T_1 and T_2. To bound T_2, we observe that

$$T_2 \leq E\left\{\sup_{\delta\,:\,\|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k},\ \|\delta\|\leq m(t,n)} \xi^\top\delta\right\}, \tag{23}$$

by Lemma 16. Hence, by Lemma B.1 from Guntuboyina et al. (2017b) and (16),

$$T_2 \leq C_k m(t,n)\left\{\frac{\sqrt{n}V^*}{m(t,n)}\right\}^{1/(2+2k)} + C_k m(t,n)\{\log(en)\}^{1/2}.$$

To bound T_1, let $v_1, \ldots, v_{k+1}$ an orthonormal basis of $\mathcal{R}^\perp$, where $\|v_j\|_\infty \leq (2/n)^{1/2}$, for $j = 1, \ldots, k+$

1, as in the proof of Lemma 13. Then, for any $\delta \in R^n$ with $D^2(\delta) \leq t^2$,

$$\begin{aligned}
\xi^\top P_{\mathcal{R}^\perp}\delta &\leq \left|\sum_{j=1}^{k+1} \delta^\top v_j \cdot \xi^\top v_j\right| \\
&\leq \sum_{j=1}^{k+1} \left|\delta^\top v_j\right| \cdot \left|\xi^\top v_j\right| \\
&\leq (k+1)\left(\max_{j=1,\ldots,k+1} |\xi^\top v_j|\right)\left(\max_{j=1,\ldots,k+1} |\delta^\top v_j|\right) \\
&\leq (k+1)\left\{\max_{j=1,\ldots,k+1} |\xi^\top v_j|\right\}\left\{\left(\frac{2}{n}\right)^{1/2} t^2 + t,\right\},
\end{aligned} \tag{24}$$

where the last inequality follows as (22). Therefore,

$$T_1 \leq (k+1)\left\{\left(\frac{2}{n}\right)^{1/2} t^2 + t,\right\} \sum_{j=1}^{k+1} \mathbb{E}\left(|\xi^\top v_j|\right) \leq C(k+1)^2 \left\{\left(\frac{2}{n}\right)^{1/2} t^2 + t\right\}, \tag{25}$$

for some positive constant $C > 0$, and where the last inequality follows since $\xi^\top v_j$ are sub-Gaussian random variables with variance 1. The conclusion follows combining (25), (23) and Lemma B.1 from Guntuboyina et al. (2017a). □

B.2 Proof of Theorem 1

Proof. We start by defining $\tilde{D}^2(\delta)$, for $\delta \in R^n$, as

$$\tilde{D}^2(\delta) = \sum_{i=1}^{n} \min\left\{\frac{|\delta_i|}{L}, \frac{|\delta_i|^2}{L^2}\right\}. \tag{26}$$

Then clearly,

$$\min\{L, L^2\}\tilde{D}^2(\delta) \leq D^2(\delta) \leq \max\{L, L^2\}\tilde{D}^2(\delta). \tag{27}$$

Now let $\hat{\delta} = \hat{\theta} - \theta^*$. Suppose that

$$\frac{1}{n} D^2(\hat{\delta}) > \frac{\eta^2}{n} := c_1 n^{-(2k+2)/(2k+3)},$$

for some constant $c_1 > 0$. Then,

$$0 > \inf_{\delta \in \mathcal{F}(\eta)} \hat{M}_n(\theta^* + \delta), \tag{28}$$

where

$$\mathcal{F}(\eta) = \left\{\delta \,:\, \|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k},\, \max\{L, L^2\}\tilde{D}^2(\delta) \geq \eta^2\right\}.$$

Next, let $r^2 = \max\{L, L^2\}\tilde{D}^2(\hat{\delta})$. Then define $g : [0,1] \to R$ as $g(t) = \max\{L, L^2\}\tilde{D}^2(t\hat{\delta})$. Clearly, g is a continuous function with $g(0) = 0$, and $g(1) = r^2$. Therefore, there exists $t_{\hat{\delta}} \in [0,1]$ such that $g(t_{\hat{\delta}}) = \eta^2$. Hence, by the convexity of $\hat{M}_n$ and by (28), we obtain that

$$0 > \inf_{\delta \in \tilde{\mathcal{F}}(\eta)} \hat{M}_n(\theta^* + \delta), \tag{29}$$

where

$$\tilde{\mathcal{F}}(\eta) = \left\{ \delta \,:\, \|\Delta^{(k+1)}\delta\|_1 \leq \frac{2V^*}{n^k},\, \max\{L, L^2\}\tilde{D}^2(\delta) = \eta^2 \right\}.$$

Therefore,

$$\begin{aligned} 0 &> \inf_{\delta \in \tilde{\mathcal{F}}(\eta)} \{M_n(\theta^* + \delta) - M_n(\theta^*)\} - \sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\} \\ &\geq \tfrac{\min\{L, L^2\}}{\max\{L, L^2\}} \eta^2 - \sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\}, \end{aligned} \tag{30}$$

where the second inequality follows by Lemma 8. However, for $\epsilon > 0$, we define

$$\gamma := \epsilon^{-1} \left[C(k+1)^2 \left\{ \left(\frac{2}{n}\right)^{1/2} \eta^2 + \eta, \right\} + C_k m(\eta, n) \left\{ \frac{n^{1/2} V^*}{m(\eta, n)} \right\}^{1/(2+2k)} + C_k m(\eta, n) \{\log(en)\}^{1/2} \right],$$

and notice that

$$\begin{aligned} \mathrm{pr}\left[\sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\} \geq \gamma \right] &\leq \gamma^{-1} E \left[\sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\} \right] \\ &\leq \gamma^{-1} E \left\{ \sup_{v \in K\,:\, D^2(v - \theta^*) \leq \eta^2} (M_n - \hat{M}_n)(v) \right\} \\ &\leq \gamma^{-1} \Big[C(k+1)^2 \left[\left(\tfrac{2}{n}\right)^{1/2} \eta^2 + \eta \right] + \\ &\quad C_k m(\eta, n) \left\{ \tfrac{n^{1/2} V^*}{m(\eta, n)} \right\}^{1/(2+2k)} + \\ &\quad C_k m(\eta, n) \{\log(en)\}^{1/2} \Big], \\ &\leq \epsilon, \end{aligned}$$

where the first inequality holds by Markov's inequality, the second by (27), the third from Lemma 17 proceeding as in the proof of Lemmas 10–11, and the last due to our choice of γ.

Therefore, with probability at least $1 - \epsilon$, if $D^2(\hat{\delta}) \geq \eta^2$, then

$$\begin{aligned} 0 > \tfrac{\min\{L, L^2\}}{\max\{L, L^2\}} \eta^2 - \epsilon^{-1} \Big[& C(k+1)^2 \left\{ \left(\tfrac{2}{n}\right)^{1/2} \eta^2 + \eta \right\} + C_k m(\eta, n) \left\{ \tfrac{n^{1/2} V^*}{m(\eta, n)} \right\}^{1/(2+2k)} \\ & + C_k m(\eta, n) \{\log(en)\}^{1/2} \Big], \end{aligned}$$

but the right hand side is positive for a large enough value of c_1 in (26). This proves the claim. □

C Theorem 2

C.1 Auxiliary lemmas for Theorem 2

Throughout we write

$$K = \left\{ \theta \in R^n \,:\, \|\Delta^{(k+1)}\theta\|_1 \leq \frac{V^*}{n^k} \right\}.$$

Lemma 18. *Let $\delta = a(v - \theta^*)$, where $v \in K$ and $a \geq 0$. Then*

$$\|P_{\mathcal{R}}\delta\|_\infty \leq a\tilde{C}_k V^*,$$

where $\tilde{C}_k$ is the same constant from Lemma 15.

Proof. This follows from Lemma 15 since $\delta = a(v - \theta^*)$. □

Lemma 19. *Let $\delta \in R^n$, with $\delta = a(v - \theta^*)$ for some $v \in K$ and $a \in [0, 1]$. Suppose that $D^2(\delta) \leq t^2$, then*

$$\|P_{\mathcal{R}}\delta\| \leq \max\{V^* t, (V^*)^{1/2} t\} c(k, L),$$

for some constant $c(k, L) > 0$.

Proof. We proceed in two cases. If $a < t/n^{1/2}$, then by Lemma 18,

$$\|P_{\mathcal{R}}\delta\| \leq n^{1/2}\|P_R\delta\|_\infty \leq n^{1/2} a\tilde{C}_k V^* \leq \tilde{C}_k V^* t. \tag{31}$$

If $t/n^{1/2} \leq a \leq 1$, then $\min\{a, a^2\} = a^2$, $\min\{a^{1/2}, a\} = a$, and

$$D^2(v - \theta^*) \leq \frac{\max\{L, L^2\}}{\min\{a, a^2\}} D^2(\delta) \leq \frac{\max\{L, L^2\} t^2}{a^2} =: \frac{(t')^2}{a^2}.$$

Therefore, by Lemma 16,

$$\begin{aligned}
\|P_{\mathcal{R}}\delta\| &= a\|P_{\mathcal{R}}(v - \theta^*)\| \\
&\leq a \max\left\{(\tilde{C}_k V^*)^{1/2}, 1\right\} \max\left\{(2L)^{1/2}, \tfrac{2}{(L)^{1/2}}\right\} \left(\left[1 + \left\{2\gamma\left(\tfrac{t'}{a}, n\right)\right\}^{1/2} + 4(k+1)\right] \tfrac{t'}{a} + \right. \\
&\quad \left. \tfrac{4(k+1)(t')^2}{a^2\sqrt{n}} \right) \\
&= \max\left\{(\tilde{C}_k V^*)^{1/2}, 1\right\} \max\left\{(2L)^{1/2}, \tfrac{2}{(L)^{1/2}}\right\} \left(\left[1 + \left\{2\gamma\left(\tfrac{t'}{a}, n\right)\right\}^{1/2} + 4(k+1)\right] t + \tfrac{4(k+1)(t')^2}{a\sqrt{n}} \right),
\end{aligned} \tag{32}$$

and as in Lemma 13,

$$\gamma\left(\frac{t'}{a}, n\right) = 2(k+1)\left\{\frac{t'}{an^{1/2}} + \frac{(t')^2}{an^2}\right\} \leq 4(k+1)\max\{L, L^4\}. \tag{33}$$

The claim follows combining (31)–(33). □

Lemma 20. *Let*

$$\tilde{K}_t = \left\{\delta \in R^n \,:\, \delta = a(v - \theta^*),\, a \in [0, 1],\, v \in K,\, D^2(\delta) \leq t^2\right\}, \tag{34}$$

and the tangent cone of K defined as

$$T_K(\theta^*) = \textit{Closure}\left\{\delta \in R^n \,:\, \delta = a(v - \theta^*),\; v \in K,\; a \geq 0\right\}.$$

Then

$$\begin{aligned}
E\left\{\sup_{v \in \tilde{K}_t} (M_n - \hat{M}_n)(v + \theta^*)\right\} \;\leq\; & 2t\max\{V^*, (V^*)^{1/2}\}\, c(k, L)\, E\left\{\sup_{\delta \in T_K(\theta^*),\, \|\delta\| \leq 1} \xi^\top \delta\right\} + \\
& C(k+1)^2\left\{\left(\tfrac{2}{n}\right)^{1/2} t^2 + t\right\},
\end{aligned}$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher variables independent of $\{y_i\}_{i=1}^n$, and $C > 0$ is a positive constant.

Proof. Notice that

$$\begin{aligned} E\left\{\sup_{v\in\tilde{K}_t}(M_n-\hat{M}_n)(v+\theta^*)\right\} &\leq 2E\left(\sup_{v\in\tilde{K}_t}\xi^\top v\right) \\ &\leq 2\,E\left(\sup_{\delta\in\tilde{K}_t}\xi^\top P_{\mathcal{R}^\perp}\delta\right) \\ &\quad +2\,E\left(\sup_{\delta\in\tilde{K}_t}\xi^\top P_{\mathcal{R}}\delta\right) \\ &=: T_1+T_2. \end{aligned}$$

where the first inequality follows proceeding as in the proof of Lemmas 10–11. We now proceed to bound T_1 and T_2. To bound T_2, we observe that

$$T_2 \leq 2\,E\left\{\sup_{\delta\in T_K(\theta^*),\ \|\delta\|\leq\max\{V^*,(V^*)^{1/2}\}tc(k,L)}\xi^\top\delta\right\} = 2\max\{V^*,(V^*)^{1/2}\}tc(k)\,E\left\{\sup_{\delta\in T_K(\theta^*),\ \|\delta\|\leq 1}\xi^\top\delta\right\}, \tag{35}$$

where the first inequality holds by Lemma 19. An upper bound on T_1 follows as in (25). This completes the proof. □

C.2 Proof of Theorem 2

Proof. First, by appendix B.2 in Guntuboyina et al. (2017a) and (16), we have that

$$E\left\{\sup_{\delta\in T_K(\theta^*),\ \|\delta\|\leq 1}\xi^\top\delta\right\} \leq c_{k+1}\left\{(s+1)\log\left(\frac{en}{s+1}\right)\right\}^{1/2},$$

for some positive constant c_{k+1}. Hence, by Lemma 20, we have

$$\begin{aligned} E\left\{\sup_{v\in\tilde{K}_t}(M_n-\hat{M}_n)(v+\theta^*)\right\} &\leq 2\max\{V^*,(V^*)^{1/2}\}\,tc(k,L)\,c_{k+1}\left\{(s+1)\log\left(\tfrac{en}{s+1}\right)\right\}^{1/2} \\ &\quad +C(k+1)^2\left\{\left(\tfrac{2}{n}\right)^{1/2}t^2+t\right\}, \end{aligned} \tag{36}$$

where $\tilde{K}_t$ was defined in (34). The conclusion follows by proceeding as in the proof of Theorem 1, and exploiting (36). □

D Theorem 3

D.1 Auxiliary lemmas for Proof of Theorem 3

Throughout we assume that Assumptions 1–2 hold with

$$K=\left\{\theta\,:\,\|\Delta^{(k+1)}\theta\|_1\leq\frac{V^*}{n^k}\right\}. \tag{37}$$

Also, for $\delta\in R^n$ we write $D(\delta)=\{D^2(\delta)\}^{1/2}$ with $D^2(\cdot)$ defined as in Lemma 8. Furthermore, we use the notation M_n and $\hat{M}_n$ from Definition 1.

Lemma 21. *There exists a constant A satisfying*

$$A \asymp n^{(2k+1)/(2k+3)} (\log n)^{1/(2k+3)} \|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}$$

such that

$$\sup_{x \in \mathrm{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \le n^{-k}} \frac{a^\top x - An^{-k}}{D(x)} = O_{\mathrm{pr}}\left(\tilde{B}\right),$$

where $a = (a_1, \ldots, a_n)$ is a vector with independent coordinates satisfying

$$\mathrm{pr}(a_i = \tau) = 1 - \tau, \quad \mathrm{pr}(a_i = \tau - 1) = \tau, \quad \textit{for } i = 1, \ldots, n, \tag{38}$$

and $\tilde{B}$ satisfies

$$\tilde{B} \asymp n^{(2k+1)/(4k+6)} (\log n)^{1/(4k+6)} \|\Delta^{(k+1)}\theta^*\|_1^{2/(4k+6)}.$$

Proof. Let x be such that $x \in \mathrm{row}\{\Delta^{(k+1)}\}$ and $\|\Delta^{(k+1)}x\|_1 \le n^{-k}$, then by Lemma 15 there exists a constant $\tilde{C} > 0$ independent of x such that

$$\|x\|_\infty \le \tilde{C}.$$

Then

$$\begin{aligned} D(x) &= \{c_\tau(\textstyle\sum_{i:|x_i|>L} |x_i| + \sum_{i:|x_i|\le L} |x_i|^2)\}^{1/2} \\ &\ge \left(\tfrac{c_\tau}{\tilde{C}}\right)^{1/2} \|x\|. \end{aligned}$$

Hence,

$$\begin{aligned} \sup_{x \in \mathrm{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \le n^{-k}} \frac{a^\top x - An^{-k}}{D(x)} &\le \left(\frac{\tilde{C}}{c_\tau}\right)^{1/2} \sup_{x \in \mathrm{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \le n^{-k}} \frac{a^\top x - An^{-k}}{\|x\|} \\ &= \left(\frac{\tilde{C}}{c_\tau}\right)^{1/2} \sup_{x \in \mathrm{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \le n^{-k}} \frac{a^\top (n^k x) - A}{\|n^k x\|} \\ &= \left(\frac{\tilde{C}}{c_\tau}\right)^{1/2} \sup_{x \in \mathrm{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \le 1} \frac{a^\top x - A}{\|x\|}, \end{aligned}$$

and the claim follows by Corollary 7 in Wang et al. (2016). □

Lemma 22. *Let $a = (a_1, \ldots, a_n)$ is a vector with independent coordinates satisfying (38). Let $\mathcal{R} = \mathrm{row}\{\Delta^{(k+1)}\}$ and $\mathcal{R}^\perp$ its orthogonal complement. Then*

$$\sup_{x \in R^n} \frac{a^\top P_{\mathcal{R}^\perp} x}{\frac{D^2(x)}{n} + D(x)} = O_{\mathrm{pr}}(1),$$

where $P_{\mathcal{R}^\perp}$ denotes the orthogonal projection onto $\mathcal{R}^\perp$.

Proof. This follows immediately as (24). □

Lemma 23. *Suppose that $\|\Delta^{(k+1)}\theta^*\|_1 = O(n^{-k})$. Let $\epsilon \in (0, 1)$ then there exists a choice*

$$\lambda = \Theta\left\{n^{(2k+1)/(2k+3)} (\log n)^{1/(2k+3)} \|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}\right\}$$

such that for a constant $C_0 > 0$, we have that, with probability at least $1 - \epsilon/4$,

$$\kappa(\hat{\theta}-\theta^*) \in \mathcal{A} := \left\{\delta : \|\Delta^{(k+1)}\delta\| \leq C_0 \max\left\{\frac{1}{n^k}, \gamma_1 D^2(\delta), \|\Delta^{(k+1)}\delta\|_1 + A^{-1}(a^*)^\top P_{R^\perp}(\tilde{\theta}-\theta^*)\right\}\right\},$$

for all $\kappa \in [0,1]$, where

$$\gamma_1 := \left(\frac{\max\{L,\sqrt{L}\}}{\min\{L,\sqrt{L}\}}\right)^2 \frac{B^2 n^k}{A^2},$$

$B \asymp \tilde{B}$ (B dependents on ϵ), with A and $\tilde{B}$ defined as in Lemma 21, and with $a_i^* = \tau - 1\{y_i \leq \theta_i^*\}$ for $i = 1, \ldots, n$.

Proof. Let B such that $B \asymp \tilde{B}$ where $\tilde{B}$ is as in Lemma 21, and such that

$$\sup_{x \in \text{row}\{\Delta^{(k+1)}\} : \|\Delta^{(k+1)}x\|_1 \leq n^{-k}} \frac{a^\top x - An^{-k}}{D(x)} \leq B, \tag{39}$$

happens with probability at least $1 - \epsilon/4$. From here on, we suppose that (39) holds.

Now pick $\kappa \in [0,1]$ be fixed, and let $\tilde{\delta} = \kappa(\hat{\theta} - \theta^*)$. Then by the optimality of $\hat{\theta}$ and the convexity of (2) we have that

$$\sum_{i=1}^n \rho_\tau(y_i - \tilde{\theta}_i) + \lambda\|\Delta^{(k+1)}\tilde{\theta}\|_1 \leq \sum_{i=1}^n \rho_\tau(y_i - \theta_i^*) + \lambda\|\Delta^{(k+1)}\theta^*\|_1,$$

where $\tilde{\theta} = \theta^* + \tilde{\delta}$. Then as in the proof of Lemma 3 from Belloni et al. (2011),

$$0 \leq \lambda\left[\|\Delta^{(k+1)}\theta^*\|_1 - \|\Delta^{(k+1)}\tilde{\theta}\|_1\right] + (\tilde{\theta}-\theta^*)^\top a^*. \tag{40}$$

Next, suppose that $\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 > n^{-k}$. Hence, by (39)

$$\begin{aligned}
(\tilde{\theta}-\theta^*)^\top a^* &= (a^*)^\top P_{\mathcal{R}}(\tilde{\theta}-\theta^*) + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \\
&= \left\{(a^*)^\top P_{\mathcal{R}} \tfrac{(\tilde{\theta}-\theta^*)}{n^k\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1}\right\} n^k\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \\
&\leq \left\{BD\left(\tfrac{\tilde{\theta}-\theta^*}{n^k\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1}\right) + An^{-k}\right\} n^k\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \\
&\leq \tfrac{\max\{L,\sqrt{L}\}}{\min\{L,\sqrt{L}\}} Bn^{k/2}\{\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1\}^{1/2} D\left(\tilde{\theta}-\theta^*\right) + A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 + \\
&\quad (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*)
\end{aligned} \tag{41}$$

where the second inequality holds by the definition of $D(\cdot)$ and by (27).

If

$$A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 < \frac{\max\{L,L^{1/2}\}}{\min\{L,L^{1/2}\}} Bn^{k/2}\{\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1\}^{1/2} D\left(\tilde{\theta}-\theta^*\right).$$

then

$$\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 \leq \left(\frac{\max\{L,L^{1/2}\}}{\min\{L,L^{1/2}\}}\right)^2 \frac{B^2 n^k D^2(\tilde{\theta}-\theta^*)}{A^2} \tag{42}$$

If

$$A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 \geq \frac{\max\{L,L^{1/2}\}}{\min\{L,L^{1/2}\}} Bn^{k/2}\{\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1\}^{1/2} D\left(\tilde{\theta}-\theta^*\right),$$

then

$$(\tilde{\theta}-\theta^*)^\top a^* \quad \leq \quad 2A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \tag{43}$$

Hence, choosing $\lambda = 3A$, and combining (40) with (43),

$$\begin{aligned} A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 &\leq \lambda\left\{\|\Delta^{(k+1)}\theta^*\|_1 - \|\Delta^{(k+1)}\tilde{\theta}\|_1\right\} + \\ &\quad 3A\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \\ &\leq 6A\|\Delta^{(k+1)}\theta^*\|_1 + (a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*) \end{aligned}$$

,

with the second inequality follows by the triangle inequality. Therefore,

$$\|\Delta^{(k+1)}(\tilde{\theta}-\theta^*)\|_1 \leq \max\left\{\frac{1}{n^k}, 6\|\Delta^{(k+1)}\theta^*\|_1 + A^{-1}(a^*)^\top P_{\mathcal{R}^\perp}(\tilde{\theta}-\theta^*), \left(\frac{\max\{L,L^{1/2}\}}{\min\{L,L^{1/2}\}}\right)^2 \frac{B^2 n^k D^2(\tilde{\theta}-\theta^*)}{A^2}\right\}.$$

□

D.2 Proof of Theorem 3

Proof. Let $\epsilon \in (0,1)$ and suppose that the following events hold

$$\begin{aligned} \Omega_1 &= \left\{\kappa(\hat{\theta}-\theta^*) \in \mathcal{A}, \quad \forall \kappa \in [0,1]\right\}, \\ \Omega_2 &= \left\{\sup_{x\in R^n} \frac{(a^*)^\top P_{\mathcal{R}^\perp} x}{\frac{D^2(x)}{n} + D(x)} \leq E\right\}, \end{aligned} \tag{44}$$

which by Lemmas 22–23 happen with probability at least $1-\epsilon/2$ for some constant E, and with $\mathcal{A}$ as in Lemma 23. Furthermore, we denote by C_1 a positive constant such that $\|\Delta^{(k+1)}\theta^*\|_1 \leq C_1 n^{-k}$. This constant exists since $\theta^* \in K$.

Then suppose that

$$\frac{1}{n}D^2(\hat{\delta}) > \frac{\eta^2}{n}, \quad \text{with} \quad \eta = c_0 n^{1/(4k+6)} (\log n)^{1/(4k+6)}, \tag{45}$$

for some large enough constant $c_0 > 1$ such that

$$A = c_0 n^{(2k+1)/(2k+3)} (\log n)^{1/(2k+3)} \|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}$$

in Lemma 21, and $\lambda = 3A$.

Next, notice that by (45),

$$0 > \inf_{\delta\in\mathcal{F}(\eta)} \left[\hat{M}_n(\theta^*+\delta) + \lambda\left\{\|\Delta^{(k+1)}(\theta^*+\delta)\|_1 - \|\Delta^{(k+1)}\theta^*\|_1\right\}\right], \tag{46}$$

where

$$\mathcal{F}(\eta) = \left\{\delta \in \mathcal{A} \,:\, \max\{L,L^2\}\tilde{D}^2(\delta) \geq \eta^2\right\}.$$

Then by the convexity of the function $\hat{M}_n(\theta^*+\cdot) + \lambda\|\Delta^{(k+1)}(\theta^*+\cdot)\|_1$, Lemma 23, and the same argument in the proof of Theorem 1, we obtain that

$$0 > \inf_{\delta\in\mathcal{G}(\eta)} \left[\hat{M}_n(\theta^*+\delta) + \lambda\left\{\|\Delta^{(k+1)}(\theta^*+\delta)\|_1 - \|\Delta^{(k+1)}\theta^*\|_1\right\}\right], \tag{47}$$

where

$$\mathcal{G}(\eta) = \left\{\delta \in \mathcal{A} \,:\, \max\{L, L^2\}\tilde{D}^2(\delta) = \eta^2\right\}.$$

Therefore,

$$\begin{aligned} 0 \;>\; & \inf_{\delta\in\mathcal{G}(\eta)} M_n(\theta^*+\delta) - \sup_{\delta\in\mathcal{G}(\eta)} \left\{M_n(\theta^*+\delta) - \hat{M}_n(\theta^*+\delta)\right\} \\ & - \sup_{\delta\in\mathcal{G}(\eta)} \lambda\left|\|\Delta^{(k+1)}(\theta^*+\delta)\|_1 - \|\Delta^{(k+1)}\theta^*\|_1\right|, \\ \geq\; & \tfrac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2 - \sup_{\delta\in\mathcal{G}(\eta)} \left\{M_n(\theta^*+\delta) - \hat{M}_n(\theta^*+\delta)\right\} \\ & - \sup_{\delta\in\mathcal{G}(\eta)} \lambda\left|\|\Delta^{(k+1)}(\theta^*+\delta)\|_1 - \|\Delta^{(k+1)}\theta^*\|_1\right|, \end{aligned} \tag{48}$$

where the second inequality follows by Lemma 8 and by (26).

Next, define

$$\mathcal{H}(\eta) = \{\delta \in \mathcal{A} \,:\, D(\delta) \leq \eta\}$$

and notice that for $\gamma \asymp n^{1/(2k+3)}$, to be chosen later, we have that

$$\begin{aligned} \mathrm{pr}\left[\sup_{\delta\in\mathcal{H}(\eta)} \left\{M_n(\theta^*+\delta) - \hat{M}_n(\theta^*+\delta)\right\} \geq \gamma\right] \;&\leq\; \gamma^{-1}\mathrm{pr}\left[\sup_{\delta\in\mathcal{H}(\eta)} \left\{M_n(\theta^*+\delta) - \hat{M}_n(\theta^*+\delta)\right\}\right] \\ &\leq\; 4\gamma^{-1}E\left(\sup_{\delta\in\mathcal{H}(\eta)} \sum_{i=1}^n \xi_i\delta_i\right), \end{aligned} \tag{49}$$

where the first inequality holds by Markov's inequality, the second as in the proof of Lemmas 10–11.

However, notice that $\delta \in \mathcal{H}(\eta)$ implies that for some constants $a_1, \tilde{C} > 0$ independent of c_0,

$$\begin{aligned} \|\Delta^{(k+1)}\delta\|_1 \;&\leq\; C_0 \max\left\{\tfrac{1}{n^k}, \gamma_1 D^2(\delta), \|\Delta^{(k+1)}\theta^*\|_1 + A^{-1}(a^*)^\top P_{R^\perp}(\tilde{\theta}-\theta^*)\right\} \\ &\leq\; C_0 \max\left\{\tfrac{1}{n^k}, \gamma_1 D^2(\delta), \|\Delta^{(k+1)}\theta^*\|_1 + A^{-1}\left(\tfrac{D^2(\delta)}{n} + D(\delta)\right)\right\} \\ &\leq\; C_0 \max\left\{\tfrac{1}{n^k}, \left(\tfrac{\max\{L,\sqrt{L}\}}{\min\{L,\sqrt{L}\}}\right)^2 \tfrac{B^2 n^k}{A^2}\eta^2, \tfrac{C_1}{n^k} + A^{-1}\left(\tfrac{\eta^2}{n} + \eta\right)\right\} \\ &\leq\; a_1\Bigg[\max\Bigg\{\tfrac{1}{n^k}, \frac{n^{(2k+1)/(2k+3)}(\log n)^{1/(2k+3)}\|\Delta^{(k+1)}\theta^*\|_1^{2/(2k+3)} n^k n^{1/(2k+3)}(\log n)^{1/(2k+3)}}{n^{(4k+2)/(2k+3)}(\log n)^{2/(2k+3)}\|\Delta^{(k+1)}\theta^*\|_1^{-(4k+2)/(2k+3)}}, \\ &\qquad \tfrac{C_1}{n^k} + \frac{c_0 n^{1/(2k+3)}(\log n)^{1/(2k+3)}}{n^{1+(2k+1)/(2k+3)}(\log n)^{1/(2k+3)}\|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}} + \\ &\qquad \frac{n^{1/(4k+6)}(\log n)^{1/(4k+6)}}{n^{(2k+1)/(2k+3)}(\log n)^{1/(2k+3)}\|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}}\Bigg\}\Bigg] \\ &\leq\; \tilde{C}n^{-k}, \end{aligned} \tag{50}$$

where the first inequality follows from the definition of Ω_1, the second because we are assuming that Ω_2 holds, the third by the definition of $\mathcal{H}(\eta)$, the fourth by definition of A and B, and the last by (45).

Therefore, defining

$$\mathcal{L}(\eta) \;=\; \left\{\delta \,:\, \|\Delta^{(k+1)}\delta\|_1 \leq \tilde{C}n^{-k}, \quad \text{and} \quad D(\delta) \leq \eta\right\}.$$

we have that for some constant $a_2 > 0$ independent of c_0,

$$\begin{aligned}
\text{pr}\left[\sup_{\delta\in\mathcal{H}(\eta)}\left\{M_n(\theta^*+\delta)-\hat{M}_n(\theta^*+\delta)\right\}\geq\gamma\right] &\leq 4\gamma^{-1}E\left(\sup_{\delta\in\mathcal{L}(\eta)}\sum_{i=1}^{n}\xi_i\delta_i\right)\\
&\leq 4\gamma^{-1}\left[C(k+1)^2\left\{\left(\frac{2}{n}\right)^{1/2}\eta^2+\eta\right\}+C_k+\right.\\
&\quad C_k m(\eta,n)\left\{\frac{n^{1/2}\tilde{C}n^{-k}}{m(\eta,n)}\right\}^{1/(2+2k)}+\\
&\quad \left.C_k m(\eta,n)\{\log(en)\}^{1/2}\right]\\
&\leq a_2\gamma^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)}\\
&= \frac{\epsilon}{2},
\end{aligned} \tag{51}$$

where the first inequality follows by (49) and (50), the second by Lemma 17, the third by definition of η, and the last by choosing $\gamma = 2a_2\epsilon^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)}$. Hence, if (45) holds , with probability at least $1-\epsilon$, by (48) and (51) we have that

$$\begin{aligned}
0 &> \frac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2 - a_2\gamma^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)} - \sup_{\delta\in\mathcal{G}(\eta)}\lambda\left|\|\Delta^{(k+1)}(\theta^*+\delta)\|_1-\|\Delta^{(k+1)}\theta^*\|_1\right|,\\
&\geq \frac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2 - a_2\gamma^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)} -\\
&\quad \sup_{\delta\in\mathcal{G}(\eta)}\lambda\|\Delta^{(k+1)}\delta\|_1\\
&\geq \frac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2 - a_2\gamma^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)} -\\
&\quad 3c_0n^{(2k+1)/(2k+3)}(\log n)^{1/(2k+3)}\|\Delta^{(k+1)}\theta^*\|_1^{-(2k+1)/(2k+3)}\left[\tilde{C}n^{-k}\right]\\
&\geq \frac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2 - a_2\gamma^{-1}c_0^{1-1/(2k+2)}n^{1/(2k+3)} -\\
&\quad 3c_0n^{(2k+1)/(2k+3)}(\log n)^{1/(2k+3)}n^{(2k^2+k)/(2k+3)}\left[\tilde{C}n^{-k}\right]\\
&> 0
\end{aligned}$$

for large enough choice of c_0 in (45). It follows that for large enough c_0 we have that

$$\frac{1}{n}D^2(\hat{\delta}) \leq \frac{\eta^2}{n}$$

with probability at least $1-\epsilon$. □

E Proof Theorem 5

Proof. Throughout we use the notation from Definition 1 and Lemma 8. Then let $\hat{\delta} = \hat{\theta} - \theta^*$ and suppose that

$$\frac{1}{n}D^2(\hat{\delta}) > \frac{\eta^2}{n} := \frac{c_1(\log n)^2}{n^{1/2}},$$

for some constant $c_1 > 0$. Then,

$$0 > \inf_{\delta\in\mathcal{F}(\eta)}\hat{M}_n(\theta^*+\delta), \tag{52}$$

where

$$\mathcal{F}(\eta) = \left\{\delta\,:\,\|\nabla\delta\|_1\leq 2V,\ \max\{L,L^2\}\tilde{D}^2(\delta)\geq\eta^2\right\}.$$

Then, as in the proof of Theorem 1,

$$0 > \inf_{\delta \in \tilde{\mathcal{F}}(\eta)} \hat{M}_n(\theta^* + \delta), \tag{53}$$

where

$$\tilde{\mathcal{F}}(\eta) = \left\{ \delta \,:\, \|\nabla \delta\|_1 \leq 2V,\, \max\{L, L^2\} \tilde{D}^2(\delta) = \eta^2 \right\}.$$

Hence,

$$\begin{aligned} 0 &> \inf_{\delta \in \tilde{\mathcal{F}}(\eta)} \{M_n(\theta^* + \delta) - M_n(\theta^*)\} - \sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\} \\ &\geq \tfrac{\min\{L, L^2\}}{\max\{L, L^2\}} \eta^2 - \sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\}. \end{aligned} \tag{54}$$

Now, for $\gamma > 0$ we have that

$$\begin{aligned} \mathrm{pr}\left\{ \sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left[M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right] \geq \gamma \right\} &\leq \gamma^{-1} E \left[\sup_{\delta \in \tilde{\mathcal{F}}(\eta)} \left\{ M_n(\theta^* + \delta) - \hat{M}_n(\theta^* + \delta) \right\} \right] \\ &\leq \gamma^{-1} E \left\{ \sup_{\delta \,:\, D^2(\delta) \leq \eta^2,\, \|\nabla \delta\|_1 \leq 2V} (M_n - \hat{M}_n)(\delta + \theta^*) \right\} \end{aligned} \tag{55}$$

where the first inequality holds by Markov's inequality, and the second by (27).

Next, we have that for independent Rademacher random variables $\xi_1, \ldots, \xi_n$ independent of y it holds that

$$\begin{aligned} E \left\{ \sup_{\delta \,:\, D^2(\delta) \leq \eta^2,\, \|\nabla \delta\|_1 \leq 2V} (M_n - \hat{M}_n)(\delta + \theta^*) \right\} &\leq 4E \left(\sup_{\delta \,:\, D^2(\delta) \leq \eta^2,\, \|\nabla \delta\|_1 \leq 2V} \sum_{i=1}^n \xi_i \delta_i \right) \\ &\leq 4E \left(\sup_{\delta \,:\, D^2(\delta) \leq \eta^2,\, \|\nabla \delta\|_1 \leq 2V,\, \bar{\delta} = 0} \sum_{i=1}^n \xi_i \delta_i \right) + \\ &\quad 4E \left(\sup_{\delta \,:\, D^2(\delta) \leq \eta^2,\, \delta \in \mathrm{Span}\{\mathbf{1}\}} \sum_{i=1}^n \xi_i \delta_i \right) \\ &\leq 4E \left(\sup_{\delta \,:\, \|\nabla \delta\|_1 \leq 2V,\, \bar{\delta} = 0} \sum_{i=1}^n \xi_i \delta_i \right) + \\ &\quad 4E \left(\sup_{u \,:\, D^2(u\mathbf{1}) \leq \eta^2,} u \sum_{i=1}^n \xi_i \right) \\ &\leq 4C \left\{ V \log n \log\left(1 + 2Vn^2\right) + 1 \right\} + \\ &\quad 4E \left(\sup_{u \,:\, D^2(u\mathbf{1}) \leq \eta^2,\, |u| \leq L} u \sum_{i=1}^n \xi_i \right) + \\ &\quad 4E \left(\sup_{u \,:\, D^2(u\mathbf{1}) \leq \eta^2,\, |u| > L} u \sum_{i=1}^n \xi_i \right), \end{aligned} \tag{56}$$

where $C > 0$ is a constant, and the last inequality follows from Lemma 4.5 in Chatterjee and Goswami

(2019) and (38). Therefore,

$$
\begin{aligned}
E\left\{\sup_{\delta\,:\,D^2(\delta)\le\eta^2,\,\|D\delta\|_1\le 2V}(M_n-\hat{M}_n)(\delta+\theta^*)\right\} &\le 4C\left\{V\log n\log\left(1+2Vn^2\right)+1\right\}+\\
&\quad 4\left(1+\frac{\eta^2}{n}\right)\mathbb{E}\left(\left|\sum_{i=1}^n\xi_i\right|\right), \qquad (57)\\
&\le 4C\left\{V\log n\log\left(1+2Vn^2\right)+1\right\}+\\
&\quad 4C_1\left(L+\frac{\eta^2}{n}\right)n^{1/2},
\end{aligned}
$$

where the first inequality follows from (56) and definition of $D(\cdot)$, and where the second inequality holds by the basic properties of symmetric random walks.

Therefore, if

$$
\gamma := \epsilon^{-1}\left[4C\left\{V\log n\log\left(1+2Vn^2\right)+1\right\}+4C_1\left(L+\frac{\eta^2}{n}\right)n^{1/2}\right],
$$

then with probability at least $1-\epsilon$, $D^2(\hat{\delta})\ge\eta^2$ implies

$$
0>\frac{\min\{L,L^2\}}{\max\{L,L^2\}}\eta^2-\epsilon^{-1}\left[4C\left\{V\log n\log\left(1+2Vn^2\right)+1\right\}+4C_1\left(L+\frac{\eta^2}{n}\right)n^{1/2}\right],
$$

but the right hand side is positive for a large enough value of c_1 in (26). This proves the claim. □

F Theorem 6

F.1 Auxiliary lemmas for proof of Theorem 6

Throughout we assume that Assumption 4–5 hold.

Definition 2. We define the empirical loss function $\hat{M}_n : R^p \to R$, as

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

where

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

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

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

We also write $K=\{\theta\in R^p : \|\theta\|_1\le s\}$.

Lemma 24. *Suppose that (13) Assumption 4 holds. Then there exists a constant c_τ such that for all $\delta\in\mathbb{R}^n$, we have*

$$
M_n(\theta^*+\delta)-M_n(\theta^*)\ge D^2(X^T\delta),
$$

where D^2 is the function defined in Lemma 8.

Proof. Notice that as in Equation B.3 of Belloni et al. (2011),

$$\begin{aligned} M_{n,i}(\theta+\delta) - M_{n,i}(\theta^*) &= \sum_{i=1}^{n} \int_0^{x_i^\top \delta} \left\{ F_{y_i}(x_i^\top \theta_i^* + z) - F_{y_i}(x_i^\top \theta_i^*) \right\} dz \\ &= \sum_{i=1}^{n} \mathbf{1}_{\{|x_i^\top \theta^*| \le L\}} \int_0^{x_i^\top \delta} \left\{ F_{y_i}(x_i^\top \theta_i^* + z) - F_{y_i}(x_i^\top \theta_i^*) \right\} dz + \\ &\quad \sum_{i=1}^{n} \mathbf{1}_{\{|x_i^\top \theta^*| > L\}} \int_0^{x_i^\top \delta} \left\{ F_{y_i}(x_i^\top \theta_i^* + z) - F_{y_i}(x_i^\top \theta_i^*) \right\} dz \qquad (58) \\ &\ge \sum_{i=1}^{n} \mathbf{1}_{\{|x_i^\top \theta^*| \le L\}} \int_0^{x_i^\top \delta} f_{y_i}(u(x_i^\top \theta_i^*, z)) z dz + \\ &\quad \sum_{i=1}^{n} \mathbf{1}_{\{|x_i^\top \theta^*| > L\}} \int_{\frac{L \operatorname{sign}(x_i^\top \delta)}{2}}^{x_i^\top \delta} \left\{ F_{y_i}(x_i^\top \theta_i^* + z) - F_{y_i}(x_i^\top \theta_i^*) \right\} dz \end{aligned}$$

where $u(x_i^\top \theta^*, z)$ is a point between $x_i^\top \theta^* + z$ and $x_i^\top \theta^*$. The claim the follows proceeding as in the proof of Lemma 8, exploiting (58). □

F.2 Proof of Theorem 6

Proof. Let $\{\xi_i\}_{i=1}^n$ be independent Rademacher random variables independent of $\{(y_i)\}_{i=1}^n$. By Lemma 24 and with the same argument that was used to prove Corollary 9, we obtain that

$$\begin{aligned} \tfrac{1}{n} E\left[D^2 \{ X^T (\hat{\theta} - \theta^*) \} \right] &\le \frac{4}{n} E\left\{ \sup_{v \in K} \sum_{i=1}^{n} \xi_i x_i^\top (v - \theta^*) \right\} \\ &\le \frac{8s}{n} E\left(\sup_{v \in K_1} \sum_{i=1}^{n} \xi_i x_i^\top v \right) \\ &= \frac{8s}{n} E\left(\sup_{v \in X K_1} \xi^\top v \right). \end{aligned}$$

where

$$K_1 = \{ v : \|v\|_1 \le 1 \}.$$

By Assumption 5 and proof of Theorem 2.14 in Rigollet and Hütter (2015), there exists a constant C such that

$$E\left(\sup_{v \in X K_1} \xi^\top v \right) \le C (n \log p)^{1/2}.$$

The claim of the theorem then follows. □

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.

Alexandre Belloni, Mingli Chen, Oscar-Hernan Madrid-Padilla, and Zixuan (Kevin) Wang. High dimensional latent panel quantile regression with an application to asset pricing. https://arxiv.org/pdf/1912.02151.pdf, 2019.

Dries F Benoit and Dirk Van den Poel. Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. Expert Systems with Applications, 36 (7):10475–10484, 2009.

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.

Brian S Cade and Barry R Noon. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment, 1(8):412–420, 2003.

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

Alex Coad and Rekha Rao. Innovation and market value: a quantile regression analysis. Economics Bulletin, 15(13), 2006.

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.

Yanqin Fan and Ruixuan Liu. A direct approach to inference in nonparametric and semiparametric quantile models. Journal of Econometrics, 191(1):196–216, 2016.

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.

Xuming He, Pin Ng, and Stephen Portnoy. Bivariate quantile smoothing splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(3):537–550, 1998.

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.

Joel L Horowitz and Sokbae Lee. Nonparametric estimation of an additive quantile regression model. Journal of the American Statistical Association, 100(472):1238–1249, 2005.

Peter J Huber. Robust estimation of a location parameter. The Annals of Statistics., page 73–101, 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.

Mi-Ok Kim et al. Quantile regression with varying coefficients. The Annals of Statistics, 35(1):92–108, 2007.

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

Charles A Knight and David D Ackerly. Variation in nuclear dna content across environmental gradients: a quantile regression analysis. Ecology Letters, 5(1):66–76, 2002.

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.

Youjuan Li, Yufeng Liu, and Ji Zhu. Quantile regression in reproducing kernel hilbert spaces. Journal of the American Statistical Association, 102(477):255–268, 2007.

Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In Advances in Neural Information Processing Systems, pages 6884–6893, 2017.

Yufeng Liu and Yichao Wu. Stepwise multiple quantile regression estimation using non-crossing constraints. Statistics and its Interface, 2(3):299–310, 2009.

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

Nicolai Meinshausen. Quantile regression forests. Journal of Machine Learning Research, 7(Jun):983–999, 2006.

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.

Petr Pata and Jaromir Schindler. Astronomical context coder for image compression. Experimental Astronomy, 39(3):495–512, 2015.

Claudia Perlich, Saharon Rosset, Richard D Lawrence, and Bianca Zadrozny. High-quantile modeling for customer wallet estimation and other applications. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 977–985, 2007.

Monica Pratesi, M Giovanna Ranalli, and Nicola Salvati. Nonparametric m-quantile regression using penalised splines. Journal of Nonparametric Statistics, 21(3):287–304, 2009.

Jeffrey S Racine and Kevin Li. Nonparametric conditional quantile estimation: A locally weighted quantile kernel approach. Journal of econometrics, 201(1):72–94, 2017.

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.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. To appear, Neural Information Processing Systems, 2016.

Alastair JR Sanderson, Trevor J Ponman, and Ewan O'Sullivan. A statistically selected chandra sample of 20 galaxy clusters–i. temperature and cooling time profiles. Monthly Notices of the Royal Astronomical Society, 372(4):1496–1508, 2006.

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

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. Journal of machine learning research, 7(Jul):1231–1264, 2006.

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

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B, 67(1):91–108, 2005.

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

Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. Annals of Statistics, 39 (3):1335–1371, 2011.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. The Annals of Statistics, 40 (2):1198–1232, 2012.

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.

Conrad Wasko and Ashish Sharma. Quantile regression for investigating scaling of extreme precipitation with temperature. Water Resources Research, 50(4):3608–3614, 2014.

Keming Yu. Smoothing regression quantile by combining k-nn estimation with local linear kernel fitting. Statistica Sinica, pages 759–774, 1999.

Chong Zhang, Yufeng Liu, and Yichao Wu. On quantile regression in reproducing kernel hilbert spaces with the data sparsity constraint. The Journal of Machine Learning Research, 17(1):1374–1418, 2016.

Weihua Zhao, Heng Lian, Hua Liang, et al. Quantile regression for the single-index coefficient model. Bernoulli, 23(3):1997–2027, 2017.